SocialCattle: IoT-based Mastitis Detection and Control through Social Cattle Behavior Sensing in Smart Farms

Yunhe Feng*, Haoran Niu*, Fanqi Wang*, Susan Ivey†, Jayne Wu*, Hairong Qi*, Raul Almeida†, Shigetoshi Eda‡, Qing Cao*

* Department of Electrical Engineering and Computer Science, University of Tennessee
Email:*{yfeng14, hniu1, fwang20}@vols.utk.edu, *{jwu10, hqi, cao}@utk.edu,
† Department of Animal Science, University of Tennessee, Email:†{ivey, ralmeida}@utk.edu,
‡ Department of Forestry, Wildlife and Fisheries, University of Tennessee, Email:‡seda@utk.edu

***Abstract*—Effective and efficient animal disease detection and control have drawn increasing attention in smart farming in recent years. It is crucial to explore how to harvest data and enable data-driven decision making for rapid diagnosis and early treatment of infectious diseases among herds. This paper proposes an IoT-based animal social behavior sensing framework to model mastitis propagation and infer mastitis infection risks among dairy cows. To monitor cow social behaviors, we deploy portable GPS devices on cows to track their movement trajectories and contacts with each other. Based on those collected location data, we build directed and weighted cattle social behavior graphs by treating cows as vertices and their contacts as edges, assigning contact frequencies between cows as edge weights, and determining edge directions according to contact spatial-temporal information. Then, we propose a flexible probabilistic disease transmission model, which considers both direct contacts with infected cows and indirect contacts via environmental contamination, to estimate and forecast mastitis infection probabilities. Our model can answer two common questions in animal disease detection and control: 1) which cows should be given the highest priorities for an investigation to determine whether there are already infected cows on the farm; 2) how to rank cows for further screening when only a tiny number of sick cows have been identified. Both theoretical and simulation-based analytics of in-the-field experiments (17 cows and more than 70-hours data) demonstrate the proposed framework's effectiveness. In addition, somatic cell count (SCC) mastitis tests validate our predictions as correct in real-world scenarios.**

***Index Terms*—IoT, social cattle behavior sensing, propagation modeling, agriculture 4.0 and smart farming**

I. Introduction

PRODUCTION of high quality milk is the most important task of modern dairy operations [1]. However, critical biosecurity challenges, such as transmissible diseases, not only affect the health of cows and sustainability of dairy farms, but also the quality of end products [2]. One type of disease that has drawn considerable attention in recent years is mastitis [3], which affects all areas of the dairy industry: from animal health, to lost milk production and lower product quality. In fact, it has been considered one of the most significant diseases of dairy herds, and has huge effects on farm economics. Cows can contact mastitis-causing bacteria through either environment or other cows, and no simple solutions are available for its control to date.

In this paper, we develop an IoT-based integrated solution to address this problem, where we aim to achieve cost-effective mastitis transmission control using a decentralized platform with novel sensing and interference algorithms to identify the most vulnerable cows for further screening. As such a screening process is labor-intensive and costly, our goal is to maximize the likelihood of successful identification while minimizing false positives. Our core contributions involve the development of a complete methodology for sensing cows' social behavior, inferring the social interaction graphs of cows, modeling disease transmission based on such social graphs, and inferring the most likely candidates for detailed screening. Such a methodology allows us to identify how the potential disease may have transmitted on a farm in a posterior manner. As output, our methodology performs a forecast and recommends a short list of cows for manual screening using additional validation methods, based on which we either conclude the population is free from the disease, or a set of most likely propagation paths and infected cows.

Background: Mastitis is a disease commonly found on dairy farms that is caused by microorganisms, usually bacteria, that invades the udder and multiplying in the milk-producing tissues, producing toxins that result in inflammation. The immune system of cows will respond and fight the infection with an increase in the number of immunocytes, referred to as somatic cells. The number of somatic cells in milk, i.e., somatic cell count (SCC), is an indication of inflammation [4], [5]. An elevation of SCC above 200,000 cell/ml is generally considered abnormal and indicates inflammation in the udder. Mastitis will lead to reduced milk production and lower milk quality. Good-quality milk not only lasts longer, tastes better, but is also more nutritious. On the other hand, milk with a high SCC is arguably associated with indirect health risks to the consumer. The National Mastitis Council (NMC) estimates that mastitis costs dairy producers in the United States over $2 billion annually.

Based on its severity, mastitis can be classified into subclinical and clinical forms. Subclinical mastitis is challenging to detect due to the absence of any visible indications, while it can still lead to abnormally high SCC in milk and can be up to 40 times more common than clinical cases of the illness.

Subclinical mastitis has major cost implications. Studies found that 70 to 80% of mastitis losses were due to subclinical mastitis. On the other hand, commonly used approaches to measure the SCC count are both costly and time-consuming procedures to test the produced milk, hence, not scalable to large herd size and farm-scale practices. Currently, although subclinical mastitis is the predominant form in most herds and is the most costly to a producer, most producers know neither the prevalence rate in their herd nor which cows are infected, and therefore, do not attempt to treat the infected cows.

Approach: Our key research challenge to be answered is that with the help of IoT instrumentation, whether it is possible for us to detect subclinical mastitis by testing just a few cows that are most vulnerable to disease transmission, so that we can infer the health status of the entire farm. This way, preventive measures can be taken early. However, existing sensing and monitoring approaches, such as those based on cameras or manual calibration methods [6], [7], are either too costly to deploy on a large scale, or do not provide the needed accuracy to generate a short list of candidates for further testing. In this paper, we propose an integrated sensing, modeling, and inference framework. Specifically, we deploy sensors on cows to track their behavioral models, based on which we infer their contact history. Based on this history, we infer the most likely disease transmission paths if the candidate cows are indeed identified as positive for disease monitoring. On the other hand, if the whole farm is actually healthy and free of diseases, our methods will validate this with the least amount of effort and testing overhead.

To our best knowledge, the proposed framework is the first IoT based instrumentation platform that will help us understand the social behavior of cows on farm environments. Our results demonstrate the success of the overall approach with validations and results. Our major contributions of this approach are listed as follows:

First, to our knowledge, we are the first to propose the use of social behavior networks of cows for disease tracking applications, where we passively reconstruct such social behavior history for our application. This demonstrates a new potential area of social networks among animals in semi-controlled environments. This study also paves the way for us to uncover the disease transmission paths in later studies.

Second, we design and evaluate novel inference algorithms to reconstruct the history of social interactions purely based on passive data. Our algorithms take both direct and indirect contact caused infection into account, and can be adapted to support different types of configurable combinations, such as the number of assumed sick cows, the disease transmission probability, and the distance thresholds to determine contact.

Third, we demonstrate the correctness and effectiveness of our results through extensive experimental results, including not only simulations, but also in-the-field deployments. Our study also supports the results with a real-world case study that accurately detected one cow with SCC testing results. This demonstrates the correctness of the inference models for the disease transmissions and their analysis results.

The remaining of this paper is organized as follows. Section II and III describes the related work and an overview of system design. Section IV and V present the implementation and the evaluation results respectively. Section VI concludes the paper.

II. Related Work

In this section, we summarize the state-of-the-art practices in smart farming deployments.

Smart Farming Deployments: Smart farming refers to the deployment of information and communication technologies to modern agriculture practices. Recent technologies and techniques are rapidly taken advantage of using satellite imagery [8], [9], agricultural robots [10], [11], large deployments of sensor nodes [12], [13], and unmanned aerial vehicles or drones for aerial imagery and actuation [14]–[17]. IoT based smart farming solutions are applied in a wide variety of domains, such as precision agriculture, greenhouse control, and livestock monitoring [18].

Precision agriculture technologies aim to enhance agricultural productivity by providing smart automation, improving and optimizing agricultural production environments and planning. Different IoT sensors are developed for the applications, including climate condition monitoring [19], soil pattern analysis [20], disease monitoring [21], plant and harvest time optimization [22], among others.

Mastitis Detection in Dairy Cows: IoT based livestock monitoring solutions aim to improve dairy products and livestock conditions by attaching different monitoring sensors to the animals to obtain their performance. Due to the increasing demand for early identification of disease symptoms, IoT sensing devices have been deployed to monitor the animal temperature [23], heart rate [18], and physical gestures [24] to prevent animals suffering from any disease. However, our work is different in that we are the first to use location data to build passive social network graphs of cows for disease tracking and preventing.

One important step of the overall procedure of our methodology is to manually screen cows once most likely candidates have been identified. In practice, this labor-intensive step is done by measuring the SCC counts of milk produced. On January 1, 2012, the U.S. dairy industry began transitioning to a producer-level milk sampling program for SCC compliance with EU regulations for products exported to the EU. The impact of this regulation will be profound because this change will hold individual farms accountable to stricter quality criteria.

Commonly used methods of mastitis detection include SCC estimation, electrical conductivity (EC), and identification of the causative microorganisms [25]. Methods have been developed for SCC estimation, such as direct microscopic somatic cell counting, California mastitis test (CMT), Portacheck, Fossomatic SCC, or DeLaval cell counter, among others [26]. The gold standard to determine SCC is to count the somatic cells with methylene blue staining. This is time consuming and technically demanding. Alternatively, many farms measure individual cow SCC on a monthly basis by the Dairy Herd Improvement Association (DHIA) which can be costly and offers limited effectiveness. About 60 to 70% of environmental pathogen infections exist for less than 30 days and are not

easy to be detected, so some cases could go unnoticed with monthly tests. For the inline test, the only commonly used method is to test the EC content of milk [27], [28]. It is known that mastitic milk has a higher EC than normal milk. Most automatic milking units today are equipped with EC sensors that flag cows with clinical mastitis but cannot effectively identify cows with subclinical mastitis. However, it is fair to say that the existing methods are rather limited in terms of speed, sensitivity, frequency, and/or expense [29]. Therefore, it will be extremely useful if continuous farm-wide monitoring is conducted to provide clues as to whether the mastitis is of the contagious or environmental type.

Comparison with Similar Approaches: There have been limited deployments efforts in using IoT technologies to forecast herd farm diseases. In [30], the authors developed a platform for similar purpose of detecting mastitis using IoT integration. In their approach, data from a range of measurement devices, cattle collars, milking station and feed wagon are integrated into a cloud infrastructure. At the milking robot, they used sensors to measure the conductivity of the milk as a change in conductivity arises from an increase in milk Na+ concentration as a consequence of infection. To improve detection, they also collected accelerometer derived data from the Afimilk Silent Herdsman collar to provide an early indication of the onset of mastitis which informs an early intervention action. The combination of the two measurements provides corroboration between two radically different sensor modalities and provides an improvement in the measurement reliability and accuracy. This approach is very different from ours in the types of sensors used and in the data aggregation algorithm. Further, they still need to manually test each cow for disease detection without developing any real capabilities for forecasting, as they do not collect any interaction history nor cow trajectories.

In another effort of using IoT to improve cows' health [31], the authors discover that in almost all cases they observed, mastitis manifests itself in a very sharp decrease in rumination as measured using sensors. By measuring this metric alone, in eight out of eleven mastitis cases, the alarm was generated before the mastitis was diagnosed. However, the paper is different from ours considerably as they do not provide complete details on the false alarm rate, and whether their experience can be validated on farms with different management protocols.

III. System Design

The proposed system consists of four components, i.e., cow social behavior tracking, social graph building, disease propagation modeling, and screening list recommendation, as shown in Figure 1.

A. Cow Social Behavior Tracking

We leverage portable battery-powered GPS devices to monitor cows' trajectories and track their social behaviors. Specifically, two GPS devices with the same signal scanning frequency are assigned to each cow to track its movement. Even if when one of the two devices failed to work, cow trajectories can still be collected by the other. Suppose Δt data points are lost in a GPS trajectory $\vec{g} = \langle g_{t-1}, g_t, g_{t+\Delta t+1}, g_{t+\Delta t+2} \rangle$, where $g_t = (lon_t, lat_t)$ is the GPS location collected at time t, and $\langle g_{t+1}, \ldots, g_{t+\Delta t} \rangle$ are missing. Assuming that the missing Δt points are evenly distributed between g_t and $g_{t+\Delta t+1}$ (i.e., cows walk along a straight line at a constant velocity), we estimate the missing point g_{t+i} as follows:

$$g_{t+i} = (\frac{lon_{t+\Delta t+1} - lon_t}{\Delta t+1} * i + lon_t, \frac{lat_{t+\Delta t+1} - lat_t}{\Delta t+1} * i + lat_t) \quad (1)$$

where $i \in \{1, 2, \ldots, \Delta t\}$. Besides location g_t, modern GPS tracking devices, e.g., GPS SiRF star III chipsets [32] used in our studies, usually return an estimated horizontal position error (EHPE) e_t associated with g_t. Accordingly, we use the averaged $\frac{e_t + e_{t+\Delta t+1}}{2}$ to approximate e_{t+i}.

Next, it is necessary to detect and weed location outliers because external factors, such as occlusions and GPS signal instability, might introduce noisy data into cow trajectories. We rely on both EHPEs and cow walking speeds to clean data. As a built-in location error estimation reported by GPS devices, an EHPE measures the inaccuracy of an individual collected location. We design an EHPE threshold θ_e to filter out those GPS data points with a high EHPE. For more details of the θ_e setting and EHPE distributions in our study, see Section V.

Besides EHPEs, we leverage cow walking speeds to identify noisy locations. For location g_t collected at time t, we estimate cow walking speeds for its following n time steps (i.e., from time $t+1$ to $t+n$). If g_t is an outlier, its distance to normal data points is more likely to be large, leading to a high walking speed. Considering noisy data may exist at the following n time steps, we remove the largest speed to reduce such effects. Then we calculate the averaged speed for the rest $n-1$ time steps:

$$\bar{v} = \frac{\sum_{i=1}^{n} \frac{d(g_{t+i}, g_t)}{t+i-t} - \max_{1 \le i \le n} \frac{d(g_{t+i}, g_t)}{t+i-t}}{n-1} \quad (2)$$

where $d(g_{t+i}, g_t)$ represents the Euclidean distance between GPS locations g_{t+i} and g_t. If $\bar{v}$ is higher than a speed threshold θ_v, location g_t is regarded as a noise data point. Otherwise, g_t is a normal data point. The setting of θ_v can be found in Section V.

Finally, we perform data fusion. After fixing missing data and removing noise, we merge the two GPS trajectories generated by the same cow into one. Since the two GPS locations at the same time may have different EHPEs, it is reasonable to assign a larger weight to the location with a lower EHPE. Suppose we have two GPS locations $g_t^1 = (lon_t^1, lat_t^1)$ and $g_t^2 = (lon_t^2, lat_t^2)$ collected at time t with EHPEs of e_t^1 and e_t^2 respectively. To ensure the location with a lower EHPE is assigned a higher weight in data fusion, we express the merged weighted average latitude and longitude coordinate (lon_t, lat_t) is expressed as:

$$g_t = (lon_t, lat_t) = (\frac{lon_t^1 e_t^2 + lon_t^2 e_t^1}{e_t^1 + e_t^2}, \frac{lat_t^1 e_t^2 + lat_t^2 e_t^1}{e_t^1 + e_t^2}) \quad (3)$$

Note that if either of g_t^1 and g_t^2 is missing, we take the existing one to represent the merged data.

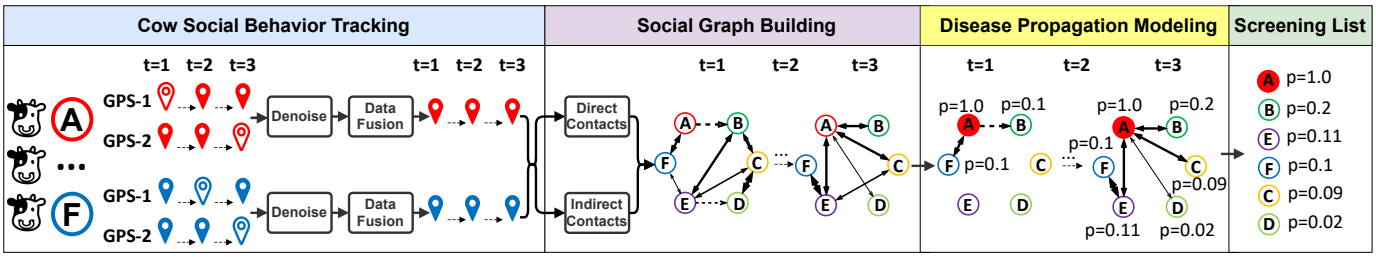


Fig. 1. Framework overview. We deploy two GPS devices on each cow to track its movement. The two collected trajectories will be merged after fixing missing data (represented by unfilled location markers) and removing noise data. Then we build cow social behavior networking graphs by treating cows as vertices, and direct (solid line) and indirect (dashed line) contacts as edges. Base on social behavior graphs, a probabilistic disease transmission model is proposed to estimate disease propagation in cow herds. Finally, we rank all cows' infection probabilities and suggest a list of cows for screening in order of decreasing priority.

B. Social Network Modeling

To model cow social behaviors, we build a weighted directed graph $G = (V,\ E,\ W(E))$ by treating cows as vertices, contacts between cows as edges, and contact frequencies as edge weights. In graph G, $V = \{v_i | i = 1, 2, \cdots, N\}$ is a set of vertices where N is the total number of cows; $E \subseteq \{(v_i, v_j) | (v_i, v_j) \in V^2 \wedge v_i \neq v_j\}$ is a set of directed (from v_i to v_j) edges; $W(E) = \{w_{v_i,v_j} | (v_i, v_j) \in E \wedge w_{v_i,v_j} \in \mathbb{R}^+\}$ is a set of edge weights. Instead of building one graph G for each time step, we build one graph G by aggregating all cow contacts within a long time period T to reduce computing overhead. For example, our GPS devices collect one location point every second, but we build one graph G per minute rather than per second.

Suppose we have two preprocessed GPS trajectories $\vec{g^a} = \langle g_1^a, g_2^a, \cdots, g_T^a \rangle$ and $\vec{g^b} = \langle g_1^b, g_2^b, \cdots, g_T^b \rangle$ for cow A and cow B respectively, where g_t^a and g_t^b are latitude and longitude coordinates at time t, and T is the length of total time steps to build G. We propose the following models to identify both direct and indirect contacts between cow A and cow B to establish edges.

Direct Contact Model: It is intuitive to determine a contact between cow A and cow B if the Euclidean distance between g_t^a and g_t^b is below a distance θ_d, i.e., $d(g_t^a, g_t^b) < \theta_d$. Although the above method is straightforward and efficient, it may suffer from insufficient robustness due to the instability of GPS signals and EHPEs. Therefore, we propose a sliding-window contact model, which incorporates all point-to-point distances within a time sliding-window of $2\tau + 1$ rather than a single distance calculated at time t. Note that the sliding-window contact model only considers the $d(g_t^a, g_t^b)$ when we set τ as 0. Specifically, we calculate the average point-to-point distance from the time $t - \tau$ to the time $t + \tau$:

$$\bar{d} = \frac{\sum_{i=t-\tau}^{t+\tau} d(g_i^a, g_i^b)}{2\tau + 1} \tag{4}$$

If the average distance $\bar{d}$ is below θ_d, cow A and cow B interacts with each other at time t. Let v_a and v_b represent the vertices of cow A and B in G. Then edge (v_a, v_b) and (v_b, v_a) are added into G. If cow contacts are observed for n times among the T time steps, we set w_{v_a,v_b} and w_{v_b,v_a} as n.

Indirect Contact Model: Direct contact requires two cows to stay physically near enough at the same time step. However, cow living environments also play a key role in the transmission of infectious diseases. For example, a sick cow A stays in one place and contaminates its surrounding environment. Later, cow B, who comes and stays close to the contaminated place for a long time, might be infected through the environment. We define such contacts as indirect contacts in this paper.

One of the challenges involved in recognizing indirect contacts is to localize where cows stay without moving. GPS devices used in our study switch automatically into hibernation mode if the cow stops moving longer than a certain consecutive time period, and re-enter data collecting mode when the cow starts to move again. When processing the GPS data, we set latitude and longitude coordinates during the hibernation mode as $(0, 0)$. For a GPS trajectory $\vec{g} = \langle g_1, g_2, ..., g_n \rangle$, we use a time period δ threshold to identify consecutive inactivity. If more than or equal to δ consecutive $(0, 0)$ are found, a staying will be determined. Then we use the most recent k GPS data points prior to inactivity to infer exact staying locations. For example, we regard the latest non-hibernation data point as the staying location when k is set as 1. When k is larger than 1, instead an average location is used. For each staying, we also record its start time t_s and end time t_e. The details of how to infer staying locations are illustrated in Algorithm 1.

Algorithm 1 Infer Staying Location

```
procedure StayLocation(g⃗, δ, k)
    s⃗ ← [ ]                                  ▷ a list of pairs of staying location and time
    c ← 0                                    ▷ initialize the number of consecutive (0, 0)
    for i = 1 → |g⃗| do                       ▷ |g⃗| is the length of trajectory
        if g_i = (0, 0) then
            c ← c + 1                        ▷ find a (0, 0) location
        else
            if c >= δ then                   ▷ find a staying location
                t_s ← i − c                  ▷ the start time of staying
                t_e ← i − 1                  ▷ the end time of staying
                ḡ ← (Σ_{t=i−c−k}^{i−c−1} g_t) / k   ▷ the staying location
                append (ḡ, (t_s, t_e)) to s⃗
                c ← 0                        ▷ re-count consecutive (0, 0)
    return s⃗                                ▷ staying locations and associated time
```

Using Algorithm 1, we obtain staying locations and associated time information $\vec{s_a}$ and $\vec{s_b}$ for cow A and cow B respectively. Next, we present how to identify indirect interactions and how to add corresponding weighted directed edges in cow social graph G based on $\vec{s_a}$ and $\vec{s_a}$. Similar to the direct contact model, we use vertices v_a and v_b to represent cow A and cow B in G. Then we propose a threshold θ_c to determine whether staying locations of two cows are close enough for disease transmission. A directed edge (v_a, v_b) with

a weight of w_{v_a,v_b} will be added to the social graph G when the staying location distance between cow A and cow B are below θ_c and cow A comes first. The w_{v_a,v_b} is estimated by cow A's staying duration $t_e^a - t_s^a$ and the time interval between the departure of cow A and the arrival of cow B, i.e., $t_s^b - t_e^a$. We also introduce an environment-based disease transmission probability decay rate ρ to estimate the weight as:

$$w_{v_a,v_b} = (t_e^a - t_s^a)e^{-\rho(t_s^b - t_e^a)} \tag{5}$$

where t_s^a (t_s^b) and t_e^a (t_e^b) are the start and end staying time of cow A (cow B). The definition of w_{v_a,v_b} keeps consistent with the two following facts: (i) a longer staying of cow A leads to a higher infection probability for cow B, and (ii) the environment-based infectivity decreases over time. More details of identifying indirect contacts are illustrated in Algorithm 2.

Algorithm 2 Identify Indirect Contact

procedure INDIRECTCONTACT($\vec{s_a}, \vec{s_b}, G, \theta_c, \rho$)
  for each s_a in $\vec{s_a}$ **do**
    for each s_b in $\vec{s_b}$ **do**
      $(g^a, (t_s^a, t_e^a)) \leftarrow s_a$ ▷ staying info. for cow A
      $(g^b, (t_s^b, t_e^b)) \leftarrow s_b$ ▷ staying info. for cow B
      if $d(g^a, g^b) < \theta_c$ **then** ▷ indirect contact exists
        if $t_e^a <= t_s^b$ **then** ▷ cow B follows cow A
          $w_{v_a,v_b} \leftarrow (t_e^a - t_s^a)e^{-\rho(t_s^b - t_e^a)}$
          add edge (v_a, v_b) with w_{v_a,v_b} to G
        else ▷ cow A follows cow B
          $w_{v_b,v_a} \leftarrow (t_e^b - t_s^b)e^{-\rho(t_s^a - t_e^b)}$
          add edge (v_b, v_a) with w_{v_b,v_a} to G
  return G ▷ return updated social behavioral graph

C. Disease Propagation Model

As mentioned before, we build one graph for a long time period, such as one minute. Based on a list of cow social behavior graphs $\vec{G}$, we propose a disease transmission model to explore how diseases propagate among cows over time. Suppose we have two cows v_i and v_j and there exists an edge (v_i, v_j) with a weight w_{v_i,v_j} in G_t, which is the graph built during time period t. We set the disease transmission probability for each unit contact (i.e., $w_{v_i,v_j} = 1$) as r. The infection probability of cow v_i at time t is denoted as $p_{v_i}^t$. At time period $t+1$, we update $p_{v_j}^{t+1}$ as:

$$p_{v_j}^{t+1} = p_{v_j}^t + (1 - p_{v_j}^t) * p_{v_i}^t * w_{v_i,v_j} * r \tag{6}$$

where $p_{v_j}^t$ is the infection probability of cow v_j at time period t, and second term is the infection probability contributed by the edge (v_i, v_j) at time period $t+1$. We assume a subset of cows $V' \subseteq V$ are sick at the beginning (i.e., $t = 0$) and update infection probability of each cow as shown in Algorithm 3.

D. Screening List Recommendation

It is laborious to screen all cows to check whether they are infected with mastitis. Instead, we rank cows according to their expected infection probabilities and recommend cows most likely to be infected. For each cow v_i, we assume it is infected at the beginning (i.e., $V' = \{v_i\}$) and use the proposed propagation model in Algorithm 3 to calculate the corresponding final infection probabilities $P^{V'} = \{P_{v_j}^{V'} | v_j \in V\}$ where $P_{v_j}^{V'}$ is the infection probability of cow v_j given cow v_i is sick. Then we calculate the average infection probability $\bar{p}_{v_i}$ of cow v_i with the assumption that each cow is sick at the beginning. Finally, we rank $\bar{p} = \{\bar{p}_{v_i} | v_i \in V\}$ in descending order and suggest cows for further screening. Algorithm 4 shows more details of generating the recommended cow list.

Algorithm 3 Propagation Model

procedure DISEASETRANSMISSION($\vec{G}, V, V', r$)
  for each v_i in V **do** ▷ all cows
    if $v_i \in V'$ **then** ▷ sick cows
      $p_{v_i}^0 \leftarrow 1$ ▷ initialize prob. of sick cows as 1
    else ▷ healthy cows
      $p_{v_i}^0 \leftarrow 0$ ▷ initialize prob. of healthy cows as 0
  for $t = 0 \rightarrow |\vec{G}| - 1$ **do** ▷ update prob. over time
    for each (v_i, v_j) in G_t **do**
      $p_{v_j}^{t+1} = p_{v_j}^t + (1 - p_{v_j}^t) * p_{v_i}^t * w_{v_i,v_j} * r$
  $P^{V'} = \{p_{v_i}^{|\vec{G}|} | v_i \in V\}$
  return $P^{V'}$ ▷ infection prob. of all cows at time T

Algorithm 4 Generate Screening List

procedure GENERATESCREENLIST($\vec{G}, V, r$)
  for each v_i in V **do** ▷ calculate prob. if cow v_i is sick
    $V' = \{v_i\}$ ▷ cow v_i is sick at time $t = 0$
    $P^{V'} \leftarrow DiseaseTransmission(\vec{G}, V, V', r)$
  for each v_i in V **do** ▷ iterate each cow
    $sum = 0$
    for each v_j in V **do** ▷ iterate each sick cow
      $V' = \{v_j\}$ ▷ cow v_j is sick at time $t = 0$
      $sum = sum + P_{v_i}^{V'}$ ▷ add corresponding prob.
    $\bar{p}_{v_i} = \frac{sum}{|V|}$ ▷ average final probability of cow v_i
  sort $\bar{p} = \{\bar{p}_{v_i} | v_i \in V\}$ by value descending
  return $\bar{p}$ ▷ infection prob. at time T for all cows

IV. IMPLEMENTATION

In this section, we present the system implementation including GPS hardware deployment and software development.

A. GPS Device Deployment

We use portable GPS tracking devices equipped with a 65nm SiRF III GPS chipset to collect cow trajectories. Two types of GPS devices – i-gotU GT-600 and iTrail Logger H6000 – are deployed because of a series of advantages like small sizes (1.8″ x 1.6″ x 0.5″), light weights (1.3oz) and a long battery life (>100 hours). In addition, the two devices are water-resistant and get into hibernation after several minutes of inactivity, making them more practical on dairy farms. When deploying GPS devices on cows, we first power on a device and put it into a pouch, which could prevent accidental button pushes and improve water resistance. Then, we attach the pouch to cow collars. Both the two GPS devices have built-in storage memory, where the collected data is automatically saved. The data was processed in an offline manner. We detached GPS devices from cows and exported the GPS data into local computers for further analysis.

1) Data Visualization and Preprocessing: We develop a cow trajectory visualization system, cow social behavior graph visualization tools, and a disease transmission modeling simulator. It is essential to develop a flexible GPS data visualization system to understand cow moving behaviors. We implement

such a visualization system through Google Maps JavaScript APIs. Our system is capable of plotting any given latitude and longitude coordinates, drawing cow trajectories of specific time periods and cow identifiers.

2) Disease Transmission Modeling Simulator: Besides the theoretical analysis of disease transmissions in Algorithm 3, we implement a disease propagation simulator to validate the correctness of our theoretical propagation model. Instead of assigning an infection probability to each cow, we monitor the infection status (it can be either infected or uninfected) of cows. When launching the simulator, we label the sick cow status as infected and the rest as healthy. Then, we updated cow statuses step by step based on their contacting information. For an edge (v_i, v_j) in cow social graph G, we generate a random number $x \in [0, 1]$. If x is smaller than the transmission probability r and cow v_i has been infected, the status of cow v_j will be set as infected. Otherwise, the status of cow v_j keeps unchanged. For each initialized setting, e.g., assumed sick cows and infection probabilities, we run the simulator for a large number of rounds. Then we calculate the average infection probabilities of each cow using the number of infected status over the total number of running iterations.

V. Evaluation

In this section, we first present the experimental setup and the collected dataset. Next, we give all parameter settings used in data preprocessing, social behavior graph building, and disease propagation modeling. We also compare the proposed disease propagation theoretical model with simulators to prove its correctness. Furthermore, we show how to suggest cows for further screening in two common mastitis detection scenarios. Finally, one real-world SCC based case study is conducted to demonstrate the effectiveness of our framework.

A. Experimental Setup and Dataset

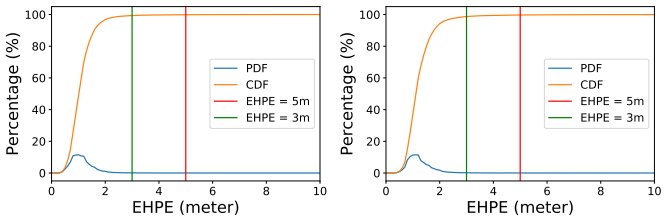
We conducted in-the-field experiments in a span of fourteen days on a university-owned dairy farm. With the help of farm staff, two GPS devices, which could be either i-gotU GT-600and or iTrail Logger H6000, were deployed on the collar of each cow (see Figure 2). In our experiments, 17 cows from 6 different pens were involved. We combined pen IDs and cow IDs to identify unique cows in our study. For example, pen4_1 is the first cow from the fourth pen. The scanning frequency of all GPS devices was set as 1, i.e., collecting one data point per second. In total, we collected more than 70-hours cow movement data.

B. Parameter Settings

When preprocessing raw GPS data, we set the missing data threshold θ_m as 5. It means that if one GPS device lost more than 5 consecutive data points, we treat these data are true positive lost data. Otherwise, we recover the false positive missing data based on Equation 1. The average false positive missing data percentage of all GPS devices is 0.72%, with a standard deviation of 1.47%. The aggregated EHPE distributions of the two types of GPS devices used in our experiments are illustrated in Figure 3. More than 98.56% collected location data have an EHPE within 3 meters,

(a) Dairy cattle shed (b) GPS devices deployed on collar

Fig. 2. GPS device deployment on a university-hosted dairy cattle farm.



(a) i-gotU GT-600 (b) iTrail Logger H6000

Fig. 3. EHPE distributions of i-gotU GT-600 and iTrail Logger H6000 GPS devices. More than 98.56% collected location data have an EHPE within 3 meters, and 99.68% within 5 meters.

and 99.68% within 5 meters. So we set the EHPE filtering threshold θ_e as 3 meters, removing those locations with an EHPE higher than 3 meters because of the high uncertainty. In addition, we set $n = 10$ and $\theta_v = 5$ m/s in Equation 2 to remove noisy data by speed. After data cleaning, we find the mean conflict percentage between the dual devices on the same cow is 12.54%, with a standard deviation of 11.51%.

When building cow social behavior graph G, we aggregated and identified direct and indirect contacts within one minute. In direct contact model, we set the time sliding-window size as 3, i.e., $\tau = 1$ in Equation 4. We illustrated the impacts of different direct contact distance thresholds θ_d in our case studies. In the indirect contact model, the consecutive inactive time period threshold δ is set as 60 seconds and the number of most recent points prior to inactivity k in Algorithm 1 is set as 15, because we found these settings achieved the highest accuracy when inferring the pen of cows based on cow inactive behaviors. We set the indirect contact distance threshold $\theta_c = 1$ and environment-based disease transmission probability decay rate $\rho = 0.01$ in Algorithm 2 by default.

C. Theoretical Analysis VS Simulated Analysis

It is time-consuming to simulate the disease propagation using the entire collected real-world dataset. Our dataset contained 17 cows and more than 3900 time steps in minute, which requires a huge number of random numbers to be generated for each round of simulation. In addition, thousands of rounds of simulations have to be run to ensure accurate final infection probabilities. Therefore, we randomly pick up 30-minutes collect data to compare the theoretical and simulated results with different disease propagation settings. Specifically, each cow or any arbitrary two cows are assumed as sick before launching the simulation. We calculate the root-mean-square error (RMSE) between simulated and theoretical infection probabilities of other cows with different transmission

probabilities, i.e., $r = 0.0001$ or $r = 0.0002$, after certain simulation rounds. As shown in Figure 4, the RMSE between simulated and theoretical results decreases with the increase of simulation rounds. When the number of simulation rounds achieves 10000, the RMSE is less than 0.001, indicating the correctness of our proposed theoretical propagation model.

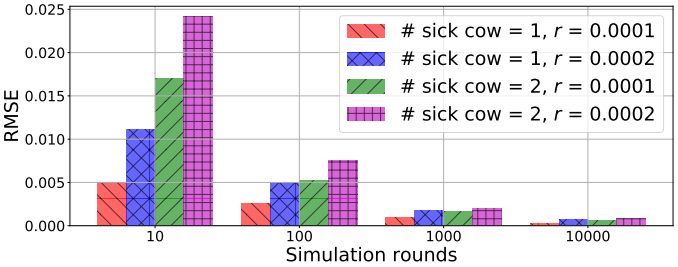


Fig. 4. RMSE between simulated and theoretical results with different number of assumed infected cows and transmission probabilities. We set $\theta_d = 5$, $\theta_c = 1$, $\rho = 0.01$, and disease transmission probability $r = 0.0001$.

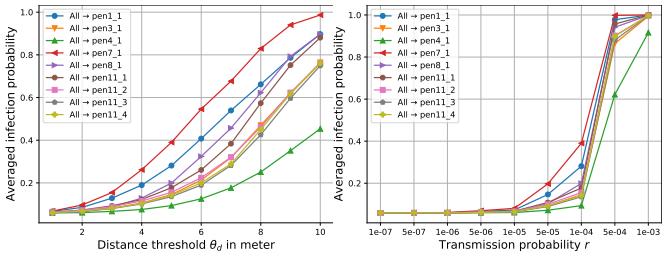
D. Recommend Cows for Further Screening in Two Scenarios

The proposed framework is flexible to offer suggestions on which animals should be screened with a high priority in the following two scenarios: 1) sick cows are unknown, but we want to determine whether there exist infected cows on the farm; 2) sick cows are known, and we want to figure out which cows are more likely to be infected. We use *pen_x→pen_y* to represent mastitis is transmitted from cow pen_x to cow pen_y.

Scenario 1: Sick Cows Are Unknown. Without knowing which cows are infected, it is reasonable to assume that each cow is sick with the same likelihood. Following Algorithm 4, we generate a screening list by ranking cows based on their averaged infection probabilities when each cow is assumed as sick at the beginning. The effects of both distance thresholds θ_d and transmission probabilities r on averaged probabilities are demonstrated in Figure 5a and Figure 5b, where the legend of *All→pen_x* represents the disease is transmitted from each cow to cow pen_x. The cow pen7_1 and cow pen4_1 are the most and least likely to be infected cows in the two figures. Therefore, cow pen7_1 will be assigned the highest screening priority to check whether exist infected cows on the farm. It is interesting to note that cows from the same pen (see cow pen11_1, pen11_2, pen11_3, and pen11_4 in Figure 5a and Figure 5b) illustrate similar infection probabilities.

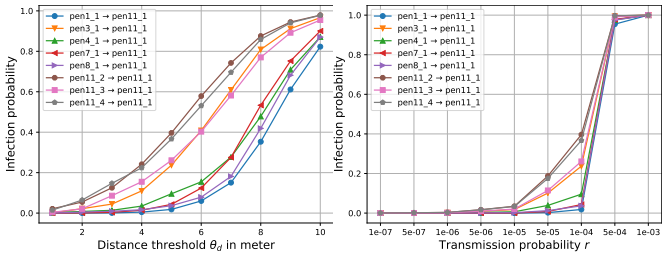
Scenario 2: Sick Cows Are Known. When some cows have already been diagnosed with mastitis, we apply the propagation model in Algorithm 3 to figure out more infected cows. By treating each diagnosed cow as sick at the beginning, we calculate the theoretical infection probabilities of other cows. We used different distance thresholds θ_d to build contact graphs and different transmission probabilities r to estimate the propagation speed between cows.

Figure 6a and Figure 6b illustrate the infection probabilities of cow pen11_1 when different cows are diagnosed with mastitis. We can see both a larger θ_d and a larger r cause a higher infection probability due to more interactions and faster propagation. Besides the two factors of θ_d and r, we find cows located in the same pen are more likely to be infected.



(a) Different θ_d and $r = 0.0001$ (b) Different r and $\theta_d = 5$

Fig. 5. Infection probabilities of each cow to suggest screening priorities when sick cows are unknown (All → pen_x). We set indirect contact distance threshold $\theta_c = 1$ and the decay rate $\rho = 0.01$.



(a) Different θ_d and $r = 0.0001$ (b) Different r and $\theta_d = 5$

Fig. 6. Infection probabilities of pen11_1 when other cows are diagnosed with mastitis (pen_x → pen11_1). We set $\theta_c = 1$ and the decay rate $\rho = 0.01$.

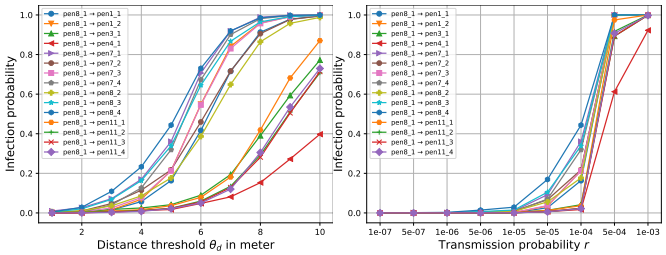
As demonstrated in both Figure 6a and Figure 6b, the cow pen11_1 has a higher infection probability if cow pen11_2, pen11_3 and pen11_4 which shared the same pen with cow pen11_1, were sick. Note the pen 3 is a dump pen where cows are located for investigation. The cow pen3_1 was first located in the pen 3 and then was sent back to the pen 11, which explained why the cow pen11_1 had a high infection probability when the cow pen3_1 was sick at the beginning.

E. Real-world SCC Case Study

To demonstrate the effectiveness of the proposed disease propagation model in real life, we conduct the comparison of our predictions to ground truth as shown in Somatic cell count (SCC) tests. Specifically, if the SCC, i.e., the number of cells present in 1 ml (about a quarter of a teaspoon) of milk, was larger than 200,000 cells/ml, the cow was regarded as infected [33]. During our GPS data collection period, the cow pen8_1 was detected to have mastitis because of its SCC of 800,000 cells/ml. After data collection, we observed the SCC of the cow pen8_4 increased from 123,000 to 348,000. In our model, we assume the pen8_1 is sick at beginning and calculate the final infection probabilities of all other cows. As shown in Figure 7, our model suggests the cow pen8_4 for further screening with the highest priority. The prediction results are hence consistent with the ground truth, demonstrating the effectiveness of the proposed overall approach.

VI. Conclusion

In this paper, we design, implement, and evaluate an IoT-based cattle social behavior sensing framework to detect and prevent mastitis among dairy cows. To make the collected trajectories more robust and reliable, data fusion techniques are adopted to merge data from multiple sources, and noisy



(a) Different θ_d and $r = 0.0001$ (b) Different r and $\theta_d = 5$

Fig. 7. Infection probabilities of other cows when pen8_1 was diagnosed with mastitis (pen8_1 $\rightarrow$ pen_x). We set $\theta_c = 1$ and the decay rate $\rho = 0.01$.

data are filtered out by GPS hardware-reported errors. Then we propose configurable directed and weighted social behavior graphs, based on which we develop a probabilistic disease transmission model to forecast and detect cows with a high infection risk for further screening. Both theoretical and simulation-based analytics of in-the-field experimental data demonstrate that the proposed framework is useful and effective. Finally, additional SCC based tests demonstrate that our approach achieves consistent results on predicting infected cows with the ground truth results.

ACKNOWLEDGEMENT

This work was supported by the US NSF CPS/USDA NIFA (Grant No. 2017-67007-26150).

REFERENCES

[1] A. Villarroel, D. A. Dargatz, V. M. Lane, B. J. McCluskey, and M. D. Salman, "Suggested outline of potential critical control points for biosecurity and biocontainment on large dairy farms," *Journal of the American Veterinary Medical Association*, vol. 230, no. 6, pp. 808–819, 2007.

[2] F. Maunsell and G. A. Donovan, "Biosecurity and risk management for dairy replacements," *Veterinary Clinics of North America: Food Animal Practice*, vol. 24, no. 1, pp. 155–190, 2008.

[3] H. Seegers, C. Fourichon, and F. Beaudeau, "Production effects related to mastitis and mastitis economics in dairy cattle herds," *Veterinary research*, vol. 34, no. 5, pp. 475–491, 2003.

[4] H. Barkema, Y. Schukken, T. Lam, M. Beiboer, H. Wilmink, G. Benedictus, and A. Brand, "Incidence of clinical mastitis in dairy herds grouped in three categories by bulk milk somatic cell counts," *Journal of dairy science*, vol. 81, no. 2, pp. 411–419, 1998.

[5] M. Green, L. Green, Y. Schukken, A. Bradley, E. Peeler, H. Barkema, Y. De Haas, V. Collis, and G. Medley, "Somatic cell count distributions during lactation predict clinical mastitis," *Journal of Dairy Science*, vol. 87, no. 5, pp. 1256–1264, 2004.

[6] M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E. M. Aggoune, "Internet-of-things (iot)-based smart agriculture: Toward making the fields talk," *IEEE Access*, vol. 7, pp. 129551–129583, 2019.

[7] K. Haseeb, I. U. Din, A. Almogren, and N. Islam, "An energy efficient and secure iot-based WSN framework: An application to smart agriculture," *Sensors*, vol. 20, no. 7, p. 2081, 2020.

[8] C. Krintz, R. Wolski, N. Golubovic, B. Lampel, V. Kulkarni, B. Sethuramasamyraja, B. Roberts, and B. Liu, "Smartfarm: Improving agriculture sustainability using modern information technology," in *KDD Workshop on Data Science for Food, Energy, and Water*, 2016.

[9] C. Corbari, R. Salerno, A. Ceppi, V. Telesca, and M. Mancini, "Smart irrigation forecast using satellite landsat data and meteo-hydrological modeling," *Agricultural Water Management*, vol. 212, pp. 283–294, 2019.

[10] H. Gan and W. Lee, "Development of a navigation system for a smart farm," *IFAC-PapersOnLine*, vol. 51, no. 17, pp. 1–4, 2018.

[11] N. V. Reddy, A. Reddy, S. Pranavadithya, and J. J. Kumar, "A critical review on agricultural robots," *International Journal of Mechanical Engineering and Technology*, vol. 7, no. 4, pp. 183–188, 2016.

[12] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems engineering*, vol. 114, no. 4, pp. 358–371, 2013.

[13] T. Ojha, S. Misra, and N. S. Raghuwanshi, "Wireless sensor networks for agriculture: The state-of-the-art in practice and future challenges," *Computers and Electronics in Agriculture*, vol. 118, pp. 66–84, 2015.

[14] P. Tripicchio, M. Satler, G. Dabisias, E. Ruffaldi, and C. A. Avizzano, "Towards smart farming and sustainable agriculture with drones," in *2015 International Conference on Intelligent Environments*, pp. 140–143, IEEE, 2015.

[15] P. Lottes, R. Khanna, J. Pfeifer, R. Siegwart, and C. Stachniss, "Uav-based crop and weed classification for smart farming," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3024–3031, IEEE, 2017.

[16] T. Moribe, H. Okada, K. Kobayashl, and M. Katayama, "Combination of a wireless sensor network and drone using infrared thermometers for smart agriculture," in *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1–2, IEEE, 2018.

[17] U. R. Mogili and B. Deepak, "Review on application of drone systems in precision agriculture," *Procedia computer science*, vol. 133, pp. 502–509, 2018.

[18] M. S. Farooq, S. Riaz, A. Abid, K. Abid, and M. A. Naeem, "A survey on the role of iot in agriculture for the implementation of smart farming," *IEEE Access*, vol. 7, pp. 156237–156271, 2019.

[19] J. M. Talavera, L. E. Tobón, J. A. Gómez, M. A. Culman, J. M. Aranda, D. T. Parra, L. A. Quiroz, A. Hoyos, and L. E. Garreta, "Review of iot applications in agro-industrial and environmental fields," *Computers and Electronics in Agriculture*, vol. 142, pp. 283–297, 2017.

[20] B. Keswani, A. G. Mohapatra, A. Mohanty, A. Khanna, J. J. Rodrigues, D. Gupta, and V. H. C. de Albuquerque, "Adapting weather conditions based iot enabled smart irrigation technique in precision agriculture mechanisms," *Neural Computing and Applications*, vol. 31, no. 1, pp. 277–292, 2019.

[21] S. Zhang, X. Chen, and S. Wang, "Research on the monitoring system of wheat diseases, pests and weeds based on iot," in *2014 9th International Conference on Computer Science & Education*, pp. 981–985, IEEE, 2014.

[22] T. F. Khan and D. S. Kumar, "Ambient crop field monitoring for improving context based agricultural by mobile sink in wsn," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1431–1439, 2020.

[23] X. Shi, X. An, Q. Zhao, H. Liu, L. Xia, X. Sun, and Y. Guo, "State-of-the-art internet of things in protected agriculture," *Sensors*, vol. 19, no. 8, p. 1833, 2019.

[24] K. Saravanan and S. Saraniya, "Cloud iot based novel livestock monitoring and identification system using uid," *Sensor Review*, 2018.

[25] A. Schepers, T. Lam, Y. Schukken, J. Wilmink, and W. Hanekamp, "Estimation of variance components for somatic cell counts to determine thresholds for uninfected quarters," *Journal of Dairy Science*, vol. 80, no. 8, pp. 1833–1840, 1997.

[26] C. Viguier, S. Arora, N. Gilmartin, K. Welbeck, and R. OKennedy, "Mastitis detection: current trends and future perspectives," *Trends in biotechnology*, vol. 27, no. 8, pp. 486–493, 2009.

[27] B. K. Bansal, J. Hamann, N. T. Grabowski, and K. B. Singh, "Variation in the composition of selected milk fraction samples from healthy and mastitic quarters, and its significance for mastitis diagnosis," *Journal of Dairy Research*, vol. 72, no. 2, pp. 144–152, 2005.

[28] C. Le Maréchal, R. Thiéry, E. Vautor, and Y. Le Loir, "Mastitis impact on technological properties of milk and quality of milk productsa review," *Dairy Science & Technology*, vol. 91, no. 3, pp. 247–282, 2011.

[29] E. Bendixen, M. Danielsen, K. Hollung, E. Gianazza, and I. Miller, "Farm animal proteomicsa review," *Journal of proteomics*, vol. 74, no. 3, pp. 282–293, 2011.

[30] C. Michie, I. Andonovic, C. Davison, A. Hamilton, C. Tachtatzis, N. Jonsson, C.-A. Duthie, J. Bowen, and M. Gilroy, "The internet of things enhancing animal welfare and farm operational efficiency," *Journal of Dairy Research*, vol. 87, no. S1, pp. 20–27, 2020.

[31] O. Unold, M. Nikodem, M. Piasecki, K. Szyc, H. Maciejewski, M. Bawiec, P. Dobrowolski, and M. Zdunek, "Iot-based cow health monitoring system," in *International Conference on Computational Science*, pp. 344–356, Springer, 2020.

[32] G. Morris and L. M. Conner, "Assessment of accuracy, fix success rate, and use of estimated horizontal position error (ehpe) to filter inaccurate data collected by a common commercially available gps logger," *PloS one*, vol. 12, no. 11, 2017.

[33] G. Pighetti, S. Piper, and K. H. Campbell, *Using DHI Reports to Troubleshoot Mastitis*, 2019.