Using Computer Vision to Detect and Analyze Sheep Health and Activity

JOSH MANDZAK, TANNER THORNTON, and COBY WHITE, University of Tennessee, USA

Precision livestock farming (PLF) is an emerging topic in the agricultural industry. PLF aims to bring animals "closer" to the farmer in ever growing farm sizes. As consumer perceptions and policy changes have pushed their industry to focus more on animal welfare. It has been proposed that the use of PLF technology (sensors, cameras, robots, etc.) can help bridge this task. One problem, however, is that many of these technologies are too expensive, too hard to use, or not durable enough for farm use. This project uses a more cost-friendly camera system combined with various machine learning algorithms to determine activity and behavior in video samples of 23 sheep held in a controlled setting. The sheep in this study were subjected to a treatment plan in which each animal (except for the control group) got a different dose of endotoxin. Endotoxin is known to cause stress and illness in sheep; however, endotoxin is a treatable infection but very little has been done to examine the early detection signs of it. Our product will create models to first predict whether a sheep is laying or standing, then use the results to identify patterns and classify level of illness.

MAIN COMPONENT: "Developing a Model"
SECONDARY COMPONENT: "Evaluation"

Additional Key Words and Phrases: Machine Learning, Artificial Intelligence, Convolutional Neural Networks

ACM Reference Format:
Josh Mandzak, Tanner Thornton, and Coby White. 2022. Using Computer Vision to Detect and Analyze Sheep Health and Activity. 1, 1 (December 2022), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Throughout recent history, many advancements have been made in the medical community. Several of these technological innovations have come with the advances of artificial intelligence and machine learning as humans better learn how to use these tools to aid in the detection, diagnosis, and treatment of medical ailments. However, many of these research efforts have been human-oriented, such as using machine learning to identify compounds that could be used to fight Covid-19 [Hack and Papka 2020]. Less research has been done in using technologies such as machine learning and artificial intelligence to aid in the health of animals, in this case specifically sheep. One specific case of this can be discussed with the detection of endotoxin infection within sheep. Currently, very little research has been done to aid in early detection of this illness, despite the fact that it is a treatable infection and it would benefit the animals to be treated as soon as possible. This paper aims to discover if early detection of such an illness is possible with the use of machine learning.

The hypothesis this paper makes is that there is some connection between the standing vs sitting activity level of sheep and level of illness. Currently, this kind of detection is done using accelerometers, which are costly both in terms of time used to set up, monitor, and update the device, as well as monetarily costly. Thus, the first step of this research will be identifying if low cost cameras can be used in combination with computer vision techniques (namely convolutional neural networks) to determine when a sheep is standing vs when a sheep is sitting. Ideally, the results from this portion of the research would be good enough to use as results for step two of this research. However, for the purposes of time, the ground truth data gathered from the accelerometers will be used to complete step two.

The second step in this research will be attempting to build a model to classify level of illness within a sheep based on activity levels of sitting vs standing. This portion of the research will use the ground truth data collected by accelerometer to build this model.

The decision was made to use accelerometer data to complete part two to enable the group to parallelize the two tasks. This means that both steps of the projects will be worked on concurrently, allowing work to be divided.

1.1 Background

The worlds population has been rapidly increasing since the 1st agricultural revolution, as population has increased agriculture has adapted to meet the growing food needs of the world. Now the industry is faced with a challenge to have the ability to feed 9 billion people by the year 2050. The livestock industry has also experienced two other major challenges; growing welfare concerns from consumers, and growing labor shortages. To solve this problem many researchers have suggested implementing precision livestock farming technologies(PLF). PLF aims to allow the farmer to monitor the welfare of their animals in real-time at all times. PLF could prove very effective in providing a solution to the welfare concerns through increased transparency to consumers and increased to reaction time to adverse events such as disease or injury[Berckmans 2014]. PLF also provides the opportunity to reduce labor costs and increase efficiency as the system collects data over time. Technology in the U.S. livestock industry is severely lacking compared to what is available in other industries and on European markets[Berckmans 2014]. Specifically in the sheep industry there is research beginning to be done, however, mainly in European universities. The current use of cameras have the ability to determine sheep behavior such as lying and standing, however, the studies in which this has been done have used expensive sensor-camera combos, infrared cameras, or a Yolov5 model[Buller et al. 2020]. This prompted us to look into the possibility of expanding upon these methods by using a simple RGB camera and some deep learning and machine learning techniques.

Authors' address: Josh Mandzak, jmandzak@vols.utk.edu; Tanner Thornton, tthorn15@vols.utk.edu; Coby White, cwhit152@vols.utk.edu, University of Tennessee, Knoxville, TN, USA.

XXXX-XXXX/2022/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DATA

As part of another research project within the University of Tennessee, Knoxville, sheep have already been given varying levels of endotoxin as part of a controlled experiment and had their activity level of sitting vs standing collected and verified by way of accelerometer. The sheep in this study were assigned to 5 different groups with varying levels of endotoxin, with group 1 receiving the largest concentration of endotoxin and group 5 representing the control group. The animals were first equipped with accelerometers 2 weeks before the start of the study to acclimate the sheep not only to the monitor, but the handling process of the sheep that would take place weekly. Since HOBO monitors are already validated for use in sheep, this data will be used as the ground truth for part 1 of this research (creating a computer vision model to detect sitting vs standing), and used in part 2 (creating a model to detect level of illness based on activity level). Camera data has also already been collected as part of this research, with multiple terabytes of video already stored and ready for analysis. A total of 4 cameras and 8 sheep had results that were used to train the model. During the video processing, a frame was outputted every 30 seconds to match with the HOBO data. Also, because each camera was placed directly overhead and in between two sheep pens, images were cropped in half during processing and each image was outputted to their respective sheep numbers.

3 PROPOSED METHODS

The main component for this project is developing a model. This was chosen as the primary component, as multiple models will have to be created due to the 2 part nature of this work (classifying sitting+standing and classifying sickness). This will require both deep learning (CNNs) and shallow learning (Random Forest, etc.) to be used, which means many models will be created, thus making this the primary focus.

The secondary focus of this project will be evaluation. For the first part, the main aspect of evaluation will be in hyperparameter tuning with the CNNs created, trying to find the best fit for the data. Evaluation will also be used to determine if the model is accurate enough to be used for part 2. Evaluation will be used in part 2 when tuning models or trying out different models to see what gives the best accuracy in detecting level of illness.

4 EXPECTED RESULTS

By then end of this project, multiple things will have been learned. In the topic of research, two questions will hopefully be answered or at least understood in more depth.

1. Can computer vision techniques use low-cost cameras to replace accelerometers in monitoring sheep activity level?

2. Can activity levels of sitting vs standing be used to accurately classify levels of illness within a sheep?

In terms of general technological learning, experience will be gained in working with computer vision topics, namely convolutional neural networks and the frameworks used to implement them. Experience will also be gained in evaluating various models and data preparation when using the data collected to determine sickness level.

5 PROPOSED TIMELINE

Week 1 should be used to ensure data is collected and able to be processed. Week 2 should be used to begin work with stage 1 of the research and begin creating convolutional neural networks to classify images of sheep as either standing or sitting. Week 3 should be used to finish up any work with stage 1, evaluate the results of stage 1, and preprocess data to be used by models in stage 2. Week 4 should be spent creating and evaluating models to perform level of sickness detection in sheep. Week 5 should be used to finish up any work leftover, write the report, and generate plots of data to be used in a presentation and within the report itself.

6 RESULTS

6.1 Part 1: Sit/Stand Detection

In order to best create a model for sitting/standing detection using images, multiple convolutional neural networks (CNN) were created and evaluated in order to create the best possible model for this detection. Ultimately, the goal of the CNNs was to be good enough to replace accelerometers as a tool of collecting this kind of data. The threshold for accuracy needed to replace accelerometers as described by multiple experts in the field was something around ninety percent accuracy.

In an effort to achieve this goal, a grid search was performed where various hyperparameters were tested on separate CNN models to evaluate effectiveness. The hyperparameters tuned in this process were the number of filters used in each convolutional layer, the kernel size of each convolutional layer, and the optimizer used for the model as a whole. The actual values of each hyperparameter can be seen in figure 1. Each model was trained for a maximum of fifty epochs, with early stopping being employed with a patience value of three. Early stopping was used to prevent overfitting, as each model converged to it's optimal level at very different epochs. Tensorboard was also used to track the models as they trained and plot loss and accuracy as a function of epochs.

Number of Convolutional Filters	[32, 16, 8]	[64, 32, 16]	
Kernel Sizes for Convolution	[3, 3, 3]	[2, 2, 2]	
Optimizer	Stochastic Gradient Descent	Adam	RMSprop

Fig. 1. Hyperparameter Configurations

The models themselves were trained on 6,400 images of 8 different sheep, and tested on 1,600 images of the same sheep. More images of the sheep were collected as part of the original research experiment, but the large amount of data proved to be a limitation for local machines. In addition to this, the 8,000 images proved to be more than enough data for the networks to attain desirable accuracy, as discussed later.

The models themselves were composed of three convolutional layers all using relu as the activation function, each followed by a max pooling layer of kernel size (2,2). Finally, the models used a flatten layer to reduce the dimensionality of the output, followed by a dense layer using sixty-four neurons and relu activation. The models finished with a dense layer of a single neuron using the sigmoid activation function. This single neuron allowed the models to be compiled using binary cross entropy loss, with a 1 representing standing and a 0 representing sitting. The best accuracies of each model can be seen in figure 2

Ultimately, model tuning was not especially useful as all models performed far above expectations. The goal of this project was to make a camera model replace the accelerometer model with an ideal accuracy of 90 percent. This accuracy was almost immediately surpassed by all models. There are a few key distinctions to make when examining the accuracies however. In almost all cases, using a kernel size of 3x3 is superior to a kernel size of 2x2, as some the 2x2 models exhibited uncharacteristically low performance metrics. The other key distinction to make can be noted in figures 3 and 4, which show the accuracy/epoch curve and loss/epoch curve for the validation data of each model, respectively. While the colors generated are repeated and thus no label was given to these charts, it can clearly be seen that there are two groups of curves for each image, where one group learns slowly and the other group learns much faster. All of the slow learners use SGD as their optimizer, whereas Adam and RMSprop allow the model to converge in almost half the time.

6.2 Part 2: Illness Detection

As stated previously, the research project conducted where the sheep were actually given levels of endotoxin had already been performed, and ground truth data was provided for these sheep in the form of which sheep received which dosage, as well as accelerometer data collected about each sheep. In order to ensure this illness detection task could be performed with a camera providing the input data and not an accelerometer, only sit/stand data was used from the accelerometer rather than all of the accelerometer measurements. This will allow models created from this research to be used when the input is sit/stand outputs generated by camera data interpretations rather than accelerometers, which is shown in part 1 of this research.

Data for sitting/standing was windowed, with the length of each window being an entire day. This window was chosen as it would be real-world applicable for a sheep owner to receive updates daily about the activity of their sheep. Feature engineering was performed after windowing on the data, where sit/stand numbers were used to generate the following eleven metrics:

(1) Total sits
(2) Total stands
(3) Total changes
(4) Mean
(5) Mode
(6) Longest sit
(7) Longest stand
(8) Average length of sitting bouts
(9) Average length of standing bouts
(10) Standard deviation of sitting bouts
(11) Standard deviation of standing bouts

As this data created numbers of widely different ranges, the data was standardized using a Standard Scaler provided by the library SciKit-Learn.

Evaluation occurred over the course of several models, using a few different input configurations. Six different inputs were used for each model tested. For all models except the last two, there were three potential outputs: not sick, moderately sick, and very sick. The final two models used a binary model instead, opting for not sick and sick. The actual experiment contained five different levels of sickness. The models that use three output labels grouped illness level 1 as not sick, 2 and 3 as moderately sick, and 4 and 5 as very sick. The binary models classified 1 and 2 into not sick, and 3, 4, and 5 as sick. All six model configurations are described below

(1) Using only 6 sheep providing raw sit/stand data as input
(2) Using only 6 sheep providing created features as input
(3) Using all sheep providing raw sit/stand data as input
(4) Using all sheep providing created features as input
(5) Using all sheep providing raw sit/stand data as input and sick/not sick output
(6) Using all sheep providing created features as input and sick/not sick output

Each model was tested using 10-fold cross validation, and confusion matrices and accuracies were reported for each model. Since there were so many models tested, not all confusion matrices are shown in the figures. The accuracies for each of the 6 scenarios are reported in figure 5. As can be noted, the random forest models accounted for 3/6 best accuracies for each scenario, while the SVM models accounted for the three best the random forest did not. The accuracies from the first four scenarios (ones with three outcome labels) show better performance than random, which would be 33 percent, but are overall not incredibly high. This disappointment continues acknowledging that when more data is included, the accuracy actually gets significantly worse. A possible reason for this could be that a single global threshold was used on the accelerometer data of all the sheep to see if it was sitting or standing. Due to how each sheep could have the accelerometer on in a slightly different way, it could be possible that ground truth data is not always accurate. This could potentially throw off the accuracy of the models when learning.

A bright spot in this chart shows that feature engineering improved the accuracy in all scenarios, showing that the features extracted contained useful ideas that help predictive power that correlated with the sickness label.

In addition to the accuracy matrix, the confusion matrices for the best model for each scenario can be viewed in figures 6 and 7 (split into two figures for space reasons).

7 LIMITATIONS

7.1 Part 1: Sit/Stand Detection

The main limitations experienced in this study were time and amount of data. Processing time to convert videos to images and read in these images as NumPy arrays and storing them took a substantial amount of time and space, which in turn limited the amount of time to test for other features. This limitation was avoided by using a

	Convolutional Filters: [32, 16, 8]		Convolutional Filters: [64, 32, 16]	
Optimizer: SGD	95%	92.81%	95.38%	96%
Optimizer: Adam	96.63%	91.75%	97%	95.56%
Optimizer: RMSprop	96%	96.44%	97.12%	96.19%

Note: First accuracy listed uses [3,3,3] for kernel sizes, second uses [2,2,2]

Fig. 2. Part 1 Accuracies

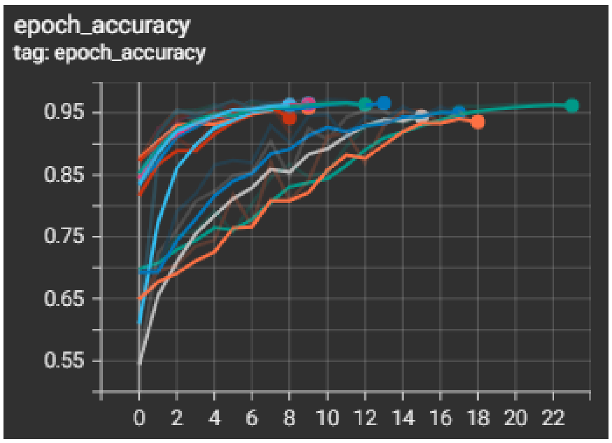

Fig. 3. Part 1 Accuracy/Epoch Curves

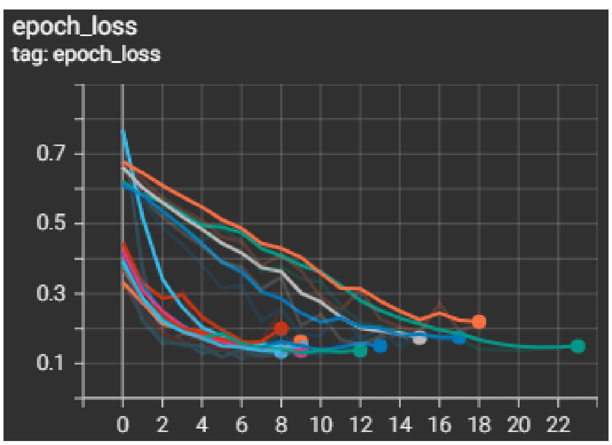

Fig. 4. Part 1 Loss/Epoch Curves

small subset of the dataset to train and evaluate the model. Over two-hundred thousand images could've been used for this model and in a larger experiment with more resources perhaps these images could be used, but for the space and time constraints for this research it was determined 8,000 images would meet the requirements and goals of the research.

7.2 Part 2: Illness Detection

The illness detection models used the accelerometers as ground truth data, and used far more sheep and data points as only a single variable was required to be held (sitting vs standing) so more data could be held by the program in a single run-through. Though more data is usually considered a beneficial factor, some problems did arise from this. One problem that appeared in multiple sheep was that the accelerometer was placed on the sheep upside down, which changed the output of the accelerometer. This highlighted a greater problem, which is that a global threshold for stating whether or not a certain accelerometer value defines sitting or standing is less accurate the more accelerometers used. This is usually due to human error in placing and setting the accelerometer. Unfortunately, this can result in data that was used as the ground truth data to potentially have incorrect values. This hypothesis is supported by the fact that machines were able to learn more and perform better when examining a small subset of sheep rather than all of the sheep used for the experiment. This problem will hopefully be alleviated by switching to cameras to determining sitting or standing, which don't have to manually set (other than initial set up) and are much easier to use.

Another limitation of this experiment is that only sitting or standing is used as an input to perform feature engineering on. There are likely several other aspects of data that can be collected more easily from a camera than accelerometer, such as amount of food eaten, total movement, water consumed, etc. It is hypothesized that including more data such as these examples could provide better results for determining illness level in sheep and providing better illness detection for sheep owners.

8 DISCUSSION

8.1 Part 1: Sit/Stand Detection

The results from this study provide great insight into the usability of cameras for animal research use. In order to follow ethical guidelines set by the animal care and use act, it is important for researchers in animal science to reduce the amount of animals used in studies. Because of this, it is very common in the field of animal science to equip animals (and their surroundings) with accelerometers, cameras, sensors etc. on any research project even if no data will be immediately used[Berckmans 2014]. This allows researchers to reduce the amount of future use of animals to gather the same

	Raw Data Mini Set	Feature Engineered Mini Set	Raw Data Full Set	Feature Engineered Full Set	Raw Data Full Set Binary Sickness	Feature Engineered Full Set Binary Sickness
RandomForest	43.75%	63.04%	39.92%	43.61%	58.05%	64.41%
XGBoost	35.36%	60.00%	35.87%	43.26%	56.34%	60.76%
SVM	41.25%	61.25%	50.66%	50.69%	61.69%	63.09%
KNN	33.39%	61.07%	33.24%	40.91%	45.97%	57.70%

Fig. 5. Part 2 Accuracies

type of data. Unfortunately many sensors used not easily accessible, usable, or durable enough for animal research. For instance, in the case of the HOBO accelerometer a researcher must be able to remove the accelerometer weekly in order to offload the data. An example accelerometer like one used in this research can be seen in figure 8. This not only adds an extra labor requirement for the project but also exposes the animal to a potential stressful situation which could lead to injury.

This also opens the project up to potential instances of human error as seen in this study through the improper placement of the accelerometers. This brings into question results determined using a global threshold during analysis. Unfortunately, this global threshold method is what determined the ground truth data for this research. Given the extremely large amount of data and lack of time, it was infeasible to go through and manually verify each instance of sitting/standing for each sheep. It was deemed good enough that the expected accuracy of the accelerometer is greater than 90 percent.

Though detection of lying and standing behaviors can already be determined using expensive camera sensor combos and using the YOLOv5 model, our model provides a method that can be done using a normal RGB camera using deep learning (CNN) techniques which would both be new methods for this field of study. The model developed in this project allows researchers to gather and analyze the same data that would be gathered from a HOBO accelerometers without the labor and animal stress being problems. It was determined that this task can be performed with any basic camera system and placement of the cameras overhead does not seem to effect the model. This allows for researchers to install the cameras in areas that will be least invasive to the animals, equipment, and other project elements. The camera also does not need to collect data in RGB format, as black and white images were used for the actual training and testing of the model.

8.2 Part 2: Illness Detection

Prior to this study it was hypothesized that it would be very difficult to create an illness detection model based on the amount of data,types of data, and time available for this project. Our hypothesis was confirmed with the results, however, multiple implications for future work could be seen during the training of the model.First, it was expected that the addition of data contributes to improving the model's accuracy. Unfortunately this was not the case, possibly due to human error in the placement of the HOBO monitors. The use of cameras in future studies will hopefully alleviate this problem and allow for better data to be used in the training of the model.

Secondly, the model accuracy was also observed to improve through the use of feature engineering, which implies that careful feature selection could be important in the creation of an accurate model. It could also be assumed that the addition of another behavior parameter (other than lying and standing) could improve the accuracy of the model. Extra behaviors could be gathered fairly easily if switching to a camera based method of input rather than an accelerometer based one. The camera system could also be supplemented by other cameras or sensors if there were behaviors found to be linked with illness level that needed to be detected in some other way. The prospect of a model for illness detection via cameras or sensors is very important for animal agriculture. Not only does this remove the need for extra hired labor, thus reducing labor cost, it also allows for treatment to be given to an animal early in the infection process, which should reduce expenses on long term treatments.

9 FUTURE WORK

The first part of this project presents very little future work to be made. Small improvements could be attempted with accuracy measurements, or model size could be reduced to make the final product even smaller. Finally, some manual verification could be utilized to ensure that the accelerometer ground truth data is correct and true.

The future work of this project mainly centers around the improvement of the Illness detection model. Firstly the addition of more data gathered via a camera system should immediately help improve the model accuracy. However, it is assumed that the model accuracy will still not be an acceptable percent until more behavioral

Fig. 6. Part 2 Confusion Matrices

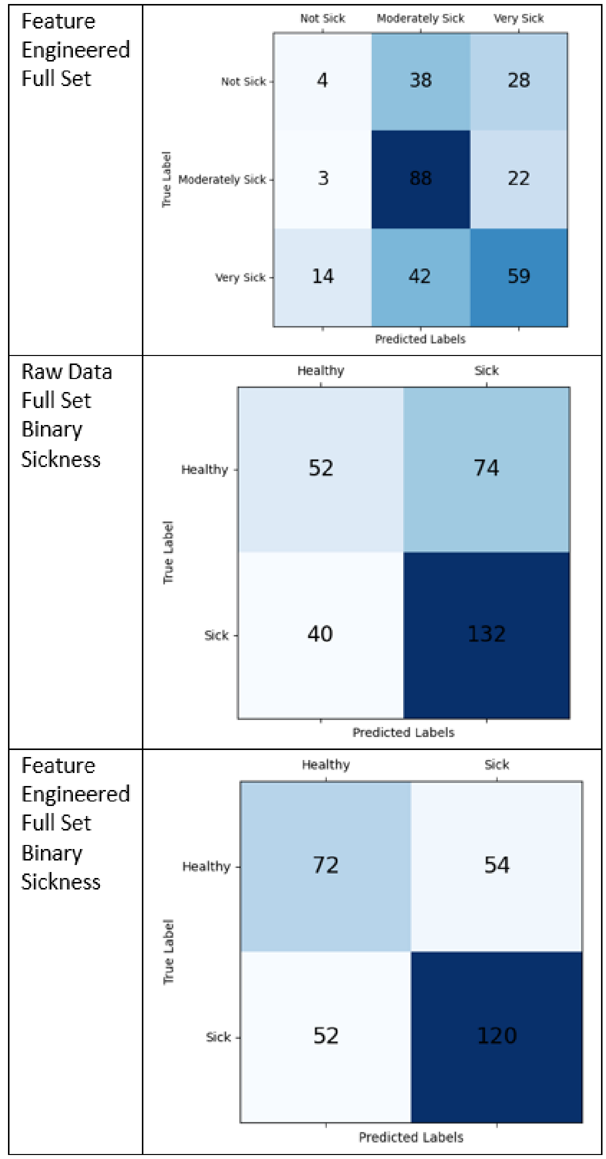

Fig. 7. Part 2 Confusion Matrices cont.

parameters are included. These parameters could include: distance traveled, time spent eating, time spent ruminating, time spent with enrichment items, and many more. One future application for this model could also be used to accomplish recent requirements for farmers to prove animals raised on their farms are raised in a good state of welfare. The European Union has already passed a law concerning this issue and is actively looking into Animal-based measures to base their assessments on. It is reasonable to believe that as this model develops and more behavior parameters are added in, this model could then collect, analyze and provide a welfare score and summary for the farmer. This could then be provided to the government to stay in compliance with guidelines. Overall the future of image processing in agriculture seems to have many applications and can lead to more transparency within the industry, reduced cost on labor and general time spent on farm, increased disease detection thus better disease management, and increased efficiency.

10 CONCLUSION

The results from this study demonstrate the potential for the use of smart technology in animal agriculture, especially cameras. This study has demonstrated the effectiveness of cameras in determining lying and standing in sheep, thus removing the need for accelerometers to collect this data. Though the accuracy for part 2 was not at a usable percentage, we can conclude that through the use of more targeted feature engineering, and with the inclusion of more accurate data, this model can be improved upon to detect illness in sheep. The

Fig. 8. Accelerometer

improvement of this model provides multiple implications including early disease detection/prevention, and a transparent system that can be used to meet the needs of consumers welfare concerns. This conclusion is drawn from the fact that the illness detection models achieved higher scores than random guessing, which shows there is some causal link between sitting/standing and level of illness.

Ultimately, part one of the project with using cameras to replace accelerometer can be deemed an overwhelming success, while part two of the project using sit/stand data to predict level of illness needs more research before it could be effectively used in the field.

REFERENCES

Daniel Berckmans. 2014. Precision livestock farming technologies for welfare management in intensive livestock systems. *Rev. Sci. Tech* 33, 1 (2014), 189–196.

Henry Buller, Harry Blokhuis, Kees Lokhorst, Mathieu Silberberg, and Isabelle Veissier. 2020. Animal welfare management in a digital world. *Animals* 10, 10 (2020), 1779.

James J Hack and Michael E Papka. 2020. The US high-performance computing consortium in the fight against COVID-19. *Computing in Science & Engineering* 22, 6 (2020), 75–80.