Modeling Environmental Conditions Based on Audio in Poultry Houses

Chloe Pomeroy
Georgia Institute of Technology
cpomeroy6@gatech.edu

Christopher Sniffen
Georgia Institute of Technology
csniffen3@gatech.edu

Abstract

Large-scale poultry farms often have difficulty manually monitoring the health of their livestock due to the volume of animals. AudioT attempts to address this by evaluating chicken health in near real time from audio data collected in the poultry houses. It is hoped that by quickly detecting changes in chicken behavior, AudioT can help farmers optimize their operations and revenue while improving the treatment of the poultry themselves. As one step towards this goal, AudioT asked the team to build on prior work from others to develop a model that can accurately discern between night and day using AudioT's proprietary dataset. The team explored, labelled, and modeled a large subset of the AudioT data using several different approaches, substantially advancing AudioT's ability to model night and day from raw audio files and setting new performance benchmarks for this classification task.

1. Introduction and Motivation

In large-scale farming, the number of animals involved makes it impossible for a farmer to manually monitor the health of their livestock, even with hired help. This issue is particularly acute in poultry farming where a single poultry house can house 24,000 chickens, and a single farmer might operate a handful of houses.

AudioT aims to solve this problem by recording and analyzing the audio in the chicken houses, specifically focusing on the vocalizations of the birds. The idea is that anomalies in the bird vocalizations can help to inform the farmers when the flock is unhealthy. Since chickens make different vocalizations as they age, they aim to create different anomaly detection models for different growth stages of the flock. Being able to accurately identify issues with the flock in near real-time would not only benefit the efficiency and revenue of the farm but also the treatment of the chickens themselves.

The main obstacle AudioT has experienced in their work with anomaly detection models is that noisy farming equipment such as ventilation fans, the feeding auger, and other equipment is being picked up in the audio data and being identified as anomalous. To remedy this, AudioT's planned approach is to somehow identify these farming machinery noises and create different noise removal models for each type of equipment, so that the bird vocalizations are not overwhelmed by the background noise. As a step in this direction, the team was asked to attempt to model the day and night transitions, since it's much easier to label and identify these transitions manually in the AudioT data. It is hoped that the same methods we develop for modeling day and night will transfer to more specialized models that combine day and night states with equipment states, such as an active ventilation fan in the background. This is the area the practicum team has been working on for the semester.

2. Background

2.1. Data

The AudioT data is collected from active large-scale poultry houses by agreement with the owners. For each monitored poultry house, AudioT collects lossless audio from a series of 15 microphones hung from the irrigation piping at regular intervals through the house. These condenser microphones are connected to hardened RaspberryPi units and collect data continuously, storing the audio in one minute increments as .flac files. The recordings are then uploaded to cloud storage with a naming convention that encodes the relative location of the microphone within the house (i.e. 'microphone six') and the time of recording. For this project, the team was instructed to focus on data from one partner organization, which was collected between 2021 and 2024 in a variety of different houses.

2.2. Domain Knowledge

This project required the team to develop a working knowledge of poultry farming, specifically large-scale broiler farming. AudioT provided a wealth of information on how chickens develop, key environmental variables that can impact chicken health, the types of equipment used to regulate large-scale chicken houses, and even some of the economic considerations for the AudioT product. In addi-

tion, the team had to get familiar with the audio processing methods used by AudioT and the limitations of the audio recording equipment installed in the chicken houses. Since AudioT stores its data in the Amazon Web Services cloud, we used the AWS command line interface and console to interact with the raw audio files. AudioT recommended we use the open source Audacity software for listening to audio files, and we were encouraged to perform our analysis and experimentation in Jupyter notebooks. These tools enabled us to easily build upon the work of prior teams and made it easy for us to replicate earlier results.

2.3. Prior Work

In the previous semester, other OMSA students had worked on creating a Hidden Markov Model to predict night and day based on the bird vocalizations. Night and day in this context does not rely on the outside environment, but rather is artificial based on the farm and how they configure the lights within the poultry houses. Using just one microphone of data, the prior team was able to accurately predict night and day within 99% accuracy for the latter half of the flock's development, which allowed AudioT to train different anomaly detection models based on the current state, since vocalizations differ so drastically between night and day. We were asked to extend this initial work to a more general set of conditions, such as other microphones, flocks, and houses.

3. Exploratory Data Analysis

AudioT provided the team with access to over ten terabytes of audio data in cloud storage from just one partner organization as described above. Since the audio files are not labeled beyond the date and time of recording, one of the first tasks for the team was to identify useful data among the massive set of data available. Since the team did not have access to cloud compute, this process involved downloading audio to a local machine or processing the audio in memory so as to visually evaluate the data for consistency. The average broiler chicken develops over approximately seven weeks, and AudioT the team was able to quickly identify a few key flocks of data to work with. It was immediately apparent that the data contained periods of equipment failure, missing data, and inconsistent recording across the multiple mics in a given house.

Exploratory analysis on the data generally took one of three forms. First, we might simply download and listen to the audio files. However, since a single mic for a single flock comprises over one thousand hours of audio, this was not practical for high level exploratory analysis, it was generally more useful when examining outliers or anomalies in the data. Second, we made use of a web application provided by AudioT that provided a visualization of an entire flock at a time in the form of a two dimensional heatmap.

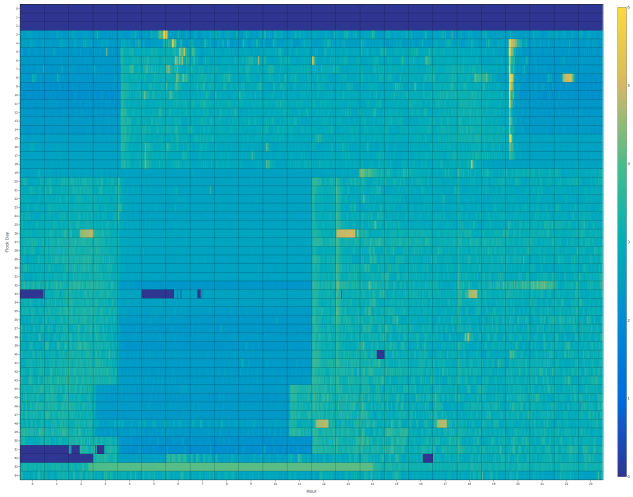

Figure 1. The AudioT anomaly dashboard showing audio dropouts (dark blue) and the day-night pattern for the full lifespan of a flock raised in 2023. Each row of the image is audio, by minute, for a day. The grid columns separate the audio by hour.

This heatmap allowed the team to quickly identify bad data (mic dropouts) and get a sense of the day night trend. Figure 1 shows a screenshot of the web application with a subset of one microphone from one flock raised in 2023. It's easy to identify microphone malfunctions as the dark blue, and the day/night pattern is also clear, with one week that shifts the daylight hours by an hour. This inconsistency in daytime hours stems from the fact that the poultry houses are artificially illuminated, with the 'daylight' hours controlled by the farmer. Whether intentionally or as a result of equipment malfunction, we found that in many cases the daylight schedule shifted for at least part of any given flock's data, complicating efforts to label the data accurately.

The third and most common method for exploratory analysis of the AudioT data was to process each minute of audio into a set of thirteen numerical features using AudioT's proprietary library. These features could then be organized, visualized, and evaluated at the flock level. It was this form of exploration that helped the team most in considering the limitations of the data and potential modeling approaches. In figure 2 two full flocks of data are depicted, with all thirteen generated features. While the day and night pattern is visible to the human eye in both sets of data, the range of values for the features is significantly different. In both flocks there appear to be three successive stages that the flock goes through over the course of its development, with the first stage being especially chaotic. The feature values during the middle stage appear to have less variance, while the final stage has the variance of the first stage with a more sharply defined structure. In data from other flocks, these three stages were less sharply marked. Note that in the 2023 data depicted, there is roughly a week of data missing.

Figure 2. Visualization of Two Full Flocks.

4. Approach & Methodology

Our team felt strongly that this use case would lend itself well to a classification model, as opposed to the Hidden Markov Model (HMM) approach that our predecessors took. To us, it seemed that the states we were predicting were deterministic rather than probabilistic, which HMMs excel at. We unfortunately did not have access to labelled data to try classifying the different types of farming equipment, so we pursued modeling night and day as two classes with a variety of classification techniques including Support Vector Machines, Naive Bayes Classifiers, and Multi-Layer Perceptron Classifiers.

We were also interested in feature engineering or adding additional features to the data unrelated to the actual audio, such as ambient temperature and humidity. However, temperature and humidity data corresponding with our audio data were not available for use.

4.1. Feature Generation

Based on our observations from the exploratory analysis, we selected eleven microphones covering three flocks raised in three different years as our base dataset to use for modeling and evaluation. The data selected is described in table 1. Since not all of the data we used was available in the AudioT web application, we generated the full thirteen features from each minute of audio for all microphones and all flocks using AudioT's proprietary library, and then generated visualizations to give us context on the various flocks.

Year	Mic Pos	Begin	End	Obs (Min)
2021	04	11 May	22 Jun	60k
2023	05	19 Nov	31 Dec	60k
2023	12	19 Nov	31 Dec	60k
2023	13	19 Nov	31 Dec	60k
2023	14	19 Nov	31 Dec	60k
2024	00	01 Feb	06 Mar	50k
2024	00	25 Apr	20 Jun	80k
2024	01	25 Apr	20 Jun	80k
2024	02	25 Apr	20 Jun	80k
2024	09	25 Apr	20 Jun	80k
2024	10	25 Apr	20 Jun	80k
2024	11	25 Apr	20 Jun	80k

Table 1. Subset of AudioT Data Selected for Modeling

4.2. Using Predicted HMM States as Labels

In this project, the main obstacle we faced was attempting to label the data accurately. As noted earlier, we didn't have access to labels for the various farming equipment we were looking to classify. We also lacked ground truth samples for the different types of equipment. This meant that when we found a 'new' sound, we had to work with our AudioT colleagues to try to identify what the source of the sound was. This was both time-consuming and inconsistent.

In order to avoid the data labelling issues, we attempted to build on prior teams unsupervised work by using the state predictions from their HMM as labels for three states, and

then use these labels to train a classifier. The rationale behind this was that the previous group's models had been very successful for modelling day and night, and so they had attempted to extend their work to large numbers of states that they hoped would further distinguish between different noise states. However, when they increased the number of states there was no real way to verify if these were successful or not, due to a lack of a ground truth labels for any states beyond night and day. Training a separate classifier allowed us to analyze whether the extra state identified by the Hidden Markov Model was identifying a real pattern in the data that the classifier could also identify.

To do this, we first trained a Hidden Markov Model using the previous semester's code for the 2024-04 flock, and produced predictions for the entirety of the flock as well as for the entirety of the 2024-02 flock. For both of these flocks we used the same house and microphone 00 to minimize any variation due to differences in microphone placement. Next, we created a labelled dataset using the predictions for the 2024-04 flock as the training set, and another labelled dataset using the predictions for the 2024-02 flock as the testing set. Using these datasets, we trained a 3-class XGBoost model and evaluated both the training and testing accuracy, which can be seen in Table 2.

		Testing	
		2024-02	2024-04
Training	2024-02	0.856	0.369
	2024-04	0.5911	0.855

Table 2. XGBoost training and testing accuracies using HMM states as labels

It's clear from these results that the HMM states between flocks may not generalize well, although we can't be completely certain unless we compare with additional data, since these results could reflect anomalies unique to these particular flocks. Additionally, we can see in Figure 4 that though the classifier output for the training flock has a similar distribution as the ground truth, the testing flock has a much different distribution with no observations predicted as class one. This same pattern can be seen in the testing flock no matter which flock is used for training. This implies that the extra state from the HMM (class 1) is not transferable between the two flocks. Further, although it does indicate some pattern that we are able to identify with the classifier, the pattern identified as 'Class 1' in the HMM model trained on the 2024-04 flock does not correspond to the pattern identified as 'Class 1' in the HMM model trained on the 2024-02 flock. This was one of our fundamental worries about using Hidden Markov models for this purpose; that since the states are hidden, there is no guarantee the states we predict will be capturing the same patterns and events between flocks.

4.3. Data Labelling for Day/night Classification

To extend our work with classifiers, we decided to implement a separate data labelling strategy for day/night classification. We used the visualizations of the generated features to guide us in labelling the data with a binary 'daytime' feature containing a value of one for daytime and zero for nighttime. In all cases, we were forced to remove data where either the day/night shifts were inconsistent or it was unclear where the shifts were. Figure 3 depicts roughly two weeks of data of two features from a single microphone overlaid with the labels for that data. The purple line represents the labels (ten is day and two is night for ease of viewing). The vertical black line in the middle separates the early data that does not clearly show day/night structure, from the subsequent data where the pattern is more clear. We generally found that audio files from the first week of a flock's development did not show a discernible day/night pattern, and our AudioT colleagues confirmed that the first week of a flock's development is notoriously hard to model.

4.4. Multi-Layered Perceptron Classifiers

As indicated earlier, the team felt that treating the day/night prediction task as a classification task might work better than the HMM approach. In order to test our theories, we evaluated two classification approaches: Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP) and see whether we could improve on the accuracy of the HMM's unsupervised approach. Once we had labeled the data as described above, we built a training and test set from the 2021 and 2023 data, scaled the features, and performed some initial modeling without parameter tuning to baseline the approach. For this initial modeling, we used a random-

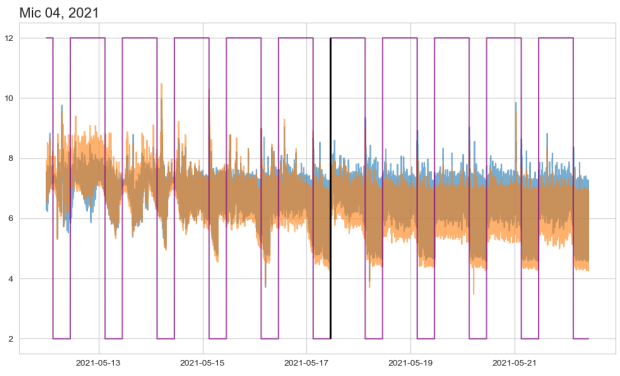

Figure 3. Two weeks of labelled data, (single mic, two features, 2021 flock).

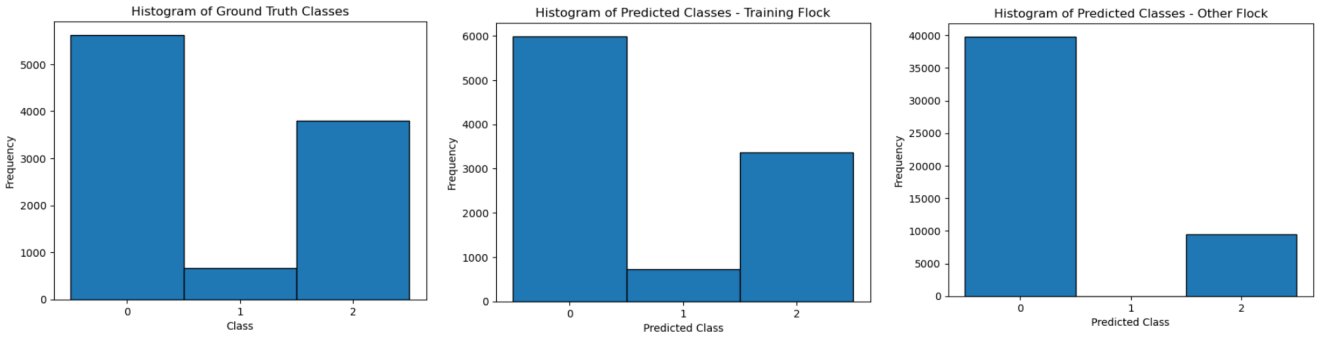

Figure 4. Histograms showing the class distribution for the XGBoost predictions, where the training flock is 2024-04 and testing flock is 2024-02

ized 70-30 training-test split over just two microphones of data from the two separate years.

The initial results were encouraging with both models scoring above ninety percent accuracy as depicted in the confusion matrices in figure 6. The MLP consistently scored several points higher, even with more careful tuning and additional data, so we eventually dropped the SVM approach to focus on improving the accuracy of the MLP classifier. While we were able to achieve accuracy scores over 98% with multiple hidden layers, we found that a properly parameterized MLP classifier with a single hidden layer of 100 nodes could easily reach 96% accuracy, with less chance of overfitting.

An early concern with the MLP approach was that given the limited labeled data available, the model would overfit one flock at the expense of the other when multiple flocks were represented in the training and test dataset. However, with two flocks in the training data, we found that the accuracy was very similar for randomized test data from both flocks 0.969 over 15,367 samples for the 2021 data and 0.972 over 10,857 samples for the 2023 data. Our other main concern was whether the models would generalize to data from a completely different flock. This concern ended up being very valid. Even after careful hyperparameter tuning and including all of the available data from the 2021 and 2023 flocks, our MLP's performance against the 2024 data ranged from 0.519 to 0.850, depending on the microphone tested, for an MLP with an accuracy score of 0.979 overall against the held out test set from the original two flocks.

We were curious as to whether including the 2024 flock in the training data would allow it to continue to perform accurately against all three flocks. Indeed, as soon as we added 2024 data to the training set, the accuracy for data from the microphones the model had seen jumped significantly. Prior to including the 2024 data in the training set, overall model accuracy on microphone nine from the 2024 flock was a dismal 37.3%. After including the data in the training set, the overall accuracy jumped to 94.9% as seen in 7. Interestingly though, we found that the accuracy was

Microphone	Accuracy (no 2024 data)	Accuracy (with 2024 Data)
0	77.3	73.9
1	76.9	68.4
2	84.2	87.3
9*	37.3	95.0
10*	85.0	95.2
11*	56.4	93.6

Figure 5. Accuracy scores for microphones in the 2024 flock, model trained with and without 2024 data. Microphones with an asterisk (*) were included in the 2024 training data.

still much lower for 2024 microphones that the model had never seen, as shown in 5. In our opinion, this indicates that the generalization problem was not necessarily about the differences between flocks, it appears to be due to either the placement or behavior of different microphones. We hypothesize that the model is very sensitive to the relative amplitude (volume) of sound from machinery, which is highly dependent on the distance of the machinery from the microphone. Even in very similar poultry houses, we expect that small differences in the placement of microphones or the location of fans and feed augers make it hard to generalize the high level patterns we're trying to model.

In addition to testing the performance and generalization against the thirteen standard features that are generated by AudioT's proprietary library, we experimented with adding engineered features to see whether they improved performance. Some of this testing was directly at the request of AudioT, other tests were simply experiments that we came up with as a team. Tests included generating rolling one hour minimum and maximum values for features, which we had hoped might help the models cope with drastically different amplitudes among different microphones. We also tested cyclical values to help the model 'learn' a 24 hour day. In that case, we took the timestamps and generated

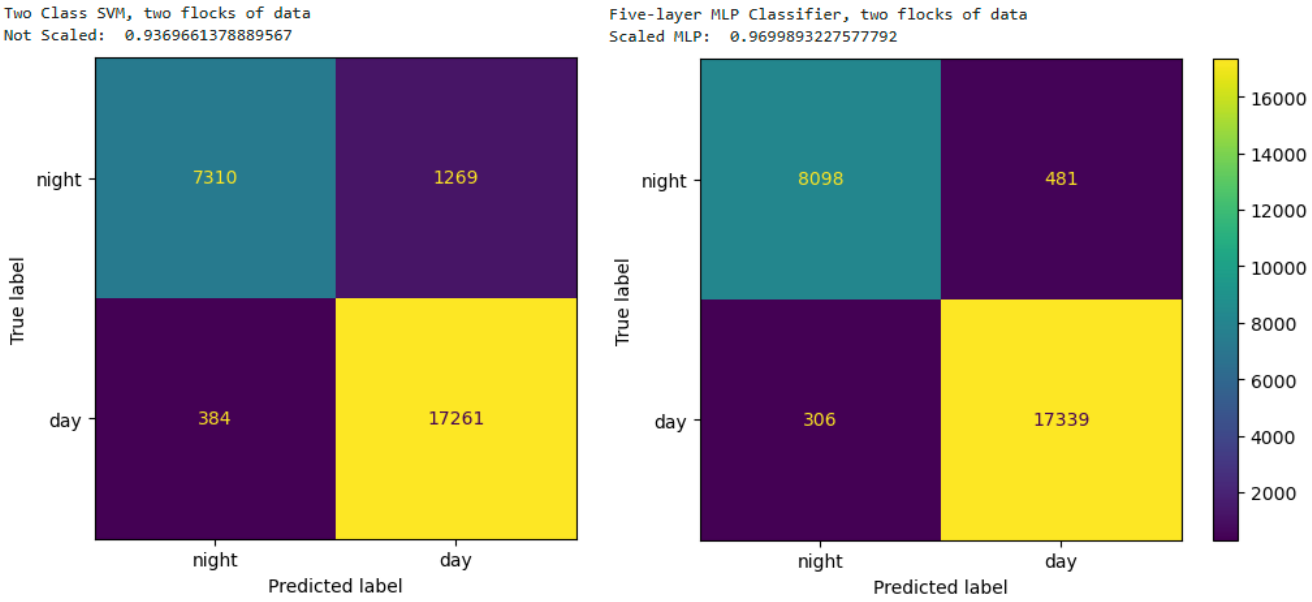

Figure 6. Confusion Matrices and Accuracy Scores for Initial SVM and MLP Models on Two Microphones from 2021 and 2023.

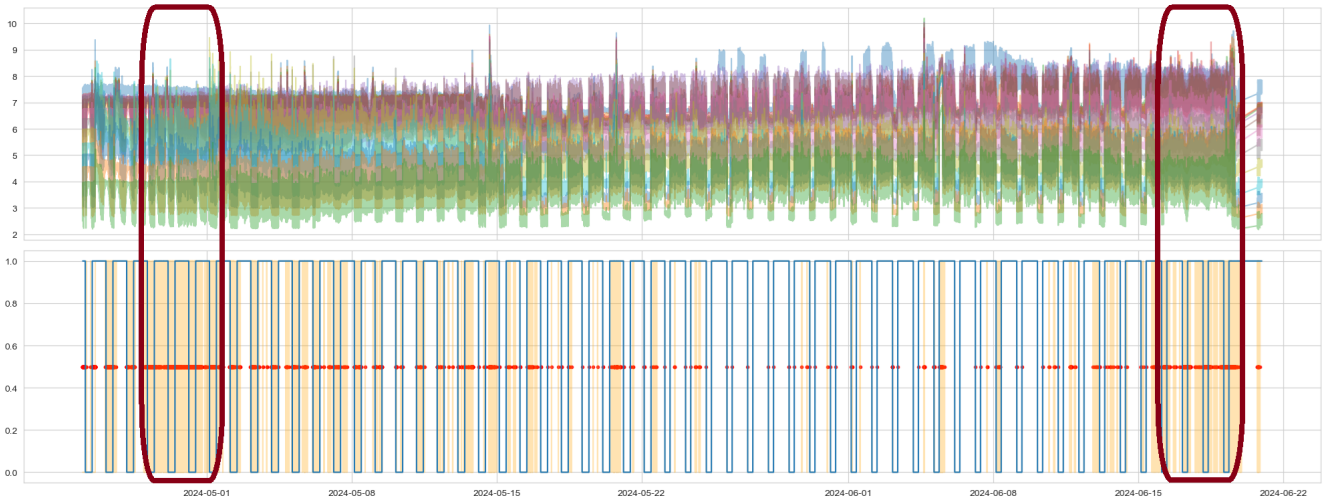

Figure 7. Evaluation of model on microphone nine from the 2024 flock after inclusion in the training data. The top graphic depicts the 13 features from a single microphone, the bottom picture depicts the day/night 'ground truth' in blue, the predicted values in orange, and the errors (points at which the blue and orange values differ) in red. Overall accuracy for the model on this microphone prior to including 2024 data in training was 37.3%, after inclusion the overall accuracy jumped to 94.9%. The maroon rounded rectangles highlight the higher error rates at times when the day/night shift is irregular.

a repeating half-sine wave covering 24 hours. Finally, we gave the model access to a rolling window of feature data comprising the past three minutes of feature values for any given observation. In none of these cases did this feature engineering substantially improve either the classifier performance or the ability of the classifier to generalize among different microphones and flocks.

4.5. Using PCA to Incorporate Multiple Mics

Since the biggest hurdle with our models seemed to be generalizing between flocks, we started to investigate using Principal Components Analysis (PCA) to allow us to incorporate multiple microphones into a single set of features. Our previous models had been very dependent on which microphones we used due to the amplitude of the sounds being an important factor in prediction. Because of this, we thought it would be useful to take in the features from

all microphones for a given flock, and then use PCA to reduce the dimensions back down, which would allow us to get a full picture of all sounds captured in the house. Then we could use these principal components to train a classification model and determine if we get better generalization performance.

First, we took 4 microphones for the 04-2024 flock in the TRF0 house (Mic0, Mic1, Mic9, Mic10) and created various PCA models with the combinations shown in Table 3.

	Mic00	Mic01	Mic09	Mic10
Recorder 0	x	x		
Recorder 1			x	x
Mic 0 and 9	x		x	
Mic 1 and 10		x		x
All	x	x	x	x

Table 3. PCA Models and which mics were used to train them

We decided to utilize the first three principal components which captured between 96% and 98% of the variance depending on the model. Looking at these principal components over time for weeks 2-3 of this flock's development, we can see that both of the mixed models (Mic 0 and 9 and Mic 1 and 10) have very similar patterns, with the Recorder models having the most differences from each other. This implies that there is a significant difference between the data from different areas of the house, since Recorder 0 is near the front and Recorder 1 is further back in the house.

Since the principal components for both of the mixed models are so similar, we may also be able to conclude that we don't need to have data from all the mics, as long as we have data from a variety of areas of the house. This is an important finding since an issue that AudioT often runs into is malfunctioning microphones, but we've shown that the variance can still accurately be captured if you just have one working microphone from each recorder. Further work would need to be done to confirm this, but we were limited in this analysis due to the computation time needed to generate features for more microphones and flocks. We also had wanted to try using these PCA datasets to train classification models, but would have been unable to verify the generalization issue due to not having data from a variety of mics for another flock to test against. This would be an interesting area for further research.

5. Challenges and Limitations

We experienced two main challenges with this project: the lack of labelled data and the amount of computation time needed to generate the features. In addition, there were some minor things we would have liked to have had, such as temperature and humidity data at similar granularity to the audio data, a diagram describing the physical placement of the microphones and machinery in the poultry houses, and clean audio samples of the farming equipment noises that AudioT wants to be able to model.

Since we were interested in using classification models for this problem, the lack of labelled data was a significant impediment for us. We worked around this by focusing on night/day classification, experimenting with forms of automatically labelling, such as using the HMM predictions, and manually listening to audio files and observing their wave forms. With the focus shifting from HMMs to classification and with the other team who worked alongside us focusing on identifying the farming equipment and labelling the data by doing deeper analysis of the waveforms, we think AudioT is in a good place to create labelled datasets to take this work further.

Although previous teams had worked on this project before and generated the features, these were not accessible to us and so when starting the project we were forced to spend time generating these features ourselves. We could do this using AudioT's proprietary code, so it wasn't difficult so much as computationally intensive - simply generating the features for a single microphone and a single flock could take 24 hours or more depending on the computers processing power and internet service. As such, we were limited with the amount of data we could actually use in our modelling experiments, even though we had access to a huge amount of valuable audio data. We would have loved to have been able to have features from all 15 mics for 3 flocks, so that we could test whether the additional information gained from the full set of microphones would improve the model's ability to generalize to other datasets. To prevent this from continuing to be a limitation for future students working with AudioT, we've uploaded the features for all the flocks and microphones we generated to an S3 bucket in their AWS environment, which will be accessible for others to use without needing to process the data themselves.

6. Conclusion

Ultimately, the MLP classifier with a single hidden layer of 100 notes and a rolling window of three minutes for all 13 features (39 total features for each observation) was the best performing and most adaptable approach among the methods we tested. The main drawback of this model is the fact that it does not generalize well to other flocks, though further testing with additional data is necessary to clarify exactly how limited the model is. We feel that a far larger and more general training set that mixes in multiple flocks, houses, and microphones will be necessary to build and thoroughly evaluate a robust and general model. However, we feel confident that our limited experiments have proven the viability of MLP classifiers for achieving Au-

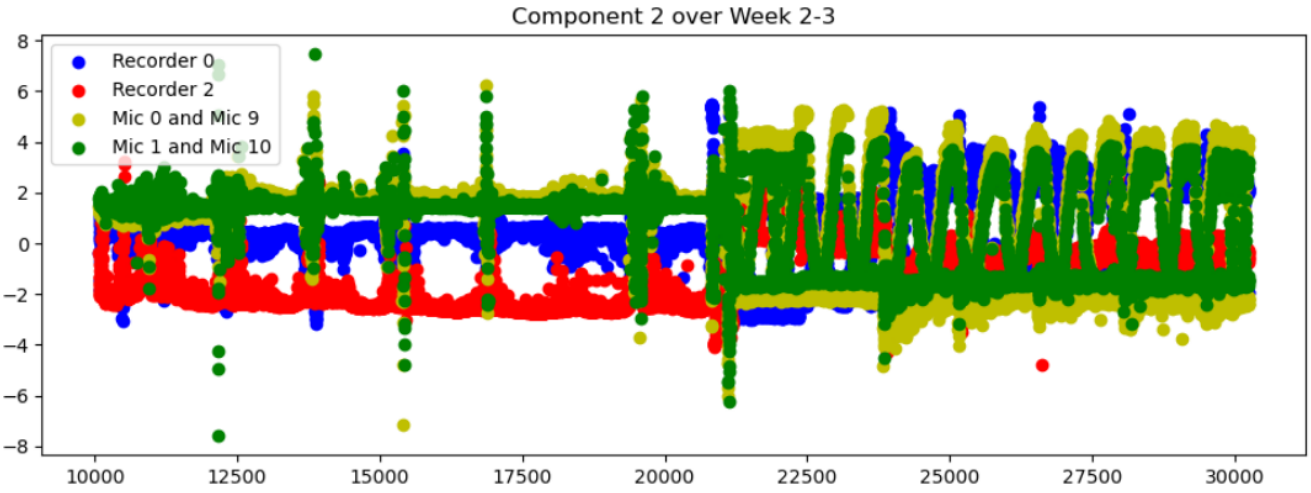

Figure 8. This figure shows the second principal component for the various PCA Models. We can see that the two mixed models are very similar whereas the two recorders are the most different.

dioT's goals. Moving beyond day and night to model microphone dropouts and machine noises will take a more intensive data labelling effort, but it appears to be an achievable set of goals. While labelling would necessarily be an expensive task, we feel the benefits for AudioT make it worthwhile.

Beyond improving the quality and quantity of data, we see many additional avenues for future work, such as furthering the PCA work and incorporating this into the classification models. This appears to be a promising method for working with equipment malfunctions and could ultimately be a path towards better generalization as well. Working with dimensions beyond audio such as temperature or humidity could allow more precise labelling and help tune the anomaly detection AudioT is seeking to create by ensuring that the audio patterns actually correspond to environmental changes. Finally, if all fifteen microphones for a single house are available, treating the thirteen features for fifteen microphones as a single dataset may provide a much more comprehensive and detailed picture than working with microphones piecemeal, as we have done. We leave these ideas and opportunities for future work.

7. Work Division

All team members contributed equally to this project. The work division table can be seen in figure 9.

Task	Description	Team Member Contributions
Downloading and extracting features	Running feature extraction for a flock (~2 months worth of minute-by-minute data)	Both Chloe and Chris extracted features for 2 flocks each. Chloe identified which data to use based on microphones had the most complete data.
Reproducing previous experiments	Trying to replicate the previous semester's work of identifying clear night/day states using our extracted data	Both Chloe and Chris replicated previous experiments with data from different flocks/microphones. Chris validated that the models overfit and do not generalize well and Chloe looked at dealing with missing data.
Alternative Models	Testing alternative model approaches	Chris worked on creating a proof of concept with a small neural network and Chloe worked with supervised HMMs
Feature Engineering	Experimenting with the features	Chloe experimented with adding data from multiple mics and using PCA for dimensionality reduction
Model Engineering	Experimenting with different numbers of states and hyperparameter tuning	Chris experimented with engineered features and techniques for making his MLP generalize better to new flocks.
Midterm Report	Creating the powerpoint presentation	Chris created the presentation and all the initial slides. Chloe looked over these and made a few content changes, as well as adding the contributions table.
Final Report	Creating the final report	Chloe and Chris both wrote the report, each contributing their own thoughts, experiments and findings.

Figure 9. Distribution of work