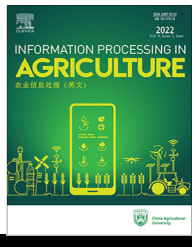
Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE xxx (xxxx) xxx

journal homepage: www.keaipublishing.com/en/journals/information-processing-in-agriculture/



Deep learning based classification of sheep behaviour from accelerometer data with imbalance

Kirk E. Turner [a,b], *Andrew Thompson* [c], *Ian Harris* [d], *Mark Ferguson* [d], *Ferdous Sohel* [a,b,*]

[a] *Discipline of Information Technology, Murdoch University, Murdoch, WA 6150, Australia*
[b] *Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, WA 6150, Australia*
[c] *College of Science, Health, Engineering and Education, Murdoch University, Murdoch, WA 6150, Australia*
[d] *neXtgen Agri, Saint Martins, Christchurch 8022, New Zealand*

ARTICLE INFO

Article history:
Received 2 September 2021
Received in revised form
29 March 2022
Accepted 12 April 2022
Available online xxxx

Keywords:
Sheep behaviour classification
Data synthesis
Class imbalance
Grazing sheep

ABSTRACT

Classification of sheep behaviour from a sequence of tri-axial accelerometer data has the potential to enhance sheep management. Sheep behaviour is inherently imbalanced (e.g., more *ruminating* than *walking*) resulting in underperforming classification for the minority activities which hold importance. Existing works have not addressed class imbalance and use traditional machine learning techniques, e.g., Random Forest (RF). We investigated Deep Learning (DL) models, namely, Long Short Term Memory (LSTM) and Bidirectional LSTM (BLSTM), appropriate for sequential data, from imbalanced data. Two data sets were collected in normal grazing conditions using jaw-mounted and ear-mounted sensors. Novel to this study, alongside typical single classes, e.g., *walking*, depending on the behaviours, data samples were labelled with compound classes, e.g., *walking_grazing*. The number of steps a sheep performed in the observed 10 s time window was also recorded and incorporated in the models. We designed several multi-class classification studies with imbalance being addressed using synthetic data. DL models achieved superior performance to traditional ML models, especially with augmented data (e.g., 4-Class + Steps: LSTM 88.0%, RF 82.5%). DL methods showed superior generalisability on unseen sheep (i.e., F1-score: BLSTM 0.84, LSTM 0.83, RF 0.65). LSTM, BLSTM and RF achieved sub-millisecond average inference time, making them suitable for real-time applications. The results demonstrate the effectiveness of DL models for sheep behaviour classification in grazing conditions. The results also demonstrate the DL techniques can generalise across different sheep. The study presents a strong foundation of the development of such models for real-time animal monitoring.

1. Introduction

Monitoring the behaviour patterns of individual sheep has the potential to inform multiple areas of animal production, from profiling physiological state, to grazing management practices [1,2]. However, monitoring sheep behaviour has chal-

* Corresponding author at: Discipline of Information Technology, Murdoch University, 90 South Street, Murdoch, WA 6150, Australia.
E-mail address: F.Sohel@murdoch.edu.au (F. Sohel).

Peer review under responsibility of China Agricultural University.
https://doi.org/10.1016/j.inpa.2022.04.001

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

lenges due to the vast spatial area and the substantial number of animals involved. The classification of sheep behaviour through observation is not a highly technical skill, but it is a labour intensive one, and cannot be practically performed at scale. The use of sensor technology can overcome this gap to provide real-time information about the productivity, health, and well-being of the animals. Monitoring sheep behaviour through tri-axial accelerometers provides a cost and power-efficient method for performing the monitoring, aided by the recent miniaturisation of the technology. Taking advantage of these recent advances, the sheep industry is looking to apply this to "optimise production, reduce costs and enhance sustainability" [3].

Accelerometers measure gravitational and inertial acceleration due to movement. They collect data across three axes and each dimension is recorded simultaneously, representing three-dimensional movement in the sampled data [4]. One method taking advantage of the data generated from accelerometers is the use of machine learning to classify the behaviours of the sheep based upon the accelerometer data. However, the balance of the data influences the classification process as the imbalances in the data are learned through the training process. Recent studies have shown the class imbalance issues when evaluating sheep behaviour classification using accelerometer data. Fogarty et al. reported low recall (the proportion of correctly predicted classifications) for *walking* behaviours (65.6% for *walking* versus 90.3% for *grazing*) [3], while Barwick et al. reported poor recall for *lying* behaviours (6% for *lying* versus 88% for *grazing*) [5]. While there are other contributing factors, such as similarity between behaviours, in both Fogarty et al. and Barwick et al., the authors attribute the results to data set imbalance [3,5]. Importantly, often the under-represented behaviours hold significance. *Grazing* and *ruminating* are considered the most important behaviours for ruminants [6]. However, in our study *ruminating* only made up 8.6% of the samples. Ideally, the imbalance could be addressed at the data acquisition stage. However, as alluded to in Fogarty et al. [3], the large areas, difficult terrain and the limited time spent expressing certain behaviours (e.g., *walking*), compared to *grazing* and *resting*, leads to natural imbalance in the data. Therefore, alternative methods are needed to overcome the class imbalance present in the data.

A mix of statistical [7] and machine learning models have been used for classification of sheep behaviours. The machine learning models include ensemble learning methods [8-10], decision tree algorithms [2,3,11], instance-based algorithms [3,9-11], as well as dimensionality reduction algorithms [3,5,9,12,13]. Although Deep Learning (DL) techniques have been successfully applied in other applications, there is a lack of studies making use of DL. Studies have shown that DL can produce better results given the correct models and data sets for time series data (which includes accelerometer data) [14,15]. In contrast to classification of sheep behaviour studies, DL is more prevalent in the human activity recognition (HAR) field, which also makes extensive use of accelerometer data. There is significant research making use of Long Short Term Memory (LSTM), a Recurrent Neural Network (RNN) implementation, shown to work well with time series data [16]. Additionally, similar, or related classes are predicted with higher accuracy when a bidirectional LSTM is used [17]. Sheep classification have similar behaviours that are harder to distinguish (e.g., *lying* and other behaviours [5]). Additionally, transitions between states occur within the same time window, resulting in mixed signals within the same sample. The transition directions are not specified in the data. Therefore, the bidirectional nature will aid in the correct classification of the behaviours displaying transitions. Feature selection in the sheep behaviour studies has been a mixture of statistical [8-12] techniques, but limited to Random Forest (RF) in terms of Machine Learning methods [2,3,13]. DL has not featured among the methods, but studies have shown successful application of Convolutional Neural Networks (CNN) in the context of LSTM classification models [18-20].

There are various forms of class imbalance, either in the number of instances, or the density of the instances, but these differences result in the machine learning techniques overfitting the majority classes and densities [21]. The class imbalance issue can be addressed at the separate phases of the classification pipeline: feature selection, classification and in the data preparation phases. In this study, the aim was to compare classifiers in the context of the class imbalance, and the influence of synthetic data on the classifiers. Therefore, the focus is on addressing the class imbalance in the data preparation phase. One form of oversampling (creating new samples) is the generation of synthetic data. Rather than reusing the existing data in duplicate, the feature space of the data is used to generate data samples that match the space. There are multiple mechanisms for performing the generation of the synthetic data, but one of the most popular is Synthetic Minority Oversampling Techniques (SMOTE) [22,23].

SMOTE, drawing inspiration from techniques applied to handwritten character recognition, applies changes to the feature space rather than the data space. Examples are generated by interpolating between several minority class examples that are nearest neighbours [22,24]. There are over 100 variants of SMOTE, with differences in how the new samples are distributed [23]. Polynom-fit-SMOTE [25] is based on curve fitting methods that find the coefficients of a polynomial that fit the minority instances with different topology options, such as 'star', 'polynomial', 'bus' and 'mesh'. These generate samples that are relatively far apart and therefore the synthetic data is more scattered within the decision boundaries of the minority class [21]. Kova´cs [21] performed an empirical comparison of 85 different variants against 104 imbalanced data sets, concluding that the polynom-fit-SMOTE was the best performer for an unseen set of data showing imbalance issues. Baseline SMOTE has been used to address class imbalance issues for many instances of accelerometer data [26-29], but as noted by Kova´cs [21] many comparisons for data synthesis are performed against baseline SMOTE where advances have been made with other variants. Therefore, our comparison will focus on polynom-fit-SMOTE.

The objectives of this study were to evaluate (i) the addition of synthetic data and (ii) DL techniques to the classification of sheep behaviour. Automating sheep behaviour classification is the first step in improving productivity through physiological state profiling. However, class imbalance, inherent from natural behaviour bias and sampling

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

techniques, reduces the accuracy of behaviour classification as the machine learning algorithms inherit the bias from the data set. An evaluation of the literature shows only rudimentary techniques of addressing the class imbalance have been applied to the sheep behaviour classification task, and DL techniques have not been thoroughly investigated. Therefore, our study contributes to the understanding of the classification of sheep behaviour, providing a foundation for the development of real-time monitoring systems.

The rest of the paper is organised as follows: Section 2 describes the research methodology while Section 3 presents the results and analysis. Discussions are given in Section 4 with conclusions in Section 5.

2. Materials and methods

2.1. Animals and research site

All procedures described were performed according to the guidelines of the Australian Code of Practice for the Use of Animals for Scientific Purposes 2013 and received approval from the Murdoch University Animal Ethics Committee (R3039/18). Two experiments were completed at the Muresk Institute Farm, near Northam in Western Australia (31° 44′59″S, 116°40′13″E). The two experiments represented different grazing scenarios; (i) Muresk Dry Pasture - Merino ewes (18 months of age) grazed a 3-hectare field of dry annual pasture for 7 days from 7th to 13th December 2018; (ii) Muresk Stubble - Merino ewes grazed a 3-hectare field of barley crop residues for 7 days from 1st to 8th February 2019. In both cases the total amount of plant biomass grazed exceeded 2 000 kg dry matter/ha.

Both experiments involved a total of 30 ewes. A subset of the ewes (Muresk Dry Pasture: 9 sheep and Muresk Stubble: 10 sheep, disjoint from Muresk Dry Pasture) were fitted with jaw mounted ActiGraph sensors (ActiGraph, Pensacola, Florida, USA) and ear mounted Axivity sensors (Axivity Ltd, Newcastle, UK) for the seven days (Fig. 1, Fig. 2).

Fig. 1 – Sheep in a pen at Muresk Institute Farm, wearing the full set of sensors.

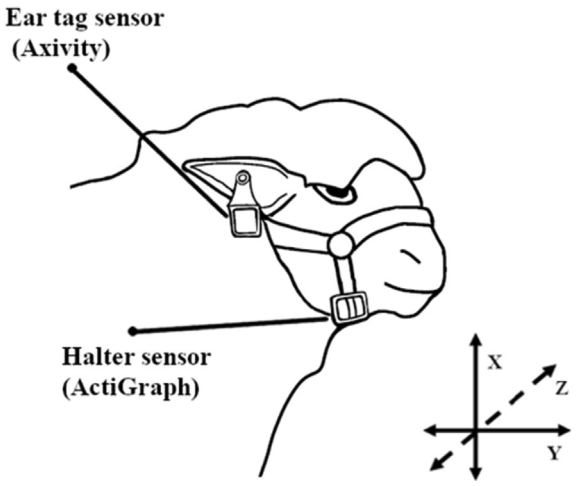


Fig. 2 – Schematic drawing of halter mounted ActiGraph sensor under the jaw and the ear mounted Axivity sensor. The three axes (x, y, z) are simultaneously sampled, representing three-dimensional movement.

Each of the ewes was also branded on each side with a unique paint brand to enable the ewes to be identified in the video recordings.

2.2. Data set description

The ActiGraph and Axivity sensors recorded tri-axial accelerometer data. The ActiGraph sensors were sampled at 30 Hz and the Axivity sensors at 25 Hz. The sensors ran for four days after a 3-day adaptation period, while the sheep were also observed during daylight hours through video recordings. A subset of the videos was subdivided into ten second blocks, and observations were made to allocate behaviours to the activities in the ten second blocks. The behaviours recorded were *sitting*, *standing*, *walking*, *grazing* and *ruminating*. These can be subdivided into two categories: Movement (*sitting*, *standing*, *walking*), and Feeding (*grazing*, *ruminating*). A sheep could undertake multiple activities within the ten second block, such as *sitting* and *grazing*, or be a single behaviour, such as *standing*. However, the Feeding categories were always recorded in connection with a movement behaviour.

The combining of observations resulted in thirteen separate categories. The breakdown of observation combinations is shown in Table 1. The overall class imbalance can be seen in the 'Muresk Dry Pasture' column, with significant differences between the highest (9 757 for *sitting*) and lowest (1 for *walking_grazing_ruminating*) observations.

The classification pipeline is shown in Fig. 3. It has four key modules: data preparation, data augmentation, classification and evaluation. Detailed descriptions of the modules are given below.

2.3. Data preparation

The raw data was provided as a series of comma separated values (CSV) files: two for each sheep. One containing the ActiGraph jaw mounted data, and the other containing the Axivity ear tag data. Each row contained a sheep identifier,

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

Table 1 – Breakdown of observation combinations. Sheep could undertake multiple activities within a 10 s window, forming compound classifications. Data for two sample sheep (from Muresk Dry Pasture) are provided, as well as the totals for the Muresk Dry Pasture (9 sheep) and Muresk Stubble (10 sheep).

Behaviours	Sheep #7	Sheep #9	Muresk Dry Pasture	Muresk Stubble
sitting	1 024	2 068	9 757	2 469
standing_grazing	815	426	6 652	5 253
standing	858	526	6 295	6 392
standing_ruminating	237	197	2 088	1 684
sitting_ruminating	73	197	1 498	836
walking	151	72	907	1 715
standing_walking_grazing	68	63	888	5 253
walking_grazing	56	34	653	218
standing_walking	63	34	380	1 470
standing_walking_ruminating	7	1	39	89
walking_ruminating	11	2	14	14
sitting_walking	0	0	4	11
standing_walking_grazing_ruminating	0	0	3	3
walking_grazing_ruminating	0	0	1	1

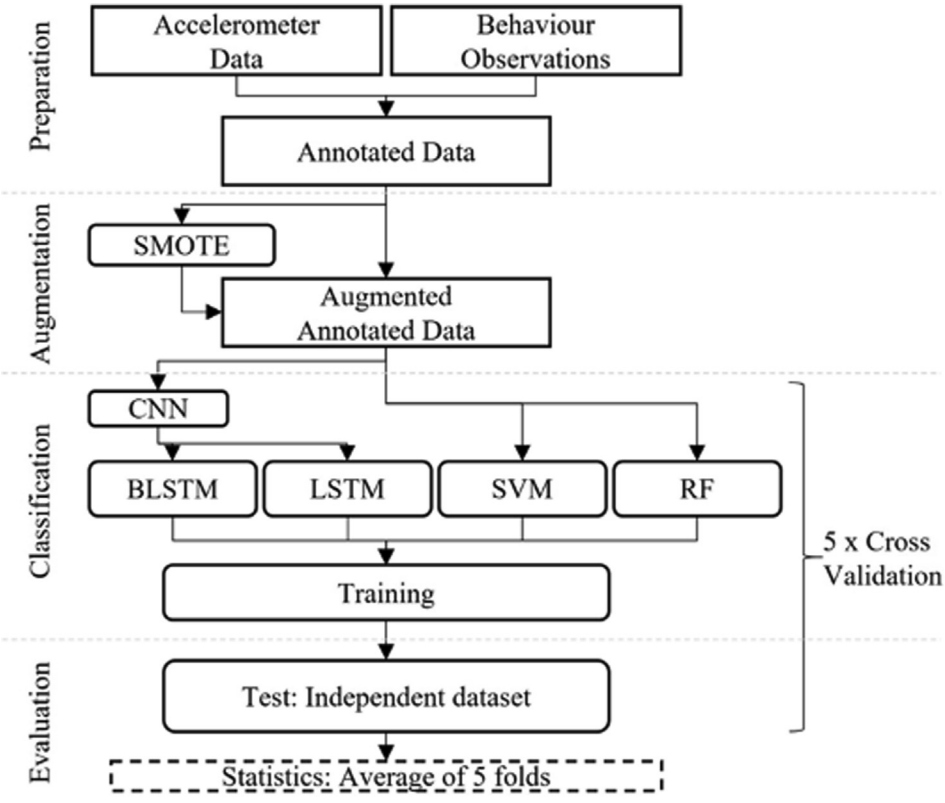


Fig. 3 – The classification pipeline showing the four phases of the classification process. The annotated data was augmented with SMOTE synthetic data, and input into four classifiers. For LSTM and BLSTM, the raw accelerometer data was passed through a CNN.

timestamp, the observed behaviours, the additional observations and finally the accelerometer data for the time window.

Using Pandas and Numpy Python libraries the two files for each sheep were loaded and merged as follows:

- The data sets were merged by the timestamps to produce one large set of columnar data. While merging the files, the observations were validated to ensure they matched between the two sources.

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

- The observed behaviours were encoded into discrete labels, matching the grouped behaviours. For example, if *standing* and *grazing* were observed, they were given the label *standing_grazing*.
- The data with labels that had fewer than 30 samples were removed from the data set. These represented classifications that had too few data points to be useful to classify.
- The unused columns (study name, sheep identifier, timestamp, observed behaviours and additional observations, including the number of steps except for the *4-Class + Steps* data set) were not included in the final data set for analysis.

Seven data sets were created from the observed behaviours, using both Muresk Dry Pasture (MDP) and Muresk Stubble (MS) experiments as detailed below:

(i) *9-Class for MDP*: The unique, combined 13 classes with distinct classes for each combination, with the extreme minority classes removed. This resulted in 9 classes.
(ii) *3-Class for MDP*: A data set with 3 distinct classes: *grazing*, *ruminating* and *other*, where *other* was all the other behaviours that did not have ruminating or grazing. The 4 samples that were labelled as ruminating and grazing were dropped.
(iii) *4-Class for MDP*: A data set with 4 distinct classes: *grazing*, *ruminating*, *walking*, and *other*, where the combination of ruminating and grazing with walking were allocated to the feeding category.
(iv) *4-Class + Step for MDP*: A second data set with the 4 distinct classes *grazing*, *ruminating*, *walking*, and *other* as above. In addition to the accelerometer data, the number of steps taken by the sheep, as observed from the video footage, was also included in the data for classification.
(v) *4-Class + Steps for MS*: The 4-Class + Steps test for MDP was replicated with the accelerometer data from Muresk Stubble.
(vi) *4-Class + Steps Leave one class out for MDP*: The 4-Class + Steps test was replicated with 9-fold cross validation by training on 8 sheep, and testing on the 1 left out.
(vii) *4-Class + Steps generalisation for both MDP and MS*: The generalisation of the classification methods for the best performing classifiers were performed by training on one experiment, and tested with the second. This was replicated in both directions training on MDP and testing on MS, and then training on MS, and testing on MDP.

The prepared data was split into 5-fold cross-validation sets, ensuring all samples had an opportunity to be excluded from the training set. These sets were marked as the baseline sets, and were saved to CSV files for use with each classifier. Likewise, after the addition of SMOTE synthetic data, the augmented data sets were saved to CSV files, and served as the input for each classifier.

2.4. Data augmentation

Polynom-fit-SMOTE, as implemented by the smote-variants Python library, was used to generate synthetic samples that match the data distribution. Two different topologies were tested in preliminary tests, 'star' and 'mesh', with the results for 'mesh' being slightly better.

2.5. Classification

The baseline and SMOTE augmented data sets were used to train the classification models. The trained models were then evaluated against the test sets of each fold. There were four different classification methods defined as follows:

2.5.1. Long Short Term Memory (LSTM)

LSTM is a form of RNN that is more robust, overcoming problems RNNs have with long term dependencies [30]. Accelerometer data is sequential. Without the memory capabilities of the LSTM, the sequential nature would be lost, losing information important for classification of the behaviours [16]. A combination of a CNN and LSTM was implemented using TensorFlow in Python using the Keras libraries to implement the models. The CNN was used for feature selection, configured based upon the research of Deep and Zheng [19], where they had a similar CNN and LSTM hybrid model for HAR data. The data was collated as an array of the accelerometer data: 900 data points for the ActiGraph data, and 750 for the Axivity data. This data, when combined, produced an array of 1 650 columns. The array was input into two 1D convolution layers, with a kernel size of 6, a filter size of 128, and using ReLU activation. The output of the convolutions was then passed through a dropout layer, to prevent over-fitting. Next, to reduce the complexity, the data was passed through a maxpooling layer with a pool size of 2. Finally, the feature selection was performed by passing the output of the maxpooling layer, through a dense layer, selecting 115 features. The number of features was determined through testing with different values to find the optimal value with the training data. The output of the feature selection was then fed into a LSTM layer to perform the classification. The results of the LSTM layer were then passed to a dense layer using softmax activation, to select for the number of classes in the training set.

2.5.2. Bidirectional Long Short Term Memory (BLSTM)

BLSTM introduces a bidirectional pass over the data, forwards and backwards. Given the 10 s epoch, a sheep may undertake multiple behaviours within the same epoch. There are transitions in both directions within the same epoch, e.g., *standing_walking* records both *standing* to *walking* and *walking* to *standing* transitions. Given the bidirectional nature of the transitions, applying the bidirectional LSTM may improve the classification results. The implementation for this classification method followed the same design as the CNN LSTM form, however a bidirectional LSTM was used instead of a single direction LSTM.

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

2.5.3. Support vector Machines (SVM)

SVM is a machine learning technique based on statistical learning theory [31], which classifies by minimizing real error, developing a hyperplane that separates the samples into two categories. The aim is to separate the two points as much as possible to minimise the inaccuracies in predictions [32]. SVM is a binary classifier, but can be applied to multiclass problems by applying the one-versus-one technique [33]. SVM classification was performed using Scikit-learn's [34] implementation. A radial basis function (RBF) kernel was used with a regularisation parameter of 1.0, and gamma to set scale.

2.5.4. Random Forest (RF)

RF is a type of ensemble learning where the feature set is randomly split for use in decision trees. The results of the individual trees are then combined into a classification decision. The random nature of the feature selection splits, leads to a more robust model that is resistant to overfitting [35]. RF classification was performed using Scikit-learn's implementation [34], with 100 trees.

2.6. Evaluation metrics

The evaluation of the results was performed using a confusion matrix which maps the actual values and the predicted values to determine the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) results of the classification. From these values the following were calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Accuracy by Eq. (1), the percentage of samples correctly classified, gives an overall performance indication of the classification. Precision by Eq. (2) measures what fraction of predictions that gave a positive class, were actually positive. Recall by Eq. (3) (also referred to as sensitivity) indicates what fraction, of all positive samples, were correctly predicted as positive by the classifier. F1-Score by Eq. (4) combines the precision and recall into one value for comparison, in the form of the harmonic mean.

The following metrics were used to aggregate the results:

(i) Weighted Average takes the metric and calculates the average weighted by the support (the number of occurrences of each class in the test set).
(ii) Macro Average is the unweighted mean. Support is not taken into account.
(iii) Minor Average takes the metric and calculates the average weighted by the inverse of the support. This promotes the minority classes.

3. Results

3.1. Test environment

All tests were performed on an AMD Ryzen 5 3600 6-Core with 32 GB RAM, with an NVIDIA GeForce RTX 2060 with 6 GB RAM, running Linux Mint 20 with kernel version 5.4.0–60.

3.2. Classification

The overall accuracy results show that LSTM, augmented with SMOTE data, produced the highest accuracy for the small class sets (3-Class: 88.5%, 4-Class: 87.3%;Table 2). However, for the entire 9 classes, RF with SMOTE data was the highest (72.4%).

The addition of classes resulted in a reduction in accuracy. Attempting to distinguish *walking* from the other categories from the 3-Class to 4-Class studies resulted in a drop from 88.5% to 87.3%. Examining the confusion matrix, the accuracy of *walking* was much lower, showing considerable confusion with *grazing* as well as the other (*sitting* and *standing*) classes (Fig. 4 (a)). *Walking* had the smallest number of samples for the data set, with 1 291 samples, versus 3 639 for *ruminating*, 8 193 for *grazing* and 16 052 for other. Examining the raw accelerometer data for the failures, the mixture of inactivity and activity within the recorded data can result in failures to detect the *walking* activity. Fig. 5 shows the raw accelerometer data of a false negative result where a similar *walking* activity is seen for the first 5 to 7.5 s of the epoch before stilling into inactivity. Even with a majority of the time showing activity it resulted in an incorrect prediction. However, once the number of steps was added to the classification process, the accuracy of the *walking* improved from 0.43 to 0.69 (Fig. 4 (b)). There is still some confusion with *grazing*, but distinguishing *walking* from other activities drops from 0.31 to 0.09.

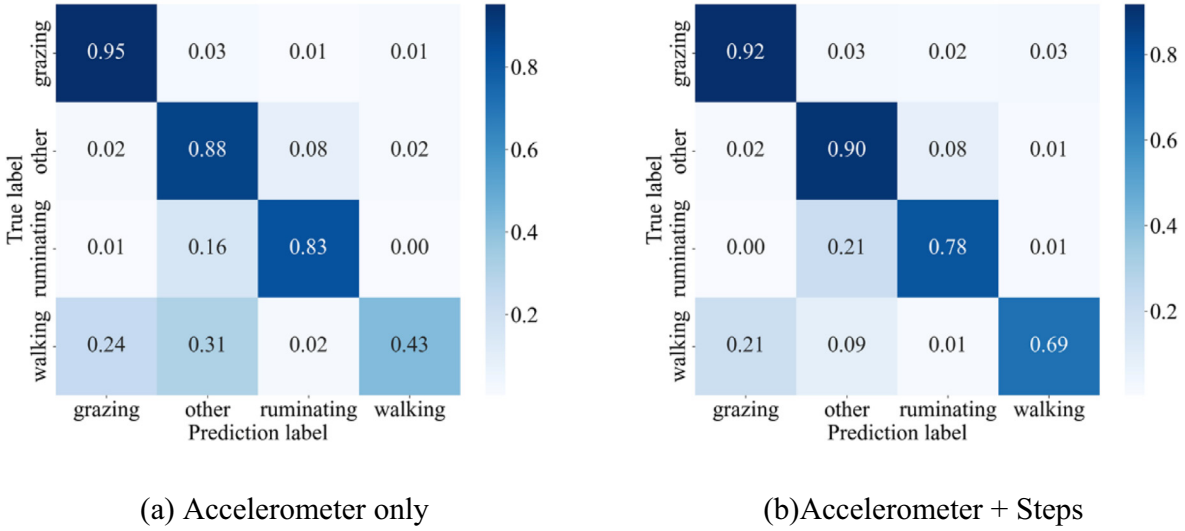
Comparing the DL methods, BLSTM accuracy was close to LSTM. Examining the details of the individual metrics for the 9-Class study, BLSTM was ahead of LSTM for precision, while having equivalent recall and F1-Score (Table 3). With the greater number of classes, and class imbalance present, BLSTM resulted in slightly better results for the minority classes. The use of BLSTM does introduce a training cost with the training and evaluation time for BLSTM taking 61% longer to complete than LSTM (for the 9-Class study). RF was the most efficient to train and evaluate, while SVM was the worst (Table 4). However, given the total time of testing for LSTM was 14.5 s for 28 988 samples, the average inference time is 0.5 ms, supporting real-time inferences.

The 4-Class + Steps tests on the Muresk Dry Pasture experiment were replicated with a second flock, Muresk Stubble, to validate the results. The overall F1-Score decreased for the second set of sheep, but maintained the performance order, with BLSTM and LSTM outperforming RF and SVM (Table 5). However, RF was closer to BLSTM and LSTM than in the first experiment (LSTM = 0.83, RF = 0.82). Combining the data sets, LSTM maintained the performance from the Muresk Dry Pasture results, with an F1-Score of 0.88. The performance for detecting *walking* and *ruminating* dropped for the second

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

Table 2 – Classification accuracy (%) for Muresk Dry Pasture. The results are shown for each classifier, taking the average of the 5 × Cross validation folds for the baseline, and SMOTE augmented data sets.

	3-Class		4-Class		4-Class + Steps		9-Class	
	Baseline	SMOTE	Baseline	SMOTE	Baseline	SMOTE	Baseline	SMOTE
BLSTM	78.7	87.7	73.9	85.8	82.9	87.4	61.7	70.4
LSTM	83.3	88.5	75.9	87.3	81.8	88.0	61.5	70.5
SVM	77.6	76.8	74.7	74.1	77.0	76.0	59.6	58.4
RF	83.9	83.0	81.2	81.2	81.7	82.5	71.3	72.4



(a) Accelerometer only (b)Accelerometer + Steps

Fig. 4 – The normalised confusion matrix for LSTM with SMOTE data for Muresk Dry Pasture. In both cases a 4-class study was undertaken, with only accelerometer data present in (a), and the steps observations, in addition to the accelerometer data, shown in (b). The addition of the steps improved the results for walking, reducing the confusion with the other movement behaviours.

experiment, having difficulty distinguishing *grazing* and *walking* behaviours (Fig. 6).

3.3. *Generalisation across data*

To demonstrate the generalisability performance of the DL methods, we performed two sets of tests: i) Leave one class out cross validation on the Muresk Dry Pasture data set and ii) Train and test across data sets, i.e., train the techniques on one data set and test on the other one. The results of the Leave one class out cross validation approach are shown in Table 6, while the results of the second study are shown in Table 7. The results in Table 6 show that BLSTM was best able to generalise the results when training within the same set, the additional pass over the data resulting in a slight improvement over LSTM. However, the DL methods outperformed RF and SVM by a larger margin, RF and SVM showing a larger loss of performance (SMOTE Weighted Average F1-Score reduction; BLSTM: 0.04, LSTM: 0.05, SVM: 0.12, RF: 0.17). The results in Table 7 show that the DL methods can be generalised between the experiments. There was some loss in performance for DL, but the loss is much smaller than compared to SVM and RF. Additionally, the performance is improved when the training sample size is larger, as is seen when training on Muresk Stubble, and testing on Muresk Dry Pasture, as Muresk Stubble has the larger data set.

3.4. *Data synthesis*

Data synthesis improved the classification results. For the 4-Class + Steps classification, the improvements with the addition of the SMOTE synthetic data can be seen with improvements across three of the four classes (Fig. 7, note Fig. 4 (b), Fig. 6 (a) and Fig. 7 (b) are the same confusion matrix). The improvements for *walking* were dramatic, improving from just 0.20 to 0.69. However, this was not a universal improvement. Considering the accuracy results for the 3-class study for the individual sheep, SMOTE improved the results for LSTM for Sheep #7 by 9.8%, whereas the results decreased for Sheep #9 by 11.1% (Table 8). Similarly, for SVM and RF in the 3-Class study the addition of synthetic data results in a drop in accuracy. However, looking at the breakdown of the changes with the 9-Class study for Sheep #9, the synthetic data does improve the results for the minority classes (minor average), at the cost of the majority classes (Table 9). Additionally, the use of SMOTE does result in improved precision

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

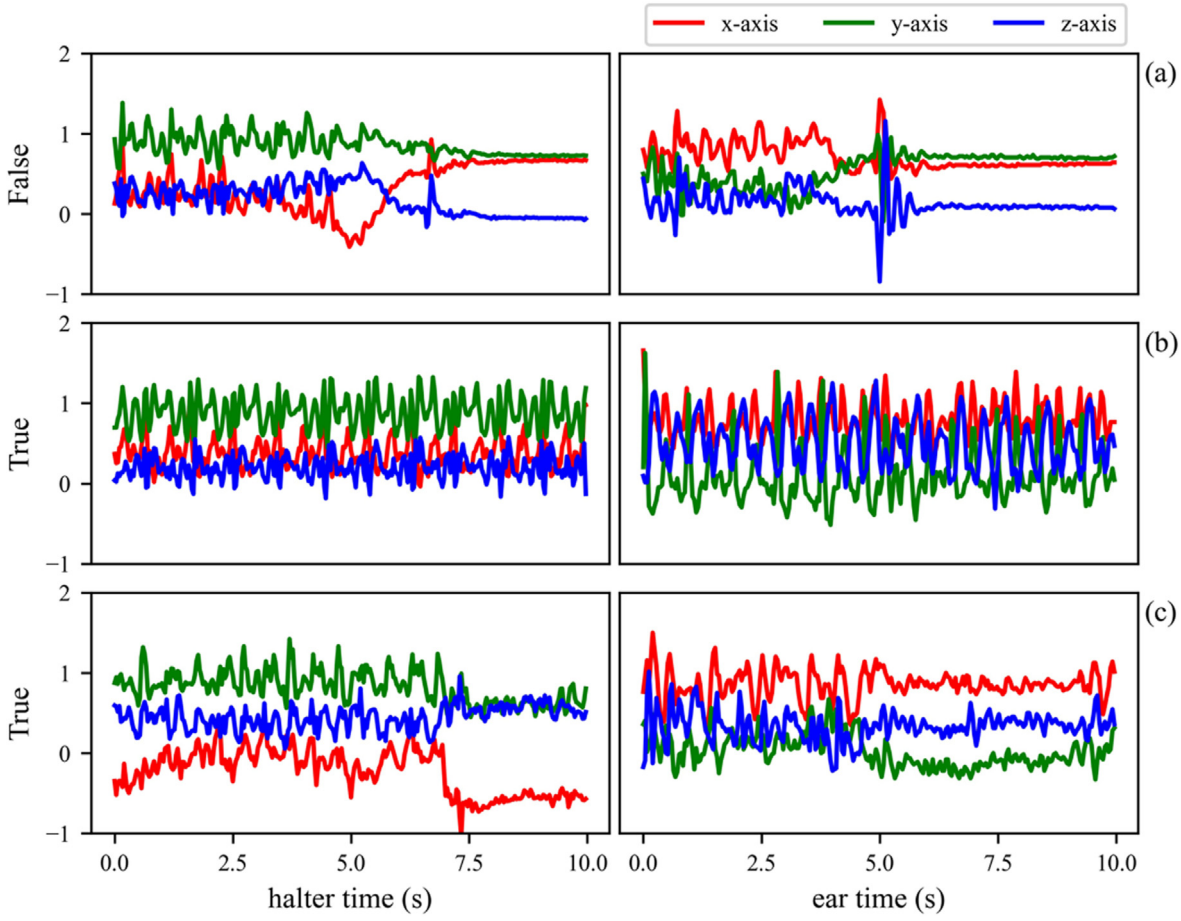


Fig. 5 – Sample raw accelerometer accelerometer data for *walking* observations for Muresk Dry Pasture. The left column represents the accelerometer data captured by the halter sensor. The right column represents the data from the ear sensor. (a) was incorrectly identified as *grazing* rather than *walking*. (b) and (c) are samples of correctly identified *walking* activities.

Table 3 – 9-Class results summary for Muresk Dry Pasture (9 sheep). The best results for each category are highlighted in bold. The results are split into summary statistics with the weighted average showing the results focused on the majority class, the macro average showing the balanced results and the minor average showing the results highlighting the minority classes.

	BLSTM Baseline	SMOTE	LSTM Baseline	SMOTE	SVM Baseline	SMOTE	RF Baseline	SMOTE	Support
Weighted avg									
Precision	0.56	**0.70**	0.56	0.69	0.51	0.52	0.67	0.69	28 988
Recall	0.62	0.70	0.61	0.70	0.60	0.58	0.71	**0.72**	28 988
F1-score	0.58	0.69	0.58	0.69	0.52	0.53	0.66	**0.70**	28 988
Macro avg									
Precision	0.32	**0.45**	0.33	**0.45**	0.27	0.29	0.44	0.44	28 988
Recall	0.30	**0.46**	0.31	0.45	0.23	0.26	0.33	0.39	28 988
F1-score	0.29	**0.43**	0.31	0.42	0.21	0.25	0.33	0.39	28 988
Minor avg									
Precision	0.25	**0.35**	0.26	**0.35**	0.21	0.23	0.34	0.34	28 988
Recall	0.23	**0.35**	0.24	**0.35**	0.18	0.20	0.25	0.30	28 988
F1-score	0.23	**0.33**	0.24	0.32	0.17	0.20	0.26	0.30	28 988

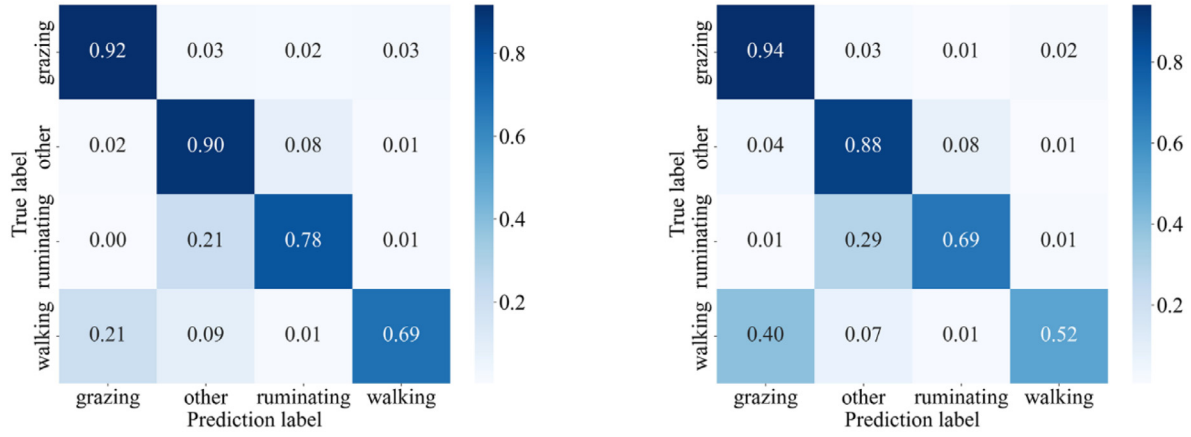
Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

Table 4 – 9-Class full study running times (seconds) for Muresk Dry Pasture. Training for all 5 folds. Testing was the time to evaluate all 28 988 samples.

	Training		Testing	
	Baseline	SMOTE	Baseline	SMOTE
BLSTM	1731.5	8314.6	24.7	24.5
LSTM	1200.1	5157.2	14.4	14.5
SVM	1716.0	11102.5	472.2	1384.3
RF	272.5	1371.8	1.0	1.2

Table 5 – A comparison of the weighted average F1-Scores for 4-Class + Steps augmented with SMOTE synthetic data for Muresk Dry Pasture and Muresk Stubble.

	Muresk Dry Pasture			Muresk Stubble			Combined
	Sheep #7	Sheep #9	9 Sheep	Sheep #5	Sheep #8	10 Sheep	
BLSTM	0.80	0.78	0.88	0.81	0.85	0.83	0.87
LSTM	0.73	0.77	0.88	0.58	0.76	0.83	0.88
SVM	0.77	0.82	0.72	0.79	0.80	0.70	0.69
RF	0.84	0.88	0.82	0.84	0.90	0.82	0.83



(a) Muresk Dry Pasture (b) Muresk Stubble

Fig. 6 – The normalised confusion matrix for the SMOTE augmented LSTM classification of the 4-Class + Steps studies. Muresk Dry Pasture (a) shows a stronger performance with the *ruminating* and *walking* classes, compared to Muresk Stubble (b).

with the single sheep tests. In the case of *walking* for RF, the use of SMOTE improved the precision from 0.00 to 0.62 (Table 10).

4. Discussion

4.1. *Classification*

One focus of this study was to compare DL classification to alternative machine learning techniques. With the smaller number of classes, LSTM and BLSTM outperformed both RF and SVM for the full 9 sheep study. Additionally, the DL methods showed to generalise better between the two flocks of sheep, maintaining a higher performance over RF and SVM, when training on one experiment and testing on the other. However, the success of DL appears to be a factor of the data set size, with the LSTM and BLSTM showing inferior performance to both RF and SVM with the single sheep tests and performing better in the generalisation tests when there was more data. This is also reflected in the results for the 9-Class test. Creating unique classes by combining the behaviours, reduced the number of samples with which to train each class. This resulted in poorer performance overall in the 9-Class test. RF was the top performer in this case, but DL with the assistance of the synthetic data is approaching the results of RF. Further investigation with a larger study

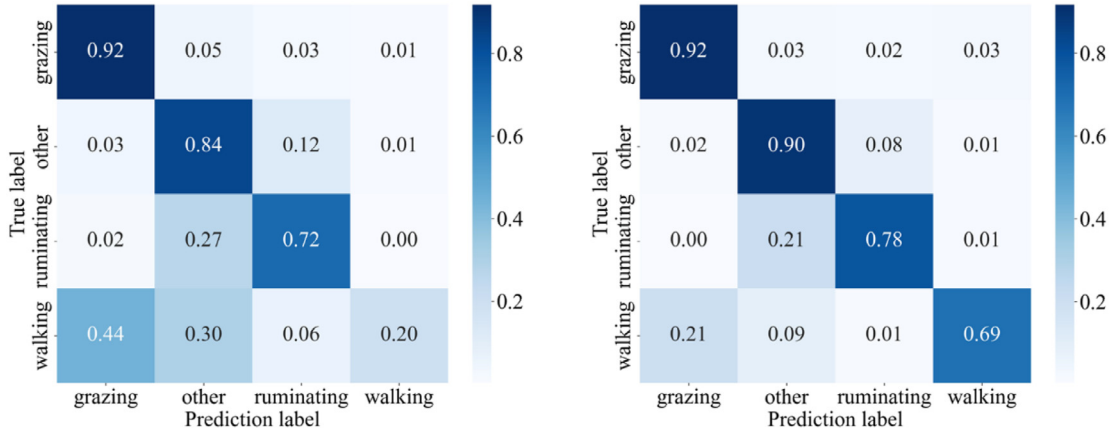
Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

Table 6 – Muresk Dry Pasture 9 × Cross validation Leave one class out 4-Class + Steps test. The methods were trained on 8 sheep and tested on 1 sheep that was excluded from the training data.

	BLSTM Baseline	SMOTE	LSTM Baseline	SMOTE	SVM Baseline	SMOTE	RF Baseline	SMOTE	Support
Weighted avg									
Precision	0.76	**0.85**	0.83	0.84	0.58	0.60	0.58	0.66	29 175
Recall	0.77	**0.84**	0.82	0.83	0.64	0.64	0.60	0.64	29 175
F1-score	0.76	**0.84**	0.82	0.83	0.58	0.60	0.55	0.65	29 175
Macro avg									
Precision	0.74	**0.78**	0.74	0.74	0.54	0.55	0.58	0.54	29 175
Recall	0.63	**0.77**	0.68	0.75	0.44	0.47	0.35	0.58	29 175
F1-score	0.67	**0.77**	0.68	0.74	0.45	0.48	0.35	0.55	29 175
Minor avg									
Precision	0.42	**0.44**	0.42	0.42	0.32	0.32	0.33	0.30	29 175
Recall	0.36	**0.44**	0.39	0.43	0.25	0.27	0.20	0.33	29 175
F1-score	0.38	**0.44**	0.39	0.42	0.27	0.28	0.21	0.31	29 175

Table 7 – The generalisation tests between the two experiments, Muresk Dry Pasture (MDP) and Muresk Stubble (MS). The 4-Class + Steps model was trained on one experiment and tested on the second experiment. There were 19 distinct sheep, 9 for MDP, and 10 for MS. This was performed both ways for each classifier.

Classifier	Train	Test	Precision	Recall	F1-Score	Support
BLSTM	MDP	MS	0.75	0.63	0.61	24 402
BLSTM	MS	MDP	0.81	0.67	0.67	29 175
LSTM	MDP	MS	0.78	0.69	0.71	24 402
LSTM	MS	MDP	0.85	0.71	0.74	29 175
SVM	MDP	MS	0.30	0.35	0.30	24 402
SVM	MS	MDP	0.50	0.29	0.32	29 175
RF	MDP	MS	0.31	0.35	0.33	24 402
RF	MS	MDP	0.50	0.33	0.37	29 175



(a) Baseline (b) SMOTE augmented

Fig. 7 – The normalised confusion matrix for the LSTM classification of the Muresk Dry Pasture 4-Class + Steps study. The baseline result (a) was performed with the raw accelerometer data. The addition of SMOTE synthetic data (b) shows improved classification and less confusion for the minority classes.

would be useful to determine if DL can show superior performance over RF with more data.

Breaking down the individual components in the 9-Class test, the DL methods showed to be more beneficial to the minority classes, providing some successful classification results where there were none for RF. Given the different advantages of RF and the DL methods, RF performing better for the majority classes, and DL for the minority classes,

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

Table 8 – Muresk Dry Pasture 3-Class accuracy results by data set size (%). Sheep #7 contained 3 356 samples, Sheep #9 3 590 samples, and the full study 29 175 samples. Sheep #7 and Sheep #9 have similar sizes, but different class distributions.

	Sheep #7		Sheep #9		Full Study	
	Baseline	SMOTE	Baseline	SMOTE	Baseline	SMOTE
BLSTM	75.3	77.9	72.0	78.3	78.7	87.7
LSTM	65.8	75.6	85.9	74.8	83.3	88.5
SVM	82.2	78.0	86.1	84.9	77.6	76.8
RF	86.7	84.5	92.5	89.3	83.9	83.0

Table 9 – Muresk Dry Pasture 9-Class results summary for Sheep #9. The best results for each category are highlighted in bold. The results are split into summary statistics with the weighted average showing the results focused on the majority class, the macro average showing the balanced results and the minor average showing the results highlighting the minority classes.

	BLSTM		LSTM		SVM		RF		
	Baseline	SMOTE	Baseline	SMOTE	Baseline	SMOTE	Baseline	SMOTE	Support
Weighted avg									
Precision	0.74	0.79	0.61	0.74	0.71	0.77	0.82	**0.83**	3 588
Recall	0.78	0.72	0.64	0.67	0.78	0.75	**0.86**	0.84	3 588
F1-score	0.74	0.74	0.55	0.69	0.72	0.75	**0.84**	0.83	3 588
Macro avg									
Precision	0.39	0.45	0.37	0.43	0.37	0.39	0.45	**0.49**	3 588
Recall	0.38	0.44	0.19	0.42	0.29	0.42	0.46	**0.50**	3 588
F1-score	0.36	0.42	0.20	0.38	0.28	0.37	0.45	**0.48**	3 588
Minor avg									
Precision	0.30	0.33	0.28	0.32	0.28	0.29	0.34	**0.37**	3 588
Recall	0.28	0.33	0.14	0.31	0.22	0.32	0.34	**0.37**	3 588
F1-score	0.27	0.31	0.16	0.28	0.21	0.28	0.34	**0.36**	3 588

Table 10 – 9-Class precision for Muresk Dry Pasture Sheep #9. The precision results for the individual classes are shown for each of the classifiers, with SMOTE synthetic data (S) and the baseline (B) without synthetic data.

	BLSTM		LSTM		SVM		RF		
	B	S	B	S	B	S	B	S	Support
sitting	0.91	0.97	0.64	0.90	0.83	0.93	0.94	0.97	2 068
sitting, ruminating	0.40	0.49	0.84	0.46	0.33	0.33	0.85	0.60	168
standing	0.63	0.49	0.61	0.40	0.65	0.70	0.77	0.73	526
standing, grazing	0.62	0.79	0.70	0.81	0.68	0.65	0.69	0.71	426
standing, ruminating	0.34	0.50	0.41	0.47	0.75	0.60	0.77	0.74	197
standing, walking	0.00	0.04	0.00	0.05	0.00	0.00	0.00	0.03	34
standing, walking, grazing	0.00	0.04	0.00	0.19	0.00	0.00	0.00	0.00	63
walking	0.61	0.71	0.10	0.57	0.07	0.28	0.00	0.62	72
walking, grazing	0.00	0.02	0.00	0.06	0.00	0.00	0.00	0.00	34

one potential optimisation for the classification implementation would be to employ both methods in the training process. RF could be used to detect the majority behaviours, and LSTM to identify the minority behaviours. This way the training process will work to the strengths of the classification models, resulting in an overall higher performance. At inference time, both models can be evaluated with a confidence factor to determine the appropriate classification where the model inferences are in conflict. The combination would lead to better overall accuracy for both majority and minority classes.

While a greater number of classes may result in better predictions for subsequent analysis, the inferior performance, with the 9-Class classification, degrades confidence in the outcomes. The importance of detecting behaviours for specific purposes, for example accurately predicting *ruminating* and *grazing* behaviours in order to predict food intake, may prove both more practical and more valuable. As seen by the 3-Class test, a much higher accuracy for the *ruminating* and *grazing* behaviours can be achieved. The introduction of *walking* in the 4-Class test resulting in a reduction of accuracy, shows that there is a cost in introducing more classes.

Walking is harder to classify than the other movement behaviours. Even as part of the labelling task this can be challenging. For example, determining the difference between steps taken while *grazing*, as distinct from steps taken while

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

walking, is hard. This may require specifying a threshold on the number of steps, or distance travelled. The tests in this study also showed that the models have difficulty predicting the *walking* behaviour from accelerometer data only, confusing it with *grazing* and other behaviours. However, the addition of the steps to the classification improved the success in correctly identifying the *walking* behaviour. The data collected here is not representative of a real world usage scenario, as the number of steps were obtained via manual observation. However, it demonstrates that additional data to the accelerometer data would be beneficial in identifying the *walking* behaviour. For example, studies have shown that gyroscope-based sensors assist in identifying behaviours [10], while inclusion of tri-axial magnetism data helped behaviour detection in goats, particularly for the minority classes [11]. Therefore, future research should examine the use of integrated sensors, such as 6-axis sensors that incorporate more forms of data.

The validation test of the 4-Class + Steps on Muresk Stubble showed a drop in performance in Muresk Stubble, when compared to Muresk Dry Pasture. This is presenting as a confusion between the *grazing* and *walking* behaviours. Looking at the distribution of the classes, Muresk Dry Pasture has less overlap of *grazing* and *ruminating* behaviours with the *walking* behaviour, with Muresk Dry Pasture overlapping in just 5% of cases, and Muresk Stubble overlapping in 19% of cases (Table 1). This shows that the overlapping of the classes does present a challenge in terms of accurately identifying the behaviours, and that further research needs to be performed to optimise the process of detecting *walking*. In addition to adding additional data sources, there may be benefits to splitting the classification process into two, one for determining the Feeding behaviours (*ruminating* and *grazing*), and one for determining the movement behaviours. As there is the potential overlap for these behaviours, producing separate results may give a better indication of the sheep's health, or physiological state. The movement behaviours could be broken down into an activity/inactivity classification, depending on the purpose [3]. Additionally, this study used a fixed window size of 10 s. Using a smaller time window (5 s) [2], a mix of time window sizes [9], or a sliding window [5] may result in better classification of *walking* behaviour.

Comparing the DL methods, BLSTM and LSTM showed similar results, with LSTM providing generally better overall performance, and BLSTM being slightly ahead of LSTM for the minority class classification. However, given the computation costs involved with BLSTM, LSTM is likely the better candidate. Given the slight improvement in BLSTM for the 9 class tests, it may prove the better option where there is a small differentiation between the class feature spaces. The forward and reverse traversal of the data may prove more beneficial in this case in detecting the smaller differences.

4.2. *Data synthesis*

Examining the results in the context of the SMOTE data augmentation, the addition of synthetic data generally helped improve the performance for the DL methods. However, for RF and SVM, the addition of the synthetic data resulted in either a reduction in accuracy or a very small improvement. This was demonstrated in the 9-Class test for Sheep #9. With the limited data in the single sheep data set and a highly dominant majority class, the additional synthetic data leads to greater confusion with the majority class, reducing the overall success of the classification. With Sheep #9, the majority class (*sitting*) is so dominant (57.6% of samples), the influence on the weighted average by a reduction on the recall for that class, results in an overall performance loss. With the addition of the SMOTE data, the model learns from more examples of the minority cases which led to confusion over similar samples to the majority class. This results in a drop in the majority class recall, which, because of the dominant position of the majority class results in a quick degradation of the weighted average recall.

Therefore, the level of class imbalance needs to be considered before applying the synthetic data. Further study needs to be performed to see if a reduction of the synthetic data, with the aim of maintaining some, but not extreme, imbalance, improves performance over balancing out the entire data set to equal proportions.

5. Conclusion and future work

In this study, we performed a study of DL based sheep behaviour analysis on accelerometer data collected under grazing conditions. In particular, we considered the challenges of class imbalance in the data set. Additionally, this is a first study of classification where multiple behaviours are present in the observation time window, represented as compound behaviours. Previous works have examined other machine learning methods, but have not explored DL. For comparative purposes we also used two alternative machine learning techniques. Altogether, four classification techniques, LSTM, BLSTM, SVM and RF, were analysed in this study. The comparison was made by adjusting the data set with synthetic data generated using polynom-fit-SMOTE and performing classification training with the augmented data set. These tests were performed on four behaviour sets, a three class test, a four class test, a four class test with additional steps data, and a nine class test. The tests were carried out on individual sheep, as well as the full nine sheep study.

The use of synthetic data showed benefits for the DL methods. For RF, benefits were seen for the minority classes, but reduced performance in classifying strongly dominant majority classes. For the three and four class tests, the DL methods were the best performers, regardless of data set size, with the addition of the step data making noticeable improvements to the *walking* classification. From the improvement with step data, we can infer that the models will benefit from diverse sensor data sources. For the nine class test, RF was the best performing classifier, but this was largely due to its success in classifying the majority classes. The DL methods, with more data, showed comparable performance to RF, and showed better performance when classifying the minority classes. Generalisability tests, i.e., train on one set of sheep and test on a different set, demonstrated the DL techniques can generalise effectively.

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

5.1. *Future work*

Further research is required explore the possibilities of utilising a combination of the DL and RF classification techniques to provide a more accurate prediction on imbalanced sheep behaviour data. Further study is also required to correctly predict *walking* behaviour, adding alternative data sources, simplifying the categories, or separating the feeding and movement behaviours into separate models. Additionally, examining effects of modifying the time window size for detecting *walking* behaviour should be undertaken.

Finally, the average inference times suggest that real-time classification should be possible. However, future research is required to validate this conclusion, particularly on a microcontroller. The CNN-LSTM used in this study cannot currently be deployed on TensorFlow Lite for Microcontrollers as it does not support all required operations for the LSTM layers. Therefore, future research will need to look at alternative implementations of LSTM that can be deployed in such an environment.

CRediT authorship contribution statement

Kirk E. Turner: Conceptualization, Methodology, Data curation, Investigation, Writing – original draft, Writing – review & editing. **Andrew Thompson:** Methodology, Resources, Data curation, Writing – review & editing, Project administration. **Ian Harris:** Methodology, Data curation, Writing – review & editing. **Mark Ferguson:** Funding acquisition, Resources, Data curation. **Ferdous Sohel:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work received funding from Meat and Livestock Australia, Murdoch University and the Western Australian Department of Training and Workforce Development. Steve Wainewright and staff at the Muresk Institute farm are thanked for their assistance in conducting this experiment. The efforts of all staff that assisted with sheep measurements including capture of video recordings is also appreciated.

REFERENCES

[1] Frost AR, Schofield CP, Beaulah SA, Mottram TT, Lines JA, Wathes CM. A review of livestock monitoring and the need for integrated systems. Comput Electron Agric 1997;17(2):139–59.

[2] Alvarenga FAP, Borges I, Palkovič L, Rodina J, Oddy VH, Dobos RC. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. Applied Animal Behaviour Science 2016;181:91–9.

[3] Fogarty ES, Swain DL, Cronin GM, Moraes LE, Trotter M. Behaviour classification of extensively grazed sheep using machine learning. Comput Electron Agric 2020;169:105175.

[4] Brown DD, Kays R, Wikelski M, Wilson R, Klimley AP. Observing the unwatchable through acceleration logging of animal behavior. Anim Biotelem 2013;1:1–16.

[5] Barwick J, Lamb DW, Dobos R, Welch M, Schneider D, Trotter M. Identifying sheep activity from tri-axial acceleration signals using a moving window classification model. Remote Sensing 2020;12:1–13.

[6] Hancock J. Studies in grazing behaviour of dairy cattle: II. Bloat in relation to grazing behaviour. The Journal of Agricultural Science 1954;45(1):80–95.

[7] Giovanetti V, Decandia M, Molle G, Acciaro M, Mameli M, Cabiddu A, et al. Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer. Livestock Science 2017;196:42–8.

[8] Walton E, Casey C, Mitsch J, Vázquez-Diosdado JA, Yan J, Dottorini T, et al. Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour. R Soc open sci 2018;5(2):171442.

[9] Hu S, Ingham A, Schmoelzl S, McNally J, Little B, Smith D, et al. Inclusion of features derived from a mixture of time window sizes improved classification accuracy of machine learning algorithms for sheep grazing behaviours. Comput Electron Agric 2020;179:105857.

[10] Mansbridge N, Mitsch J, Bollard N, Ellis K, Miguel-Pacheco G, Dottorini T, et al. Feature selection and comparison of machine learning algorithms in classification of grazing and rumination behaviour in sheep. Sensors (Switzerland) 2018;18(10):3532.

[11] Sakai K, Oishi K, Miwa M, Kumagai H, Hirooka H. Behavior classification of goats using 9-axis multi sensors: The effect of imbalanced datasets on classification performance. Comput Electron Agric 2019;166:105027.

[12] Guo L, Welch M, Dobos R, Kwan P, Wang W. Comparison of grazing behaviour of sheep on pasture with different sward surface heights using an inertial measurement unit sensor. Comput Electron Agric 2018;150:394–401.

[13] Barwick J, Lamb DW, Dobos R, Welch M, Trotter M. Categorising sheep activity using a tri-axial accelerometer. Comput Electron Agric 2018;145:289–97.

[14] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. Data Min Knowl Disc 2019;33(4):917–63.

[15] Jiang W. Time series classification: nearest neighbor versus deep learning models. SN Applied Sciences 2020;2:1–17.

[16] Ramasamy Ramamurthy S, Roy N. Recent trends in machine learning for human activity recognition—A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2018;8:1–11.

[17] Hernández F, Suárez LF, Villamizar J, Altuve M. Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network. 2019 22nd Symposium on Image, Signal Processing and Artificial Vision, STSIVA 2019 - Conference Proceedings 2019.

[18] Wang J, Chen Y, Hao S, Peng X, Hu L. Deep learning for sensor-based activity recognition: A survey. Pattern Recogn Lett 2019;119:3–11.

[19] Deep S, Zheng X. Hybrid Model Featuring CNN and LSTM Architecture for Human Activity Recognition on Smartphone Sensor Data. Proceedings - 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies, PDCAT 2019 2019:259–64.

[20] Sainath TN, Vinyals O, Senior A, Sak H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2015;2015-Augus:4580–4.

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001

[21] Kovács G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Applied Soft Computing Journal 2019;83:105662.
[22] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002;16:321–57.
[23] Fernández A, García S, Herrera F, Chawla NV. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research 2018;61:863–905.
[24] Guo X, Yin Y, Dong C, Yang G, Zhou G. On the class imbalance problem. Proceedings - 4th International Conference on Natural Computation, ICNC 2008 2008;4:192–201.
[25] Gazzah S, Ben ANE. New oversampling approaches based on polynomial fitting for imbalanced data sets. DAS 2008 - Proceedings of the 8th IAPR International Workshop on Document Analysis Systems 2008:677–84.
[26] Raziff ARA, Sulaiman N, Mustapha N, Perumal T. Smote and OVO multiclass method for multiple handheld placement gait identification on smartphone's accelerometer. J Engineering Applied Sciences 2017;12:374–82.
[27] Khojasteh SB, Villar JR, Chira C, González VM, de la Cal E. Improving fall detection using an on-wrist wearable accelerometer. Sensors (Switzerland) 2018;18:1–28.
[28] Sundararajan K, Georgievska S, te Lindert BHW, Gehrman PR, Ramautar J, Mazzotti DR, et al. Sleep classification from wrist-worn accelerometer data using random forests. Sci Rep 2021;11(1).
[29] Javed AR, Sarwar MU, Khan S, Iwendi C, Mittal M, Kumar N. Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition. Sensors (Switzerland) 2020;20:1–18.
[30] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[31] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Networks 1999;10:988–99.
[32] Sultana J, Usha Rani M, Farquad MAH. Deep Learning Based Recommender System Using Sentiment Analysis to Reform Indian Education. International Conference On Computational and Bio Engineering 2019:143–50.
[33] Rocha A, Klein Goldenstein S. Multiclass from binary: Expanding One-versus-all, one-versus-one and ECOC-based approaches. IEEE Trans Neural Networks Learn Syst 2014;25 (2):289–302.
[34] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.
[35] Breiman L. Random forests. Machine Learning 2001;45:5–32.

Please cite this article as: K. E. Turner, A. Thompson, I. Harris et al., Deep learning based classification of sheep behaviour from accelerometer data with imbalance, Information Processing in Agriculture, https://doi.org/10.1016/j.inpa.2022.04.001