Inclusion of features derived from a mixture of time window sizes improved classification accuracy of machine learning algorithms for sheep grazing behaviours

Hu, Shuwen, Ingham, Aaron, Schmoelzl, Sabine, McNally, Jody, Little, Bryce, Smith, Daniel, Bishop-Hurley, Greg, Wang, You-Gan and Li, Yutao

Hu, S., Ingham, A., Schmoelzl, S., McNally, J., Little, B., Smith, D., Bishop-Hurley, G., Wang, Y.-G. and Li, Y. (2020). Inclusion of features derived from a mixture of time window sizes improved classification accuracy of machine learning algorithms for sheep grazing behaviours. *Computers and Electronics in Agriculture*, *179*, Article 105857. https://doi.org/10.1016/j.compag.2020.105857

Inclusion of features derived from a mixture of time window sizes improved classification accuracy of machine learning algorithms for sheep grazing behaviours

Shuwen Hu[1], Aaron Ingham[2], Sabine Schmoelzl[3], Jody McNally[3], Bryce Little[2], Daniel Smith[4], Greg Bishop-Hurley[2], You-Gan Wang[1] and Yutao Li[2*]

[1]School of Mathematical Sciences, Queensland University of Technology, Gardens Point, QLD 4001, Australia

[2]CSIRO Agriculture & Food, 306 Carmody Road, St Lucia, QLD 4067, Australia

[3]CSIRO Agriculture & Food, New England Highway, Armidale, NSW 2350, Australia

[4]CSIRO DATA61, Private Bag 12, Hobart, Tasmania 7001, Australia

*Corresponding author: Yutao Li. CSIRO Agriculture & Food, 306 Carmody Road, St Lucia, QLD 4067, Australia. E-mail address: Yutao.Li@csiro.au.

Highlights

- Simultaneous inclusion of features derived from mixed time window sizes of sensor signal data significantly improved the sheep behaviour classification accuracy, in comparison to those from a single unique time window size.
- Using features derived from time windows of different lengths provided key information needed to accurately identify different behaviours that involve multiple movements of unequal duration.
- Using Random Forest and a mixed window size approach significantly improved the ability of identifying the walking behaviour, only accounted for 1% of the ground truth data.

Abstract:

Inertial motion sensors located on the animal have been used to study the behaviour of ruminant livestock. The time window size of segmented signal data can significantly affect the classification accuracy of animal behaviours. To date, there have been no studies evaluating the impact of a mixture of time window size features on the accuracy of animal behaviour classification. In this study, data was collected from accelerometers attached to the neck of 17 Merino sheep over a period of two days. We also recorded a ground truth dataset of behaviour recordings (grazing, ruminating, walking, and standing) over the same time period, We then investigated the ability of three machine learning approaches, Random Forest (RF), Support Vector Machine (SVM) and linear discriminant analysis (LDA), to accurately classify sheep behaviour. Our results clearly show that simultaneous inclusion of features derived from time windows of mixed sizes, ranging from 2-15 seconds, significantly improved the behaviour classification accuracy, in comparison to those determined from a single unique time window size. Of the three ML methods applied here, the Random Forest approach yielded the best results. Together our results show that including features obtained from mixed window sizes improved the classification accuracy of sheep behaviours.

Keywords: Mixture of time window sizes; Features ranking; Classification algorithm; Machine learning; Sheep behaviour; Accelerometer;

1. Introduction

Animal behaviour can be used to provide a mechanism for the early detection and quantitative assessment of animal health status (Martiskainen et al., 2009). Grazing and ruminating are two important behaviours for ruminants and continuous monitoring of animal eating behaviour provides vital information about ruminant health, productivity and welfare (Mansbridge et al., 2018). Traditional methods of animal monitoring are based on direct observation by human operators or assessment of video recording, both are labour-intensive, time-consuming, and prone to human error (Alvarenga et al., 2016). The rapid development of sensor technologies provides great opportunities for remotely monitoring animals in a range of applications (Brown et al., 2013; Schmoelzl et al., 2016). Sensor devices, especially those using an accelerometer that measures inertial acceleration associated with movement (usually on three different axes), can give a good insight into individual animal behaviour patterns (Fogarty et al., 2020).

To date, several studies have used sensors to study the behaviour of ruminant animals, especially in cattle. Greenwood et al. (2017) investigated the possibility of predicting pasture intake based on behaviour classification. They developed a simple algorithm to predict pasture intake of individual cattle. Rahman et al. (2018) compared tcattle behaviour classification using the information from sensors located on different parts of the body (collar, halter and ear) and found that different sensor placement can still achieve good classification accuracy providing that the feature variation between the training and testing animals is very small. For sheep behaviour classification, Guo et al. (2018) compared grazing behaviour of sheep on pasture with different sward surface heights using an inertial measurement unit sensor. They found that a high accuracy (> 95%) of identifying grazing behaviour from non-grazing behaviour could be achieved for all epochs (5s, 10s and 15s) with 10s being the best, regardless of sward surface heights. There are also commercially available monitoring systems that may be used to capture feeding behaviours for dairy cattle, such as Lely (Bar and Solomon, 2010) and MooMonitors (Verdon et al., 2018). However, these automatic systems cannot be directly applied to other species such as sheep as there are likely differences in accelerometer signal patterns between species. Further, the sensor may be impractical or unsafe for deployment on sheep due to limitations involving sensor size, shape, weight or method of attachment (Mansbridge et al., 2018)

Recently machine learning algorithms have become very popular and offer great potential in animal behaviour classification, largely because of their abilities in dealing with high dimension datasets (such as sensor data) and provide high prediction accuracy for complex phenotypes. For example, Dutta et al. (2015) applied six supervised machine learning methods, binary tree, LDA, naive Bayes, k nearest neighbour (kNN) and adaptive neuro fuzzy inference system (ANFIS), to classify five major cattle behaviours (Grazing, Ruminating, Resting, Walking and other behaviour). They achieved a high accuracy (96%) of classification by using the bagging ensemble classification with tree learner. For sheep behaviour classification, Mansbridge et al. (2018) found that RF performed the best when compared with SVM, kNN and adaptive boosting (AdaBoost) for sheep data collected by accelerometer/gyroscope sensor attached to the ear and collar. Guo et al. (2018) also reported that the LDA classifier was the best performer compared to binary tree, naive Bayes, kNN and ANFIS on classifying grazing activities.

Window size for signal segmentation is one of the crucial factors influencing on activity recognition. Banos et al. (2014) evaluated the impact of different window sizes (0.25 s – 7 s) on human activity classification with accelerometer data and found that the interval 1-2 s was the best trade-off between speed and accuracy of recognition. In sheep, Walton et al. (2018) investigated the effects of sensor position (ear and collar), sampling frequency (8, 16 and 32 Hz) of triaxial accelerometer and gyroscope sensor, and window size (3, 5 and 7 s) on behaviour classification, and concluded that the combination of 16Hz with 7 s window would produce the benefits of energy efficient and reasonable classification accuracy (91-93%) in a real-time sheep monitoring system. Smith et al. (2016) built a separate classifier for each of five cattle behaviours (grazing, walking, ruminating, resting and "other") using a "one vs all" ensemble on 24 Holstein-Friesian dairy cows. Of nine window sizes evaluated (1.5, 2.5, 5, 7.5, 10, 15, 20, 25, 30 s), they found that "the grazing, resting and rumination behaviours produced their highest mean F-score for the longest window of the study (30 s)".

Since there is no consensus about which time window size for signal segmentation and machine learning methods should be used, one obvious question is whether the features derived from a mixture of different time window sizes can be used to improve classification accuracy of livestock grazing behaviour. In this study, we aimed to: (1) Determine if animal behaviour classification could be improved by the simultaneous inclusion of features calculated fromtime windows of different sizes; (2) Compare the performance of three machine learning methods, RF, SVM and LDA in behaviour classification; (3) Investigate if new features from cumulative effects can improve the classification performance.

2. Materials and methods

2.1. Experiment design and data collection

Data collection was conducted on animals enrolled in a grazing trial, according to the Australian Code for the Use and Care of Animals in Research and Teaching, and approved protocols were approved by the CSIRO Armidale Animal Ethics Committee (Animal Research Authority 18-13). A total of 20 Merino ewes, habituated to human presence, were kept in a square mixed sward pasture paddock of 70 m x 70 m. A subgroup of 17 animals was randomly chosen for device deployment and behavioural annotation. Devices were attached around the neck of the animals with an elasticated strap (Fig 1) for a period of 48 hours. The sensor datasets were collected from Actigraph wGT3X-BT fitted with collars around the neck of Merino ewes, each containing a triaxial microelectromechanical systems (MEMS) accelerometer. The accelerometer sampled at a frequency of 30 Hz. The X-axis aligned approximately with the vertical or dorsoventral direction, the Y-axis with the craniocaudal direction, and the Z axis with the transverse or mediolateral direction (see Fig. 1 for illustration).

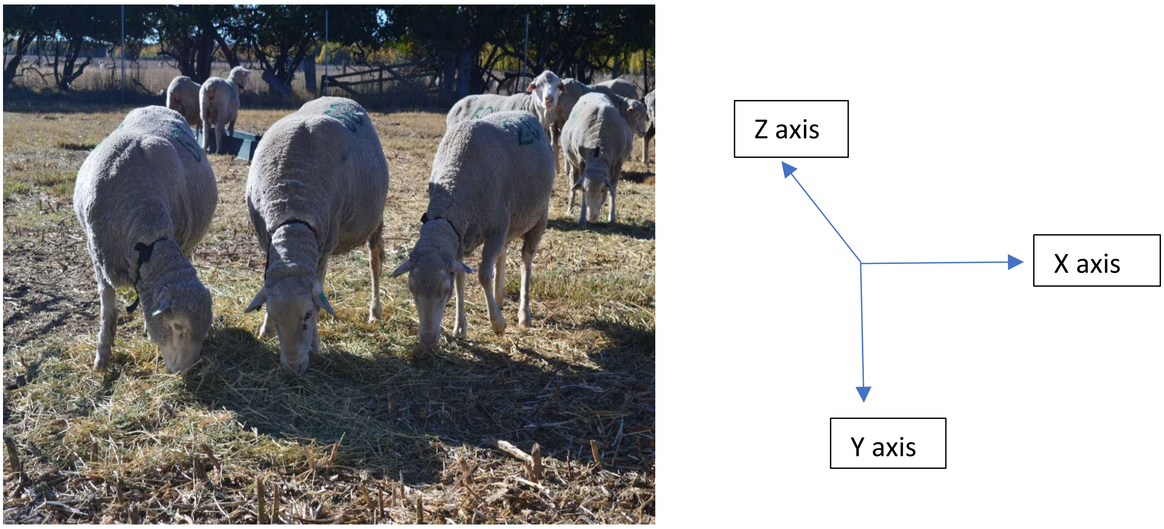
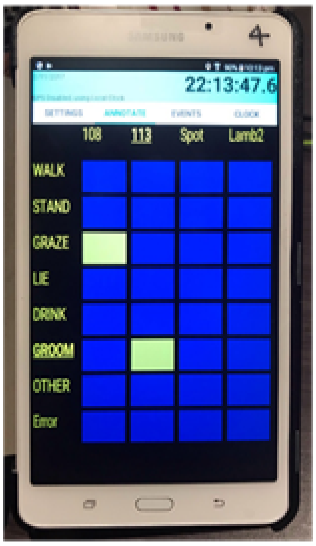


Fig. 1. Location of sensor and its orientation on sheep.

Annotation of behaviours was performed by direct observation of animals within the paddock environment. Trained operators were equipped with tablet devices with a custom-designed annotation application (CSIRO AnnoLOG v 1.0.23, (Little, 2018)) installed on a Samsung Galaxy Tab A 7.0” (Samsung, Seoul, Korea). The application allowed users to record time stamped behaviours during the recording period (Fig. 2a), and the output of the behaviour log was presented in tabular form (see Fig. 2b). Four behaviours (Grazing, Ruminating, Walking and Standing) were recorded. Recording time differed between animalsand included approximately 30 minutes of annotated behaviour information for each sheep.

a)



b)

Clock date	Clock start time	Animal ID	Behaviour	Clock end time
21/02/2019	8:43:54 AM	1	Graze	8:49:10 AM
21/02/2019	8:49:10 AM	1	Walk	8:49:16 AM
21/02/2019	8:49:16 AM	1	Stand	8:49:47 AM
21/02/2019	8:49:47 AM	1	Walk	8:50:08 AM
21/02/2019	8:50:08 AM	1	Stand	8:51:01 AM
21/02/2019	8:51:01 AM	1	Walk	8:51:05 AM
21/02/2019	8:51:05 AM	1	Graze	8:56:15 AM
21/02/2019	8:56:15 AM	1	Walk	8:56:26 AM
21/02/2019	8:56:26 AM	1	Stand	8:58:16 AM
21/02/2019	8:58:16 AM	1	Walk	8:58:28 AM
21/02/2019	8:58:28 AM	1	Stand	8:59:08 AM
21/02/2019	8:59:08 AM	1	Stand	9:00:26 AM

Fig. 2. a) Operator interface of the annotation tool CSIRO AnnoLOG v. 1.0.23. b) Tabular output of annotated behaviours for animal ID1.

2.2 Consolidation of sensor and ground truth datasets

We aligned the raw sensor data and the behaviour observations together via the time stamps (i.e. windows, every 1/30 second). A total of 1,052,475 data points was obtained. Fig. 3 demonstrates the distribution of four different behaviour classes from 17 Merino sheep. Note that Walking has a small representation.

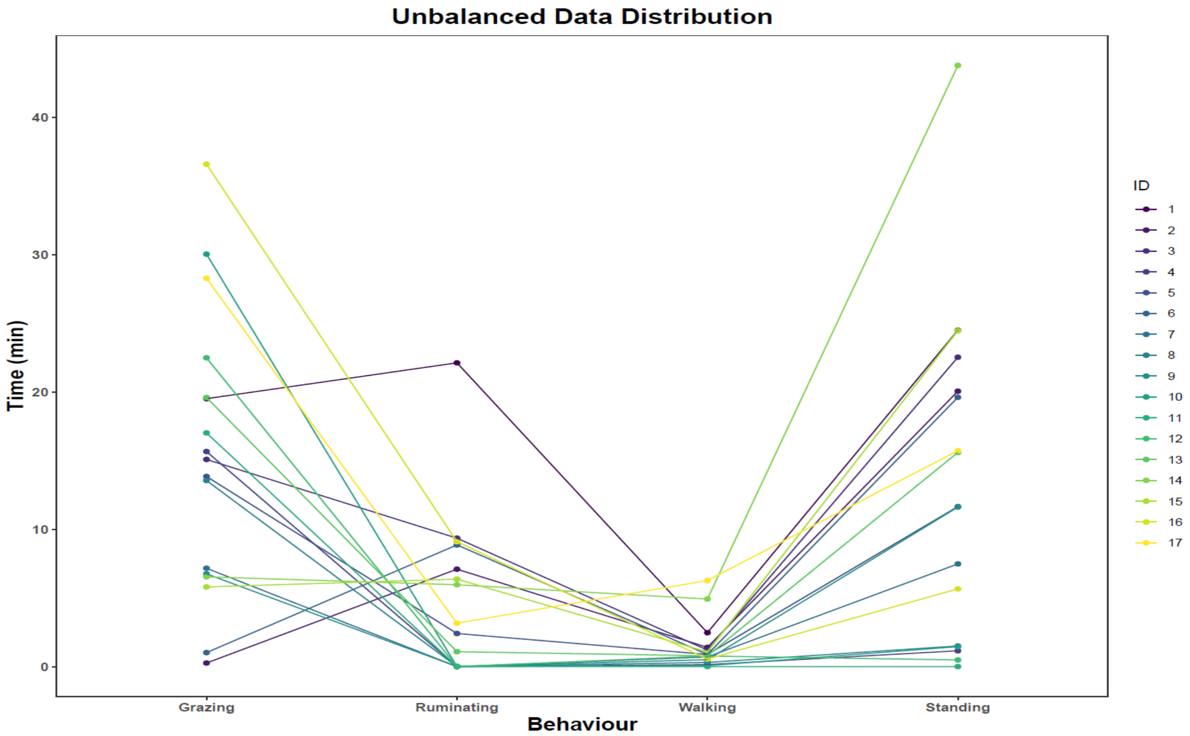


Figure 3. Unbalanced Behaviour Dataset.

2.3 Feature extraction from sensor data

In this study, five commonly used time window sizes, 1s, 2s, 5s, 10s and 15s, were applied for signal segmentation (Banous et al., 2014; Smith et al., 2016; Walton et al., 2018). For each time window, six basic statistical features, minimum, maximum, mean, standard deviation, skewness and kurtosis, were computed for each of three axes (X, Y and Z) acceleration data.

Instead of applying the features from five different time window sizes in isolation, we developed a new method that enabled us to conduct the classification analysis with all features from different time windows together in the same dataset (see Table 1). In brief, a time window of 1s was used as the basis. A total number of 28,425 intervals (bins) were generated with the 1s window. Using the average of acceleration magnitude values for X-axis (a_i) at 1s as an example, the corresponding average values for all 1s intervals were denoted as $a_1, a_2, \ldots a_{28425}$. When using 2s time window, a total number of intervals will be 14,213 with the average values being $b_1, b_2, \ldots b_{14213}$. To combine the features of average values from 1s and 2s windows together in the same dataset, individual average values of 2s window were used twice to meet 1s window requirement (see the second column in Table 1). By doing so, the average value from a 2s window remained unchanged for the 1st and 2nd 1s bins. The same concept applies to the average values derived for the time window sizes of 5s, 10s and 15s (See Table 1) as well as other statistical features (i.e. minimum, maximum, standard deviation, skewness and kurtosis values for X-axis acceleration magnitude measurement).

Table 1 Illustration of deriving average features from mixed time window sizes.

	time window				
Number of bins	1s	2s	5s	10s	15s
1	a_1	b_1	c_1	d_1	e_1
2	a_2	b_1	c_1	d_1	e_1
3	a_3	b_2	c_1	d_1	e_1
4	a_4	b_2	c_1	d_1	e_1
5	a_5	b_3	c_1	d_1	e_1
6	a_6	b_3	c_2	d_1	e_1
…	…	…	…	…	…
…	…	…	…	…	…
28423	a_{28423}	b_{14212}	c_{5685}	d_{2843}	e_{1895}
28424	a_{28424}	b_{14212}	c_{5685}	d_{2843}	e_{1895}
28425	a_{28425}	b_{14213}	c_{5685}	d_{2843}	e_{1895}

a, b, c, d and e are the average values of acceleration magnitude values from X axis.

Apart from six basic statistics features, we also explored new features of cumulative effects of raw data X_t, Y_t and Z_t. Table 2 illustrates how the cumulative effects are derived for the corresponding features named Xsum, Xvelocity, Xsummean, Xsum2, Xdis and Xsum2mean, using the time series data from X-axis acceleration magnitude measurements. The same methods were also applied for the computation of the features for the Y and Z axes acceleration magnitude measurements. In addition, the squared acceleration magnitude (acc) (Rahman et al. 2018) that considers the joint effects of X, Y and Z measurements were also included in this study. As the Y-axis detected the motion from front to back and the Z-axis detected the motion from the side to side, the interaction between Y and Z axis measurements were further examined using the features named dyz and interyz. For each window size and each statistics feature, there were 24 different metrics (including X, Y and Z axis data). A total of 720 features were generated.

Table 2: Illustration of calculation of additional features of cumulative effects of X-axis Acceleration magnitude measurement. *T*: the total number of intervals for a given time window; *t*: a particular interval.

Time	X	Xsum	Xvelocity	Xsummean	Xsum2	Xdis
1	x_1	x_1	x_1/30	x_1/1	x_1	x_1/30
2	x_2	$x_1 + x_2$	$(x_1 + x_2)/30$	$(x_1 + x_2)/2$	$2x_1 + x_2$	$(2x_1 + x_2)/30$
…	…	…	…	…	…	…
…	…	…	…	…	…	…
T	x_T	$(\sum_{t=1}^{T} x_t)$	$(\sum_{t=1}^{T} x_t)$/30	$(\sum_{t=1}^{T} x_t)$/T	$Tx_1 + (T-1)x_2 + \cdots + x_T$	$(Tx_1 + (T-1)x_2 + \cdots + x_T)/30$

Table Cont'd

Time	X	Xsum2mean	acc	dyz	interyz
1	x_1	x_1/1	$\sqrt{x_1^2 + y_1^2 + z_1^2}$	$\sqrt{y_1^2 + z_1^2}$	$y_1 z_1$
2	x_2	$(2x_1 + x_2)/2$	$\sqrt{x_2^2 + y_2^2 + z_2^2}$	$\sqrt{y_2^2 + z_2^2}$	$y_2 z_2$

…	…	…	…	…	…
…	…	…	…	…	…
T	x_T	$(Tx_1 + (T-1)x_2 + \cdots + x_T)/T$	$\sqrt{x_T^2 + y_T^2 + z_T^2}$	$\sqrt{y_T^2 + z_T^2}$	$y_T z_T$

2.4 Machine learning (ML) algorithms for classification

2.4.1 RF

RF is a tree-based ensemble method that builds a large collection of decision trees using training datasets, and validates predictions using testing datasets (Breiman, 2001). The library ranger in R (Wright and Ziegler, 2017) was applied for determining hyperparameters in RF. The final parameters were: mtry = 27, Ntree = 12 and default values for all other data.

2.4.2 SVM

SVM constructs a linear partition of the high-dimensional space into two subspaces for classification or regression (James et al., 2013). Intuitively, as the larger margins tend to provide lower classification errors, a good separation is obtained by a hyperplane that has the largest distance to the nearest training data. In this study, both linear and radial kernel functions were applied. The caret function in R (Kuhn, 2008) was applied. The final parameters for the analysis were: the average cost and sigma being 115 and 0.0014 for a radial function, and the average cost of 1 for a linear function.

2.4.3 LDA

LDA is a discriminant analysis that can separate a dataset into two or more classes. LDA assumes that the data within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all classes (James et al., 2013). In this study, we used the LDA classifier as a benchmark to compare the classification performance of RF and SVM.

For all three methods, a five-fold stratified cross-validation scheme was applied. That is, the dataset from 17 sheep was randomly partitioned into 5 subsets. Each subset was in turn used as a test dataset while the other 4 subsets were used as the training dataset. .

2.5 Performance of the classification

Similar to the approach previously used for cattle behaviour classification (Rahman et al., 2018), we chose four metrics, namely overall accuracy, precision, recall (also called sensitivity) and F1-score, to assess the classification performance of individual ML algorithms. For each target behaviour (e.g. Grazing, Ruminating, Walking, or Standing), a binary classifier was defined as the target behaviour (e.g. Grazing class) against a combined class of all remaining behaviour classes (Non-Grazing class). The calculation of four metrics can be found in Rahman et al. (2018).

3. Results

3.1 Behaviour classification performance using individual unique time window sizes

Table 3 presents the effects of different window sizes on the classification performance of three ML methods, RF, SVM (with linear kernel), SVM (with radial kernel) and LDA classifiers, when ignoring cumulative effects, squared acceleration magnitude and interaction effects between Y and Z axes. When the window size increased from 2s, 5s, 10s to 15s, the classification performance of ML classifiers for grazing behaviour showed a continued improvement except in SVM (radial kernel). The 15 second window size gave the best classification performance. Among all methods, RF showed the highest F1-score (0.876 – 0.889). However, when considering ruminating, increasing the window size resulted in a reduction in performance in both RF and SVM (radial kernel) with 2s being the best window size. SVM (linear kernel) and LDA had no ability or a weak power to identify the ruminating behaviour (Table 3). In all cases, none of the ML methods had the ability to recognize walking behaviour regardless of the time window applied, except a very weak power identified by RF (0.173) at 2s window.

Table 3. Effects of individual time window sizes on the behaviour recognition performance F1-score, when using RF, SVM and LDA. NA – not available.

		F1 score			
Method	Behaviour	2 sec	5 sec	10 sec	15 sec
RF	Grazing	0.876	0.879	0.886	0.889
	Ruminating	0.655	0.582	0.545	0.550
	Walking	0.173	NA	NA	NA
	Standing	0.794	0.782	0.783	0.781
SVM (linear kernel)	Grazing	0.832	0.850	0.864	0.875
	Ruminating	NA	NA	NA	NA
	Walking	NA	NA	NA	NA
	Standing	0.706	0.721	0.731	0.741
SVM (radial kernel)	Grazing	0.839	0.844	0.842	0.846
	Ruminating	0.286	0.289	0.254	0.224
	Walking	NA	NA	NA	NA

	Standing	0.720	0.719	0.705	0.692
LDA	Grazing	0.823	0.838	0.851	0.863
	Ruminating	0.129	0.057	NA	NA
	Walking	NA	NA	NA	NA
	Standing	0.697	0.709	0.719	0.724

When comparing the overall accuracies of behaviour classification for each of the three ML methods (Fig. 4), the RF performed best of all classifiers regardless of time window size. However, the window size did impact on the performance of SVM and LDA classifiers. The smaller window size (≤ 5s), SVM (radial kernel) performed better than SVM (linear kernel) and LDA. For both SVM (linear kernel) and LDA, increasing the window size improved the accuracy value, with 15s giving the highest value (Fig. 4). In general, SVM (linear kernel) produced higher accuracy than LDA, and it outperformed SVM (radial kernel) when the window size was ≥10s.

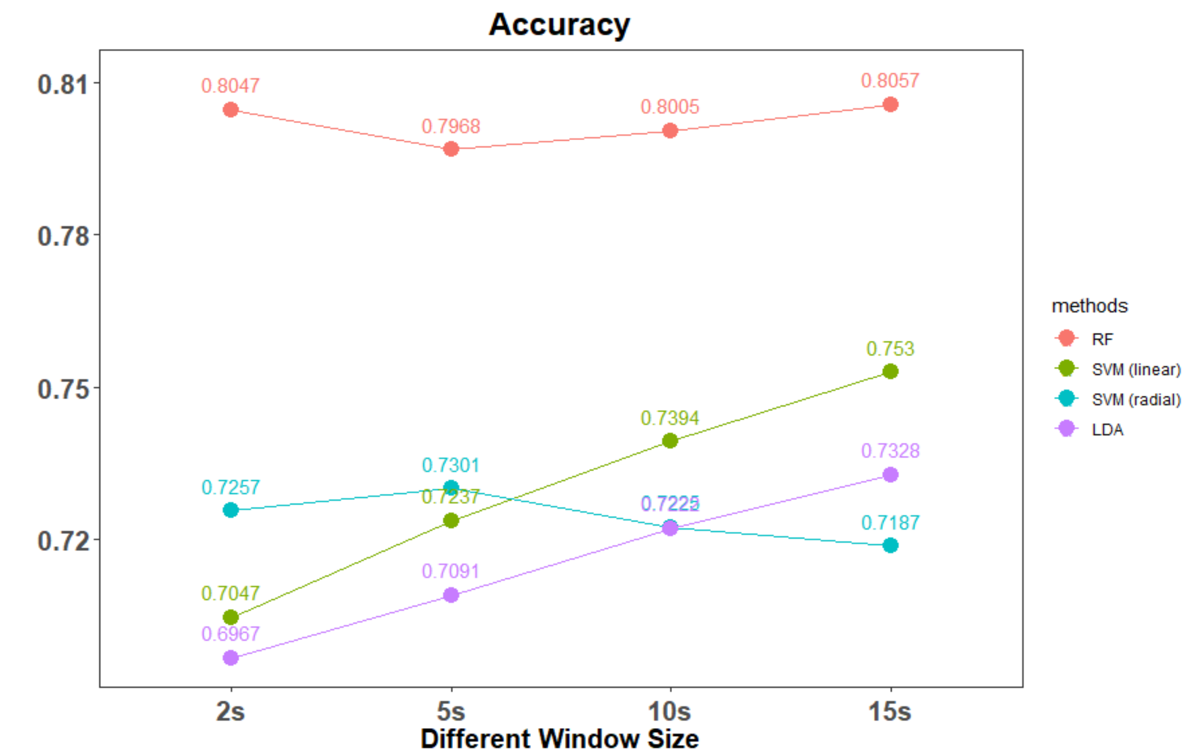


Fig. 4. The change of overall accuracy for individual ML methods with different time window sizes.

3.2 Classification of behaviours using mixed time window sizes

Fig. 5 illustrates the composition of different behaviours in the newly formed mixed time window dataset. Grazing behaviour accounted for the largest percentage (47.0%) of the data and the walking behaviour had the least representation at 1% of data.

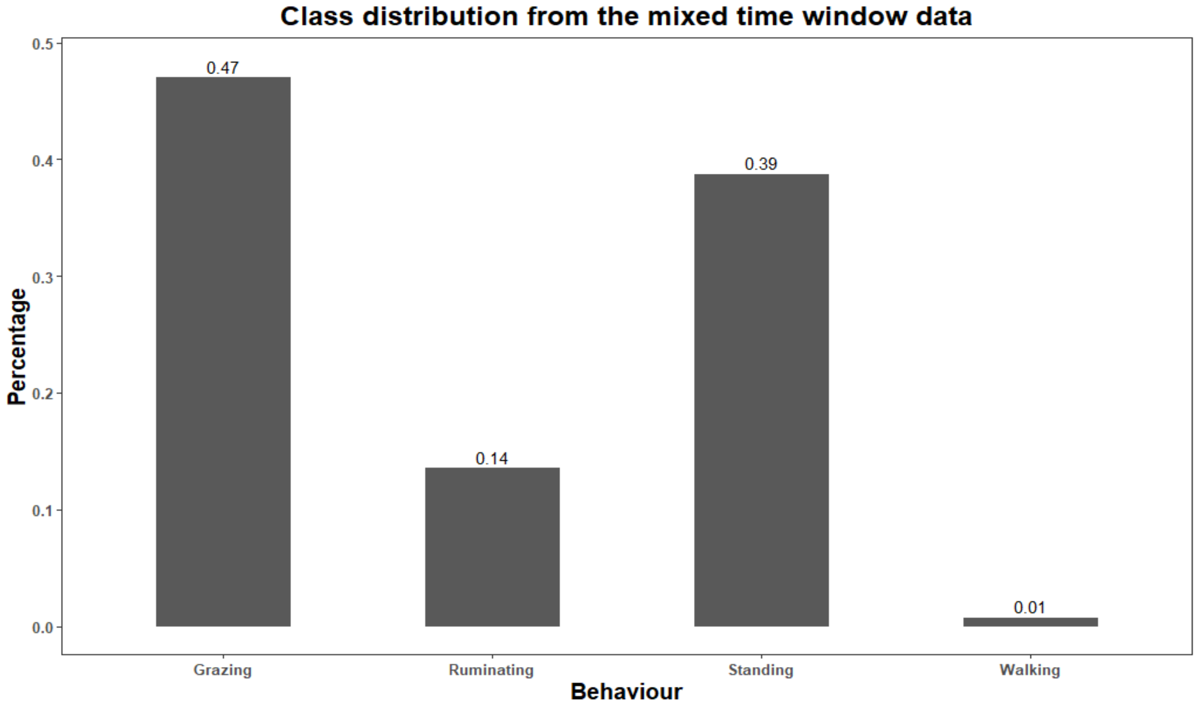


Fig. 5. The percentage distribution of four behaviours in the mixed time window data.

3.2.1 RF

Of the 720 features examined with the mixed window size dataset, the top 36 features (with the highest Gini index values) identified by the RF are shown in Fig. 6. Among the top 36 features, the features derived from different window sizes (1s, 2s, 5s, 10s and 15s) all contributed to the classification accuracy of different behaviours. One other noteworthy observation was that the features derived from using the cumulative effects (i.e. with “sum” in the labels) were also among the top contributing features.

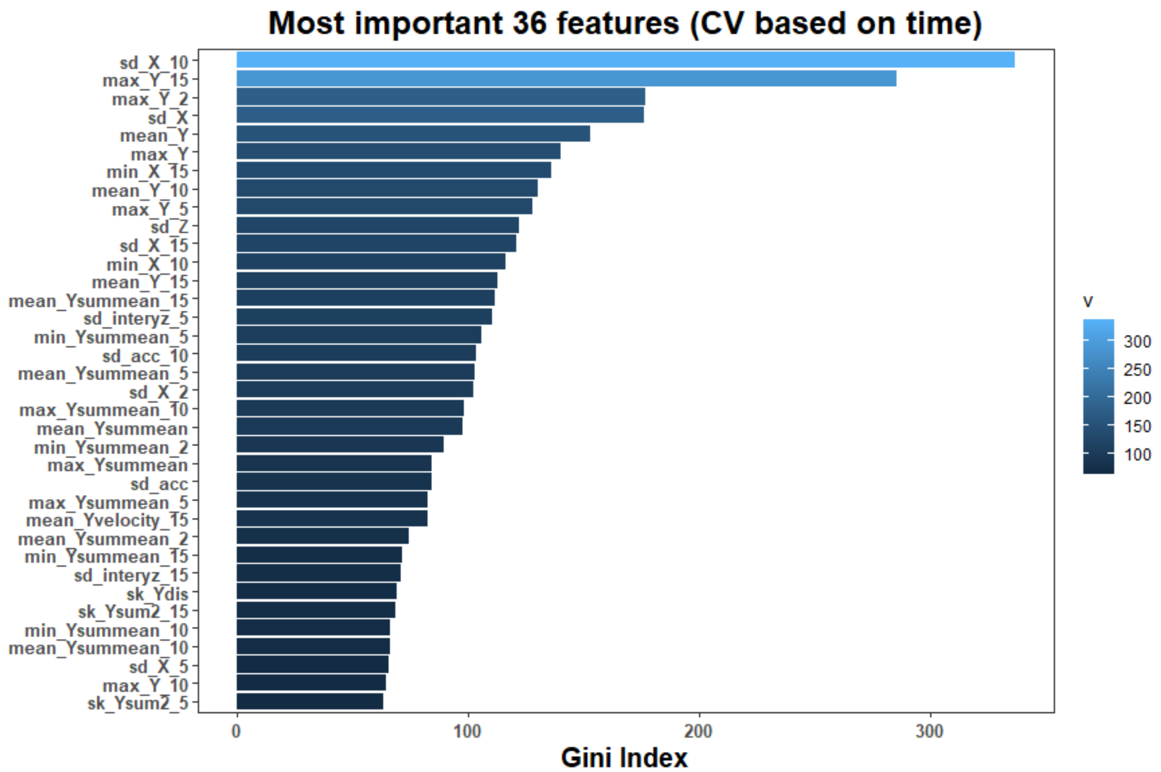


Fig. 6. The list of top 36 features selected by RF.

Next, we compared the behaviour classification performance of different subsets of the top features from RF with that of 720 features. The top 3 features produced an overall accuracy of 0.946 (Fig. 7(a)). The accuracy value increased to 0.986 with the top 9 features and to almost 1 (0.999) with the top 27 features. Similar trends were observed irrespective of the performance metric used (See Figs 7b, 7c and 7d). For all four behaviour classes, the RF identified the individual behaviour classes with high precision (>0.970), sensitivity (Recall > 0.950) and F1-score (>0.970) when either the top 9 or 27 features were applied (Fig. 7). Even with the top three features, the results show that for the most difficult behaviour to classify – walking, there was 0.910 (precision), 0.780 (sensitivity) or 0.850 (F1-score) achieved.

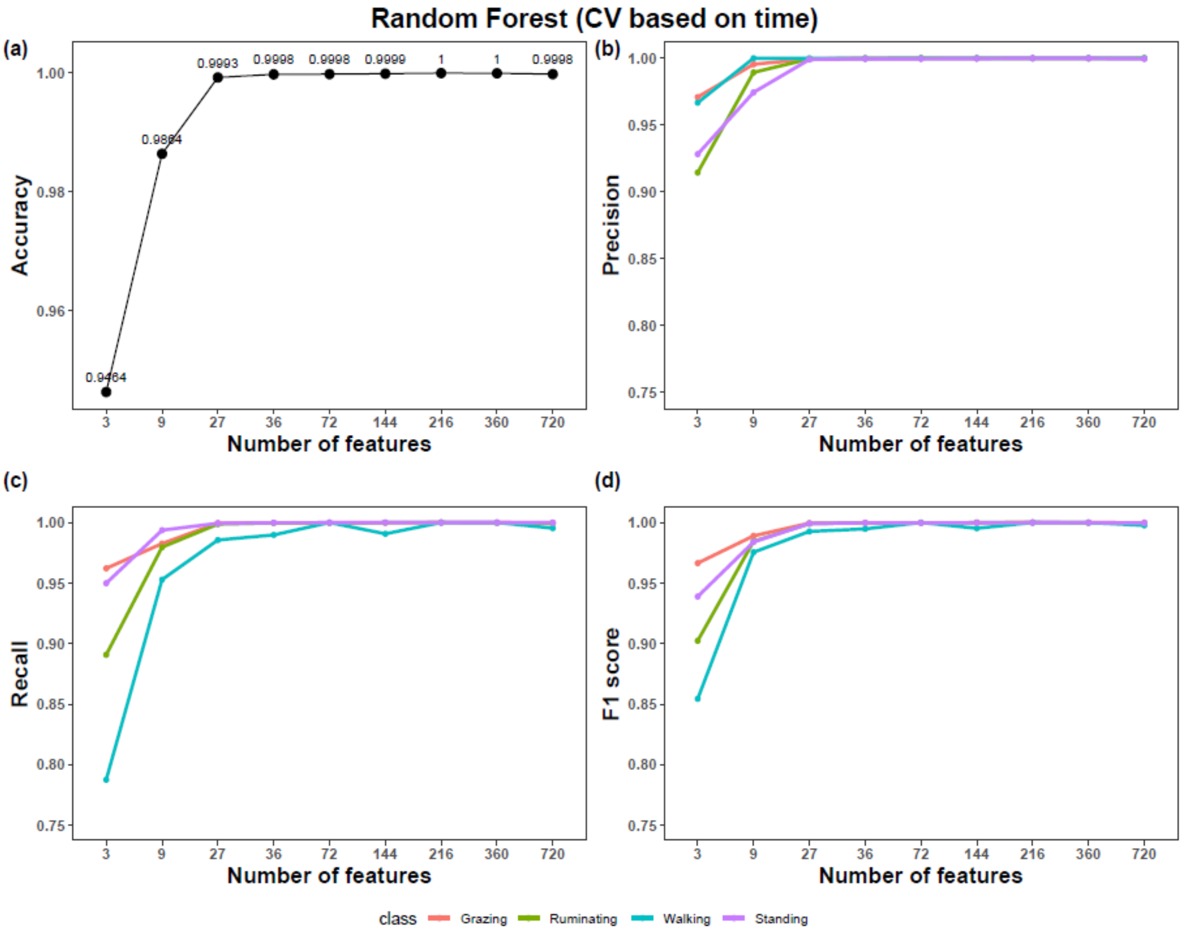


Fig. 7. The overall accuracy, precision, recall and F1 score values from using the different number of top ranked features chosen from RF for behaviour classification.

3.2.2 SVM

The SVM classifier was evaluated with both linear and radial kernels. The overall accuracy of SVM (linear kernel) was largely dependent on the number of features applied for classification (Fig. 8a). While the lowest accuracy was 0.741 with three top features, the highest accuracy (0.933) was achieved with all the 720 features.

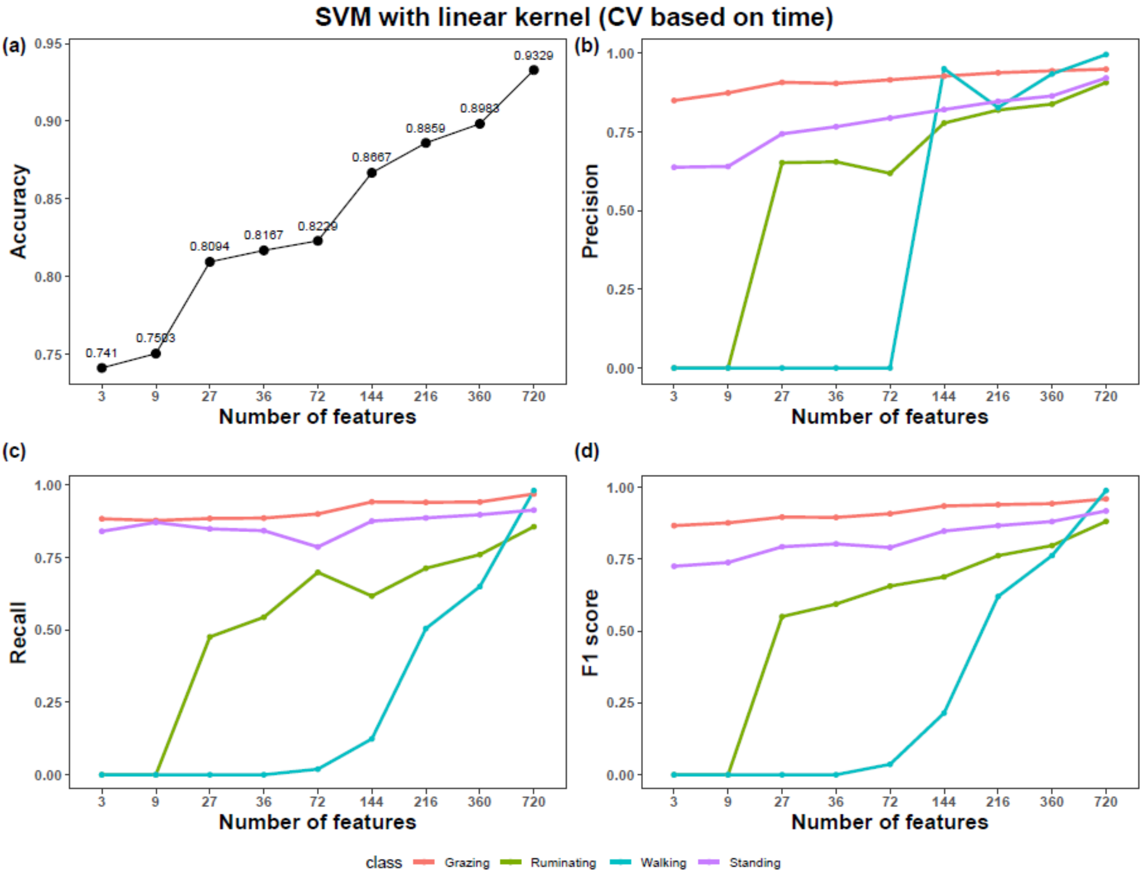


Fig. 8. The over accuracy, precision, recall and F1 score values from the different number of top ranked features chosen from SVM with a linear kernel function and the mixed time window sizes.

When investigating the classification performance of SVM (linear kernel) for individual behaviour classes Fig. 8), increasing the number of features slightly improved the classification performance for grazing and standing behaviours(Figs. 8b-8d) . However, for walking and ruminating, the change in the number of features had a significant impact on the classification performance (Fig. 8d). For example, SVM (linear kernel) had no or little power to correctly classify the walking behaviour until the number of features reached more than 144 (Fig. 8d, F1 score = 0.210).

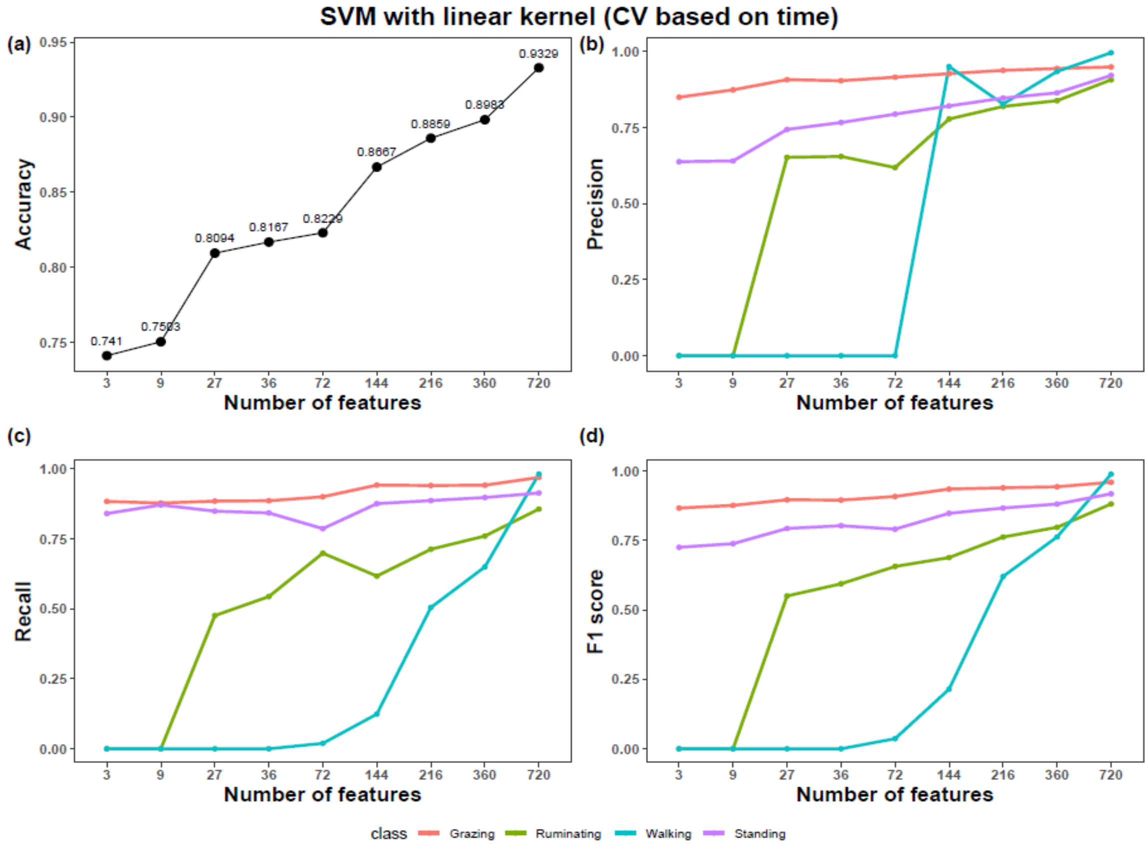


Fig. 8. The overall accuracy, precision, recall and F1 score values from the different number of top ranked features chosen from SVM with a linear kernel function and the mixed time window sizes.

When comparing SVM (radial kernel) with SVM (linear kernel), the overall accuracy was significantly improved by 8.40% with top three features, and by 17.70% with 27 features (Fig. 9a vs Fig. 8a). SVM (radial kernel) performed well in classification of all individual behaviours when the number of features was more than 27. This can be demonstrated by high precision (> 0.900), reasonable sensitivity (> 0.750) and medium to high F1 score (> 0.850) in Fig 9. The most noticeable results are for the walking behaviour.

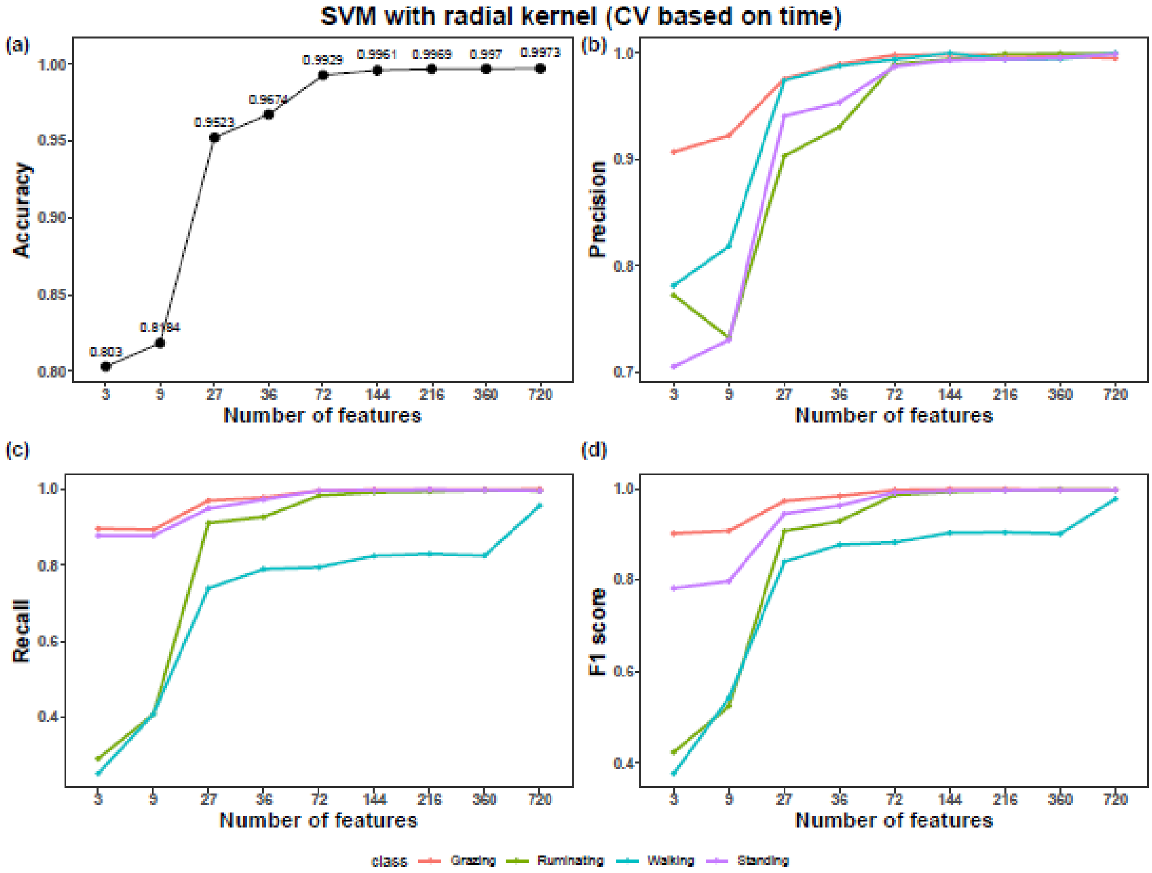


Fig. 9. The overall accuracy, precision, recall and F1 score values from the different number of top ranked features chosen from SVM with a radial kernel function and the mixed time window sizes

3.2.3 LDA

Fig. 10 presents the classification results from the LDA classifier. The overall accuracy from the LDA classifier followed a similar trend as the SVM classifier with linear kernel (Figure 10(a) vs Figure 8(a)). However, the SVM (linear kernel) still gave 2.750% (with 3 features) - 7.010% (with 720 features) better accuracy values than the LDA classifier. When comparing the classification performance for individual behaviour, again, LDA had a very similar performance to the SVM with linear kernel for the grazing, ruminating, and standing behaviours. For the walking behaviour, although the LDA classifier had an overall low precision (<0.300), low recall (<0.500) and low F1 score (<0.375), surprisingly, it did show some ability to recognise the walking behaviour when the number of features were less than 72. This was a stark contrast to the SVM classifier with linear kernel for the walking behaviour.

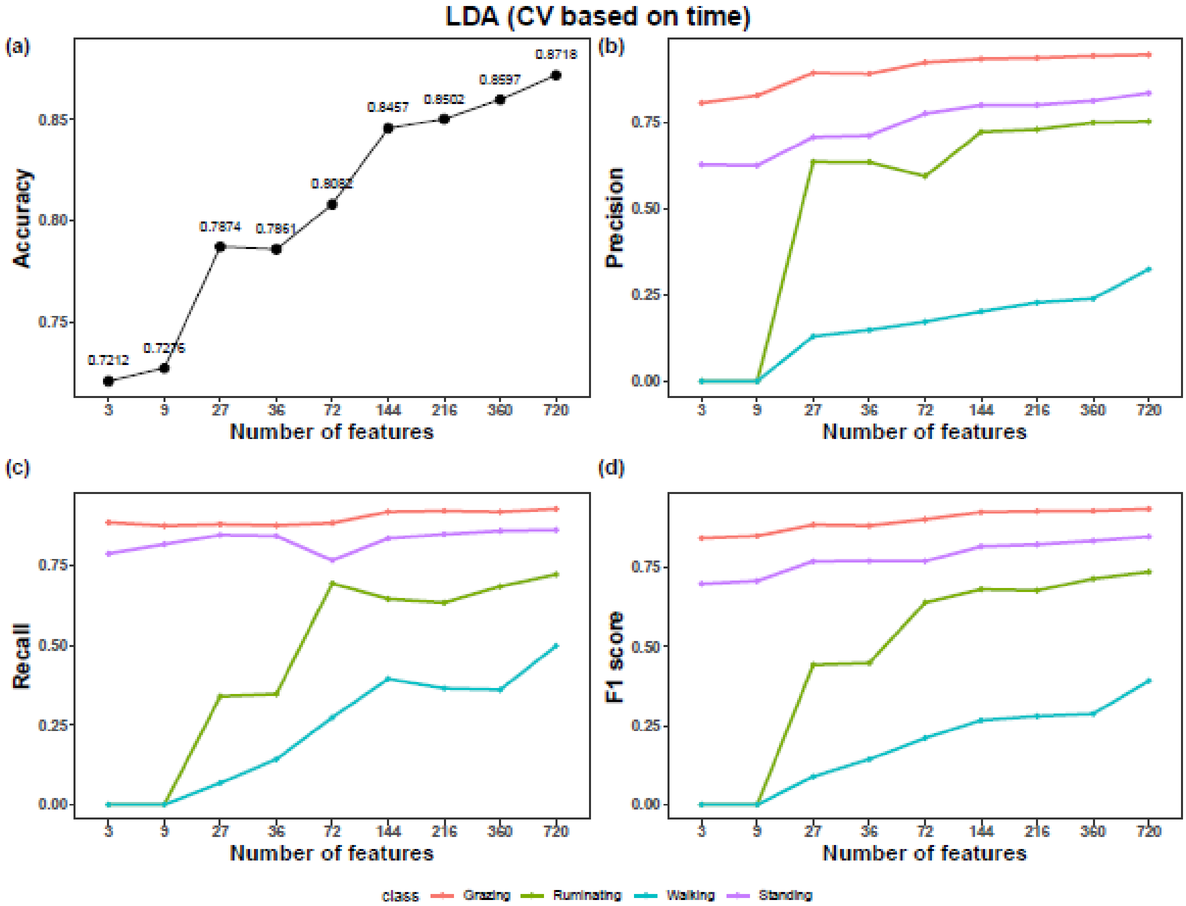


Fig. 10. The overall accuracy, precision, recall and F1 score values from the different number of top ranked features chosen from LDA with the mixed time window sizes.

4. Discussion

Accurate classification of animal behaviour from sensor derived data is influenced by a number of factors, including experimental design, sensor placement position, data sample rate, signal segmentation window size, feature selection and different machine learning methods applied (Banos et al., 2014; Rahman et al., 2018; Walton et al., 2018; Mansbridge et al., 2018;

Fogarty et al., 2020). Among them, window size, feature selection and analytical methods play critical roles.

4.1. Window size

To our knowledge, this is the first study that evaluated the classification of animal behaviour using features derived from a mixed window size approach. It is a common practice to spend extensive amount of time to evaluate a range of time window sizes in order to identify an optimal size that produces reasonable classification accuracy from sensor datasets. In this study, developing a new combined window size approach enabled us not only to minimise subjective selection of window sizes, but also to systematically and simultaneously consider the features from multiple window sizes together to capture the irregular duration of animal behaviours, that occur both within and across behaviours.

Most importantly, we have demonstrated that using mixed window sizes in combination with the ML method RF, significantly improved the classification accuracies for all behaviour classes, especially for walking and ruminating. The overall accuracy, precision, recall and F1 score for individual behaviour, when mixed window sizes were applied, were in strong contrast to those when individual time window size was applied. This is largely due to the features derived from different window sizes being inter-related. The correlations between these features of different window sizes provide additional information for ML methods to correctly identify individual animal behaviours. The biological basis for this can be explained by the need to classify specific short duration movements as a component of a longer movement. For instance, grazing might involve the lowering of the head and several biting events. Only classifying the lowering of the head would likely lead to inaccurate classification as would the classification of biting-like behaviours alone. The combination of the two is however likely to be much more informative. In contrast, when analysing the datasets with the features from one window size only, many feature correlations were not accounted for, especially in the cases where mixed behaviours occurred in a given time window (i.e. unequal length of animal behaviour), errors were expected to arise.

4.2. Feature selection

To date an optimum number of features that can be used for classifying animal behaviours has varied greatly between studies, depending on the nature of sensor and behaviour data and analytical methods applied. For example, by analyzing the data collected by collar mounted motion sensors, Guo et al. (2018) found that the top 5 features, mean of accelerometer Z-axis, entropy of accelerometer Y-axis, entropy of accelerometer Z-axis, mean of gyroscope X-axis and mean of gyroscope Y-axis, can be used in classifying the grazing versus non-grazing activities in sheep. Mansbridge et al. (2018) identified 39 being the optimum number of features that can be used successfully in the classification of eating behaviors in sheep with a high accuracy (91% for ear and 92% for collar data). These features ranged from dominant frequency, zero crossings, signal area, spectral entropy, to basic statistics such as mean, min, max, standard deviation, and kurtosis.

In this study for each axis acceleration magnitude measurements, apart from the common features derived from 6 basic statistics and the squared acceleration magnitude (acc), we also evaluated the effects of new features of cumulative effects of measurements for a given time window size on classification performance. The reasons for using these features include: 1) to properly evaluate the efficiency of the new approach - a mixed time window sizes, it is crucial to compare the new method with conventional methods using commonly used features. 2) using new features from the accumulative effects was to examine if they could better capture actual change of motion movements for mixed behaviours. When applying the RF with the mixed window sizes, of 720 features, we found 9 top ranking features that contributed the most in the classification accuracy were all related to basic statistics (e.g. SD_X_10, max_Y_15 Max_Y_2, sd_X, mean_Y, max_Y, min_X_15, mean_Y_10 and max_Y_5, see Figure 6) of X and Y-axis measurements. The X and Y-axis in this study aligned with upward and downward, and front to back movements of the neck, respectively hence kinetically related to grazing and ruminating behaviors. Among the 27 top ranking features, there were also 9 features derived from the average of the accumulative effects of Y axis (Ysummean, Fig. 6) for different window sizes. This indicates that the average cumulative values of Y axis from mixed window sizes may have better-reflected changes of acceleration, therefore contribute to additional improvement in the accuracy of behaviour classification.

Feature selection can also be impacted by where signal information sensor placement position is one of the key factors that impact classification accuracy of sheep behaviour (Barwick et al., 2018).

4.3. Machine learning methods

Different machine learning methods could yield different outcomes when dealing with unbalance behaviour classes. Among three machine learning algorithms (RF, SVM and LDA) evaluated, the RF classifier performed the best with over 99% accuracy, followed by the SVM (radial kernel), the SVM (linear kernel) and LDA classifiers. RF has been known to produce good classification accuracies in sheep behaviour (Alvarenga et al., 2016), especially classification of grazing and rumination behaviour (Walton et al., 2018; Mansbridge et al., 2018). This is mainly due to its great ability in handling non-linearly correlated data and robustness to noise (Mansbridge et al., 2018). SVM (radial kernel) performed better than the SVM (linear kernel) and LDA classifiers, also because its radial kernel function can non-linearly separate the sensor signals associated with irregular length of individual behaviours. There are other machine learning methods that can also be applied to provide good classification accuracy of sheep behaviour, depending on the sources of signal information of sensor placement (e.g. ear, collar, or leg, Barwick et al., 2018), and trade-off between energy consumption and classification accuracy (Le Roux et al, 2018). Future work needs to be carried out with ensemble classifiers in which several different classifiers are trained simultaneously and their classification decisions can be combined at the end (D'Este et al., 2014; Dutta et al., 2015).

All the results presented in this study were obtained using a five-fold stratified cross-validation scheme based on time, rather than the cross-validation based on individual sheep. The initial analysis using a five-fold cross-validation approach based on subsets of sheep produced much worse results than that of a stratified cross-validation approach (see supplementary results I). The primary reason was due to the small number of animals used in this study and the big variation between individual sheep behaviours in feature space. Therefore, it is difficult to obtain the consistent results in test datasets when different subsets of sheep were used as training datasets. To minimize the impact of individuality on classification accuracy of animal behaviours in future sensor application, it will be crucial to: 1) obtain results from larger numbers of animals; and 2) explore and validate results using a number of repeated k-fold CV to improve the prediction on both population and individual results

Limitations of the study that may have influenced the results, include thesmall number of animals and limited paddock space (70 m x 70 m) used during the experiment. However, the study aimed to serve as a proof of concept that incorporation of features calculated across time windows of different lengths has the potential to improve classification accuracy. We believe this principle has been demonstrated and the broader applicability of the approach can be tested in future trials involving larger numbers of animals.

4. Conclusions

This study demonstrated that the sheep behaviours of grazing, ruminating, walking, and standing can be differentiated with a high accuracy using ML algorithm RF and a mixed window size approach. One clear benefit of applying the RF, mixed window approach was the ability to accurately classify walking behaviour, that only accounted for 1% of the ground truth data, when conventional approaches failed. One possible explanation for this outcome is that behaviour classification requires the information contained in features derived from time windows of different length to provide the context needed for accurate identification.

Acknowledgements

We would like to acknowledge the financial support SH received from the Queensland University of Technology. We thank Dr Peter Hunt for allowing us to deploy devices on animals enrolled in his experimental work.

References

Alvarenga, F.A.P., Borges, I., Palkovic, L., Rodina, J., Oddy, V.H., & Dobos, R.C. (2016). Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. *Applied Animal Behaviour Science*, 181, 91–99.

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., & Rojas, I. (2014). Window size impact in human activity recognition. *Sensors*, 14, 6474-6499.

Bar, D., & Solomon, R. (2010, March 2-5). Rumination collars: What can they tell us. [Paper presentation]. The First North American Conference on Precision Dairy Management, Toronto, ON, Canada.

Barwick, J., Lamb, D., Dobos, R., Schneider, D., Welch, M., & Trotter, M. (2018). Predicting Lameness in Sheep Activity Using Tri-Axial Acceleration Signals. *Animals (Basel)*, 8(1),12.

Brown, D.D., Kays, R., Wikelski, M., Wilson, R., & Klimley, A.P. (2013). Observing the unwatchable through acceleration logging of animal behavior. *Animal Biotelemetry*, 1, 20.

D'Este, C., Timms, G., Turnbull, A., & Rahman, A. (2014). Ensemble aggregation methods for relocating models of rare events. *Engineering Applications of Artifical Intelligence,* 34, 58-65.

Dutta, R., Smith, D., Rawnsley, R., Hurley, G.B., Hills, J., Timms, G., & Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*, 111, 18-28.

Fogarty E.S., Swain D.L., Cronin G.M., Moraes L.E., & Trotter M. (2020). Behaviour classification of extensively grazed sheep using machine learning. *Computers and Electronics in Agriculture*, 169: 105-175.

Gonzalez, L.A., Bishop-Hurley, G.J., Handcock, R.N., & Crossman, C. (2015). Behavioural classification of data from collars containing motion sensors in grazing cattle. *Computers and Electronics in Agriculture*, 110, 91-102.

Guo, L., Welch, M., Dobos, R., Kwan, P., & Wang, W. (2018). Comparison of grazing behaviour of sheep on pasture with different sward surface heights using an inertial measurement unit sensor. *Computers and Electronics in Agriculture*, 150, 394-401.

Greenwood, P.L., Paull, D.R., McNally, J., Kalinowski, T., Ebert, D., Little, B., Smith, D.V., Rahman, A., Valencia, P., Ingham, A.B., & Bishop-Hurley, G.J. (2017). Use of sensor-determined behaviours to develop algorithms for pasture intake by individual grazing cattle. *Crop and Pasture Science*, 68, 1091-1099.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer: Berlin, Germany.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal for Statistical Software,* 28(5): 26.

Le Roux

Little, B. (2018). CSIRO AnnoLOG (1.0.23) [Mobile application software]. Retrieved from https://play.google.com/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer: Heidelberg, Germany.

Mansbridge, N., Mitsch, J., Bollard, N., Ellis, K., Miguel-Pacheco, G. G., Dottorini, T., & Kaler, J. (2018). Feature Selection and Comparison of Machine Learning Algorithms in

Classification of Grazing and Rumination Behaviour in Sheep. *Sensors*, 18(10), 3532. doi:10.3390/s18103532

Martiskainen, P., Jarvinen, M., Skon, J.-P., Tiirikainen, J., Kolehmainen, & M., Mononen, J. (2009). Cow behavior pattern recognition using a three-dimensional accelerometer and support vector machines. *Applied Animal Behaviour Science*, 119, 32–38.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for thefeature selection and classification of spectral data. *BMC Bioinformatics*, 10, 213.

Rahman, A., Smith, D.V., Little, B., Ingham, A.B., Greenwood, P.L., & Bishop-Hurley, G.J. (2018). Cattle behaviour classification from collar, halter, and ear tag sensors. *Information Processing in Agriculture*, 5, 124-133.

Smith, D., Rahman, A., Bishop-Hurley, G.J., Hills, J., Shahriar, S., Henry, D., Rawnsley, R. (2016). Behavior classification of cows fitted with motion collars: decomposing the multi-class classification problem into a set of binary problems. *Computers and Electronics in Agriculture*, 131, 40-50.

Verdon, M., Rawnsley, R., Raedts, P., & Freeman, M. (2018). The Behaviour and Productivity of Mid-Lactation Dairy Cows Provided Daily Pasture Allowance over 2 or 7 Intensively Grazed Strips. *Animals*, 8, 115.

Walton, E., Casey, C., Mitsch, J., Vazquez-Diosdado, J.A., Yan, J., Dottorini, T., Ellis, K.A., Winterlich, A., & Kaler, J. (2018). Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour. *Royal Society Open Science*, 5, 171442. http://dx.doi.org/10.1098/rsos.171442.

Watanabe, N., Sakanoue, S., Kawamura, K., & Kozakai, T. (2008). Development of anautomatic classification system for eating, ruminating and resting behavior of cattle using an accelerometer. *Japanese Society of Grassland Science*, 54, 231–237.

Wright, M. N. & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal for Statistical Software,* 77(1): 17.

Yoshitoshi, R., Watanabe, N., Kawamura, K., Sakanoue, S., Mizoguchi, R., Lee, H.-J., & Kurokawa, Y. (2013). Distinguishing cattle foraging activities using an accelerometry-based activity monitor. *Rangeland Ecology& Management*, 66 (3), 382–386.

Appendix A. Supplementary material

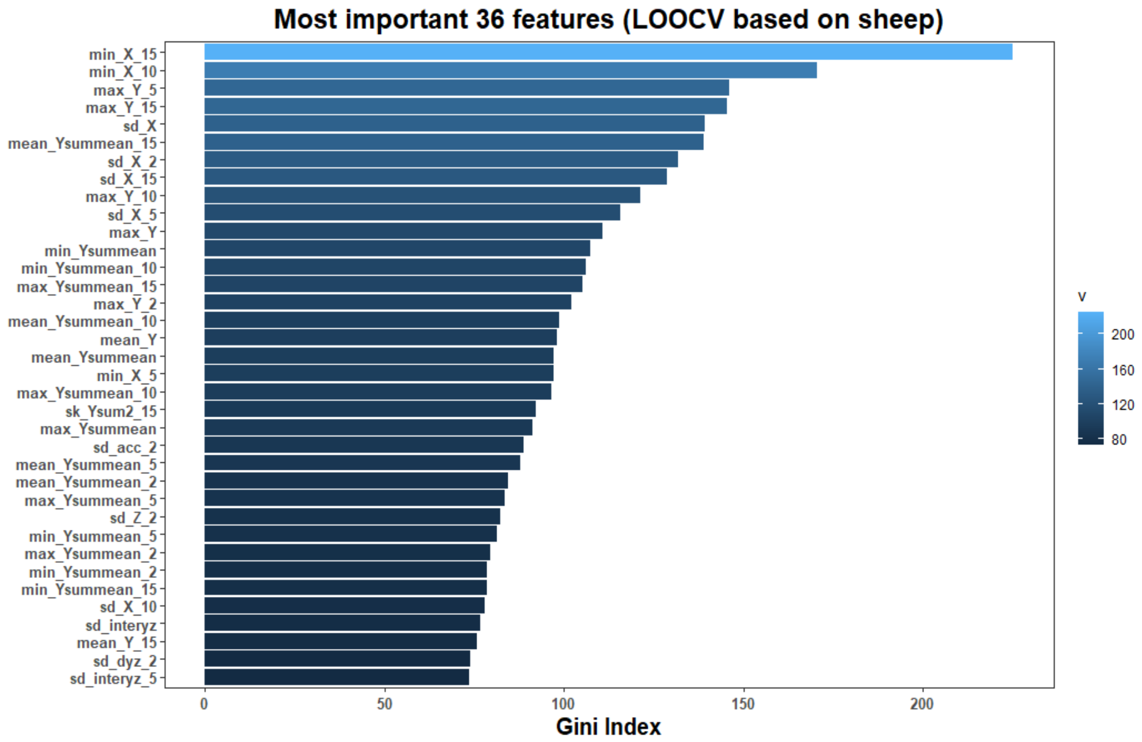


Fig. 1s. The list of 36 of the most important features selected by RF based on the leaving one animal out cross-validation scheme.

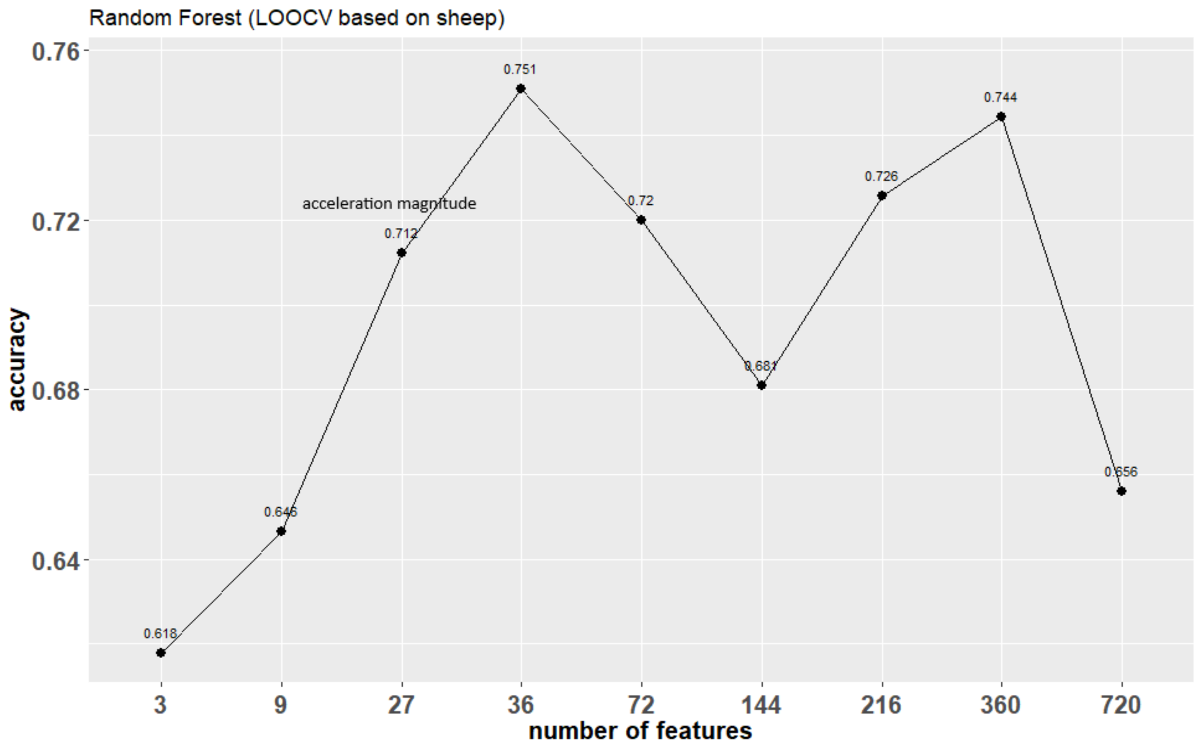


Fig. 2s. The over accuracy values from the different number of top ranked features chosen from Random Forest (RF) when using the mixed time window approach and leaving one animal out cross-validation scheme.