# XPM: An explainable-by-design pattern-based estrus detection approach to improve resource use in dairy farms

Issei Harada, Kevin Fauvel, Thomas Guyet, Véronique Masson, Alexandre Termier, Philippe Faverdin

The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)
Workshop AIAFS

# XPM: An Explainable-by-Design Pattern-Based Estrus Detection Approach to Improve Resource Use in Dairy Farms

**Issei Harada[1], Kevin Fauvel[1], Thomas Guyet[1], Véronique Masson[1], Alexandre Termier[1], Philippe Faverdin[2]**

[1]Univ Rennes, Inria, CNRS, IRISA, France
[2]PEGASE, INRAE, AGROCAMPUS OUEST, France

**Abstract**

A powerful automatic detection of estrus, the only period when the cow is susceptible to pregnancy, is a key driver to help farmers with reproduction management and subsequently to improve milk production resource use in dairy farms. Automatic solutions to detect both types of estrus (behavioral and silent estrus) based on the combination of affordable phenotyping data (activity, body temperature) exist, but they do not provide faithful explanations to support their alerts and in ways that farmers can understand based on the behaviors they could observe in animals. In this paper, we first propose XPM, a novel pattern-based classifier to detect both types of estrus with real-world affordable sensor data (activity, body temperature) which supports its predictions with perfectly faithful explanations. Then, we show that our approach performs better than a commercial reference in estrus detection, driven by the detection of silent estrus. Finally, we present the explainability of our solution which stems from the communication to the farmers the presence and/or absence of a limited number of patterns determinant of estrus detection, therefore reducing solution mistrust and supporting farmers' decision-making.

## 1 Introduction

Improve resource use in dairy farms is one of the most important steps towards meeting both food production and environmental goals (Searchinger et al. 2018). The detection of determining events for milk production like estrus, the only period when the cow is susceptible to become pregnant, is crucial for an optimal resource use. Reproduction issue is the most prevalent reason for cow culling (Bascom and Young 1998), and reproduction performance directly impacts milk production (Inchaisri et al. 2010).

Traditionally, estruses are detected visually by observing cow behaviors as cows significantly increase their activity during estrus (Silper et al. 2015; Gaillard et al. 2016). However, the visual detection rate is generally less than 50% for dairy cows (Peralta, Pearson, and Nebel 2005) due to physiological reasons (e.g., 35% of the estruses are not associated with obvious behavioral signs - silent estruses (Palmer et al. 2010), and some signs are expressed during the night (Kerbrat and Disenhaus 2004)). These observations call for the development of automatic estrus detection solutions to support farmers. The gold standard is estrus estimation using progesterone dosage in milk (Cutullic et al. 2011; Tenghe et al. 2015). However, the cost of this solution limits its extensive application. Commercial solutions based on affordable sensor data (activity, body temperature) have been developed. Nonetheless, their adoption rate remains moderate (Steeneveld and Hogeveen 2015) as these solutions suffer from insufficient performance and from a lack of explanations supporting alerts.

Most of existing studies about the application of data science techniques to improve automatic estrus detection (Dolecheck et al. 2015; Minegishi, Heins, and Pereira 2019; Ma et al. 2020) present detection algorithms from a performance-only perspective and do not discuss their explainability. In addition, none of these studies uses the currently recognized method for behavioral and silent estrus identification as labels (progesterone profiles), so their estrus labeling methods are not exhaustive (maximum 65% of total estruses), therefore the high detection performance reported do not reflect real-world conditions. A recent study (Fauvel et al. 2019) proposes an explainability method (SHAP (Lundberg and Lee 2017) - post-hoc model-agnostic method) along with an ensemble detection method (LCE - a "black-box" model, i.e. a complicated-to-understand model) and uses the method of progesterone dosage in milk as the reference to obtain an exhaustive estrus labeling. However, the explanations from a post-hoc model-agnostic method like SHAP cannot be perfectly faithful with respect to the original model (Rudin 2019). Faithfulness is critical to reduce solution mistrust from the farmers as it corresponds to the level of trust an end-user can have in the explanations of model predictions, i.e. the level of relatedness of the explanations to what the model actually computes.

Therefore, *(i)* we propose a new eXplainable-by-design Pattern-based classifier for Multivariate time series (XPM) to detect both types of estrus (behavioral/silent) with combined real-world affordable sensor data (activity, body temperature), which provides perfectly faithful explanations in ways that farmers can understand based on the patterns they could observe in animals. A pattern-based detection, i.e. small conjunctions of symbols (Fournier-Viger et al. 2017), is more informative to the user than other explainable machine learning methods based on subseries (e.g., shapelet methods (Karlsson, Papapetrou, and Boström 2016)) as they provide information about the relevant relations among the elements of the subseries (e.g., order, time gap). Then, *(ii)* we show that XPM performs better than a commercial reference in estrus detection, driven by the detection of silent estrus. Finally, *(iii)* we present the limited set of patterns needed for these detections and how it can be used to support farmers' decision-making.

## 2 Related Work

**Pattern-Based Classification** Multiple studies have proposed different classification methods based on pattern features, including itemset-based approaches (Cheng et al. 2007, 2008) and sequence-based classification (Buza and Schmidt-Thieme 2010; Fradkin and Mörchen 2014). An *itemset* can be defined as a group of symbols and *sequences* are ordered group of symbols. We excluded to mine more elaborated patterns (e.g., chronicles (Dauxais et al. 2019)) due to their limited interest on the short time series we consider (e.g., 4 days with one timestamp per day).

According to the results published and our experiments, sequence-based classifiers outperform item-based ones on average on time series data. The state-of-the-art sequence-based classifiers are not adopted for explainability reasons. The adoption of support vector machines and Bayesian networks to perform classification in (Buza and Schmidt-Thieme 2010) limits the comprehensibility of how the patterns are used in the model output. Then, Fradkin and Mörchen (2014) classify based on discriminant sequential patterns, i.e. patterns that are characteristic of a class. But, the discriminant sequence mining task extracts patterns that can also occur in other classes than those which are initially discriminated. Thus, the classification task can lead to unclear explanations supporting predictions, specifically communicating to the user the discriminative patterns of other classes than the one the model is predicting. So, in this study, we mine frequent patterns without considering the class information. In addition, as stated in (Fradkin and Mörchen 2014), direct methods (2 stages: pattern mining with class label, classification) can reduce the number of patterns generated but can also lead to significantly worse performance compared to indirect methods (3 stages: unsupervised pattern mining, feature selection, classification). Therefore, a new indirect pattern-based classifier using frequent sequential patterns for estrus detection is proposed in section 3.

**Estrus Detection** There are a couple of studies about the application of data science techniques to improve automatic estrus detection based on affordable sensor data (Dolecheck et al. 2015; Fauvel et al. 2019; Minegishi, Heins, and Pereira 2019; Ma et al. 2020). Dolecheck et al. (2015) based the study on time series data of activity, using visual detection as the ground truth (65% of all estruses). Three machine learning techniques are tested on a limited dataset of 18 estruses (18 cows): random forest, linear discriminant and a multilayer perceptron. Minegishi, Heins, and Pereira (2019) learnt a logistic regression on activity variables using the combination of an automatic estrus detection solution (collar-mounted activity meter) and visual detection as the ground truth (total dataset: 1,462 estruses, 300 cows). Ma et al. (2020) trained a long short-term memory network along with a convolution neural network as estrus detection solution based on activity data (40 cows with 6 estruses labeled visually). However, we cannot compare the detection results of our algorithm to the ones from these three studies using affordable activity measurements because the labeling method used is not exhaustive (visual detection). As far as we have seen, (Fauvel et al. 2019) is the only study adopting an exhaustive labeling (dataset: 125 cows - 671 estruses labeled using progesterone dosage in milk). Therefore, we have limited our baselines to (Fauvel et al. 2019) and the commercial solution performance.

## 3 XPM

In this section, we first present our proposed approach XPM and then detail its explainability.

**XPM Presentation** Estrus detection can be formulated as a classification problem, where the input is sensor data and the output is a class (estrus/non-estrus). More specifically, the problem is an instance of multivariate time series classification. We have a set of co-evolving time series (7 variables), recorded simultaneously by 2 sensors (activity meter, thermobolus), which form a multivariate time series (MTS). As illustrated in Figure 1, our new indirect and eXplainable Pattern-based approach for MTS classification (XPM) is composed of the following steps:

- *Discretization*: we apply SAX (Lin et al. 2003) on each variable. SAX transforms a time series into a string using an alphabet of predefined size. Each symbol of the alphabet corresponds to an interval of a variable values set by the algorithm, therefore SAX symbols can be interpreted (e.g., alphabet of size 3, i.e. 3 intervals: {low, medium, high}). We give in Figure 4 (Appendix A.1) a discretization example on a rumination time series. Alphabet size per variable is a hyperparameter of XPM. We limit the alphabet size to [1,10] to obtain readable patterns and we set 3 sizes of alphabet according to the types of variables (alphabet 1: continuous variable - temperature, alphabet 2: integer variables - other and over activity, alphabet 3: binary variables - the remaining variables);
- *Pattern Mining*: we extract two types of patterns for comparison - frequent itemsets with Eclat algorithm (Zaki 2000) and frequent sequences with BIDE algorithm (Wang and Han 2004). As presented in section 2, frequent itemsets are groups of symbols occuring in at least a predefined percentage (support) of the time series. For example, the itemset $\{=-, +, ++\}$ present in the rumination time series of Figure 4 (Appendix A.1) is frequent if it occurs in at least 20% of the time series of the training set. Frequent sequences correspond to frequent ordered groups of symbols. The type of pattern (itemsets, sequences) and support are hyperparameters of XPM. We restrict the support to [10%,50%] for itemsets and [3%,9%] for sequences to not only mine high frequency patterns of length 1;
- *Encoding*: we encode in a matrix which patterns (as columns) are present in which MTS (as rows) to form the input data of the classifier (see Figure 5 in Appendix A.2);
- *Feature Selection*: before classification, we perform a feature selection to keep a limited and explainable set of patterns. We use a filter method to select the k-best patterns according to a score (Chi-Square). We choose the Chi-Square because it is suited for feature selection on booleans data relative to classes. The number of patterns is a hyperparameter of XPM and we limit its range to [10,40];
- *Classification*: finally, we classify the MTS using a decision tree to keep the explainability on the classifier predictions. The explainability provided by the decision tree classifier is detailed in the next section.

**Explainability** The explainability of our approach stems from the communication to the farmers the presence and/or absence of a limited number of patterns determinant of estrus detection. Patterns are communicated to the farmers fol-

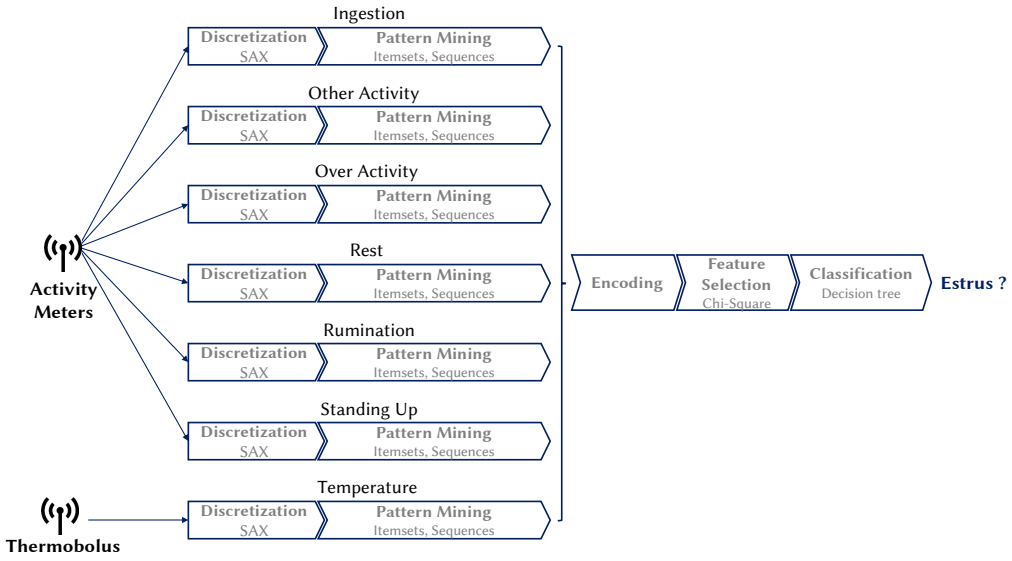


Figure 1: Pipeline of the pattern-based classifier XPM.

lowing a decision tree to classify estrus. We present in this section how to read a pattern and a decision tree. Figure 2 shows an example of a one node with two leaves decision tree trained on a dataset of 600 MTS (300 estruses/300 non-estruses). The node is composed of the pattern $++++$ on the variable over activity. In an alphabet of size 9 $\{----, ---, --, -, =, +, ++, +++, ++++\}$, $++++$ refers to the interval of highest values relative to the variable over activity. Therefore, we observe that when a high over activity (pattern $++++$ on over activity) is observed in a MTS (pattern present), which is the case for 199 MTS, most of the MTS correspond to estruses (184 over 199, error rate: 8%). In this case, the decision tree predicts the class estrus: the most represented class in the leaf. Blue filled nodes mostly contain estruses and grey filled nodes mostly non-estruses. When the pattern $++++$ on over activity is not observed in a MTS, the decision tree predicts non-estrus but with a higher error rate ($116/401 = 29\%$). Additional patterns could refine the decision and reduce the error of the tree. The explainability results are presented in section 5.

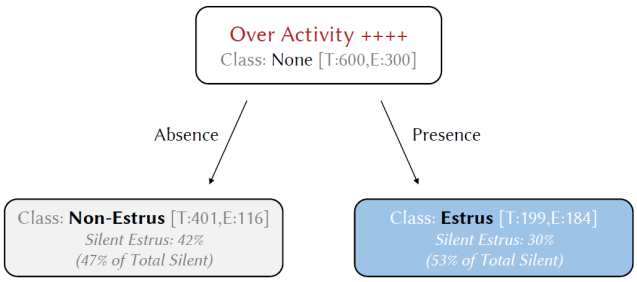


Figure 2: Explainability - decision tree example with one node and two leaves. Abbreviations: T - total number of samples, E - number of estruses.

## 4 Evaluation

In this section, we present our evaluation method.

**Dataset** Our dataset is a real-world dataset collected during an experiment conducted at the dairy research farm of Méjusseaume (48°06' N, 1°47' W, Brittany, France) from 2014 to 2018. The composition of the 5-fold cross-validation and external validation datasets are presented in Table 1. Additional information about the experimental setting and labeling is available in Appendix B.1.

Table 1: Dataset Split. Abbreviation: Ext Val - External Validation.

| | Folds | | | | | | Ext Val |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | All | |
| Estrus | 99 | 100 | 100 | 100 | 100 | 499 | 321 |
| Silent % | 32 | 38 | 34 | 30 | 36 | 34 | 44 |
| Lactation 1 | 64 | 55 | 61 | 58 | 61 | 299 | 193 |
| Silent % | 36 | 36 | 41 | 26 | 43 | 36 | 50 |
| Lactation 2+ | 35 | 45 | 39 | 42 | 39 | 200 | 128 |
| Silent % | 26 | 40 | 23 | 36 | 26 | 31 | 34 |

**Benchmark** We evaluate the performance of XPM in comparison with a reference Commercial Solution - CS (Medria Heatphone), the current state-of-the-art estrus detection algorithm LCE (Fauvel et al. 2019) and a variant of XPM (XPM-Derivatives). XPM-Derivatives corresponds to XPM on a dataset augmented by the derivatives of each variable. Górecki and Łuczak (2013) show that using derivatives can be helpful in time series classification. Derivatives correspond to the value difference of a variable compared to the previous day (see Appendix B.2 for an illustration of a dataset augmented by the derivatives of each variable).

**Hyperparameters Setting** Most of the hyperparameters of XPM presented in section 3 and summarized in Appendix B.3 (alphabet sizes, number of patterns, time series length, patterns, support) are determined by grid search on the validation sets of the cross-validation (5-fold cross-validation 60/20/20 train/validation/test split). On the same validation sets, decision tree hyperparameters are determined by the hyperopt algorithm (Bergstra, Yamins, and Cox 2013).

**Performance Calculation** We do not make assumptions on dairy herd management, meaning that we do not have a preference between reducing false positives (false estrus alerts) and false negatives (estruses not detected) so we have optimized the F1-score, the harmonic mean between precision and recall. More details about the performance calculation are presented in Appendix B.4.

## 5 Results and Discussions

**Detection Performance** Firstly, we have evaluated our XPM approach on the cross-validation dataset. We observe a better F1-score with less variability across folds of our XPM approach based on sequences than itemsets on both behavioral and silent estruses (see Table 2 in Appendix C.1 - total: 74.5% $\pm$ 2.2 versus 73.6% $\pm$ 3.9, behavioral: 78.2% $\pm$ 3.1 versus 77.4% $\pm$ 4.2, silent: 57.7% $\pm$ 1.3 versus 56.4% $\pm$ 2.9). Thus, according to our experiments, the most informative patterns (sequences) have to be selected for estrus detection. Moreover, patterns extracted from a dataset with derivatives allow additional information, for example information about the activity level of a cow, as well as variation of activity compared to the day before. We observe that mining sequences on both raw variables and derivatives improves F1-score compared to sequences on raw variables (total: XPM-Derivatives 75.4% versus XPM 74.5%). Therefore, based on the cross-validation, the best performing approach is XPM-Derivatives with Sequences (alphabet 1 size: 7, alphabet 2 size: 9, alphabet 3 size: 8, time series length: 4 days, patterns: sequences, feature selection: 20, support: 3%).

Then, we observe that on the external validation dataset our approach has a better F1-score than the commercial solution (62.4% versus 60.9%), based on a better estrus coverage (recall: XPM-Derivatives with Sequences 53.0% versus CS 49.1%, precision: XPM-Derivatives with Sequences 75.9% versus CS 80.0%). The higher performance of our approach is driven by the detection of silent estrus (F1-score: 39.2% versus 0.0%). We infer that the lower detection performance of our pattern-based classifier on behavioral estrus than the commercial solution (F1-score: 56.3% versus 81.8%) is due to the non-discriminative patterns mined. We mined patterns based on a frequency criteria without considering the type of estrus. Therefore, selected patterns are mostly as frequent in behavioral as in silent estrus, which prevents to fully characterize behavioral estrus. We can also compare the performance of our approach to the existing study (Fauvel et al. 2019). The ensemble method developed (LCE) shows better F1-score than XPM-Derivatives with Sequences (71.6% $\pm$ 0.4). Nonetheless, the ensemble approach cannot support its predictions with perfectly faithful explanations as it relies on a post-hoc model-agnostic explainability method (SHAP), which can prevent LCE adoption as faithfulness is critical to reduce solution mistrust from farmers.

**Explainability** Figure 8 (see Appendix C.2) shows the decision tree corresponding to the best configuration determined by cross-validation. Firstly, we observe that the presence of a steep peak in over activity (root node: pattern $\langle =, ---- \rangle$ on over activity derivative) leads to the identification of 49% of estruses of the training set (244 over 499 estruses) with a low error rate (3% - 7 non-estruses over 251 samples). The simultaneous presence of a pattern confirming the first one with a steep decrease followed by a rise in rest (pattern $\langle ---, =+ \rangle$ on rest derivative) leads to a refinement of the detection and identifies 38% of all estruses (191 over 499 estruses) with an error rate of 0.5% (1 non-estrus over 192 samples). In the case of the absence of this confirming pattern (pattern $\langle ---, =+ \rangle$ on rest derivative), the presence of a steep rise of temperature leads to the identification of estrus with a low error rate of 2.5% (1 non-estrus over 40 samples). Then, in the absence of the two patterns relative to over activity (pattern $\langle =, ---- \rangle$ on over activity derivative and pattern $\langle -, ++++ \rangle$ on over activity), a low rumination (pattern $\langle --- \rangle$ on rumination) with a low rest (pattern $\langle --- \rangle$ on rest) confer the estrus (error rate of 21% - 9 non-estruses over 43 samples). Additional information about the decision tree is available in Appendix C.2.

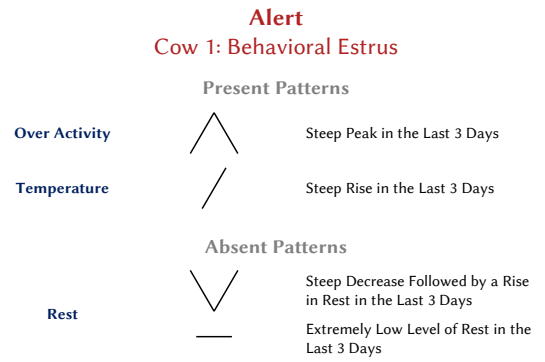


Figure 3: Example of a behavioral estrus alert with the corresponding patterns detected.

The explainability of our approach stems from the communication to the farmers the presence and/or absence of a limited number of patterns determinant of estrus detection. Patterns are communicated to the farmers following the decision tree shown in Figure 8 (Appendix C.2). On a daily basis, our solution informs the farmers about the cows in estrus, the type of estrus detected with its associated probability and the patterns which have generated the alerts. In case of behavioral estrus, these patterns allow the farmers to confirm the patterns visually sensed. In addition, in case of silent estrus, our solution allows the farmers to save time looking for non visually verifiable behavioral signs and tells them that a potential insemination would be performed on a silent estrus. For example, Figure 3 illustrates the level of information that a farmer could receive with an estrus alert. The interface contains the animal identifier, the type of estrus and the patterns detected. Our solution predicts that cow 1 is in behavioral estrus on the day of the alert based on the detection in the last 3 days of a steep peak of over activity, a steep rise in temperature and the absence of 2 patterns on rest (a steep decrease followed by a rise in rest, an extremely low level of rest). As indicated in the decision tree of our approach, the combination of these 2 patterns and the absence of the patterns on rest always lead to an estrus (0% error rate - 0 non-estrus over 21 samples), with a vast majority of behavioral estruses ($<$10% of silent estruses).

## 6 Conclusion

Our study confirms the potential of our pattern-based classifier XPM to improve the estrus detection performance of commercial solutions based on the combination of affordable sensor data (activity, body temperature), while providing perfectly faithful explanations in ways that farmers can understand based on the sequential patterns they could observe in animals. With regard to future work, we plan to study a three class classification setting (non-estrus/behavioral estrus/silent estrus) with discriminative patterns to gain further insights on silent estrus detection, as the patterns mined in silent estrus are roughly as frequent as those in behavioral estrus in the two class setting of this study (non-estrus/estrus). We also plan to work on an approach with a broader data heterogeneity combining the detection performance of our previous study (Fauvel et al. 2019) and the explainability of this pattern-based approach.

## Acknowledgments

We thank all technical staff of INRAE Méjusseaume dairy farm who helped managing and monitoring this long-term experimentation. We also thank Medria for its collaboration by providing activity and temperature sensor data. This work was supported by the French National Research Agency under the Investments for the Future Program (ANR-16-CONV-0004), the Inria Project Lab Hybrid Approaches for Interpretable AI (HyAIAI), the French national project Deffilait (ANR-15-CE20-0014) and APIS-GENE.

## References

Adriaens, I.; Huybrechts, T.; Geerinckx, K.; Daems, D.; Lammertyn, J.; Ketelaere, B. D.; Saeys, W.; and Aernouts, B. 2017. Mathematical Characterization of the Milk Progesterone Profile as a Leg Up to Individualized Monitoring of Reproduction Status in Dairy Cows. *Theriogenology*, 103: 44–51.

Adriaens, I.; Saeys, W.; Huybrechts, T.; Lamberigts, C.; François, L.; Geerinckx, K.; Leroy, J.; Ketelaere, B. D.; and Aernouts, B. 2018. A Novel System for On-Farm Fertility Monitoring Based on Milk Progesterone. *Journal of Dairy Science*, 101(9): 8369–8382.

Bascom, S.; and Young, A. 1998. A Summary of the Reasons Why Farmers Cull Cows. *Journal of Dairy Science*, 81(8): 2299–2305.

Bergstra, J.; Yamins, D.; and Cox, D. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*.

Buza, K.; and Schmidt-Thieme, L. 2010. Motif-Based Classification of Time Series with Bayesian Networks and SVMs. *Advances in Data Analysis, Data Handling and Business Intelligence*, 105–114.

Chanvallon, A.; Coyral-Castel, S.; Gatien, J.; Lamy, J.; Ribaud, D.; Allain, C.; Clément, P.; and Salvetti, P. 2014. Comparison of Three Devices for the Automated Detection of Estrus in Dairy Cows. *Theriogenology*, 82(5): 734–741.

Cheng, H.; Yan, X.; Han, J.; and Hsu, C. 2007. Discriminative Frequent Pattern Analysis for Effective Classification. In *Proceedings of the 23rd International Conference on Data Engineering*.

Cheng, H.; Yan, X.; Han, J.; and Yu, P. 2008. Direct Discriminative Pattern Mining for Effective Classification. In *Proceedings of the 24th International Conference on Data Engineering*.

Cutullic, E.; Delaby, L.; Gallard, Y.; and Disenhaus, C. 2011. Dairy Cows' Reproductive Response to Feeding Level Differs According to the Reproductive Stage and the Breed. *Animal*, 5(5): 731–740.

Dauxais, Y.; Gross-Amblard, D.; Guyet, T.; and Happe, A. 2019. Discriminant Chronicle Mining. In *Advanced Knowledge and Data Mining*, 89–118.

De Silva, A.; Anderson, G.; Gwazdauskas, F.; McGilliard, M.; and Lineweaver, J. 2015. Interrelationships With Estrous Behavior and Conception in Dairy Cattle. *Journal of Dairy Science*, 64(12): 2409–2418.

Dolecheck, K.; Silvia, W.; Heersche, G.; Chang, Y.; Ray, D.; Stone, A.; Wadsworth, B.; and Bewley, J. 2015. Behavioral and Physiological Changes Around Estrus Events Identified Using Multiple Automated Monitoring Technologies. *Journal of Dairy Science*, 98(12): 8723–8731.

Fauvel, K.; Masson, V.; Fromont, É.; Faverdin, P.; and Termier, A. 2019. Towards Sustainable Dairy Management - A Machine Learning Enhanced Method for Estrus Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fournier-Viger, P.; Lin, J.; Vo, B.; Chi, T.; Zhang, J.; and Le, H. 2017. A survey of Itemset Mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4).

Fradkin, D.; and Mörchen, F. 2014. Mining Sequential Patterns for Classification. *Knowledge and Information Systems*, 45: 731–749.

Friggens, N.; Bjerring, M.; Ridder, C.; Højsgaard, S.; and Larsen, T. 2008. Improved Detection of Reproductive Status in Dairy Cows Using Milk Progesterone Measurements. *Reproduction in Domestic Animals*, 43(2): 113–121.

Gaillard, C.; Barbu, H.; Sørensen, M.; Sehested, J.; Callesen, H.; and Vestergaard, M. 2016. Milk Yield and Estrous Behavior During Eight Consecutive Estruses in Holstein Cows Fed Standardized or High Energy Diets and Grouped According to Live Weight Changes in Early Lactation. *Journal of Dairy Science*, 99(4): 3134–3143.

Górecki, T.; and Łuczak, M. 2013. Using Derivatives in Time Series Classification. *Data Mining and Knowledge Discovery*, 26(2): 310–331.

Gwazdauskas, F.; Lineweaver, J.; and McGilliard, M. 1983. Environmental and Management Factors Affecting Estrous Activity in Dairy Cattle. *Journal of Dairy Science*, 66: 1510–1514.

Inchaisri, C.; Jorritsma, R.; Vos, P.; Weijden, G. V. D.; and Hogeveen, H. 2010. Economic Consequences of Reproductive Performance in Dairy Cattle. *Theriogenology*, 74: 835–846.

Karlsson, I.; Papapetrou, P.; and Boström, H. 2016. Generalized Random Shapelet Forests. *Data Mining and Knowledge Discovery*, 30: 1053–1085.

Kerbrat, S.; and Disenhaus, C. 2004. A Proposition for an Updated Behavioural Characterisation of the Oestrus Period in Dairy Cows. *Applied Animal Behaviour Science*, 87(3-4): 223–238.

Lin, J.; Keogh, E.; Lonardi, S.; and Chiu, B. 2003. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

Lopez, H.; Satter, L.; and Wiltbank, M. 2004. Relationship Between Level of Milk Production and Estrous Behavior of Lactating Dairy Cows. *Animal Reproduction Science*, 81: 209–223.

Lundberg, S.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Ma, N.; Pan, L.; Chen, S.; and Liu, B. 2020. NB-IoT Estrus Detection System of Dairy Cows Based on LSTM Networks. In *IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*.

Martin, O.; Friggens, N.; Dupont, J.; Salvetti, P.; Freret, S.; Rame, C.; Elis, S.; Gatien, J.; Disenhaus, C.; and Blanc, F. 2013. Data-Derived Reference Profiles With Corepresentation of Progesterone, Estradiol, LH, and FSH Dynamics During the Bovine Estrous Cycle. *Theriogenology*, 79: 331–343.

Mayo, L.; Silvia, W.; Ray, D.; Jones, B.; Stone, A.; Tsai, I.; Clark, J.; Bewley, J.; and Heersche, G. 2019. Automated Estrous Detection Using Multiple Commercial Precision Dairy Monitoring Technologies in Synchronized Dairy Cows. *Journal of Dairy Science*, 3: 2645–2656.

Minegishi, K.; Heins, B.; and Pereira, G. 2019. Peri-Estrus Activity and Rumination Time and its Application to Estrus Prediction: Evidence from Dairy Herds Under Organic Grazing and Low-Input Conventional Production. *Livestock Science*, 221: 144–154.

Palmer, M.; Olmos, G.; Boyle, L.; and Mee, J. 2010. Estrus Detection and Estrus Characteristics in Housed and Pastured Holstein-Friesian Cows. *Theriogenology*, 74(2): 255–264.

Peralta, O.; Pearson, R.; and Nebel, R. 2005. Comparison of Three Estrus Detection Systems During Summer in a Large Commercial Dairy Herd. *Animal Reproduction Science*, 87(1-2): 59–72.

Petersson, K.; Gustafsson, H.; Strandberg, E.; and Berglund, B. 2006. Atypical Progesterone Profiles and Fertility in Swedish Dairy Cows. *Journal of Dairy Science*, 89(7): 2529–2538.

Ranasinghe, R.; Nakao, T.; Yamada, K.; and Koike, K. 2010. Silent Ovulation, Based on Walking Activity and Milk Progesterone Concentrations, in Holstein Cows Housed in a Free-Stall Barn. *Theriogenology*, 73: 942–949.

Reith, S.; and Hoy, S. 2012. Relationship Between Daily Rumination Time and Estrus of Dairy Cows. *Journal of Dairy Science*, 95(11): 6416–6420.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1: 206–215.

Searchinger, T.; Waite, R.; Hanson, C.; Ranganathan, J.; Dumas, P.; and Matthews, E. 2018. *Creating a Sustainable Food Future*. World Resources Institute.

Silper, B.; Madureira, A.; Kaur, M.; Burnett, T.; and Cerri, R. 2015. Comparison of Estrus Characteristics in Holstein Heifers by 2 Activity Monitoring Systems. *Journal of Dairy Science*, 98(5): 3158–3165.

Steeneveld, W.; and Hogeveen, H. 2015. Characterization of Dutch Dairy Farms Using Sensor Systems for Cow Management. *Journal of Dairy Science*, 98(1): 709–717.

Tenghe, A.; Bouwman, A.; Berglund, B.; Strandberg, E.; Blom, J.; and Veerkamp, R. 2015. Estimating Genetic Parameters for Fertility in Dairy Cows from In-Line Milk Progesterone Profiles. *Journal of Dairy Science*, 98(8): 5763–5773.

Van Eerdenburg, F.; Karthaus, D.; Taverne, M.; Merics, I.; and Szenci, O. 2002. The Relationship Between Estrous Behavioral Score and Time of Ovulation in Dairy Cattle. *Journal of Dairy Science*, 85(5): 1150–1156.

Van Vliet, J.; and Eerdenburg, F. V. 1996. Sexual Activities and Oestrus Detection in Lactating Holstein Cows. *Applied Animal Behaviour Science*, 50(1): 57–69.

Wang, J.; and Han, J. 2004. BIDE: Efficient Mining of Frequent Closed Sequences. In *Proceedings of the 20th International Conference on Data Engineering*, 79–90.

Yániz, J.; Santolaria, P.; Giribet, A.; and López-Gatius, F. 2006. Factors Affecting Walking Activity at Estrus During Postpartum Period and Subsequent Fertility in Dairy Cows. *Theriogenology*, 66(8): 1943–1950.

Zaki, M. 2000. Scalable Algorithms for Association Mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3): 372–390.

Zebari, H.; Rutter, S.; and Bleach, E. 2018. Characterizing Changes in Activity and Feeding Behaviour of Lactating Dairy Cows During Behavioural and Silent Oestrus. *Applied Animal Behaviour Science*, 206: 12–17.

# Appendices

## XPM Presentation

### A.1 Discretization

We present in this section a discretization example on a rumination time series of length 7 with an alphabet of size 8 $\{---, --, -, =-, =+, +, ++, +++\}$. Based on the intervals defined by SAX, the discretization output of the time series is: $+++, +++, --, +, =-, ++, ++$.

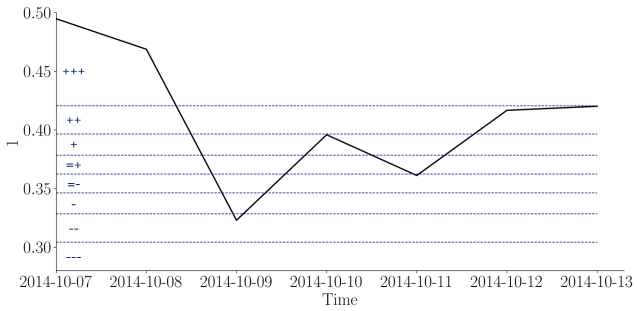


Figure 4: SAX discretization example on a rumination time series of length 7 with an alphabet of size 8.

### A.2 Encoding

We show in this section how we encode in a matrix which patterns (as columns) are present in which MTS (as rows) to form the input data of the classifier.

**Encoding**: $m^*(p+2)$

| MTS ID | Pattern 1 | Pattern 2 | Pattern 3 | ... | Pattern p-2 | Pattern p-1 | Pattern p | Label |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | | 1 | 0 | 0 | Non-Estrus |
| 2 | 1 | 0 | 0 | | 0 | 0 | 1 | Estrus |
| 3 | 0 | 0 | 0 | | 0 | 1 | 0 | Non-Estrus |
| 4 | 1 | 0 | 1 | | 0 | 0 | 1 | Estrus |
| . | | | | | | | | . |
| . | | | | | | | | . |
| m-2 | 0 | 1 | 0 | | 0 | 1 | 0 | Non-Estrus |
| m-1 | 1 | 0 | 0 | | 0 | 0 | 1 | Estrus |
| m | 0 | 1 | 0 | | 1 | 1 | 0 | Non-Estrus |

Figure 5: Encoding matrix example. Abbreviations: ID - identifier, m - number of Multivariate Time Series (MTS) samples, p - number of patterns mined.

## Evaluation

### B.1 Dataset

**Experimental Setting** An experiment was conducted at the dairy research farm of Méjusseaume (48°06' N, 1°47' W, Brittany, France) from 2014 to 2018. This experiment enrolled 162 Holstein cows (214 lactations) housed in free stalls. The first 3 years dataset (125 different cows) is used for cross-validation and the last year is used for external validation (61 cows). In the external validation dataset, 24 cows are also in the cross-validation dataset but within a new lactation and 37 are different. Cows calved between August and September. At the beginning of this experiment, there were 50% of primiparous among all cows. Primiparous correspond to cows in their first lactation. The parity, i.e. the average lactation number, were 1.8 with a standard error of 0.1. Milking occurred twice daily at 7:00-9:00 and 16:00-18:00. Delivery of the total mixed ration containing on average 65% of maize silage, 10% of dehydrated alfalfa pellets and 25% of concentrate occurred twice daily at 09:00 and 17:00. This experiment was carried out in accordance with the guidelines for animal research of the French Ministry of Agriculture (décret NOR AGRG 1231951D) and approved by the "Comité National de Réflexion Ethique sur l'Expérimentation Animale" (Authorization of the French Ministry of Higher Education, Research and Innovation reference APAFIS 3122-2015112718172611).

**Data Collection** Data were collected by affordable activity meters and body temperature sensors. Each cow was equipped with a collar-mounted activity meter (HeatPhone and FeedPhone - Medria Technologies, Châteaubourg, France) and a temperature sensor in the reticulorumen (Thermobolus - Medria Technologies, Châteaubourg, France). The dataset consists of seven Medria numeric variables with a 5-minute frequency (*rumination*, *ingestion*, *rest*, *standing up*, *over activity*, *other activity* and *temperature*). *Temperature* takes into account the cooling effect of water ingestion by the cows. The variables have different types: *temperature* is a continuous variable; *over activity* and *other activity* are integer variables that relate to the intensity of the activity (values [0,10]); the remaining ones are binary variables. The values of activity variables at each timestamp correspond to the dominant activity during each 5-minute period. Time series are 24hr aggregated (average), which is sufficient for an estrus alert system and a timely insemination. In addition, the 24hr aggregation allows the mining of patterns that are not affected by the intraday sequence of animals activities (e.g., moment of the day a cow is drinking), which is irrelevant to estrus. No other preprocessing has been done to the data collected. We assume that the processing operated by Medria on raw data to generate variables is stable during our experiment.

Based on its good performance compared to other solutions (Chanvallon et al. 2014) and its international market presence, we consider that Medria estrus detection system is a reasonable basis of comparison and constitutes our baseline. The estrus alerts of Medria Heatphone are called the commercial solution (CS).

**Gold Standard** Our study covers both estrus types (behavioral and silent). Therefore, we labeled estrus by measuring the progesterone concentration in whole milk, the costly gold standard for an exhaustive estrus identification (Martin et al. 2013). This non-invasive method for the cow induces commonly accepted errors (Adriaens et al. 2018). Milk progesterone measurements are subject to a variability, partly caused by the measurement technique and calibration method (Adriaens et al. 2017), the sampling technique, or the fat content in the milk sample (Friggens et al. 2008).

Milk samples were collected from each cow twice a week on Tuesdays and Thursdays and were immediately frozen at -20°C until the dosage. We used the enzyme-linked immunosorbent assay technique (kit ELISA Ridgeway Science Ltd). Then, with preserved and frozen milk, the separation of basic concentrations of progesterone to estrus period has been determined based on the quantile method (Petersson et al. 2006; Cutullic et al. 2011).

Figure 6 shows an example of the output from the quantile method applied on the progesterone profile of a cow on a one month time period, and how this output is used to label the


Figure 6: Example of the output from the quantile method (Petersson et al. 2006), applied on the progesterone profile of a cow on a one month time period, with the corresponding non overlapping and sliding window labeling in the case of time series length of 4 days. Quantile method output: 1 - day of an estrus period, 0 - day of a non-estrus period.

corresponding time series, in this case of 4-day length (our final configuration). We adopt two types of labeling: first, a non-overlapping labeling on the cross-validation dataset to find patterns on clearly identified periods, which improves the relevancy of patterns used subsequently for classification; second, a sliding window labeling on the external validation dataset to evaluate the performance of our approach on real-world conditions, i.e. as a daily monitoring solution. The non-overlapping labeling consists of an equal number of estrus and non-estrus time series to avoid class imbalance, with the estrus MTS ending in the middle of the estrus period and the non-estrus MTS preceding the estrus one. On both labeling, the label of the last day is used to label the whole time series.

**Definition of Behavioral and Silent Estrus** In our analysis of the results, we distinguish between the patterns characteristic of behavioral and silent estrus. Silent estrus is defined by the absence of obvious behavioral sign and represent around 35% of all estrus (Kerbrat and Disenhaus 2004; Palmer et al. 2010; Ranasinghe et al. 2010). In our dataset, an estrus is marked as behavioral estrus when either a visual detection or a CS alert occurred. Commercial solution alerts are based on an increase in over activity. Staff observation generating visual estrus alerts occurred 4 times a day for about 30-minute period (before milking and during rest period at 7:00, 11:00, 15:30, 22:00). And, the estruses not marked as behavioral estruses are silent estruses. As presented in the next section, our dataset composition is aligned with the literature with regard to silent estrus proportion (cross-validation: 34%, external validation: 44%).

**Composition** The composition of the cross-validation and external validation datasets are presented in Table 1. The cross-validation dataset was split into five folds. The split has kept the same number of estruses in each fold (100). This split does not lead to an overfit on a particular animal. It has been studied in (Fauvel et al. 2019) that a dataset split on animals has a negative impact on detection performance (-2.6% impact on F1-score). We observe in our dataset that more estruses occur in the first lactation (60%, external validation: 60% of all estruses), and that cows in the first lactation (primiparous) experience a higher proportion of silent estrus compared to cows in higher lactations - multiparous (36% vs 31%, external validation: 50% vs 34%). We do not consider that it could bias our study. In the literature, the effect of parity is unclear on the estrus detection performance. Some authors reported greater estrus intensity for older cows (De Silva et al. 2015; Gwazdauskas, Lineweaver, and McGilliard 1983) while others reported a greater activity for primiparous (Van Vliet and Eerdenburg 1996; Peralta, Pearson, and Nebel 2005; Yániz et al. 2006) or no difference (Van Eerdenburg et al. 2002; Lopez, Satter, and Wiltbank 2004).

## B.2 Derivatives

Figure 7 illustrates a dataset augmented by the derivatives of each variable.

| Animal ID | Timestamp | Attribute 1 | Attribute 1 Der | . . | Attribute 7 | Attribute 7 Der |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | - | | 39.24 | - |
| 1 | 2 | 1 | 0 | | 38.9 | -0.34 |
| 1 | 3 | 0 | -1 | | 38.78 | -0.12 |
| 1 | 4 | 0 | 0 | | 39.1 | 0.32 |
| 1 | 5 | 1 | 1 | | 39.19 | 0.09 |
| . | . | . | . | . . | . | . |
| 1 | T-1 | 1 | 0 | | 39.23 | 0.08 |
| 1 | T | 0 | -1 | | 39.34 | 0.11 |

Figure 7: MTS sample with derivatives for one animal of our dataset. For each timestamp, the 7 variables of our dataset with their derivatives are represented. Abbreviations: Der - derivatives, ID - identifier, T - time series length.

## B.3 Hyperparameters

The hyperparameters of XPM are:

- Alphabet sizes: 3 alphabets are defined, according to the types of variables (alphabet 1: continuous variable - temperature, alphabet 2: integer variables - other and over activity, alphabet 3: binary variables - the remaining variables), with sizes in [1,10];
- Time series length: it ranges from 4 to 21 days - the length of a regular ovarian cycle ([4,21]);
- Patterns: type of patterns (itemsets, sequences);
- Number of patterns: number of patterns kept during the feature selection ([10,40]);

- Support: minimum frequency of patterns (itemsets: [10%,50%], sequences: [3%,9%]);
- Decision tree hyperparameters: depth of the tree [1,ln(number of patterns)], minimum number of samples at a leaf [2,number of patterns].

### B.4 Performance Calculation

We do not give a preference to reduce false positives (false estrus alerts) or false negatives (estruses not detected) so we have optimized the F1-score, the harmonic mean between precision and recall. The *precision* measures the proportion of actual positives among those predicted positives, and the *recall*, also called sensitivity, measures the proportion of actual positives that are correctly identified as such. Adopting a conservative approach, we decided to aggregate our model daily predictions by the maximum of the daily predictions on estrus/non-estrus period to calculate the classification performance. Based on a 5-fold cross-validation 60/20/20 train/validation/test split, the algorithm is selected on the best F1-score on validation sets. We present two levels of performance. First, we show the F1-score with precision set to the same as CS (78%, threshold: 0.4) to allow comparison between the approaches. The threshold corresponds to the value from which the pattern-based classifier class probabilities predict estrus. The second level is the F1-score across all possible calibrations (threshold range: 0.3-0.75) which corresponds to the average performance of our solution. We observe that for high thresholds (threshold $>$ 0.75), our pattern-based classifier performance is unstable with a significant decrease in estrus detection rate (recall below 70%). In addition, for low thresholds (threshold $< 0.3$), our classifier is equivalent to a random classifier. So, we decided to adopt a F1-score calculation based on the average of F1-score on threshold range 0.3-0.75, which corresponds to the plausible range of calibration for dairy management and shows a detection performance closer to real conditions.

## Results

### C.1 Detection Performance

We present in Table 2 the F1-score on test sets of our approach on the cross-validation and external validation datasets.

### C.2 Explainability

Figure 8 presents the decision tree corresponding to the best configuration determined by cross-validation and presented in section 5.

Concerning the different types of estrus (silent/behavioral), some patterns among the 20 patterns are characteristic of behavioral estrus. Three patterns are three times more frequent in behavioral estrus than in silent estrus MTS: pattern $\langle -, ++++ \rangle$ on over activity, pattern $\langle --, = \rangle$ on over activity derivative and pattern $\langle -, = \rangle$ on over activity derivative. One of them (pattern $\langle -, ++++ \rangle$ on over activity - a prolonged high over activity) is used in the decision tree as a splitting variable and the subsequent leaf contains, as expected, one of the lowest silent estrus proportion of the leaves (12.5%). Nonetheless, we observe that most of the patterns used in the decision tree are present in the same proportion in behavioral as in silent estrus MTS. There is no pattern among the 20 patterns that is characteristic of silent estrus (pattern at least twice more frequent in silent than in behavioral estrus). Therefore, our pattern-based approach mostly relies on the identification of patterns that are as much associated to behavioral estrus as to silent estrus (pattern $\langle =, ---- \rangle$ on over activity derivative, pattern $\langle + \rangle$ on temperature derivative, pattern $\langle --- \rangle$ on rest, pattern $\langle ---, =+ \rangle$ on rest derivative). It allows us to correctly classify a bit more than half of the estrus type not detected by the commercial solution (51% of silent estrus). In order to obtain some patterns characteristic of behavioral and silent estrus and enhance their detection performance, it would be interesting to work on a three class classification setting (non-estrus/behavioral estrus/silent estrus) with discriminative patterns.

The patterns observed corroborate some observations from previous studies. First, over activity is significantly higher in estrus than in non-estrus. According to the device used, activity measured in steps and neck movements increases on the day of estrus from 69 to 170% in (Mayo et al. 2019). Jonsson et al. (2011) show that during estrus, the period of time cows spent lying decreases as a result of increased activity and restlessness. Concerning the low rumination pattern, a study from (Reith and Hoy 2012) reveals that rumination reduces on the day of estrus from 7.2 to 5.9 h/d. Mayo et al. (2019) confirm this reduction by publishing a $-2$ to $-16$% change in rumination time on the day of estrus for both neck and ear-based technologies. Additionally, the period of 4 days of our best configuration is aligned with the study of (Zebari, Rutter, and Bleach 2018) which demonstrates that on the day of behavioral estrus, the number of the steps are higher compared to 3 days before and 3 days after estrus.

Table 2: Comparison[1] of estrus detection F1-score[2] with 95% confidence interval and statistical significance[3] on test sets and external validation. Abbreviation: CS - Commercial Solution.

| | Cross-Validation | | | External Validation | |
|---|---|---|---|---|---|
| | **XPM with Itemsets** | **XPM with Sequences** | **XPM-Derivatives with Sequences** | **XPM-Derivatives with Sequences** | **CS** |
| Total | 73.8 ± 3.6 | 75.2 ± 1.5 | 78.1 ± 0.9 | 62.4 | 60.9 |
| | (73.6 ± 3.9) | (74.5 ± 2.2) | (75.4 ± 1.6) | - | - |
| Behavioral | 76.1 ± 3.9 | 78.8 ± 2.5 | 79.1 ± 1.3 | 56.3 | 81.8* |
| | (77.4 ± 4.2) | (78.2 ± 3.1) | (77.6 ± 1.6) | - | - |
| Silent | 56.2 ± 2.5 | 58.4 ± 1.2 | 55.2 ± 1.5 | 39.2 | 0.0*** |
| | (56.4 ± 2.9) | (57.7 ± 1.3) | (57.1 ± 1.8) | - | - |
| Lactation 1 | 73.3 ± 3.5 | 74.3 ± 1.7 | 76.4 ± 1.3 | 58.9 | 57.3 |
| | (72.9 ± 3.8) | (73.6 ± 2.3) | (74.4 ± 2.3) | - | - |
| Lactation 2+ | 74.3 ± 4.3 | 76.1 ± 2.2 | 80.4 ± 2.2 | 66.9 | 65.3 |
| | (74.1 ± 4.5) | (75.1 ± 2.7) | (76.4 ± 1.8) | - | - |

[1] Methods compared: XPM based on itemsets/sequences/sequences with derivatives and the commercial solution.
[2] Two levels of performance are presented. The first line corresponds to the performance based on the same total precision as CS (78%, XPM threshold set to 0.4). The second line with parenthesis shows the average performance across all possible calibrations (threshold range: 0.3-0.75).
[3] The P-value represents the 5*2-fold cross-validation paired t-test result of CS compared to XPM-derivatives with Sequences (*P<0.05, ***P<0.01).

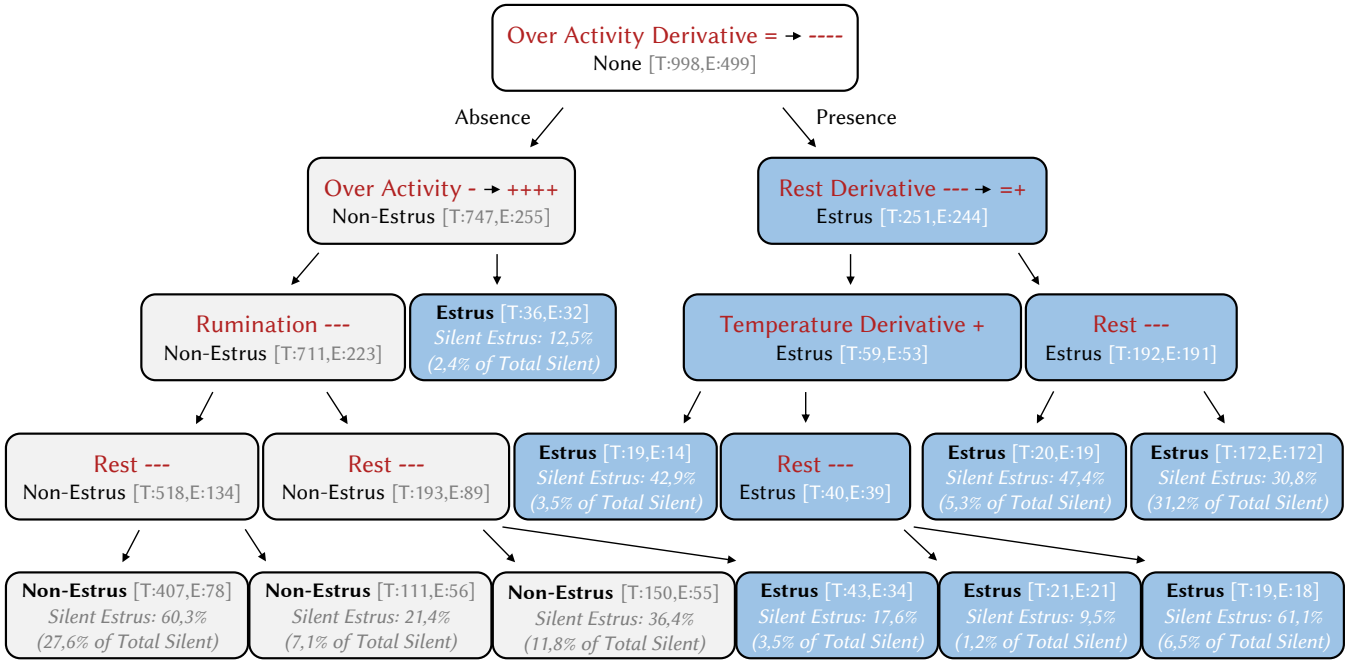


Figure 8: Decision tree determined by cross-validation. Abbreviations: T - total number of samples, E - number of estruses.