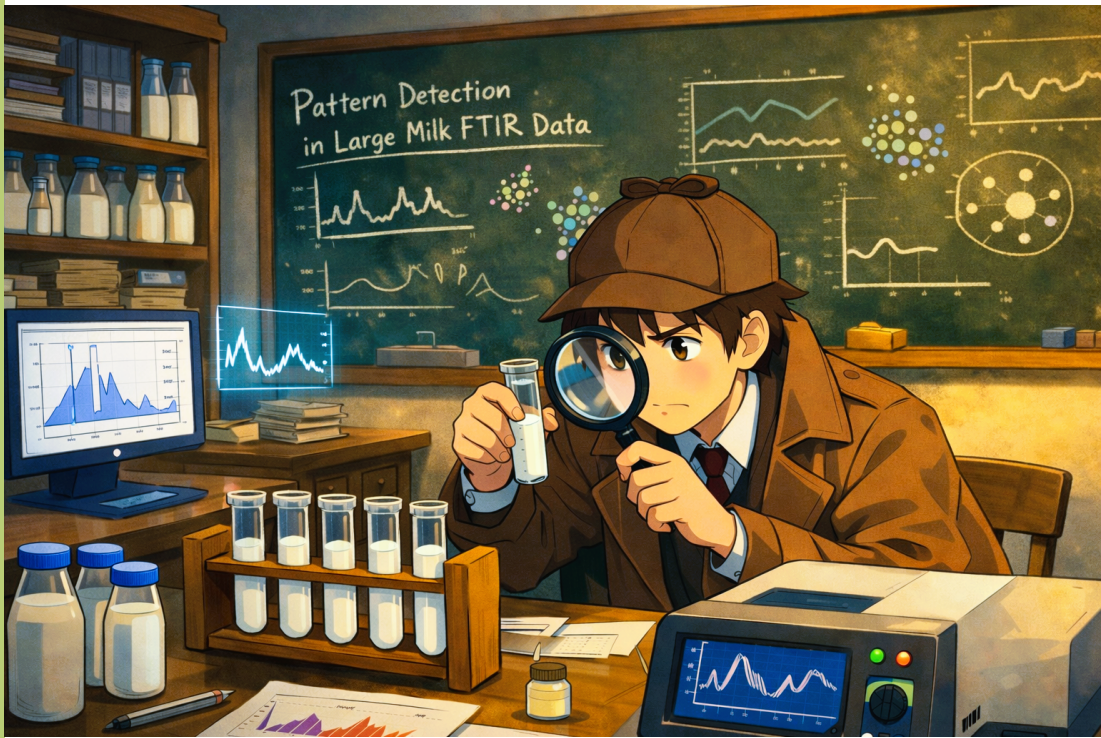


Health pattern detection in large-scale milk mid-infrared spectrally derived databases through unsupervised learning



Sébastien Franceschini

COMMUNAUTÉ FRANÇAISE DE BELGIQUE
UNIVERSITÉ DE LIÈGE - GEMBLOUX AGRO-BIO TECH

Health pattern detection in large-scale milk mid-infrared spectrally derived databases through unsupervised learning

Sébastien Franceschini

Dissertation originale présentée en vue de l'obtention du grade de docteur en
sciences agronomiques et ingénierie biologique

Promotrice:

Pr. Hélène Soyeurt

Année civile: 2026

During the preparation of this manuscript, the author used ChatGPT 4.5 and 5.5 (OpenAI, San Francisco, United States), Paperpal by Editage (Cactus Communications, Mumbai, India), and Thesify (Thesify, Renens, Switzerland) as writing assistants. ChatGPT was used to challenge ideas, refine phrasing, improve readability, and generate or modify images. Thesify was used as a reviewer and editor to identify missing elements in the manuscript's flow, while Paperpal was used for grammar correction. The author reviewed and edited all generated content and takes full responsibility for the content of this publication. All the images generated by AI are mentioned. ChatGPT generated the cover image.

Abstract

Animal health is a key challenge for sustainable dairy farming, impacting productivity, profitability, drug use, and societal views on livestock production. Monitoring animal health on a large scale is difficult due to costly, heterogeneous, and often unavailable clinical diagnoses. Meanwhile, Fourier-transform mid-infrared (FT-MIR) spectrometry is globally used in milk recording and payment systems, producing extensive spectral data that can provide insights into animal physiology, metabolism, and herd management.

This thesis aimed to explore new methodologies for processing large-scale milk FT-MIR databases to develop data-driven monitoring tools for dairy farming. Specifically, it investigated unsupervised learning approaches to identify recurrent clusters in milk data. Unlike conventional supervised methods relying on predefined diagnoses or reference values, this methodology aimed to exploit the data's intrinsic structure to generate new hypotheses and monitoring indicators. Two main applications were developed: a cow-level approach using Holstein milk data and a herd-level approach based on bulk tank milk analyses. Clustering methodologies applied to FT-MIR-derived variables enabled the identification of distinct patterns related to management practices, metabolism disorders, and udder health.

Beyond methodology, this thesis emphasizes that validating unsupervised monitoring systems requires more than methodological performance. A multilayered validation framework was proposed, structured around five dimensions: cluster stability, transferability, predictability, biological validity, and practical utility. Stability refers to reproducibility across repeated analyses, while transferability evaluates robustness across populations, lactations, regions, or systems. Predictability assesses the capacity to predict new observations from different contexts, and biological validity examines whether clusters correspond to meaningful physiological or management-related phenomena. Practical utility considers operational feasibility in field conditions and return on investment.

The final validation phase involves on-farm deployment, incorporating a feedback loop where farmers and veterinarians verify system-generated alerts. This real-time validation would enable continuous model refinement through reinforcement learning, enhancing system adaptability and accuracy over time. Successful implementation requires robust data and computational infrastructure for rapid processing and alert dissemination. By integrating these concepts, this thesis promotes the development of precision dairy farming alert systems that are both scientifically rigorous and operationally effective. Overall, the thesis demonstrates the potential of combining FT-MIR spectrometry, big data, and unsupervised learning to develop scalable and interpretable monitoring tools for precision dairy farming.

Résumé

La santé animale constitue un enjeu majeur pour la durabilité des élevages laitiers, avec des répercussions sur la productivité, la rentabilité, l'utilisation des médicaments et la perception sociétale de la production animale. Le suivi de la santé animale à grande échelle demeure difficile en raison du coût, de l'hétérogénéité et de la fréquence des diagnostics cliniques. Pourtant, la spectrométrie moyen infrarouge à transformée de Fourier (FT-MIR), largement utilisée pour le contrôle laitier et le paiement du lait, génère d'importantes quantités de données spectrales fournissant des informations sur la physiologie animale, le métabolisme et la gestion des troupeaux.

Cette thèse a pour objectif d'explorer de nouvelles méthodologies de traitement des spectres FT-MIR du lait à grande échelle afin de développer des outils de suivi pour l'élevage laitier. Elle s'est intéressée à des approches d'apprentissage non supervisé visant à identifier des clusters au sein des données spectrales. Contrairement aux méthodes supervisées conventionnelles reposant sur des diagnostics prédéfinis ou des valeurs de référence, cette approche exploite la structure intrinsèque des données afin de générer de nouvelles hypothèses et de nouveaux indicateurs de suivi. Deux applications principales ont été développées : une approche à l'échelle de la vache utilisant des données de lait individuel et une approche à l'échelle du troupeau basée sur le lait de tank. Ces clustering ont permis d'identifier des profils associés aux pratiques de gestion, aux désordres métaboliques et à la santé du pis.

Au-delà des aspects méthodologiques, cette thèse souligne que la validation des systèmes de suivi non supervisés nécessite davantage qu'une simple évaluation de la performance statistique. Une approche de validation multicouche a été proposée autour de cinq dimensions : la stabilité des clusters, la transférabilité, la prédictibilité, la validité biologique et l'utilité pratique. La stabilité fait référence à la reproductibilité des résultats, tandis que la transférabilité évalue leur persistance entre populations, lactations, régions ou systèmes. La prédictibilité mesure la capacité à classer de nouvelles observations, et la validité biologique examine si les clusters correspondent à des phénomènes physiologiques ou de gestion pertinents. Enfin, l'utilité pratique concerne l'utilité et la faisabilité opérationnelle en conditions de terrain.

La phase finale de validation implique le déploiement du système sur site, en intégrant une boucle de rétroaction grâce aux retours des éleveurs et des vétérinaires qui vérifient les alertes générées. Ceci nécessite cependant une infrastructure informatique robuste permettant un traitement rapide des données et une diffusion efficace des alertes. Cette validation en temps réel permettrait une amélioration continue des modèles via l'apprentissage par renforcement. En conclusion, cette thèse démontre le potentiel de la combinaison de la spectrométrie FT-MIR, du big data et de l'apprentissage non supervisé pour développer des outils de suivi destinés à l'élevage laitier de précision.

Acknowledgments

« Et toi, Sébastien, comment se passent tes études ? Tu as encore beaucoup d'examens ? Tu comptes toujours les vaches ? » Je ne compte plus le nombre de fois que ces questions m'ont été posées durant ma thèse. Pourtant, elles soulignent le soutien sans faille que j'ai reçu de mon entourage durant ce long périple. Sans ce soutien, ce document n'existerait probablement pas. Une thèse est une aventure semée d'embûches, et les aventures sont toujours plus belles quand elles sont partagées.

Pour commencer, je souhaiterais remercier ma promotrice, Hélène Soyeurt. Ta confiance, ton encadrement, ton optimisme et ton soutien sans faille m'ont toujours aidé dans les moments de doute. Malgré ton emploi du temps chargé, tu as toujours su trouver le temps pour une réunion improvisée. J'ai beaucoup apprécié travailler avec toi. Preuve en est : je rempile !

Merci également à Nicolas Gengler de m'avoir accordé ta confiance et offert un poste de doctorant dans ton équipe, malgré le fait que je ne connaissais rien à l'élevage bovin laitier en Wallonie. Ce temps passé dans le laboratoire de génétique animale m'a beaucoup appris. De plus, tu as toujours eu des remarques très pertinentes pour améliorer mon travail.

Ensuite, je souhaite remercier mon comité de thèse : Sylvie Vanderick, Clément Grelet, Alain Gillon et Yves Brostaux, pour ces nombreuses réunions qui ont permis d'arriver au terme de ce doctorat. J'espère que mes présentations sur le clustering étaient de plus en plus claires au fil de ces six années.

Merci également à mon jury : Frédéric Dehareng, Philippe Lejeune, Laura Symul et Miel Hostens, d'avoir pris le temps de lire cette thèse avec intérêt et d'en avoir discuté longuement. Vos commentaires m'ont permis d'identifier certaines lacunes de ma recherche.

Merci à tout l'axe Modélisation et Développement pour cet environnement de travail si agréable. Nos pauses-café, nos barbecues et nos tartes ont été autant d'occasions de discuter et de profiter ensemble. Merci également à tous les TFEistes qui sont venus rejoindre notre équipe et qui ont participé de près ou de loin à ces recherches. Ne baissons pas les bras : un jour, le Fac Trophy sera nôtre.

Merci à toute l'équipe de zootechnie pour son accueil lors de mon arrivée à Gembloux. Particulièrement à Hélène Wilmot, qui m'a fait visiter ces vieux murs de Gembloux : merci pour tous ces fous rires. Notre compétition de Shifumi n'est pas finie !

Merci également à la team Padel : Arno, Loïc et Loze, pour ces midis remplis de Par3 et de Par4. Notre bandeja s'est bien améliorée depuis nos débuts.

Merci à Miel Hostens et à Cornell University de m'avoir accueilli pendant six mois aux États-Unis. Ce voyage m'a beaucoup apporté, tant professionnellement que personnellement. Merci également à toute l'équipe de Bovi-Analytics : Meike, Sonam, Hanlu, Sumit, Puchun, Rebecca, Haowen et Enhong, pour tous ces bons moments.

Cette aventure américaine aurait été différente sans vous.

Merci pour toutes ces belles rencontres faites à travers les congrès, les conférences et les différents projets : Julie, Marie-Nguyet, Octave, Huzaifa, Carlo, Jérémie et plein d'autres encore. Je garde de merveilleux souvenirs de ces moments.

Merci à toute ma famille d'être là, tout simplement. Je me rends compte que j'ai de la chance d'avoir une telle famille, même si, entre nous... trois sœurs, c'était parfois galère ! Le petit garçon qui posait toujours trop de questions en a fait du chemin. Papa, Maman, merci de m'avoir toujours soutenu malgré un parcours atypique : biologie, statistique, puis ingénierie biologique. Cela n'aurait jamais été possible sans vous ! Je tiens également à remercier toute ma belle-famille pour leur soutien et leur intérêt.

Merci à mes amis : Alexis, Arthur, François, Nicolas, Robin, Ulysse, Valentine et Victor, pour tous ces bons moments. Merci pour toutes ces soirées gaming, ces vacances et ces (trop ?) nombreux verres qui m'ont fait oublier le boulot. Je peux enfin dire que la première étape du plan est finalisée.

Merci également à la coloc : Alexis, Amélie, Marie et Maxime ! Quelle expérience que cette colocation en plein COVID. Je garde d'excellents souvenirs de ces moments partagés.

Enfin, merci à Claire, mon amour ! Tu m'as toujours soutenu inconditionnellement. Notre vie a bien évolué durant ces six années de thèse. Entre la colocation, notre appartement, Tanuki, la maison et, enfin, William, nous traçons notre chemin dans la vie à notre rythme. Merci pour tout cet amour, merci pour toute cette énergie positive ! Je t'aime.

Récemment, notre famille s'est agrandie. Je mentirais en disant que finir une thèse en étant jeune papa a été une expérience facile. Pourtant, William, si tu lis ces mots, merci d'avoir été un bébé incroyable. Je ne suis pas sûr que j'aurais été capable de finir ma rédaction dans d'autres conditions.

Contents

1	General introduction	1
1.1	Context	3
1.1.1	The European vision on the future of agriculture	3
1.1.2	The current state of dairy farming in Wallonia	4
1.2	Assessing animal health	5
1.2.1	Cost of health-related event	6
1.2.2	Monitoring technologies	6
1.2.3	Fourier-Transform Mid infrared	8
1.2.4	Modelling predictions FT-MIR spectrometry	11
1.2.5	FT-MIR spectrometry for health monitoring	13
1.2.6	Existing tools and their limits	14
1.3	Objectives: Toward data-driven welfare through unsupervised learning	16
2	Unsupervised learning techniques in precision dairy farming	27
2.1	Chapter outline	29
2.2	Abstract	29
2.3	Introduction	29
2.4	An unsupervised learning taxonomy	32
2.5	Dimensionality reduction	33
2.6	Clustering	34
2.6.1	Hierarchical	35
2.6.2	Partitional	38
2.6.3	Other categories	41
2.7	Examples of unsupervised learning algorithms	43
2.7.1	Partitional clustering algorithms: k-means	43

2.7.2	Partitional clustering algorithms: DBSCAN	45
2.7.3	Partitional clustering algorithms: Mean-shift	46
2.7.4	Partitional clustering algorithms: FASTCLUS	47
2.7.5	Partitional clustering algorithms: Affinity Propagation	48
2.7.6	Partitional clustering algorithms: Gaussian Mixture Model . .	48
2.7.7	Anomaly detection: Isolation Forest	49
2.8	Guidelines for validation and interpretation	51
2.8.1	Validation	51
2.8.2	Interpretation	52
2.9	Combining unsupervised learning and KDD	55
2.9.1	Selection	57
2.9.2	Preprocessing and transformation	57
2.9.3	Data mining: getting the best of the two approaches	59
2.10	Case studies in dairy farming	60
2.10.1	Biomarker focused studies	61
2.10.2	Health disorder-focused studies	61
2.10.3	Detection of oestrus	63
2.10.4	Cow behaviour	63
2.10.5	Livestock monitoring: herd level	63
2.10.6	Classification of farms/respondents to surveys	65
2.11	Conclusion and future trends	66
2.12	Where to look for further information	66
2.13	Acknowledgment	67
3	A first unsupervised learning methodology	81
3.1	Chapter outline	83
3.2	Abstract	83
3.3	Introduction	84
3.4	Material and methods	87
3.4.1	Calibration Data	87
3.4.2	Clustering Method	88
3.4.3	Cluster Interpretation	89
3.4.4	Expanding the Result to the Original Data Set	89
3.4.5	Validation	90

3.5	Results and discussion	93
3.5.1	Developing the Clusters	93
3.5.2	Predicting the Cluster for the Calibration Data Set	94
3.5.3	Validating Against Reference Values	97
3.6	Conclusions	100
3.7	Acknowledgments	100
3.8	Unpublished Supplementary Results	100
4	Reproducibility of FT-MIR predicted clusters	109
4.1	Chapter outline	111
4.2	Interpretive Summary	111
4.3	Highlights	112
4.4	Abstract	113
4.5	Study	113
4.6	Notes	121
4.7	Unpublished Supplementary Results	121
5	Detection of dairy herd management issues	127
5.1	Chapter Outline	129
5.2	Simple Summary	129
5.3	Abstract	130
5.4	Introduction	130
5.5	Materials and Methods	133
5.5.1	Belgian Dataset	133
5.5.2	Canadian Dataset	136
5.5.3	Unsupervised Learning	139
5.5.4	Hierarchical Clustering	139
5.5.5	Cluster Prediction	140
5.5.6	Interpretation	141
5.6	Results and Discussion	141
5.6.1	Clustering on Belgian Dataset	141
5.6.2	Cluster Prediction on Belgian and Canadian Datasets	144
5.6.3	Cluster Interpretation	148
5.6.4	From Clusters to Probabilities	154
5.6.5	Practical Implications and Study Limitations	155

5.7	Conclusions	155
5.8	Acknowledgments	156
6	General discussion	163
6.1	Introduction	165
6.1.1	The previous works on unsupervised learning in milk FT-MIR data	166
6.1.2	The scarcity of animal health references	167
6.1.3	A new approach to process FT-MIR data	169
6.1.4	Challenges arising from the approach	170
6.2	Process of multilayered validation for PLF	171
6.2.1	Cluster Stability	174
6.2.2	Cluster Transferability	175
6.2.3	Cluster Predictability	178
6.2.4	Cluster Biological Explainability	179
6.2.5	Cluster Practical Utility	181
6.2.6	Synthesis of the multilayered validation	185
6.3	Framework for a feedback-driven monitoring	185
6.3.1	Data acquisition, standardization, and automated processing	187
6.3.2	Data governance and confidentiality	188
6.3.3	Automated pattern detection, reporting, and feedback integration	188
6.3.4	Validation as an embedded iterative process	189
6.4	Conclusion	190
6.4.1	Limitations	191
6.4.2	Perspectives	192
6.4.3	Critical perspective on my thesis	193
A	Descriptive statistics of Chapter 4	203
B	Validation Efforts of Clustering Results	205

List of Figures

1.1	Summary of farm to fork	3
1.2	Evolution of the average number of dairy cows per farm in Wallonia	5
1.3	Average total cost of the major diseases	7
1.4	FT-MIR spectra	10
1.5	Standard processes for FT-MIR predictions	11
1.6	Standard process for the development of an equation	12
2.1	Unsupervised learning taxonomy	32
2.2	Dimensionality reduction	35
2.3	Different clustering techniques	36
2.4	Divisive and Agglomerative Hierarchical Clustering	36
2.5	Hierarchical Clustering	37
2.6	Concept of k-means clustering	40
2.7	Comparison of clustering algorithms	43
2.8	Knowledge Discovery in Database	56
3.1	Dendrogram of the clustering on 50,000 centroids	93
3.2	Barplot of the LSM by clusters	95
3.3	Scatterplot of the two first latent variables	96
4.1	Graphical Abstract	112
4.2	Methodology	116
4.3	Boxplots of the distributions of the different clusters	119
5.1	Fatty acid synthesis pathways	131
5.2	Workflow of the analysis	134
5.3	Dendrogram of the seven clusters	142

5.4 Projection of subset samples on the first two principal components . 143

5.5 Projection of Canadian data onto Belgian database 145

5.6 Evolution of cluster probabilities in time 154

6.1 Methodology from Tedde (2023) 166

6.2 Shift in the development of equations 168

6.3 Methods for large dataset clustering 177

6.4 Additive genetic variance explained for different traits of energy balance 180

6.5 Metabolic disorder alerts by days in milk 183

6.6 IT framework for monitoring tool development 186

List of Tables

- 1.1 Performance of multiples β -hydroxybutyrate equations 13
- 2.1 Examples of similarities metrics. 39
- 3.1 Published prediction models for health biomarkers 86
- 3.2 Performance of the equations used 87
- 3.3 Descriptive statistics of the mid-infrared-based predictors, milk yield, and SCC from the calibration data set ($n = 740,454$ records). 91
- 3.4 Descriptive statistics for the reference measurements of health-related traits from the validation sets 92
- 3.5 Descriptive statistics for the probability of being in a cluster for the validation records 97
- 3.6 Correlations between the probability of being in a cluster and reference measurements of biomarkers related to animal health 99
- 3.7 Performances of prediction model for the udder health cluster 101
- 4.1 Assignment of the 100 clusters (5 clusters per repetition $\times$ 5 repetitions per lactation dataset $\times$ 4 lactation datasets) to the five overall groups. 118
- 4.2 Comparison of event occurrence between good and bad alerts. 122
- 5.1 Descriptive statistics for the 39 FT-MIR-predicted traits used in the study calculated from the available Belgian and Canadian spectra. . . 137
- 5.2 Herd characteristics for 3006 Canadian farms. 138
- 5.3 Proportions of the seven clusters for the observations from the subset ($N = 27,322$). 142
- 5.4 Confusion matrix related to random forest algorithm applied on 31 predicted fatty acids ($N = 27,322$). 144
- 5.5 Percentages of Belgian and Canadian records by cluster. 145

5.6	Cluster transitions between 2 successive controls, expressed in %, for data collected in Belgium and Canada.	147
5.7	Means by cluster of the phenotypes predicted by FT-MIR from bulk tank milk	149
5.8	Percentages of organic and conventional farms in the various clusters, in % of herds present in the cluster.	152
5.9	Percentages of different feeding in the various clusters, in % of herds present in the cluster.	152
5.10	Means and standard deviations by cluster of management related data for 3006 Quebec herds.	153
6.1	Comparison of different frameworks related to multilayered validation.	173
6.2	Confusion matrix for a binary classification	183
A.1	Descriptive Statistics of the datasets from Chapter 4	204

Acronyms

R^2	Coefficient of Determination.
AI	Artificial Intelligence.
AMS	Automated Milking System.
ANOVA	Analysis of Variance.
AP	Affinity Propagation.
AUC	Area Under the Curve.
AWE	Walloon Breeders Association.
AWU	Annual Working Unit.
BHB	β -hydroxybutyrate.
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies.
C18:1cis9	oleic acid.
CA	Correspondence Analysis.
CAP	Common Agricultural Policy.
CURE	Clustering Using Representatives.
DBSCAN	Density-Based Spatial Clustering.
DHI	Dairy Herd Improvement.
DIANA	Divisive Analysis clustering.
DIM	Days In Milk.
DMI	Dry Matter Intake.
EB	Energy Balance.
EC	European Commission.
EEA	European Environment Agency.
EMR	European Milk Recording.
EU	European Union.

FA	Fatty Acid.
FAO	Food and Agriculture Organization.
FPCM	Fat- and Protein-Corrected Milk.
FT-MIR	Fourier Transform mid-infrared.
GA	Genetic Algorithm.
GAN	Generative Adversarial Networks.
GH	standardized Mahalanobis distance.
GLM	Generalized Linear Model.
GLRM	Generalized Low Rank Model.
GMM	Gaussian Mixture Model.
HDBSCAN	Hierarchical Density-Based Spatial Clustering.
ICA	Independent Components Analysis.
ICAR	International Committee for Animal Recording.
IDF	International Dairy Federation.
IoT	Internet of Things.
IT	Information Technology.
KDD	Knowledge Discovery in Database.
KDE	kernel density estimation.
LDA	Latent Dirichlet Allocation.
LDH	Lactate Deshydrogenase.
LEDs	logit-transformed Energy Deficit Score.
LLM	Large Language Models.
LOF	Local Outlier Factor.
LSM	Least Squares Means.
LSTM	Long Short-Term Memory.
MDS	Multidimensional Scaling.
ML	Machine Learning.
MRO	Milk Recording Organization.
MUFA	Monounsaturated Fatty Acid.
NAGase	N-acetyl- β -D-glucosaminidase.
NEFA	Non Esterified Fatty Acids.
P4	Progesterone.

PCA	Principal Component Analysis.
PDF	Precision Dairy Farming.
PLF	Precision Livestock Farming.
PLS	Partial Least Square.
PLS-DA	Partial Least Square Discriminant Analysis.
PUFA	Polyunsaturated Fatty Acid.
RF	Random Forest.
RMSE	Root Mean Square Error.
RPD	Ratio Performance to Deviation.
SCC	Somatic Cells Count.
SCS	Somatic Cells Score.
SFA	Saturated Fatty Acid.
SPW	Service Public de Wallonie.
SVD	Singular Value Decomposition.
t-SNE	t-Distributed Stochastic Neighbor Embedding.
THI	Temperature Humidity Index.
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction.
USDA	United States Department of Agriculture.
VAE	Variational auto encoders.

General introduction

1.1. Context

1.1.1. The European vision on the future of agriculture

In 2019, the European Commission (EC) launched the European Green Deal, a strategic policy framework developed in response to growing societal concerns about climate change and environmental degradation. The objectives of the Green Deal are first to cut carbon emissions by 50% by 2030 and then to make the European Union (EU) the first climate-neutral continent by 2050, integrating sustainability across all sectors of the economy while maintaining competitiveness (EC 2019).

Among these sectors, agriculture plays a pivotal role in this transition, with livestock production contributing approximately 14.5% of global anthropogenic greenhouse gas emissions (Gerber et al. 2013), including substantial enteric methane emissions (FAO 2023). Moreover, livestock farming shapes ecosystems through land use, resource management, and the maintenance of rural landscapes essential for biodiversity (Tscharntke et al. 2005).

The agri-food component of the Green Deal is the Farm to Fork Strategy (Figure 1.1), outlining how the Common Agricultural Policy (CAP) should promote healthier, fairer, and more environmentally friendly food systems in the EU (EC 2020). By reorienting financial aid and research priorities, this strategy seeks to encourage more sustainable production models, reduce environmental footprints, and improve resilience throughout the food chain, from primary producers to consumers (Arabska 2021). This resilience is particularly crucial, as farmers are among the most vulnerable to climate change, given their strong dependence on weather-sensitive production systems and the increasing frequency of extreme climate events (FAO 2021).



Figure 1.1: Summary of the farm to fork strategy. Figure from EC (2020).

According to the Eurobarometer, this transition toward more sustainable food systems is widely supported by the European population, with three-quarters of the respondents agreeing to change the way we produce to fight climate change, even if it means that EU agriculture will be less competitive (EC 2024b). However, the reforms associated with the CAP have not been met with equal enthusiasm among farmers. For European farmers, this policy implies a paradigm shift toward practices that must simultaneously ensure productivity, animal welfare, and efficient resource use, while remaining internationally competitive. As confirmed by European Environment

Agency (EEA), the challenge lies in reconciling the economic sustainability of farms with societal expectations for climate-smart and ethically responsible production systems (FAO 2020; EEA 2022; EAA 2024). Over recent decades, sustainability has evolved from a normative principle into a measurable framework guiding research, policy, and farm management practices (Pretty 2007), with the concept of planetary boundaries (Rockström et al. 2009). Many agricultural producers perceive these environmental and social ambitions as additional constraints rather than opportunities, particularly in contexts marked by economic uncertainty, administrative complexity, and rising production costs (Mennig 2024).

The persistent gap between policy objectives and field-level feasibility exemplifies the challenges of sustainability. This multidimensional concept encompasses environmental, economic, and social dimensions, which are referred to as the three pillars of sustainable development. These dimensions are interdependent and sometimes conflicting. For example, economic growth can exert pressure on ecosystems, whereas environmental preservation may challenge short-term profitability or social equity. In the agricultural sector, this complexity is particularly evident, as producers must balance productivity, animal welfare, and environmental impact in the context of fluctuating markets and evolving societal expectations (EAA 2024). As European farmers, many Walloon farmers consider the new environmental conditionalities of the CAP, while aligned with the objectives of the Green Deal, as misaligned with the current socio-economic difficulties they face (SPW 2025b).

1.1.2. The current state of dairy farming in Wallonia

Therefore, understanding how these European and national policy ambitions translate into the economic realities of Walloon farms is essential to assess the region's capacity to engage in the transition toward sustainable production. Across the EU, the agricultural sector has experienced a steady decline in the number of farms, with a decrease of 24.8% between 2010 and 2020, a trend accompanied by farm enlargement, land concentration, and growing dependence on volatile input and energy markets (EC 2024a). These structural changes have eroded the resilience of smaller farms. The state of agriculture in Wallonia is similar to the global situation in Europe. In the report about the state of Walloon agriculture, the Service Public de Wallonie (SPW) highlights that the average farm size has increased over the past two decades. However, profit margins have remained narrow, and generational renewal is insufficient to ensure long-term continuity (SPW 2025a).

With 55.2% of farms having cattle and 41.8% utilized agricultural areas for permanent grasslands, cattle farming is one of the major agricultural sectors in Wallonia. However, only 13.1% of the farms focuses only on dairy farming, showing that Walloon farms are often beef or mixed crop-livestock (SPW 2025c). Walloon dairy farming is also characterized by relatively small-scale, family operated companies, where herd sizes

(69 cows on average in 2023) are modest compared with those found in more intensive European or North American production systems (283 cows on average in the United States)(USDA 2025). As the European trend for farm size, the average herd size in Wallonia continues to increase (Figure 1.2).

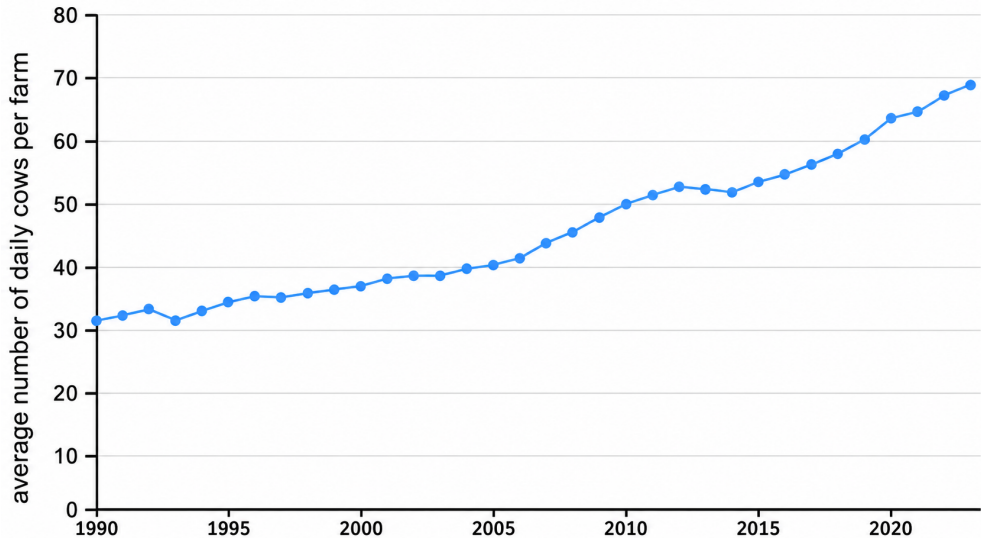


Figure 1.2: Evolution of the average number of dairy cows per farm in Wallonia (SPW 2025a).

The average dairy farm labor income per Annual Working Unit (AWU) in Wallonia is 43,179€/AWU. However, a large disparity is observed among Walloon specialized dairy farms, with 10% of the farms earning more than €100,000/AWU, while an equivalent proportion reports a negative income (SPW 2025a). This difference is associated with the milk production costs of high- and low-performing farms, which are 48.48€/100 L and 74.88€/100 L of milk, respectively. One explanation could be farm management system, but the SPW (2025a) also highlights the geographical area or the farmer's health, which is independent of management. Those numbers show that while dairy farmers are in a tough position, there are opportunities for profitability by increasing production efficiency.

1.2. Assessing animal health

Several strategies should be considered to optimize production efficiency, including increase in production, enhancing feed efficiency or reducing costs associated with animal health events. This thesis focuses on improving animal health through monitoring, which has the advantage of also improving welfare. Moreover, health-related costs, such as veterinary treatments, reduced milk yield during illness, and extended calving intervals, represent a significant but often underestimated component of production

costs.

1.2.1. Cost of health-related event

On a global scale, Rasmussen et al. (2024) estimated annual global losses from 12 major dairy cattle diseases at approximately 65 billion US\$ annually, with 18 billion US\$ associated with subclinical ketosis, 13 billion US\$ with clinical mastitis, and 9 billion US\$ with subclinical mastitis, showing that subclinical diseases account for a large part of the losses. Puerto et al. (2021) showed that mastitis reduced the cumulative milk yield by 382 to 989 kg per lactation in primiparous cows, depending on the stage of lactation. Hagnestam et al. (2007) estimated a loss between 0 to 705 kg in primiparous cows and 0 to 902 kg for multiparous cows. In other words, mastitis can impose huge losses on a farm per year. Turning to metabolic disorders, a systematic review by Cainzos et al. (2022) estimated that the economic cost per case of ketosis in dairy cows ranged widely from 19€ to 812€ depending on the study context but highlighted that all studies considered reported a significant impact of ketosis on the dairy industry. Liang et al. (2017) estimated the average total cost per case of the seven major diseases from 77 US\$ for a ketosis to 639.51 US\$ for a left-displaced abomasum (Figure 1.3).

Hence, when considering the economic sustainability of dairy systems, particularly when margins are small, assessing animal health is a key aspect. Monitoring animal health allows for early or subclinical disease detection but also disease prevention. It enables better overall management, which is a central lever influencing farm margins.

Beyond their direct economic impact, health issues compromise the overall sustainability of the production system through animal welfare and environmental efficiency. Therefore, improving herd health management through prevention or monitoring constitutes both an economic opportunity and a pathway toward more sustainable dairy production, aligning animal welfare and farm profitability. While prevention is also essential, this thesis focuses on monitoring opportunities that are highly relevant in the context of increasing herd size. However, improved health monitoring can also provide a better understanding of health issues, which could ultimately contribute to more effective prevention.

1.2.2. Monitoring technologies

In the context of supporting dairy farms in Wallonia, various monitoring technologies offer pathways to improve animal health, productivity, and sustainability without necessarily imposing large additional capital costs. With the rise of Agriculture 4.0, the Internet of Things, and the increasing availability of large on-farm and inter-farm databases, a wide range of data-driven monitoring strategies are emerging, enabling farmers to make more informed and proactive management decisions (Tangorra et al. 2024; Hostens et al. 2025).

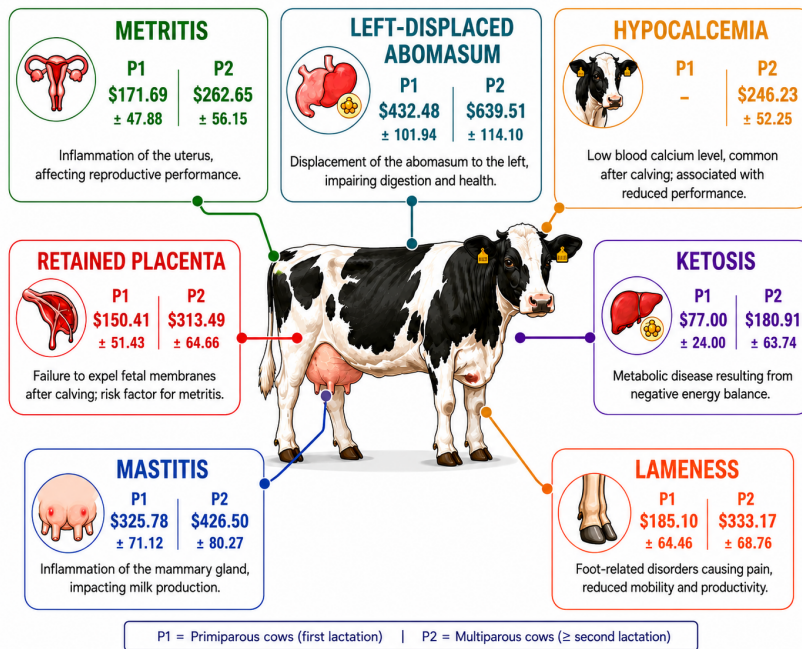


Figure 1.3: Average total cost of the major diseases. Generated by ChatGPT using data from Liang et al. (2017)

Key monitoring options include daily milk yield (Salamone et al. 2022), rapid on-farm tests with quick tests for mastitis or metabolic disorders (Zeineldin et al. 2021), automated milking systems (Cogato et al. 2021), sensor-based systems for activity, rumination, or temperature sensors (Liseune et al. 2021), and computer-vision systems that monitor behaviors such as lying or standing patterns via cameras (Borges Oliveira et al. 2021). Decision tools based on such data already exist such as MmmooOgle (Puurs-St-Amans, Belgium) mainly for feed management or RumiSense (Rumexperts, Faimés, Belgium) for sensors and weather.

Another potential option, currently underused for the specific purpose of monitoring, is the Fourier Transform mid-infrared (FT-MIR) spectral analysis of milk with tools such as PROFILab (Lactanet 2022), CetoLait (awé-Elevéo 2021), MastiMIR (Rienès et al. 2022a), and RumiSense (Rumexperts, Faimés, Belgium), which rely on fine milk composition.

A robust on-farm monitoring tool for dairy cows must fulfil several key criteria to be truly effective and applicable. First, validity and reliability are essential: the tool must accurately detect or measure the trait of interest (e.g., health, feeding, production) under real-world farm conditions, with acceptable sensitivity, specificity, and error margins (Stygar et al. 2021). Second, it must provide continuous or timely information rather than infrequent snapshots, enabling the early detection of deviations and

supporting proactive decision-making (Silva et al. 2023). Third, the system should be non-invasive, cost-effective, and user-friendly, so it can integrate seamlessly into farm workflows rather than creating an additional burden (Bell et al. 2018). Finally, it must produce actionable insights; that is, the output must be interpretable and linked to management decisions (feed adjustment, health intervention, reproductive planning) rather than simply raw data (Simoni et al. 2024). In summary, a monitoring tool should provide reliable information at a relevant time without being a burden for the farmer or animal.

While the current use of FT-MIR technology does not meet all these criteria, mostly frequent data acquisition for milk recording, its use is promising because it leverages data already collected through the milk recording and milk payment scheme and hence does not require substantial extra hardware investment on the farm. In a small-farm, margin-tight context, the key constraint is using already existing data or infrastructure and avoiding major new costs. In that respect, the monitoring options that leverage existing milk recording data or on-farm data are most aligned with both economic and environmental sustainability, because they minimize additional investment and thus require fewer resources.

Research has shown that milk FT-MIR spectra have been used beyond routine schemes to derive proxies of health and indicators of cows and herds (Soyeurt 2023; Ramirez Mauricio et al. 2025). Thus, by exploiting the existing milk-recording infrastructure, farms may gain improved information at a marginal cost. Sensor systems and computer vision offer similar advantages with continuous monitoring of behavior and health indicators (e.g., lameness and lying time) but are more costly to implement.

Therefore, given the structural characteristics of Walloon dairy farms, prioritizing monitoring options that (1) use existing data, (2) integrate with existing workflows, and (3) provide actionable insights for health, welfare, feed, and environment is a prudent strategy. These options could collectively help reconcile the three pillars of sustainability: supporting profitability, reducing resource use and emissions, and improving animal welfare and the social acceptability of dairy production.

1.2.3. Fourier-Transform Mid infrared

The determination of major milk components is a cornerstone of modern dairy production, as it is used for both milk payment systems and herd performance evaluation. Historically, the need to remunerate producers according to milk quality parameters, such as fat and protein content, has driven the development of rapid and reliable analytical techniques capable of processing thousands of samples per day (Dehareng 2023).

Among these, FT-MIR spectrometry has become the reference technology worldwide for routine milk analysis. FT-MIR spectrometry measures the absorption or transmission of infrared light by a milk sample across specific wavelengths in the

mid-infrared region (typically 900–4000 cm^{-1}). The law of Beer-Lambert defines the relation between the absorption and the concentration in milk, which allow for a quantification of different components (Dehareng 2023). Each wavelength in the mid-infrared region is associated with specific molecular vibrations such as C-H, O-H or C=O groups; reflecting the chemical composition of milk (Figure 1.4). However, these absorption bands are often broad and overlapping. Consequently, quantitative estimation of major milk components such as fat, protein, and lactose is achieved through chemometric calibration models that exploit the multivariate structure of the milk spectra. Because of its direct relation with the milk composition, FT-MIR spectra vary according to many factors such as animal breed or health status, feed and farm management or environmental effect such as the weather (Bahadi et al. 2021; Soyeurt et al. 2006b; Coppa et al. 2021).

Historically, predictions were limited to major milk components, which were directly estimated by the instrument using equations provided by the manufacturer (Figure 1.5). However, advances in chemometric modelling combined with the export of the spectra outside of the instruments have extended their use to the prediction of minor constituents, such as individual fatty acids (Soyeurt et al. 2006a) or minerals (Christophe et al. 2021), and even indirect traits that are physiologically or metabolically linked to milk composition, such as dry matter intake (Tedde et al. 2021), methane (Vanlierde et al. 2021), or metabolic status (Grelet et al. 2019). Thus, the strategy evolved from collecting major milk components directly from the instrument to exporting the spectra to develop prediction models which are called equations (Figure 1.5). This strategy allows the building of large database of milk FT-MIR spectra.

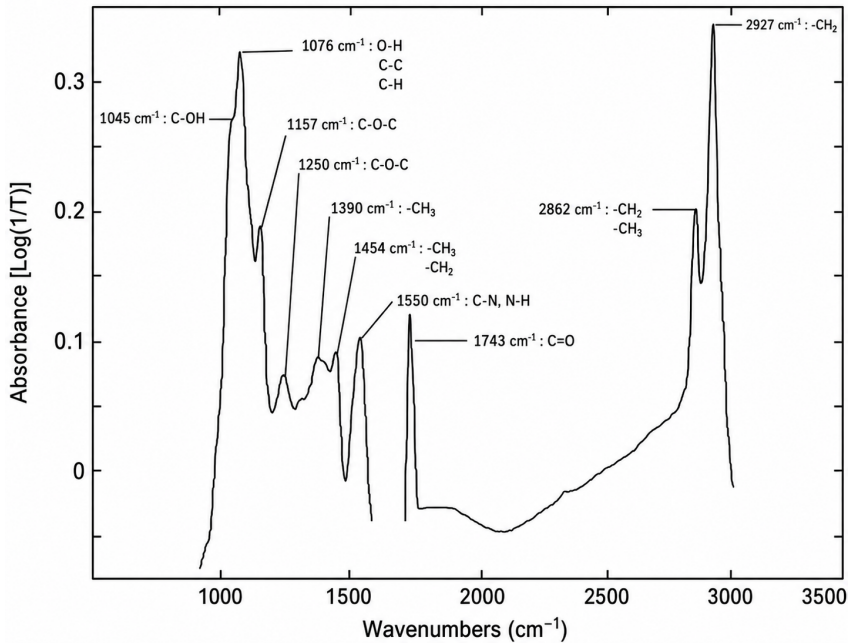


Figure 1.4: FT-MIR spectra (Grelet et al. 2015)

Nowadays, milk FT-MIR spectrometry is routinely implemented in two main contexts: milk recording, performed within Dairy Herd Improvement (DHI) programs to support genetic and management evaluations, and milk payment analysis, conducted by dairy laboratories to determine the price and quality of delivered milk. In both cases, the analytical process is subject to rigorous standardization and quality control to prevent analytical drift within laboratories and ensure comparable results across laboratories. Internationally, the methodology is defined by the International Organization for Standardization (ISO 9622:1999 and ISO 9622:2013), the International Dairy Federation (IDF) and the International Committee for Animal Recording (ICAR), which provides detailed guidelines, reference materials, and certification procedures to ensure data harmonization across laboratories and countries. These efforts were important to guarantee that the fat and protein contents predicted using FT-MIR were reliable and comparable. This reliability is essential because these contents are used to fix the milk price.

However, the increasing number of phenotypes predicted through FT-MIR and the export of the spectra outside of the machine quickly became a problem for a traditional calibration where every lab should measure the same samples at the same time for every phenotype. The process becomes logistically complex and can not ensure accurate and independent calibration for each phenotypes. Instead, it was decided to directly standardize the spectra provided by the instrument based on this

principle: A milk sample should provide the same FT-MIR spectra independently of the instrument or the laboratory through spectral standardization (Figure 1.5).

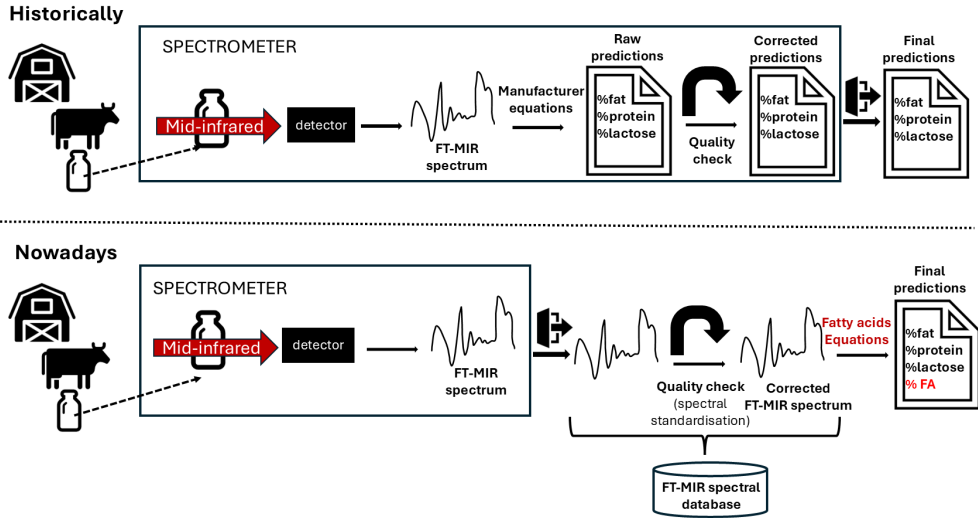


Figure 1.5: Standard processes for FT-MIR predictions. FA = Fatty Acids

Researchers have shown that this is achievable, even though additional steps are required to ensure that spectral information is comparable between laboratories, instruments, and manufacturers and remains comparable over time, which is essential for monitoring purposes (Grelet et al. 2015; Grelet et al. 2017). Having spectra generated in different regions, being comparable and reliable is the foundation for further applications, such as health monitoring and sustainability assessment within the dairy sector. Spectral standardization is a critical prerequisite for data comparability. Techniques such as direct standardization or piecewise direct standardization use transfer samples analyzed on both a reference and secondary spectrometer to align the spectra mathematically. Multiple methods already exist such as the European Milk Recording (EMR) correction developed by Grelet et al. (2015, 2017) or the methodology of Bonfatti et al. (2017) or Mensching et al. (2025). Once the infrared spectrum of a milk sample is acquired and standardized, previously developed prediction models are applied to estimate the traits of interest, or new models are developed for additional applications.

1.2.4. Modelling predictions FT-MIR spectrometry

The development of a new prediction model, also called an equation, follows several methodological steps to ensure accuracy and robustness. First, a representative and sufficiently variable calibration dataset must be assembled, including milk samples collected from different breeds, production systems, and physiological stages. De-

pending on the variable to be predicted, each sample is analyzed using reference methods to determine real value of the target variables (e.g., fatty acid composition, minerals, or other metabolites). The FT-MIR spectra from the same milk samples are also recorded under standardized conditions. The variability captured in both the reference and spectral data is essential, as calibration equations perform poorly when extrapolated beyond their calibration domains (Grelet et al. 2021). Before modelling, the spectra are pre-processed using techniques such as baseline correction, smoothing, normalization, or derivation, to minimize instrumental noise and enhance informative spectral regions (Rinnan et al. 2009; Tiplady et al. 2020; Zhang et al. 2021). A wide range of statistical and machine learning algorithms can then be used to link spectra to reference data, including most usually Partial Least Square (PLS) regression but other algorithms can also be used like random forests, and artificial neural networks, each offering different balances between interpretability and flexibility (De Marchi et al. 2014; Soyeurt et al. 2020). Model performance is first evaluated by internal cross-validation (e.g., k-fold or leave-one-out) to fix the hyperparameters and to assess the predictive stability within the calibration set, and subsequently by external validation using independent samples to ensure robustness and transferability across populations or instruments. The folds and external validation should be as independent as possible from the rest of the dataset to avoid data leakage. A common choice is to stratify at herd level (Grelet et al. 2021). The Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Ratio Performance to Deviation (RPD) are typically reported as performance metrics. Once validated, these models can be integrated into routine FT-MIR analytical pipelines for large-scale phenotyping, genetic evaluation, and health monitoring applications in dairy systems. The whole process is presented in Figure 1.6.

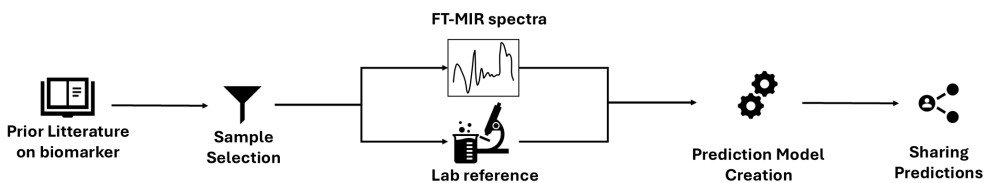


Figure 1.6: Standard process for the development of an equation

Through this combination of chemometric modelling and international standardization, FT-MIR spectroscopy has evolved from a routine compositional tool into a high-throughput phenotyping platform. It now enables the prediction of a broad range of traits, including technological properties (e.g., coagulation ability), nutritional quality (e.g., fatty acid profile), metabolic status, and environmental indicators, such as methane emissions (Vanlierde et al. 2021) or nitrogen-use efficiency (Grelet et al. 2020). This diversity of predicted traits, while not always accurately predicted, explains

why FT-MIR spectroscopy, among other technologies, is increasingly viewed as a cornerstone of data-driven dairy farming.

1.2.5. FT-MIR spectrometry for health monitoring

To assess animal health status from milk spectra, FT-MIR spectrometry has progressively extended beyond the quantification of major milk constituents to encompass more complex traits, such as minor milk components and indirect traits. Milk reflects numerous systemic processes, and its composition changes in response to metabolic stress, inflammation, or infection (Xu et al. 2019; Mota et al. 2021; Rienesl et al. 2022b; Grelet et al. 2024). Bahadi et al. (2021) even showed that FT-MIR spectra could reflect housing conditions and welfare status. Consequently, researchers have identified biomarkers and developed calibration models to predict health-related proxies, such as β -hydroxybutyrate (BHB), acetone, citrate, lactoferrin, and lactate dehydrogenase, directly from the milk spectrum (Grelet et al. 2016; Soyeurt et al. 2020). These models enable noninvasive pseudo-diagnosis of major disorders, such as ketosis, mastitis, and energy imbalance, providing a potentially powerful tool for large-scale monitoring.

However, the link between biomarkers and disease for individual diagnosis remains methodologically and practically challenging. The performances of the prediction models are not always reliable depending on the biomarker. Table 1.1 shows the performance of the different equations published for BHB. Those performances directly depend on the signal strength in the FT-MIR spectra, the reference method for the biomarker, the reference population and the quantity and quality of the diagnosis. Obtaining accurate diagnoses on farms is inherently difficult because the clinical data are often incomplete, heterogeneous, or inconsistently recorded. Moreover, hundreds of observations are required to develop a robust equation which make it costly to collect but also unethical to induce or even to measure if specific protocol is required such as blood sampling. Even when data are available, diagnostic definitions and thresholds for specific disorders, such as subclinical ketosis or mastitis, vary widely across studies and regions, resulting in low reproducibility (Bonfatti et al. 2019; Grelet et al. 2019).

Table 1.1: Performance of multiples β -hydroxybutyrate equations

Study	N	R2c	R2v	Range (mmol/L)
Pralle et al. (2018) ¹	3629	0.46 ³	0.56	0.01-5.4
Bonfatti et al. (2019)	1910	0.56	0.50	0.15-4.00
Mota et al. (2021)	449	0.92	0.77	0.22-0.99
Grelet et al. (2016) ²	558	0.71 ³	0.63	0.045- 1.596
Grelet et al. (2024) ²	7166	0.72	0.58	0.002-1.989

¹ = using test-day variables. ² milk β -hydroxybutyrate. ³ = accuracy of cross-validation.

Moreover, one biomarker is not always sufficient to detect a disease, and the available value is only FT-MIR predictions with an associated error. For example, blood BHB cut-offs for subclinical ketosis range from 1.0 to 1.4 mmol/L in the literature, depending on the sampling protocol and reference test (Leblanc 2010; Suthar et al. 2013), whereas the root mean square error associated with the equation from Pralle et al. (2018) is between 0.16 and 0.17 mmol/L. This variability introduces bias into supervised model development, as predictions become dependent on context-specific and sometimes subjective disease definitions. In conclusion, finding one trusted biomarker for animal health status and diseases prediction from milk FT-MIR spectra remains one limiting factor of monitoring dairy cows using spectrometers.

Animal health is inherently multidimensional. Disorders such as metabolic imbalance or inflammation cannot be fully captured by a single biomarker, as they arise from complex physiological interactions involving energy, protein, and metabolic pathways. For example, metabolic imbalance often begins with insufficient energy intake, leading to fat mobilization and increased production of non-esterified fatty acids, which are subsequently converted into ketone bodies. All those changes are reflected in the milk composition ensuring a strong signal in FT-MIR spectra. Furthermore, this metabolic stress can impair immune function and ultimately increase susceptibility to diseases such as mastitis or metritis. Consequently, using one biomarker or its MIR-based prediction as a diagnostic proxy may be too simplistic and fail to capture the broader physiological state of the animal. This limitation highlights the need for more holistic and data-driven approaches. Because the diagnosis is rarely available and the definition not always clear, some researchers naturally turn to unsupervised learning approaches where a response variable is not required (De Koster et al. 2019). However, they used small dataset with references instead of FT-MIR predictions.

1.2.6. Existing tools and their limits

Tools and metrics using multiples biomarkers already exist. As mentioned before, PROFILab (Lactanet, Sainte-Anne-de-Bellevue, Canada), CetoLait (AWé, Ciney) or MastiMIR (LKV Baden-Württemberg, Stuttgart) are monitoring tools based on the spectral information (Lactanet 2022; awé-Elevéo 2021; Rienesl et al. 2022a). However, they have some limits. Currently, PROFILab only reports the predicted values for different traits, mainly fatty acids, and compares the values to the averages for regions, breeds or management systems. While it is highly informative, it requires trained people able to understand those values and the complex interactions in order to give advices to the farmers. Moreover, there are no automated alerts systems. CetoLait developed an alert system for ketosis based on four FT-MIR predictions: the fat to protein ratio, the oleic acid, the acetone and the BHB in milk. This alert was created using diagnoses but the number of records available is unknown. MastiMIR also developed an alert and the tool uses the entire spectra. However, the biomarkers

used as diagnoses to train the alert model is a threshold on the Somatic Cells Count (SCC), which is directly related to mastitis and not predicted by FT-MIR. Thus, SCC may not encompass the whole complexity of mastitis such as the different possible origins(Rienesl et al. 2022a; Lin et al. 2025).

So the problem of an alert monitoring system using FT-MIR is summarized in these four points:

1. References and diagnoses data are very limited.
2. Diseases are rare events and complex to measure.
3. Some FT-MIR predictions are not accurate.
4. Using directly FT-MIR predictions in routine monitoring is complex and require formation.

In this context, the integration of multiple FT-MIR-predicted traits and classical performance indicators using unsupervised learning methods offers a promising alternative. The diagnoses are no longer required in the first-step, the amount of records allow for the observation of rare events, the many FT-MIR predictions should capture the multidimensional aspect of diseases and the low accuracies could be balanced by the other predictions. Indeed, sick animals are expected to have higher or lower values than healthy animals for multiples different predicted traits simultaneously. This different data pattern should be detectable in large database. By analyzing large-scale milk recording data without requiring predefined diagnostic labels, unsupervised algorithms, such as hierarchical clustering or anomaly detection, can detect latent patterns or anomalies that correspond to health issues as researchers shown by Shao et al. (2022) using a large dataset of electronic medical records for human health or Sogandi (2024) using association rules. This strategy leverages the routine availability of FT-MIR spectra within DHI programs, thus enabling large-scale, low-cost health monitoring without the need for additional sampling or laboratory analyses. Such approaches are particularly suited to sustainability goals because they improve animal welfare and productivity while maximizing the informational value of existing data resources.

Ultimately, the application of FT-MIR spectrometry for health monitoring illustrates a shift in dairy data science, from the measures of major milk component, to the pursuit of single-biomarker diagnostics, to the exploration of multivariate alerts and now to health issues detection using the whole pupulation data integrating a lot of variability. This shift not only enhances the biological interpretability of FT-MIR data but also supports more integrative monitoring of livestock systems.

1.3. Objectives: Toward data-driven welfare through unsupervised learning

Building on the potential of FT-MIR spectrometry as a large-scale phenotyping tool, this thesis explores how unsupervised learning methods can extract new insights from the routinely collected milk spectral data. As discussed, the challenge of accurately defining health diagnoses and the multidimensional nature of animal health call for analytical strategies that do not rely on predefined labels or simplistic thresholds. Unsupervised approaches enable the identification of latent structures and emergent patterns in milk spectral predictions, offering a complementary perspective to traditional supervised models (Shao et al. 2022; Sogandi 2024). These methods provide an opportunity to transform routinely collected FT-MIR spectra, originally intended for compositional analysis or milk payment, into actionable indicators of animal and herd health status.

This approach is innovative as it is a new way to consider the large amount of FT-MIR data collected in routine and create simple alerts to observe anomalies in the database. By integrating these alerts with the existing pipelines, this would also allow for quick detection giving opportunities for samples collection, vet's diagnoses or farmer's feedbacks. Thus the research question becomes: Using the large amount of FT-MIR spectra and their predictions, is it possible to detect complex health issue patterns through unsupervised learning approaches and bring back those patterns as simpler alerts on-farm to improve management but also collect supplementary information?

After introducing the global context of this thesis in this first chapter, the second chapter presents the conceptual and methodological framework of unsupervised learning in the context of precision dairy farming. This book chapter establishes a theoretical basis for subsequent applications by reviewing clustering, dimensionality reduction, and validation techniques suited for large, high-dimensional datasets. The third chapter presents the first empirical application of this framework, using milk recording data to uncover health-related patterns among dairy cows under routine field conditions in Wallonia, Belgium. The fourth chapter extends this approach across multiple lactations and develops methods to assess the robustness and transferability of clustering-based health indicators under different conditions. The fifth chapter applies a similar methodology to milk payment data, where a higher sampling frequency offers the possibility of near-continuous monitoring of herd dynamics and milk quality. Finally, the general discussion presented in the sixth chapter integrates the findings of these previous chapters to evaluate the relevance of unsupervised learning as a monitoring and decision-support tool in the dairy sector while addressing the practical and conceptual challenges of implementing such data-driven systems in farm management.

Through this structure, the thesis aims to determine whether unsupervised learning

applied to FT-MIR based phenotypes and associated routine data detects the different data patterns of sick dairy cows or herds, to assess whether it is useful as a health-monitoring system for dairy farmers and lay the foundation for future use in routine. This system would contribute to bridging the gap between research and on-farm decision-making, supporting more efficient, and welfare-oriented dairy systems by maximizing the informational value of data already available within the existing infrastructures.

Bibliography

- Arabska, E. (2021). "From farm to fork: Human health and well-being through sustainable agri-food systems". In: *Journal Of Life Economics* 8.1, pp. 11–27. ISSN: 2148-4139. DOI: 10.15637/jlecon.8.1.02.
- awé-Elevéo (2021). *Cétolait : outil d'analyse du risque d'acétonémie dans les troupeaux laitiers*. PDF brochure. awé. URL: <https://www.awenet.be/awe/UserFiles/file/asbl/lait/Services/Folder-CETOLAIT-2021.pdf> (visited on 01/23/2026).
- Bahadi, M., A. A. Ismail, and E. Vasseur (2021). "Fourier Transform Infrared Spectroscopy as a Tool to Study Milk Composition Changes in Dairy Cows Attributed to Housing Modifications to Improve Animal Welfare". In: *Foods* 10.2. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 450. DOI: 10.3390/foods10020450.
- Bell, M. J. and G. Tzimiropoulos (2018). "Novel Monitoring Systems to Obtain Dairy Cattle Phenotypes Associated With Sustainable Production". In: *Frontiers in Sustainable Food Systems* 2. Publisher: Frontiers. ISSN: 2571-581X. DOI: 10.3389/fsufs.2018.00031.
- Bonfatti, V., S.-A. Turner, B. Kuhn-Sherlock, T. D. W. Luke, P. N. Ho, C. V. C. Phyn, and J. E. Pryce (2019). "Prediction of blood β -hydroxybutyrate content and occurrence of hyperketonemia in early-lactation, pasture-grazed dairy cows using milk infrared spectra". In: *Journal of Dairy Science* 102.7, pp. 6466–6476. ISSN: 1525-3198. DOI: 10.3168/jds.2018-15988.
- Bonfatti, V., D. Vicario, A. Lugo, and P. Carnier (2017). "Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle". In: *Journal of Dairy Science* 100.7, pp. 5526–5540. ISSN: 0022-0302. DOI: 10.3168/jds.2016-11667.
- Borges Oliveira, D. A., L. G. Ribeiro Pereira, T. Bresolin, R. E. Pontes Ferreira, and J. R. Reboucas Dorea (2021). "A review of deep learning algorithms for computer vision systems in livestock". In: *Livestock Science* 253, p. 104700. ISSN: 1871-1413. DOI: 10.1016/j.livsci.2021.104700.

- Cainzos, J. M., C. Andreu-Vazquez, M. Guadagnini, A. Rijpert-Duvivier, and T. Duffield (2022). "A systematic review of the cost of ketosis in dairy cattle". In: *Journal of Dairy Science* 105.7. Publisher: Elsevier, pp. 6175–6195. ISSN: 0022-0302. DOI: 10.3168/jds.2021-21539.
- Christophe, O. S., C. Grelet, C. Bertozzi, D. Veselko, C. Lecomte, P. Höeckels, A. Werner, F.-J. Auer, N. Gengler, F. Dehareng, and H. Soyeurt (2021). "Multiple Breeds and Countries' Predictions of Mineral Contents in Milk from Milk Mid-Infrared Spectrometry". In: *Foods* 10.9. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 2235. DOI: 10.3390/foods10092235.
- Cogato, A., M. Brščić, H. Guo, F. Marinello, and A. Pezzuolo (2021). "Challenges and Tendencies of Automatic Milking Systems (AMS): A 20-Years Systematic Review of Literature and Patents". In: *Animals : an Open Access Journal from MDPI* 11.2, p. 356. ISSN: 2076-2615. DOI: 10.3390/ani11020356.
- Coppa, M., B. Martin, S. Hulin, J. Guillemin, J. V. Gauzentes, A. Pecou, and D. Andueza (2021). "Prediction of indicators of cow diet composition and authentication of feeding specifications of Protected Designation of Origin cheese using mid-infrared spectroscopy on milk". In: *Journal of Dairy Science* 104.1, pp. 112–125. ISSN: 0022-0302. DOI: 10.3168/jds.2020-18468.
- De Koster, J., M. Salavati, C. Grelet, M. A. Crowe, E. Matthews, R. O'Flaherty, G. Opsomer, L. Foldager, and M. Hostens (2019). "Prediction of metabolic clusters in early-lactation dairy cows using models based on milk biomarkers". In: *Journal of Dairy Science* 102.3, pp. 2631–2644. ISSN: 0022-0302. DOI: 10.3168/jds.2018-15533.
- De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa (2014). "Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits". In: *Journal of Dairy Science* 97.3. Publisher: Elsevier, pp. 1171–1186. ISSN: 0022-0302. DOI: 10.3168/jds.2013-6799.
- Dehareng, F. (2023). "Spectrométrie moyen infrarouge dans le lait : contribution à l'étude des facteurs principaux visant à accroître la qualité des modèles de prédiction et à assurer leur contrôle". HDL: <https://orbi.uliege.be/handle/2268/300713>. PhD Thesis. ULiège - Gembloux Agro-Bio Tech, Gembloux, Belgium.
- EAA (2024). *Solutions for restoring Europe's agricultural ecosystems*. European Environment Agency Briefing. LU: Publications Office. URL: <https://data.europa.eu/doi/10.2800/1811894> (visited on 10/14/2025).
- EC (2019). *The European Green Deal*. In collab. with E. Commission. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640> (visited on 10/13/2025).
- EC (2020). *From farm to fork: our food, our health, our planet, our future : the European Green Deal*. eng. LU: Publications Office. DOI: 10.2875/653604.

- EC (2024a). *Europeans, agriculture and the CAP: report*. LU: European Commission. Directorate General for Agriculture and Rural Development. URL: <https://data.europa.eu/doi/10.2762/5962409> (visited on 10/19/2025).
- EC (2024b). *Key figures on the European food chain: 2023 edition*. European Commission, Statistical Office of the European Union. URL: <https://data.europa.eu/doi/10.2785/265789> (visited on 10/13/2025).
- EEA (2022). *Transforming Europe's food system: assessing the EU policy mix*. LU: Publications Office. URL: <https://data.europa.eu/doi/10.2800/295264> (visited on 10/14/2025).
- FAO (2020). *The State of Food and Agriculture 2020: Overcoming Water Challenges in Agriculture*. Rome: Food and Agriculture Organization of the United Nations (FAO). ISBN: 978-92-5-133441-6. DOI: 10.4060/cb1447en.
- FAO (2021). *The impact of disasters and crises on agriculture and food security: 2021*. Food and Agriculture Organization of the United Nations (FAO). ISBN: 978-92-5-134071-4. DOI: 10.4060/cb3673en.
- FAO (2023). *Methane emissions in livestock and rice systems. Sources, quantification, mitigation and metrics*. English. Rome, Italy: FAO. ISBN: 978-92-5-138148-9. DOI: 10.4060/cc7607en.
- Gerber, P. J., H. Steinfeld, B. Henderson, A. Mottet, C. Opio, J. Dijkman, A. Falcucci, and G. Tempio (2013). *Tackling Climate Change through Livestock: A Global Assessment of Emissions and Mitigation Opportunities*. Rome: Food and Agriculture Organization of the United Nations (FAO). ISBN: 978-92-5-107920-1. URL: <https://www.fao.org/3/i3437e/i3437e.pdf>.
- Grelet, C., C. Bastin, M. Gelé, J. -. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng (2016). "Development of Fourier transform mid-infrared calibrations to predict acetone, β -hydroxybutyrate, and citrate contents in bovine milk through a European dairy network". In: *Journal of Dairy Science* 99.6, pp. 4816–4825. ISSN: 0022-0302. DOI: 10.3168/jds.2015-10477.
- Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng (2015). "Standardization of milk mid-infrared spectra from a European dairy network". In: *Journal of Dairy Science* 98.4, pp. 2150–2160. ISSN: 0022-0302. DOI: 10.3168/jds.2014-8764.
- Grelet, C., E. Froidmont, L. Foldager, M. Salavati, M. Hostens, C. P. Ferris, K. L. Ingvarsen, M. A. Crowe, M. T. Sorensen, J. A. Fernandez Pierna, A. Vanlierde, N. Gengler, and F. Dehareng (2020). "Potential of milk mid-infrared spectra to predict nitrogen use efficiency of individual dairy cows in early lactation". In: *Journal of Dairy Science* 103.5, pp. 4435–4445. ISSN: 0022-0302. DOI: 10.3168/jds.2019-17910.
- Grelet, C., T. Larsen, M. Crowe, D. Wathes, C. Ferris, K. Ingvarsen, C. Marchitelli, F. Becker, A. Vanlierde, J. Leblois, U. Schuler, F. Auer, A. Köck, L. Dale, J. Sölkner, O. Christophe, J. Hummel, A. Mensching, J. Fernández Pierna, H. Soyeurt, M. Calmels,

- R. Reding, M. Gelé, Y. Chen, N. Gengler, and F. Dehareng (2024). “Prediction of key milk biomarkers in dairy cows through milk mid-infrared spectra and international collaborations”. In: *Journal of Dairy Science* 107.3, pp. 1669–1684. ISSN: 00220302. doi: 10.3168/jds.2023-23843.
- Grelet, C., J. A. F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng (2017). “Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models”. In: *Journal of Dairy Science* 100.10, pp. 7910–7921. ISSN: 0022-0302. doi: 10.3168/jds.2017-12720.
- Grelet, C., A. Vanlierde, M. Hostens, L. Foldager, M. Salavati, K. L. Ingvarsten, M. Crowe, M. T. Sorensen, E. Froidmont, C. P. Ferris, C. Marchitelli, F. Becker, T. Larsen, F. Carter, and F. Dehareng (2019). “Potential of milk mid-IR spectra to predict metabolic status of cows through blood components and an innovative clustering approach”. In: *Animal* 13.3, pp. 649–658. ISSN: 1751-7311. doi: 10.1017/S1751731118001751.
- Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Steevens, N. Gengler, and F. Dehareng (2021). “Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions”. In: *Methods*. ISSN: 1046-2023. doi: 10.1016/j.ymeth.2020.07.012.
- Hagnestam, C., U. Emanuelson, and B. Berglund (2007). “Yield Losses Associated with Clinical Mastitis Occurring in Different Weeks of Lactation”. In: *Journal of Dairy Science* 90.5, pp. 2260–2270. ISSN: 0022-0302. doi: 10.3168/jds.2006-583.
- Hostens, M., S. Franceschini, M. van Leerdam, H. Yang, S. Pokharel, E. Liu, P. Niu, H. Zhang, S. Noor, K. Hermans, M. Salamone, and S. Sharma (2025). “The future of big data and artificial intelligence on dairy farms: A proposed dairy data ecosystem”. In: *JDS Communications*. ISSN: 2666-9102. doi: 10.3168/jdsc.2025-0843.
- Lactanet (2022). *PROFILab : Fact Sheet and Potential Solutions*. PDF fact sheet. Lactanet. URL: https://lactanet.ca/wp-content/uploads/2022/02/1.-MAJ_Aide-Memoire_ANG.pdf (visited on 01/23/2026).
- Leblanc, S. (2010). “Monitoring Metabolic Health of Dairy Cattle in the Transition Period”. In: *Journal of Reproduction and Development*. Journal of Reproduction and Development 56 (S), pp. 529–535. doi: 10.1262/jrd.1056S29.
- Liang, D., L. M. Arnold, C. J. Stowe, R. J. Harmon, and J. M. Bewley (2017). “Estimating US dairy clinical disease costs with a stochastic simulation model”. In: *Journal of Dairy Science* 100.2, pp. 1472–1486. ISSN: 0022-0302. doi: 10.3168/jds.2016-11565.
- Lin, D. and J. A. A. McArt (2025). “Prediction and classification of metritis and mastitis in Holstein cows using transition milk spectra under different modeling strategies”. In: *Journal of Dairy Science* 108.12, pp. 13752–13763. ISSN: 0022-0302. doi: 10.3168/jds.2024-26217.
- Liseune, A., D. V. den Poel, P. R. Hut, F. J. C. M. van Eerdenburg, and M. Hostens (2021). “Leveraging sequential information from multivariate behavioral sensor

- data to predict the moment of calving in dairy cattle using deep learning". In: *Computers and Electronics in Agriculture* 191, p. 106566. ISSN: 0168-1699. DOI: 10.1016/j.compag.2021.106566.
- Mennig, P. (2024). "A never-ending story of reforms: on the wicked nature of the Common Agricultural Policy". In: *npj Sustainable Agriculture* 2.1. Publisher: Nature Publishing Group, p. 20. ISSN: 2731-9202. DOI: 10.1038/s44264-024-00027-z.
- Mensching, A., J. Braunleder, E. Bohlson, S. Schierenbeck, and R. Reents (2025). "Daily standardization of routinely collected milk mid-infrared spectra from dairy herd improvement testing in a statistical framework". In: *Journal of Dairy Science* 108.7, pp. 7202–7223. ISSN: 0022-0302. DOI: 10.3168/jds.2024-25482.
- Mota, L. F. M., S. Pegolo, T. Baba, F. Peñagaricano, G. Morota, G. Bittante, and A. Cecchinato (2021). "Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data". In: *Journal of Dairy Science* 104.7, pp. 8107–8121. ISSN: 0022-0302. DOI: 10.3168/jds.2020-19861.
- Pralle, R. S., K. W. Weigel, and H. M. White (2018). "Predicting blood β -hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network". In: *Journal of Dairy Science* 101.5, pp. 4378–4387. ISSN: 0022-0302. DOI: 10.3168/jds.2017-14076.
- Pretty, J. (2007). "Agricultural sustainability: concepts, principles and evidence". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1491. Publisher: Royal Society, pp. 447–465. DOI: 10.1098/rstb.2007.2163.
- Puerto, M. A., E. Shepley, R. I. Cue, D. Warner, J. Dubuc, and E. Vasseur (2021). "The hidden cost of disease: I. Impact of the first incidence of mastitis on production and economic indicators of primiparous dairy cows". In: *Journal of Dairy Science* 104.7. Publisher: Elsevier, pp. 7932–7943. ISSN: 0022-0302. DOI: 10.3168/jds.2020-19584.
- Ramirez Mauricio, M. A., M. Berton, N. Amalfitano, D. Giannuzzi, S. Pegolo, S. Raniolo, M. Nocetti, R. Negrini, M. Coppa, B. Martin, S. Schiavon, L. Gallo, E. Sturaro, and A. Cecchinato (2025). "Leveraging milk mid-infrared spectroscopy to authenticate animal welfare, farming practices, and dairy systems of Parmigiano Reggiano cheese". In: *Journal of Dairy Science* 108.3, pp. 2642–2657. ISSN: 00220302. DOI: 10.3168/jds.2024-25466.
- Rasmussen, P., H. W. Barkema, P. P. Osei, J. Taylor, A. P. Shaw, B. Conrady, G. Chaters, V. Muñoz, D. C. Hall, O. O. Apenteng, J. Rushton, and P. R. Torgerson (2024). "Global losses due to dairy cattle diseases: A comorbidity-adjusted economic analysis". In: *Journal of Dairy Science* 107.9. Publisher: Elsevier, pp. 6945–6970. ISSN: 0022-0302. DOI: 10.3168/jds.2023-24626.
- Rienesl, L., N. Khayatdadeh, A. Köck, C. Egger-Danner, N. Gengler, C. Grelet, L. M. Dale, A. Werner, F.-J. Auer, J. Leblois, and J. Sölkner (2022a). "Prediction of Acute

- and Chronic Mastitis in Dairy Cows Based on Somatic Cell Score and Mid-Infrared Spectroscopy of Milk”. In: *Animals* 12.14. Publisher: Multidisciplinary Digital Publishing Institute, p. 1830. ISSN: 2076-2615. DOI: 10.3390/ani12141830.
- Rienesl, L., M. Marginter, P. Stückler, A. Köck, C. Egger-Danner, and J. Sölkner (2022b). “Use of differential somatic cell count, somatic cell score, and milk mid-infrared spectral analysis for monitoring mastitis in dairy cows during routine milk recording”. In: *Livestock Science* 264, p. 105050. ISSN: 1871-1413. DOI: 10.1016/j.livsci.2022.105050.
- Rinnan, Å., F. v. d. Berg, and S. B. Engelsen (2009). “Review of the most common pre-processing techniques for near-infrared spectra”. In: *TrAC Trends in Analytical Chemistry* 28.10, pp. 1201–1222. ISSN: 0165-9936. DOI: 10.1016/j.trac.2009.07.007.
- Rockström, J., W. Steffen, K. Noone, Å. Persson, F. S. Chapin, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sörlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. A. Foley (2009). “A safe operating space for humanity”. In: *Nature* 461.7263. Publisher: Nature Publishing Group, pp. 472–475. ISSN: 1476-4687. DOI: 10.1038/461472a.
- Salamone, M., I. Adriaens, A. Vervae, G. Opsomer, H. Atashi, V. Fievez, B. Aernouts, and M. Hostens (2022). “Prediction of first test day milk yield using historical records in dairy cows”. In: *animal* 16.11, p. 100658. ISSN: 1751-7311. DOI: 10.1016/j.animal.2022.100658.
- Shao, W., X. Luo, Z. Zhang, Z. Han, V. Chandrasekaran, V. Turzhitsky, V. Bali, A. R. Roberts, M. Metzger, J. Baker, C. La Rosa, J. Weaver, P. Dexter, and K. Huang (2022). “Application of unsupervised deep learning algorithms for identification of specific clusters of chronic cough patients from EMR data”. In: *BMC Bioinformatics* 23.3, p. 140. ISSN: 1471-2105. DOI: 10.1186/s12859-022-04680-4.
- Silva, F. G., C. Conceição, A. M. F. Pereira, J. L. Cerqueira, and S. R. Silva (2023). “Literature Review on Technological Applications to Monitor and Evaluate Calves’ Health and Welfare”. In: *Animals : an Open Access Journal from MDPI* 13.7, p. 1148. ISSN: 2076-2615. DOI: 10.3390/ani13071148.
- Simoni, A., F. König, K. Weimar, A. Hancock, C. Wunderlich, M. Klawitter, T. Breuer, M. Drillich, and M. Iwersen (2024). “Evaluation of sensor-based health monitoring in dairy cows: Exploiting rumination times for health alerts around parturition”. In: *Journal of Dairy Science* 107.8, pp. 6052–6064. ISSN: 0022-0302. DOI: 10.3168/jds.2023-24313.
- Sogandi, F. (2024). “Identifying diseases symptoms and general rules using supervised and unsupervised machine learning”. In: *Scientific Reports* 14.1. Publisher: Nature Publishing Group, p. 17956. ISSN: 2045-2322. DOI: 10.1038/s41598-024-69029-8.

- Soyeurt, H., Dardenne, P., Gillon, A., Croquet, C., V, erick, S., Mayeres, P., Bertozzi, C., Gengler, and N. (2006a). "Variation in fatty acid contents of milk and milk fat within and across breeds". In: *J. Dairy Sci.* 89 4858–4865. DOI: 10 . 3168/ jds . S0022-0302(06)72534-6.
- Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler (2006b). "Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry". In: *Journal of Dairy Science* 89.9, pp. 3690–3695. ISSN: 0022-0302. DOI: 10 . 3168/ jds . S0022-0302(06)72409-2.
- Soyeurt, H., C. Grelet, S. McParland, M. Calmels, M. Coffey, A. Tedde, P. Delhez, F. Dehareng, and N. Gengler (2020). "A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra". In: *Journal of Dairy Science* 103.12, pp. 11585–11596. ISSN: 0022-0302. DOI: 10 . 3168/ jds . 2020-18870.
- Soyeurt, H. (2023). "Fourier transform mid-infrared milk screening to improve milk production and processing". In: *JDS Communications* 4.2, pp. 61–64. ISSN: 2666-9102. DOI: 10 . 3168/ jds . 2022-0294.
- SPW (2025a). *État de l'Agriculture Wallonne*. Belgium: SPW ARNE. URL: <https://etat-agriculture.wallonie.be/home.html> (visited on 03/27/2025).
- SPW (2025b). *Rapport annuel de performance 2024 du Plan stratégique de la PAC 2023–2027 (RAP 2024)*. Namur, Belgique: Service public de Wallonie. URL: https://agriculture.wallonie.be/files/accueil/Politique/Plan%20strat%C3%A9gique%20PAC/RAP2024_valid%C3%A9_SFC2021.pdf.
- SPW (2025c). *Marge brute, excédent brut et revenus de l'exploitation spécialisée en bovins laitiers*. Etat de l'Agriculture Wallonne. URL: <https://etat-agriculture.wallonie.be/contents/indicatorsheets/EAW-D3c.html> (visited on 10/25/2025).
- Stygar, A. H., Y. Gómez, G. V. Berteselli, E. Dalla Costa, E. Canali, J. K. Niemi, P. Llonch, and M. Pastell (2021). "A Systematic Review on Commercially Available and Validated Sensor Technologies for Welfare Assessment of Dairy Cattle". In: *Frontiers in Veterinary Science* 8. Publisher: Frontiers. ISSN: 2297-1769. DOI: 10 . 3389/ fvets . 2021 . 634338.
- Suthar, V. S., J. Canelas-Raposo, A. Deniz, and W. Heuwieser (2013). "Prevalence of subclinical ketosis and relationships with postpartum diseases in European dairy cows". In: *Journal of Dairy Science* 96.5, pp. 2925–2938. ISSN: 0022-0302. DOI: 10 . 3168/ jds . 2012-6035.
- Tangorra, F. M., E. Buoio, A. Calcante, A. Bassi, and A. Costa (2024). "Internet of Things (IoT): Sensors Application in Dairy Cattle Farming". In: *Animals : an Open Access Journal from MDPI* 14.21, p. 3071. ISSN: 2076-2615. DOI: 10 . 3390/ ani 14 213071.
- Tedde, A., C. Grelet, P. N. Ho, J. E. Pryce, D. Hailemariam, Z. Wang, G. Plastow, N. Gengler, E. Froidmont, F. Dehareng, C. Bertozzi, M. A. Crowe, H. Soyeurt, and on behalf of the GpluSE Consortium (2021). "Multiple Country Approach to Improve

- the Test-Day Prediction of Dairy Cows' Dry Matter Intake". In: *Animals* 11.5. Publisher: Multidisciplinary Digital Publishing Institute, p. 1316. ISSN: 2076-2615. DOI: 10.3390/ani11051316.
- Tiplady, K. M., T. J. Lopdell, M. D. Littlejohn, and D. J. Garrick (2020). "The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle". In: *Journal of Animal Science and Biotechnology* 11.1, p. 39. ISSN: 2049-1891. DOI: 10.1186/s40104-020-00445-2.
- Tscharntke, T., A. M. Klein, A. Kruess, I. Steffan-Dewenter, and C. Thies (2005). "Landscape perspectives on agricultural intensification and biodiversity – ecosystem service management". In: *Ecology Letters* 8.8, pp. 857–874. ISSN: 1461-023X, 1461-0248. DOI: 10.1111/j.1461-0248.2005.00782.x.
- USDA (2025). *Milk Cost of Production Estimates | Economic Research Service*. URL: <https://www.ers.usda.gov/data-products/milk-cost-of-production-estimates> (visited on 10/25/2025).
- Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt (2021). "Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra". In: *Journal of the Science of Food and Agriculture* 101.8, pp. 3394–3403. ISSN: 1097-0010. DOI: 10.1002/jsfa.10969.
- Xu, W., A. T. Van Knegsel, J. J. Vervoort, R. M. Bruckmaier, R. J. Van Hoeij, B. Kemp, and E. Saccenti (2019). "Prediction of metabolic status of dairy cows in early lactation with on-farm cow data and machine learning algorithms". In: *Journal of Dairy Science* 102.11, pp. 10186–10201. ISSN: 00220302. DOI: 10.3168/jds.2018-15791.
- Zeineldin, M., A. A. Elolimy, P. R. K. Reddy, M. Abdelmegeid, M. Mellado, M. M. M. Y. Elghandour, and A. Z. M. Salem (2021). "On-Farm Point-of-Care Diagnostic Technologies for Monitoring Health, Welfare, and Performance in Livestock Production Systems". In: *Sustainable Agriculture Reviews 54: Animal Biotechnology for Livestock Production 1*. Ed. by V. K. Yata, A. K. Mohanty, and E. Lichtfouse. Cham: Springer International Publishing, pp. 209–232. ISBN: 978-3-030-76529-3. DOI: 10.1007/978-3-030-76529-3_7.
- Zhang, L., C. Li, F. Dehareng, C. Grelet, F. Colinet, N. Gengler, Y. Brostaux, and H. Soyeurt (2021). "Appropriate data quality checks improve the reliability of values predicted from milk mid-infrared spectra". In: *Animals* 11.2, pp. 1–17. DOI: 10.3390/ani11020533.

Unsupervised machine learning techniques in precision dairy farming

Adapted from: Nickmilder, C., Soyeurt, H., Alexakis, I. Przybylski, T., Gertz, M. and Franceschini, S. (2026). Unsupervised machine learning techniques in precision dairy farming. Invited book chapter in Advances in precision dairy and beef farming technologies (In press)

2.1. Chapter outline

The introduction presents the context of dairy farming in Wallonia and suggests health monitoring for early disease and subclinical disease detection. We hypothesize that using unsupervised learning methodologies on Fourier Transform mid-infrared (FT-MIR) predictions could overcome the limitations of current tools. However, the field of unsupervised learning encompasses many different types of algorithms designed for various purposes and with different computational complexities. The objective of this chapter is to introduce the concepts of unsupervised learning, different popular algorithms and their use cases in Precision Dairy Farming (PDF), and finally Knowledge Discovery in Database (KDD). These concepts are essential for understanding this thesis, as the proposed methodology could be defined as using unsupervised learning for KDD. This chapter was written later than Chapter 3, Chapter 4 and Chapter 5 so the algorithms presented were not considered at the time of those scientific papers.

2.2. Abstract

With the increasing volume and variety of data available in the dairy and beef industries, KDD has become increasingly important. KDD typically involves five key steps: data selection, data preprocessing, data transformation, data mining, and evaluation. Throughout this process, a wide range of machine learning methodologies, both supervised and unsupervised, can be applied. The main distinction between these approaches lies in the presence or absence of a target variable to explain or predict. Unsupervised learning aims to explore and structure datasets without a response variable, by identifying redundancies, patterns, or correlations among variables. This chapter reviews the main categories of unsupervised learning algorithms and their associated taxonomy, explaining how they work, when they should be used, and how they can be integrated with supervised methods within the broader KDD process. Finally, practical use cases in precision dairy farming are presented to illustrate their application.

Key words: unsupervised learning, knowledge discovery in databases, precision farming

2.3. Introduction

A modern livestock production system should have a low impact on its environment, have high welfare standards for animals, be economically profitable, and should produce sufficient quantities of high-quality food to meet consumer demand. To meet these conditions, animals and their environment should ideally be monitored continuously so that farmers can intervene quickly if necessary. Such a task may

have appeared impossible a decade ago, but advances in digital devices, their inter-connectivity and the overall digitalization of agriculture has enabled a whole new “digital ecosystem” which is now being used as a matter of course on many farms. Large amounts of digital data can be collected in the barn or in the field, remotely or from sensors on or even within animals themselves. Satellite images can be collected to assess field conditions; animal feed, milk and meat can be analysed using chemical and/or spectroscopic methods; while soil and meteorological conditions can be collected (e.g. using a meteorological station on the farm). Whilst there remain challenges in collecting accurate data using suitably robust and validated methods with standardised protocols (i.e. gold standard), there is now an increasing focus on how best to aggregate and analyse the huge and growing volume of data that is now becoming available to farmers and others in order to produce actionable insights that can be used to improve production systems.

The twentieth century saw major advances in statistics, together with incredible advances in computer science and computer performance. This enabled the emergence of a new branch of statistics, commonly called machine learning, defined as a field of computer science that studies algorithms and techniques used to solve complex problems (Rebala et al. 2019). Machine learning techniques have enabled automated searching for significant patterns in large volumes of data, in a way that was virtually impossible to achieve with formal statistical models such as the analysis of variance or linear modelling.

Machine learning models are part of a larger group of Artificial Intelligence (AI) models (Panesar 2019), defined as algorithms that train machines to be “intelligent” (Rebala et al. 2019). However, it is important to be careful of potential confusion in terminology. AI is often understood as equivalent to Large Language Models (LLM) such as ChatGPT, Claude or Le Chat. However, the whole spectrum of AI is broader than just LLMs which are often based on techniques such as neural networks. It is important to be aware that, despite the impressive achievements of deep neuronal networks and especially LLMs in recent years, they are not always suitable for different data analysis tasks.

Neuronal network-based methods are by definition (universal approximation theorem) “generalists” (Cybenko 1989). In contrast, more specific “specialist” Machine Learning (ML) methods (which rely on techniques such as cost functions and hyperparameterization) have the potential to undertake a specific data analysis task more accurately as well as more rapidly and with less computing power (and cost) required. Machine learning techniques provide a framework for iterative optimization along a gradient to minimize the cost function (i.e. a measure of how effectively a model is achieving the kind of target outputs a user is looking for) and maximize the model’s fit to the ground truth (i.e. how it matches to verified data that can be used as a benchmark to train machine learning models and validate results). In a similar

way to neuronal network-based methods, machine learning algorithms also learn an optimal set of weights to discriminate a ground truth within complex datasets but do so using specialized algorithms to avoid the computational overhead associated with neural networks.

There are a number of general data characteristics that may determine which ML technique may be most suitable. These include: the size of the dataset, the nature of data inputs (features) and their relation or the absence of responses to predict.

In terms of modelling, ML can be split into two distinct kinds of learning:

1. Supervised learning
2. Unsupervised learning

In supervised learning data contains examples of inputs (i.e. features) and outputs (i.e.labels): these training pairs help guide algorithms in inferring what desired output values to look for. There is a quantitative or qualitative response associated to each record allowing the model to be trained to estimate this target against specific quantitative and/or qualitative features.

In unsupervised learning the dataset has no response variable. Unsupervised learning is “a kind of machine learning strategy utilized for drawing inferences from the datasets containing input data without any labelled responses” (Chander et al. 2021). In contrast to supervised learning, which aims to optimise a model to accurately predict an outcome, unsupervised learning focuses on uncovering structure within a dataset without predefined labels. It simplifies data to highlight patterns or inherent features in the data in order to make it more analysable (e.g. in relating a particular feature to an outcome). A key approach is the optimization of a mathematically-derived quantity that captures an essential property of the underlying data. A simple example is the ratio of variance within and between groups created. More complex features may involve abstract representations in a latent space (i.e. a simplified representation of multi-dimensional data such as pixels in an image), which can be derived through advanced mathematical transformations and projections. UL is often considered as exploratory data analysis to present data in a form that can then be analysed more easily in subsequent steps (James et al. 2023), although its applications extend well beyond this. Indeed, some detected patterns may have immediate practical relevance.

This chapter will discuss these algorithms and their application within the scope of dairy science. As the chapter shows, supervised and unsupervised learning can also be inter-related. The first sections will present the taxonomy of unsupervised learning and the concepts behind specific algorithms. Guidelines to addressing difficulties encountered in unsupervised learning are explored. Finally, examples from animal production using UL are presented later in the chapter. The chapter aims to explain unsupervised learning in particular within machine learning and its relevance in

modern dairy farming. To achieve this, the chapter combined a traditional manual review process with the inclusion of AI research engines to help the authors find overlooked viewpoints as highlighted by Rane et al. (2024).

2.4. An unsupervised learning taxonomy

This section discusses a taxonomy of different approaches to unsupervised learning based on their individual computational approaches. The aim is to give the reader a general understanding and a simplified overview of the approaches available before starting a detailed description of the most important algorithms. However, unsupervised learning algorithms do not have one unique taxonomy. As discussed later, they can be categorized variously according to their objectives, their outputs or their computational approaches. The taxonomy presented here is inspired by Patel (2019) outlining two main types of unsupervised learning:

1. Clustering
2. Dimensionality reduction

The classification shown in Figure 2.1 is based on Ezugwu et al. (2022).

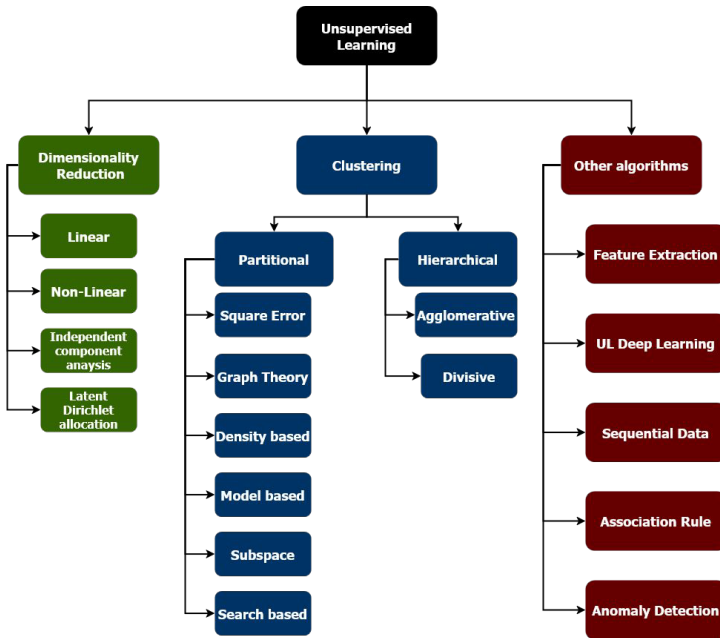


Figure 2.1: Unsupervised learning taxonomy based on Patel (2019) and Ezugwu et al. (2022).
UL=Unsupervised Learning

Though most approaches can be classified under the categories of clustering and dimensionality reduction, other approaches exist for specific objectives such as anomaly detection or feature extraction, or for specific types of data such as sequential data, or specific methodologies such as Deep Learning or the Association Rule. Whilst these are mentioned, the main focus is on clustering techniques. For each section a figure will be included to explain the concept with a theoretical example from milk recording. In this example, there are nine milk samples, three for each cow and each milk sample is analysed using FT-MIR spectrometry. Each milk sample contains around a thousand different absorbance values for different wavelengths.

2.5. Dimensionality reduction

This family of algorithms aims to reduce the dimensionality of a dataset while retaining the most relevant features and the intrinsic structure of the data. This approach is particularly useful for high dimensional datasets (i.e. data with a large number of variables or features compared to the number of observations). This is characteristic, for example, of data from spectroscopy. The reduction is achieved through projection of the data into lower-dimensional space, which suppresses random noise and non-essential information. This family of algorithms is further divided into four categories (Patel 2019):

1. linear
2. non-linear
3. Independent Components Analysis (ICA)
4. Latent Dirichlet Allocation (LDA)

Principal Component Analysis (PCA) is the most common linear approach (Figure 2.2). This approach uses Singular Value Decomposition (SVD) to simplify complex, high-dimensional datasets by creating smaller sets of variables called principal components. These are linear combinations of original data features optimized to represent the highest amount of data variability (Greenacre et al. 2022). Many variations of PCA are used to offset potential limitations of this approach, one of which is the requirement of continuous features as input. In this case, Correspondence Analysis (CA) allows for categorical features as input while using chi-square distances from a contingency table (Greenacre et al. 2022).

In some cases, a linear projection is not enough to capture the complexity of the dataset. A non-linear approach is then required to interrogate more complex data. Those methods, also called manifold learning, aim to capture those non-linear relations through different techniques (Meilă et al. 2024). As an example, isometric mapping, also called isomap, could represent three-dimensional features on a two-dimensional

plane; it computes nonlinear pairwise geodesic distances between observations before projecting them in a low-dimensional space (Izenman 2012). Among those methods, t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten et al. 2008) and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) are also worth mentioning. Both algorithms aim at reducing dimensionality but also retaining key local relationships between observations. UMAP has the advantage of being repeatable (McInnes et al. 2020). A non-linear variant of PCA is the Generalized Low Rank Model (GLRM) Method (Udell et al. 2016), extending PCA by approximating a data matrix as the product of two low-rank matrices under customizable loss functions and regularizers (Udell et al. 2016). All these techniques can be used to visualize and to identify underlying patterns in ‘messy’ data which is particularly the case with measuring phenomena in agricultural settings which are, by definition, particularly complex and dynamic environments.

The other two categories are less commonly used. ICA is often used as a preprocessing method to extract features from a blended signal. The general idea is to detect features that are uncorrelated and to consider those that are statistically independent (Stone 2002). LDA is a type of dimensionality reduction statistical model. The objective is to model count data as mixtures of topics. A typical use-case is identifying a specific research topic within a large text document (Blei et al. 2003). All those dimensionality reduction algorithms could be used for data visualization or as a first preprocessing step in a data analytics pipeline but other applications exist.

2.6. Clustering

Clustering approaches can be helpful when analysing large datasets where data is likely to fall into natural groups (i.e. groups of data with shared characteristics) related to specific underlying processes. Unsupervised learning can easily be confused with clustering because it is such a commonly-used technique. However, while clustering is a key part of unsupervised learning, other unsupervised learning categories exist and should not be overlooked. Clustering algorithms aim to divide a dataset into groups, commonly called clusters, optimized to explain dataset variability. Within clustering algorithms, a commonly categorisation is between (Jain et al. 1999; Pitafi et al. 2023):

1. hierarchical clustering
2. partitional clustering

Hierarchical clustering involves arranging groups with similar data points into a cluster hierarchy. Clusters are iteratively formed, either through merging from individual to population levels of records (i.e. agglomerative clustering also called bottom-up clustering) or dividing groups from the population level to the individual (i.e., divisive clustering also called top-down clustering) (Ezugwu et al. 2022).

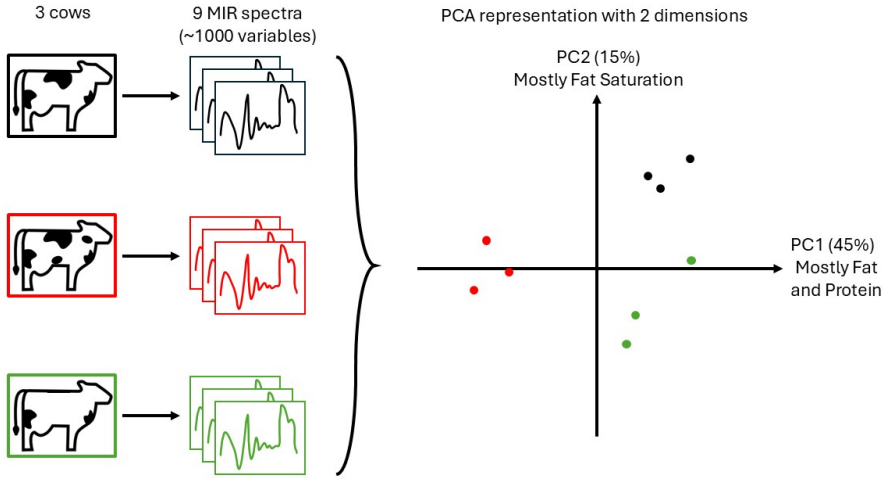


Figure 2.2: Concept of dimensionality reduction and its use for visualization. Nine high dimensional spectra are simplified into two principal components (i.e. linear combinations of the absorbances) representing already 60% of the variability of those spectra which is mostly the fat and protein content.

In contrast partitional clustering organizes groups without a hierarchical structure, potentially avoiding the intrinsic pitfall of distance computation required for hierarchical clustering (Ezugwu et al. 2022). Partitional strategies organize patterns into clusters (i.e. one partition), whilst hierarchical clustering consists of a sequence of nested data partitions (Fred et al. 2000). Figure 2.3 from Gao et al. (2023) illustrates this difference. The classification for this section comes from Ezugwu et al. (2022) with some slight changes to simplify it for non-specialist readers.

2.6.1. Hierarchical

Hierarchical clustering can be useful for certain types of data, particularly data with an inherently hierarchical structure. Algorithms in this category aim to establish a hierarchical structure among data based on a measure of dissimilarity or similarity between clusters (Ran et al. 2023). These algorithms are used to solve clustering problems where relations between observations are important. For example, Xu et al. (2020) used this approach to analyse social behaviour among 10 yearling cows, while Soyeurt et al. (2022) used this approach observe bulk tank milk samples over different months.

To group data, it is essential to define a method for measuring elements and their distances relative to one another, allowing the determination of group membership

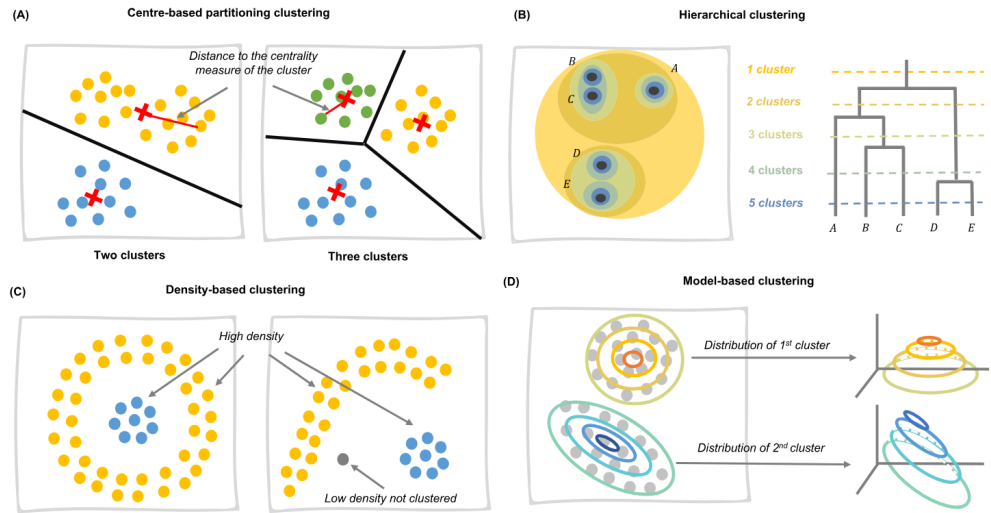


Figure 2.3: Different techniques of clustering and their concept. In our classification, (A), (C) and (D) are different techniques from the partition-based approach and (B) is hierarchical clustering. Figure from Gao et al. (2023).

(Murtagh et al. 2014). The initial version of hierarchical clustering consisted of grouping records iteratively based on similarity. In most algorithms, this similarity is a distance between observations or groups of observations.

As shown in Figure 2.4, two main approaches are often used (Murtagh et al. 2017; Patel 2019; Ezugwu et al. 2022; Gao et al. 2023):

1. Agglomerative hierarchical clustering
2. Divisive hierarchical clustering

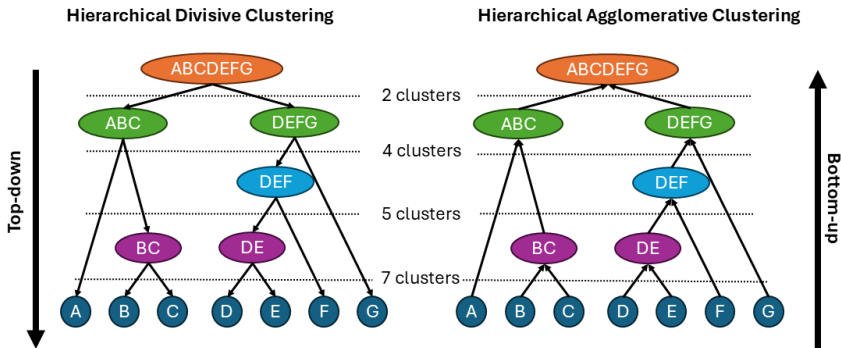


Figure 2.4: Divisive and agglomerative hierarchical clustering and their dendrograms

The first is the agglomerative or bottom-up approach where each observation is considered individually and linked with other observations (Figure 2.5). Linked similar observations are grouped to make new clusters. A dendrogram can also easily be produced to represent the relationships among clusters as output.

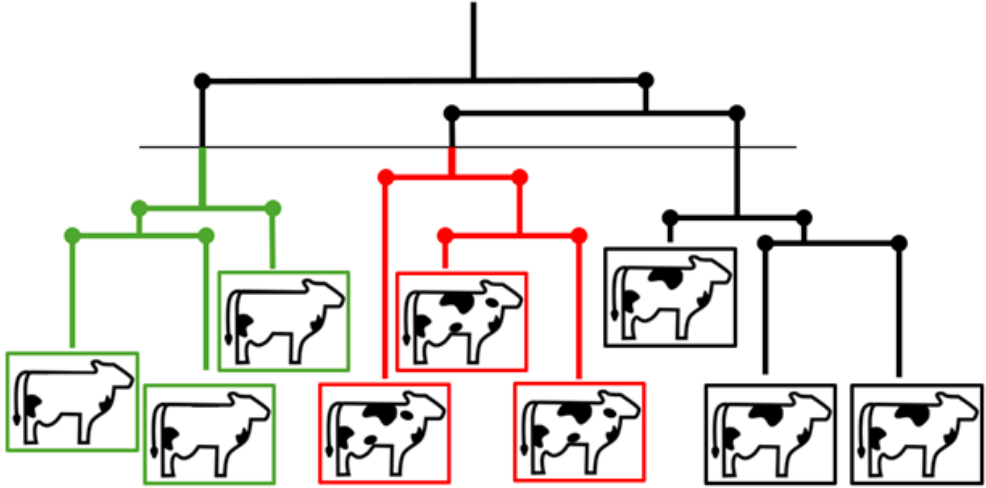


Figure 2.5: Agglomerative Hierarchical Clustering. The nine milk samples are grouped together based on their similarity. At each step, two samples or groups of samples are merged.

In this example, the dendrogram representing those merges is divided into three groups. Because the main difference between the nine samples was the different cows, it is most likely to observe three groups representing the different cows.

The agglomerative hierarchical clustering algorithm has two main hyperparameters to tune, one regarding similarity measurement and the other regarding the method to compute distance between observations and clusters, or between clusters (Murtagh et al. 2012; Murtagh et al. 2017). Many algorithms are based on hierarchical clustering with enhancements to overcome limitations in the original method, including Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) (Zhang et al. 1996) or Clustering Using Representatives (CURE) (Guha et al. 1998).

The second approach is the divisive or top-down approach, where all observations are considered as a single group and splits are made iteratively based on similarity until clusters of one observation are created (Chander et al. 2021; Mahdi et al. 2021). One such algorithm is Divisive Analysis clustering (DIANA) (Kaufman 2005).

The two hyperparameters underpinning such algorithms are similarity measurements and how these similarities are handled for groups of observations. To compute the (dis-)similarity between and within records/groups, some similarity metrics can be estimated. The most common is Euclidean distance but other distances exist such as

Manhattan distance or Gower distance. Table 2.1 gives some examples of similarities metrics found in the literature. These distances are often sensitive to the feature scale, hence the need for standardisation or using feature transformation into another reference space (or manifold).

For some similarities, the algorithm needs a solution to compute the distances between groups of observation as it is not simply the distance between two observations anymore. Multiple methods exist such as the single-linkage, the complete-linkage and the average-linkage. The single-linkage computes the distance between the two closest observations from two distinct groups while the complete-linkage computes the distances between the two farthest observations. Finally, the average-linkage computes the average distance between all observations.

Some advantage of these algorithms is that there is no need to specify the number of clusters in advance, the clustering results are visualized easily using a dendrogram, and a flat partition can be achieved at varying levels of granularity by making cuts at different heights in the dendrogram (Ran et al. 2023). The height at which the cut is made is then a reflexion of the variability represented. A certain percentage of variability within and between clusters helps set the global threshold for cluster determination. This might be expressed as a percentage of the variance explained or inertia measurement.

However, this kind of clustering is computationally demanding, especially due to computation of the distance matrices when performing using a large-scale dataset. Indeed, the memory required is the square of the number of observations, even though some smart implementations divide this requirement by a factor near to two by using triangular distance matrices instead of square ones. This has led to approaches using smaller datasets sampled from a global database. These include a divide and conquer approach (Franceschini et al. 2022) or an iterative clustering approach (Narita et al. 2018). The divide and conquer approach consist of drawing samples from the population (also called batches), attributing them to the nearest centroids, recompute the centroids of the clusters and repeating the sampling. Iterative clustering consists in considering batches one after another and integrating previous clusters centroids/representatives in the following run. As the amount of data is decreased, the processing time and power requirements both decrease too (Ran et al. 2023).

2.6.2. Partitional

Partitional clustering approaches can be preferable to hierarchical approaches for their computational speed, scalability and suitability for data with more clearly-identifiable clusters. Partition-based clustering algorithms divide observations into clusters, each called a data partition. Approaches are classified as hard-partitioning or soft-partitioning, depending on whether the output is a membership value or has

Table 2.1: Examples of similarities metrics.

Name	Formula	Explanations
Euclidean distance*	$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$	Length of the straight line between two points corresponding to the displacement along the Cartesian axes.
Manhattan distance*	$d = \sum_{i=1}^n x_i - y_i $	Sum of the absolute differences of the Cartesian coordinates.
Relative entropy	$d = \sum P(x) \log \frac{P(x)}{Q(x)}$	Mathematical expectation of information content of a random variable (Bonilla et al. 2021), can be used to get the amount of information detected (Shahinfar et al. 2014). It is the goodness of fit of a distribution respectively to another distribution.
Gower distance	$d = \frac{\sum ws}{\sum w}$ For binary, $s = 0$ or 1 $s = 1 - \frac{ x - y }{range}$ for continuous feature	Distance capable of dealing with mixed types of features. It can include feature weights (w). Depending on the implementation, the computation of the pair's similarities might differ.
Inertia*	$I = \frac{1}{n} \sum_{i=1}^n d(x_i, center)^2$	It corresponds to the sum of squared distance of points to their nearest cluster centre. Often adopted in k-means clustering to explain the variability covered by the clusters.

Similarities with a * are sensitive to feature scale.

the probability of belonging to a cluster. Multiples categories of algorithms exist, depending on the approach used to determine the group (Ezugwu et al. 2022). Some algorithms optimize a squared error, by trying to maximize the ratio between the inter- and intra-cluster variability. The most common squared error algorithm is the k-means algorithm shown in Figure 2.6 (Bock 2007). This algorithm divides a dataset by iteratively computing the average distance between the centre of a cluster (called a centroid) and the observations in the cluster, then updating the value of the centre and attributing new clusters to observations until it converges. The result is one partition of the dataset into k clusters. Other algorithms could use the density of the data to determine whether observations should be grouped together Figure 2.3. Those algorithms work similarly to algorithms for anomaly detection except they try to find groups instead of anomalies. Groups are formed in high density areas with cutoffs defined by low density areas. One such algorithm is Density-Based Spatial Clustering (DBSCAN) (Ester et al. 1996).

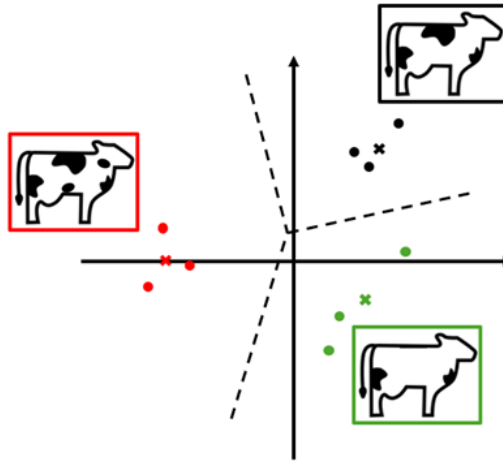


Figure 2.6: k-means clustering projected on a Principal Component Analysis. The nine samples have around 1000 variables. To avoid the curse of dimensionality, they are projected into Principal Component, then a k-means clustering is performed to find clusters that are optimizing the distances between the clusters (i.e. the coloured crosses) compared to the distances inside the cluster (i.e. the distances between the cross and the observation). This optimization is equivalent to find the minimum squared-error.

Not to be confused with density-based clustering, model-based clustering algorithms use probabilistic distributions of the data Figure 2.3. The method considers data as observations from underlying distributions and tries to find the most likely way of dividing the observations through Expectation Maximization (Fraley et al. 1998). These algorithms include the Gaussian Mixture Model (GMM).

Subspace clustering algorithms aim to avoid the ‘curse of dimensionality’ inherent

to high dimensionality datasets. This well-known problem in data analysis is where attempts to organise data in many dimensions (by feature or attribute) results in insufficient data in any dimension to produce meaningful results. It can also result in overfitting (i.e. fitting to irrelevant background ‘noise’ rather than any real underlying pattern). This is achieved by performing feature transformation or feature selection. These algorithms search for a subspace, among many possible combinations, where observations are easily divided into groups. Different algorithms have different ways to explore possible subspaces (Parsons et al. 2004). One such algorithm is CLIQUE (Agrawal et al. 1998).

Graph-theoretic clustering algorithms work with the same concept of grouping similar observations. Differences depend on the input used to perform clustering. In this approach, the input is a graph with observations called nodes and relationships between them called edges (Schaeffer 2007). A later section in this chapter includes an example of study from Bonora et al. (2018) using this methodology on time series data from an Automated Milking System (AMS) to study herd patterns. One example of a graph-based model, more specifically message-passing-based clustering, is Affinity Propagation (AP) (Frey et al. 2007) also discussed later in the chapter. Search-based clustering, also called automated clustering algorithms were created to solve the issue in which most clustering algorithms have hyperparameters to optimize and thus require optimization (Agarwal et al. 2011). These hyperparameters correspond to values used to control the learning process. However, optimization can be slow without a priori assumptions or larger volumes of data (Ezugwu 2020; Ezugwu et al. 2022). Those algorithms are often inspired by natural processes such as the Genetic Algorithm (GA) based on evolutionary processes.

2.6.3. Other categories

This section reviews the categories that we cannot classify directly in dimensionality reduction or clustering categories but that are worth mentioning. Such algorithms can be suitable for specific types of dataset in animal monitoring such as sequential data, data that present an inherent order between the records, as is typically the case with time series (e.g. tracks animal body temperature over time). Sequential data requires specific algorithms as most dimensionality reduction and clustering algorithms treat successive observations as independent rather than linked, thus losing a key aspect of such data. One example is hidden Markov Model, which is specifically designed to handle sequential data such as time series (Ghahramani 2004). At each step of the sequence, the algorithm will consider that there is a hidden state from which the observation comes from and then search for that state. At the next step, the observation could come from the same state or from another state related to the previous one by a probability to switch from one to another (Mor et al. 2021). As an example, Pastell et al. (2018) used this approach to estimate cow feeding time from

indoor monitoring data.

Anomaly detection algorithms aim to detect whether a new or following observation follows the same rules as other observations, whether based on a threshold, distribution, density, distance or dynamic pattern. Those algorithms are also called outlier, fraud or novelty detection methods. The detection of outliers is a very important aspect before starting the training of a model because their presence can impair global prediction performance. While the most basic detection method is the use of quantile distribution, there are numerous methods for global or local, parametric or non-parametric, and univariate or multivariate detection (Samariya et al. 2023). Most anomaly detection algorithms are based on a common methodology: identifying a profile for nominal (non-anomalies) data points, then attributing the anomaly tag to the data points that correspond the least to that profile. Anomalies are thus defined as data points that have unexpected characteristics, which implies defining what expected characteristics are based upon process knowledge. In dairy science and more precisely in FT-MIR spectrometry applied to milk analysis, a common method is Global Hotelling Distance also called Standardized Mahalanobis Distance which is a common multivariate approach in chemometrics (De Maesschalck et al. 2000; Zhang et al. 2021).

The 52 anomaly detection algorithms described by Samariya et al. (2023) include the Local Outlier Factor (LOF) and the Isolation Forest methods are discussed below. DBSCAN (mentioned earlier as a density-based algorithm), can also be useful for outlier detection. LOF is a density-based algorithm using the assumption that outliers are found in low density areas. The algorithm computes the density around an observation and compares it to the densities of the k -closest observation. If the ratio is significantly different, the observation will be considered as outlier (Alghushairy et al. 2021). The Isolation Forest works by isolating observations. The algorithm builds many random decision trees with observations divided based on randomly selected variables and thresholds. Since anomalies are typically rare and differ significantly from normal observations, they tend to be isolated faster, meaning they require fewer splits to separate them from the rest of the data.

Finally, there are the deep learning algorithm like Variational auto encoders (VAE) and Generative Adversarial Networks (GAN). Those algorithms can perform anomaly detection but also conventional clustering and can even handle sequential data when paired with architectures such as Long Short-Term Memory (LSTM). However, those rely on deep neural networks to divide the dataset. As neural networks are based on a completely different paradigm, they are not covered in this chapter.

The overall taxonomy used so far allows to characterize algorithms and models. However, this categorisation is quite theoretical. In many data processing pipelines, different algorithms are combined to bypass limitations and create more robust models. For instance, it is possible to achieve cluster- belonging probability from hard clus-

tering by adding a supervised layer, or a DBSCAN algorithm could be implemented with a hierarchical aspect to create a Hierarchical Density-Based Spatial Clustering (HDBSCAN)(McInnes et al. 2017).

2.7. Examples of unsupervised learning algorithms

2.7.1. Partitional clustering algorithms: k-means

As mentioned in the introduction, this chapter focuses on clustering algorithms, as most applications in the field are based on the detection of specific groups of records, e.g. detecting sick cows. We focus on some of the most utilized partitional clustering algorithms, as well as various emerging algorithms attracting particular interest. These are illustrated in Figure 2.7.

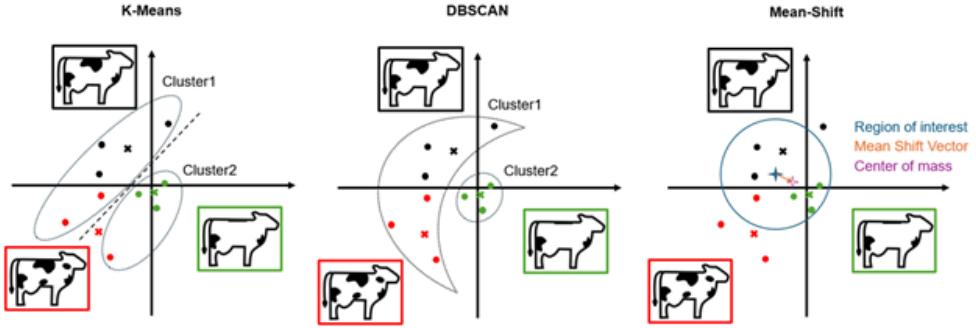


Figure 2.7: Illustration of k-means, Density Based Spatial Clustering of Applications with Noise (DBSCAN) and mean-shift.

Of the partitional clustering algorithms, the most used algorithm is k-means, originating from the work of Lloyd (1982) and (MacQueen 1967). Examples of use include detecting milk adulteration (Sowmya et al. 2021), poor metabolic adaptation syndrome (Tremblay et al. 2018) and oestrus detection (Shahriar et al. 2015).

This algorithm requires the number of clusters to be fixed before starting the clustering process and primarily uses inertia for similarity measurement. First, it is necessary to set a number k of clusters we want to study. k points are then chosen randomly from the dataset. In some cases, the initialization of the cluster centroids can be set arbitrarily based on original sample probability distributions. These are considered cluster centres. All points are then assigned to the nearest cluster centre. These point clouds are subsequently processed to extract a new centre of mass, i.e. the average position of the points along each axis. These centres are referred to as centroids. If extreme values are present in the training dataset, it may be preferable to use the median instead of the mean to compute the cluster centres. This is called k-medoids

instead of k-means (Oehm et al. 2022). This process of assigning points to the nearest cluster is repeated until the process stabilizes and the clusters move only within a tolerable margin.

The main advantage of this algorithm is its speed and lower memory requirements compared to hierarchical clustering, making it suitable for use with a large-scale dataset. In the worst-case scenario for classical k-means clustering, the computational complexity is $O(n^{(k+2)/p})$, with O being the complexity, n the number of records, k the number of clusters, and p the number of features (Arthur and Vassilvitskii, 2006). In practice, the complexity is more like $O(k \times n \times p \times T)$, where T is the number of iterations. The “ T ” highlights a drawback of this algorithm – it is overly sensitive to local minima as the algorithm will not explore alternative combinations, hence the need to repeat the process several times to ensure cluster stability. Since the number of clusters k is central to cluster identification, it should be closely aligned to the end goal. All the features entering the k-means create a hyper-space, often called a manifold. In this multi-dimensional space, clusters are assumed to be “regular and convex”, i.e. it might result in unexpected situations when data are not normally distributed along the features, exhibit non-homoscedasticity, or are uneven in group size within clusters, as the Euclidean distance is the most used and can become less reliable in highly dimensional spaces.

The major disadvantage of this algorithm is that the number of clusters must be defined in advance. As it might not be possible to predict the best number of clusters, this results in running the algorithm iteratively with a growing number of clusters, and then applying common optimum finding methods to limit the number of clusters.

Another approach to determine the optimal number of clusters involves graphical analysis of the cluster repartition as it is not possible to make a dendrogram. Whilst it is feasible to simply monitor the cluster belonging to the original features distribution, multivariate approaches using combinations of features/manifolding, as explained above in the dimension reduction section, make it easier to discriminate between clusters. For example, PCA can help identify feature repartition and influence (Tremblay et al. 2018). From a graphical perspective, the silhouette score is a commonly used criterion to determine the optimal cluster number. This can be defined as the ratio between the difference of the mean intra-cluster distance and the mean nearest-cluster distance, and the maximum of these two values. It is the mathematical equivalent to assessing on a graph if the clusters are well grouped together and do not overlap. It has been used by Ma et al. (2024), Rebuli et al. (2024) and Tremblay et al. (2018) to detect misclassifications. Advanced refinements of these approaches relate to a current topic of interest, the automation of determining the optimal number of clusters (Shi et al. 2020).

Based on the explained variability, the traditional determination of the optimal set of clusters in k-means is the “elbow curve”. In this approach, k-means clustering is run

iteratively on a varying predefined number of clusters and the percentage of variation explained is recorded. A graphical observation of the behaviour of the degree of variation explained against the number of clusters is then used to detect the point at which the slope changes abruptly – the “elbow” of the curve. However, identifying the elbow can be highly dependent on the observer’s perspective.

2.7.2. Partitional clustering algorithms: DBSCAN

Another example of partitional clustering is DBSCAN. This has been used by Michelena et al. (2025) to detect oestrus. This algorithm uses density-based clustering and is designed to cluster data of various shapes, even in the presence of outliers and noise, and from large datasets (Ester et al. 1996; Singh et al. 2022). One advantage it has over other traditional clustering methods is that DBSCAN can identify clusters of any shape, not just spherical or convex ones, as illustrated in Figure 2.7 (Singh et al. 2022). It effectively handles noise and outliers by identifying them as points that do not belong to any cluster, enhancing clustering accuracy in real-world noisy datasets. Unlike k-means, DBSCAN does not require the number of clusters to be predefined, which is useful when this information is unknown.

In DBSCAN, the density of each sample is defined by the number of neighbouring points within a specified distance, controlled by the epsilon (ϵ) hyperparameter. For clustering, DBSCAN classifies each sample into three categories:

1. Core points corresponding to samples with a minimum number of neighbouring points (determined by the MinPts hyperparameter)
2. border points related to samples with fewer neighbours than MinPts but still within the neighbourhoods of a core point
3. noise representing samples that are neither core nor border points

To achieve this classification, the algorithm first randomly selects a core point and assigns it to a new cluster. Then, all core and border points within its ϵ -neighbourhood are added to the same cluster. Next, for each new point added, the previous step is repeated to find all densely connected points. This continues until no more points can be added. If unassigned core points remain, the process starts over with a new cluster. Upon completion, the algorithm forms clusters and identifies outliers (noise), making DBSCAN useful not only for clustering but also for anomaly detection.

The performance of DBSCAN relies heavily on tuning its hyperparameters, ϵ and MinPts. For MinPts, various heuristic approaches have been proposed, indicating that a good value for this parameter is twice the dimension of the dataset. For ϵ , k-nearest neighbour distance plots are often used, where the “elbow” point indicates a suitable value (Michelena et al. 2025). A disadvantage of this algorithm is that it struggles when

there is varied density in a dataset and is strongly reliant on user-defined parameters, which can be a limitation (Singh et al. 2022).

2.7.3. Partitional clustering algorithms: Mean-shift

This algorithm is thought to originate from Fukunaga et al. (1975), whilst the clustering dedicated application of the algorithm is considered to originate from Comaniciu et al. (2002). Applied to image analysis, this algorithm is often related to image segmentation. In the context of dairy farming, it has been used by Yujuan et al. (2017) to monitor the rumination of cows.

According to the typology defined earlier, this class of clustering algorithms is partitional. Compared to k-means, it does not require the number of clusters to be defined before the fine-tuning and uses the combination of bandwidth and kernel for similarity measurement. This algorithm attempts to find the modes/peaks in the data distribution by iteratively recomputing the centres of gravity of the distribution. First, every record is considered as a potential cluster centre. Second, the algorithm defines a window/distance/radius/neighbourhood on the manifold to inspect. Third, for each record, the algorithm investigates the window around the data point and computes the centroid of all the points within this radius. Fourth, it shifts each record to this associated recomputed centroid. This implies the displacement of points to higher density regions, hence the name “mean-shift”. Finally, the process of defining a window until its association to a centroid is repeated until the algorithm converges, i.e. the displacement caused by the shifting of records is below a certain threshold. The local maxima highlighted with this method are the clusters. A post-processing filter might be needed to eliminate near duplicates centres to form the final set of centroids.

From a more theoretical point of view, the size of the neighbourhood to inspect is called the bandwidth (h). This is used within a kernel density estimation (KDE) function. This KDE function aims to reduce the impact of data points by giving more weight to points closer to the individual study points. Often, this KDE is a gaussian kernel. More information on kernel choice and its constraints can be found in Comaniciu et al. (2002).

This algorithm relies heavily on the metric of the feature space – therefore, it is preferable that all input features are converted into a Euclidean space in order to reduce constraints on the bandwidth parameters and allows the use of a consistent kernel size for all mean-shift procedures (Comaniciu et al. 2002). This algorithm uses a gradient approach for the displacement of the neighbourhood in question. As a result, it may be sensitive to regions that are not true local maximum likelihood but rather areas influenced by a common background noise. Another potential issue affecting convergence is the presence of “orphan” points that lack neighbouring data points. These may be ignored or flagged as outliers. Similar to the k-means algorithm, this

requires a parameter to be manually tuned – in this case the bandwidth to achieve optimal discrimination. Given the wide variation in KDE choice, the bandwidth, the method to identify cluster members and the potential sampling, defining the algorithm's complexity is not straightforward. Assuming the adoption of the scikit learn implementation, a python package for machine learning (Pedregosa et al. 2011), complexity is typically in the order of $O(T \times n^2)$ with T being the number of records and n the number of samples considered.

To accelerate cluster identification, it is possible to resample the input data with a regular grid and process these sub-spaces independently first. Afterwards, these sub-spaces can then be regrouped to reiterate and identify the global centres.

2.7.4. Partitional clustering algorithms: FASTCLUS

The so-called proc FASTCLUS procedure in the SAS® statistical software package is a lesser-known variant of the classical k-Means approach. The FASTCLUS procedure was inspired by the Hartigan (1975) leader algorithm and the MacQueen (1967) k-means algorithm (SAS, 2014, pdf. P. 2393). This algorithm was utilised by Alessio et al. (2021) to determine relationships in bulk tank milk data.

In this implementation, k-means is optimized for large datasets and speed. By default, FASTCLUS uses Euclidean distance, meaning that centers are based on least squared mean estimation, which is equivalent to k-means clustering. However, whereas k-means requires several iterations to converge, the FASTCLUS method provides good, if not necessarily optimal, clusters in as few as two or three iterations (SAS, 2014, pdf. P. 2394). This is largely due to its initialization method. The FASTCLUS procedure employs a method known as nearest centroid sorting (Anderberg 1973) to select initial cluster seeds in a way that promotes separation. This differs from traditional k-means, which often selects initial seeds randomly. However, this initialization method is sensitive towards outlier data, which can be beneficial if this method is to be used for outlier detection but may be disadvantageous if robustness against outliers is required. The procedure is designed to run well with datasets consisting of more than 100 observations, can handle missing values, and requires the variables to be standardized.

Overall, Proc FASTCLUS is fast and efficient but limited in flexibility, as is common with most SAS procedures. Its main advantage lies in its integration within the SAS ecosystem, ensuring well-tested procedures and seamless compatibility with other SAS procedures. On the downside, its sensitivity towards outliers must be considered, and SAS procedures tend not to incorporate the latest developments and methods in statistics.

2.7.5. Partitional clustering algorithms: Affinity Propagation

AP is an innovative clustering algorithm introduced by Frey et al. (2007) that operates by exchanging messages between data points to identify exemplars (i.e. representative points) and assign other points accordingly. It has been used to extrapolate whole crop lines from image-derived centroids by Li et al. (2024) and to detect subclinical mastitis in dairy sheep farms by Kiouvrekis et al. (2024).

Unlike k-means or k-medoids, Affinity Propagation does not require the number of clusters to be specified in advance. Instead, it takes as its input a similarity matrix between pairs of points, where each similarity $s(i,k)$ reflects how well point k represents point i . The core of the algorithm is based on the iterative exchange of two types of messages between data points – the responsibility message, $r(i,k)$, indicates how well-suited point k is to serve as the exemplar for point i , considering other potential exemplars; and the availability message, $a(i,k)$, reflects how appropriate it would be for point i to choose k as its exemplar, based on how much support k receives from other points. These messages are updated according to simple rules, and at each iteration, each point is assigned to the exemplar k that maximizes the sum of $r(i,k)+a(i,k)$.

Rather than imposing a fixed number of clusters, the affinity propagation uses a real-valued input $s(k,k)$ for each data point k , known as the preference. The higher this value, the more likely the point is to be chosen as an exemplar (i.e., cluster representative). The resulting number of clusters therefore depends on both the specified preferences and the message-passing process. If all data points are considered equally likely to become exemplars, a common preference value can be assigned. Adjusting this value allows control over the number of clusters – for example, using the median of the input similarities often yields a moderate number of clusters, while using the minimum tends to produce fewer clusters.

The algorithm offers several advantages, including not requiring a predefined number of clusters or their centres, relatively low computational complexity, consistent results across multiple runs, and a smaller amount of squared errors compared to other clustering methods.

2.7.6. Partitional clustering algorithms: Gaussian Mixture Model

GMM is a soft-clustering method whereby the data is fitted to belong to n Gaussian distributions (McLachlan et al. 2019). It has been used by Toan et al. (2019) to detect oestrus from a neck-mounted accelerometer. The term “soft” refers to the fact that each observation is not assigned to a single cluster outright, but instead has a probability of belonging to each distribution in the mixture. This approach builds on the Expectation-Maximization (EM) algorithm (Dempster et al. 1977), which iteratively refines the

model as follows.

In the Expectation step (E-step), the algorithm estimates the probability of membership for each data point in one of the n Gaussian distributions, represented by a latent (unobserved) class variable using Bayes Theorem. In the Maximization step (M-step), these estimated class memberships are used to update the parameters of the Gaussian distributions – such as their means, variances, and mixture weights – to maximize the overall likelihood of the data. The updates are weighted by the probabilities, meaning that each data point influences the parameter estimates in proportion to its likelihood of belonging to each component. By iteratively repeating the E-step and M-step, both the distribution parameters and the class membership estimates improve progressively, leading to convergence at a (local) maximum of the likelihood function. Observation can be with a probability of $p=60\%$ part of Cluster A but with a probability of $p=45\%$ also part of Cluster B, and so on. This identifies which data points are confidently attributed to a certain group, and which are in the grey zone where two or more predicted groups overlap (those points will have close highest probability values).

GMM is often used in clustering tasks when it is suspected that several groups are present in the data, but no assumption can be made on the exact boundary, either because it is unknown, or the structure of the data intrinsically exposes overlap (e.g. if animal behaviour is analysed). GMM has several advantages and is especially well-suited for clustering tasks where it is known that the data is approximately normally distributed among each group, since the Gaussian distribution is explicitly assumed in the model. Because it fits several distinct distributions with their own mean and covariance, it can deal with intertwining data groups or groups that have quite different shapes or distance between their respective data points. The noise within the data does not influence the result as much as in other “hard clustering” methods. The drawbacks of GMM are that each step of the EM requires the computation of the covariance matrix of each distribution. This can be time-consuming, especially in cases of large datasets or high dimensionality. It is also prone to overfitting, and while it resolves some issues found in k-means, other issues remain, such as sensitivity to initialization and the risk of reaching local optima during fitting. Overall, GMM is a more “realistic” approach than k-means as it models the probability distribution over all n Gaussians but may complicate interpretation of results if discreet clusters are desired.

2.7.7. Anomaly detection: Isolation Forest

The Isolation Forest assumes that anomalies are few and different and has been used by Michelena et al. (2025) to detect oestrus. This underlying assumption leads to a reflection on anomaly detection – why focus on the numerous nominal data points when the goal is to highlight a few unusual ones? Liu et al. (2008) developed a

simple algorithm aimed at isolating such points in the feature space, regardless of the nominal points. To do so, they introduced a building brick called the “Isolation Tree”, a binary tree that randomly partitions the data. At the root node, a feature is randomly selected, and a split value inside the data’s range is also chosen randomly. Left and right subtrees are then created, based on whether the data point’s value for the selected feature falls below or above the split value. The same algorithm is run recursively in the subtrees until a stopping criterion is met. Because anomalies are different, the path through the isolation tree from the root node to a leaf node is typically shorter for anomalies than for nominal nodes. In other words, identifying leaves at the end of short paths is a proxy to identifying anomalies. Due to the randomness involved in tree building, the process is repeated multiple times to obtain convergence for each record’s path length. This results in an Isolation Forest – an ensemble of isolation trees that collectively contribute to the mean path length.

Tweaks have been highlighted by Liu et al. (2008), demonstrating that the algorithm is capable of detecting outliers from a restricted dataset size, and since anomalies are defined by the shortest path, tree depth can be limited to decrease the computational load. These tweaks allow the method to achieve linear time complexity and memory usage proportional to the dataset size. Liu et al. (2008) created a comparison between their method and other commonly used anomaly detection algorithms, with a focus on large datasets. The reported advantages of Isolation Forest included fast execution, high Area Under the Curve (AUC) scores, and suitability for high dimensional data. However, the Isolation Forest works best alongside a feature selector to reduce dimensions.

Despite its efficiency and accuracy, Hariri et al. (2021) point out a limitation related to how the data is partitioned within Isolation Tree nodes, in terms of how the Isolation Forest determines anomaly scores in the whole space. In the initial method, the splitting was performed on one feature at a time, resulting in score regions aligned with the feature directions in the feature hyperspace. In a theoretical application to dairy cows, this would be akin to observing a biomarkers diagnosis, one biomarker at a time, rather than considering the animal’s holistic state, potentially biasing to the most common illness diagnosis.

To circumvent this issue, Hariri et al. (2021) proposed to extend the Isolation Forest by using a decision plane to split the data, allowing the borders of the regions to go in all directions in the hyperspace. Aside from this consideration, the algorithm maintains the same workflow. As a result, the accuracy of this new algorithm, deemed the Extended Isolation Forest, is slightly improved with minimal additional processing time and memory requirements.

2.8. Guidelines for validation and interpretation

2.8.1. Validation

Whilst unsupervised learning is advantageous in that it does not require a response variable in the optimisation process, it unfortunately makes the validation process more challenging once an experiment is completed. In contrast to supervised learning, in which evaluation typically involves comparing predicted outputs to known target values (i.e., ground truth), unsupervised learning relies solely on various metrics that reveal structures, patterns, or similarities within the data.

Because unsupervised learning methods rely solely on the intrinsic structure of the data, they do not provide direct explanations for the discovered patterns. While this lack of explicit explanation can be an issue as it is a black box, the fact that the model does not rely on a priori knowledge from the researcher is also a huge opportunity as it avoids confirmation bias. Outputs are driven entirely by the data structure.

If underlying causal information were available to the researcher, supervised learning would be more appropriate, particularly when the dataset is sufficiently large. Consequently, evaluating unsupervised learning outcomes is inherently more challenging, since ground truth labels are typically not available in the training data.

In addition, there is no universal agreement on key modelling decisions, such as the choice of distance metric or the number of clusters. While similar issues exist in supervised learning, they are typically less pronounced due to the availability of ground truth, which provides clearer guidance when selecting and evaluating algorithms. As a result, the suitability of the algorithm and design choices largely depends on the characteristics of the data and its inherent heterogeneity. It is therefore common practice to apply multiple methods and parameter settings to a dataset, followed by evaluation and validation (Gao et al. 2023).

A common evaluation strategy is to assess the relevance and interpretability of the discovered clusters. This can be approached by:

External validation by comparing clusters with external information available in another dataset, quantities similar to a ground truth (e.g., diagnostic measurements on a subset of records), or metrics such as the Rand Index or the Jaccard Coefficient.

Internal validation by examining the relationships between clusters and the original input features used to generate them, often relying on goodness-of-fit indicators.

Relative validation by criteria that assess which model or set of parameters produces the most meaningful or high-quality clusters. A variety of measures have been proposed based on cluster compactness (homogeneity within clusters),

separation (distance between clusters), representativeness (how well clusters reflect the underlying data distribution), connectedness (similarity between clustered points and their nearest neighbours), and stability (consistency of clustering results across different data subsets).

Combinations of these have also been suggested as evaluation criteria. Before interpreting the clustering results, however, it is essential to assess their reproducibility. In some cases, small perturbations in the underlying sample can lead to substantially different results, indicating low reliability.

Resampling and cross-validation are therefore integral to the validation process, for example, to choose the appropriate number of clusters, as many methods are sensitive to minor variations in the data and the choice of random seed, which increases the risk of overfitting (Gao et al. 2023).

Liu et al. (2022) highlight the concept of stability as a strategy for evaluating the performance and reproducibility of data clustering. The core idea involves generating perturbed datasets that remain close to the original, then applying the clustering process. If the clustering is stable, the original groupings will be retained across these perturbed versions. In practice, this means that if the clustering process is repeated using a different (but similar) or sub-sampled original dataset, similar or comparable clustering outcomes should be obtained. It is worth mentioning that using a part of the same dataset as validation, whilst technically an external validation, is not entirely independent and does not guarantee that it will work on any similar dataset. Numerous methods have been developed for cluster stability, including resampling for stability estimation and cluster validation via data splitting and subsampling, as well as alternative methods that diverge from these classic approaches. Ensuring consistency is a key indicator of the reliability and robustness of unsupervised learning results.

2.8.2. Interpretation

While unsupervised learning provides powerful and flexible tools for discovering hidden patterns within data, interpreting the results remains challenging. Nevertheless, it is possible to attempt an indirect inference of the potential causes behind the observed patterns by leveraging external knowledge or auxiliary datasets. The following ideas and procedures offer guidelines for interpretation and validation of results.

A useful starting point is the application of general descriptive statistics for cluster interpretation. For example, threshold comparisons grounded in scientific knowledge, or the distribution of data, can be informative. Aungier et al. (2015) used such comparisons relative to blood concentrations to characterise endocrine profiles. Another approach involves calculating correlations between features to assess their “proximity”. Most of the time, it implies a Pearson correlation (based purely on numerical

values). If the data are not normally distributed, a common alternative is the Spearman correlation, which is based on ranked values. Van Hoeij et al. (2019) applied this to analyse the link between plasma metabolites and hormones.

A more advanced approach is the Analysis of Variance (ANOVA), which can be used to compare least-square means of specific variables in-between clusters. Ho et al. (2019) used ANOVA to assess whether populations differed with respect to different measured traits and traits predicted from FT-MIR. However, this approach is sensitive to dataset size, performing well on sufficiently large datasets but potentially resulting in excessive statistical power, increasing the likelihood of detecting statistically significant yet practically irrelevant effects.

To solve the issue for large-scale datasets, an accessible way to interpret clusters is to estimate the Least Squares Means (LSM) of each cluster's variables using a simple linear model. By calculating the LSM of each variable within each cluster, one can compare, for example, concentrations of chemical compounds or metabolites within and in-between clusters. By then judging the observed (or not observed) average differences and absolute average concentrations, the scientist can apply their specialist knowledge to understand the causal relationships between the clusters (and therefore the patterns in the data). After clustering, LSM are calculated using the following formula:

$$LSM_j = \bar{Y} + \sum_{i=1}^k \beta_i X_{ij} \quad (2.1)$$

where LSM_j is the Least-Square Mean for cluster j

$\bar{Y}$ is the overall mean of the response variable

β_i are the estimated regression coefficients of variable i

X_{ij} represents the presence (1) or absence (0) of variable i in cluster j

As the aim of the LSM calculation is to compare concentrations and/or their differences, it is advisable in this context to remove the intercept from calculation of the linear model. Although this step may affect the statistical analysis, it greatly facilitates subsequent interpretation.

Interpreting results in high-dimensional spaces poses significant challenges, especially if results are to be presented graphically. To facilitate intuitive visualisation, it is often necessary to reduce the number of dimensions while preserving the interpretability and structure of the data. This process is closely related to dimensionality reduction, as discussed previously. Projecting the data onto feature axes or onto a latent space – such as that generated by PCA – can reveal meaningful patterns that are otherwise hidden in higher dimensions. For example, Brotzman et al. (2015)

visualised clusters by projecting them onto the first two principal components of PCA, thereby assessing both the separation of clusters and their interrelationships. Similarly, Multidimensional Scaling (MDS) was employed by Xu et al. (2020) to convert pairwise distances into a Cartesian coordinate system, facilitating intuitive cluster interpretation. As highlighted in the previous section, t-SNE can also reveal apparent clusters within a dataset. However, its use is not without controversy – studies such as Wattenberg et al. (2016) caution against a direct interpretation of the clusters, and Linderman et al. (2017) emphasize the need for careful analysis to avoid misrepresentation of the underlying structure. Despite these concerns, t-SNE can produce meaningful visualizations under certain conditions. A more recent alternative, UMAP retains the advantages of t-SNE while offering improved preservation of the global structure of the data. This enables a more accurate and holistic understanding of the dataset’s architecture, thereby enhancing the interpretability of the clustering results.

A newer methodology for interpretation and validation involves leveraging the capabilities of LLM. These models, which have demonstrated impressive performance across a wide range of natural language processing tasks, are increasingly being explored for their potential to support unsupervised learning objectives, including semantic clustering, feature extraction, and representation learning. Their ability to capture complex contextual relationships in high-dimensional textual data offers new opportunities for data interpretation and pattern discovery, particularly in domains where labelled data are scarce. While the expertise of scientists remains the go-to gold standard for evaluation of scientific results, some results can be so abstract that interpretation becomes an arduous task. In addition, expert interpretation often requires years of experience and personal insight, which may only be applicable to a limited number of specialists in a respective field of science. Both issues can be eased by using LLMs. Since modern LLMs are trained on incredibly large bodies of data, including scientific papers, they can identify patterns related to real-world experiments. Although they remain, to an extent, a “black box”, particularly when it comes to questions such as, “How does a LLM store its information?” and “Where does the reasoning originate?” (Prince 2024), factual evidence shows their incredible capabilities, even in terms of “understanding” quite complex and abstract tasks.

To enable LLMs to reason about outputs from unsupervised learning, it is necessary to translate cluster patterns into a structured format that they can interpret. One effective approach is to use the LSM method, which facilitates the transformation of identified clusters into a form that resembles the types of data found within the pre-training corpus of most LLMs. The rationale for using LSM in this context is clear. In line with good scientific practices, results are often summarized in tables presenting descriptive statistics such as means and standard deviations for the variables of interest. These forms of representation are not only human but also machine-readable, particularly by LLMs trained to process and analyse both natural language and quantitative data. As such, LLMs are well-suited to assist in the interpretation of

LSM-derived summaries and in deriving domain-specific insights.

The procedure for querying an LLM using unsupervised learning outputs involves three essential steps. The first and most crucial is to define a clear and precise context for the task. This includes specifying the scientific domain (e.g., animal science, agronomy, dairy production), detailing the experimental setup, including variables measured, timeframes, and population characteristics, and providing any contextual information that may inform the interpretation of patterns. This step is fundamental to limiting the degrees of freedom available to the LLM, thereby enhancing the relevance, accuracy, and domain alignment of its outputs. Constraining the model's scope improves both the reliability and interpretability of the resulting explanations. Next, a clear and explicit task must be posed to the LLM. An example of an effective prompt is: "Please interpret the anonymized results of our experiment and explain what the findings mean for a dairy farmer. In addition, suggest practical actions the farmer could take to address potentially unfavourable situations." This step helps to ensure that the LLM provides contextualized, actionable, and domain-specific interpretations rather than generic summaries. This would benefit from LLMs trained for animal science, such as the model that Liu et al. (2025) developed. By articulating a specific intent, this prompt structure improves the model's capacity to generate meaningful reasoning and recommendations. Finally, the results should be presented in a concise and structured format. This may involve a simple table that includes the list of variables, the respective cluster identifiers, and the corresponding statistical estimates, such as mean, standard deviation, or other statistical descriptors. By following these three steps (defining the framework, formulating a targeted query, and structuring the data appropriately), researchers can effectively harness LLMs to derive interpretable and actionable insights from unsupervised learning outputs.

2.9. Combining unsupervised learning and KDD

We have already identified the different categories of algorithms that can be used to develop unsupervised learning. It is important to mention where these algorithms fit in a classical data pipeline, also known as KDD. This concept can be divided into five essential steps Figure 2.8 starting with "step 0", which consists of precisely defining the objective of the study. Success is not only linked to the choice of algorithm but also to many decisions made before and after the data mining process. Figure 2.8 illustrates examples of operations available at each KDD step. This section reviews those steps and emphasises the use of UL algorithms, enabling discussion of certain aspects of data analysis not previously covered. In particular, the combination of supervised and unsupervised learning opens up many new data processing approaches. This section aims to show how unsupervised algorithms are integrated into the process and how they complement traditional supervised learning methods.

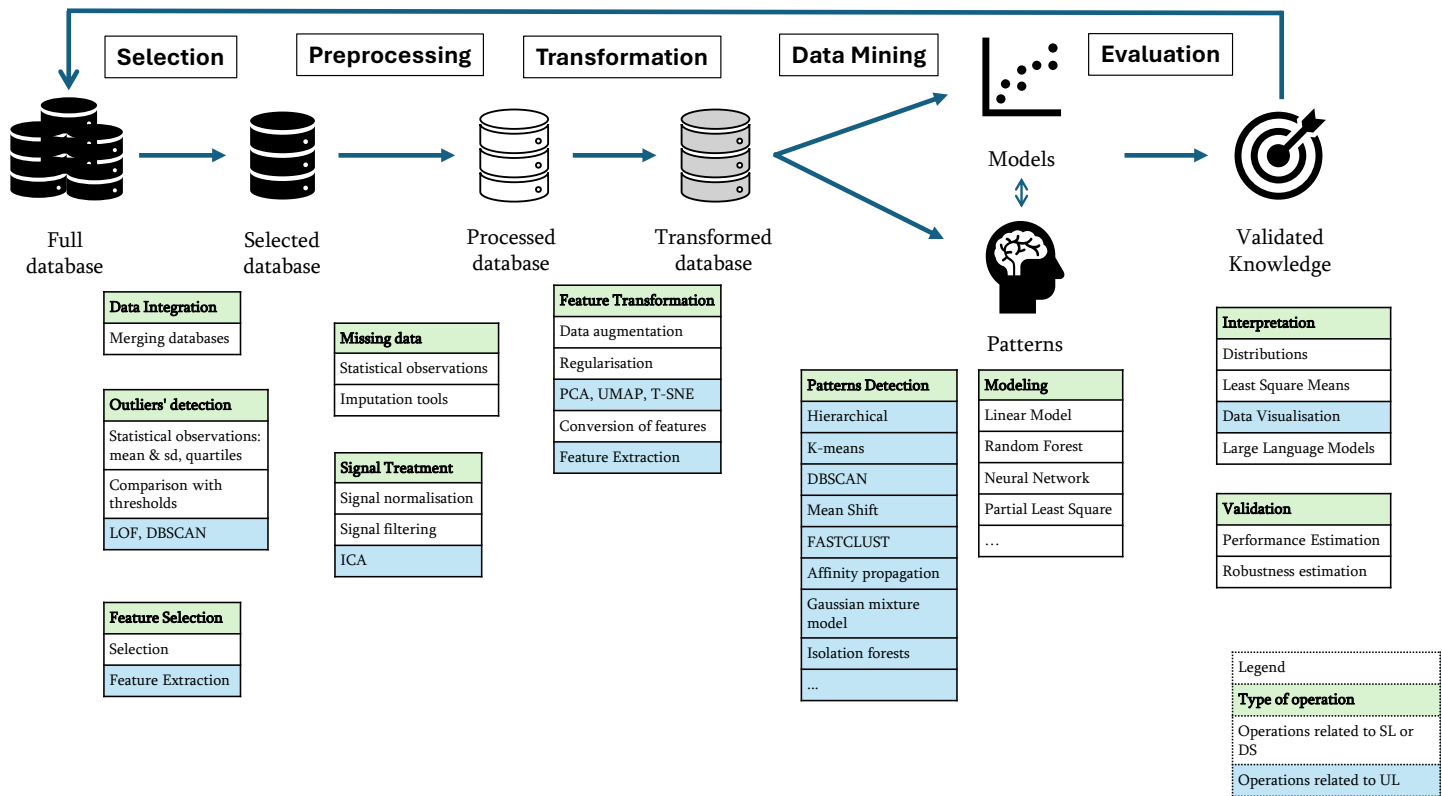


Figure 2.8: Knowledge Discovery in Database pipelines with examples of operations available at each step.

2.9.1. Selection

The first step is selection. At this stage, the priority is to integrate data from different sources and databases. Machine learning models often treat datasets as tabular, with each record occupying one line, and each column representing the monitored features. For more complex datasets, some algorithms relax this concept of “one-line-equals-one-record” by replacing the “line” with a processing stream, e.g. photos in image processing.

The next task is selecting the variables (or derived features) necessary for the study. This requires a precise inventory of the available variables. Some variables may be unusable due to poor reliability or a high proportion of missing data. At this stage, domain knowledge or consultation with experts is crucial to identify the most relevant variables. Feature selection and extraction algorithms can also be useful.

Finally, outlier detection is essential to ensure that algorithms operate “within specifications”. Outliers are values that are extremely high or low relative to the expectation from the underlying distribution and may therefore be irrelevant to the study.

These extreme values can cause several issues in the modelling and/or interpretation of results. Many algorithms – particularly those based on the Gaussian distribution – are sensitive to extreme or outlying values. While this sensitivity is not a weakness itself (as it often allows for the detection of subtle data structures and can in fact improve model fit), it becomes problematic when the data contains a substantial number of outliers. In such cases, results may be misleading or less accurate, as the presence of outliers violates the underlying statistical assumptions.

This highlights a fundamental tension between algorithmic strengths and vulnerabilities in terms of the introduction of errors. Features that make an algorithm powerful can also render it fragile, depending on the context. Finding an optimal combination of data quality/properties and matching algorithm specifications is therefore essential to minimise overall errors by accounting for such trade-offs. However, formalising such a selection process is methodologically challenging since it is hard to see which datapoints are outliers and why, and if the resulting interactions are harmful. The choice of strategy frequently relies on the statistician’s intuition and experience.

In general, outlier detection methods often employ statistical indicators such as quantiles of a distribution or fixed thresholds of “good” data points. UL anomaly detection algorithms, such as LOF (Satola et al. 2024) or DBSCAN (Shine et al. 2018), can also prove effective.

2.9.2. Preprocessing and transformation

These KDD steps are related but differ in their objectives. Preprocessing focuses on cleaning the dataset of its potential inconsistencies, while transformation optimizes

the dataset for the data mining stage. Commonly available preprocessing operations include noise treatment (particularly in signal treatment) and handling missing data. UL is less commonly used here, although an algorithm such as ICA can extract a specific signal from multiple merged signals. For missing data, several strategies can be carried out, such as deleting lines where data is missing, replacing the missing data with a variable such as the mean, median, or most represented class, or by using any other imputation method such as k-Nearest Neighbour.

The transformation step is related to feature type conversion, scaling or the normalization of the features, and dimensionality reduction. Some algorithms cannot handle categorical features, requiring conversion into numerical features. Vanlierde et al. (2021) used the conversion of categorical features (breed) into dummy features. In other cases, the opposite process is more beneficial, i.e. converting numerical into categorical features. Grelet et al. (2024) highlighted that, for some key biomarkers predicted from milk, quantitative performances are disappointing although trends appear to be modelled. For biomarkers that display highly skewed distribution, it can be necessary to transform quantitative measures into levels and perform discriminant analysis on these categories. However, as mentioned, this is only relevant for management purposes that require levels instead of precise values.

For feature transformation, it is common to observe features that are not uniformly or normally distributed along the range of values. A way to bypass the associated modelling complexity is to transform the features. Ghaffari et al. (2020) used log10 transformation of blood metabolite concentration, followed by a Pareto scaling for heteroscedasticity correction through skewness reduction. It is also possible to use a natural logarithm to transform input feature distribution into a normal distribution. A typical example used in milking studies is Somatic Cells Count (SCC). As this is a marker of inflammatory response, it often explodes and thus there is a need to rationalize it, with for example the application of a log transform, or the computation of a Somatic Cells Score (SCS), as demonstrated by Satoła et al. (2024).

Dimensionality reduction is a final key objective, and UL is relevant with feature extraction algorithms and dimensionality reduction algorithms. In tabular datasets, features might be interpreted as dimensions of a multi-dimensional space. As some algorithms are less efficient in highly dimensional datasets – known as “the curse of dimensionality” – aggregating the dimensions is beneficial. This curse might be defined as the likelihood of finding observations alone in single sub-spaces of a manifold, categorizing the lonely observation as the “typical example” for that part of the hyper-space, and thus training the algorithm to be too specific. Some methods commonly seen in the literature include:

PCA this consists of linearly combining the features to create a new set of orthogonal coordinates, capturing the maximum variation in the data. Brotzman et al. (2015) used it for feature reduction before performing clustering, whilst Todde et al.

(2016) used it to reduce the feature sets before k-means clustering.

MDS this consists of arranging points in a space based on pairwise similarities or dissimilarities of entities. The rank order of distances should match the rank order of dissimilarities. It is similar to PCA but adapted to also account for categorical data. In practice, it does not scale well on large datasets and therefore its use is restricted to simpler tasks.

t-SNE a non-linear dimensionality reduction technique. It is poor at conserving the global structure of the data and should therefore be used only for tasks that focus on local characteristics of the dataset.

UMAP this is a variation of t-SNE, based with constraints on the original features (McInnes et al. 2020). Simankov et al. (2025) define it as an algorithm aiming to find a low-dimensional representation of a dataset while preserving its global structure by first constructing a high-dimensional graph and then, through embedding optimisation, creating a lower dimensional graph that maintains the pairwise similarities.

VAE based on neural networks, these are another way to aggregate feature information and are fed the original information as the information to process and the information modelled. This mirroring effect is enabled through multiple layers and can be seen as a symmetrical hourglass, with one end being the original data, and the other a “perfect” reproduction of these original data, with funnels going both ways. The tightest point (the bottleneck) is where all the information can be modelled and summarized. After training it for perfect reproduction, the enlarging funnel is removed from the model and the predictions created by the “bottleneck” are considered to be the new summarized record score in a new latent space.

2.9.3. Data mining: getting the best of the two approaches

While the other steps are essential for building a relevant model, this is the step in which knowledge is extracted from the database. Although the data mining step is itself a full process inside the KDD, it can be summarized in two parts. The first part is modelling, i.e. supervised learning, where models are used either for better understanding of a process, or to make predictions. Many supervised algorithms are available and information about the structure of the dataset or the response variable is required to choose the most appropriate one. This is covered in the section on supervised learning. The second part is pattern recognition, i.e. unsupervised learning, which primarily consists of different clustering approaches. These approaches have been covered in a previous section and will not be discussed here. However, this subsection highlights an aspect that is often overlooked – although supervised

learning and unsupervised learning are carried out separately, they can be combined sequentially to strengthen the general approach. For example, in a large database, patterns may be identified but not validated. These patterns can then be predicted using a supervised model and applied to a new database where reference values related to the interpreted patterns are available. Combining unsupervised and supervised learning is a powerful approach, especially useful given the increasing volume of data generated today. Conversely, moving from supervised to unsupervised approaches can reveal significant differences in prediction performance in a model created by underlying groups not identified beforehand. These combinations, often referred to as stacking, are presented in the following case studies. The final step of KDD, evaluation, has been discussed in the section on unsupervised learning and will not be further developed here. The evaluation of supervised models, often referred to as validation, is covered in the chapter on supervised modelling.

2.10. Case studies in dairy farming

Having explored the different methodologies, this section illustrates concrete examples of unsupervised learning applications in dairy farming. Although this chapter focuses primarily on UL, very few studies rely exclusively on unsupervised learning algorithms (denoted by a “U” before the example). It is widespread practice to include supervised algorithms (examples beginning with “U+S”) within processing pipelines. Because unsupervised learning does not rely on reference outputs, there are no accuracy/transferability measurement metrics to report for studies using only unsupervised learning. In contrast such metrics can be reported for U+S studies.

In terms of stacking, the most common approach consists of first detecting patterns using unsupervised learning and then developing a supervised model to infer these patterns in new datasets, enabling the possibility of near-real time application. For example, Xu et al. (2019) identified the metabolic status of Holstein-Friesian cows using k-means on biomarkers such as free fatty acids, glucose, β -hydroxybutyrate, insulin and insulin growth factor 1, which are costly to measure. They then predicted these statuses using supervised machine learning algorithms on cheaper on-farm data such as dry period length, parity and milk production traits. In this study, the output was a classification, but it could also be expressed as the probability of belonging to a specific cluster.

Another use case for stacking unsupervised and supervised learning is the interpretation of complex processes. In such cases, clustering is not the final objective but rather a means of understanding the underlying process. For example, Sowmya et al. (2021) combined supervised and unsupervised learning to detect milk adulteration using spectrometry. K-means were used to identify emerging groups and assess whether they were associated with adulteration, before predicting the adulteration using supervised machine learning. Beyond group identification, a predictive super-

vised algorithm enables transferability to new records, as shown by (Soyeurt et al. 2022), where Partial Least Square Discriminant Analysis (PLS-DA) was applied after clustering to infer cluster membership in new populations. Stacking is not limited to unsupervised algorithms followed by supervised applications. It can be mixed for pre-processing, using a multistep approach. Foldager et al. (2020), for example, used random forest algorithms to predict blood energy-related bioindicators (such as Non Esterified Fatty Acids (NEFA), cholesterol, etc.) from FT-MIR milk analysis in dairy cows. Based on these predicted values, energy balance clusters were created using a k-means algorithm and subsequently used to predict physiological imbalances.

The objective of this section is to present case studies in which the methodologies described above were applied to solve challenges in Precision Livestock Farming (PLF).

2.10.1. Biomarker focused studies

“U+S” Franceschini et al. (2022) utilized 740,454 milk recording events from 114,536 primiparous Holstein cows to detect patterns in FT-MIR predicted traits, including body weight, minerals, β -hydroxybutyrate (BHB), citrate and acetone. Pattern detection was carried out using Ward hierarchical clustering, before applying PLS-DA to predict the identified clusters. One of these clusters was associated with metabolic disorders, demonstrating that FT-MIR spectra can be used as an individual monitoring tool. A similar methodology was applied to 774,781 bulk tank milk fatty acid profiles to detect potential herd-level issues, benefiting from a higher data frequency (Franceschini et al. 2025). Although no labels for metabolic disorders were available, these studies demonstrate that strong signals exist in milk spectra. However, these studies are limited by the lack of direct feedback from farmers.

“U+S” Tremblay et al. (2018) acquired 790 observations from 312 cows between May 2015 and February 2016 to detect poor metabolic adaptation syndrome. They used blood BHB, milk fat to protein ratio, and blood NEFA (although additional features, such as FT-MIR spectra, were available), and performed a PCA on these features and then performed a k-means cluster analysis on the first two PC. Five clusters appeared, distinguishing different levels of poor metabolic adaptation syndrome, while additional features were used for cluster interpretation. Longer-term perspectives included the use of these clusters to develop predictive models for poor metabolic adaptation syndrome and to improve individual and herd metabolic health management. Regarding the validation strategy, they did not partition data beyond dataset filtering.

2.10.2. Health disorder-focused studies

“U+S” Ghaffari et al. (2020) primarily used supervised classifying algorithms, sequential minimal optimization, Random Forest (RF), alternating decision tree, and naïve Bayes-updatable, to detect over-conditioning in dairy cows around calving, using blood metabolomic profiles. All algorithms achieved accuracies over 70%. The

unsupervised component of the study relates to the clustering of the variables (i.e. the serum metabolites) using Ward Linkage for the heatmaps. Although the study applied mostly supervised learning, it illustrates that unsupervised learning can be used in very specific use-cases to better understand the relation between metabolites.

“U+S” Van Hoeij et al. (2019) used SAS PROC Cluster, PROC CORR, and PROC TREE to investigate the relationship between metabolic status and behaviour in 81 postpartum Holstein-Friesian cows, using computerized feeders and accelerometers. The study consisted of a Ward clustering using blood metabolite concentration (e.g. free fatty acids, BHB, glucose, insulin-like growth factor-1) to define metabolic status; a correlation analysis between plasma metabolites and the clusters to validate those clusters; and the creation of a Generalized Linear Model (GLM) to assess the plasma concentration of metabolites or behaviour based on integration into one of the metabolic status clusters, the dry period length, and the parity (capped at three). The study highlighted differences in behaviour between cows with better metabolic status and those with poorer status.

“U+S” Oehm et al. (2022) analysed 606 farms in Germany to detect parasites in dairy cows, using UL to avoid introducing bias. Based on a stratified sampling of farms across three regions and herd sizes occurring between 2016 and 2019, they performed clustering based on production traits in addition to antibody status using partitioning around medoids. Clustering was carried out separately for each of the study regions and each of the parasites. They successfully identified bulk tank milk contaminated by at least one of two parasites (*F. hepatica* & *O. ostertagi*). Random forest was then used to assess variable importance and better understand the underlying process.

“U+S” To study hoof disorders in dairy cattle in Korea, Espiritu et al. (2025) collected data from 877 cattle across 15 farms. The 15 farms were clustered based on farm management practices, distinguishing two groups – one with higher disease prevalence and poor conditions, and the other with lower prevalence and better management. The study utilized Gower’s Distance and Ward’s linkage method. They then computed the correlation between the different management practices and the hoof diseases to validate the hypothesis, demonstrating that floor elevation transitions, aggressive trimming, shallow bedding, and wet flooring strongly correlated with hoof disorders.

“U+S” Kiouvrekis et al. (2024) carried out detection of subclinical mastitis in dairy sheep (2,220 ewes across 110 farms) in Greece. They associated the prevalence of subclinical mastitis in farms with management, temporal and meteorological data. Assuming a bi-modality regarding the affected and non-affected populations, they explored the application of Affinity propagation, birch, hierarchical, k-means and spectral clustering. On top of this detection, they also explored supervised algorithms such as decision trees and k-nearest neighbours. The study highlighted that farmers were not always aware of the health status of their sheep, with the ultimate goal of improving awareness and increasing farmers’ attention to subclinical mastitis.

2.10.3. Detection of oestrus

“U” Shahriar et al. (2015) used Fast Fourier Transform and k-means clustering to analyse accelerometer data acquired from the collars of 19 cows between October and December 2013, aiming to predict heat events in pasture-based dairy cows. Regarding the accelerometer data, mobility was defined as the sum of absolute accelerations along each axis, while movement intensity was defined as the square root of the sum of squared accelerations. Longer-term perspectives included the refinement of oestrus detection methods, particularly by reducing false positives to enable farm management deployment. Regarding the validation strategy, they did not partition data beyond the dataset filtering.

“U” Toan et al. (2019) combined Internet of Things (IoT) technologies and machine learning with data acquired from a low-frequency neck-mounted accelerometer to detect oestrus. They applied unsupervised learning alongside a novel heuristic post-processing algorithm to address inaccuracies caused by the low-frequency. A GMM was trained on synthetic data, improving precision from 0.28 to 0.94, compared to a traditional threshold model. They then applied the model to seven cows as a proof of concept, noting that further labelled data are required.

“U” Michelena et al. (2025) applied three unsupervised anomaly detection techniques – DBSCAN, Isolation Forest and Local outlier factor – to sensor data for oestrus detection in dairy cattle. The data consisted of five behaviours – resting, rumination, grazing, eating and playing. Using 22 Holstein-Friesian cows from two dairy farms, the study found that cow-level models achieved F1-Scores above 90%, while herd-level models achieved 62.22% and 75.27%. These results suggest that individualized models are essential for improving performance in anomaly detection.

2.10.4. Cow behaviour

“U” McVey et al. (2020) focused on animal behaviour monitoring using ear-tag accelerometers and milking order. The study analysed data over a six-month period for 200 Holstein cows. Based solely on milking order, three main groups emerged – cows at the beginning of the queue, those at the end, and those moving around the middle. The findings were consistent with the ear-tag data, in terms of segmentation of the population into three sub-groups. “U+S” Yujuan et al. (2017) monitored cow rumination during July 2013 using the application of mean-shift to 65,400 dairy cow images, combined with jaw tracking. However, this rumination monitoring approach was hindered by behaviours such as raising or turning of the cows’ heads.

2.10.5. Livestock monitoring: herd level

“U” Bonora et al. (2018) developed and tested a process stack to analyse the various time series recorded in automated milking systems of dairy cows. Their study focused

on a single farm with 88 cows. Data related to cows, milk production and cow collar accelerometers were monitored from June to September 2015. Their end-goal was to provide automatic grouping of the cows, based on production and behavioural features, to support livestock farm management. To achieve this, they performed multiple k-means clustering and, using graph representation, created networks of cows. This approach enabled them to highlight three clusters of cows in terms of productive characteristics. As the study was primarily exploratory and group detection oriented, all data entered the pipeline without the use of an external validation set.

“U” Achour et al. (2019) monitored the activity of dairy cows using data acquired via an inertial measurement unit (accelerometers combined with gyroscopes and magnetometers) attached to the back of eight dairy cows during December 2017. They used data related to three cows for model development and all data for validation. To monitor cow behaviour and better correlate behaviour with signals, they used an “automaton” – a pattern validator with nodes representing the states and the possible transitions validated, or, if forbidden, indicating a problem in the expected behavioural pattern. Building on this pattern detection, they developed a decision tree to classify behaviours into crisp clusters, reporting accuracies close to 98-99

“U” Xu et al. (2020) used location monitoring from ten cattle yearlings to identify the social structure of a small group via unsupervised learning (k-means and hierarchical clustering), applied on a conversion of their dataset based on MDS. More specifically, the study focused on “affinity”, defined as a measure derived from the ranking of pairwise distance. This methodology enabled a better understanding of social interactions within the group.

“U+S” Ismail et al. (2019) analysed images of dairy cows (no information on the study period was available), using 800 images for calibration and 200 for validation. Their processing stack included image resizing and gray-scaling, applying region-based convolutional neural networks to detect cows. They then explored four clustering methods – mixture model, hierarchical clustering, k-means, and DBSCAN – in order to distinguish herds and identify outliers. Hyper-parameters were explored (notably epsilon and k) and the impact of Euclidean and Manhattan distance were also studied regarding the formation of clusters.

“U+S” To identify grass-fed cows in 2,449 Walloon farms in Belgium, Soyeurt et al. (2022) used the FT-MIR spectra alongside the date of milk payment analysis, as most Walloons cows are in pasture during the summer. They predicted the grass-fed cows using PLS-DA and then studied the average value per farm/year to discover patterns in management practices.

“U+S” Kogo (2016) presented in their master thesis an automated approach to detect the presence of outliers in feed intake data, recorded from dairy cows from late August 2015 to late January 2016. This resulted in a dataset of 180,372 observations (primarily time-related) of 78 cows. The analysis was based on a GMM.

“U+S” Alessio et al. (2021) used a combination of methods to determine relationships in bulk tank milk composition, analyzing 154,781 samples from southern Brazil. Once clusters were defined through a PROC FASTCLUS, a PROC STEPDISC was used to understand which variables most strongly discriminated between clusters. Their findings indicated that an increase in somatic cell count and total bacterial count is associated with a decrease in lactose.

2.10.6. Classification of farms/respondents to surveys

“U” Todde et al. (2016) characterized 285 Italian dairy cow farms through the use of k-means on PCA projected features, reflecting general information such as herd size, animal categories, land use and ownership, milk quality and production. More precisely, the study aimed to identify gradients of insensitivity, mechanisation and structurality, based on detailed descriptions of cultivated crops, farm structures, equipment and machinery.

“U” Hyland et al. (2018) applied hierarchical clustering based on PCA scores to analyse factors influencing dairy farmers’ adoption of best management practices. Data were collected between May and August. They identified groups among 256 dairy farmers using with varying levels of use of a management tool – the Spring Rotation Planner. They highlighted two main groups within their dataset – low adopters, characterised by a strong sense of resource constraints, and high adopters.

“U” Vázquez-González et al. (2024) used hierarchical clustering combined with principal component factor analysis to analyse 86 survey responses, with the aim of classifying cow dairy farms in Cantabria between November 2016 and February 2017. The resulting topology was based on six pillars – milk production, land base, livestock, facilities and machinery, family labour and workforce and economic factors. This approach identified four main production typologies, differing in levels of agricultural activity, profitability and diversification, along with a fifth “group” comprising singular cases.

“U+S” Brotzman et al. (2015) explored trends in performance characteristics across 4,190 US dairy herds (3,078 after filtering, all used together without partitioning). Based on milk composition, cow description and herd repartition, they applied PCA to 22 standardised features and identified the 16 variables most strongly associated with the first six principal components. These entered the cluster analysis consisting of Ward’s partitioning method and Fourier plots. They identified six herd groups differentiated by performance, infection risk, inter-lactation timing and transitioning. A key perspective highlighted in the study is the relationship between herd management and performance cluster attribution.

2.11. Conclusion and future trends

Compared with supervised learning, relatively few studies focus exclusively on the application of unsupervised learning. This may reflect the common perception that a model must include a target variable to be practically useful. However, as demonstrated throughout this chapter, unsupervised learning offers considerable potential – not only for discovering novel patterns in data but also for improving data quality through cleaning, transforming features to reduce dimensionality, and facilitating the visualization of both inputs and outputs.

As highlighted, when large databases are available, unsupervised learning can be especially powerful, particularly when combined with supervised learning, to generate actionable insights for dairy farmers and the broader dairy industry. Such large-scale data sources are now more accessible than ever, and the required computational resources are widely available, enabling the extraction of meaningful knowledge at scale. In our view, further progress will depend on the integration of heterogeneous data sources, which could unlock new opportunities for precision dairy farming and innovation within the dairy value chain. Importantly, this perspective is not limited to dairy systems and can be extended to other types of livestock and agricultural production systems.

This chapter has aimed to provide an overview of the main techniques in unsupervised learning and their practical applications in dairy science. However, given the breadth and complexity of the subject, it is not possible to cover all aspects comprehensively within a single chapter. In addition to the key references cited herein, readers seeking further detail are encouraged to consult the broader literature for deeper insights into specific methods and use cases.

2.12. Where to look for further information

- For the theory of unsupervised learning:

Hinton, G. and Sejnowski, T.J. (1999). *Unsupervised Learning: Foundations of Neural Computation*. Cambridge: MIT Press (Computational Neuroscience Series).

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer (Information Science and Statistics).

- For the implementation of unsupervised learning algorithms in Python:

Patel, A.A. (2019). *Hands-on Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*. First edition. Sebastopol, CA: O'Reilly Media.

- Scikit-learn Python package and its documentation:

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

2.13. Acknowledgment

The authors would like to thank the EU Interreg NWE Holicow project for their funding.

Bibliography

- Achour, B., M. Belkadi, R. Aoudjit, and M. Laghrouche (2019). "Unsupervised automated monitoring of dairy cows' behavior based on Inertial Measurement Unit attached to their back". In: *Computers and Electronics in Agriculture* 167, p. 105068. ISSN: 01681699. DOI: 10.1016/j.compag.2019.105068.
- Agarwal, P., M. A. Alam, and R. Biswas (2011). "Issues, Challenges and Tools of Clustering Algorithms". In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1110.2610.
- Agrawal, R., J. Gehrke, D. Gunopulos, and P. Raghavan (1998). "Automatic subspace clustering of high dimensional data for data mining applications". In: *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*. SIGMOD '98. New York, NY, USA: Association for Computing Machinery, pp. 94–105. ISBN: 978-0-89791-995-1. DOI: 10.1145/276304.276314.
- Alessio, D. R. M., J. P. Velho, C. M. McManus, D. A. Knob, F. R. Vancin, G. V. Antunes, M. Busanello, F. De Carli, and A. T. Neto (2021). "Lactose and its relationship with other milk constituents, somatic cell count, and total bacterial count". In: *Livestock Science* 252, p. 104678. ISSN: 18711413. DOI: 10.1016/j.livsci.2021.104678.
- Alghushairy, O., R. Alsini, T. Soule, and X. Ma (2021). "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams". In: *Big Data and Cognitive Computing* 5.1. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 1. ISSN: 2504-2289. DOI: 10.3390/bdcc5010001.
- Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Probability and mathematical statistics: a series of monographs and textbooks 19. Burlington: Elsevier Science. 376 pp. ISBN: 978-1-4831-9139-3.
- Aungier, S., J. Roche, P. Duffy, S. Scully, and M. Crowe (2015). "The relationship between activity clusters detected by an automatic activity monitor and endocrine changes during the periestrous period in lactating dairy cows". In: *Journal of Dairy Science* 98.3, pp. 1666–1684. ISSN: 00220302. DOI: 10.3168/jds.2013-7405.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent Dirichlet allocation". In: *Journal Of Machine Learning Research* 3.4. Conference Name: 18th International Conference on Machine Learning Num Pages: 30 Place: Brookline Publisher: Microtome Publ

- Web of Science ID: WOS:000184926200015, pp. 993–1022. ISSN: 1532-4435. DOI: 10.1162/jmlr.2003.3.4-5.993.
- Bock, H.-H. (2007). “Clustering Methods: A History of k-Means Algorithms”. In: *Selected Contributions in Data Analysis and Classification*. Ed. by P. Brito, G. Cucumel, P. Bertrand, and F. De Carvalho. Red. by H. -. Bock, W. Gaul, M. Vichi, P. Arabie, D. Baier, F. Critchley, R. Decker, E. Diday, M. Greenacre, C. Lauro, J. Meulman, P. Monari, S. Nishisato, N. Ohsumi, O. Opitz, G. Ritter, M. Schader, and C. Weihs. Series Title: Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 161–172. ISBN: 978-3-540-73558-8. DOI: 10.1007/978-3-540-73560-1_15.
- Bonilla, J., D. Vélez, J. Montero, and J. T. Rodríguez (2021). “Fuzzy Clustering Methods with Rényi Relative Entropy and Cluster Size”. In: *Mathematics* 9.12, p. 1423. ISSN: 2227-7390. DOI: 10.3390/math9121423.
- Bonora, F., S. Benni, A. Barbaresi, P. Tassinari, and D. Torreggiani (2018). “A cluster-graph model for herd characterisation in dairy farms equipped with an automatic milking system”. In: *Biosystems Engineering* 167, pp. 1–7. ISSN: 15375110. DOI: 10.1016/j.biosystemseng.2017.12.007.
- Brotzman, R., N. Cook, K. Nordlund, T. Bennett, A. Gomez Rivas, and D. Döpfer (2015). “Cluster analysis of Dairy Herd Improvement data to discover trends in performance characteristics in large Upper Midwest dairy herds”. In: *Journal of Dairy Science* 98.5, pp. 3059–3070. ISSN: 00220302. DOI: 10.3168/jds.2014-8369.
- Chander, S. and P. Vijaya (2021). “Unsupervised learning methods for data clustering”. In: *Artificial Intelligence in Data Mining*. Elsevier, pp. 41–64. ISBN: 978-0-12-820601-0. DOI: 10.1016/B978-0-12-820601-0.00002-1.
- Comaniciu, D. and P. Meer (2002). “Mean shift: a robust approach toward feature space analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5, pp. 603–619. ISSN: 01628828. DOI: 10.1109/34.1000236.
- Cybenko, G. (1989). “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of Control, Signals, and Systems* 2.4, pp. 303–314. ISSN: 0932-4194, 1435-568X. DOI: 10.1007/BF02551274.
- De Maesschalck, R., D. Jouan-Rimbaud, and D. L. Massart (2000). “The Mahalanobis distance”. In: *Chemometrics and Intelligent Laboratory Systems* 50.1, pp. 1–18. ISSN: 0169-7439. DOI: 10.1016/S0169-7439(99)00047-7.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1977.tb01600.x.
- Espiritu, H. M., S.-w. Kwon, S.-s. Lee, and Y.-i. Cho (2025). “Surveillance of hoof disorders in Korean dairy cattle and the correlation of farm condition risk factors

- to their prevalence". In: *BMC Veterinary Research* 21.1, p. 169. ISSN: 1746-6148. DOI: 10.1186/s12917-025-04628-9.
- Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, pp. 226–231. (Visited on 04/14/2025).
- Ezugwu, A. E. (2020). "Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study". In: *SN Applied Sciences* 2.2, p. 273. ISSN: 2523-3971. DOI: 10.1007/s42452-020-2073-0.
- Ezugwu, A. E., A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu (2022). "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects". In: *Engineering Applications of Artificial Intelligence* 110, p. 104743. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2022.104743.
- Foldager, L., C. Gaillard, M. T. Sorensen, T. Larsen, E. Matthews, R. O'Flaherty, F. Carter, M. A. Crowe, C. Grelet, M. Salavati, M. Hostens, K. L. Ingvarsten, and M. A. Krogh (2020). "Predicting physiological imbalance in Holstein dairy cows by three different sets of milk biomarkers". In: *Preventive Veterinary Medicine* 179. ISSN: 0167-5877. DOI: 10.1016/j.prevetmed.2020.105006.
- Fraley, C. and A. E. Raftery (1998). "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis". In: *The Computer Journal* 41.8, pp. 578–588. ISSN: 0010-4620. DOI: 10.1093/comjnl/41.8.578.
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt (2022). "Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?" In: *Journal of Dairy Science* 105.8, pp. 6760–6772. ISSN: 0022-0302. DOI: 10.3168/jds.2022-21975.
- Franceschini, S., C. Fastré, C. Nickmilder, D. E. Santschi, D. Warner, M. Bahadi, C. Bertozzi, D. Veselko, F. Dehareng, N. Gengler, and H. Soyeurt (2025). "Detection of Dairy Herd Management Issues Using Fatty Acid Profiles Predicted by Mid-Infrared Spectrometry". In: *Animals* 15.11. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 1575. ISSN: 2076-2615. DOI: 10.3390/ani15111575.
- Fred, A. L. N. and J. M. N. Leitão (2000). "Partitional vs Hierarchical Clustering Using a Minimum Grammar Complexity Approach". In: *Advances in Pattern Recognition*. Ed. by F. J. Ferri, J. M. Iñesta, A. Amin, and P. Pudil. Red. by G. Goos, J. Hartmanis, and J. Van Leeuwen. Vol. 1876. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 193–202. ISBN: 978-3-540-67946-2. DOI: 10.1007/3-540-44522-6_20.
- Frey, B. J. and D. Dueck (2007). "Clustering by Passing Messages Between Data Points". In: *Science* 315.5814. Publisher: American Association for the Advancement of Science, pp. 972–976. DOI: 10.1126/science.1136800.

- Fukunaga, K. and L. Hostetler (1975). "The estimation of the gradient of a density function, with applications in pattern recognition". In: *IEEE Transactions on Information Theory* 21.1, pp. 32–40. ISSN: 1557-9654. DOI: 10.1109/TIT.1975.1055330.
- Gao, C. X., D. Dwyer, Y. Zhu, C. L. Smith, L. Du, K. M. Filia, J. Bayer, J. M. Menssink, T. Wang, C. Bergmeir, S. Wood, and S. M. Cotton (2023). "An overview of clustering methods with guidelines for application in mental health research". In: *Psychiatry Research* 327, p. 115265. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2023.115265.
- Ghaffari, M. H., A. Jahanbekam, C. Post, H. Sadri, K. Schuh, C. Koch, and H. Sauerwein (2020). "Discovery of different metabotypes in overconditioned dairy cows by means of machine learning". In: *Journal of Dairy Science* 103.10, pp. 9604–9619. ISSN: 00220302. DOI: 10.3168/jds.2020-18661.
- Ghahramani, Z. (2004). "Unsupervised Learning". In: *Advanced Lectures on Machine Learning*. Ed. by O. Bousquet, U. Von Luxburg, and G. Rätsch. Vol. 3176. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 72–112. ISBN: 978-3-540-23122-6. DOI: 10.1007/978-3-540-28650-9_5.
- Greenacre, M., P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina (2022). "Principal component analysis". In: *Nature Reviews Methods Primers* 2.1. Publisher: Nature Publishing Group, pp. 1–21. ISSN: 2662-8449. DOI: 10.1038/s43586-022-00184-w.
- Grelet, C., T. Larsen, M. Crowe, D. Wathes, C. Ferris, K. Ingvarsten, C. Marchitelli, F. Becker, A. Vanlierde, J. Leblois, U. Schuler, F. Auer, A. Köck, L. Dale, J. Sölkner, O. Christophe, J. Hummel, A. Mensching, J. Fernández Pierna, H. Soyeurt, M. Calmels, R. Reding, M. Gelé, Y. Chen, N. Gengler, and F. Dehareng (2024). "Prediction of key milk biomarkers in dairy cows through milk mid-infrared spectra and international collaborations". In: *Journal of Dairy Science* 107.3, pp. 1669–1684. ISSN: 00220302. DOI: 10.3168/jds.2023-23843.
- Guha, S., R. Rastogi, and K. Shim (1998). "CURE: an efficient clustering algorithm for large databases". In: *ACM SIGMOD Record* 27.2, pp. 73–84. ISSN: 0163-5808. DOI: 10.1145/276305.276312.
- Hariri, S., M. C. Kind, and R. J. Brunner (2021). "Extended Isolation Forest". In: *IEEE Transactions on Knowledge and Data Engineering* 33.4, pp. 1479–1489. ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2019.2947676. arXiv: 1811.02141[cs].
- Hartigan, J. A. (1975). *Clustering algorithms*. A Wiley publication in applied statistics. New York: Wiley. 351 pp. ISBN: 978-0-471-35645-5.
- Ho, P., V. Bonfatti, T. Luke, and J. Pryce (2019). "Classifying the fertility of dairy cows using milk mid-infrared spectroscopy". In: *Journal of Dairy Science* 102.11, pp. 10460–10470. ISSN: 00220302. DOI: 10.3168/jds.2019-16412.

- Hyland, J. J., K. Heanue, J. McKillop, and E. Micha (2018). "Factors influencing dairy farmers' adoption of best management grazing practices". In: *Land Use Policy* 78, pp. 562–571. ISSN: 02648377. DOI: 10.1016/j.landusepol.2018.07.006.
- Ismail, Z. H., A. K. K. Chun, and M. I. Shapiai Razak (2019). "Efficient Herd – Outlier Detection in Livestock Monitoring System Based on Density – Based Spatial Clustering". In: *IEEE Access* 7, pp. 175062–175070. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2952912.
- Izenman, A. J. (2012). "Introduction to manifold learning". In: *WIREs Computational Statistics* 4.5, pp. 439–446. ISSN: 1939-0068. DOI: 10.1002/wics.1222.
- Jain, A. K., M. N. Murty, and P. J. Flynn (1999). "Data clustering: a review". In: *ACM Computing Surveys* 31.3, pp. 264–323. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/331499.331504.
- James, G., D. Witten, T. Hastie, R. Tibshirani, and J. Taylor (2023). "Unsupervised Learning". In: *An Introduction to Statistical Learning*. Series Title: Springer Texts in Statistics. Cham: Springer International Publishing, pp. 503–556. ISBN: 978-3-031-38746-3. DOI: 10.1007/978-3-031-38747-0_12.
- Kaufman, L. (2005). *Finding groups in data: an introduction to cluster analysis*. In collab. with P. J. Rousseeuw. Wiley series in probability and mathematical statistics. Hoboken, NJ: Wiley. 342 pp. ISBN: 978-0-470-31748-8.
- Kiouvrekis, Y., N. G. C. Vasileiou, E. I. Katsarou, D. T. Lianou, C. K. Michael, S. Zikas, A. I. Katsafadou, M. V. Bourganou, D. V. Liagka, D. C. Chatzopoulos, and G. C. Fthenakis (2024). "The Use of Machine Learning to Predict Prevalence of Subclinical Mastitis in Dairy Sheep Farms". en. In: *Animals* 14.16, p. 2295. ISSN: 2076-2615. DOI: 10.3390/ani14162295.
- Kogo, G. (2016). "Analyzing Automatic Cow Recordings to Detect the Presence of Outliers in Feed Intake Data Recorded from Dairy Cows in Lovsta Farm". PhD thesis. Dalarna University. URL: <https://urn.kb.se/resolve?urn=urn:nbn:se:du-24522>.
- Li, H., X. Tian, Z. Ruan, S. Liu, W. Ren, Z. Su, R. Gao, and Q. Kong (2024). "Seedling Stage Corn Line Detection Method Based on Improved YOLOv8". In: *Smart Agriculture* 6.6, pp. 72–84.
- Linderman, G. C. and S. Steinerberger (2017). *Clustering with t-SNE, provably*. Version Number: 1. DOI: 10.48550/ARXIV.1706.02582.
- Liu, E., H. Yang, S. Sharma, M. B. v. Leerdam, P. Niu, M. J. VandeHaar, and M. Hostens (2025). "Agents are all you need: Pioneering the use of agentic artificial intelligence to embrace large language models into dairy science". English. In: *Journal of Dairy Science* 108.12, pp. 14038–14049. ISSN: 0022-0302. DOI: 10.3168/jds.2025-26775.
- Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008). "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008 Eighth IEEE International Conference

- on Data Mining (ICDM). Pisa, Italy: IEEE, pp. 413–422. ISBN: 978-0-7695-3502-9. DOI: 10.1109/ICDM.2008.17.
- Liu, T., H. Yu, and R. H. Blair (2022). “Stability estimation for unsupervised clustering: A review”. In: *WIREs Computational Statistics* 14.6, e1575. ISSN: 1939-5108, 1939-0068. DOI: 10.1002/wics.1575.
- Lloyd, S. (1982). “Least squares quantization in PCM”. In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. ISSN: 1557-9654. DOI: 10.1109/TIT.1982.1056489.
- Ma, J., A. Kok, E. E. A. Burgers, R. M. Bruckmaier, R. M. A. Goselink, J. J. Gross, B. Kemp, T. J. G. M. Lam, A. Minuti, E. Saccenti, E. Trevisi, F. Vosseveld, and A. T. M. van Knegsel (2024). “Time profiles of energy balance in dairy cows in association with metabolic status, inflammatory status, and disease”. In: *Journal of Dairy Science* 107.11, pp. 9960–9977. ISSN: 1525-3198. DOI: 10.3168/jds.2024-24680.
- Maaten, L. v. d. and G. Hinton (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. ISSN: 1533-7928. URL: <http://jmlr.org/papers/v9/vandemaaten08a.html> (visited on 04/29/2025).
- MacQueen, J. (Jan. 1, 1967). “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Vol. 5.1. University of California Press. (Visited on 04/29/2025).
- Mahdi, M. A., K. M. Hosny, and I. Elhenawy (2021). “Scalable Clustering Algorithms for Big Data: A Review”. In: *IEEE Access* 9, pp. 80015–80027. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3084057.
- McInnes, L., J. Healy, and S. Astels (2017). “hdbscan: Hierarchical density based clustering”. In: *The Journal of Open Source Software* 2.11, p. 205. ISSN: 2475-9066. DOI: 10.21105/joss.00205.
- McInnes, L., J. Healy, and J. Melville (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426[stat].
- McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). “Finite Mixture Models”. In: *Annual Review of Statistics and Its Application* 6 (Volume 6, 2019). Publisher: Annual Reviews, pp. 355–378. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-031017-100325.
- McVey, C., F. Hsieh, D. Manriquez, P. Pinedo, and K. Horback (2020). “Mind the Queue: A Case Study in Visualizing Heterogeneous Behavioral Patterns in Livestock Sensor Data Using Unsupervised Machine Learning Techniques”. In: *Frontiers in Veterinary Science* 7, p. 523. ISSN: 2297-1769. DOI: 10.3389/fvets.2020.00523.
- Meilä, M. and H. Zhang (2024). “Manifold Learning: What, How, and Why”. In: *Annual Review of Statistics and Its Application* 11 (Volume 11, 2024). Publisher: Annual Reviews, pp. 393–417. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-040522-115238.

- Michelena, Á., A. Díaz-Longueira, P. Novais, D. Simić, Ó. Fontenla-Romero, and J. L. Calvo-Rolle (2025). “Comparative analysis of unsupervised anomaly detection techniques for heat detection in dairy cattle”. In: *Neurocomputing* 618, p. 129088. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2024.129088.
- Mor, B., S. Garhwal, and A. Kumar (2021). “A Systematic Review of Hidden Markov Models and Their Applications”. In: *Archives of Computational Methods in Engineering* 28.3, pp. 1429–1448. ISSN: 1886-1784. DOI: 10.1007/s11831-020-09422-4.
- Murtagh, F. and P. Contreras (2012). “Algorithms for hierarchical clustering: an overview”. In: *WIREs Data Mining and Knowledge Discovery* 2.1, pp. 86–97. ISSN: 1942-4795. DOI: 10.1002/widm.53.
- Murtagh, F. and P. Contreras (2017). “Algorithms for hierarchical clustering: an overview, II”. In: *WIREs Data Mining and Knowledge Discovery* 7.6, e1219. ISSN: 1942-4795. DOI: 10.1002/widm.1219.
- Murtagh, F. and P. Legendre (2014). “Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?”. In: *Journal of Classification* 31.3, pp. 274–295. ISSN: 1432-1343. DOI: 10.1007/s00357-014-9161-z.
- Narita, K., T. Hochin, and H. Nomiya (2018). “Incremental Clustering for Hierarchical Clustering”. In: *2018 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*. 2018 5th International Conference on Computational Science/ Intelligence and Applied Informatics (CSII), pp. 102–107. DOI: 10.1109/CSII.2018.00025.
- Oehm, A. W., A. Springer, D. Jordan, C. Strube, G. Knubben-Schweizer, K. C. Jensen, and Y. Zablotzki (2022). “A machine learning approach using partitioning around medoids clustering and random forest classification to model groups of farms in regard to production parameters and bulk tank milk antibody status of two major internal parasites in dairy cows”. In: *PLOS ONE* 17.7. Ed. by R. L. Smith, e0271413. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0271413.
- Panesar, A. (2019). *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. Berkeley, CA: Apress L. P. 1 p. ISBN: 978-1-4842-3799-1.
- Parsons, L., E. Haque, and H. Liu (2004). “Subspace clustering for high dimensional data: a review”. In: *SIGKDD Explor. Newsl.* 6.1, pp. 90–105. ISSN: 1931-0145. DOI: 10.1145/1007730.1007731.
- Pastell, M. and L. Frondelius (2018). “A hidden Markov model to estimate the time dairy cows spend in feeder based on indoor positioning data”. In: *Computers and Electronics in Agriculture* 152, pp. 182–185. ISSN: 0168-1699. DOI: 10.1016/j.compag.2018.07.005.
- Patel, A. A. (2019). *Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data*. First edition. OCLC: on1066070019. Sebastopol, CA: O’Reilly Media. 337 pp. ISBN: 978-1-4920-3564-0.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M.

- Brucher, M. Perrot, and É. Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. ISSN: 1533-7928. URL: <http://jmlr.org/papers/v12/pedregosa11a.html> (visited on 05/06/2025).
- Pitafi, S., T. Anwar, and Z. Sharif (2023). “A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms”. In: *Applied Sciences* 13.6, p. 3529. ISSN: 2076-3417. DOI: 10.3390/app13063529.
- Prince, S. J. D. (2024). *Understanding deep learning*. Cambridge, Massachusetts London, England: The MIT Press. 1 p. ISBN: 978-0-262-04864-4.
- Ran, X., Y. Xi, Y. Lu, X. Wang, and Z. Lu (2023). “Comprehensive survey on hierarchical clustering algorithms and the recent developments”. In: *Artificial Intelligence Review* 56.8, pp. 8219–8264. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10366-3.
- Rane, J., Ö. Kaya, S. K. Mallick, and N. L. Rane (2024). *Smart farming using artificial intelligence, machine learning, deep learning, and ChatGPT: Applications, opportunities, challenges, and future directions*. Deep Science Publishing. DOI: 10.70593/978-81-981271-7-4.
- Rebala, G., A. Ravi, and S. Churiwala (2019). “Machine Learning Definition and Basics”. In: *An Introduction to Machine Learning*. Ed. by G. Rebala, A. Ravi, and S. Churiwala. Cham: Springer International Publishing, pp. 1–17. ISBN: 978-3-030-15729-6. DOI: 10.1007/978-3-030-15729-6_1.
- Rebuli, B., L. Ozella, F. Masia, E. Vrieze, and M. Giacobini (2024). “Exploring Milk Productivity and Energy Output Dynamics across Lactation Periods in Automatic Milking Systems through Multi-Algorithm Clustering”. In: Accepted: 2024-10-23T07:50:40Z. ITA. URL: <https://iris.unito.it/handle/2318/2027989> (visited on 04/29/2025).
- Samariya, D. and A. Thakkar (2023). “A Comprehensive Survey of Anomaly Detection Algorithms”. In: *Annals of Data Science* 10.3, pp. 829–850. ISSN: 2198-5812. DOI: 10.1007/s40745-021-00362-9.
- Satoła, A. and K. Satoła (2024). “Performance comparison of machine learning models used for predicting subclinical mastitis in dairy cows: Bagging, boosting, stacking, and super-learner ensembles versus single machine learning models”. In: *Journal of Dairy Science* 107.6, pp. 3959–3972. ISSN: 00220302. DOI: 10.3168/jds.2023-24243.
- Schaeffer, S. E. (2007). “Graph clustering”. In: *Computer Science Review* 1.1, pp. 27–64. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2007.05.001.
- Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke, and K. Weigel (2014). “Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms”. In: *Journal of Dairy Science* 97.2, pp. 731–742. ISSN: 00220302. DOI: 10.3168/jds.2013-6693.

- Shahriar, M. S., D. Smith, A. Rahman, D. Henry, G. Bishop-Hurley, R. Rawnsley, M. Freeman, and J. Hills (2015). "Heat event detection in dairy cows with collar sensors: An unsupervised machine learning approach". In: *2015 IEEE SENSORS*. 2015 IEEE Sensors. Busan: IEEE, pp. 1–4. ISBN: 978-1-4799-8203-5. DOI: 10.1109/ICSENS.2015.7370528.
- Shi, C., B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu (2020). *A Quantitative Discriminant Method of Elbow Point for the Optimal Number of Clusters in Clustering Algorithm*. DOI: 10.21203/rs.3.rs-58011/v2.
- Shine, P., M. Murphy, J. Upton, and T. Scully (2018). "Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms". In: *Computers and Electronics in Agriculture* 150, pp. 74–87. ISSN: 01681699. DOI: 10.1016/j.compag.2018.03.023.
- Simankov, N., R. Tahzima, S. Massart, and H. Soyeurt (2025). "Illuminating The Path To Enhanced Resilience Of Machine Learning Models Against The Shadows Of Missing Labels". In: *IEEE Journal of Biomedical and Health Informatics*, pp. 1–11. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2025.3558846.
- Singh, H. V., A. Girdhar, and S. Dahiya (2022). "A Literature survey based on DB-SCAN algorithms". In: *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). ISSN: 2768-5330, pp. 751–758. DOI: 10.1109/ICICCS53718.2022.9788440.
- Sowmya, N. and V. Ponnusamy (2021). "Development of Spectroscopic Sensor System for an IoT Application of Adulteration Identification on Milk Using Machine Learning". In: *IEEE Access* 9, pp. 53979–53995. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3070558.
- Soyeurt, H., C. Gerards, C. Nickmilder, J. Bindelle, S. Franceschini, F. Dehareng, D. Veselko, C. Bertozzi, N. Gengler, A. Marvuglia, A. Bayram, and A. Tedde (2022). "Prediction of Indirect Indicators of a Grass-Based Diet by Milk Fourier Transform Mid-Infrared Spectroscopy to Assess the Feeding Typologies of Dairy Farms". In: *Animals* 12.19, p. 2663. ISSN: 2076-2615. DOI: 10.3390/ani12192663.
- Stone, J. V. (2002). "Independent component analysis: an introduction". In: *Trends in Cognitive Sciences* 6.2. Publisher: Elsevier, pp. 59–64. ISSN: 1364-6613, 1879-307X. DOI: 10.1016/S1364-6613(00)01813-1.
- Toan, T. V., R. Nishikawa, L. T. Thanh, M. Takemoto, T. Van Hoai, H. T. T. Binh, and H. Nakajo (2019). "Cow estrus detection with low-frequency accelerometer sensor by unsupervised learning". In: *Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019*. the Tenth International Symposium. Hanoi, Ha Long Bay, Viet Nam: ACM Press, pp. 342–349. ISBN: 978-1-4503-7245-9. DOI: 10.1145/3368926.3369683.
- Todde, G., L. Murgia, M. Caria, and A. Pazzona (2016). "A multivariate statistical analysis approach to characterize mechanization, structural and energy profile

- in Italian dairy farms”. In: *Energy Reports* 2, pp. 129–134. ISSN: 23524847. DOI: 10.1016/j.egyr.2016.05.006.
- Tremblay, M., M. Kammer, H. Lange, S. Plattner, C. Baumgartner, J. Stegeman, J. Duda, R. Mansfeld, and D. Döpfer (2018). “Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis”. In: *Journal of Dairy Science* 101.8, pp. 7311–7321. ISSN: 00220302. DOI: 10.3168/jds.2017-13582.
- Udell, M., C. Horn, R. Zadeh, and S. Boyd (2016). “Generalized Low Rank Models”. In: *Foundations and Trends® in Machine Learning* 9.1, pp. 1–118. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/22000000055.
- Van Hoeij, R., A. Kok, R. Bruckmaier, M. Haskell, B. Kemp, and A. Van Kneegsel (2019). “Relationship between metabolic status and behavior in dairy cows in week 4 of lactation”. In: *Animal* 13.3, pp. 640–648. ISSN: 17517311. DOI: 10.1017/S1751731118001842.
- Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt (2021). “Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra”. In: *Journal of the Science of Food and Agriculture* 101.8, pp. 3394–3403. ISSN: 1097-0010. DOI: 10.1002/jsfa.10969.
- Vázquez-González, I., E. García-Suárez, F. Ruiz-Escudero, and A. Isabel García-Arias (2024). “A combined multi-variate statistical analysis to establish dairy farm typologies in Cantabria”. In: *Computers and Electronics in Agriculture* 221, p. 109007. ISSN: 01681699. DOI: 10.1016/j.compag.2024.109007.
- Wattenberg, M., F. Viégas, and I. Johnson (2016). “How to Use t-SNE Effectively”. In: *Distill* 1.10, 10.23915/distill.00002. DOI: 10.23915/distill.00002.
- Xu, H., S. Li, C. Lee, W. Ni, D. Abbott, M. Johnson, J. M. Lea, J. Yuan, and D. L. M. Campbell (2020). “Analysis of Cattle Social Transitional Behaviour: Attraction and Repulsion”. In: *Sensors* 20.18, p. 5340. ISSN: 1424-8220. DOI: 10.3390/s20185340.
- Xu, W., A. T. Van Kneegsel, J. J. Vervoort, R. M. Bruckmaier, R. J. Van Hoeij, B. Kemp, and E. Saccenti (2019). “Prediction of metabolic status of dairy cows in early lactation with on-farm cow data and machine learning algorithms”. In: *Journal of Dairy Science* 102.11, pp. 10186–10201. ISSN: 00220302. DOI: 10.3168/jds.2018-15791.
- Yujuan, C., H. Dongjian, F. Yinxi, and S. Huaibo (2017). “Intelligent monitoring method of cow ruminant behavior based on video analysis technology”. In: *International Journal of Agricultural and Biological Engineering* 10.5, pp. 194–202. ISSN: 1934-6344. DOI: 10.25165/j.ijabe.20171005.3117.
- Zhang, L., C. Li, F. Dehareng, C. Grelet, F. Colinet, N. Gengler, Y. Brostaux, and H. Soyeurt (2021). “Appropriate data quality checks improve the reliability of values predicted from milk mid-infrared spectra”. In: *Animals* 11.2, pp. 1–17. DOI: 10.3390/ani11020533.

Zhang, T., R. Ramakrishnan, and M. Livny (1996). “BIRCH: an efficient data clustering method for very large databases”. In: *ACM SIGMOD Record* 25.2, pp. 103–114. ISSN: 0163-5808. DOI: 10.1145/235968.233324.

Can unsupervised learning methods applied to Milk Recording Big Data bring new insights into dairy cow health?

Adapted from: Franceschini, S., Grelet, C., Leblois, J., Gengler, N., the GplusE consortium, and Soyeurt, H. (2022). Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health? Journal of Dairy Science, 105(8), 6760–6772. <https://doi.org/10.3168/jds.2022-21975>

3.1. Chapter outline

After reviewing methodologies used in unsupervised learning, this chapter presents the first research work of the thesis: the application of clustering approaches to a large milk recording database composed of Fourier Transform mid-infrared (FT-MIR)-based predictions. The interest in unsupervised learning emerged from two major research questions that arose during the work:

1. How can cows with potential health issues be identified when no reliable reference diagnoses are available?
2. Are the numerous traits predicted from milk FT-MIR spectra truly providing distinct information, despite being derived from the same spectral signal?

These two questions reflect important limitations of the traditional FT-MIR prediction framework: the lack of reference phenotypes and the potential redundancy of spectral information. Indeed, the number of FT-MIR-based predictions has continuously increased over recent years, while all models are developed from the same spectral information. Consequently, an important question arises regarding the extent to which these predictions capture distinct biological information or reflect overlapping physiological signals, and whether using these predictions together is relevant

Rather than considering FT-MIR predictions as isolated variables, this work explored whether their joint structure could reveal underlying physiological or pathological states shared across animals. A hierarchical clustering approach was therefore applied to a large dataset of FT-MIR predictions in order to investigate latent structures within the data. The objective was to develop first a proof of concept of the health pattern detection from FT-MIR predictions and to identify the limitations of the approach.

3.2. Abstract

Among the dairy sector's current concerns, the assessment of global animal health status is a complex challenge. Its multidimensionality means that global monitoring tools are rarely considered. Instead, specific disease detection is often studied separately and, due to financial and ethical issues, uses small-scale datasets focusing on few biomarkers. Several studies have already been conducted using milk FT-MIR spectroscopy to detect mastitis and lameness or to quantify health-related biomarkers in milk or blood. Those studies are relevant but they focus mainly on one biomarker or disease. To solve this issue and the small-scale data set, in this study, we proposed a holistic approach using big data obtained from milk recording, including milk yield, somatic cell count, and 27 FT-MIR-based predictors related to milk composition and animal health status. Using 740,454 records collected from 114,536 first-parity Holstein cows in southern Belgium, we performed repeated unsupervised learning

algorithms based on Ward's agglomerative hierarchical clustering method to find potential interesting patterns. A divide-and-conquer approach was used to overcome the limitation of computational resources in clustering a relatively large dataset. Five groups of records were identified. Differences observed in the fourth group suggested a relationship to metabolic disorders. The fifth group seemed to be related to mastitis. In a second step, we performed Partial Least Square Discriminant Analysis (PLS-DA) to predict the probability of belonging to those specific groups for the entire data set. The obtained global accuracy was 0.77 and the balanced accuracy (i.e., the mean between sensitivity and specificity) of discriminating the fourth and fifth groups was 0.88 and 0.96, respectively. Then, a validation of the interpretation of those groups was performed using 204 milk and blood reference records. The predicted probability associated with the metabolic disorders issue had significant correlations of 0.54 with blood β -hydroxybutyrate (BHB), 0.44 with blood nonesterified fatty acids, -0.32 with blood glucose, -0.23 with milk glucose-6-phosphate, and 0.38 with milk isocitrate. In contrast, the predicted probability of belonging to the mastitis group had correlations of 0.69 with milk lactate dehydrogenase, 0.46 with milk N-acetyl- β -d-glucosaminidase, -0.18 with milk free glucose, and 0.16 with milk glucose-6-phosphate. Consequently, these results suggest that the obtained quantitative traits indirectly reflect some of the main health disorders in dairy farming and could be used to monitor dairy cows on a large scale. By using unsupervised learning on large-scale milk recording data and then validating the pattern using reference laboratory measures, we propose a new approach to quickly assess dairy cow health status.

Key words: big data, animal health, unsupervised learning, milk, mid-infrared

3.3. Introduction

Monitoring dairy cow health is complex because it involves many aspects such as disease, energy balance, and heat stress. However, dairy cow health, being a part of animal welfare, is important for dairy farming sustainability. Stakeholders (producers, policymakers, and consumers) need warranties about the ethics of the product. In particular, the welfare status of cows is of increasing concern (Fraser 1995; Blokhuis et al. 2003).

Currently, cow health monitoring tools mostly rely on milk and blood analyses, which provide health related biomarkers. However, defining health indicators is a lengthy and costly process because it requires many potential biomarkers to be scanned among an animal population that has health issues (Foldager et al. 2020; Krogh et al. 2020). Depending on the issue, the process can raise ethical questions when the disorder is induced in animals or the measurement is invasive. In other cases, detecting animals with health issues is complicated because the farms participating in the studies are already aware of the problem, making the sample unrepresentative and biased toward healthy animals.

Animal scientists have developed strategies to predict health indicators using milk FT-MIR spectrometry to avoid ethical issues and the cost of routine measurements of biomarkers. This spectroscopic technology has many advantages as it is noninvasive and reflects part of the animal's metabolism by assessing the global milk composition. Moreover, milk FT-MIR spectra, along with other dairy production phenotypes, are collected by Dairy Herd Improvement (DHI) organizations on average once a month in participating herds. From the FT-MIR spectra, statistical models can predict biomarkers of interest such as acetone, BHB, citrate, lactoferrin, and Somatic Cells Count (SCC) to assess and, therefore, monitor dairy cow health (Soyeurt et al. 2012; Luke et al. 2019). This technique enables routine individual assessment of cow health status on farms after applying appropriate prediction equations related to the biomarkers to the recorded spectral data.

A standard procedure to obtain prediction models consists of collecting blood and milk samples from as many healthy and sick animals as possible, analyzing those milk samples using FT-MIR spectrometry to record the spectral data, and measuring the contents of the studied biomarkers in milk and blood using certified chemical methodologies. Afterward, supervised machine learning methods such as Partial Least Square (PLS) regression can be used to build the model, predicting the desired biomarker (Soyeurt et al. 2020). This process has been largely used in the last decades to increase the number of phenotypes predicted by FT-MIR and not only to predict health issues (De Marchi et al. 2014; Bastin et al. 2016; Gengler et al. 2016). Table 1 presents the characteristics of some models published in the literature predicting different health biomarkers naturally present in milk and blood. A large range of cross-validation accuracy can be observed: Coefficient of Determination (R^2) varied between 0.21 and 0.88 (Table 3.1). More information on key factors affecting the quality of predictions is available in Grelet et al. (2021).

Model performance and validation accuracy can sometimes be poorer for several reasons. First, for practical, ethical, or financial issues, the calibration data set may not be representative enough, with too few animals or herds. However, to maximize the prediction quality, the modeling procedure requires sampling among the whole animal population to maximize the variability of the calibration data used to develop the equation (Grelet et al. 2021). Thus, making a prediction outside of the calibration variability is detrimental to the prediction quality due to inner extrapolation. In addition, maximizing the calibration set for health indicators requires both healthy and sick animals to be included. Sick animals are relatively scarce, which explains why obtaining high BHB levels is more difficult, as mentioned by Grelet et al. (2016), Fleming et al. (2017), and Bonfatti et al. (2019). This unbalanced situation between sick and healthy animals can affect the final prediction accuracy of the developed models.

Finally, most of the metabolites of interest have a very low concentration in milk or are even predicted in another matrix (e.g., blood). These indirect predictions rely

more on global milk changes than on measuring a specific molecule. Consequently, the disorder cannot be simulated artificially, such as spiking milk with molecules of interest, because it would interfere with the natural relationships existing between the milk components. All these aspects explain the difficulty in developing prediction models related to health indicators. To solve these issues, it would be interesting to combine different health biomarker predictions, different models derived from different populations, and usual production traits and indicators to provide insights into cow health status (e.g., milk production or BW).

Table 3.1: Published prediction models for health biomarkers

Trait	Unit	No. of records	No. of cows	No. of herds	R ² cv	Reference
Energy balance	MJ	1,535	378	1	0.67–0.78	McParland et al. 2016
	MJ/d	11,525–30,409	1,101	1	0.42–0.77	Smith et al. 2019
Energy intake	MJ/d	11,940–30,565	1,101	1	0.39–0.75	Smith et al. 2019
	MJ	1,535	378	1	0.64–0.88	McParland et al. 2016
Blood BHB	mmol/L	1,914	826	55	0.21–0.43	Belay et al. 2017
	mmol/L	3,629	1,013	3	0.27–0.46	Pralle et al. 2018
	mmol/L	380	241	-	0.70	Grelet et al. 2016
	mmol/L	1,910	542	2	0.50–0.58	Bonfatti et al. 2019
Blood glucose	mmol/L	380	241	-	0.44	Grelet et al. 2019
Blood IGF-1	mg/L	387	241	-	0.61	Grelet et al. 2019
Blood NEFA	uEq/L	380	241	-	0.39	Grelet et al. 2019
Milk lactoferrin	mg/L	2,499	-	>9	0.71	Soyeurt et al. 2012
	mg/L	5,541	-	>9	0.60	Soyeurt et al. 2020

R²cv = accuracy of cross-validation. NEFA = Nonesterified fatty acids.

In their review of infrared spectrometry as a high-throughput phenotyping technology, Bresolin et al. (2020) suggest using larger data sets and modern data-mining approaches to improve the models. In the current study, the innovative aspect consists of realizing an unsupervised learning approach that profits from the large-scale data set collected from DHI and combining several FT-MIR-based predictions and usual dairy traits to discriminate groups of cows that have a different data pattern. This approach allows the combination of several indicators of health issues (i.e., multidimensional aspect). Furthermore, the exploration of a large DHI database ensures the consideration of high variability and warrants the representativeness of the sample set. Therefore, the context of this study can be considered as a “big data” case; the 5 Vs representing aspects of big data are found: volume, variety, veracity, velocity, and value (Gandomi et al. 2015).

Moreover, unsupervised methods have the advantage of not introducing any bias about the definition of health. Among those methods, clustering, and particularly hierarchical clustering, is used as it allows patterns to be observed at any scale of the database and also cuts the data into groups. Previous studies have considered multiple reference health measures together and even clustering, but they were conducted using small-scale reference data sets (Leblanc 2010; Grelet et al. 2019; Atashi et al.

2020) and not from a routine DHI database. Therefore, the current approach presents the distinction of having no reference analysis of health biomarkers to construct the discrimination between healthy and sick animals. However, this discrimination will be validated using real reference measurements to attest to the relevancy of the discovered groups.

Table 3.2: Performance of the equations used

Equation ²	Unit	N	R_c^2	R_{cv}^2	SE_c	SE_{cv}	RPD_{cv}	Reference
BW	kg	721	0.51	0.47	51	53	1.37	Tedde et al. 2021
CH ₄	g/d	1,089	0.73	0.68	53	57	1.78	Vanlierde et al. 2021
Sodium	mg/kg of milk	1,019	0.48	0.44	49.25	50.98	1.34	Christophe et al. 2021
Calcium	mg/kg of milk	1,094	0.83	0.82	51.79	53.38	2.34	Christophe et al. 2021
Phosphorus	mg/kg of milk	1,083	0.76	0.75	57.36	58.71	1.99	Christophe et al. 2021
Magnesium	mg/kg of milk	1,124	0.73	0.72	6.35	6.53	1.88	Christophe et al. 2021
Potassium	mg/kg of milk	1,090	0.57	0.55	85.79	88.14	1.48	Christophe et al. 2021
Lactoferrin	mg/L	924	0.74	0.71	125	131	1.86	Soyeurt et al. 2012
EB	%	1,010	0.48	0.43	4.84	5.04	1.33	Grelet et al. 2017
DMI2	kg/d	–	–	0.71	–	5.70	1.83	Grelet et al. 2021
NUE	%	1,093	0.55	0.52	3.05	3.17	1.44	Grelet et al. 2020
Citrate	mmol/L	498	0.90	0.89	0.68	0.72	3.04	Grelet et al. 2016
Acetone	mmol/L	201	0.69	0.62	0.19	0.21	1.63	Grelet et al. 2016
BHB	μ mol/L	419	0.75	0.71	0.13	0.14	1.85	Grelet et al. 2016
RFI	kg/d	1,008	0.52	0.48	2.73	2.87	1.39	Grelet et al. 2017
RFI2	kg/d	1,013	0.46	0.43	2.77	2.89	1.31	Grelet et al. 2017
SFA	g/dL of milk	1,790	0.99	0.99	0.07	0.07	10.22	Grelet et al. 2014
C18:1 <i>cis</i> -9	g/dL of milk	1,781	0.95	0.95	0.06	0.06	4.35	Grelet et al. 2014
MUFA	g/dL of milk	1,793	0.97	0.97	0.06	0.06	5.83	Grelet et al. 2014
PUFA	g/dL of milk	1,799	0.78	0.77	0.02	0.02	2.10	Grelet et al. 2014

R_c^2 = coefficient of determination of calibration; R_{cv}^2 = coefficient of determination of cross validation;
 SE_c = standard error in calibration; SE_{cv} = standard error in cross validation; RPD_{cv} = ratio of
performance to deviation in cross validation. EB = energy balance; DMI2 = predicted DMI; NUE =
nitrogen use efficiency; RFI = residual feed intake (alternate calculation via RFI2); C18:1 *cis*-9 = oleic acid

3.4. Material and methods

All data treatments and analyses were performed using R software (version 4.0.2; www.r-project.org). All data were collected as per standard routine by the Walloon Breeders Association (AWE), which is accredited to perform such collection.

3.4.1. Calibration Data

The data set to be used with unsupervised learning methods, hereafter called the calibration data set, contained 740,454 DHI records collected by Elevéo (AWE groupe, Ciney, Belgium) in 1,077 herds located in southern Belgium between January 2012 and March 2020 from 114,536 primiparous Holstein cows with Days In Milk (DIM) ranging

between 5 and 365 d. Milk analysis was performed by the milk laboratory Comité du Lait (Battice, Belgium) on Foss Milkoscan FT6000, FT+, and FT7 standardized spectrometers. Among the large collection of possible phenotypes for this study, we kept only milk yield, milk FT-MIR spectra, SCC, and contents of fat, protein, urea, and lactose predicted by the spectrometer. Furthermore, predictive equations developed from previous studies Table 3.2 were applied to the recorded spectral data to predict 20 additional biomarkers related to health issues. Although initially predicted in grams per deciliter (g/dL) of milk, the contents of Saturated Fatty Acid (SFA), oleic acid (C18:1cis9), Monounsaturated Fatty Qcid (MUFA), and Polyunsaturated Fatty Acid (PUFA) were expressed in grams per 100 grams of fat using the fat content predicted by the spectrometer, as this unit is more relevant for assessing the biological pathways behind fatty acid production. Then, Dry Matter Intake (DMI) and Fat-and Protein-Corrected Milk (FPCM) were computed using the following equations (National Research Council 2001; Yan et al. 2011):

$$DMI = (0.372FPCM + 0.0968BW^{0.75})(1 - \exp(-0.192WOL - 0.70746)) \quad (3.1)$$

$$FPCM = (0.337 + 0.116FAT)MY + 0.06PROT \quad (3.2)$$

where BW is the predicted BW using the equation of Tedde et al. (2021), WOL is the week of lactation, MY is milk yield, and FAT and PROT are fat and protein contents expressed as percentages. The consumption index was defined as the ratio of DMI to FPCM. Finally, the calibration data set used for the clustering contained 29 traits that were directly or indirectly related to health issues (Table 3.3).

3.4.2. Clustering Method

Patterns were found in the calibration data set using hierarchical clustering. This agglomerative algorithm works step by step to combine the closest records or groups of records until the whole data set is considered. Ward's agglomerative hierarchical clustering method with Euclidean distance was chosen because this is widely used to distinguish groups in a multivariate Euclidean space (Ward 1963). However, this type of clustering has the constraint of requiring the computing of a distance matrix between every observation or cluster at each step. So, for n records, the algorithm computes $n * (n - 1)/2$ distances. For 740,454 records, this represents 274×10^9 distances, which would make the computation too slow, and storing the distance matrix in memory would be impossible with standard computational infrastructure. The volume of data is a common hindrance in big data methodologies, and many approaches exist to solve this issue (Wang et al. 2016). In this study, we decided to apply a divide-and-conquer approach, which consisted of dividing the data into more

manageable subsets, analyzing them separately, and recombining them afterward. The calibration data were first split randomly into 100 subsets of around 7,500 records. For each subset, the records were clustered, and the obtained dendrogram was cut to divide the subset into 500 groups. A mean sample (i.e., a centroid) was created for each group by averaging all features. Therefore, 50,000 centroids (i.e., 500 centroids from 100 subsets) were computed to summarize the information contained in the initial data set. Those centroids were then clustered using the same hierarchical clustering method. As the approach depends on the splitting of the data set, the procedure was replicated 10 times to estimate the robustness of the final clustering obtained. From this final clustering, the inertia (i.e., the ratio between the between-class and within-class variance) of the dendrogram was estimated to define the best number of groups (i.e., selection of several groups before a large drop of inertia gain). This can be done by looking at the graph of the inertia according to the number of groups.

3.4.3. Cluster Interpretation

Standard statistical methods to test the significance of the effects, such as Analysis of Variance (ANOVA), are useless here because of the number of records. There are too many observations, so the tests are too powerful, making every effect significant. Moreover, the main objective was to interpret the clusters to understand what they represent. Thus, the differences between clusters were represented by calculating their Least Squares Means (LSM) per group, which were visualized using bar plots on a standardized scale.

3.4.4. Expanding the Result to the Original Data Set

For the previous step, the large data set was simplified into 50,000 centroids clustered into several groups. To expand this result to the whole data set, a cluster prediction model was needed. A PLS-DA model was built from the centroids using the 29 biomarkers. It is a predicting discriminant model using linear combinations of the original variables called latent variables, which are optimized to best represent the variability of the response and the original variables. The output of the model is the probability of belonging to a specific cluster. The hyperparameter of the model is the number of latent variables, which was fixed based on the performances obtained from a 10-fold crossvalidation. As our aim was to develop a model able to predict every cluster, but more specifically the clusters associated with sick animals, we compared the global accuracy and the balanced accuracy for each cluster, which is the mean between sensitivity and specificity. The PLS-DA model is sensitive to unbalanced data as it will most likely predict the most observed cluster. In a case where the proportions of observations in each cluster vary widely, it is recommended to generate a balanced training data set. The considered up-sampling approach consists of sampling with replacement of the original data set but with the constraint to reach 10,000 records for

each cluster. Moreover, we expected that a quantitative trait would better represent the continuum of the animal health status instead of a binary division (healthy vs. sick). Compared with the k-nearest neighbors algorithm, PLS-DA gives more quantitative information: the scores on specific latent variables and the probability of being in a cluster.

3.4.5. Validation

The innovative aspect of our approach was to use a large-scale database to find patterns potentially related to animal health to keep the maximum variability existing in the population. Therefore, no reference measurements for health status were used. This is why validation was performed using reference biomarkers measured in milk and blood samples. This approach aimed to validate the interest of obtained clusters and the potential of an associated quantitative trait such as a PLS-DA probability of belonging to a cluster.

The validation health data set was created from the European Union FP7 Genotype plus Environment project (<http://www.gpluse.eu>). The data set contained 380 records collected between 2013 and 2018 for primiparous and multiparous Holsteins with DIM between 10 and 50. As the clusters are defined from primiparous Holsteins, a first data set containing only primiparous cows was created but only 50 records were usable as SCC measures were missing. Therefore, a second data set containing 204 records from primiparous and multiparous cows was developed to increase the number of observations available. Several biomarkers were measured in individual milk and blood samples: cholesterol, fructosamine, BHB, Non Esterified Fatty Acids (NEFA), Progesterone (P4), and glucose in blood and glucose-6-phosphate, free glucose, isocitrate, urea, N-acetyl- β -D-glucosaminidase (NAGase), and Lactate Deshydrogenase (LDH) in milk (Table 3.4). De Koster et al. (2019) describe the data collection and the applied protocol in more detail. All phenotypes included in the calibration data set were also collected to allow prediction of the clusters, PLS-DA probabilities, and scores on latent variables for those observations.

To validate the relationships between the probabilities or the scores on latent variables and the reference values measured in the laboratory, Pearson correlations were calculated between those quantitative and reference traits. Then, statistical t-tests for association between paired samples were performed to assess the significance of the estimated correlation values.

Table 3.3: Descriptive statistics of the mid-infrared-based predictors, milk yield, and SCC from the calibration data set ($n = 740,454$ records).

Abbreviation	Unit	Minimum	Maximum	Mean	SD	CV
Milk	kg/d	8.10	78.00	23.72	5.91	0.25
pFAT	g/dL of milk	21.10	69.90	39.66	5.99	0.15
pPROT	g/dL of milk	20.30	51.50	33.59	3.27	0.10
Lactose	g/dL of milk	3.53	5.68	4.84	0.17	0.03
Urea	mmol/L	10.00	1,260.00	249.32	79.66	0.32
SCC	cells $\times$ 1,000/mL	10.00	3,490.00	139.60	297.49	2.13
Weight	kg	401.49	934.12	620.16	65.20	0.11
CH ₄	g/d	100.00	699.99	306.31	58.38	0.19
Na	mg/kg of milk	200.02	499.98	321.90	34.52	0.11
Ca	mg/kg of milk	900.00	1,500.00	1,154.47	97.32	0.08
P	mg/kg of milk	700.62	1,299.95	997.53	83.77	0.08
Mg	mg/kg of milk	80.00	130.00	100.68	7.73	0.08
K	mg/kg of milk	1,100.03	1,700.00	1,494.43	90.87	0.06
Lactoferrin	mg/L	35.11	800.34	162.11	88.15	0.54
DMI2	kg/d	14.00	33.00	24.50	3.20	0.13
DMI	kg/d	7.65	38.51	19.85	2.67	0.13
EB	%	-19.38	10.00	0.33	3.54	10.89
NUE	%	-2.37	27.10	11.34	2.14	0.19
Citrate	mmol/L	2.84	17.00	8.64	1.36	0.16
Acetone	mmol/L	0.01	0.30	0.06	0.03	0.46
BHB	μ mol/L	38.20	399.82	158.07	42.24	0.27
RFI	kg/d	-16.09	1.50	-3.46	2.18	-0.63
RFI2	kg/d	-13.44	5.68	-2.10	2.02	-0.96
FPCM	kg/d	6.50	74.25	23.52	5.53	0.23
CI	kg of DMI/kg	0.36	2.24	0.88	0.18	0.21
SFA	g/100 g of fat	30.50	178.75	70.39	4.66	0.07
C18:1 <i>cis</i> -9	g/100 g of fat	7.67	47.63	20.47	3.68	0.18
MUFA	g/100 g of fat	10.39	66.65	28.17	4.21	0.15
PUFA	g/100 g of fat	1.26	8.81	3.70	0.57	0.15

pFAT = fat content; pPROT = protein content; DMI2 = predicted DMI; EB = energy balance; NUE = nitrogen use efficiency; RFI = residual feed intake (alternate calculation via RFI2); FPCM = fat- and protein-corrected milk; CI = consumption index; C18:1 *cis*-9 = oleic acid.

Table 3.4: Descriptive statistics for the reference measurements of health-related traits from the validation sets

Measure ¹	Unit	Lactation 1 (n = 50)				All lactations (n = 204)			
		Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD
<i>Blood</i>									
Cholesterol	g/L	1.64	5.57	3.22	0.89	1.64	7.33	3.70	1.12
Fructosamine	μmol/L	192.83	296.07	263.71	19.81	192.83	298.77	259.62	18.35
BHB	mmol/L	0.24	2.14	0.57	0.40	0.21	4.29	0.63	0.48
LogBHB	log(mmol/L)	−0.62	0.33	−0.31	0.22	−0.68	0.63	−0.27	0.22
NEFA	μEq/L	95.77	1,884.36	518.48	422.37	26.07	2,757.01	606.17	424.85
P4	ng/mL	0.12	11.07	1.89	2.23	0.02	12.92	2.48	2.80
Glucose	mmol/L	2.79	4.70	3.89	0.39	2.02	4.70	3.61	0.41
<i>Milk</i>									
Glu6P	mmol/L	0.05	0.44	0.20	0.09	0.05	0.44	0.17	0.08
GluFree	mmol/L	0.08	0.41	0.23	0.09	0.03	0.58	0.23	0.10
BHB	μmol/L	20.00	241.00	61.00	44.81	20	735	65.33	60.97
Isocitrate	mmol/L	0.09	0.32	0.19	0.05	0.07	0.32	0.17	0.05
Urea	mmol/L	0.95	6.80	3.37	1.46	0.47	7.17	3.11	1.38
NAGase	units/L	0.56	6.67	2.18	1.18	0.56	10.78	1.92	1.08
LDH	units/L	0.78	8.02	3.27	1.68	0.64	39.88	3.10	3.23

¹ NEFA = nonesterified fatty acid; P4 = progesterone; Glu6P = glucose-6-phosphate; GluFree = free glucose; NAGase = N-acetyl-β-d-glucosaminidase; LDH = lactate dehydrogenase.

3.5. Results and discussion

3.5.1. Developing the Clusters

As mentioned previously, standard hierarchical clustering methodologies could not be used and a specific approach was developed. As this approach involved randomness, the process was repeated 10 times to assess its robustness. To determine the similarity, we looked at whether the groups obtained and the division order between them were equivalent. Nine dendrograms out of 10 repetitions were similar. Thus, only one representative dendrogram is shown in Figure 3.1. Based on inertia, the dendrogram was cut into 5 groups. The graph is structured into 2 large groups divided into one sub-group containing cluster 1 and cluster 4 and another sub-group divided into 3 sub-groups (clusters 2, 3, and 5; Figure 3.1). Among these 3 sub-groups, cluster 5 was very small and well separated from the others. The distribution of the centroids among the clusters was uneven, with 15,271 centroids in cluster 1 (30.54%), 19,573 in cluster 2 (39.15%), 5,452 in cluster 3 (10.90%), 6,856 in cluster 4 (13.71%), and 2,848 in cluster 5 (5.70%). Clusters 1 and 2 covered most of the records, with 69.69% of observations.

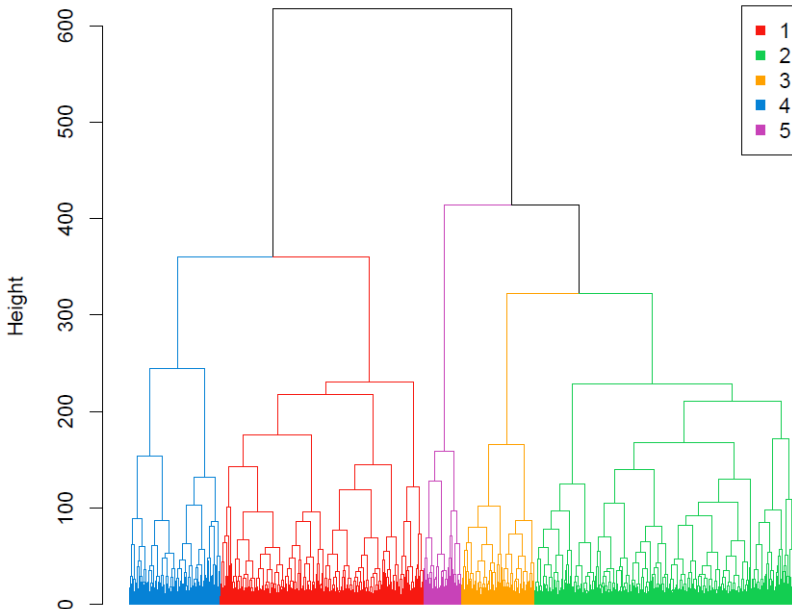


Figure 3.1: Dendrograms for the 50,000 centroids. Height indicates the value of the Ward's criterion for each agglomeration of centroids or groups of centroids. Agglomerations with low heights represent centroids, or groups of centroids, that are close. Clusters are identified by color.

As mentioned earlier, testing for the significance of the effect is irrelevant here.

Instead, the LSM of the 29 studied traits were computed and are illustrated in Figure 3.2 on a standardized scale. From the visual analysis, 2 clusters differed from the others: cluster 4 and cluster 5. Clusters 1 and 4 had similar behavior, with both means of each prediction being positive or negative. However, the means of cluster 4 were larger than those of cluster 1, suggesting that clusters 1 and 4 represented the same biological phenomena but with severe cases being in cluster 4. Those groups were characterized by negative values for the predictions of protein content, body weight, methane, phosphorus, DMI, Energy Balance (EB), residual feed intake, consumption index, and SFA. There were positive means for the predictions of sodium, nitrogen use efficiency, citrate, acetone, BHB, and the predicted percentages of C18:1 cis-9, MUFA, and PUFA in milk fat. Based on those results, we hypothesized that cluster 4 referred to cows with metabolic disorders induced by negative EB, up to ketosis. Although we mainly used predictions in this study, the results are similar to those of Gross et al. (2011), who observed reduced milk production with a lower percentage of protein, combined with a lower BW and DMI for Holstein cows in negative EB. Moreover, cluster 4 was related to lower content of SFA and higher contents of MUFA and PUFA. Vranković et al. (2017) and Fiore et al. (2020) showed that similar changes in milk fatty acid composition, with decrease of de novo fatty acids, are associated with negative EB and body fat mobilization.

Cluster 5 was characterized mostly by a large positive value for SCC. Moreover, Na and lactoferrin means were high, and that of lactose was low, suggesting that cluster 5 was related to mastitis issues (Figure 3.2). Although we used predictions for Na and lactose, these findings can be corroborated by previous works. Bansal et al. (2005) made the same observation for cows with subclinical mastitis using reference measurements. Fernando et al. (1985) also mentioned the use of Na as an indicator of subclinical mastitis. Lactose as an indicator of subclinical mastitis was identified by Ebrahimie et al. (2018). Cluster 3 was defined by low production of milk associated with high estimated means for fat, protein, and minerals.

3.5.2. Predicting the Cluster for the Calibration Data Set

At this stage, we discovered 2 potentially interesting groups for disease detection (i.e., cluster 4 and cluster 5). However, those clusters were only defined using the centroids. Therefore, we needed to expand the results of the clustering to the whole data set. As each of the centroids was defined from 500 groups for each of the 100 subsets, this could be performed by recovering the records for each centroid. However, this method prevents the extrapolation of the results to future records. Instead, a PLS-DA model was built to allow discrimination of the found clusters from the 50,000 centroids. As the number of centroids by cluster was greatly unbalanced, which risks impairing the predictions, we used an up-sampling approach based on bootstrapping, wherein we sampled with higher probabilities for the smaller cluster. As a result, the training

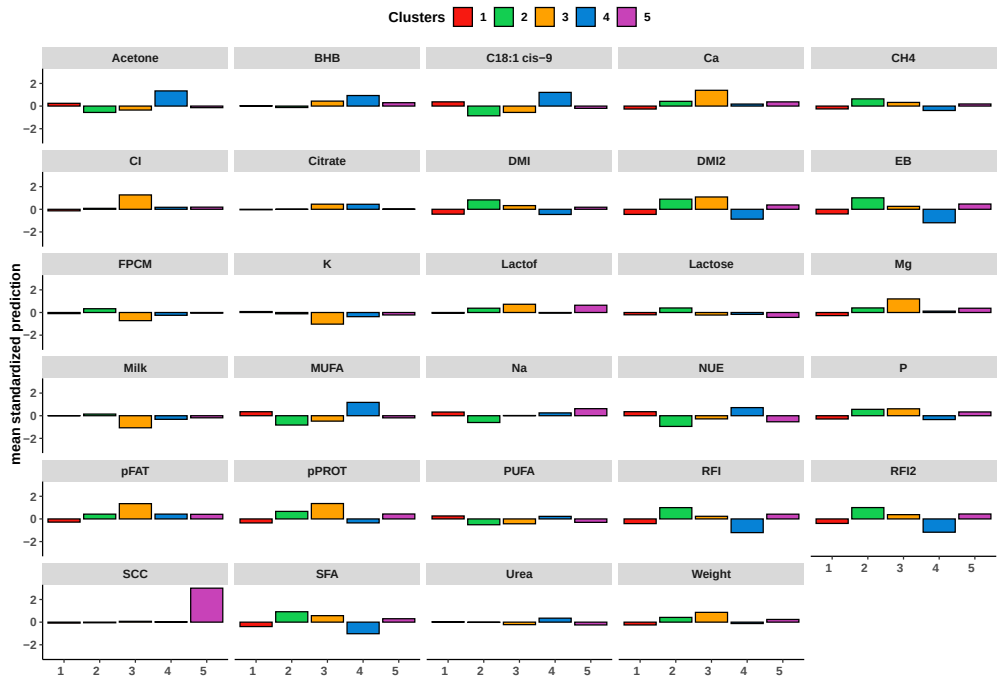


Figure 3.2: Bar plots of the least squares means of the 29 biomarkers for the 50,000 centroids.

C18:1 cis-9 = oleic acid; CI = consumption index; EB = energy balance; FPCM = fat- and protein-corrected milk; Lactof = lactoferrin; NUE = nitrogen use efficiency; pPROT = protein content; pFAT = fat content; RFI = residual feed intake (alternate calculation via RFI2); DMI2 = predicted DMI.

data set was balanced with around the same number of observations for each cluster, creating a model that was not biased toward the prediction of the most represented clusters. The number of latent variables was fixed at 6, as the highest overall cross-validation accuracy was found for this model (0.77). The accuracy decreased to 0.70 when the model was applied to the original centroid data. The balanced accuracy of discriminating cluster 4 and cluster 5 was 0.88 and 0.96, respectively. With use of the PLS-DA model, we were able to extend the prediction of clusters to the entire data set. We observed that cluster 1 was assigned to 21.39% of the records, cluster 2 to 40.89%, cluster 3 to 18.01%, cluster 4 to 16.62%, and cluster 5 to 3.1%. This repartition is more or less the same as the repartition observed using the 50,000 centroids, confirming the model, with a lower percentage of records belonging to cluster 4 and cluster 5. This is expected as those groups should be related to cows with metabolic disorders or mastitis. Although we had clusters for all records, these categorical variables do not fully represent the reality of a sick animal. Animal health status is a continuous variable from healthy to sick. By observing the PLS-DA latent variables, we can observe this continuum and the overlap between clusters (Figure 3.3). Therefore, the probabilities of being in cluster 4 or cluster 5 seemed to be more relevant compared with the previous hard clustering, where each observation belongs to only one cluster. However, it is important to verify whether the found clusters were still pertinent using reference biomarker measurements.

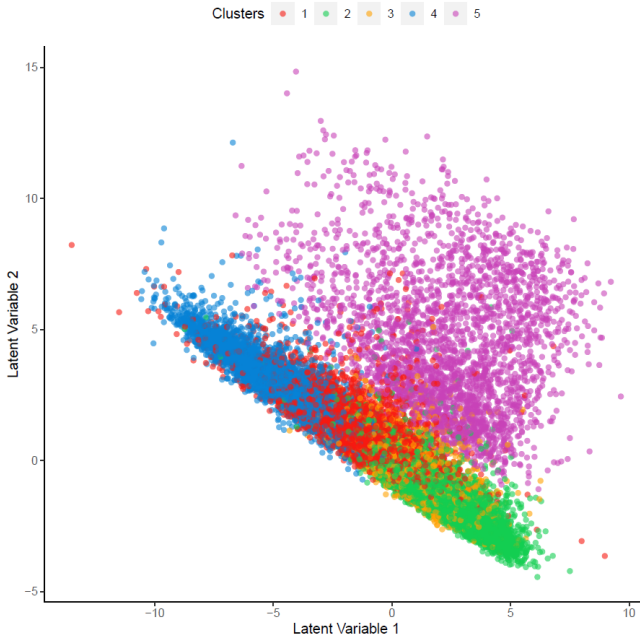


Figure 3.3: Scatterplot representing the distribution of the 740,454 records on the first and second latent variables from the partial least squares discriminant analysis.

3.5.3. Validating Against Reference Values

Based on the GplusE project, it was possible to create a data set containing reference measurements of biomarkers related to animal health. As all predictors needed to predict the clusters were available for these records, we were able to estimate for each record the probability of being in a cluster from the PLS-DA (Table 3.5). As the variabilities from the data set composed of primiparous Holsteins were low, especially for the probability related to cluster 5, we assume that this small data set contained very few cases of mastitis. If we consider a proportion of 6% for cluster 5, we would expect 3 observations for this cluster. To solve this issue, we extended the result to subsequent lactations. The extension allowed the consideration of a larger number of observations, where more cows with health issues could be expected. This example shows the difficulty in finding representative reference data when working with health data.

Table 3.5: Descriptive statistics for the probability of being in a cluster for the validation records

Measure	Lactation 1 (n = 50)				All lactations (n = 204)			
	Min	Max	Mean	SD	Min	Max	Mean	SD
Cluster 1	0.12	0.31	0.25	0.03	0.03	0.35	0.25	0.04
Cluster 2	0.09	0.27	0.18	0.04	0.01	0.33	0.19	0.05
Cluster 3	0.12	0.39	0.33	0.05	0.03	0.33	0.18	0.04
Cluster 4	0.12	0.55	0.22	0.09	0.04	0.55	0.20	0.08
Cluster 5	0.12	0.27	0.18	0.02	0.12	0.88	0.18	0.09

As all of the PLS-DA probabilities were quantitative, Pearson correlations could be computed between these outputs and the reference contents of biomarkers measured in blood and milk. Table 3.6 shows these correlations and their significance. The correlations confirmed the previous interpretation of the clusters. Among the significant correlations, the probability associated with cluster 4 was positively correlated with blood and milk BHB, isocitrate, and NEFA and negatively correlated with blood fructosamine, blood glucose, milk free glucose, and milk glucose-6-phosphate. As reported in many studies, these biomarkers reflect EB issues up to ketosis through a lower level of circulating glucose in blood, reflected in the long term through a lower content of fructosamine, and leading to release of NEFA because of body fat mobilization and ultimately production of ketone bodies in the liver (Ingvarsen 2006; Suthar et al. 2013; Esposito et al. 2014). Larsen (2014) and Larsen et al. (2015) also proved that free glucose, glucose-6-phosphate, and isocitrate are related to EB issues or risk of diseases.

For the correlation between the probability of belonging to cluster 5 and the biomark-

ers, we observed that they differed between the first and the subsequent lactations. We assumed that this was related to the low variability in the first-lactation data set, suggesting that there were no severe mastitis issues in the 50 observations. However, we do not have the on-farm information about mastitis so we cannot validate this assumption. Thus, we also considered the correlations from the other data set. The probability of belonging to cluster 5 was negatively correlated with free glucose in milk but not in blood, and positively correlated with NAGase and LDH. Similarly, Larsen et al. (2010) and Kitchen et al. (1980) showed that LDH and NAGase were relevant indicators of mastitis.

As shown by the validation step based on reference laboratory data, the probabilities were strongly correlated with relevant biomarkers of health issues. Thus, our results highlighted that even using predictions from models with low or fair quality (9 models had an $R_{cv}^2 < 0.7$), trends and patterns can be found in a large data set and validated through external data. As mentioned by McParland et al. (2016) or Bonfatti et al. (2017), the current results validate the idea that models with low or fair accuracy can be valuable for breeding and monitoring if they can highlight trends or genetic correlation through large-scale data sets. Moreover, the measures could be exploited in a longitudinal approach by computing the variance, the mean value for lactation, or the expected curves. The measures, frequencies, and means of measures could be indicators of the resilience of the animals. However, the method requires more frequently collected spectral data, not a monthly test, to be used as a continuous monitoring tool. The current approach has the limitation of requiring previous biomarker models to constitute the clusters.

However, the cluster prediction method, being based only on DHI records, opens up many avenues to use already available data. The method offers a way to summarize information from multiple models, to find consistency between all the predictions, and ultimately to detect health patterns in large databases.

Table 3.6: Correlations between the probability of being in a cluster and reference measurements of biomarkers related to animal health

Reference	Lactation 1 (n = 50)					All lactations (n = 204)				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
<i>Blood</i>										
Cholesterol	0.05	0.33*	-0.17	-0.09	0.02	0.01	0.28***	-0.12	-0.06	-0.07
Fructosamine	0.00	0.09	0.13	-0.15	0.19	0.03	0.18*	0.01	-0.19**	0.03
BHB	-0.51***	-0.46***	-0.38**	0.70***	-0.40**	-0.34***	-0.31***	-0.10	0.54***	-0.06
NEFA	-0.25	-0.45***	-0.55***	0.67***	-0.36**	-0.14*	-0.39***	-0.13	0.44***	-0.01
P4	0.13	0.21	-0.05	-0.09	-0.11	0.06	0.07	-0.13	0.05	-0.06
Glucose	0.25	0.49***	0.35*	-0.60***	0.42**	0.25***	0.28***	0.08	-0.32***	-0.06
<i>Milk</i>										
Glu6P	0.12	-0.05	0.45***	-0.32*	0.29*	-0.08	-0.08	0.28***	-0.23***	0.16*
GluFree	0.04	0.23	-0.45*	0.16	-0.26	0.31***	0.45***	-0.28***	-0.15*	-0.18**
BHB	-0.60***	-0.34*	-0.17	0.53***	-0.21	-0.34***	-0.24***	-0.04	0.35***	0.04
Isocitrate	-0.46***	-0.61***	0.01	0.50***	-0.16	-0.32***	-0.34***	0.32***	0.38***	-0.10
Urea	-0.15	-0.16	-0.29*	0.32*	-0.15	0.05	0.12	-0.22**	0.03	-0.03
NAGase	-0.09	-0.36**	0.19	0.10	0.02	-0.39***	-0.46***	-0.07	0.06	0.46***
LDH	-0.18	-0.29*	0.34*	-0.07	0.43**	-0.52***	-0.39***	-0.16*	-0.13	0.69***

¹ NEFA = nonesterified fatty acid; P4 = progesterone; Glu6P = glucose-6-phosphate; GluFree = free glucose; NAGase = N-acetyl- β -d-glucosaminidase; LDH = lactate dehydrogenase.

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

3.6. Conclusions

This research aimed to consider the multidimensionality of animal health and its associated traits to detect potential health issues in a routine environment. Instead of using reference laboratory data, the innovative aspect of this study was to make use of a DHI data set containing records related to usual dairy traits and FT-MIR-based predictions that have direct or indirect links with health issues, to discover groups of cows with different data patterns. Even though the method had some computational and validation limitations, this novel approach allowed detection of global patterns from a large data set that is more representative of the population. The results are promising, showing that the approach is worth developing. Two clusters were shown to represent cows with metabolic disorders or that have mastitis issues. From the probability of belonging to 1 of those 2 clusters estimated through a PLS-DA model, we propose a quantitative gradient of sickness related to metabolic disorders or mastitis. These traits are directly usable in many applications; for example, as a tool to monitor the animal health status of every productive cow during routine milk recording or to develop new estimated breeding values due to the large amount of directly available data.

3.7. Acknowledgments

The authors acknowledge the support of the Walloon Government (Service Public de Wallonie, Namur, Belgium) through the ScorWelCow project (Grant agreement D31-1390) and their co-financing of the INTERREG NWE HappyMoo project (Grant agreement NWE 730), which is also acknowledged for its support of this research. The GplusE project has received funding from the European Union's Seventh Framework Program (Brussels, Belgium) for research, technological development, and demonstration, under Grant agreement no. 613689. The views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. Also acknowledged are the use of the computation resources of the University of Liège–Gembloux Agro-Bio Tech (Gembloux, Belgium) provided by the technical platform Calcul et Modélisation Informatique (CAMI) of the TERRA Teaching and Research Centre, partly supported by the National Fund for Scientific Research (F.R.S.-FNRS, Brussels, Belgium) under Grants No. T.0095.19 (PDR “DEEPSELECT”) and J.0174.18 (CDR “PREDICT-2”). The authors have not stated any conflicts of interest.

3.8. Unpublished Supplementary Results

This publication does not mention the prediction of those clusters or the prediction of the cluster membership (i.e. the probability) through supervised model using FT-

MIR. However, it was performed *a posteriori* as a preliminary study for the Chapter 4. While the prediction of the metabolic disorder cluster is discussed in Chapter 4, the results are not mentioned for the udder health cluster. The performances in prediction for this cluster are highly dependent on the information given as input of the model (Table 3.7). The SCC seems to contain most of the information even though the addition of the spectra increases slightly the performance.

Table 3.7: 10-fold cross-validation results using `caret` and 212 spectral wavelengths (Results not published).

Model	Variables	LV	$R^2_{cv} \pm SE$
LM	SCC	/	0.9302 ± 0.0009
PLS	FT-MIR and SCC	15	0.9882 ± 0.0005
PLS	FT-MIR	15	0.1841 ± 0.0031

LV = number of latent variables, SE = Standard Error, LM = Linear Model.

Bibliography

- Atashi, H., M. Salavati, J. De Koster, M. A. Crowe, G. Opsomer, and M. Hostens (2020). “Genome-wide association for metabolic clusters in early-lactation Holstein dairy cows”. In: *Journal of Dairy Science* 103.7, pp. 6392–6406. ISSN: 0022-0302. DOI: 10.3168/jds.2019-17369.
- Bastin, C., L. Théron, A. Lainé, and N. Gengler (2016). “On the role of mid-infrared predicted phenotypes in fertility and health dairy breeding programs¹”. In: *Journal of Dairy Science* 99.5, pp. 4080–4094. ISSN: 0022-0302. DOI: 10.3168/jds.2015-10087.
- Belay, T. K., B. S. Dagnachew, Z. M. Kowalski, and T. Ådnøy (2017). “An attempt at predicting blood β -hydroxybutyrate from Fourier-transform mid-infrared spectra of milk using multivariate mixed models in Polish dairy cattle”. In: *Journal of Dairy Science* 100.8, pp. 6312–6326. ISSN: 0022-0302. DOI: 10.3168/jds.2016-12252.
- Blokhuis, H. J., R. B. Jones, R. Geers, M. Miele, and I. Veissier (2003). “Measuring and monitoring animal welfare: Transparency in the food product quality chain”. In: *Animal Welfare* 12, pp. 445–455.
- Bonfatti, V., S.-A. Turner, B. Kuhn-Sherlock, T. D. W. Luke, P. N. Ho, C. V. C. Phyn, and J. E. Pryce (2019). “Prediction of blood β -hydroxybutyrate content and occurrence of hyperketonemia in early-lactation, pasture-grazed dairy cows using milk infrared spectra”. In: *Journal of Dairy Science* 102.7, pp. 6466–6476. ISSN: 1525-3198. DOI: 10.3168/jds.2018-15988.
- Bonfatti, V., D. Vicario, A. Lugo, and P. Carnier (2017). “Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle”. In: *Journal of Dairy Science* 100.7, pp. 5526–5540. ISSN: 0022-0302. DOI: 10.3168/jds.2016-11667.
- Bresolin, T. and J. R. R. Dórea (2020). “Infrared Spectrometry as a High-Throughput Phenotyping Technology to Predict Complex Traits in Livestock Systems”. In: *Frontiers in Genetics* 11. ISSN: 1664-8021.
- Christophe, O. S., C. Grelet, C. Bertozzi, D. Veselko, C. Lecomte, P. Höeckels, A. Werner, F.-J. Auer, N. Gengler, F. Dehareng, and H. Soyeurt (2021). “Multiple Breeds and Countries’ Predictions of Mineral Contents in Milk from Milk Mid-

- Infrared Spectrometry". In: *Foods* 10.9. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 2235. DOI: 10.3390/foods10092235.
- De Koster, J., M. Salavati, C. Grelet, M. A. Crowe, E. Matthews, R. O'Flaherty, G. Opsomer, L. Foldager, and M. Hostens (2019). "Prediction of metabolic clusters in early-lactation dairy cows using models based on milk biomarkers". In: *Journal of Dairy Science* 102.3, pp. 2631–2644. ISSN: 0022-0302. DOI: 10.3168/jds.2018-15533.
- De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa (2014). "Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits". In: *Journal of Dairy Science* 97.3. Publisher: Elsevier, pp. 1171–1186. ISSN: 0022-0302. DOI: 10.3168/jds.2013-6799.
- Ebrahimie, E., F. Ebrahimi, M. Ebrahimi, S. Tomlinson, and K. R. Petrovski (2018). "A large-scale study of indicators of sub-clinical mastitis in dairy cattle by attribute weighting analysis of milk composition features: highlighting the predictive power of lactose and electrical conductivity". In: *Journal of Dairy Research* 85.2. Publisher: Cambridge University Press, pp. 193–200. ISSN: 0022-0299, 1469-7629. DOI: 10.1017/S0022029918000249.
- Esposito, G., P. C. Irons, E. C. Webb, and A. Chapwanya (2014). "Interactions between negative energy balance, metabolic diseases, uterine health and immune response in transition dairy cows". In: *Animal Reproduction Science* 144.3, pp. 60–71. ISSN: 0378-4320. DOI: 10.1016/j.anireprosci.2013.11.007.
- Fernando, R. S., S. L. Spahr, and E. H. Jaster (1985). "Comparison of Electrical Conductivity of Milk with Other Indirect Methods for Detection of Subclinical Mastitis". In: *Journal of Dairy Science* 68.2, pp. 449–456. ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(85)80844-4.
- Fiore, E., F. Blasi, M. Morgante, L. Cossignani, T. Badon, M. Ganesella, B. Contiero, and M. Berlanda (2020). "Changes of milk fatty acid composition in four lipid classes as biomarkers for the diagnosis of bovine ketosis using bioanalytical Thin Layer Chromatography and Gas Chromatographic techniques (TLC-GC)". In: *Journal of Pharmaceutical and Biomedical Analysis* 188, p. 113372. ISSN: 0731-7085. DOI: 10.1016/j.jpba.2020.113372.
- Fleming, A., F. S. Schenkel, J. Chen, F. Malchiodi, V. Bonfatti, R. A. Ali, B. Mallard, M. Corredig, and F. Miglior (2017). "Prediction of milk fatty acid content with mid-infrared spectroscopy in Canadian dairy cattle using differently distributed model development sets". In: *Journal of Dairy Science* 100.6, pp. 5073–5081. ISSN: 0022-0302. DOI: 10.3168/jds.2016-12102.
- Foldager, L., C. Gaillard, M. T. Sorensen, T. Larsen, E. Matthews, R. O'Flaherty, F. Carter, M. A. Crowe, C. Grelet, M. Salavati, M. Hostens, K. L. Ingvarlsen, and M. A. Krogh (2020). "Predicting physiological imbalance in Holstein dairy cows by three different sets of milk biomarkers". In: *Preventive Veterinary Medicine* 179. ISSN: 0167-5877. DOI: 10.1016/j.prevetmed.2020.105006.

- Fraser, D. (1995). "Science, Values and Animal Welfare: Exploring the 'Inextricable Connection'". In: *Animal Welfare* 4.2, pp. 103–117.
- Gandomi, A. and M. Haider (2015). "Beyond the hype: Big data concepts, methods, and analytics". In: *International Journal of Information Management* 35.2, pp. 137–144. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2014.10.007.
- Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M. -. Vanrobays, A. Lainé, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne (2016). "Capitalizing on fine milk composition for breeding and management of dairy cows". In: *Journal of Dairy Science* 99.5, pp. 4071–4079. ISSN: 0022-0302. DOI: 10.3168/jds.2015-10140.
- Grelet, C., C. Bastin, M. Gelé, J. -. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng (2016). "Development of Fourier transform mid-infrared calibrations to predict acetone, β -hydroxybutyrate, and citrate contents in bovine milk through a European dairy network". In: *Journal of Dairy Science* 99.6, pp. 4816–4825. ISSN: 0022-0302. DOI: 10.3168/jds.2015-10477.
- Grelet, C., E. Froidmont, L. Foldager, M. Salavati, M. Hostens, C. P. Ferris, K. L. Ingvarlsen, M. A. Crowe, M. T. Sorensen, J. A. Fernandez Pierna, A. Vanlierde, N. Gengler, and F. Dehareng (2020). "Potential of milk mid-infrared spectra to predict nitrogen use efficiency of individual dairy cows in early lactation". In: *Journal of Dairy Science* 103.5, pp. 4435–4445. ISSN: 0022-0302. DOI: 10.3168/jds.2019-17910.
- Grelet, C., A. Vanlierde, M. Hostens, L. Foldager, M. Salavati, K. L. Ingvarlsen, M. Crowe, M. T. Sorensen, E. Froidmont, C. P. Ferris, C. Marchitelli, F. Becker, T. Larsen, F. Carter, and F. Dehareng (2019). "Potential of milk mid-IR spectra to predict metabolic status of cows through blood components and an innovative clustering approach". In: *Animal* 13.3, pp. 649–658. ISSN: 1751-7311. DOI: 10.1017/S1751731118001751.
- Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Steevens, N. Gengler, and F. Dehareng (2021). "Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions". In: *Methods*. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2020.07.012.
- Grelet, C., J. Fernandez Pierna, H. Soyeurt, F. Dehareng, N. Gengler, and P. Dardenne (2014). "Creation of universal MIR calibration by standardization of milk spectra: example of fatty acids". In: Wageningen Academic Publishers. ISBN: 978-90-8686-248-1.
- Grelet, C., A. Vanlierde, F. Dehareng, E. Froidmont, and G. Consortium (Aug. 30, 2017). "Prediction of energy status of dairy cows using MIR milk spectra". In: URL: <https://orbi.uliege.be/handle/2268/224000> (visited on 09/17/2021).
- Gross, J., van Dorl, H.A., Bruckmaier, R.M., Schwarz, and F.J. (2011). "Milk fatty acid profile related to energy balance in dairy cows". In: *J. Dairy Res.* 78 479–488. DOI: 10.1017/S0022029911000550.

- Ingvartsen, K. L. (2006). "Feeding- and management-related diseases in the transition cow: Physiological adaptations around calving and strategies to reduce feeding-related diseases". In: *Animal Feed Science and Technology*. Feed and Animal Health 126.3, pp. 175–213. ISSN: 0377-8401. DOI: 10.1016/j.anifeedsci.2005.08.003.
- Kitchen, B. J., G. Middleton, I. G. Durward, R. J. Andrews, and M. C. Salmon (1980). "Mastitis Diagnostic Tests to Estimate Mammary Gland Epithelial Cell Damage". In: *Journal of Dairy Science* 63.6, pp. 978–983. ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(80)83035-9.
- Krogh, M. A., M. Hostens, M. Salavati, C. Grelet, M. T. Sorensen, D. C. Wathes, C. P. Ferris, C. Marchitelli, F. Signorelli, F. Napolitano, F. Becker, T. Larsen, E. Matthews, F. Carter, A. Vanlierde, G. Opsomer, N. Gengler, F. Dehareng, M. A. Crowe, K. L. Ingvartsen, and L. Foldager (2020). "Between- and within-herd variation in blood and milk biomarkers in Holstein cows in early lactation". In: *Animal: An International Journal of Animal Bioscience* 14.5, pp. 1067–1075. ISSN: 1751-732X. DOI: 10.1017/S1751731119002659.
- Larsen, T. and K. M. Moyes (2015). "Are free glucose and glucose-6-phosphate in milk indicators of specific physiological states in the cow?" In: *Animal* 9.1, pp. 86–93. ISSN: 1751-7311. DOI: 10.1017/S1751731114002043.
- Larsen, T., C. M. Røntved, K. L. Ingvartsen, L. Vels, and M. Bjerring (2010). "Enzyme activity and acute phase proteins in milk utilized as indicators of acute clinical E. coli LPS-induced mastitis". In: *Animal* 4.10, pp. 1672–1679. ISSN: 1751-7311. DOI: 10.1017/S1751731110000947.
- Larsen, T. (2014). "Fluorometric determination of free and total isocitrate in bovine milk". In: *Journal of Dairy Science* 97.12, pp. 7498–7504. ISSN: 0022-0302. DOI: 10.3168/jds.2014-8018.
- Leblanc, S. (2010). "Monitoring Metabolic Health of Dairy Cattle in the Transition Period". In: *Journal of Reproduction and Development*. Journal of Reproduction and Development 56 (S), pp. 529–535. DOI: 10.1262/jrd.1056S29.
- Luke, T. D. W., S. Rochfort, W. J. Wales, V. Bonfatti, L. Maret, and J. E. Pryce (2019). "Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra". In: *Journal of Dairy Science*. Journal of Dairy Science 102.2, pp. 1747–1760. ISSN: 0022-0302. DOI: 10.3168/jds.2018-15103.
- McParland, S. and D. P. Berry (2016). "The potential of Fourier transform infrared spectroscopy of milk samples to predict energy intake and efficiency in dairy cows¹". In: *Journal of Dairy Science* 99.5, pp. 4056–4070. ISSN: 0022-0302. DOI: 10.3168/jds.2015-10051.
- National Research Council (2001). *Nutrient Requirements of Dairy Cattle: Seventh Revised Edition, 2001*. Washington, DC: The National Academies Press. 405 pp. ISBN: 978-0-309-06997-7. DOI: 10.17226/9825.

- Pralle, R. S., K. W. Weigel, and H. M. White (2018). "Predicting blood β -hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network". In: *Journal of Dairy Science* 101.5, pp. 4378–4387. ISSN: 0022-0302. DOI: 10.3168/jds.2017-14076.
- Smith, S. L., S. J. Denholm, M. P. Coffey, and E. Wall (2019). "Energy profiling of dairy cows from routine milk mid-infrared analysis". In: *Journal of Dairy Science* 102.12, pp. 11169–11179. ISSN: 0022-0302. DOI: 10.3168/jds.2018-16112.
- Soyeurt, H., C. Bastin, F. G. Colinet, V. M. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Scheifers, J. Vandenplas, K. Weigel, M. Coffey, L. Th  ron, J. Detilleux, E. Reding, N. Gengler, and S. McParland (2012). "Mid-infrared prediction of lactoferrin content in bovine milk: potential indicator of mastitis". In: *Animal*. Animal 6.11, pp. 1830–1838. ISSN: 1751-7311. DOI: 10.1017/S1751731112000791.
- Soyeurt, H., C. Grelet, S. McParland, M. Calmels, M. Coffey, A. Tedde, P. Delhez, F. Dehareng, and N. Gengler (2020). "A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra". In: *Journal of Dairy Science* 103.12, pp. 11585–11596. ISSN: 0022-0302. DOI: 10.3168/jds.2020-18870.
- Suthar, V. S., J. Canelas-Raposo, A. Deniz, and W. Heuwieser (2013). "Prevalence of subclinical ketosis and relationships with postpartum diseases in European dairy cows". In: *Journal of Dairy Science* 96.5, pp. 2925–2938. ISSN: 0022-0302. DOI: 10.3168/jds.2012-6035.
- Tedde, A., C. Grelet, P. N. Ho, J. E. Pryce, D. Hailemariam, Z. Wang, G. Plastow, N. Gengler, Y. Brostaux, E. Froidmont, F. Dehareng, C. Bertozzi, M. A. Crowe, I. Dufrasne, G. C. Group, and H. Soyeurt (2021). "Validation of Dairy Cow Bodyweight Prediction Using Traits Easily Recorded by Dairy Herd Improvement Organizations and Its Potential Improvement Using Feature Selection Algorithms". In: *Animals* 11.5. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 1288. DOI: 10.3390/ani11051288.
- Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt (2021). "Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra". In: *Journal of the Science of Food and Agriculture* 101.8, pp. 3394–3403. ISSN: 1097-0010. DOI: 10.1002/jsfa.10969.
- Vrankovi  , L., J. Aladrovi  , D. Octenjak, D. Bijeli  , L. Cvetni  , and Z. Stojevi   (2017). "Milk fatty acid composition as an indicator of energy status in Holstein dairy cows". In: *Archives Animal Breeding* 60.3. Publisher: Copernicus GmbH, pp. 205–212. ISSN: 0003-9438. DOI: 10.5194/aab-60-205-2017.

- Wang, C., M.-H. Chen, E. Schifano, J. Wu, and J. Yan (2016). "Statistical methods and computing for big data". In: *Statistics and its interface* 9.4, pp. 399–414. ISSN: 1938-7989. DOI: 10.4310/SII.2016.v9.n4.a1.
- Ward, J. H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301. Publisher: Taylor & Francis, pp. 236–244. ISSN: 0162-1459. DOI: 10.1080/01621459.1963.10500845.
- Yan, M.-J., J. Humphreys, and N. M. Holden (2011). "An evaluation of life cycle assessment of European milk production". In: *Journal of Environmental Management* 92.3, pp. 372–379. ISSN: 0301-4797. DOI: 10.1016/j.jenvman.2010.10.025.

Assessing the reproducibility of FT-MIR-derived biomarker clusters across lactations and their direct prediction from milk spectra

Adapted from: Franceschini, S., Czaplicki, S., Nickmilder, C., Leblois, J., Grelet, C., Gengler, N. and Soyeurt, H. (2026). Assessing the reproducibility of FT-MIR-derived biomarker clusters across lactations and their direct prediction from milk spectra Journal of Dairy Science Communications (In press)

4.1. Chapter outline

The Chapter 3 presents a proof of concept of health pattern discovery in large spectral Fourier Transform mid-infrared (FT-MIR) database. While the clusters identified in Chapter 3 are promising and are associated with several blood biomarkers, the methodology also shows some weaknesses such as the algorithm complexity, the difficulty to transfer the cluster model for external validation and the introduction of randomness. Indeed, the divide-and-conquer methodology used to generate them introduced a degree of randomness into the analytical process. Consequently, the resulting clusters may have been partially influenced by stochastic variation, potentially affecting their stability across repeated analyses. Moreover, the methodology was initially developed using primiparous Holstein cows, and it remains unclear whether the identified structures can be reproduced in multiparous Holstein populations.

Therefore, the objective of Chapter 4 is to assess the reproducibility of the clusters identified in Chapter 3 by repeating the clustering approach across multiple datasets and iterations. More specifically, this chapter aims to evaluate both the stability of the clustering process and the transferability of the identified patterns across lactation groups.

Furthermore, the initial clustering approach relied on a large number of intermediate FT-MIR-based predictions, potentially limiting the transferability of the methodology between stakeholders using different prediction pipelines or calibration equations. Consequently, evaluating whether cluster membership could be predicted directly from the spectra themselves represented an important step toward simplifying the analytical workflow and improving large-scale deployment. This work was conducted in collaboration with Sébastien Czaplicki during his master thesis (Czaplicki 2022).

4.2. Interpretive Summary

Fine milk composition, predicted through FT-MIR spectroscopy for Milk Recording, is a powerful tool exploitable to detect cows with health issues. Building upon our earlier work conducted on primiparous Holstein cows, this study extends the use of unsupervised learning method to multiparous cows while also estimating the robustness of the approach. The methodology developed was repeated five times across three different lactations. The correlations between the results for primiparous and multiparous cows confirm that the patterns detected are consistent and robust only for metabolic trouble detection. To facilitate practical deployment, a predictive model was developed to estimate cluster membership probabilities directly from milk FT-MIR spectra. The resulting model achieved a classification performance (Accuracy=0.92) comparable to that obtained in the previous study, reinforcing its potential for large-scale phenotyping and field application.

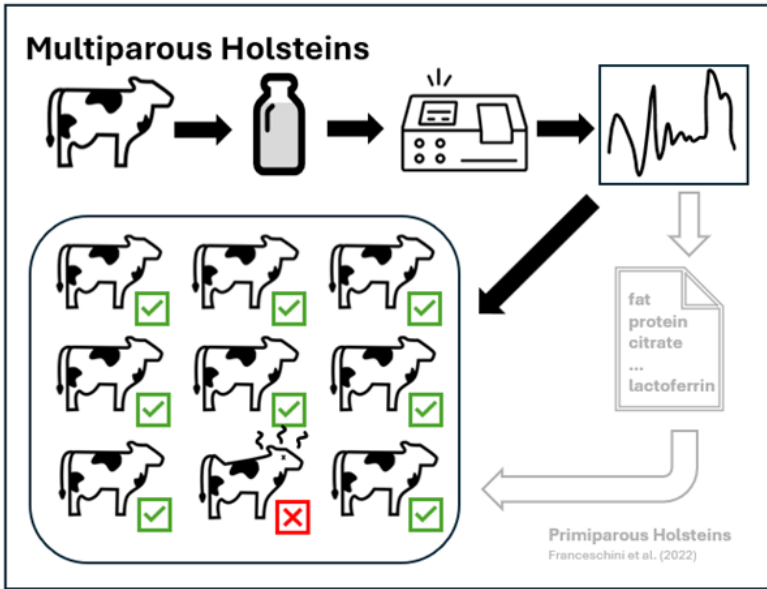


Figure 4.1: Graphical Abstract

4.3. Highlights

- Clustering through a large database of FT-MIR predictions allows for pattern detection.
- Cow metabolic status has a strong signal in FT-MIR which is consistent across lactations.
- Signal for mastitis is dependent on Somatic Cells Count (SCC) or Somatic Cells Score (SCS) and is not consistent across iterations.
- The cow metabolic status is predictable directly from the FT-MIR spectra.

4.4. Abstract

FT-MIR spectroscopy is a powerful tool to predict fine milk composition but also indirect traits such as cow health at large scale. Previously, an unsupervised learning approach applied to primiparous Holstein cows identified a cluster associated with negative energy balance and elevated concentrations of blood β -hydroxybutyrate (BHB) and milk acetone, which are recognized indicators of metabolic disruptions. This study extended the methodology to multiparous cows, representing a broader variability. A hierarchical clustering pipeline was applied to over 1,712,417 records from primiparous and multiparous Holsteins, using 27 FT-MIR-based predictions, somatic cells score and milk yield. The approach was repeated five times for four different datasets: first, second, third lactations and a merged dataset for a total of twenty iterations. Across those iterations, the metabolic cluster was consistently found, and two Partial Least Square Discriminant Analysis prediction models were estimated for each iteration using either the predictions or directly the spectra. The correlations between these predictions of all the models are between 0.82 and 1, suggesting that the models are similar. The cluster previously associated with udder health conditions was not consistently found across repetitions. To ensure transferability and facilitate the deployment of the results, a final model based on FT-MIR was estimated, presenting good performances with an accuracy in validation of 0.92. Those results prove that the metabolic cluster is extendable to other lactations and predictable directly from milk FT-MIR spectra.

4.5. Study

Disease events in dairy herds such as mastitis or ketosis can severely impair production and herd profitability while also compromising animal welfare. For instance, Puerto et al. (2021) showed that mastitic cows could lose between 382 and 989 kg in 305 days in milk cumulative milk yield compared to healthy cows, and a systematic review from Cainzos et al. (2022) highlighted the considerable impact of ketosis across multiple countries with cost per case ranging between €19 and €812. Recent estimates indicate that common dairy cattle diseases collectively cause over \$65 billion in global losses each year with subclinical ketosis, clinical mastitis and subclinical mastitis being the costliest Rasmussen et al. (2024). These numbers highlight the need for effective herd health monitoring and management tools to mitigate disease-related losses.

Early detection of health issues is vital, but current monitoring approaches face technological and economic limitations. Detecting diseases through chemical measurement of specific biomarkers for each disease is labor-intensive, costly, and sometimes invasive (e.g. blood sampling), which restricts large-scale implementation, causes discussions from an ethical perspective, and is impractical for routine use on farm.

Consequently, there is strong interest in affordable, high throughput alternative methods that can leverage existing data streams to flag cows at risk of suboptimal health. In this context, FT-MIR spectroscopy has emerged as a widely available tool. Milk Recording Organization (MRO) already use FT-MIR analysis to measure milk composition for millions of cows, providing rapid, noninvasive insight into each cow's physiology (Soyeurt et al. 2012; De Marchi et al. 2014).

Indeed, these FT-MIR spectra are used to predict an array of major and minor milk components but also health-related biomarkers such as indicators of energy balance (e.g. citrate, oleic acid), of ketosis (e.g., milk and blood BHB, milk acetone, blood Non Esterified Fatty Acids (NEFA)) or of sub-clinical mastitis (lactoferrin, N-acetyl- β -D-glucosaminidase (NAGase))(Soyeurt et al. 2012; Grelet et al. 2016; Giannuzzi et al. 2023). Nevertheless, the accuracy of each individual FT-MIR models can vary widely with coefficients of determination from 0.3 to 0.8 (Grelet et al. 2021), partly due to low concentrations in milk, indirect predictions for concentration in blood, limited data on sick cows and the skewed distributions of the concentrations in milk (Grelet et al. 2024). These constraints make it difficult to predict health conditions reliably via single biomarkers alone.

To extract more global and complementary health information from milk records, researchers have turned to multivariate data mining approaches such as clustering with the biomarkers (Grelet et al. 2019; Foldager et al. 2020). Franceschini et al. (2022) recently demonstrated that it is possible to uncover latent patterns related to cow health by clustering milk recording data (i.e., milk yield, somatic cell count, and 27 FT-MIR predicted traits) from over 100,000 primiparous Holstein cows. Their analysis identified five distinct clusters of records. Notably, one cluster was associated with metabolic disorder indicators and another with mastitis-related measures. Significant correlations were observed between the probabilities of the metabolic cluster and several reference measurements associated with energy balance and ketosis, including blood BHB (0.70), blood NEFA (0.67), blood glucose (-0.60), milk glucose-6-phosphate (-0.32), milk BHB (0.53), milk isocitrate (0.50), and milk urea (0.32). These novel cluster-derived phenotypes aggregate multiple signals and could serve as holistic health indicators for on-farm management, while being less sensitive to individual model inaccuracy. Hu et al. (2025) also showed that this metabolic cluster information was relevant for genetic purposes.

Despite the promise of this unsupervised approach, important questions remain about its robustness and practical utility. The previous study was limited to first-lactation Holsteins, and it remains unclear whether similar health-related spectral patterns emerge in later lactations. Additionally, the stability (i.e. reproducibility) of the clusters found from different datasets has not been thoroughly examined while the divide and conquer approach include randomness. Furthermore, using the clustering method in practice requires computing many intermediate predictor traits from FT-

MIR, adding complexity to the workflow.

In the present study, we hypothesized that the metabolic cluster previously identified in primiparous Holstein cows is reproducible across lactations and can be predicted directly from FT-MIR spectra. In addition, the mastitis-related cluster was also assessed across lactations even though its strongest signal was SCC which is not predicted from spectra. We aimed to address this hypothesis by evaluating the clustering approach across lactations and by testing the consistency and reproducibility of the results under repeated analyses. We also evaluated a new strategy to predict the cluster-defined phenotype directly from milk FT-MIR spectra, bypassing the need to predict multiple biomarkers which could become a limitation when the prediction equations are not available to stakeholders. Although this direct prediction approach may reduce interpretability compared with biomarker-based approaches, it could enhance the transferability and implementation of the method. Ultimately, such an approach could facilitate large-scale monitoring of cow health and contribute to improved disease management and animal welfare.

As one of the aims of this research is to compare the results of the previous research with primiparous cows detailed by Franceschini et al. (2022) to multiparous cows, the dataset and the general clustering algorithm are similar. However, additional methods are developed to assess robustness and compare the groups made by the clustering. All data treatments and analyses were performed using R software (version 4.3.2; <https://www.r-project.org/>). All data were collected as per standard routine by the Walloon Breeders Association (Elevéo, Awé groupe, Ciney, Belgium), which is accredited to perform such collection. The methodology deployed in this study is summarized in Figure 4.2, and the different steps discussed in this communication follow the structure of this workflow.

Data from this study came from the milk recording routine in the southern Belgium between 2012 and 2020. In total, there were 1,712,417 DHI records from 1,082 herds and 139,203 Holstein cows with DIM ranging between 5 and 365 days. The records were divided into three datasets based on the lactation number: 740,454 for first, 577,601 for second and 394,362 for third lactation. DHI records contain the milk yield, the SCC, the DIM, the FT-MIR spectrum. Those analyses were performed by the milk laboratory “Comité du Lait” (Battice, Belgium) on Foss Milkoscan FT6000, FT+, and FT7 standardized spectrometers. The spectral standardization method is described in Grelet et al. (2015) and Grelet et al. (2017). As advised by Grelet et al. (2021), spectra were preprocessed through a first derivative, and 212 wavelengths were selected.

The 27 FT-MIR based biomarkers used in the study were the same as those presented in Franceschini et al. (2022) and predicted with the same equations. The 28th and 29th biomarkers were the milk yield and SCC. Compared to Franceschini et al. (2022), SCC was replaced by the SCC, to avoid the potential effect of the asymmetrical distribution,

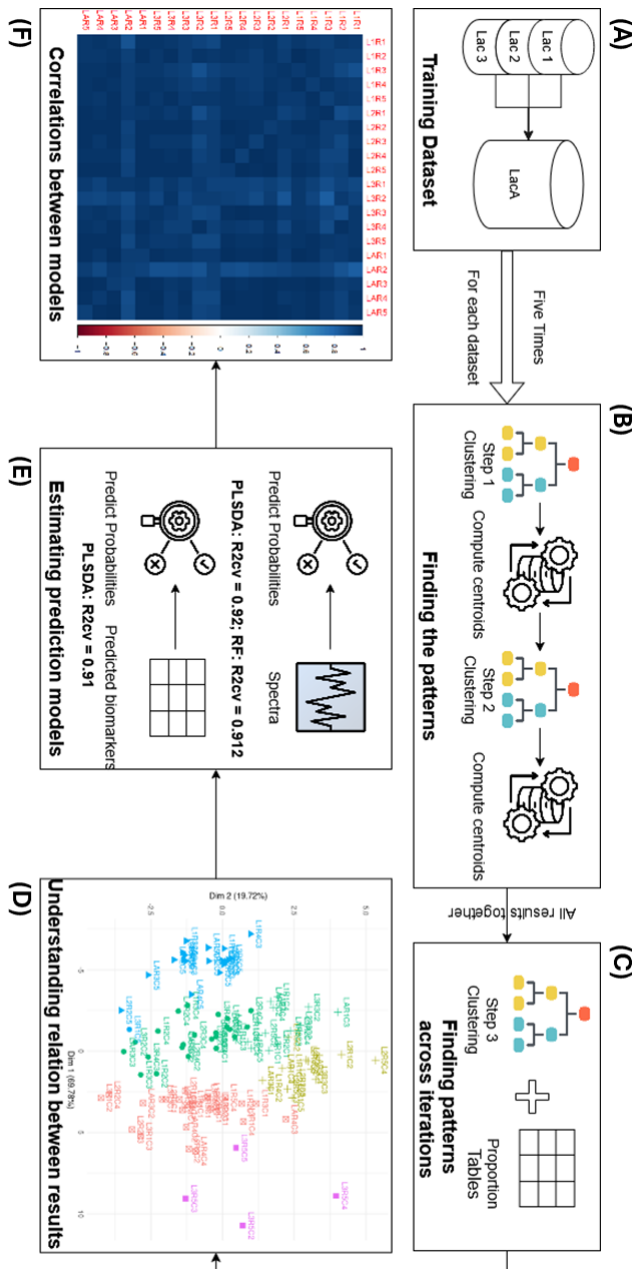


Figure 4.2: Methodology of the study

with this formula:

$$SCS = \log_2(SCC/100,000) + 3$$

Descriptive statistics for all variables are available in the appendix A. The datasets were used both separately by lactation and all together to verify whether the developed method was applicable on multiparous cow. So, four datasets were used, one for each lactation number, called L1, L2 and L3, and one more with the whole data called LA (Figure 4.2A). Moreover, the procedure was repeated five times for each dataset for a total of 20 separate analysis. This data splitting ensured that we could assess the reproducibility of the clustering approach within and between lactations while also evaluating the robustness of the identified clusters across iterations.

As developed in Franceschini et al. (2022), a two-steps hierarchical clustering method using the Euclidean distance had been performed. Indeed, it was not possible to use such a volume of data simultaneously because hierarchical clustering requires the computation of a distance matrix, which would require excessive memory allocation. Instead, the dataset was randomly split into 100 smaller datasets that were clustered separately. Then, the multiple hierarchical clustering were partitioned to obtain 600 groups of records for each. The means of each group called centroid were calculated to summarize the information. Then, a second hierarchical clustering took as input 60,000 centroids (i.e., 100 subsets x 600 centroids). Finally, we cut the final dendrogram to obtain 5 large groups of centroids as in Franceschini et al. (2022), and we computed the means for the different variables (Figure 4.2B). Those groups will be called “cluster” in the following text.

As the whole approach was repeated 5 times for each of the four datasets, we obtained 20 outputs of 5 clusters each, so a total of 100 clusters. In order to assess if these 100 clusters represented the same five clusters for each repetition, a last hierarchical clustering had been performed. The dendrogram from the final hierarchical clustering algorithm was also divided into five clusters called overall group (Figure 4.2C).

The objective of this last clustering was only to validate that the metabolic disruption and mastitis clusters from the previous study were found in every repetition. If the signals from the clusters were strong and similar in every repetition, there should be five overall groups containing 20 similar clusters each. However, the result showed a different distribution with 26 in overall group 1, 14 in overall group 2, 36 in overall group 3, 20 in overall group 4 and 4 in overall group 5 (Table 4.1). So, the different partitions from the 20 repetitions were not robust. Only overall group 4 contains 20 clusters, one for each of the five repetitions across four lactation datasets.

To identify whether this overall group 4 represented the metabolic disruption or the mastitis cluster found by Franceschini et al. (2022), we looked at a principal component analysis (Figure 4.2D) and at the predictions. Most specifically, we looked at the cluster with the highest value of acetone and BHB in milk and the lowest value of Energy Balance (EB) in each repetition (Figure 4.3). Except for one repetition (EB in the first

Table 4.1: Assignment of the 100 clusters (5 clusters per repetition $\times$ 5 repetitions per lactation dataset $\times$ 4 lactation datasets) to the five overall groups.

	Lactation 1					Lactation 2					Lactation 3					All Lactation					Total
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	
Group 1	1	1	1	1	2	1	3	1	1	1	3	1	1	1	0	1	1	1	3	1	26
Group 2	1	0	0	1	1	2	0	1	1	2	0	0	1	1	0	1	0	1	0	1	14
Group 3	2	3	3	2	1	1	1	2	2	1	1	3	2	2	0	2	3	2	1	2	36
Group 4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
Group 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	4

Group 4 in bold is consistent across repetitions and lactations. R = Repetitions

repetition of the third lactation dataset), overall group 4, which was in every repetition, was also the one with the highest value of acetone and BHB and the lowest value of EB. So, it should be associated with metabolic troubles.

A similar approach was used to find the cluster associated with mastitis, we looked at the highest value of SCC, lactoferrin and Na⁺ (Figure 4.3). However, the results were not as consistent. Overall group 2 was associated with the highest values of the three variables in 7 out of the 20 repetitions. Even if we only considered SCC, overall group 2 is only found in 10 of the repetitions. So, the mastitis cluster was not as stable. As hypothesized previously, most of the signal found in Franceschini et al. (2022) for mastitis might come from the highly skewed distribution of the SCC which might give more weight to this variable instead of the FT-MIR predictions. When the SCC was used instead of SCC, the signal of the mastitis cluster was weaker, and we could not identify the pattern across lactations and iterations. For all the reasons mentioned above, the cluster related to mastitis was not considered for the repetition analysis.

To further evaluate whether the information in the metabolic troubles cluster was identical across iterations we created 20 prediction models, one for each of the four datasets and each of the repetition, using the 29 predicted features but also 20 models using directly the FT-MIR spectra (Figure 4.2E). Indeed, using the spectra directly allows a faster and easier estimation of the cluster values and probabilities.

Each model was estimated with data used for the clustering approach. So, a model for a cluster found from the first lactation dataset was estimated using a cross-validation on the first lactation dataset. The idea was that if the clusters were similar, the predictions from the different models should be highly correlated. The 29 features were used to predict the cluster using a Partial Least Square Discriminant Analysis (PLS-DA) as in Franceschini et al. (2022). The same approach was applied to the FT-MIR spectra to develop an equation. The output for each observation was the probability to belong to the metabolic trouble cluster. To ensure a class balance between metabolic troubles and healthy observations, datasets were down sampled by sampling the same number of observations for each. Then a 10-fold cross-validation was applied to define the best number of latent variables with the train function and the oneSE approach from caret to estimate the model quality (Kuhn 2008; Breiman et al. 2017).

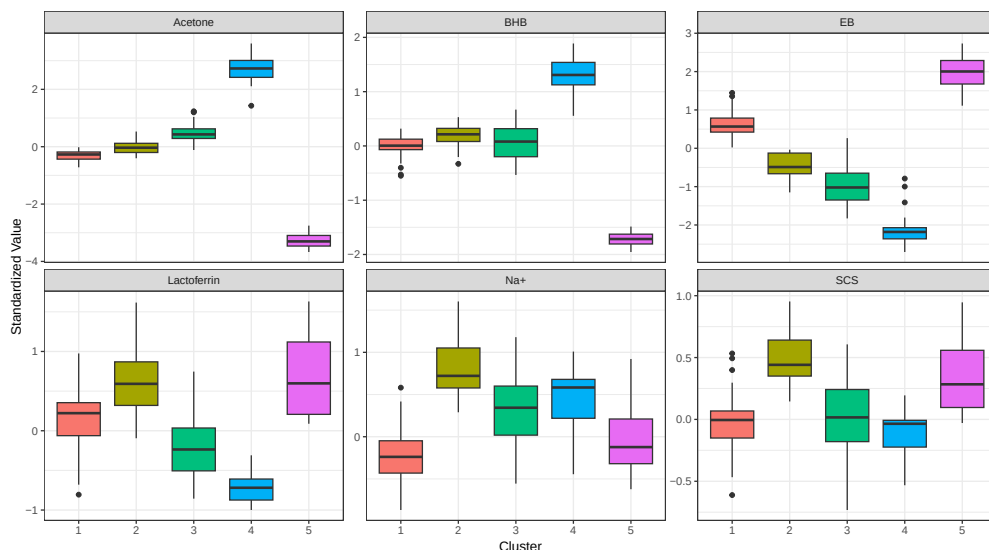


Figure 4.3: Boxplots of the distributions of the different clusters for biomarkers of interest. BHB = β -hydroxybutyrate, EB = Energy Balance, SCS= Somatic Cells Score.

The numbers of latent variables required were between 1 and 3 for models using predictions and between 10 and 19 for models using FT-MIR spectra. The retained latent variables explained between 31.4% and 48.2% of the variance in the predictor matrix for models using predictions and between 76.4 and 92.7% for models using FT-MIR spectra. The accuracies in cross-validation ranged between 0.79 and 0.93 using predictions and between 0.76 and 0.93 using FT-MIR spectra. So, the model performances were quite different between repetitions, but those performances were not comparable as the datasets were different.

After the models trained on predictions and spectra were estimated, we applied those models to the full dataset LA and computed the Pearson correlations. Those correlations are shown in the final heatmap of Figure 4.2F. This methodology allowed us to compare the results of the models and see whether they predict the same information. Overall, the weakest correlation was 0.82 and the strongest is 1.00 so the predictions for a metabolic cluster were mostly identical across lactations and repetitions. Correlations between models trained from predictions or FT-MIR spectra in the same repetitions (i.e. with the same definition of the metabolic trouble cluster) were between 0.93 and 0.95, also containing the same information. The model having the best average correlation with all the others was trained using the full dataset LA and the first repetition. This would suggest that the signal for metabolic issues was strong enough in the spectra so that the pattern could be identified in every dataset across lactation and every repetition.

The PLS-DA prediction model using spectra we kept was estimated with the full dataset LA and the first repetition. The accuracy in cross-validation for this model is 0.920 ± 0.002 . A Random Forest (RF) model was also estimated as it might capture the non-linearity while also offering a probability-like output by looking at proportion of trees voting for one cluster (Franceschini et al. 2025). This algorithm was only used on the LA results because of the computational burden of such estimation for 1,712,417 records and 212 spectral wavelengths. A 10-fold cross-validation was used to optimize the number of variables selected at each split and the minimum node size. We used Gini as the split rule; the number of trees was 500 and we limited the maximum depth to 15 to avoid a model being too complex. The best random forest uses 50 variables at each split and a minimum node size of 5. The average accuracy in cross-validation of the model is 0.912 ± 0.001 . A paired t-test was applied by pairing the differences in performance between both methods for each of the ten folds to determine whether this difference in accuracy is significant, and the result is highly significant ($P < 0.001$) even though the average difference is 0.008. So, the PLS-DA model was slightly better than the RF model. This result could be explained by the maximum depth added to simplify the RF model or by the absence of non-linearity. However, probabilities output by both algorithms were distributed quite differently with a normal distribution around 0.5 for the PLS-DA and a bimodal distribution for the RF. This difference, combined with the computational burden, could justify the use of both algorithms to fine-tune a detection threshold relevant for the farmer.

Several limitations should nevertheless be considered. First, although the metabolic cluster was reproducible across lactations and repeated iterations, the robustness of cluster assignment under changing population structures, breeds, or management systems outside southern Belgium still requires further investigation. Additional methodologies, such as introducing perturbations in the datasets and evaluating the agreement of the resulting cluster assignments, could further strengthen the reproducibility assessment. Second, the hierarchical clustering strategy relied on a divide-and-conquer approach because the computation and storage of large distance matrices generated substantial computational and memory requirements which could potentially be reduced using alternative unsupervised learning approaches. Finally, the use of Euclidean distance and distance-based clustering approaches may be influenced by non-linear relationships and correlated predictors, potentially affecting the relative importance attributed to some variables.

In conclusion, this research proposed a way to assess the generalizability across lactations and the repeatability of the approach developed while also presenting an equation to predict the metabolic trouble cluster directly from the spectra for better transferability. Those results showed that only one of the two clusters previously detected is relevant. Indeed, the metabolic trouble cluster is consistent across lactations and repetitions which highlight its strong signal in FT-MIR. This cluster could be used directly in the routine to assess metabolic issues on-farm or to screen for future

sampling. The next step would be to validate using this on-farm implementation and to assess the cluster generalizability to other countries and breeds. However, this validation requires a framework where the information is quickly streamlined to the farmer and feedback is collected. The direct prediction from FT-MIR is a first step in this direction.

4.6. Notes

This study received no external funding. The authors acknowledge the support of the AWé groupe (Ciney, Belgium) in collecting, processing and providing the dataset. During the preparation of this work, the author used Chat-GPT 4.5 to improve the flow, the grammar and the conciseness of the text. The author reviewed and edited the content as needed and takes full responsibility for the content of the publication. Ethical approval was not necessary as milk spectra used are collected by the AWé groupe in routine for Dairy Herd Improvement purpose. The authors have not stated any conflicts of interest.

4.7. Unpublished Supplementary Results

As mentioned at the end, an on-farm implementation is essential to assess its generalizability but also its usefulness on the field. The aim is to bring the metabolic trouble cluster in farm and get feedback from the farmer or the farmer's data management. However, this proves to be difficult in Belgium because of the herd sizes and the European regulation. Instead, a collaboration with the Bovi-Analytics Lab at Cornell University (Ithaca, New York State, US) was developed. The project involves the use of data from three New York State dairy farms (Aurora Ridge Dairy, Oakwood, and SunnySide) representing approximately 10,000 cows. Currently, only Aurora Ridge Dairy data were processed because we could not directly collect the farm management data from DairyComp (Valley Ag Software from URUS group, Wisconsin, USA) for Oakwood and SunnySide. Between 14/05/2025 and 14/09/2025, 27,911 herd management notes were collected, along with 18,060 milk FT-MIR spectra. The metabolic disorder cluster equation was applied to those spectra, and we compared the negative and positive alerts to the herd management notes for the following 10 days. The preliminary results are presented in Table 4.2. The proportions of management notes for negative events are systematically higher. These results are encouraging, even though the number of observations available within the selected time window was substantially reduced compared with the initial dataset, decreasing from 27,911 management notes and 18,060 spectra to 3,875 notes associated with spectra, including only 103 relevant notes, which is insufficient to confirm statistical significance.

The farmers agreed to provide observational data and on-demand feedback, which enabled the implementation of a real-time validation framework. However, the de-

Table 4.2: Comparison of event occurrence between good and bad alerts.

Alert	Total	Culture		Died		Excede		Ketosis		Mastitis		Ulcer	
	n	n	p	n	p	n	p	n	p	n	p	n	p
Good	3479	36	0.01	2	0.001	3	0.001	0	0	35	0.01	2	0.001
Bad	396	12	0.03	2	0.005	1	0.003	1	0.003	8	0.02	1	0.003

Excede = Excede Antibiotic treatment, n = number of observations, p = proportion of observations

ployment of such a system has revealed important technical challenges. In contrast to Wallonia, where spectral collection, standardization, and prediction pipelines are operational, no equivalent infrastructure is available in the United States. As a result, a dedicated framework had to be developed, including data acquisition, processing, and integration. Among these, instrument standardization has proven particularly challenging, requiring the exchange of milk samples across locations and careful calibration procedures. While the results are preliminary, this work lays the foundation of feedback-oriented validation on large US dairy farms.

Bibliography

- Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone (2017). *Classification and Regression Trees*. New York: Chapman and Hall/CRC. 368 pp. ISBN: 978-1-315-13947-0. DOI: 10.1201/9781315139470.
- Cainzos, J. M., C. Andreu-Vazquez, M. Guadagnini, A. Rijpert-Duvivier, and T. Duffield (2022). “A systematic review of the cost of ketosis in dairy cattle”. In: *Journal of Dairy Science* 105.7. Publisher: Elsevier, pp. 6175–6195. ISSN: 0022-0302. DOI: 10.3168/jds.2021-21539.
- Czaplicki, S. (2022). “Contribution à l’appréciation du bien-être animal au travers de méthodes non supervisées basées sur les données collectées dans le cadre du contrôle laitier wallon”. Unpublished master’s thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/15949>.
- De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa (2014). “Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits”. In: *Journal of Dairy Science* 97.3. Publisher: Elsevier, pp. 1171–1186. ISSN: 0022-0302. DOI: 10.3168/jds.2013-6799.
- Foldager, L., C. Gaillard, M. T. Sorensen, T. Larsen, E. Matthews, R. O’Flaherty, F. Carter, M. A. Crowe, C. Grelet, M. Salavati, M. Hostens, K. L. Ingvarsten, and M. A. Krogh (2020). “Predicting physiological imbalance in Holstein dairy cows by three different sets of milk biomarkers”. In: *Preventive Veterinary Medicine* 179. ISSN: 0167-5877. DOI: 10.1016/j.prevetmed.2020.105006.
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt (2022). “Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?”. In: *Journal of Dairy Science* 105.8, pp. 6760–6772. ISSN: 0022-0302. DOI: 10.3168/jds.2022-21975.
- Franceschini, S., C. Fastré, C. Nickmilder, D. E. Santschi, D. Warner, M. Bahadi, C. Bertozzi, D. Veselko, F. Dehareng, N. Gengler, and H. Soyeurt (2025). “Detection of Dairy Herd Management Issues Using Fatty Acid Profiles Predicted by Mid-Infrared Spectrometry”. In: *Animals* 15.11. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 1575. ISSN: 2076-2615. DOI: 10.3390/ani15111575.
- Giannuzzi, D., L. F. M. Mota, S. Pegolo, F. Tagliapietra, S. Schiavon, L. Gallo, P. A. Marsan, E. Trevisi, and A. Cecchinato (2023). “Prediction of detailed blood metabolic

- profile using milk infrared spectra and machine learning methods in dairy cattle". In: *Journal of Dairy Science*. ISSN: 0022-0302. DOI: 10.3168/jds.2022-22454.
- Grelet, C., C. Bastin, M. Gelé, J. -. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng (2016). "Development of Fourier transform mid-infrared calibrations to predict acetone, β -hydroxybutyrate, and citrate contents in bovine milk through a European dairy network". In: *Journal of Dairy Science* 99.6, pp. 4816–4825. ISSN: 0022-0302. DOI: 10.3168/jds.2015-10477.
- Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng (2015). "Standardization of milk mid-infrared spectra from a European dairy network". In: *Journal of Dairy Science* 98.4, pp. 2150–2160. ISSN: 0022-0302. DOI: 10.3168/jds.2014-8764.
- Grelet, C., T. Larsen, M. Crowe, D. Wathes, C. Ferris, K. Ingvarsten, C. Marchitelli, F. Becker, A. Vanlierde, J. Leblois, U. Schuler, F. Auer, A. Köck, L. Dale, J. Sölkner, O. Christophe, J. Hummel, A. Mensching, J. Fernández Pierna, H. Soyeurt, M. Calmels, R. Reding, M. Gelé, Y. Chen, N. Gengler, and F. Dehareng (2024). "Prediction of key milk biomarkers in dairy cows through milk mid-infrared spectra and international collaborations". In: *Journal of Dairy Science* 107.3, pp. 1669–1684. ISSN: 00220302. DOI: 10.3168/jds.2023-23843.
- Grelet, C., J. A. F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng (2017). "Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models". In: *Journal of Dairy Science* 100.10, pp. 7910–7921. ISSN: 0022-0302. DOI: 10.3168/jds.2017-12720.
- Grelet, C., A. Vanlierde, M. Hostens, L. Foldager, M. Salavati, K. L. Ingvarsten, M. Crowe, M. T. Sorensen, E. Froidmont, C. P. Ferris, C. Marchitelli, F. Becker, T. Larsen, F. Carter, and F. Dehareng (2019). "Potential of milk mid-IR spectra to predict metabolic status of cows through blood components and an innovative clustering approach". In: *Animal* 13.3, pp. 649–658. ISSN: 1751-7311. DOI: 10.1017/S1751731118001751.
- Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Steevens, N. Gengler, and F. Dehareng (2021). "Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions". In: *Methods*. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2020.07.012.
- Hu, H., S. Franceschini, P. Lemal, C. Grelet, Y. Chen, H. Atashi, K. Wijnrocx, H. Soyeurt, and N. Gengler (2025). "Exploring the relationship between predicted negative energy balance and its biomarkers of Holstein cows in first-parity early lactation". In: *Journal of Dairy Science* 108.5, pp. 5433–5447. ISSN: 0022-0302. DOI: 10.3168/jds.2024-25932.
- Kuhn, M. (2008). "Building Predictive Models in R Using the caret Package". In: *Journal of Statistical Software* 28, pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05.

- Puerto, M. A., E. Shepley, R. I. Cue, D. Warner, J. Dubuc, and E. Vasseur (2021). "The hidden cost of disease: I. Impact of the first incidence of mastitis on production and economic indicators of primiparous dairy cows". In: *Journal of Dairy Science* 104.7. Publisher: Elsevier, pp. 7932–7943. ISSN: 0022-0302. DOI: 10.3168/jds.2020-19584.
- Rasmussen, P., H. W. Barkema, P. P. Osei, J. Taylor, A. P. Shaw, B. Conrady, G. Chaters, V. Muñoz, D. C. Hall, O. O. Apenteng, J. Rushton, and P. R. Torgerson (2024). "Global losses due to dairy cattle diseases: A comorbidity-adjusted economic analysis". In: *Journal of Dairy Science* 107.9. Publisher: Elsevier, pp. 6945–6970. ISSN: 0022-0302. DOI: 10.3168/jds.2023-24626.
- Soyeurt, H., C. Bastin, F. G. Colinet, V. M. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Schefers, J. Vandenplas, K. Weigel, M. Coffey, L. Théron, J. Detilleux, E. Reding, N. Gengler, and S. McParland (2012). "Mid-infrared prediction of lactoferrin content in bovine milk: potential indicator of mastitis". In: *Animal*. Animal 6.11, pp. 1830–1838. ISSN: 1751-7311. DOI: 10.1017/S1751731112000791.

Detection of Dairy Herd Management Issues Using Fatty Acid Profiles Predicted by Mid-Infrared Spectrometry

*Adapted from: Franceschini, S., Fastré, C., Nickmilder, C., Santschi, D. E., Warner, D., Bahadi, M., Bertozzi, C., Veselko, D., Dehareng, F., Gengler, N., and Soyeurt, H. (2025). Detection of dairy herd management issues using fatty acid profiles predicted by mid-infrared spectrometry. *Animals*, 15(11), 1575. <https://doi.org/10.3390/ani15111575>*

5.1. Chapter Outline

Chapters 3 and 4 demonstrated that biologically relevant clusters could be identified from milk Fourier Transform mid-infrared (FT-MIR)-based predictions and reproduced across multiple analyses. However, an important question remained regarding their practical utility for routine herd monitoring. We believe that the practical utility, especially for monitoring rely on frequent data acquisition to study the evolution in time. However, milk recording data are typically collected every 4 to 6 weeks, which strongly limits their capacity to support near real-time monitoring of animal status, rapid management interventions, or the detection of short-term biological changes.

Therefore, the objective of Chapter 5 is to investigate whether a similar unsupervised learning approach could be applied to bulk tank milk FT-MIR data originating from milk payment systems, which are collected much more frequently. In Wallonia, bulk tank milk spectra are routinely available every 1 to 4 days, offering the possibility of monitoring the temporal evolution of herd-level biological patterns over time. Such temporal information could be particularly relevant for the detection of management changes, feeding transitions, seasonal effects, or emerging herd-level health disturbances. Although bulk tank milk does not provide information at the individual cow level, it was hypothesized that these spectra could still capture biologically meaningful signals related to herd management and global herd physiological status.

In addition, this chapter explored the transferability of the methodology across different geographical and analytical contexts through the use of Canadian data. Particular attention was also given to reducing the level of randomness introduced during the clustering process by implementing an alternative sampling strategy for handling the large-scale datasets. Together, these developments aimed to improve the robustness, scalability, temporal monitoring capacity, and practical applicability of unsupervised FT-MIR-based monitoring approaches. This work was conducted in collaboration with Claire Fastré during her master thesis (Fastré 2022).

5.2. Simple Summary

Farms generate increasing amounts of data each year. One example is the bulk tank milk composition, predicted through spectrometry, which is routinely measured for milk payment purposes. Among the different milk components, the fatty acid profile is of utmost importance because it is closely linked to animal status and farm management practices. This research aims to explore a novel application of these fatty acid profiles by developing a practical herd-monitoring tool for farmers and advisors. The methodology developed consists of an unsupervised learning method that identifies meaningful patterns in fatty acid profiles, combined with an expert-driven interpretation of those patterns. The analysis was performed using a Belgian bulk tank milk database. Seven distinct patterns were identified; among these, three

were associated with good management practices, three indicated potential risks, and one highlighted a metabolic disorder probably related to management practices. Most of these patterns were also observed in a Canadian bulk tank milk dataset, demonstrating that the method is generalizable across different regions and farming conditions. Moreover, the probabilities associated with each pattern can serve as a reliable foundation for creating practical alerts to support on-farm decision making, thereby enhancing farm management and contributing positively to animal welfare.

5.3. Abstract

This article focuses on the creation of a monitoring tool using routinely collected data from milk payment analyses. Milk samples were analyzed through Fourier Transform mid-infrared spectrometry every 1 to 3 days, and their compositions were predicted using machine learning models. Among the predicted parameters, fatty acid profiles appear to be effective indicators of animal status and management practices. In this research, these profiles were summarized using 31 fatty acids or groups of fatty acids. The methodology consists of four steps: hierarchical clustering to detect patterns in a Belgian spectral dataset (N=774,781), interpretation of the identified seven clusters, development of predictive models applied to a Canadian dataset (N = 670,165), and validation using management information collected from Canadian farms. The identified clusters revealed significant relationships with feeding management strategies and temporal evolutions, highlighting the potential to develop automated alert systems that assist farmers and advisors in herd monitoring.

Key words: fatty acids; FT-MIR spectrometry; unsupervised hierarchical clustering; dairy cows; herd management

5.4. Introduction

In 2023, the Food and Agriculture Organization (FAO) emphasized the critical role of milk in enhancing nutrition and health due to its nutritional composition and significance in global food systems FAO. 2023. Cow milk typically comprises 86.9% water, 4.6% lactose, 4.2% fat, 3.4% protein, 0.8% minerals, and 0.1% vitamins. Among the fat, 98% consists of tri-glycerides associated with Fatty Acid (FA) Jensen 2002. FAs are essential since their composition directly impacts milk quality, influencing physical, nutritional, and technological properties important for human consumption and dairy processing Chilliard et al. 2000. Moreover, the milk FA profile is a valuable proxy of the cow's metabolic status Giannuzzi et al. 2022. Although approximately 400 different FAs exist in milk, only 12 FAs constitute more than 1% each in milk fat Jensen 2002; Riuzzi et al. 2021.

Based on their origins, FAs can be categorized into three main groups: preformed

FAs (approximately 41%), mixed FAs (about 31%), and de novo FAs (around 28%). Understanding variation among these FA groups is crucial for evaluating the animal's metabolic and nutritional status, since they originate from distinct metabolic pathways which are sensitive to management practices and health conditions. Preformed and de novo FA pathways are summarized in Figure 5.1.

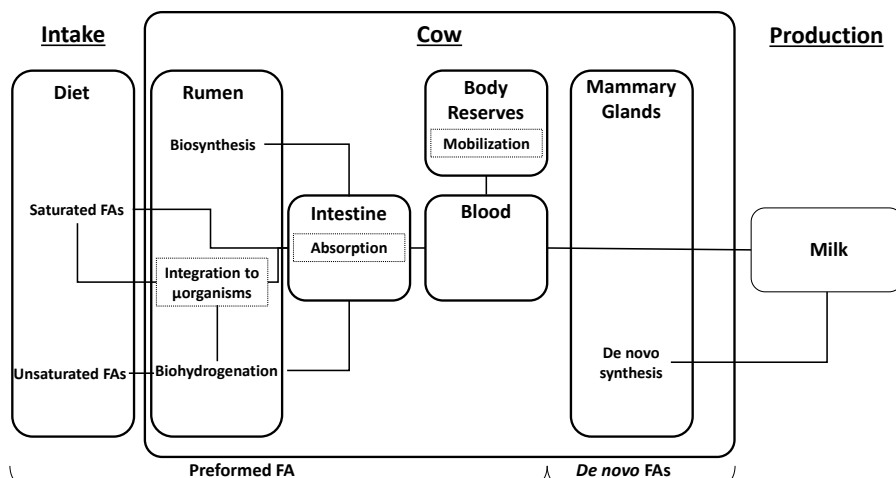


Figure 5.1: Diagram of fatty acid synthesis pathways (FAs: fatty acids, μ organism: microorganisms).

De novo FAs are synthesized directly in the epithelial cells of the mammary gland from residues of ruminal microbial fermentation (Chilliard et al. 2000). The main primers are acetate (two carbons) and butyrate (four carbons), onto which two-carbon molecules are successively added to produce FAs up to fifteen carbons long, which are immediately excreted in milk (Lindmark Månsson et al. 2008).

Preformed FAs originate from three sources: dietary intake, microbial FA synthesis in the rumen, and body fat mobilization. Dietary FAs, once digested and absorbed into the bloodstream, can enter milk unchanged or undergo microbial biohydrogenation in the rumen if polyunsaturated. Biohydrogenation saturates FA chains through microbial action in the rumen, forming intermediate unsaturated FAs before becoming fully saturated. Occasionally, intermediate products bypass the final saturation steps and are absorbed directly in their unsaturated form (Bauman et al. 2006). These ruminal processes greatly influence milk FA profiles and depend heavily on microbial populations present in the rumen.

These microbial populations can also synthesize FAs from ruminal fermentation residues (Jenkins et al. 1993). The importance of each source depends on the micro-

bial population and the proportion of lipids in the ration (Vlaeminck et al. 2006a). Biosynthesis mainly produces C16:0 and C18:0 from acetate and butyrate, but also odd- or branched-chain FAs from other primers. For example, using propionate (three carbons), microorganisms can synthesize odd-chain FAs, mainly C15:0 and C17:0. By using amino acid catabolism residues, the final products are branched-chain FAs (Vlaeminck et al. 2006a). These microbial FAs are then absorbed in the intestine, pass into the bloodstream, and are incorporated into milk fat.

The final source of preformed FAs is body fat, which can be mobilized by the animal. This source generally accounts for less than 10% of milk fatty acids, except after calving, when lactation begins and the animal's energy deficit is greatest (Fiore et al. 2020; Palmquist et al. 1993). This deficit is due to the high energy demand for calving and the onset of lactation, as well as the reduced dry matter intake by the animal at this stage of lactation (Jorjong et al. 2014). At the same time, the mammary gland loses its capacity to produce FAs *de novo*. To compensate for these phenomena, the animal will mobilize its body fat. In the first weeks of lactation, there are fewer *de novo* and more preformed FAs in the milk (Craninx et al. 2008).

Those mentioned pathways are known to be influenced by many factors (Jensen 2002; Palmquist et al. 1993; Mansbridge et al. 1997) such as breed and genetics (Soyeurt et al. 2011b); herd management, including animal density and ration (Woolpert et al. 2016; Woolpert et al. 2017); stage of lactation (Craninx et al. 2008; temperature Hammami et al. 2015; Liu et al. 2017; animal health and physiology (Mann et al. 2016); animal nutrition; and the main fermentation processes in the rumen, themselves linked to diet (Vlaeminck et al. 2006b; Vlaeminck et al. 2006c).

Thus, analyzing the changes in the FA profile of milk is a powerful tool to monitor a dairy herd routinely, whether to assess the impact of management practices or detect external factors impairing the herd, such as feed storage problems or heat stress events during periods of hot weather. However, a monitoring tool requires easy, cheap, and robust data acquisition. In this context, the FT-MIR spectrometry already applied to milk bulk tanks for milk payment is a huge opportunity. FT-MIR spectrometry is a rapid and non-destructive method that can predict several key milk FAs accurately (Soyeurt et al. 2011a; Soyeurt et al. 2006).

However, interpreting complex FA profiles routinely remains challenging due to the numerous FAs and their interactions. Traditional multivariate analyses provide global overviews but lack detail for specific patterns. Unsupervised machine learning techniques, specifically clustering and pattern recognition, may provide deeper insights into specific FA profile groupings, revealing less frequent yet significant patterns. These machine learning methodologies do not require a target to predict, which is often difficult to have at a large scale, but aim to find variables that are interacting together and that form groups of homogeneous observations. Another advantage of using an unsupervised approach is the possibility to recognize patterns on the entire

spectral database and not only on a few records for which the target is observed, as is done when using supervised learning.

Recently, Franceschini et al. (2022) demonstrated this by applying hierarchical clustering to FT-MIR predictions of individual cow milk samples. This approach was initially applied to individual cow milk samples from Dairy Herd Improvement (DHI) records using FT-MIR predictions related to animal health. In the present study, we extend this approach by utilizing spectral data obtained from milk analyses used for milk payment determination. Compared to DHI records, bulk tank milk records offer the advantage of higher sampling frequency (every 1 to 4 days) versus traditional milk recording (every 4 to 6 weeks). Furthermore, such data are available for all herds delivering milk to dairy processors, whereas DHI data are limited to herds enrolled in performance recording programs.

In conclusion, this study aims to address the complexity of interpreting the numerous FA profiles by identifying the most representative and informative ones, providing biological interpretation, and developing a predictive model. From those results, the final aim is to estimate the feasibility of establishing a decision support tool for dairy herd management, based on the FA profile of milk predicted by mid-infrared spectrometry.

5.5. Materials and Methods

The software used for data processing and analysis was the R language, version 4.2.3 (R Core Team 2022). The entire workflow from databases to results is summarized in Figure 5.2.

5.5.1. Belgian Dataset

The first database used in this study is linked to the analysis of milk samples collected within the milk payment scheme in the Southern Region of Belgium called Wallonia. Access to this database is governed by the “Futurospectre” agreement between ULiège–Gembloux Agro-Bio Tech (Gembloux, Belgium), the Walloon Research Centre (CRA-W, Gembloux, Belgium), the Walloon Breeding Association (AWé, Ciney, Belgium), and the milk laboratory Comité du Lait (Battice, Belgium). No ethical approval was needed, as the milk samples were collected during the milk payment carried out routinely.

The milk samples, taken every 1 to 3 days by the dairies from the tanks of the farms collected, were analyzed using FT-MIR spectrometry by the Comité du Lait (Battice, Belgium). All samples were analyzed with Foss MilkoScan spectrometers (Foss, Hillerød, Denmark). During this analysis, in addition to the fat and protein contents predicted by the spectrometer, spectra were also recorded in a database. The spectra were standardized using the method developed by Grelet et al. (Grelet

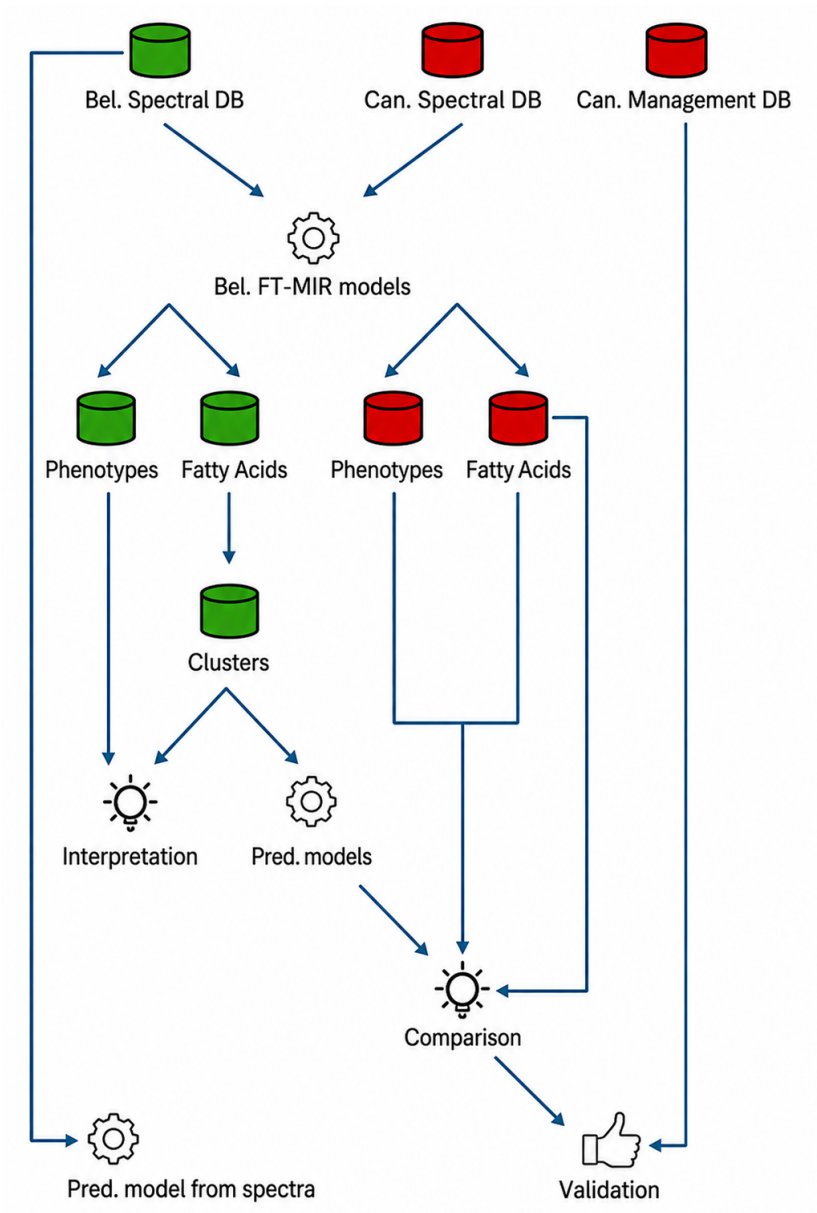


Figure 5.2: Workflow of the analysis. The colors indicate the origin of the dataset: green for Belgium and red for Canada. (Bel = Belgian, DB = database, Can = Canadian, FT-MIR = Fourier Transform mid-infrared, Pred = predictions).

et al. 2015). Different prediction equations from Grelet et al. (Grelet et al. 2021) were then applied to these spectra to extend the number of phenotypes available. These correspond to 31 phenotypes relating to FAs or FA groups (g/dL milk), milk production (kg/day), estimated fat content (g/dL milk), β -hydroxybutyrate concentration in milk (BHB; log, mmol/L plasma), energy balance, protein efficiency, free FA in blood (FFA; log, μ Eq/L plasma), and dry matter intake (kg/day).

The predicted FA concentrations have been modified to be expressed as a ratio in g/100 g fat from the predicted fat content. This unit has the advantage of reducing correlation with milk production and fat content, better reflecting the importance of different metabolic pathways responsible for producing these milk compounds.

To isolate potential heat stress effects, temperature and humidity measurements were added to the database. The zip codes of Walloon municipalities were linked to the nearest meteorological station. The map of Belgian municipalities was superimposed on that of the 30 Walloon weather stations in the AGROMET network (*Agromet.be* 2025) using the Voronoi method as presented by Nickmilder et al. (Nickmilder et al. 2023). This method involves assigning to each station a polygon of a given area on the map. Municipalities covering more than one polygon were linked to the polygon on which most of their surface area was located. Finally, the Temperature Humidity Index (THI) was calculated using the following formula:

$$THI = (0.8 \times tsa) + \left(\frac{hra}{100} \times (tsa - 14.4) \right) + 46.4 \quad (5.1)$$

where *tsa* is the ambient dry temperature (°C) and *hra* is the ambient relative humidity (%) Jubb et al. 2015. Thanks to this link between dairy records and weather stations, meteorological information could be merged with the spectral database.

Then, the resulting database was cleaned according to ICAR's recommendations for recording of dairy cattle milk data International Council on Animal Recording (ICAR) 2024. These recommendations were created for the cow level, but it is expected to work at bulk tank milk level, as it is a weighted average of the herd. Only values between 1.5% and 9% fat and between 1% and 7% protein in g/dL of milk were retained. Rows containing outliers for fat and protein, or negative values for fatty acids or missing values for one of the traits studied, were discarded.

Then, the standardized Mahalanobis distance (GH) was calculated for each record to detect potential extreme data and compare the datasets from Belgium and Canada. To calculate it, it was necessary to perform a Principal Component Analysis (PCA) on the 31 studied phenotypes, given the high correlations existing between some of them. Following this PCA, six principal components (PCs) were retained, as they explained 95.48% of the variability in the data. This analysis was performed using the FactoMineR package, version 2.4 (Lê et al. 2008). Next, the GH distance was calculated

as follows (Zhang et al. 2021):

$$GH = \left((\bar{x} - \bar{\mu})^T S^{-1} (\bar{x} - \bar{\mu}) \right) / nPC \quad (5.2)$$

where $\bar{x}$ corresponds to the vector of PCs of the observation, $\bar{\mu}$ is the vector containing the mean of each PC, T denotes the transpose, S^{-1} denotes the inverse of the variance–covariance matrix of the PCs selected by describing the matrix $\bar{x}$, and nPC is the number of PCs. No high extreme samples defined by a $GH > 5$ were observed, so we decided to keep all samples, as we wanted to observe samples with abnormal behaviors. The cleaned database included 774,781 records collected from 2835 Walloon farms between December 2018 and December 2021.

5.5.2. Canadian Dataset

The second dataset used in the present work comes from Canadian milk recording. Lactanet (Ste Anne de Bellevue, QC, Canada) provided access to spectral data from tank samples collected in Quebec from January 2020 to March 2022 from 4676 farms. The samples taken every 1 to 3 days from the tank were analyzed by Foss MilkScan spectrometers (Foss, Hillerød, Denmark). Canadian spectral data were used to predict FT-MIR phenotypes using the Belgian prediction models. Next, the predicted FT-MIR phenotypes were cleaned using the same methodology as the one applied for the Belgian dataset. Prior to any analysis, the Canadian data were projected onto the Belgian PCA used to detect outliers to ensure that their variability was indeed included in the Walloon variability and that the Walloon results are applicable.

The cleaned Canadian dataset contains 670,165 records. The estimated descriptive statistics of Belgian and Canadian datasets along with the performances of the prediction models are mentioned in Table 5.1. The differences observed are mostly related to the differences between a pasture-based system in Wallonia (Belgium) compared to a more productive system in Canada, where fat supplementation is more common (Music et al. 2022).

Table 5.1: Descriptive statistics for the 39 FT-MIR-predicted traits used in the study calculated from the available Belgian and Canadian spectra.

FT-MIR Predicted Traits ¹	Unit	Belgium (<i>N</i> = 774,781)	Canada (<i>N</i> = 670,165)	<i>R_{cv}</i>	RMSE
Fat	g/dL	4.17 ± 0.34	4.14 ± 0.27	/	/
Protein	g/dL	3.47 ± 0.19	2.62 ± 0.13	/	/
Milk yield	kg/day	26.85 ± 2.85	29.97 ± 1.54	0.69	3.48
Energy balance		-3.12 ± 3.36	-7.05 ± 1.56	0.43	1.33
Nitrogen efficiency		53.67 ± 14.12	17.54 ± 1.25	0.52	1.44
Blood BHB	log(mmol/L)	-0.82 ± 0.09	-0.78 ± 0.06	0.70	1.85
Blood free FA	μeq/L	495.15 ± 131.57	421.89 ± 84.79	0.39	344.2
Dry matter intake	kg/day	22.81 ± 2.14	24.18 ± 1.41	0.45	1.35
C4	g/100 g	2.68 ± 0.19	2.63 ± 0.13	0.93	0.008
C6	g/100 g	1.81 ± 0.12	1.83 ± 0.08	0.91	0.006
C8	g/100 g	1.18 ± 0.10	1.28 ± 0.07	0.91	0.004
C10	g/100 g	2.65 ± 0.36	3.26 ± 0.25	0.92	0.010
C12	g/100 g	3.32 ± 0.44	4.09 ± 0.33	0.93	0.011
C14	g/100 g	11.43 ± 0.87	12.32 ± 0.65	0.94	0.030
C14:1cis9	g/100 g	1.06 ± 0.11	1.17 ± 0.08	0.71	0.008
C16	g/100 g	31.33 ± 3.25	28.90 ± 1.52	0.95	0.091
C16:1	g/100 g	1.61 ± 0.17	1.56 ± 0.09	0.73	0.013
C17	g/100 g	0.64 ± 0.05	0.66 ± 0.02	0.81	0.003
C18	g/100 g	9.53 ± 1.03	9.05 ± 0.61	0.84	0.056
Total C18:1trans	g/100 g	3.12 ± 0.75	3.62 ± 0.37	0.80	0.025
C18:1cis9	g/100 g	18.56 ± 2.63	19.98 ± 1.74	0.95	0.063
Total C18:1cis	g/100 g	20.04 ± 2.78	21.57 ± 1.86	0.95	0.061
Total C18:2	g/100 g	2.10 ± 0.22	2.54 ± 0.12	0.71	0.014
C18:2cis9cis12	g/100 g	1.25 ± 0.15	1.49 ± 0.11	0.75	0.011
C18:2cis9trans11	g/100 g	0.47 ± 0.10	0.61 ± 0.05	0.74	0.010
C18:3cis9cis12cis15	g/100 g	0.76 ± 0.33	0.96 ± 0.13	0.69	0.004
Saturated FAs	g/100 g	68.49 ± 4.09	66.88 ± 2.06	0.99	0.072
Monounsaturated FAs	g/100 g	26.78 ± 3.11	27.51 ± 1.97	0.97	0.059
Polyunsaturated FAs	g/100 g	3.46 ± 0.68	4.44 ± 0.28	0.79	0.021
Unsaturated FAs	g/100 g	30.33 ± 3.57	31.87 ± 2.15	0.97	0.064
Short-chain FAs	g/100 g	8.77 ± 0.60	9.32 ± 0.42	0.93	0.025
Medium-chain FAs	g/100 g	51.68 ± 3.95	51.89 ± 2.50	0.97	0.104
Long-chain FAs	g/100 g	38.50 ± 4.12	39.04 ± 2.72	0.95	0.110
Branched FAs	g/100 g	2.27 ± 0.26	2.67 ± 0.08	0.77	0.013
Omega3	g/100 g	0.58 ± 0.12	0.71 ± 0.05	0.68	0.006
Omega6	g/100 g	2.13 ± 0.24	2.58 ± 0.14	0.74	0.014
Odd-chain FAs	g/100 g	3.82 ± 0.35	4.30 ± 0.12	0.84	0.016
Total Trans FAs	g/100 g	3.91 ± 0.93	4.59 ± 0.46	0.82	0.029
Total C18:1	g/100 g	23.15 ± 3.08	23.70 ± 1.99	0.96	0.060

¹Traits in bold will be used in the following unsupervised analysis. FAs = fatty acids; BHB = beta-hydroxybutyrate; g/dL of milk; g/100g of fat.

Table 5.2: Herd characteristics for 3006 Canadian farms.

	Unit	Mean $\pm$ SD
Number of lactation cows	Cows	66.65 $\pm$ 50.87
Days in milk	days	176.99 $\pm$ 23.64
Margin on feed costs	\$CA/cow/year	5009.77 $\pm$ 1254.85
Margin on feed costs per kg of fat	\$CA/cow/year/kg	12.53 $\pm$ 1.52
Milk yield	L/day	26.59 $\pm$ 5.05
Fat	kg/cow/day	1.06 $\pm$ 0.30
Protein	kg/cow/day	0.86 $\pm$ 0.24
Milk production at lactation peak	L/day	39.72 $\pm$ 5.76
Days in milk at lactation peak	days	44.75 $\pm$ 4.54
Somatic cells in milk	$\times 10^3$ cells/mL	182.02 $\pm$ 116.21
% of cows in the herd with somatic cells count > 200,000 cells/mL	%	18.95 $\pm$ 7.92
Urea in milk	g/mL	5.01 $\pm$ 5.90
% of cows in the herd with a urea concentration < 5 or > 12	%	7.75 $\pm$ 9.90
Transition index		236.12 $\pm$ 445.85
% of cows in the herd with a negative transition index	%	37.98 $\pm$ 18.25
Age at first calving	month	25.30 $\pm$ 2.28
Calving interval	days	409.75 $\pm$ 31.72
Management index		-303.55 $\pm$ 1218.80
Management index for fat		-14.99 $\pm$ 52.65
% of the 3006 farms in conventional farming system	%	60.98
% of the 3006 farms in organic farming system	%	13.11
% of the 3006 farms with no information about their farming system	%	25.91

In addition to milk composition information, Lactanet has made additional information available to better understand herd management. All of these data are average values per herd, calculated for the 12 months prior to April 2022, and are available for 3006 farms. These traits and their related descriptive statistics are presented in Table 5.2. The transition index corresponds to the difference between expected animal production, based on production in the previous lactation, and the projected production of these animals, based on the production in the first test. This index is used to evaluate the strategy implemented for cows during the transition period. The management index provides information at the herd level from the average animal's environment and management effect (Lactanet 2022). The index at the animal level is based on the difference between the phenotype and the genetic effect.

Information on rations given in 2021 was also available for 540 herds for which Lactanet advisors provided feeding advice. This information includes the theoretical percentage of dry matter in the ration (49.97 ± 28.97) and the theoretical percentage of corn silage in the ration (22.85 ± 17.45). The latest dataset used, supplied by Lactanet, concerns the type presence of ventilation in the barn. These data were collected by survey on 2113 farms and indicate the presence or absence of additional ventilation in summer (Yes = 75.96%, No = 21.44%, and NA = 2.60%).

5.5.3. Unsupervised Learning

Since many of the problems that can be detected are linked to multiple FA deviations, and a single FA irregularity can be associated with various problems, a multivariate approach was necessary for a more comprehensive understanding of the relationships between FAs, and consequently, to better assess the overall situation of the herd. Given the lack of diagnostic data for Walloon herds, the innovation of this work lies in studying the relationships between variables to identify groups of herds with distinct FA profiles and to explain them. This study extends the work of Franceschini et al. (2022) on individual cow milk samples. Based on the predicted phenotypes, the literature, and the more extensive Canadian data, the identified clusters were interpreted, and their practical usefulness was assessed. From those results, the final objective was to study the feasibility of developing a specific decision-making tool, based on predicted FAs, to assist dairy farmers in their daily decision making.

5.5.4. Hierarchical Clustering

A hierarchical clustering algorithm was applied to the 31 Belgian FA predictions, with the aim of grouping together samples with similar characteristics. The function used was `hclust` from the Stats package version 3.4.1 (R Core Team 2022). This algorithm groups data according to their distance from each other. In the present study, the “ward.D2” method was used to calculate these distances between groups. The algorithm sums the squares of the distances between the data, then merges

the data in such a way that the intra-group dispersion is minimized (Murtagh et al. 2014). This method requires too much computer memory to be applied to the entire dataset. Therefore, a representative subset was created in 2 steps. Firstly, records with the greatest GH distances ($GH \geq 3$) were kept, as they are the most extreme ones reflecting milk compositions different from the average. This represented 12,322 samples. Second, to equilibrate normal and more extreme samples, we selected 5000 samples randomly in every other GH strata (from 0 to 1, from 1 to 2, and from 2 to 3). So, the final subset was composed of 27,322 records. The number of clusters selected was a function of the cluster height, visible on the cluster dendrogram, as well as the distance differential with respect to the previous merge. The visualization of the first two PCs was used to illustrate the position of the various clusters. The 8 other indicators available in the database were added as additional features in this PCA to help the interpretation of these results.

5.5.5. Cluster Prediction

To facilitate the prediction of clusters in the available datasets, Partial Least Square Discriminant Analysis (PLS-DA) and Random Forest (RF) models were applied to the subset for which the cluster labels were known. The PLS-DA was estimated using the Caret package, version 6.0-90 (Kuhn 2008), while the RF model was estimated using the randomForest package, version 4.7-1.1 (Breiman 2001). PLS-DA was selected because some FA traits are correlated. The analysis was performed on 31 centered and scaled FAs, with a maximum of 30 components. The predicted cluster is the one with the highest probability of membership. A ten-fold cross-validation was used to determine the optimal number of PLS components and to assess the classification's performance. Performance was measured using global accuracy and Cohen's Kappa coefficient, which accounts for agreement occurring by chance. The random forest model was chosen to capture potential non-linear relationships. The Gini index was used as the splitting criterion, and the number of trees was set to 500. The maximum number of features was optimized through cross-validation, following the same methodology used for PLS-DA.

Since the 31 FAs were predicted by FT-MIR spectrometry, PLS-DA and RF models were also created, directly using the spectra to predict the clusters. This approach enhances the model's transferability by eliminating the FA prediction step, which may vary depending on the lab or equation model used. A first derivative was applied to the spectra, and 212 spectral points were utilized as suggested by Grelet et al. (2021). The modeling methodology was the same as that used for the model based on the 31 FAs. After cross-validation, cluster predictions using PLS-DA and RF from FAs and from spectra were performed on all available Belgian and Canadian datasets.

5.5.6. Interpretation

To interpret the clusters, the means and standard deviations of each FA and other available sources of information were calculated according to the predicted clusters. This allowed us to compare the clusters.

The means associated with each cluster are essential for understanding the underlying patterns, but the frequency of each cluster and the frequency of transitions between clusters are also crucial for interpretation. For this reason, the transition matrix between clusters was estimated. For each observation in a specific cluster at time t , we measured the probabilities of transitioning to every cluster at time $t + 1$. In this approach, transitions between clusters were considered in a binary manner: from cluster 1 to 1, from 1 to 2, and so on. This allows for consideration of the dynamics between clusters.

Since the emergence of a problem in a herd is rarely spontaneous, the transition from one cluster to another may be too abrupt as a monitoring tool. Therefore, it is important to use quantitative rather than qualitative information. In this study, the probabilities of cluster membership obtained from PLS-DA and RF were also considered as a monitoring tool. To validate their usefulness, we calculated the correlations between cluster probabilities and FAs and indicators available in all databases. These probabilities were further used to visualize the evolution of herd conditions over time.

Finally, using additional validation data provided by Lactanet on animals, herd breakdown, and feeding, the herds found in each cluster were analyzed to determine whether it was possible to identify any trends in the type of animals or management. This was carried out to explain certain clusters, but also to confirm hypotheses about the herd status in the different clusters.

5.6. Results and Discussion

5.6.1. Clustering on Belgian Dataset

Seven clusters were retained after clustering the selected subset ($N = 27,322$) based on the separation observed in the dendrogram (Figure 5.3). The distribution of observations across the clusters is strongly influenced by the GH value (Table 5.3), confirming the hypothesis that rare observations with high GH values must be considered to understand the overall picture. Moreover, the less frequent clusters associated with high GH values could be linked to stress within the herd. Indeed, it is reasonable to assume that stressed herds would exhibit more extreme patterns with lower frequencies in the dataset. However, further analysis is necessary to confirm this hypothesis.

The records for this subset were projected onto the first two PCs which together explained 74.3% of the whole dataset's variance, as shown in Figure 5.4. The first

Table 5.3: Proportions of the seven clusters for the observations from the subset ($N = 27,322$).

	Cluster						
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
All data	23.55	7.90	11.72	22.75	15.89	10.81	8.71
GH < 3	27.53	2.29	4.07	28.51	15.64	5.48	13.47
GH > 3	17.98	12.51	21.04	15.73	12.54	17.29	2.91

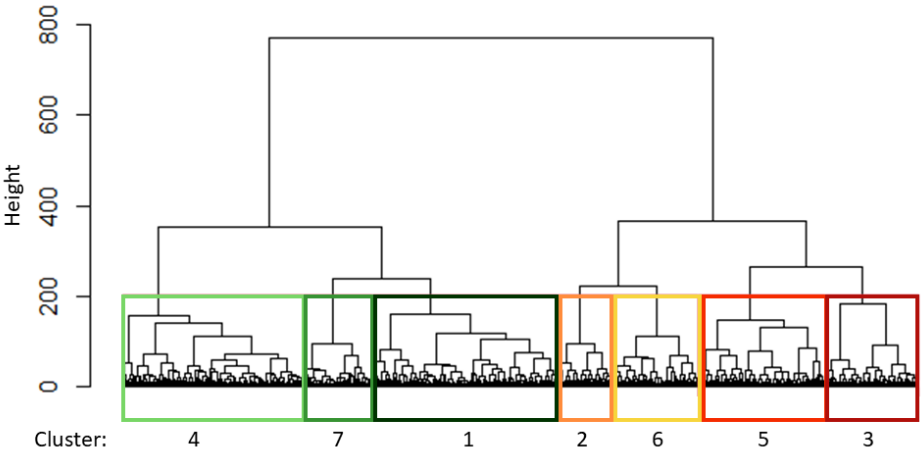


Figure 5.3: A dendrogram of the seven clusters found from the selected subset.

hypothesis about the differences in performance could be made regarding the non-linearity of the relation between some FAs and the clusters. The decrease in accuracy for RF when switching to spectra could be linked to the high dimensionality and the higher correlation between spectral wavelength absorption than between FAs.

For the following analysis of this paper, the RF model based on the 31 FAs was used to predict the cluster for the whole dataset, as it is the model with the highest global accuracy (91.81%), so most of the observations are correctly classified and misclassifications primarily occur in adjacent clusters, which can be explained by the cluster continuum shown in Figure 5.4 and Table 5.4.

Table 5.4: Confusion matrix related to random forest algorithm applied on 31 predicted fatty acids ($N = 27,322$).

Prediction	Reference						
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	5781	122	0	279	100	113	59
Cluster 2	78	1651	24	0	30	57	0
Cluster 3	0	43	2979	0	105	72	0
Cluster 4	287	0	0	5875	0	36	46
Cluster 5	50	36	111	1	4001	48	62
Cluster 6	96	33	90	36	46	2623	8
Cluster 7	53	0	0	24	59	4	2204

5.6.2. Cluster Prediction on Belgian and Canadian Datasets

Before making any predictions using the Canadian dataset, it was essential to verify that the variability in the Canadian data is encompassed within the variability in the Belgian data used for clustering. Figure 5 illustrates the projection of the Canadian data onto the first two PCs initially developed from the Belgian database, confirming that the variability in the Canadian data is included within the Belgian FA data's variability. Consequently, the RF model developed using the Belgian FA data can be applied to the Canadian data.

Using the developed RF algorithm, cluster membership was predicted for all Belgian ($N = 774,781$) and Canadian ($N = 670,165$) records. Table 5.5 presents the proportion of records by cluster for both countries, illustrating differences not only between Belgium and Canada but also compared to the GH (Table 5.4). We assumed that higher GH milk spectra are over-represented in the Belgian subset and that differences between countries may be due to the variability differences displayed in Figure 1.6.

First, there is minimal or no representation of Canadian data in Clusters 2, 3, 4, and 6. Second, the primary clusters in Belgium are Clusters 1 and 4, with Clusters 5 and 7 also occurring frequently, whereas Cluster 7 is predominant in Canada, with Clusters 1 and

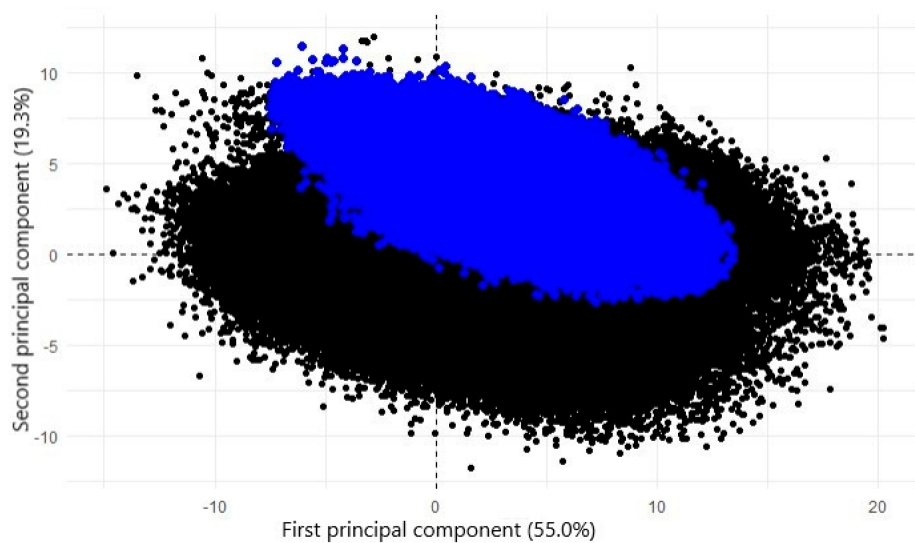


Figure 5.5: Projection of Canadian data (blue) onto graph of individuals (black) on first two principal components estimated based on whole Belgian database.

Table 5.5: Percentages of Belgian and Canadian records by cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Belgium	38.47	1.18	1.04	36.60	10.40	2.51	9.80
Canada	12.00	0.28	0.04	0.48	13.22	< 0.01	73.98

5 following. One hypothesis, therefore, is that Clusters 1 and 4 represent an average situation in Belgium, while Cluster 7 represents an average situation in Canada. As most herds in a population are expected to be well managed, this would be the normal situation. At this step, more information such as the FA profiles is required to deepen interpretations.

To further explore the dynamics between clusters, transitions from one cluster to another were computed (Table 5.6). This table gives the proportion of records going from Cluster A at time t to Cluster B at time $t + 1$. In Belgium, Clusters 1, 4, and 5 have the highest proportion—around 65%—of successive records remaining within the same cluster. Clusters 1 and 4 appear to be interrelated, as approximately 20% of the records transition between them. A significant proportion of records from Cluster 5 moves to Cluster 1. In contrast, only a small proportion of records from Clusters 1 or 4 transit to Cluster 5. Instead, records in Cluster 5 predominantly come from Clusters 2, 3, and 6. Cluster 7 is less stable and less represented than Clusters 1 and 4, but it seems to be more related to these clusters than to others. In contrast, Cluster 3 rarely transits to Clusters 1 or 4. This work was developed together with Claire Fastré during her master thesis (Fastré 2022).

Table 5.6: Cluster transitions between 2 successive controls, expressed in %, for data collected in Belgium and Canada.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Belgium	Cluster1	65.63	1.26	0.09	20.19	5.97	1.69	5.17
	Cluster2	40.24	32.52	2.32	4.97	10.86	8.90	0.20
	Cluster3	2.78	2.85	55.75	0.45	23.67	14.33	0.17
	Cluster4	22.02	0.14	0.01	69.81	0.43	0.53	7.06
	Cluster5	20.68	1.23	2.25	1.62	64.65	3.00	6.56
	Cluster6	24.01	4.21	6.55	8.02	13.18	43.42	0.62
	Cluster7	19.83	0.02	0.01	27.80	5.54	0.12	46.68
Canada	Cluster1	34.39	0.28	0.01	0.31	11.10	0.00	53.92
	Cluster2	11.51	18.63	0.70	0.00	60.87	0.00	8.30
	Cluster3	0.83	7.44	23.35	0.00	62.81	0.00	5.37
	Cluster4	7.10	0.00	0.00	35.09	0.69	0.00	57.12
	Cluster5	10.27	1.24	0.17	0.02	63.08	0.00	25.22
	Cluster6	0.00	0.00	0.00	0.00	83.33	0.00	16.67
	Cluster7	8.79	0.04	0.00	0.38	4.43	0.00	86.35

¹Each line is first defined by a cluster at time t , and then the other columns represent the probabilities of moving into the cluster at time $t + 1$. Clusters in bold represent less than 1% of the dataset.

Finally, Clusters 2 and 6 are the least stable, likely associated with temporary events. This could suggest that herds in Cluster 3 have experienced a deterioration in animal welfare, with herds remaining in this cluster longer than in others. This may indicate a problematic state and greater difficulty returning to a normal status. There are more transitions to Cluster 7 from Clusters 1 and 4, which are considered healthy clusters, than from Clusters 2, 3, 5, and 6, which are considered problematic. These observations suggest that Clusters 1 and 4 represent typical conditions in Wallonia, with Clusters 5 and 7 as intermediate clusters, and Clusters 2, 3, and 6 as potential problem clusters related to stress. If these hypotheses are correct and these clusters indeed reflect stress states, this dynamic would be expected.

In Canada, Cluster 7, which seems to be the standard, is also related to Clusters 1 and 4, even though Cluster 4 is not very represented in this dataset. Cluster 5 observations mostly stay in Cluster 5, and observations from Clusters 2, 3 and 6 go back to Cluster 5.

5.6.3. Cluster Interpretation

To deepen the interpretation of clusters, the means by cluster were calculated for phenotypes predicted by FT-MIR from bulk tank milk collected in Belgium and Canada (Table 5.6). To facilitate reading of the results, a color code was applied. The best and second-best values are in green and those with the worst values are in red. Globally, the results are similar for Belgium and Canada. The major differences between Belgium and Canada for FAs concern C18, polyunsaturated FAs, branched FAs, and odd FAs. For the supplementary variables, the protein, protein efficiency, energy balance, and blood free FAs are globally lower in Canada while the milk yield and the dry matter intake are higher. However, the performance of the equation for energy balance and blood free FAs is low, so the differences might be irrelevant. The maximum fat is also higher in Canada. Generally, Cluster 4 presents the best values in Wallonia followed by de-novo-FA-dependent Clusters 7 and Cluster 1, which have good values on average. Cluster 3 is the worst, followed by Cluster 6 and Cluster 2. Like Cluster 1, in terms of the “healthy” clusters, Cluster 5 seems to be intermediate. The same trends are observed for Canada. However, Clusters 2, 3, 4, and 6 are not very well represented in Canada. In a more controlled production and management system, these clusters may occur very rarely but may indicate more severe cases. Out of Clusters 1, 5, and 7, which represent 99.2% of the population, Cluster 5 would be the problematic cluster, with a potential transition to severe clusters.

So, Cluster 1 is the most represented in Belgium, followed by Cluster 4. These are the “standards” with better values for Cluster 4. Cluster 7, the most represented in Canada, has also good overall values. The difference in “standard” clusters between Canada and Belgium, and the small representation of some clusters, can probably be explained by the differences in farming systems between the two regions. Indeed,

Table 5.7: Means by cluster of the phenotypes predicted by FT-MIR from bulk tank milk collected in Belgium and Canada as well as temperature humidity index (THI) for Belgium. Two best values are high-lighted in shades of green, and worst in shades of red according to their rankings. Clusters for Quebec, in bold, represent less than 1% each.

	Belgium							Canada						
	C1	C2	C3	C4	C5	C6	C7	C1	C2	C3	C4	C5	C6	C7
C4	2.78	2.87	2.49	2.69	2.55	2.6	2.5	2.79	2.85	2.63	2.75	2.64	2.55	2.61
C6	1.82	1.7	1.46	1.86	1.7	1.56	1.84	1.8	1.68	1.53	1.97	1.74	1.53	1.85
C8	1.16	1.01	0.88	1.23	1.11	0.94	1.26	1.22	1.08	0.97	1.39	1.18	0.95	1.3
C10	2.53	2.06	1.78	2.77	2.51	1.87	3.06	3.07	2.53	2.23	3.6	2.9	2.38	3.36
C12	3.15	2.51	2.3	3.51	3.13	2.44	3.78	3.83	3.1	2.8	4.53	3.62	3.19	4.22
C14	11.13	9.56	9.05	12.07	10.55	9.92	11.9	11.8	10.27	9.66	13.28	11.36	11.06	12.58
C14:1c9	1.01	0.84	1	1.12	1.05	1.05	1.09	1.09	0.96	1.02	1.24	1.11	1.2	1.19
C16	31.04	27.16	24.26	34.38	25.78	30.01	29.31	28.79	25.54	23.53	32.62	26.58	29.84	29.33
C16:1	1.58	1.69	1.95	1.59	1.7	1.92	1.6	1.53	1.65	1.83	1.45	1.65	1.72	1.55
C17	0.64	0.66	0.72	0.61	0.7	0.66	0.7	0.64	0.66	0.7	0.62	0.68	0.68	0.66
C18	10.02	11.48	10.44	9.04	9.56	9.93	9.18	9.49	10.66	10.57	8.36	9.67	8.73	8.87
Total C18:1t	3.21	3.71	4.44	2.5	4.21	3.13	3.58	3.49	4.01	4.52	2.97	4.05	3.6	3.57
C18:1c9	19.06	23.82	26.28	16.56	21.81	23.26	17.98	21.18	25.25	27.26	16.46	22.62	23.21	19.31
Total C18:1c	20.61	25.69	28.17	17.97	23.39	25.07	19.27	22.9	27.22	29.29	17.8	24.36	24.78	20.85
Total C18:2	2.12	2.4	2.42	1.92	2.39	2.08	2.28	2.55	2.7	2.8	2.23	2.65	2.33	2.52
C18:2c9c12	1.28	1.47	1.2	1.22	1.2	1.21	1.26	1.56	1.61	1.55	1.34	1.5	1.41	1.47
C18:2c9t11	0.46	0.53	0.63	0.38	0.63	0.45	0.57	0.57	0.62	0.69	0.52	0.65	0.5	0.6
C18:3c9c12c15	0.72	0.8	1.38	0.5	1.33	0.79	1.08	0.85	1	1.26	0.76	1.12	1.05	0.96
SFA	68.37	63.15	57.54	71.93	61.53	64.06	67.24	66.25	61.39	58	71.84	63.48	64.81	67.59
MUFA	27.38	32.76	36.15	24.32	30.99	32.12	26.2	28.43	33.08	36.03	23.24	30.71	30.53	26.78
PUFA	3.46	4.01	4.6	2.86	4.56	3.36	4.19	4.29	4.67	5.11	3.78	4.76	4.17	4.41
UFA	30.91	36.81	40.92	27.29	35.65	35.71	30.4	32.67	37.69	41.11	27	35.41	34.67	31.11
SCFA	8.71	7.92	6.88	9.06	8.24	7.35	9.04	9.12	8.3	7.54	10.08	8.75	7.59	9.46
MCFAs	50.62	44.07	41.79	55.53	45.67	48.2	51	50.44	44.6	42.28	56.92	48.08	50.26	52.81
LCFA	39.84	47.81	49.86	34.8	43.67	44.21	38	40.64	47.12	50.03	33.48	43.3	42.18	38.01
BFA	2.21	2.2	2.66	2.11	2.7	2.26	2.54	2.58	2.6	2.77	2.49	2.73	2.6	2.67
Omega3	0.59	0.68	0.77	0.47	0.76	0.56	0.7	0.68	0.76	0.82	0.61	0.77	0.6	0.71
Omega6	2.17	2.48	2.36	1.95	2.36	2.03	2.34	2.61	2.73	2.78	2.3	2.66	2.34	2.56
Odd chain FA	3.74	3.73	4.29	3.59	4.34	3.75	4.26	4.14	4.1	4.35	4.1	4.34	4.41	4.32
Total Trans FA	4	4.58	5.54	3.16	5.32	3.88	4.56	4.31	4.92	5.62	3.78	5.11	4.33	4.54
Total C18:1	23.85	29.4	32.28	20.64	27.24	28.27	22.49	24.76	29.51	32.18	19.44	26.9	26.35	22.94
Fat	4.11	3.94	4.01	4.23	4.1	4.05	4.25	4.17	4.02	3.99	4.84	4.02	3.83	4.15
Protein	3.44	3.31	3.43	3.5	3.5	3.37	3.55	2.55	2.44	2.5	2.9	2.56	2.55	2.65
Milk yield	26.87	26.1	22.45	28.11	24.49	24.81	25.52	30.83	29.82	26.55	30.9	28.56	24.77	30.08
EB	-2.75	-5.4	-8.5	-1.43	-7.4	-4.83	-4.04	-7.87	-9.51	-10.14	-6.95	-8.22	-6.94	-6.69
NUE	56.67	56.29	31.04	58.1	38.73	41.78	49.67	18.7	19.76	19.75	17.96	18.24	20.14	17.22
Blood BHB	-0.81	-0.74	-0.71	-0.87	-0.73	-0.77	-0.79	-0.75	-0.67	-0.68	-0.89	-0.73	-0.77	-0.8
Blood free FA	526.9	678.6	714.6	407.7	590.4	629.0	520.3	449.3	597.8	622.6	293.8	514.4	496.0	400.8
DMI	22.24	19.93	19.78	23.51	22.03	20.84	24.17	23.49	21.45	21.27	27.31	23.03	24.67	24.49
THI	52.27	52.45	56.71	50.28	54.67	56.31	49.12	/	/	/	/	/	/	/

SFAs = saturated FAs, MUFAs = monounsaturated FAs, PUFAs = polyunsaturated FAs, UFAs = unsaturated FAs, SCFAs = short-chain FAs, MCFAs = medium-chain FAs, LCFAs = long-chain FAs, BFAs = branched FAs, EB = energy balance, NUE = Nitrogen Use Efficiency, BHB = β -ta-hydroxybutyrate, FFAs = free FAs, DMI = dry matter intake. THI was not calculated for the Canadian dataset, as meteorological data were not available.

Quebec has few farms practicing grazing, unlike Wallonia. In Quebec, most farms use stanchion barns for lactating cows, whereas Walloon farms generally use loose housing when the animals are not in pasture (NFACC 2020; Arnott et al. 2015). The herd size in Quebec is 73 cows, and each animal gives an average of over 9300 L of milk per year (Québec 2025). In Wallonia, there are 64 cows per herd, and the milk production is only 6600 L per year on average (SPW 2025). Moreover, those clusters are probably related to different types of feeding management. The interpretation is complex because many factors such as the type of diet, amount of concentrates, and fat supplementation are interacting with each other. The hypothesis is that Clusters 1, 4, and 7 correspond respectively to fresh grass, grass silage, and maize silage, potentially with concentrates and fat supplementation. Indeed, the grass silage diet is associated with higher C14 and C16 and lower C18 mono- and polyunsaturated FAs. The maize silage diet tends to increase the C6, C8, C10, and C12 found in Cluster 7 while the fresh grass diet favors an increase in C4 and C18:1cis9 (Chilliard et al. 2001). Among the “healthy” clusters, Cluster 1 has the highest values for those traits. Concentrates are potentially found in Clusters 4 and 7 because they increase the percentages of C10 and C12 but also trans monounsaturated FAs and C18:2cis9cis12 (Chilliard et al. 2001). Fat supplementation differs depending on the type of lipid. However, it is associated with an increase in polyunsaturated FAs and an inhibition of *de novo* FAs Chilliard et al. (2001), which can explain the difference between Wallonia and Quebec for PUFAs and the inversion of the cluster order for *de novo* FAs between Clusters 4 and 7.

The hypothesis for Clusters 2, 3, 5, and 6 is that they include herds in a state of stress or poor health. Indeed, these clusters generally present the worst values for at least five traits. However, there are differences between these four clusters. First, Cluster 3 is the worst cluster, and is associated with extreme values for most of the traits. It is difficult to interpret the exact issue, as all the traits are extreme, but it is probably related to energy balance or ruminal acidosis. Moreover, Cluster 3 also presents a low value for the predicted milk yield. Cluster 1, which is considered related to pasture or fresh grass, is similar to Clusters 2 and 6. However, Cluster 1 presents good values, unlike Clusters 2 and 6. After Cluster 3, Cluster 6 has the worse values for short-chain FAs, C16:1, and protein, and high values of THI. The low value for *de novo* FAs could indicate that there is an issue with rumen health, but for rare events, involving around 2.5% of the observations, such as dietary imbalance. Cluster 5, which has low values for C16 and high values for branched FAs, trans FAs, and odd-chain FAs, is also probably related to rumen health affected by the feeding strategy, such as being given a high-concentrate diet. This cluster is more common, involving 10% of the observations, and has low values for milk yield and energy balance, so it could be a combination with an energy balance issue. For those two clusters, farmers should monitor rumen health and adapt consequently. Cluster 2 has low values for fat, protein, unsaturated FAs, C14, C14:1c9, and DMI. It has also high values for blood free FAs and for long-chain FAs, more specifically C18 and C18:1c9, which is a sign of fat

mobilization and negative energy balance. These clusters are essential, as they are transitioning to Cluster 3, which is an abnormal herd status. At this step, on-farm information is required to obtain a deeper interpretation.

Even if we have a high percentage of missing information, increasing the difficulty of drawing general conclusions, the management indicators provided by Lactanet should be useful information to further understand the clusters. Table 5.8 shows a higher percentage of organic farms and fewer herds with additional ventilation in Clusters 2 and 6 than in Clusters 1, 3, and 5. There is a significant difference between clusters in feed intake, with a very low proportion of corn silage in the rations of Clusters 2 and 3 (Table 5.9).

If we look at the transition index of these clusters, which represents the average herd's performance at the start of lactation compared with what was expected of it, we observe that Clusters 2 and 3 have the lowest values, with negative values for Cluster 3. Cluster 5 is also slightly lower than the others. In other words, these herds showed lower performance in the first lactation than expected, suggesting a bad transition strategy between dry-off and early lactation. The milk management index, which estimates the environmental effect, shows negative values, but with very large standard deviations for all clusters. These indices are less negative for Clusters 1, 4, and 7. This trend is also found for the fat management index, suggesting that Clusters 2, 3, and 5 represent undesirable situations in Canada.

Table 5.10 shows herd indicators related to management, production, reproduction, and sanitary status averaged over twelve months. More than 50% of the Canadian herds in each cluster had data available for these different indicators. The data for Clusters 1 and 5 are very close or identical for most indicators, and they present the best values. Clusters 4 and 6 also had close, but worse values.

Table 5.8: Percentages of organic and conventional farms in the various clusters, in % of herds present in the cluster.

	Level	Cluster 1 (<i>N</i> = 4330)	Cluster 2 (<i>N</i> = 520)	Cluster 3 (<i>N</i> = 83)	Cluster 4 (<i>N</i> = 577)	Cluster 5 (<i>N</i> = 3250)	Cluster 6 (<i>N</i> = 4)	Cluster 7 (<i>N</i> = 4639)
Farm system	Organic	8.43	11.92	3.61	12.13	8.77	0.00	8.26
	Conventional	39.33	25.77	16.87	40.56	36.15	25.00	39.21
	Unknown	52.24	62.31	79.52	47.31	55.08	75.00	52.53
Additional ventilation	No	75.83	84.75	71.43	77.91	77.19	/	75.95
	Yes	21.55	13.56	28.57	18.99	20.48	/	21.44
	Unknown	2.62	1.69	0.00	3.10	2.33	/	2.61

Data available for 2227 herds out of 4675 herds for Canada, *N* = 13,403 herds × cluster.

Table 5.9: Percentages of different feeding in the various clusters, in % of herds present in the cluster.

		Cluster 1 (<i>N</i> = 498)	Cluster 2 (<i>N</i> = 27)	Cluster 3 (<i>N</i> = 2)	Cluster 4 (<i>N</i> = 66)	Cluster 5 (<i>N</i> = 319)	Cluster 6 (<i>N</i> = 0)	Cluster 7 (<i>N</i> = 529)
Feeding	% of dry matter	50.03	31.33	0.74	49.53	46.56	/	49.83
	% of corn silage	22.44	8.63	2.04	33.98	18.22	/	22.79

Data available for 2227 herds out of 4675 herds for Canada, *N* = 13,403 herds × cluster.

Table 5.10: Means and standard deviations by cluster of management related data for 3006 Quebec herds.

Traits	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 7
Number of cows in lactation	65.32 ± 41.63	46.04 ± 17.10	40.55 ± 28.32	66.22 ± 34.38	58.13 ± 33.89	66.06 ± 42.37
Days in milk	174.41 ± 20.30	177.82 ± 31.35	234.75 ± 74.50	174.26 ± 19.91	176.55 ± 22.29	174.94 ± 20.31
MFEED (\$CA/cow/year)	5134.67 ± 959.26	4643.00 ± 1056.39	3074.75 ± 1120.77	4932.20 ± 1056.97	5011.12 ± 1005.99	5129.53 ± 945.20
MFEED per fat yield (\$CA/cow/year/kg of fat)	12.59 ± 1.17	12.83 ± 1.80	11.16 ± 0.87	12.42 ± 1.30	12.59 ± 1.35	12.58 ± 1.14
Milk yield (L/cow/day)	26.88 ± 4.58	24.61 ± 4.68	18.57 ± 5.99	25.69 ± 5.26	26.32 ± 4.66	26.83 ± 4.56
Fat (kg/cow/day)	1.08 ± 0.27	0.96 ± 0.25	0.72 ± 0.18	1.07 ± 0.28	1.04 ± 0.28	1.08 ± 0.26
Protein (kg/cow/day)	0.89 ± 0.22	0.79 ± 0.21	0.62 ± 0.19	0.87 ± 0.23	0.86 ± 0.23	0.89 ± 0.22
Milk at lactation peak (L/day)	40.09 ± 5.02	37.12 ± 5.21	32.35 ± 5.92	38.46 ± 6.35	39.39 ± 5.08	40.05 ± 5.00
Days in milk at lactation peak	44.30 ± 4.37	42.85 ± 5.92	37.50 ± 10.47	42.43 ± 4.70	44.41 ± 4.73	44.25 ± 4.34
Somatic cells (10 ³ cells/mL)	175.49 ± 106.42	183.50 ± 108.72	278.25 ± 83.43	177.64 ± 110.73	181.34 ± 111.68	177.34 ± 106.55
Somatic cells > 200,000/mL (% of cows/herd)	18.57 ± 7.44	20.80 ± 8.34	31.27 ± 9.15	19.54 ± 8.33	19.41 ± 7.52	18.69 ± 7.49
Urea (g/mL)	9.36 ± 5.33	7.47 ± 5.48	5.08 ± 5.89	9.10 ± 5.71	9.22 ± 5.59	9.42 ± 5.33
Urea < 5 or > 12 g/mL (% of cows/herd)	7.50 ± 7.29	7.81 ± 7.00	6.95 ± 5.25	9.26 ± 8.72	7.98 ± 7.63	7.50 ± 7.22
Transition index	275.94 ± 406.05	53.90 ± 469.31	-406.75 ± 512.39	274.55 ± 429.90	200.70 ± 413.45	279.08 ± 401.74
Transition index < 0	35.90 ± 16.25	45.88 ± 19.16	67.5 ± 17.71	36.06 ± 17.17	38.75 ± 16.95	35.72 ± 16.05
Age at first calving (months)	24.99 ± 1.94	26.09 ± 3.27	30.15 ± 8.28	25.12 ± 2.17	25.23 ± 2.21	24.99 ± 1.92
Calving interval (days)	403.79 ± 26.51	409.22 ± 37.63	457.50 ± 95.24	404.47 ± 23.50	406.06 ± 30.01	404.24 ± 26.79
% of involuntary culling	19.43 ± 8.89	18.46 ± 8.19	20.94 ± 14.20	17.20 ± 7.79	19.83 ± 9.43	19.51 ± 8.90
% of dead cows	5.25 ± 4.63	4.78 ± 4.46	3.69 ± 4.90	5.91 ± 5.59	5.37 ± 4.94	5.33 ± 4.61
Milk MI	-289.96 ± 1223.41	-798.47 ± 1456.19	-2423.77 ± 1503.39	-358.10 ± 1256.43	-466.23 ± 1284.08	-299.49 ± 1215.20
Fat MI	-14.49 ± 52.79	-40.84 ± 62.60	-107.18 ± 55.86	-14.57 ± 55.53	-23.05 ± 54.50	-14.86 ± 52.25

MI = management index; MFEED = margin on feed costs.

5.6.4. From Clusters to Probabilities

We showed that the clusters are related to herd management and could be of interest as a monitoring tool. However, we can improve the quality of the information given by using the probability of belonging to a cluster instead of a binary value. These probabilities are already a by-product of the RF algorithm, if we consider the proportions of trees that predict an observation into each cluster. These proportions respect the properties of probability. While these probabilities are correlated, which is expected, as the clusters overlap in Figure 5.4 and the underlying biological process is not discrete, the information they provide is useful for developing a monitoring tool on a routine basis. This information is mainly related to feeding management practices at the herd level and is available every one to three days. The FT-MIR predictions of the clusters and their probabilities become time series where extreme events can be detected and sent to the farmer.

An example of such a time series is represented in Figure 5.6. Every probability varies in time, and the switches between clusters happen when the cluster with the highest probability changes. This is typically the case when a peak occurs for specific issues. However, the peak is rarely based on one milk sample, so the increase in probability could be an early method of detecting management issues. In Belgium, the most observed clusters are Clusters 1 and 4, with some potential issues detected with Clusters 2 and 5. For Canada, it is mostly Cluster 7, then Cluster 1 and Cluster 5 for some potential issues. Those probabilities could be the sources of feedback for management.



Figure 5.6: Evolution of cluster probabilities in time series for one farm in Belgium (A) and one farm in Canada (B).

5.6.5. Practical Implications and Study Limitations

The findings of this study suggest that FA profiles, summarized through a clustering approach, could be implemented as a herd-level monitoring tool for farmers. This tool would function as an alert system integrated into herd-monitoring software, where an alert would indicate the detection of an anomaly without necessarily identifying a precise cause. The alert system would not replace an expert opinion but could encourage the farmers to ask for help. Based on the cluster interpretation in this study, it appears that, even though the standard clusters differed between Belgium and Canada, likely due to different farming systems, Cluster 3 was consistently associated with abnormal FA profiles in both countries, even if Cluster 3 is not common in Canada. In Belgium, the observed dynamics between clusters showed transitions from intermediate clusters to Cluster 3, suggesting that monitoring the probability of belonging to these intermediate clusters could be used to detect early warning signals and mitigate health issues before escalation. The simplest solution would be to consider the cluster membership. So, Cluster 3 could be a red flag, while Cluster 2, 5, and 6 could be orange flags, and the others green flags. The red flag, which is severe, suggests that the farmer should contact an advisor or a veterinarian, while an orange flag suggests that they should evaluate the feed storage, the diet composition, or the fat supplementation, or check for a heat stress event and check the farm mitigation system.

However, the study presents several limitations. First, the interpretation of clusters remains complex due to the numerous interactions between FA profiles and management practices. Second, cluster validation is required using on-farm data and real-world health observations. While pattern detection from large FT-MIR prediction databases is a powerful tool, it cannot replace validation using reference measurements and expert field knowledge. Third, the alert methodology itself must be further fine-tuned to optimize the signal derived from the time series data. Fourth, practical deployment of such a system depends on a data infrastructure capable of delivering alerts rapidly enough to inform timely decision making at the farm level. These limitations could be addressed through pilot trials on commercial farms, where feedback from farmers is available when the clustering approach flags potential issues. Such trials would support better interpretation of the clusters, confirm their utility, fine-tune the alert system, and help develop an experimental real-time data pipeline that could later be scaled for routine use. This would also allow for a first use case and for an assessment of its usefulness.

5.7. Conclusions

This study extends a previously developed clustering methodology—originally designed for individual milk samples—to bulk tank milk samples. The results suggest

that the FT-MIR variability in Belgian herds encompasses that of Canadian herds, potentially reflecting a broader range of feeding systems in Belgium. Our findings indicate that this methodology could serve as a cost-effective tool for generating actionable indicators from existing milk payment data without imposing additional burdens on farmers. Notably, the transition from intermediate clusters to Cluster 3 appears promising as a reliable health-monitoring indicator. Integrating this approach into routine herd management may enhance the capacity to respond proactively to management issues. The final automated alert system should be merged with an individual level alert to improve the feedback given to farmers. However, field validation is required before widespread implementation, particularly to adapt the approach across diverse farming systems and geographical contexts.

5.8. Acknowledgments

The authors acknowledge the support of Lactanet and the University Laval for hosting one co-author during its internship in Canada. During the preparation of this manuscript/study, the author(s) used ChatGPT 4.5 as a writing assistant to refine phrasing and improve readability. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Bibliography

- Agromet.be (2025). URL: <https://agromet.be/fr/pages/home/> (visited on 03/25/2025).
- Arnott, G., C. Ferris, and N. O'Connell (2015). *A comparison of confinement and pasture systems for dairy cows: What does the science say?* AgriSearch.
- Bauman, D.E., Mather, I.H., Wall, R.J., Lock, and A.L. (2006). "Major Advances Associated with the Biosynthesis of Milk". In: *J. Dairy Sci.* 89 1235–1243. DOI: 10.3168/jds.S0022-0302(06)72192-0.
- Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.
- Chilliard, Y., A. Ferlay, and M. Doreau (2001). "Effect of different types of forages, animal fat or marine oils in cow's diet on milk fat secretion and composition, especially conjugated linoleic acid (CLA) and polyunsaturated fatty acids". In: *Livestock Production Science*. Fifth International Workshop on the Biology of Lactation in Farm Animals 70.1, pp. 31–48. ISSN: 0301-6226. DOI: 10.1016/S0301-6226(01)00196-8.
- Chilliard, Y., A. Ferlay, R. M. Mansbridge, and M. Doreau (2000). "Ruminant milk fat plasticity: nutritional control of saturated, polyunsaturated, *trans* and conjugated fatty acids". In: *Annales de Zootechnie* 49.3, pp. 181–205. ISSN: 0003-424X, 1297-9651. DOI: 10.1051/animres:2000117.
- Craninx, M., Steen, A., Van Laar, H., Van Nespen, T., Martín-Tereso, J., de Baets, B., Fievez, and V. (2008). "Effect of Lactation Stage on the Odd- and Branched-Chain Milk Fatty Acids of Dairy Cattle Under Grazing and Indoor Conditions". In: *J. Dairy Sci.* 91 2662–2677. DOI: 10.3168/jds.2007-0656.
- FAO. (2023). "Contribution of terrestrial animal source food to healthy diets for improved nutrition and health outcomes – An evidence and policy overview on the state of knowledge and gaps". In: *Rome*. DOI: 10.4060/cc3912en.
- Fastré, C. (2022). "Contribution au développement d'un outil d'aide à la décision pour l'élevage laitier à partir de la connaissance du profil en acides gras du lait prédit par la spectrométrie infrarouge". Master's thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/14927>.

- Fiore, E., F. Blasi, M. Morgante, L. Cossignani, T. Badon, M. Giancesella, B. Contiero, and M. Berlanda (2020). "Changes of milk fatty acid composition in four lipid classes as biomarkers for the diagnosis of bovine ketosis using bioanalytical Thin Layer Chromatography and Gas Chromatographic techniques (TLC-GC)". In: *Journal of Pharmaceutical and Biomedical Analysis* 188, p. 113372. ISSN: 0731-7085. DOI: 10.1016/j.jpba.2020.113372.
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt (2022). "Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?" In: *Journal of Dairy Science* 105.8, pp. 6760–6772. ISSN: 0022-0302. DOI: 10.3168/jds.2022-21975.
- Giannuzzi, D., Toscano, A., Pegolo, S., Gallo, L., Tagliapietra, F., Mele, M., Minuti, A., Trevisi, E., Ajmone Marsan, P., Schiavon, S., Cecchinato, and A. (2022). "Associations between Milk Fatty Acid Profile and Body Condition Score, Ultrasound Hepatic Measurements and Blood Metabolites in Holstein Cows". In: *Animals* 12(9), p. 1202. DOI: 10.3390/ani12091202.
- Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng (2015). "Standardization of milk mid-infrared spectra from a European dairy network". In: *Journal of Dairy Science* 98.4, pp. 2150–2160. ISSN: 0022-0302. DOI: 10.3168/jds.2014-8764.
- Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Steevens, N. Gengler, and F. Dehareng (2021). "Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions". In: *Methods*. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2020.07.012.
- Hammami, H., V. enplas, J., Vanrobays, M.-L., Rekik, B., Bastin, C., Gengler, and N. (2015). "Genetic analysis of heat stress effects on yield traits, udder health, and fatty acids of Walloon Holstein cows". In: *J. Dairy Sci.* 98 4956–4968. DOI: 10.3168/jds.2014-9148.
- International Council on Animal Recording (ICAR) (2024). "Cattle Milk Recording: Standards and Guidelines". In: *Accessed September 19*.
- Jenkins and T.C. (1993). "Lipid Metabolism in the Rumen". In: *J. Dairy Sci.* 76 3851–3863. DOI: 10.3168/jds.S0022-0302(93)77727-9.
- Jensen, R. G. (2002). "The Composition of Bovine Milk Lipids: January 1995 to December 2000". In: *Journal of Dairy Science* 85.2, pp. 295–350. ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(02)74079-4.
- Jorjong, S., van Knegsel, A.T.M., Verwaeren, J., Lahoz, M.V., Bruckmaier, R.M., De Baets, B., Kemp, B., Fievez, and V. (2014). "Milk fatty acids as possible biomarkers to early diagnose elevated concentrations of blood plasma nonesterified fatty acids in dairy cows". In: *J. Dairy Sci.* 97 7054–7064. DOI: 10.3168/jds.2014-8039.
- Jubb, T. and N. R. Perkins (2015). "Veterinary Handbook for Cattle, Sheep and Goats". In.

- Kuhn, M. (2008). "Building Predictive Models in R Using the caret Package". In: *Journal of Statistical Software* 28, pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05.
- Lactanet (2022). *Guide to the Sustainability Index*. URL: <https://lactanet.ca/en/guide-to-the-sustainability-index/> (visited on 03/21/2024).
- Lê, S., Josse, J., Husson, and F. (2008). "FactoMineR: An R Package for Multivariate Analysis". In: *J. Stat. Softw.* 25. DOI: 10.18637/jss.v025.i01.
- Lindmark Månsson and H. (2008). "Fatty acids in bovine milk fat". In: *Food Nutr. Res.* 52 1821. DOI: 10.3402/fnr.v52i0.1821.
- Liu, Z., Ezernieks, V., Wang, J., Arachchillage, N.W., Garner, J.B., Wales, W.J., Cocks, B.G., Rochfort, and S. (2017). "Heat Stress in Dairy Cattle Alters Lipid Composition of Milk". In: *Sci. Rep.* 7 961. DOI: 10.1038/s41598-017-01120-9.
- Mann, S., Nydam, D.V., Lock, A.L., Overton, T.R., McArt, and J.A.A. (2016). "Short communication: Association of milk fatty acids with early lactation hyperketonemia and elevated concentration of nonesterified fatty acids". In: *J. Dairy Sci.* 99 5851–5857. DOI: 10.3168/jds.2016-10920.
- Mansbridge, R. and J. Blake (1997). "Nutritional factors affecting the fatty acid composition of bovine milk fat". In: *Livestock Production Science* 50, pp. 95–110. DOI: 10.1016/S0301-6226(97)00038-2.
- Murtagh, F. and P. Legendre (2014). "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" In: *Journal of Classification* 31.3, pp. 274–295. ISSN: 1432-1343. DOI: 10.1007/s00357-014-9161-z.
- Music, J., S. Charlebois, A. G. Marangoni, S. M. Ghazani, J. Burgess, A. Proulx, S. Somogyi, and Y. Patelli (2022). "Data deficits and transparency: What led to Canada's 'buttergate'". In: *Trends in Food Science & Technology* 123, pp. 334–342. ISSN: 0924-2244. DOI: 10.1016/j.tifs.2022.02.005.
- NFACC (2020). *Code of Practice for the Care and Handling of Dairy Cattle: Review of Scientific Research on Priority Issues*. National Farm Animal Care Council. URL: <https://www.nfacc.ca/pdfs/codes/scientists-committee-reports/Dairy%20Cattle%20SC%20Report%202020.pdf> (visited on 03/26/2025).
- Nickmilder, C., Tedde, A., Dufrasne, I., Lessire, F., Glesner, N., Tychon, B., Bindelle, J., Soyeurt, and H. (2023). "Creation of a Walloon Pasture Monitoring Platform Based on Machine Learning Models and Remote Sensing". In: *Remote Sensing* 15(7), p. 1890. DOI: 10.3390/rs15071890.
- Palmquist, D., A. D. Beaulieu, and D. Barbano (1993). "Feed and Animal Factors Influencing Milk Fat Composition". In: *Journal of Dairy Science* 76, pp. 1753–1771. DOI: 10.3168/jds.S0022-0302(93)77508-6.
- Québec, G. du (2025). *Production laitière*. Gouvernement du Québec. URL: <https://www.quebec.ca/agriculture-environnement-et-ressources-naturelles/agriculture/industrie-agricole-au-quebec/productions-agricoles/production-lait-vache> (visited on 03/27/2025).

- R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria.
- Riuzzi, G., Davis, H., Lanza, I., Butler, G., Contiero, B., Gottardo, F., Segato, and S. (2021). “Multivariate modelling of milk fatty acid profile to discriminate the forages in dairy cows’ ration”. In: *Sci. Rep.* 11 23201. doi: 10.1038/s41598-021-02600-9.
- Soyeurt, H., Dardenne, P., Gillon, A., Croquet, C., V, erick, S., Mayeres, P., Bertozzi, C., Gengler, and N. (2006). “Variation in fatty acid contents of milk and milk fat within and across breeds”. In: *J. Dairy Sci.* 89 4858–4865. doi: 10.3168/jds.S0022-0302(06)72534-6.
- Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne (2011a). “Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries”. In: *Journal of Dairy Science* 94.4, pp. 1657–1667. issn: 0022-0302. doi: 10.3168/jds.2010-3408.
- Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. Berry, M. Coffey, and P. Dardenne (2011b). “Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries”. In: *Journal of Dairy Science* 94.4, pp. 1657–1667. issn: 00220302. doi: 10.3168/jds.2010-3408.
- SPW (2025). *État de l’Agriculture Wallonne*. Belgium: SPW ARNE. URL: <https://etat-agriculture.wallonie.be/home.html> (visited on 03/27/2025).
- Vlaeminck, B., V. Fievez, A. Cabrita, A. Fonseca, and R. Dewhurst (2006a). “Factors affecting odd- and branched-chain fatty acids in milk: A review”. In: *Animal Feed Science and Technology* 131.3, pp. 389–417. issn: 03778401. doi: 10.1016/j.anifeedsci.2006.06.017.
- Vlaeminck, B., V. Fievez, D. Demeyer, and R. Dewhurst (2006b). “Effect of Forage:Concentrate Ratio on Fatty Acid Composition of Rumen Bacteria Isolated From Ruminal and Duodenal Digesta”. In: *Journal of Dairy Science* 89.7, pp. 2668–2678. issn: 00220302. doi: 10.3168/jds.S0022-0302(06)72343-8.
- Vlaeminck, B., V. Fievez, S. Tamminga, R. Dewhurst, A. Van Vuuren, D. De Brabander, and D. Demeyer (2006c). “Milk Odd- and Branched-Chain Fatty Acids in Relation to the Rumen Fermentation Pattern”. In: *Journal of Dairy Science* 89.10, pp. 3954–3964. issn: 00220302. doi: 10.3168/jds.S0022-0302(06)72437-7.
- Woolpert, M.E., Dann, H.M., Cotanch, K.W., Melilli, C., Chase, L.E., Grant, R.J., Barbano, and D.M. (2016). “Management, nutrition, and lactation performance are related to bulk tank milk de novo fatty acid concentration on northeastern US dairy farms”. In: *J. Dairy Sci.* 99 8486–8497. doi: 10.3168/jds.2016-10998.
- Woolpert, M.E., Dann, H.M., Cotanch, K.W., Melilli, C., Chase, L.E., Grant, R.J., Barbano, and D.M. (2017). “Management practices, physically effective fiber, and ether extract are related to bulk tank milk de novo fatty acid concentration on Holstein dairy farms”. In: *J. Dairy Sci.* 100 5097–5106. doi: 10.3168/jds.2016-12046.
- Zhang, L., C. Li, F. Dehareng, C. Grelet, F. Colinet, N. Gengler, Y. Brostaux, and H. Soyeurt (2021). “Appropriate data quality checks improve the reliability of

values predicted from milk mid-infrared spectra". In: *Animals* 11.2, pp. 1–17. DOI: 10.3390/ani11020533.

6

General discussion

6.1. Introduction

Throughout this thesis, each suggested methodological step uncovered new challenges, such as handling a large distance matrix or validating using on-farm data, highlighting the intrinsic complexity of using this methodology to develop a large-scale monitoring tool for farmers. Drawing on the experience acquired during this thesis, this discussion provides methodological considerations and practical guidelines for future studies, with particular emphasis on comments from the scientific community collected through conferences and the review processes. These considerations are summarized in a new concept called multilayered validation, which provides essential steps when developing a Precision Livestock Farming (PLF) monitoring tool, especially from unsupervised learning. However, this multilayered validation requires an operational data-driven framework to reach its full potential for milk Fourier Transform mid-infrared (FT-MIR) monitoring in dairy systems.

This introduction provides some context about milk FT-MIR data as phenotyping and monitoring technologies and anchors this thesis in the previous work developed. This section also discusses the origin of the unsupervised learning methodology and the challenges of this approach.

Section 6.2 suggests the multilayered validation framework to solve the challenges highlighted in the previous section. This section is divided into five key concepts: cluster stability, transferability, predictability, biological explainability, and practical utility.

Section 6.3 discusses the operational Information Technology (IT) framework required to ensure that such a monitoring tool is validated using on-farm data and is easy to deploy for practical use. The framework focuses on the essential properties necessary for feedback loop validation and reinforcement learning.

Finally, the last section draw conclusions from this thesis, discuss about the limitations and the perspectives for future improvements in the field, and offer a critical perspective on my work.

The objective of this thesis is to experiment with new methodologies for processing FT-MIR data to identify relevant patterns and develop a health-monitoring system for farmers. In this context, this thesis represents an exploratory journey at the interface of health monitoring, unsupervised learning, and big-data analytics. This study aims to advance both the methodological use and interpretative understanding of the potential of milk FT-MIR spectroscopy in a large-scale dairy context. Building on previous work on FT-MIR spectral standardization and predictive modeling by Grelet (2019) and Tedde (2023), this thesis extends the scope from supervised prediction to unsupervised exploration of complex data for dairy cow health monitoring. Its main contributions focus on the development and application of unsupervised clustering strategies for anomaly and pattern detection, as well as a critical examination of the methodological

and practical challenges raised by such approaches when applied to routine milk recording data.

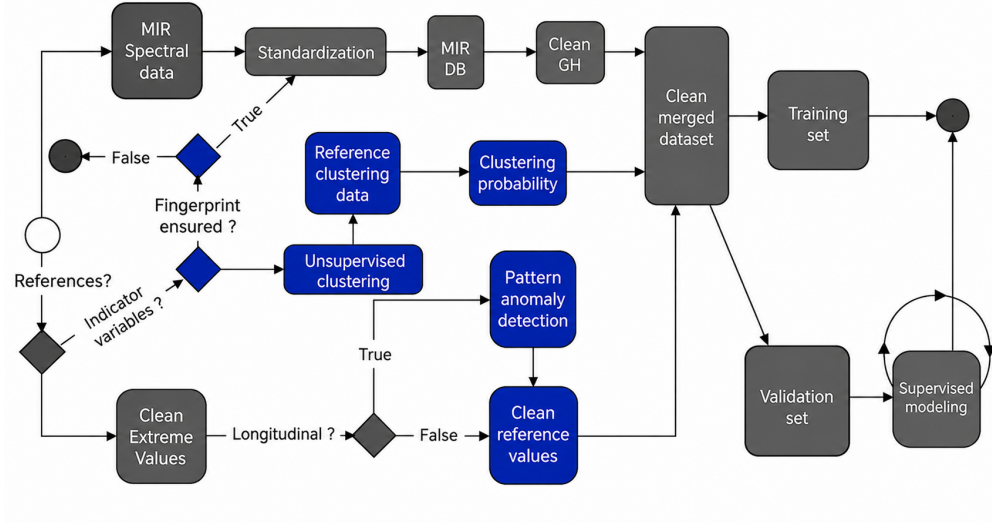


Figure 6.1: Standard supervised approach with pattern anomaly detection of longitudinal data, semi-complete and lacking reference. Figure modified from Tedde (2023).

6.1.1. The previous works on unsupervised learning in milk FT-MIR data

In his thesis, Tedde (2023) provided a comprehensive overview of the methodologies commonly applied to milk FT-MIR spectra (Figure 6.1), including pre-processing, specific modeling strategies, and data quality control procedures. He also highlighted the potential lack of reference-based information (e.g., grass diet information) and suggested moving toward unsupervised approaches in contexts where reference data are unavailable, identifying unsupervised clustering as a potential method to overcome this limitation (Tedde 2023). Pattern anomaly detection was also discussed as a means of cleaning extreme values in longitudinal data, and clustering was mentioned as a possible tool for detecting abnormal observations. Although both approaches fall under the broad concept of unsupervised learning and anomaly detection, they can be conceptually distinguished based on their objectives and the underlying definitions of anomalies. In the context of unsupervised clustering, a pattern anomaly may correspond to a group of observations sharing similar patterns, implying repetition and internal consistency; such anomalies may reflect genuine biological phenomena or systematic biases related to data acquisition or processing. In contrast, anomaly detection in the strict data-cleaning sense typically refers to isolated outliers (i.e., extreme or incoherent values) that are biologically implausible within a given context

and should be removed before analysis. This distinction is critical to avoid confounding the detection of rare but meaningful biological or management-related events with the identification of measurement errors that must be excluded during the data quality checks. In his discussion, he primarily referred to the latter category, illustrating cases in which the longitudinal data clearly violate biological plausibility, such as unrealistically large body weight changes occurring within a 24-h period (Tedde 2023).

Considering this distinction, the work presented in this thesis deliberately focuses on the other category of anomalies, namely, repeated and structured deviations that emerge at the pattern level rather than at the level of isolated observations. This focus is only allowed by the number of observations available, which permit the distinction between an anomaly and a repeated extreme event that should emerge as a rare pattern with sufficient data. While classical anomaly detection aims to remove incoherent values to ensure data consistency, the objective of this study was to explore whether such atypical patterns, when repeatedly observed across animals, herds, or time periods, could reflect meaningful biological or management-related phenomena. Such approach has been successfully applied in human medicine, where unsupervised clustering of clinical variables revealed reproducible and clinically relevant subtypes of type 2 diabetes, highlighting that atypical patient profiles may reflect distinct underlying pathophysiological mechanisms (Ahlqvist et al. 2018).

This idea was partially suggested by Tedde (2023) to create a reference when the target data are not available, but other researchers also made new “reference” using clustering on data sources available. For example, De Koster et al. (2019) estimated the metabolic status by clustering reference blood metabolites and then predicted the clusters using milk FT-MIR spectroscopy. Atashi et al. (2020) applied this strategy together with a genome-wide association study to identify single nucleotide polymorphisms related to early metabolic clusters. Other examples of unsupervised learning methods applied in precision livestock farming are presented in Chapter 2.

6.1.2. The scarcity of animal health references

In this study, animal health references were not available on a large scale to develop a monitoring tool. As discussed in the introduction, establishing a diagnosis is often costly, time-consuming, and rarely available in routine databases (Ifende et al. 2014). Moreover, ethical regulations governing animal experimentation, such as Directive 2010/63/EU (European Parliament 2010), are based on the principles of the 3Rs (Replacement, Reduction, and Refinement), which require minimizing both the number of animals used and the severity of procedures. Therefore, the use of invasive methods, such as blood sampling, may constrain the feasibility of large-scale diagnostic investigations. Instead, a common methodology is to conduct preliminary studies to highlight the association between a disease and a specific biomarker. Samples are then collected, and the biomarker is estimated using a gold-standard methodology.

An equation is then estimated using milk FT-MIR to predict the biomarker as a proxy for the disease (Figure 6.2). However, this method has several limitations.

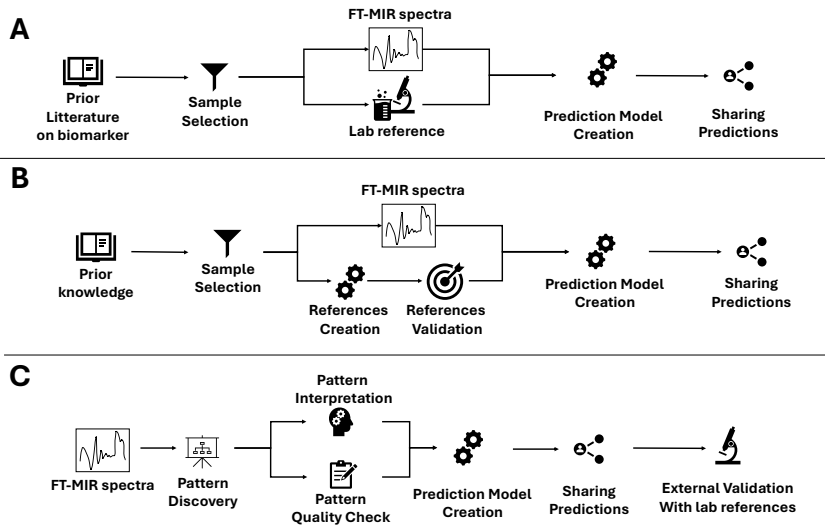


Figure 6.2: Shift in the development of new equations where references are not available easily. A is the traditional method. B is the method suggested by Tedde (2023) to create references and C is the new method developed in this thesis.

Grelet et al. (2021) highlighted that robust prediction models require sufficiently large and diverse calibration datasets, as limited or poorly representative datasets increase the risk of extrapolation and reduce predictive performance. In addition, the performance of such equations for health-related biomarkers may be limited when the trait of interest is only indirectly associated with milk composition, resulting in weak or non-specific spectral information in FT-MIR data (Grelet et al. 2024). Furthermore, relying on a single biomarker may be insufficient to capture the multidimensional nature of complex diseases.

Additionally, the definition of diagnostic thresholds for individual biomarkers remains debated, with substantial variability reported across studies depending on the population, methodology, and targeted disease (Ioannidis 2013; Ioannidis et al. 2017). For example, in mastitis detection, somatic cell count thresholds have historically varied and have not reached a universal consensus, although a threshold of 200,000 cells/mL is commonly used (Schukken et al. 2003). Benedet et al. (2019) reached similar conclusions for hyperketonemia, highlighting the challenges associated with diagnosing the condition based on blood β -hydroxybutyrate (BHB) concentrations. This highlights the context-dependent nature of the diagnosis and further complicates

its practical use (Ruegg 2017).

In practice, establishing a reliable diagnosis often requires veterinary intervention on the farm; however, subclinical diseases may still have significant impacts despite remaining undetected. Furthermore, clinical diagnosis inherently involves a degree of subjectivity, as the same disease can manifest differently in different animals (Espetvedt et al. 2013; Ruegg 2017). Consequently, records from farmers, veterinarians, or national databases do not always align (Mörk et al. 2009). One potential solution is to ask farmers to manually record these health events. This approach imposes a substantial workload and relies heavily on farmer motivation and consistency, which limits its feasibility and the quality of the data collected. In addition, this strategy introduces important sampling biases among farmers. Consequently, the collected cases are not representative of the full health spectrum but are conditional on farmers' perceptions, management styles and risk tolerance. Consequently, the notion of a single, unequivocal ground truth becomes problematic, making health reference data inherently difficult to collect, standardize, and use as reliable ground truths for model development.

6.1.3. A new approach to process FT-MIR data

For all the reasons mentioned in the previous section, milk FT-MIR spectra have a strong potential for large-scale screening of animal health status. While the collection of reliable health reference data and the definition of relevant biomarkers remain challenging, milk FT-MIR spectra are routinely available through milk recording organizations such as the European Milk Recording network (www.milkrecording.eu). Tedde (2023) suggested constructing new references from other existing data sources, as illustrated by Soyeurt et al. (2022), for instance, by leveraging temporal information (e.g., months) or by defining reference groups using thresholds for specific biomarkers (Figure 6.2). However, the approach proposed in this study differs in that it directly exploits the large number of predictions derived from milk FT-MIR spectra. The novelty of this study lies in the use of these routinely available predictions to detect health-related patterns without relying on predefined reference labels or potentially biased reference labels. This shift in objectives fundamentally changes the role of unsupervised methods: rather than serving as a data-cleaning step, clustering becomes an exploratory tool for hypothesis generation in large datasets (Mayer et al. 2025). The proposed workflow is thus structured around pattern detection, followed by interpretation and subsequent validation (Figure 6.2). This perspective raises several methodological questions and challenges regarding the application of unsupervised clustering to routine milk FT-MIR data.

6.1.4. Challenges arising from the approach

Throughout this thesis, the application of this new approach, unsupervised learning methodologies, emerged as a promising and innovative approach to enhance the use of routinely collected milk FT-MIR data for dairy farmers. However, it also raised critical methodological questions that guided further reflections and refinements. The various criticisms and remarks encountered during this thesis can be divided into five key questions.

1. Are the detected patterns stable across repeated unsupervised algorithm runs?
2. Are the identified patterns transferable across datasets, production contexts or populations?
3. Can the identified patterns be reliably predicted for new observations using the available information?
4. Are the identified patterns biologically valid and supported by external ground truths?
5. Are the identified patterns relevant to on-farm decision-making?

These questions reflect five central methodological dimensions: stability, transferability, predictability, biological explainability, and practical utility. These dimensions are often addressed separately in the literature without being integrated into a unified validation framework. Together, they form a coherent and structured approach for evaluating health monitoring tools. As mentioned in the introduction, a robust on-farm monitoring tool must fulfil several key criteria: validity and reliability under real-world conditions, continuous or timely information, and seamless integration into the farmer's workflow (Bell et al. 2018; Stygar et al. 2021; Silva et al. 2023). However, this is not sufficient to guarantee the quality of the monitoring tool.

In precision medicine, several studies have highlighted the importance of evaluating models across multiple, complementary dimensions (Ginsburg et al. 2018). From a methodological perspective, Ioannidis et al. (2009) emphasized the critical issue of reproducibility and underlined the need to assess the stability of data-driven findings. Ashley (2016) highlighted the importance of combining predictive performance with biological validity and clinical relevance when evaluating precision medicine approaches. More structured validation approaches have been proposed in biomedical research, notably through the framework developed by the Institute of Medicine et al. (2012) and the checklist introduced by McShane et al. (2013), which distinguishes successive steps from model development to clinical utility.

In the context of precision livestock farming, Rutten et al. (2013) proposed a framework describing the progression from sensor development to decision-making, emphasizing the importance of practical applicability. This framework was recently used

by Sani et al. (2026) in their review to classify the development level of sensors and show that most PLF technologies do not reach the robust decision-making level.

However, while these contributions collectively address key aspects of a valid monitoring tool, they remain fragmented within the literature. To our knowledge, no single unified framework in PLF explicitly integrates all dimensions of validation, including stability, transferability, predictive performance, biological plausibility, and practical utility.

The next section formalizes a unified multilayered validation framework by covering the methodological, biological, and applied validations that are typically treated separately, while also explaining how we tried to answer these concerns through this thesis. This multilayered validation represents this chapter's main analytical contribution. It also complements the existing framework of Norton et al. (2017), which focuses on improving the efficiency of PLF tool development. It also directly addresses the need identified by Gómez et al. (2021) in their systematic review, who highlighted that “the need for an established protocol for the validation procedures of PLF technologies can be noticed, as the measurements presented in the performance reports are very heterogeneous.”

6.2. Process of multilayered validation for PLF

In fields such as biomedical research, the development of diagnostic tools and biomarkers typically follows a stepwise process from discovery to clinical implementation. For instance, Pepe et al. (2001) proposed a five-phase approach for cancer biomarker development, later extended by Feng et al. (2020). In PLF, Rutten et al. (2013) suggested four levels of development, and overall, the Technology Readiness Level framework (*TRL Assessment Horizon Europe 2026*) describes the progressive stages of system development and maturity.

These frameworks are primarily designed to guide product development rather than to assess the robustness of data-driven results. While development phases define what should be achieved at each stage, they do not explicitly distinguish between the different dimensions that determine whether a result can be considered reliable. In contrast, the validation layers proposed in this study aim to explicitly structure these dimensions, independent of the stage of development.

Table 6.1 illustrates how existing frameworks from biomedical research, precision livestock farming, and engineering relate to the five proposed validation layers, highlighting that these dimensions are generally addressed implicitly and within sequential processes rather than as independent validation criteria. Each framework addresses specific aspects, such as reproducibility, generalization, predictive performance, biological plausibility, and practical utility. However, these dimensions are typically embedded within sequential development stages rather than explicitly defined as

independent validation criteria.

This observation supports the need for a unified multilayered validation framework, particularly in the context of unsupervised learning, where these aspects must be assessed for every pattern. The following sections examine each validation layer in detail and discuss how they were addressed in this thesis, with a particular focus on the metabolic disorder cluster identified in Chapter 3, as well as contributions from related studies.

Table 6.1: Comparison of different frameworks related to multilayered validation. Based on Pepe et al. (2001), Pepe et al. (2008), Rutten et al. (2013), *TRL Assessment Horizon Europe* (2026), Feng et al. (2020), and FDA-NIH Biomarker Working Group (2016)

Validation layer	Question addressed	FDA-NIH (2016)	BEST	Early Detection Research Network (Pepe et al. 2001)	(Rutten et al. 2013)	TRL
Stability	Are the patterns stable across repeated runs?	Analytical validity (reproducibility across sites)		Phase 2: assay reproducibility + standardization	Implicit (sensor validation)	TRL 4–5 (lab validation)
Transferability	Are the patterns valid across datasets/populations?	Context of use / population dependence		Phase 4: prospective validation	Level III: integration across systems	TRL 6–7 (relevant environment)
Predictability	Can patterns be predicted from new observations?	Clinical validity (predict outcomes)		Phase 3: ROC, sensitivity, specificity	Level II: detection algorithms	Implicit performance agreement
Biological explainability	Are the patterns biologically meaningful?	Biological plausibility		Phase 1: biological discovery	Level II: link to biological events	TRL 1–3 (concept)
Practical utility	Are patterns useful for real-world decisions?	Clinical utility		Phase 5: population impact	Level IV: decision support	TRL 8–9 (deployment)

TRL = Technology Readiness Levels

6.2.1. Cluster Stability

The first key methodological question concerns the reproducibility of the patterns identified through unsupervised learning, referred to as cluster stability: are the patterns stable across repeated algorithm runs? This question is central because many unsupervised methods introduce randomness through initialization, subsampling, or stochastic optimization, potentially leading to different solutions for the same dataset. For example, the divide-and-conquer approach presented in Chapter 3 introduces variability through random data partitions.

From a theoretical perspective, Lange et al. (2004) formalized stability as the expected dissimilarity between clustering solutions obtained from independent samples and interpreted instability as a form of classification risk. Instability may originate from the data (e.g., noise, unknown structure), clustering method (i.e., algorithm), or their interaction. Such variability is not inherently problematic, but it must be explicitly assessed, as separating randomness from biological variation is a prerequisite for a meaningful biological interpretation. In a context where external references are limited, stability across repetitions remains a key requirement for avoiding the overinterpretation of fragile patterns. Stable structures suggest reproducible and potentially meaningful patterns, whereas unstable structures should be interpreted with caution (Liu et al. 2022).

In Chapter 4, the stability of the metabolic disorder cluster was evaluated through repeated analyses, in which clustering procedures were performed five times for each of the four datasets. The results showed that the metabolic disorder cluster was consistently recovered between iterations and datasets, whereas other clusters were highly sensitive to the sampling strategy, highlighting the coexistence of robust and unstable patterns. This approach allowed distinguishing recurrent structures, which are more likely to reflect the intrinsic properties of the data, from artifacts driven by random variations.

Several procedures have been proposed to assess cluster stability, most of which rely on data perturbation strategies (Liu et al. 2022). Clustering was repeated on modified versions of the dataset, and the agreement between solutions was used as a measure of cluster stability. For instance, consensus clustering aggregates multiple clustering results obtained from resampled datasets and summarizes their agreement in a consensus matrix (Monti et al. 2003). Similarly, bootstrap-based approaches quantify the uncertainties in clustering solutions. In hierarchical clustering, Suzuki et al. (2006) proposed a multiscale bootstrap framework (R package *pvclust*), which assigns probability values to dendrogram branches based on their frequency across bootstrap replicates, thereby identifying stable and unstable branches in the dendrogram. The disadvantage of these techniques is their high computational cost for large datasets due to repeated iterations.

Complementing these approaches, Hennig (2007) proposed a cluster-wise assessment

of stability, emphasizing that stability should be evaluated at the level of individual clusters rather than the entire partition. Using similarity measures, such as the Jaccard coefficient, this approach makes it possible to distinguish consistently recovered clusters from unstable ones. This distinction is particularly relevant in exploratory analyses, where meaningful and artifactual structures may co-exist. Indeed, within the same clustering, some clusters can be highly stable, whereas others remain unstable.

Importantly, stability is required but is insufficient to establish cluster validity. While it provides essential information on the stability of clustering solutions, it must be combined with other validation layers to ensure a comprehensive assessment.

6.2.2. Cluster Transferability

The second key methodological question concerns the generalization ability of the patterns identified through unsupervised learning, referred to as transferability: are the identified patterns consistent across datasets, production contexts, or populations? By definition, unsupervised learning describes the variability present in a given dataset without relying on external references (Chander et al. 2021). Consequently, neither the methodology nor the resulting patterns guarantee that similar structures will emerge when applied to new data. In the context of unsupervised health monitoring, transferability is essential to ensure that detected patterns are useful beyond the specific dataset in which they are identified.

In this thesis, transferability was assessed by comparing clustering results across different contexts. In Chapter 4, the results showed that patterns identified in primiparous Holstein cows could be meaningfully extended to subsequent lactations, whereas the extension of bulk tank milk clusters was more limited in Chapter 5. In particular, the clustering structures observed in the Walloon dataset were only partially reflected in the Canadian dataset, highlighting that similar analytical approaches may lead to different structures depending on the dataset used. The Canadian dairy system is often more controlled, thus having lower variability in FT-MIR compared to the Walloon dataset. These results emphasize that transferability is context-dependent and cannot be assumed *a priori*.

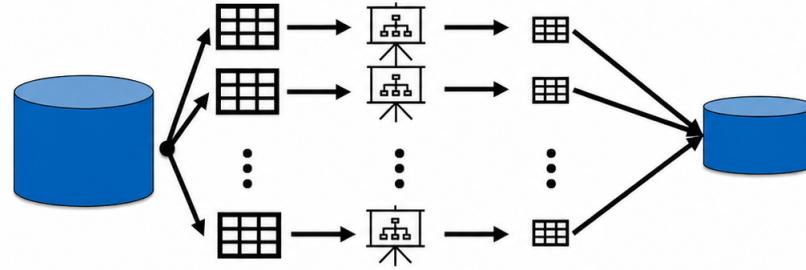
In the context of milk FT-MIR data, spectral standardization is a necessary first step toward transferability, ensuring that comparable spectra are obtained across instruments (Grelet et al. 2015; Grelet et al. 2017). However, even with standardized measurements, multiple biological and management-related factors strongly influence milk spectra, including breed, stage of lactation, feeding system, and health status (Arnould et al. 2009; De Marchi et al. 2014; Coppa et al. 2021; Grelet et al. 2021; Ramirez Mauricio et al. 2025). These factors modify the spectral signals, leading to differences in data distribution across populations. Consequently, clustering algorithms may primarily capture these dominant sources of variability rather than more subtle health-related patterns.

Several strategies have been proposed to improve generalization in unsupervised learning. A common approach consists of integrating a wide range of variability across breeds, regions, and production systems within a single large database, thereby increasing representativeness and allowing algorithms to capture more general structures (Grelet et al. 2021). However, this approach introduces two critical challenges: ensuring that the detected patterns are not dominated by major confounding factors and handling the computational burden associated with large datasets.

This first issue may be addressed by performing analyses within homogeneous subsets first or by applying data transformations, such as centering within groups, as considered by Chen et al. (2026) in the NWE Interreg HoliCow project (holi-cow.nweurope.eu), which has millions of spectra available. However, these approaches rely on prior assumptions regarding the sources of variability to control, which may bias the analysis.

The second issue is that large and heterogeneous datasets are difficult to process simultaneously because of memory and computational limitations. The divide-and-conquer strategy used in Chapter 3 ($N=740,454$) partially addresses this issue, but does not scale indefinitely. Alternative approaches such as incremental clustering have been proposed by Alexakis (2024) to handle 40 million observations (Figure 6.3). Incremental clustering methods sequentially update cluster structures as new data become available, reducing memory requirements but potentially introducing sensitivity to data ordering (Charikar et al. 2004; Narita et al. 2018; Narita et al. 2020). Similarly, grid-based clustering approaches partition the data space into discrete cells, enabling the efficient processing of large datasets at the cost of reduced resolution and sensitivity to grid definition (Tareq et al. 2022). Finally, neural network-based methods, such as autoencoders, allow batch-wise processing and dimensionality reduction, capturing the global data structure while improving scalability (Jiang et al. 2017). However, these approaches often reduce interpretability and require careful tuning, which may limit their applicability in exploratory biological studies.

Divide-and-conquer



Incremental Clustering

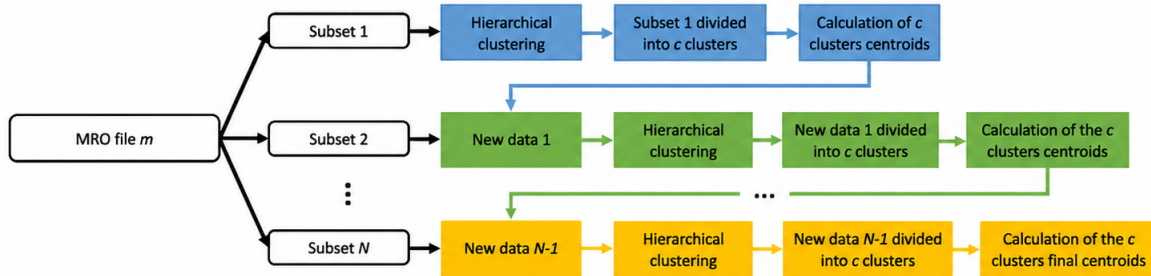


Figure 6.3: Methods to handle the volume of observations available (Franceschini et al. 2022a; Alexakis 2024).

From a validation perspective, the consistency of the clustering results across datasets can also be assessed using cluster similarity measures, such as the Jaccard coefficient, which quantifies the overlap between clusters obtained from different datasets (Fletcher et al. 2018). Although primarily used in stability analyses (Section 6.2.1), such measures can provide complementary insights into transferability by evaluating whether similar structures are recovered across independent datasets.

Overall, transferability depends on unsupervised learning methods but mostly on the datasets available. Although this is a required condition, it is not sufficient to guarantee the validity of the results.

6.2.3. Cluster Predictability

The third key methodological question concerns predictability, that is, whether the patterns identified through unsupervised learning can be reliably predicted for new observations using available information, such as milk FT-MIR spectra. In contrast to stability and transferability, which assess the robustness of patterns, predictability evaluates the ease of use. Grelet et al. (2021) comprehensively reviewed the factors influencing the quality of predictions.

In this thesis, the approach can be described as a ‘cluster-then-predict’ framework, in which patterns identified through unsupervised learning are subsequently modeled using supervised methods and FT-MIR spectra. This approach offers an easier deployment and probabilities of cluster membership which is the input required for an alert. A high predictive performance is often observed when cluster membership, which better reflects biological continuity and uncertainty, is predicted directly from milk FT-MIR predictions. This is expected because the patterns were originally identified using information derived from the spectra. However, the cluster membership can also be predicted directly from the milk FT-MIR spectra. In such cases, the predictive model learns to recognize combinations of spectral signals that define the clusters, making predictions easier to transfer to other stakeholders with different FT-MIR equations.

However, this predictability is not guaranteed when the key information used to define a pattern is not present in the milk spectrum. This limitation was illustrated by the udder-health cluster, for which somatic cell count (SCC, a variable not directly derived from FT-MIR spectra) was required to recover the cluster and achieve satisfactory prediction performance as shown in Table 3.7. The coefficients of determination in cross-validation decrease from 0.99 using FT-MIR and Somatic Cells Count (SCC) to 0.18 using only FT-MIR while a simple model using only the SCC reaches 0.93.

This observation highlights an important distinction between the existence and predictability of the patterns. The ability to predict this pattern for new observations depends on the overlap between the information used to define the pattern and the information available for prediction. Although milk FT-MIR spectra are highly

informative and capture a wide range of variations, they cannot be considered universal proxies for all biological processes. When a physiological or pathological state does not leave a detectable fingerprint in milk composition, its associated pattern is unlikely to be predicted using spectral data alone. For instance, Vanlierde et al. (2021) and McParland et al. (2024) used other variables such as milk yield, breed, parity or days in milk to increase the performances of their equation predicting methane emission.

Predictability should not be interpreted as evidence of biological relevance but rather as an indication that a detectable and exploitable signal exists within the chosen predictor space. Therefore, its success depends on the alignment between the information used to discover patterns (e.g., any feature available) and to predict them (e.g., FT-MIR spectra).

6.2.4. Cluster Biological Explainability

The absence of reference data that initially motivated the use of unsupervised learning also constitutes one of its limitations. The interpretation of detected patterns inevitably relies on prior knowledge, expert judgement, or references from previous experiments, as illustrated in Chapter 3. As a result, these interpretations should be considered hypotheses generated by unsupervised learning rather than confirmed biological truths. Without independent validation, there is a substantial risk of overinterpreting artifacts, confounding effects, or context-specific signals.

Beyond stability, transferability, and predictability, a fourth key question arises: are the identified patterns biologically valid and supported by external ground truths? Without this step, clustering outcomes remain descriptive constructs, and their relevance to animal health or management cannot be firmly established. The outputs of unsupervised learning require systematic external and biological validation before being used to support decision-making in real-world farming conditions.

In Chapter 3, the initial validation efforts focused on the metabolic disorder cluster. Some biological relevance was observed with significant correlations ($N=50$) between the probabilities of the metabolic cluster and the following: blood BHB (0.70), blood NEFA (0.67), blood glucose (0.6), milk glucose-6-phosphate (-0.32), milk BHB (0.53), isocitrate (0.50), and milk urea (0.32). However, the strength of the validation remained limited due to the relatively small number of blood and milk samples available. Moreover, the validation relied primarily on correlations between cluster membership and reference blood biomarkers, which, while informative, were insufficient to demonstrate biological causality or diagnostic performance.

To address these limitations, several complementary approaches have been developed outside the scope of this thesis. The first line of work, conducted by Cabaraux (2025), aimed to improve the biological interpretability of FT-MIR-derived clusters by integrating independent experimental datasets and previously collected diagnostic records within the NWE Interreg HoliCow consortium (holicow.nweurope.eu). This

study considered the linear and nonlinear relationships between the probabilities of each cluster and reference measures of validated biomarkers to confirm the existing link between the two. The metabolic disorder cluster from this thesis appeared to have a validated link to 14 biomarkers out of the 23 studied for ketosis (fatty acids C10, C12, C13, C14, C15, C18, C18:1cis9, blood BHB, blood glucose, blood non esterified fatty acids, energy balance, C17/C15 ratio, C18:1cis9/C15 ratio and fat-to-protein ratio) and 7 out of 22 for acidosis (fatty acids C10, C12, C13, C14, C17/C15 ratio, C18:1cis9/C15 ratio and fat-to-protein ratio). This approach provided additional evidence that the metabolic disorder cluster captures biologically coherent patterns associated with energy imbalances.

The second approach, developed by Hu (2025), investigated the genetic and genomic relevance of the detected patterns. This study evaluated the genetic correlations between cluster-derived scores and established indicators of negative energy balance and explored their association with genes involved in metabolic regulation. The results suggest that the metabolic disorder cluster, referred to as the logit-transformed Energy Deficit Score (LEDS), is strongly associated with a negative energy balance, exhibited the strongest causal association with NEB and demonstrated a high genetic ability to predict FT-MIR predicted NEB (65%) while having an heritability of 0.18, genetic and genomic correlation with FT-MIR predicted NEB of 0.85 and 0.81 and genetic and genomic correlation with NEFA of 0.79 and 0.81 (Hu et al. 2025c; Hu et al. 2025b). Hu et al. (2025a) also showed through a single-step genome-wide association that while FT-MIR predicted NEB and LEDS shared information (6 genes in common), they also have different information, 4 other genes for LEDS and 10 for FT-MIR predicted NEB (Figure 6.4). LEDS seems to collect supplementary neuronal signaling information from the energy homeostasis (Hu et al. 2025a).

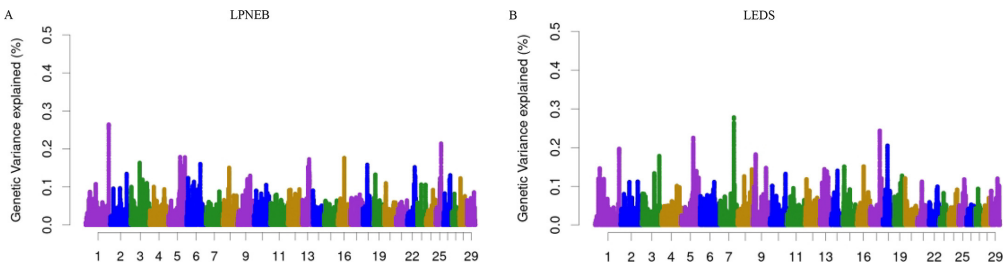


Figure 6.4: Additive genetic variance explained by windows of 50 adjacent SNPs across chromosomes for logit-transformed predicted negative energy balance (LPNEB, A) and logit-transformed energy deficiency score (LEDS, B) in early-lactation Holstein cows. Figure from Hu et al. (2025a)

Although these findings do not directly demonstrate on-farm monitoring usage, they provide strong evidence that the detected patterns reflect biologically meaningful mechanisms and may be relevant to animal breeding. Within the multilayer validation

framework proposed in this thesis, biological explainability represents the dimension that connects data-driven patterns to the physiological mechanisms. This step requires independent sources of evidence, such as experimental measurements or genetic information, to support cluster interpretation. Therefore, establishing biological validity is a necessary condition for validating the detected patterns; however, it also raises a subsequent question: whether these biologically meaningful patterns can be effectively translated into actionable tools under real farming conditions.

6.2.5. Cluster Practical Utility

In their review, Sani et al. (2026) emphasized the importance of integration and automation to achieve a farm-ready PLF monitoring tool. They argued that farmers are more interested in return on investment than the accuracy of the models. Ultimately, even when clusters are shown to be stable, transferable, predictable, and biologically relevant, their practical utility must be assessed under real-farming conditions. The final layer translates these data-driven patterns into tools that support herd monitoring and management. In practice in this thesis, these data patterns are clusters associated with predictions related to health issues. But a single cluster membership as a binary variable may not be enough. If the metabolic disorder pattern is considered, the cluster is not a diagnosis from a veterinarian but only a trend observed in the FT-MIR spectra that may be related to an issue on-farm. So, it is interesting to communicate the probability to belong in the cluster instead of the binary value to highlight the evolution in time and alert when this increase is significant. At the end, it is the alert that should be communicated to the farmer and evaluated on-farm. However, on-farm validation remains challenging.

The first difficulty lies in the collection of appropriate field data. Validating alerts requires temporally aligned FT-MIR measurements and reliable health observations by farmers or veterinarians. In practice, the relatively low frequency of milk recording, combined with the size of herds in Wallonia and the low occurrence of certain health events, resulted in a limited number of usable observations relative to the effort required. Cabaraux (2025) attempted to leverage farmer's diagnosis records from MyAweNet platform (awé, Ciney, Belgium), to validate cluster outputs. The detection rate for metabolic disorder was 66.7% but there were not enough diagnoses available to confirm this trend (N=6). The methodology seems relevant for mastitis detection, where more records were available (N=1,319), with a detection rate of 59.1% even though the quality of the diagnoses is unknown. Such datasets are often incomplete and heterogeneous, making it difficult to compute classical performance metrics.

To address these limitations, an alternative strategy consists of reversing the usual validation logic by using unsupervised patterns as triggers for the targeted feedback. Instead of relying on retrospective datasets, clusters interpreted as potentially problematic can be used to generate real-time alerts, prompting farmers to confirm whether

an observable issue exists. This approach allows validation efforts to focus on specific events rather than requiring exhaustive reference data collection. In practice, this corresponds to the implementation of a feedback loop, where data are analyzed in real time, alerts are sent to farmers, and their responses are used to assess the relevance of the detected patterns.

During the annual meeting of the European Federation for Animal Production in 2022, we suggested an alert system to exploit the different cluster probabilities simultaneously (Franceschini et al. 2022b). The idea was to flag as a red alert when the probability for a specific cluster was 10% higher than all the other clusters unrelated to health issues and an orange alert if at least one cluster had a probability within this 10% range. The classification is performed were as follow:

Algorithm 1 Example of alert system based on multiple probabilities

```

1: if Probability for metabolic cluster is the maximum then
2:   if Probability is 1.1 times higher than the others then
3:     It is a “problematic” record
4:   else
5:     It is an “at-risk” record
6:   end if
7: else
8:   if Probability for cluster 4 is only 1.1 times lower than the maximum then
9:     It is an “at-risk” record
10:  else
11:    It is a “healthy” record
12:  end if
13: end if

```

This method proved to be coherent with the thresholds for metabolic disorders in the literature (Macrae et al. 2019). Figure 6.5 shows the result by DIM. Although appealing, this feedback loop strategy introduces important methodological risks. Considering the confusion matrix of a binary classifier, there are 4 possibilities (Table 6.2). The feedback loop primarily enables the confirmation of positive alerts, whereas non-flagged observations are implicitly treated as negative. This may bias the evaluation by confounding true and false negatives, thereby limiting the interpretation of the performance metrics.

The initial motivation for this feedback loop emerged from the ScorWelCow project (grants D31-1390 and D65-1420), which aimed to assess animal welfare at the individual level. This requires identifying farms that would benefit the most from welfare assessment protocols in order to optimize the sampling effort and the variability in the dataset. Using the predicted probabilities for metabolic disorder and udder health clusters on the entire Walloon milk recording database, the average probabilities, the

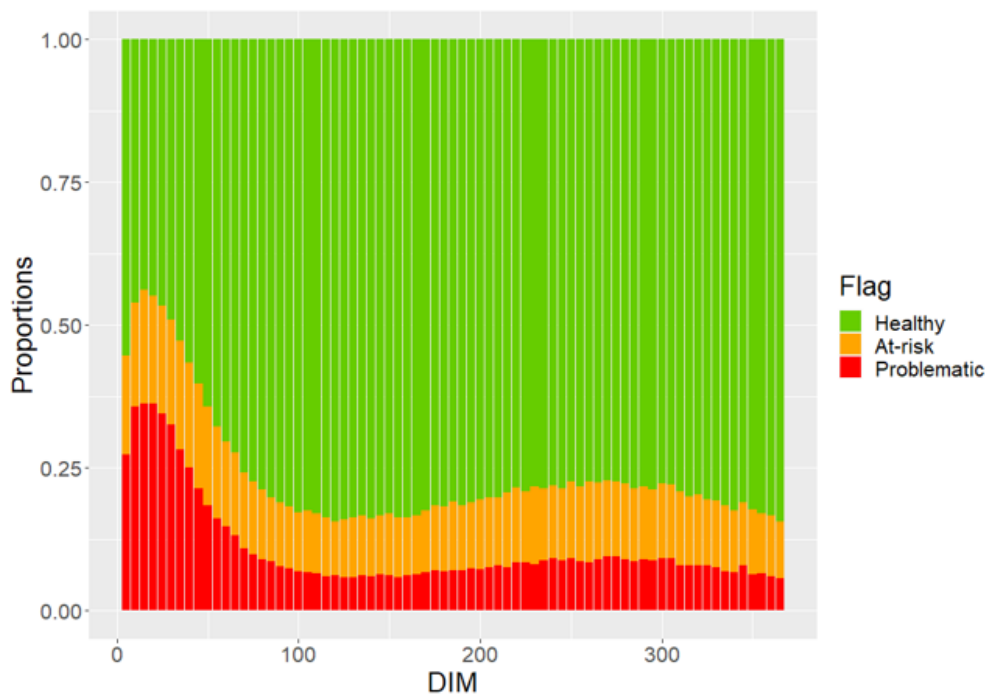


Figure 6.5: Metabolic disorder alerts by days in milk. Figure from Franceschini et al. (2022b)

Table 6.2: Confusion matrix for a binary classification

		True Class	
		Positive (P)	Negative (N)
Predicted Class	True (T)	True Positive (TP)	False Positive (FP)
	False (F)	False Negative (FN)	True Negative (TN)

variances inside the herd but also the variances between milk recording events were computed. Those values were used to establish a list of priority candidates with low average and high variances, which proved relevant in practice for the vet (Bernard Christiaens, oral communication). Building on this idea, two practical validation approaches were initiated: one in collaboration with Lactanet (Sainte-Anne-de-Bellevue, Canada) and another with the Bovi-Analytics Lab at Cornell University (Ithaca, New York State, US).

The first approach focused on the cluster outputs from Chapter 5 regarding milk payment data, which are acquired more frequently. Fastré (2022) developed herd clusters using fatty acids profiles from Walloon data then applied the cluster predictions on Canadian data. They then asked for feedback from the 4 participating farmers when observations belonged to specific clusters (Figure 5.6). The results were encouraging, with 10 alerts out of 12 being confirmed by farmers. However, the number of alerts was low because the process was not automated within the Lactanet framework (Sainte-Anne-de-Bellevue, Canada). While not on clustering results, Jacoby (2024) also researched how to develop relevant alerts using milk FT-MIR predictions and time-series.

The second approach, currently ongoing, involves the collaboration with the Bovi-Analytics Lab at Cornell University (Ithaca, New York State, US). The results presented in Table 4.2 were encouraging, milk samples with an alert were more likely related to management notes about health issue. However, it should not be considered as enough to assess practical utility because of the small numbers of cases observed. Only 3,875 management notes could be related to a FT-MIR spectrum within the previous 10 days and among those notes, only 396 were flagged as a positive alert. Among those management notes, only a few are relevant for health monitoring issue. While the frequency of acquisition of the milk spectra is partially responsible, it also proves that it is difficult to observe such events on farms as those are rare events.

Within the NWE Interreg HoliCow project (holicow.nweurope.eu), a huge spectral database (N>40millions) was collected by combining databases from every partner, and similar unsupervised learning methodologies were applied to discover patterns or clusters related to multiple topics: fertility, welfare, milk processing, environment and heat stress. Alerts were then developed using these clusters to develop a monitoring tool. During 2026 and 2027, they will collect alerts and feedback through multiples pilot farms in North West Europe countries. This initiative is designed to strengthen the link between data-driven alerts and field observations to validate the practical utility of the alerts developed. Results about alert efficiencies should be available soon.

In the longer term, such feedback loops could support the development of adaptive monitoring systems, in which model performance is progressively improved through farmer feedback, potentially leading to reinforcement learning strategies. As the event are rare, it would be more efficient to collect this information in every farm but

this requires the on-farm deployment of a first system to alert the farmer and collect feedback. However, this deployment requires infrastructures, such as WALLeSmart (<https://www.wallesmart.be>), that are only starting to be available in Wallonia.

Altogether, these elements highlight that practical utility cannot be reduced to a single validation step but requires continuous interaction between data analysis and field observations. Although feedback loop strategies offer a promising approach to bridge this gap, their implementation depends on the ability to collect, process, and interpret data in near real time. In practice, this requires a robust infrastructure to ensure efficient data flow between milk laboratories, farms, and researchers, as well as clear protocols for integrating farmer feedback into the validation process. Although promising approaches have been explored in this thesis, the full deployment of such systems remains challenging and was not achieved within its scope.

6.2.6. Synthesis of the multilayered validation

Taken together, the different validation dimensions explored in this section—stability, transferability, predictability, biological explainability, and practical utility—demonstrate that the assessment of unsupervised learning outputs cannot rely on a single criterion. Importantly, these dimensions are interdependent and require iterative evaluation as new data and evidence become available. This highlights that validation should be viewed as a dynamic and integrated process rather than a sequential set of independent steps.

In this context, the practical deployment of unsupervised learning tools depends on the ability to connect data acquisition, analysis, interpretation, and user interaction within a coherent system operating in near-real time. As discussed in the previous section, the implementation of such feedback-driven approaches remains challenging and currently limits their full deployment. Therefore, the next section introduces a structured framework designed to operationalize these principles and support the implementation of feedback-driven monitoring tools in real farming conditions.

6.3. Framework for a feedback-driven monitoring

The previous section proposed validation process for unsupervised learning outputs, highlighting that stability, transferability, predictability, biological explainability, and practical utility should be considered together rather than as independent validation steps. However, the practical utility of such outputs cannot be fully assessed if the results are limited to retrospective analyses. A monitoring tool only becomes useful when information can be generated, interpreted, communicated, and confronted with field observations within a coherent and operational system. Therefore, the question is not only whether a cluster is statistically robust or biologically plausible, but also whether the entire data flow can support its use in real-world farming conditions.

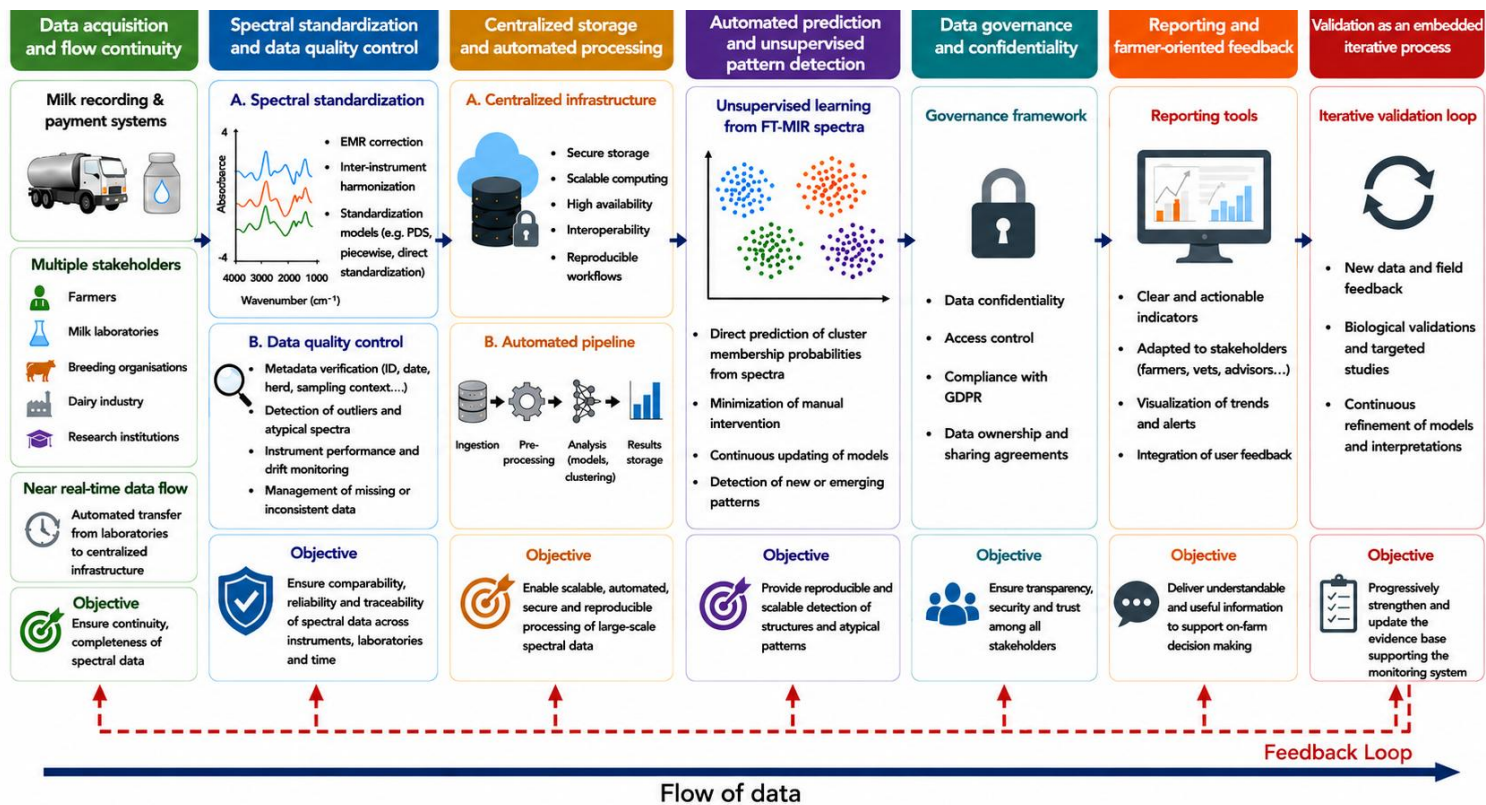


Figure 6.6: Proposed framework for unsupervised monitoring tool from FT-MIR data. Figure modified using ChatGPT.

6.3.1. Data acquisition, standardization, and automated processing

One of the major strengths of milk FT-MIR data in the context of precision livestock farming is that spectra are routinely generated through milk recording and milk payment systems. This existing infrastructure provides access to large quantities of spectral information without requiring additional invasive or expensive measurements from farmers. However, the availability of routine data does not automatically imply operational usefulness. Monitoring systems require continuous and reliable data flow because the value of an alert rapidly decreases if the information is generated too late to support observation or intervention. Consequently, operational deployment depends on the ability to continuously transfer spectra and associated metadata between laboratories, databases, and analytical pipelines while limiting delays and manual intervention as much as possible.

Once acquired, the spectra must remain comparable across instruments, laboratories, and time periods. This issue has been extensively discussed in previous milk FT-MIR studies, particularly by Grelet (2019), who demonstrated the importance of spectral harmonization procedures for enabling the transfer of models and data between instruments. Without adequate standardization, unsupervised learning algorithms may structure data according to analytical artifacts rather than biologically meaningful variability. This issue becomes even more critical as prediction models increasingly rely on larger spectral regions and indirect correlations, making them more sensitive to spectral deviations and analytical interferences (Dehareng 2023). Therefore, spectral standardization should not be viewed as a technical detail but as a prerequisite for the biological interpretation of the detected patterns.

Data quality control represents a complementary challenge. Conventional quality control procedures generally aim to eliminate atypical observations to improve consistency. However, the objective of the unsupervised approaches explored in this thesis was precisely to investigate whether repeated and coherent atypical patterns could reflect meaningful biological or management-related situations. Consequently, future monitoring systems should distinguish isolated analytical errors from structured rare observations, rather than automatically removing all outliers. This distinction is essential because some of the most informative biological signals may initially appear as atypical observations in smaller, routine datasets. The operationalization of such monitoring systems also requires centralized and reproducible analytical infrastructure capable of continuously processing large routine datasets.

In this context, cloud-based and automated analytical environments provide promising perspectives for ensuring reproducibility, scalability, and continuous updating of predictions, as discussed by Hostens et al. (2025). However, the objective here is not to discuss technical implementation details but rather to emphasize that unsupervised monitoring systems cannot function sustainably without infrastructure capable

of supporting automated data transfer, storage, and processing. Similar limitations were encountered during the collection of milk FT-MIR data from New York State dairy farms, where the absence of a standardized and automated spectral pipeline represented a major obstacle for field validation activities.

6.3.2. Data governance and confidentiality

The development of a monitoring framework based on animal and farm-level data requires a clear strategy for data governance and confidentiality. Milk spectral data, herd information, health records, and farmer feedback may contain sensitive information about farm management, productivity, and health status (FEFAC 2026). Therefore, these data cannot be freely exchanged or interpreted without clear rules defining access rights, responsibilities, and authorization procedures to access these data.

In Europe, such considerations are reinforced by the General Data Protection Regulation (European Parliament 2016) and the broader expectation that farmers should remain informed about how their data are used. Although animal and herd data are not always personal data in the strict sense, they may still be linked to identifiable farms and economic activities. Consequently, any operational framework should integrate data governance by design, rather than as an additional constraint after model development. In this context, Hostens et al. (2025) suggested the creation of embeddings produced on-farm and sent to the cloud environment to improve data confidentiality.

This includes secure storage, controlled access rights, data use traceability, and transparent communication with farmers. In this regard, initiatives such as the platform WALLESmart (<https://www.wallesmart.be>) are particularly relevant because they aim to structure data exchanges while allowing farmers to control which stakeholders can access specific categories of information. Other systems already exist such as PROFI-Lab by Lactanet (Sainte-Anne-de-Bellevue, Canada), RumiSense (Rumexperts, Faimies, Belgium) or MmmooOgle (Puurs-St-Amands, Belgium). Such systems are essential for the acceptance of data-driven monitoring tools in the sector. Without trust in data governance, even technically valid tools may fail to be adopted.

6.3.3. Automated pattern detection, reporting, and feedback integration

The analytical workflow developed in this thesis explored a monitoring strategy based on unsupervised pattern detection, rather than exclusively relying on predefined biomarkers or disease-specific models. Two applications were investigated: one using milk recording data at the individual cow level and the other using bulk tank milk data. In both cases, unsupervised learning methods were used to identify recurrent spectral profiles that were potentially associated with biological or management-related situations. Once these patterns were identified and interpreted, supervised

models were developed to predict the probability that new observations belonged to the detected clusters. This cluster-then-predict strategy provides a practical way to operationalize exploratory unsupervised learning approaches in routine monitoring systems. Figure 5.6 shows an example of the cluster probability evolution in time.

Such approaches are particularly attractive because they allow continuous monitoring without requiring the clustering procedures to be rerun each time new data become available. In addition, the direct prediction of cluster membership from spectra may simplify the analytical pipeline by bypassing intermediate biomarker predictions, potentially reducing the error propagation. More broadly, automation is not only important for efficiency but also for reproducibility, as manual interventions may introduce inconsistencies that complicate the interpretation and comparison of alerts across farms or periods.

The usefulness of a monitoring framework ultimately depends on its ability to generate understandable and actionable information for farmers, veterinarians, and advisors to use. This aspect is particularly important for unsupervised learning outputs because the clusters do not naturally correspond to predefined diagnoses. Their interpretation progressively emerges from a combination of spectral profiles, biological validation, and field observations. Consequently, reporting systems should avoid presenting clusters as definitive disease diagnoses and instead communicate them as alerts associated with previously observed biological or management-related conditions.

Importantly, reporting should not be considered a one-way communication. Farmers and veterinarians represent valuable sources of validation information because they are directly confronted with the real-world situations associated with the alerts. Their feedback may provide indications regarding health events, ration modifications, reduced intake, veterinary treatments, or management changes that occur after the alert. Although such field observations do not replace controlled biological validation, they provide realistic large-scale information capable of progressively refining the interpretation and usefulness of monitoring systems.

6.3.4. Validation as an embedded iterative process

The final objective of the framework is to provide useful alerts to farmers, veterinarians, and advisors and collect feedback to improve the system. This step is often underestimated in methodological research; however, it determines whether a monitoring tool can support decision-making (Stygar et al. 2021). Within the proposed framework, validation should not be viewed as a final step occurring after model development but rather as a continuous and embedded process.

This perspective is particularly important in the context of unsupervised learning because the interpretation of clusters is not known a priori and must emerge progressively from the accumulation of complementary evidence. Each new observation,

farmer feedback, biological measurement, or veterinary confirmation may contribute to refining our understanding of the detected patterns and their practical relevance.

Overall, the operationalization of feedback-driven unsupervised monitoring requires more than the identification of statistically coherent clusters. An integrated ecosystem is required, in which acquisition, standardization, automated processing, reporting, and validation continuously interact. Importantly, many of these components already exist within the dairy sector because FT-MIR spectra are routinely generated through milk recording and payment systems. Consequently, the main challenge is no longer the availability of data but the ability to transform these routinely collected data into reliable, interpretable, and actionable information capable of supporting adaptive monitoring systems over time.

6.4. Conclusion

This thesis explored the possibility of using unsupervised learning methods to extract new information from routinely collected milk FT-MIR data for dairy cow and herd monitoring purposes. The initial motivation was based on a simple observation: large quantities of spectra have already been collected through milk recording and milk payment systems, yet their use remains largely focused on predefined predictions of milk composition or specific biomarkers. This study investigated whether existing data could be used to detect repeated and meaningful patterns without relying on predefined health labels. This is a novelty from this approach as most biomarkers or alerts are based on smaller dataset but with diagnoses available. Other researchers started using similar unsupervised methodologies to better understand the FT-MIR and its potential as monitoring tool (Touil et al. 2026; Chen et al. 2026).

The results revealed that unsupervised learning can detect coherent structures in large FT-MIR-derived datasets. At the individual cow level, clusters related to metabolic disorders and mastitis-like profiles were identified from milk recording data. At the herd level, the use of fatty acid profiles predicted from bulk tank milk spectra allowed the identification of management-related patterns, including clusters associated with good management, potential risks, and metabolic disorder profiles. The interest of the monitoring using FT-MIR spectra is the absence of overcost for farmers already paying the milk recording. Indeed, the alert could be directly integrated in their dashboard and share with the advisor or the veterinarian. In a perfect infrastructure, the data should be available in near real-time after the spectral analysis and allow for better management. The NWE Interreg HoliCow project (holicow.nweurope.eu) is working on a solution database where the alert would be associated with a list of potential solutions for the farmer. However, the economic interest of this tool and the relevancy of the suggested solutions must still be evaluated.

6.4.1. Limitations

This thesis also demonstrated that the usefulness of unsupervised learning cannot be assessed using a single validation criterion. Some clusters may be stable but are difficult to interpret biologically. Others may be biologically plausible but less transferable across populations or production systems. Similarly, a cluster may be meaningful in retrospective analyses but difficult to use in practice if it cannot be predicted for new observations or is not reported rapidly enough to support decision-making. Therefore, this study proposes a multilayered validation process structured around stability, transferability, predictability, biological explainability, and practical utility.

Several validation efforts of similar unsupervised methodologies were performed during this thesis. The Chapter 3 validated the interpretation of metabolic and mastitis-related clusters using milk and blood reference biomarkers. The Chapter 4 evaluated cluster reproducibility and predictability across lactations, confirming the interest of the metabolic disorder cluster. Finally, Chapter 5 used the herd level information, available more frequently. The fatty acid profile clustering approach was applied to both Walloon and Canadian bulk tank milk data, and a comparison between the different contexts provided an important assessment of transferability. Field validation initiatives were also initiated through Lactanet (Sainte-Anne-de-Bellevue, Canada) and in collaboration with Cornell University (Ithaca, New York State, US) and the Bovi-Analytics team. These initiatives confirmed the relevance of feedback loop validation, while also showing that such validation requires a robust operational infrastructure. Several other validation efforts of similar unsupervised methodologies were performed around this thesis (appendix B). However, the most essential validation is the practical utility for farmer which was not completely assessed.

The frequency of FT-MIR acquisition is an important aspect of future implementation. Milk recording data provide individual cow-level information but are generally collected every four to six weeks, which limits their use for immediate health alerts. In contrast, milk payment data are collected much more frequently at the bulk tank level, often every one to four days, and therefore offer stronger potential for near-real-time herd monitoring. These two sources of information should be viewed as complementary rather than competing sources. Milk recording may support individual-level interpretation, whereas milk payment spectra may provide more frequent herd-level signals that are useful for management alerts. Therefore, the integration of these two types of information would improve the relevance of a milk FT-MIR-based monitoring tool.

Finally, milk FT-MIR spectra should not be expected to capture every aspect of animal health and farm management. Milk composition reflects animal metabolism and herd management; however, some disorders or welfare issues may be better detected through activity sensors, rumination data, body condition scoring, computer vision, veterinary records, or environmental measurements. Therefore, the future of data-

driven dairy monitoring will likely rely on the integration of multiple data sources rather than a single technology. In this context, the unsupervised methodologies developed in this thesis could be extended to multimodal datasets, where milk spectra provide one layer of information. However, the data integration step of combining different technologies should not be overlooked, as it remains challenging.

6.4.2. Perspectives

The results have multiple potential applications. At the farm level, cluster probabilities can be integrated into decision-support platforms such as WALLESmart (www.wallesmart.be), where farmers and advisors can access alerts while maintaining control over data sharing. At the sector level, these indicators can support advisory services by identifying farms or periods that require attention. In the long term, if the indicators are sufficiently validated, they may also contribute to breeding or selection objectives through the development of new phenotypes or selection indexes related to resilience, health, or management efficiency. However, such applications require careful validation to avoid overinterpreting exploratory patterns as definitive biological characteristics.

More research on the evolution of this probability in time would be highly relevant, especially to develop more refined alerts for the farmer. While a relevant cluster was identified and a supervised model was estimated to predict the probability to have a metabolic disorder. Their value for management remain unclear. A perspective could be to consider time series for multiple relevant clusters and their evolution to fine tune the alert.

An important perspective is the direct use of spectra rather than relying solely on predefined FT-MIR predictions. The approach developed in this thesis mainly used predicted biomarkers or milk components as input variables, which facilitated the biological interpretation. However, this strategy also requires prior knowledge and may miss spectral information not captured by existing equations. Direct spectra-based unsupervised learning could therefore allow the discovery of new patterns without restricting the analysis to already-known traits. This perspective is scientifically attractive but methodologically challenging because spectra contain a large number of highly correlated wavelengths and may be strongly affected by instrumental or preprocessing effects. Moreover, many distance-based unsupervised algorithms suffer from the curse of dimensionality. Therefore, future studies should investigate how dimensionality reduction, spectral pre-processing, and robust validation strategies can support direct spectral pattern discovery.

Another perspective is the development of on-farm infrared technologies, including near-infrared sensors and other rapid analytical systems. Such technologies could increase the frequency of measurements and bring the monitoring process closer to farms. However, their usefulness depends on their analytical reliability, calibration

transferability, cost, ease of use, and integration with existing data systems. As shown throughout this thesis, the value of a technology does not only depend on its ability to generate data but also on its capacity to standardize, interpret, validate, and communicate these data in a meaningful way.

6.4.3. Critical perspective on my thesis

The approach developed during this thesis may appear relatively conventional compared with more recent developments in unsupervised learning, cloud computing, and Artificial Intelligence (AI). However, bridging the gap between applying unsupervised learning to a large FT-MIR-derived database and generating an actionable alert for farmers represented a substantial challenge. The first limitation concerned the volume of data relative to the methodology used. Hierarchical clustering was initially considered as a proof of concept to assess whether unsupervised learning could provide relevant information from large-scale milk databases. Subsequent efforts therefore focused primarily on adapting the approach to large datasets and validating the identified clusters rather than on further improving the clustering methodology itself. Given the wide range of unsupervised learning methods currently available, alternative algorithms could also have been explored. In particular, anomaly or outlier detection approaches could be relevant for identifying animals or herds requiring additional attention. The longitudinal dimension of the results also deserved greater consideration, as temporal dynamics are fundamental to any monitoring application.

Developing an alert for farmers was not initially the primary objective of this thesis, but it progressively became an important perspective as the identified patterns appeared to have practical relevance. Earlier consideration of this operational objective could have strengthened the work by allowing the practical usefulness of the results to be evaluated under field conditions, although implementing such an approach through a living lab would itself have been complex. While new methodologies, algorithms, and validation strategies are scientifically relevant, one important question remains unanswered: does the proposed approach provide greater practical value than a simple threshold-based system for farmers or advisors? Ultimately, the purpose of such research is to support farm decision-making. Earlier implementation of a first operational alert could also have initiated a feedback loop in which farmer or advisor responses contributed to further validation and iterative improvement of the monitoring system.

In conclusion, this thesis shows that routinely collected milk FT-MIR data contain valuable information that remains underexploited for dairy monitoring. Unsupervised learning provides a promising method for exploring this information, identifying atypical but repeated patterns, and generating new hypotheses regarding cow health and herd management. Although the results are not sufficient to claim that such tools are ready for widespread on-farm deployment, they provide a methodological and

conceptual basis for future development. The main challenge is now to connect data acquisition, analytical methods, biological validation, and farmer feedback within operational systems capable of learning from routine use. If this challenge is addressed, milk FT-MIR-based unsupervised monitoring could contribute to the more proactive, scalable, and sustainable management of dairy herds.

Bibliography

- Ahlqvist, E., P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spéjel, H. Mulder, E. Lindholm, O. Melander, O. Hansson, U. Malmqvist, Å. Lernmark, K. Lahti, T. Forsén, T. Tuomi, A. H. Rosengren, and L. Groop (2018). “Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables”. In: *The Lancet Diabetes & Endocrinology* 6.5, pp. 361–369. ISSN: 2213-8587. DOI: 10.1016/S2213-8587(18)30051-2.
- Alexakis, I. (2024). “Utilisation de l’intelligence artificielle dans l’exploration des aptitudes à la transformation du lait de vache sur base de prédictions provenant de spectres moyen infrarouge”. fr. In: URL: <https://matheo.uliege.be/handle/2268.2/21376>.
- Arnould, V. M. -. and H. Soyeurt (2009). “Genetic variability of milk fatty acids”. In: *Journal of Applied Genetics* 50.1, pp. 29–39. ISSN: 2190-3883. DOI: 10.1007/BF03195649.
- Ashley, E. A. (2016). “Towards precision medicine”. en. In: *Nature Reviews Genetics* 17.9, pp. 507–522. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg.2016.86.
- Atashi, H., M. Salavati, J. De Koster, M. A. Crowe, G. Opsomer, and M. Hostens (2020). “Genome-wide association for metabolic clusters in early-lactation Holstein dairy cows”. In: *Journal of Dairy Science* 103.7, pp. 6392–6406. ISSN: 0022-0302. DOI: 10.3168/jds.2019-17369.
- Bell, M. J. and G. Tzimiropoulos (2018). “Novel Monitoring Systems to Obtain Dairy Cattle Phenotypes Associated With Sustainable Production”. In: *Frontiers in Sustainable Food Systems* 2. Publisher: Frontiers. ISSN: 2571-581X. DOI: 10.3389/fsufs.2018.00031.
- Benedet, A., C. L. Manuelian, A. Zidi, M. Penasa, and M. De Marchi (2019). “Invited review: β -hydroxybutyrate concentration in blood and milk and its associations with cow performance”. In: *animal* 13.8, pp. 1676–1689. ISSN: 1751-7311. DOI: 10.1017/S175173111900034X.
- Cabaraux, E. (2025). “Contribution au développement d’un outil de suivi en temps réel de la santé des vaches laitières à partir du traitement de l’information spectrale du lait.” fr. In: URL: <https://matheo.uliege.be/handle/2268.2/24271>.

- Chander, S. and P. Vijaya (2021). “Unsupervised learning methods for data clustering”. In: *Artificial Intelligence in Data Mining*. Elsevier, pp. 41–64. ISBN: 978-0-12-820601-0. DOI: 10.1016/B978-0-12-820601-0.00002-1.
- Charikar, M., C. Chekuri, T. Feder, and R. Motwani (2004). “Incremental Clustering and Dynamic Information Retrieval”. en. In: *SIAM Journal on Computing* 33.6, pp. 1417–1440. ISSN: 0097-5397, 1095-7111. DOI: 10.1137/S0097539702418498.
- Chen, Y., H. Atashi, S. Franceschini, C. Grelet, C. Nickmilder, P. Lema, K. Wijnrocx, H. Soyeurt, H. Consortium, and N. Gengler (2026). “A Data-Driven Global Health Classification for Dairy Cows Using Milk Mid-Infrared Spectra and Machine Learning”. In: *Journal of Dairy Science*. Submitted.
- Coppa, M., B. Martin, S. Hulin, J. Guillemin, J. V. Gauzentes, A. Pecou, and D. Andueza (2021). “Prediction of indicators of cow diet composition and authentication of feeding specifications of Protected Designation of Origin cheese using mid-infrared spectroscopy on milk”. In: *Journal of Dairy Science* 104.1, pp. 112–125. ISSN: 0022-0302. DOI: 10.3168/jds.2020-18468.
- De Koster, J., M. Salavati, C. Grelet, M. A. Crowe, E. Matthews, R. O’Flaherty, G. Opsomer, L. Foldager, and M. Hostens (2019). “Prediction of metabolic clusters in early-lactation dairy cows using models based on milk biomarkers”. In: *Journal of Dairy Science* 102.3, pp. 2631–2644. ISSN: 0022-0302. DOI: 10.3168/jds.2018-15533.
- De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa (2014). “Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits”. In: *Journal of Dairy Science* 97.3. Publisher: Elsevier, pp. 1171–1186. ISSN: 0022-0302. DOI: 10.3168/jds.2013-6799.
- Dehareng, F. (2023). “Spectrométrie moyen infrarouge dans le lait : contribution à l’étude des facteurs principaux visant à accroître la qualité des modèles de prédiction et à assurer leur contrôle”. HDL: <https://orbi.uliege.be/handle/2268/300713>. PhD Thesis. ULiège - Gembloux Agro-Bio Tech, Gembloux, Belgium.
- Espetvedt, M. N., O. Reksen, S. Rintakoski, and O. Østerås (2013). “Data quality in the Norwegian dairy herd recording system: Agreement between the national database and disease recording on farm”. In: *Journal of Dairy Science* 96.4, pp. 2271–2282. ISSN: 0022-0302. DOI: 10.3168/jds.2012-6143.
- European Parliament (2010). *Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the Protection of Animals Used for Scientific Purposes*. Official Journal of the European Union. URL: <https://eur-lex.europa.eu/eli/dir/2010/63/oj>.
- European Parliament (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*. Official Journal of the European Union. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

- Fastré, C. (2022). “Contribution au développement d’un outil d’aide à la décision pour l’élevage laitier à partir de la connaissance du profil en acides gras du lait prédit par la spectrométrie infrarouge”. Master’s thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/14927>.
- FDA-NIH Biomarker Working Group (2016). *BEST (Biomarkers, Endpoints, and other Tools) Resource*. eng. Silver Spring (MD): Food and Drug Administration (US). URL: <http://www.ncbi.nlm.nih.gov/books/NBK326791/> (visited on 05/25/2026).
- FEFAC (2026). *EU Code of Conduct on Agricultural Data Sharing by Contractual Agreement*. URL: <https://fefac.eu/resources/good-practices/eu-code-of-conduct-on-agricultural-data-sharing-by-contractual-agreement/> (visited on 05/08/2026).
- Feng, Z. and M. S. Pepe (2020). “Adding Rigor to Biomarker Evaluations—EDRN Experience”. en. In: *Cancer Epidemiology, Biomarkers & Prevention* 29.12, pp. 2575–2582. ISSN: 1055-9965, 1538-7755. DOI: 10.1158/1055-9965.EPI-20-0240.
- Fletcher, S. and M. Z. Islam (2018). “Comparing sets of patterns with the Jaccard index”. en. In: *Australasian Journal of Information Systems* 22. ISSN: 1326-2238. DOI: 10.3127/ajis.v22i0.1538.
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt (2022a). “Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?” In: *Journal of Dairy Science* 105.8, pp. 6760–6772. ISSN: 0022-0302. DOI: 10.3168/jds.2022-21975.
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt (2022b). “Development of a monitoring tool for metabolic imbalance issues based on Milk Recording data”. Anglais. In: URL: <https://orbi.uliege.be/handle/2268/300702> (visited on 05/07/2026).
- Ginsburg, G. S. and K. A. Phillips (2018). “Precision Medicine: From Science To Value”. In: *Health Affairs* 37.5, pp. 694–701. ISSN: 0278-2715. DOI: 10.1377/hlthaff.2017.1624.
- Gómez, Y., A. H. Stygar, I. J. M. M. Boumans, E. A. M. Bokkers, L. J. Pedersen, J. K. Niemi, M. Pastell, X. Manteca, and P. Llonch (2021). “A Systematic Review on Validated Precision Livestock Farming Technologies for Pig Production and Its Potential to Assess Animal Welfare”. English. In: *Frontiers in Veterinary Science* 8. ISSN: 2297-1769. DOI: 10.3389/fvets.2021.660565.
- Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng (2015). “Standardization of milk mid-infrared spectra from a European dairy network”. In: *Journal of Dairy Science* 98.4, pp. 2150–2160. ISSN: 0022-0302. DOI: 10.3168/jds.2014-8764.
- Grelet, C., T. Larsen, M. Crowe, D. Wathes, C. Ferris, K. Ingvarlsen, C. Marchitelli, F. Becker, A. Vanlierde, J. Leblois, U. Schuler, F. Auer, A. Köck, L. Dale, J. Sölkner, O. Christophe, J. Hummel, A. Mensching, J. Fernández Pierna, H. Soyeurt, M. Calmels,

- R. Reding, M. Gelé, Y. Chen, N. Gengler, and F. Dehareng (2024). “Prediction of key milk biomarkers in dairy cows through milk mid-infrared spectra and international collaborations”. In: *Journal of Dairy Science* 107.3, pp. 1669–1684. ISSN: 00220302. doi: 10.3168/jds.2023-23843.
- Grelet, C., J. A. F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng (2017). “Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models”. In: *Journal of Dairy Science* 100.10, pp. 7910–7921. ISSN: 0022-0302. doi: 10.3168/jds.2017-12720.
- Grelet, C. (2019). “Promoting international prediction models through standardization of milk mid-infrared spectra”. English. PhD thesis. ULiège. GxABT - Liège Université. Gembloux Agro-Bio Tech. URL: <https://orbi.uliege.be/handle/2268/236291> (visited on 04/16/2026).
- Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Steevens, N. Gengler, and F. Dehareng (2021). “Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions”. In: *Methods*. ISSN: 1046-2023. doi: 10.1016/j.ymeth.2020.07.012.
- Hennig, C. (2007). “Cluster-wise assessment of cluster stability”. In: *Computational Statistics & Data Analysis* 52.1, pp. 258–271. ISSN: 0167-9473. doi: 10.1016/j.csda.2006.11.025.
- Hostens, M., S. Franceschini, M. van Leerdam, H. Yang, S. Pokharel, E. Liu, P. Niu, H. Zhang, S. Noor, K. Hermans, M. Salamone, and S. Sharma (2025). “The future of big data and artificial intelligence on dairy farms: A proposed dairy data ecosystem”. In: *JDS Communications*. ISSN: 2666-9102. doi: 10.3168/jdsc.2025-0843.
- Hu, H. (2025). “Investigating the relationships between predicted negative energy balance and its biomarkers in early lactation dairy cows using genetic analysis”. PhD thesis. Université de Liège, Gembloux Agro-Bio Tech. URL: <https://hdl.handle.net/2268/338298>.
- Hu, H., H. Atashi, S. Franceschini, P. Lemal, C. Grelet, Y. Chen, K. Wijnrocx, H. Soyeurt, and N. Gengler (2025a). “Using single-step genome-wide association analyses to compare predicted negative energy balance and a novel energy deficiency score in early-lactation Holstein cows”. In: *JDS Communications* 6.6, pp. 792–796. ISSN: 2666-9102. doi: 10.3168/jdsc.2025-0778.
- Hu, H., S. Franceschini, P. Lemal, H. Atashi, C. Grelet, Y. Chen, K. Wijnrocx, S. Hélène, and N. Gengler (2025b). “Comparing the genetic architecture of energy balance predicted by mid-infrared spectrometry, a novel energy deficiency score, and several biomarkers”. English. In: *Journal of Dairy Science* 108.12, pp. 13499–13511. ISSN: 0022-0302. doi: 10.3168/jds.2025-26876.
- Hu, H., S. Franceschini, P. Lemal, C. Grelet, Y. Chen, H. Atashi, K. Wijnrocx, H. Soyeurt, and N. Gengler (2025c). “Exploring the relationship between predicted negative energy balance and its biomarkers of Holstein cows in first-parity early

- lactation". In: *Journal of Dairy Science* 108.5, pp. 5433–5447. ISSN: 0022-0302. DOI: 10.3168/jds.2024-25932.
- Ifende, V. I., M. Derks, G. A. Hooijer, and H. Hogeveen (2014). "Financial aspects of veterinary herd health management programmes". In: *The Veterinary Record* 175.9, p. 224. ISSN: 2042-7670. DOI: 10.1136/vr.102183.
- Institute of Medicine, Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, and Board on Health Sciences Policy (2012). *Evolution of Translational Omics: Lessons Learned and the Path Forward*. eng. Ed. by C. M. Micheel, S. J. Nass, and G. S. Omenn. Washington (DC): National Academies Press (US). ISBN: 978-0-309-22418-5. (Visited on 04/29/2026).
- Ioannidis, J. P. A. and P. M. M. Bossuyt (2017). "Waste, Leaks, and Failures in the Biomarker Pipeline". In: *Clinical Chemistry* 63.5, pp. 963–972. ISSN: 0009-9147. DOI: 10.1373/clinchem.2016.254649.
- Ioannidis, J. P. A., D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort (2009). "Repeatability of published microarray gene expression analyses". en. In: *Nature Genetics* 41.2, pp. 149–155. ISSN: 1546-1718. DOI: 10.1038/ng.295.
- Ioannidis, J. P. (2013). "Biomarker Failures". In: *Clinical Chemistry* 59.1, pp. 202–204. ISSN: 0009-9147. DOI: 10.1373/clinchem.2012.185801.
- Jacoby, P.-E. (2024). "Méthode d'agrégation d'alertes pour la gestion de troupeaux laitiers : combinaison des données longitudinales issues de l'analyse moyen infrarouge du lait et des connaissances basées sur la littérature". Unpublished master's thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/20427>.
- Jiang, Z., Y. Zheng, H. Tan, B. Tang, and H. Zhou (2017). *Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering*. DOI: 10.48550/arXiv.1611.05148. arXiv: 1611.05148[cs].
- Lange, T., V. Roth, M. L. Braun, and J. M. Buhmann (2004). "Stability-based validation of clustering solutions". eng. In: *Neural Computation* 16.6, pp. 1299–1323. ISSN: 0899-7667. DOI: 10.1162/089976604773717621.
- Liu, T., H. Yu, and R. H. Blair (2022). "Stability estimation for unsupervised clustering: A review". In: *WIREs Computational Statistics* 14.6, e1575. ISSN: 1939-5108, 1939-0068. DOI: 10.1002/wics.1575.
- Macrae, A. I., E. Burrough, J. Forrest, A. Corbishley, G. Russell, and D. J. Shaw (2019). "Prevalence of excessive negative energy balance in commercial United Kingdom dairy herds". In: *The Veterinary Journal* 248, pp. 51–57. ISSN: 1090-0233. DOI: 10.1016/j.tvjl.2019.04.001.
- Mayer, M., K. Lactaoen, G. E. Weissman, B. E. Himes, and R. A. Hubbard (2025). *Enhancing Phenotype Discovery in Electronic Health Records through Prior Knowledge-*

- Guided Unsupervised Learning*. DOI: 10.48550/arXiv.2511.02102. arXiv: 2511.02102[stat].
- McParland, S., M. Frizzarin, B. Lahart, M. Kennedy, L. Shalloo, M. Egan, K. Starsmore, and D. P. Berry (2024). “Predicting methane emissions of individual grazing dairy cows from spectral analyses of their milk samples”. In: *Journal of Dairy Science* 107.2, pp. 978–991. ISSN: 0022-0302. DOI: 10.3168/jds.2023-23577.
- McShane, L. M., M. M. Cavenagh, T. G. Lively, D. A. Eberhard, W. L. Bigbee, P. M. Williams, J. P. Mesirov, M.-Y. C. Polley, K. Y. Kim, J. V. Tricoli, J. M. G. Taylor, D. J. Shuman, R. M. Simon, J. H. Doroshow, and B. A. Conley (2013). “Criteria for the use of omics-based predictors in clinical trials”. en. In: *Nature* 502.7471, pp. 317–320. ISSN: 1476-4687. DOI: 10.1038/nature12564.
- Monti, S., P. Tamayo, J. Mesirov, and T. Golub (2003). “Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data”. en. In: *Machine Learning* 52.1, pp. 91–118. ISSN: 1573-0565. DOI: 10.1023/A:1023949509487.
- Mörk, M., A. Lindberg, S. Alenius, I. Vågsholm, and A. Egenvall (2009). “Comparison between dairy cow disease incidence in data registered by farmers and in data from a disease-recording system based on veterinary reporting”. In: *Preventive Veterinary Medicine* 88.4, pp. 298–307. ISSN: 0167-5877. DOI: 10.1016/j.prevetmed.2008.12.005.
- Narita, K., T. Hochin, Y. Hayashi, and H. Nomiya (2020). “Improvement of Incremental Hierarchical Clustering Algorithm by Re-insertion”. en. In: *Computational Science/Intelligence and Applied Informatics*. Ed. by R. Lee. Cham: Springer International Publishing, pp. 105–120. ISBN: 978-3-030-25225-0. DOI: 10.1007/978-3-030-25225-0_8.
- Narita, K., T. Hochin, and H. Nomiya (2018). “Incremental Clustering for Hierarchical Clustering”. In: *2018 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*. 2018 5th International Conference on Computational Science/ Intelligence and Applied Informatics (CSII), pp. 102–107. DOI: 10.1109/CSII.2018.00025.
- Norton, T. and D. Berckmans (2017). “Developing precision livestock farming tools for precision dairy farming”. In: *Animal Frontiers* 7.1, pp. 18–23. ISSN: 2160-6056. DOI: 10.2527/af.2017.0104.
- Pepe, M. S., Z. Feng, H. Janes, P. M. Bossuyt, and J. D. Potter (2008). “Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design”. In: *JNCI: Journal of the National Cancer Institute* 100.20, pp. 1432–1438. ISSN: 0027-8874. DOI: 10.1093/jnci/djn326.
- Pepe, M. S., R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget, and Y. Yasui (2001). “Phases of Biomarker Development for Early Detection of Cancer”. In: *JNCI: Journal of the National Cancer Institute* 93.14, pp. 1054–1061. ISSN: 0027-8874. DOI: 10.1093/jnci/93.14.1054.

- Ramirez Mauricio, M. A., M. Berton, N. Amalfitano, D. Giannuzzi, S. Pegolo, S. Raniolo, M. Nocetti, R. Negrini, M. Coppa, B. Martin, S. Schiavon, L. Gallo, E. Sturaro, and A. Cecchinato (2025). "Leveraging milk mid-infrared spectroscopy to authenticate animal welfare, farming practices, and dairy systems of Parmigiano Reggiano cheese". In: *Journal of Dairy Science* 108.3, pp. 2642–2657. ISSN: 00220302. DOI: 10.3168/jds.2024-25466.
- Ruegg, P. L. (2017). "A 100-Year Review: Mastitis detection, management, and prevention". eng. In: *Journal of Dairy Science* 100.12, pp. 10381–10397. ISSN: 1525-3198. DOI: 10.3168/jds.2017-13023.
- Rutten, C. J., A. G. J. Velthuis, W. Steeneveld, and H. Hogeveen (2013). "Invited review: Sensors to support health management on dairy farms". In: *Journal of Dairy Science* 96.4, pp. 1928–1952. ISSN: 0022-0302. DOI: 10.3168/jds.2012-6107.
- Sani, M. I., M. van der Voort, B. Tekinerdogan, and H. Hogeveen (2026). "Application of precision livestock farming: challenges and opportunities". In: *Animal Bioscience* 39.4, p. 250895. ISSN: 2765-0189. DOI: 10.5713/ab.250895.
- Schukken, Y. H., D. J. Wilson, F. Welcome, L. Garrison-Tikofsky, and R. N. Gonzalez (2003). "Monitoring udder health and milk quality using somatic cell counts". eng. In: *Veterinary Research* 34.5, pp. 579–596. ISSN: 0928-4249. DOI: 10.1051/vetres:2003028.
- Silva, F. G., C. Conceição, A. M. F. Pereira, J. L. Cerqueira, and S. R. Silva (2023). "Literature Review on Technological Applications to Monitor and Evaluate Calves' Health and Welfare". In: *Animals : an Open Access Journal from MDPI* 13.7, p. 1148. ISSN: 2076-2615. DOI: 10.3390/ani13071148.
- Soyeurt, H., C. Gerards, C. Nickmilder, J. Bindelle, S. Franceschini, F. Dehareng, D. Veselko, C. Bertozzi, N. Gengler, A. Marvuglia, A. Bayram, and A. Tedde (2022). "Prediction of Indirect Indicators of a Grass-Based Diet by Milk Fourier Transform Mid-Infrared Spectroscopy to Assess the Feeding Typologies of Dairy Farms". In: *Animals* 12.19, p. 2663. ISSN: 2076-2615. DOI: 10.3390/ani12192663.
- Stygar, A. H., Y. Gómez, G. V. Berteselli, E. Dalla Costa, E. Canali, J. K. Niemi, P. Llonch, and M. Pastell (2021). "A Systematic Review on Commercially Available and Validated Sensor Technologies for Welfare Assessment of Dairy Cattle". In: *Frontiers in Veterinary Science* 8. Publisher: Frontiers. ISSN: 2297-1769. DOI: 10.3389/fvets.2021.634338.
- Suzuki, R. and H. Shimodaira (2006). "Pvclust: an R package for assessing the uncertainty in hierarchical clustering". In: *Bioinformatics* 22.12, pp. 1540–1542. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl117.
- Tareq, M., E. A. Sundararajan, A. Harwood, and A. A. Bakar (2022). "A Systematic Review of Density Grid-Based Clustering for Data Streams". In: *IEEE Access* 10, pp. 579–596. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3134704.

- Tedde, A. (2023). “Predictions of dairy related phenotypes using milk mid infrared spectral data under different degree of information availability”. English. PhD Thesis. ULiège - Université de Liège, Gembloux Agro-Bio Tech, Gembloux, Belgium.
- Touil, T. and E. R. Paquet (2026). *Meta-clustering of milk mid-infrared spectra identifies dairy cow groups associated with negative energy balance in early lactation*. arXiv:2608.20653 [cs.LG]. DOI: 10.48550/arXiv.2608.20653.
- TRL Assessment Horizon Europe (2026). en. URL: <https://horizoneuropencpportal.eu/store/trl-assessment> (visited on 05/24/2026).
- Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt (2021). “Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra”. In: *Journal of the Science of Food and Agriculture* 101.8, pp. 3394–3403. ISSN: 1097-0010. DOI: 10.1002/jsfa.10969.

A

Descriptive statistics of Chapter 4

Table A.1: Descriptive Statistics of the datasets from Chapter 4.

	Lac1 (N=740,454)	Lac2 (N=577,601)	Lac3 (N=394,362)	LacA (N=1,712,417)
Acetone	0.06±0.03	0.06±0.03	0.06±0.03	0.06±0.03
BHB	158.07±42.24	160.58±43.43	160.55±44.67	159.49±43.23
Ca	1154.47±97.32	1168.27±101.91	1162.87±100.88	1161.06±99.9
CH4	306.31±58.38	319.12±61.2	318.67±62.38	313.48±60.6
CI	0.88±0.18	0.82±0.19	0.8±0.2	0.84±0.19
Citrate	8.64±1.36	8.48±1.4	8.4±1.43	8.53±1.39
DMI	19.85±2.67	21.05±2.82	21.52±2.91	20.64±2.87
DMI2	24.5±3.2	24.88±3.02	24.45±3.05	24.62±3.11
EB	0.33±3.54	-0.56±3.53	-1.37±3.65	-0.36±3.63
FPCM	23.52±5.53	27.13±7.43	28.58±8.14	25.91±7.19
K	1494.43±90.87	1467.92±96.47	1462.07±96.34	1478.04±95.17
Lactof	162.11±88.15	193.42±103.87	203.71±111.62	182.25±100.95
Lactose	4.84±0.17	4.75±0.18	4.71±0.19	4.78±0.18
Mg	100.68±7.73	99.69±8.04	98.47±7.89	99.84±7.92
Milk	23.72±5.91	27.07±8.05	28.58±8.79	25.97±7.68
MUFA	28.17±4.21	26.99±3.9	26.97±4.05	27.49±4.11
Na	321.9±34.52	344.83±37.98	356.84±40.46	337.68±39.85
NUE	11.34±2.14	11.49±2.17	11.69±2.21	11.47±2.17
Oleic acid	20.47±3.68	19.4±3.34	19.43±3.51	19.87±3.57
P	997.53±83.77	973.46±87.83	958.48±87.91	980.42±87.58
pFAT	39.66±5.99	40.52±6.35	40.54±6.36	40.15±6.21
pPROT	33.59±3.27	34.32±3.56	33.96±3.6	33.92±3.46
PUFA	3.7±0.57	3.65±0.58	3.64±0.57	3.67±0.58
RFI1	-3.46±2.18	-4.07±2.15	-4.62±2.23	-3.94±2.23
RFI2	-2.1±2.02	-2.4±2.01	-2.83±2.08	-2.37±2.05
SCS	2.41±1.52	2.72±1.67	3.09±1.74	2.68±1.64
SFA	70.39±4.66	71.55±4.4	71.56±4.53	71.05±4.58
Urea	249.32±79.66	243.92±80.82	241.6±80.27	245.72±80.26
Weight	620.16±65.2	621.02±65.19	617.82±65.15	619.91±65.19

CI = consumption index; DMI2 = predicted dry matter intake; EB = energy balance; FPCM = fat- and protein-corrected milk; Lactof = lactoferrin; MUFA = monounsaturated fatty acid; NUE = nitrogen use efficiency; pFAT = fat content; pPROT = protein content; PUFA = polyunsaturated fatty acid; RFI = residual feed intake (alternate calculation via RFI2); SCS = somatic cells score; SFA = saturated fatty acids.

Validation Efforts of Clustering Results

- Alexakis, I. (2024). "Utilisation de l'intelligence artificielle dans l'exploration des aptitudes à la transformation du lait de vache sur base de prédictions provenant de spectres moyen infrarouge". fr. In: URL: <https://matheo.uliege.be/handle/2268.2/21376>.
- Cabaraux, E. (2025). "Contribution au développement d'un outil de suivi en temps réel de la santé des vaches laitières à partir du traitement de l'information spectrale du lait." fr. In: URL: <https://matheo.uliege.be/handle/2268.2/24271>.
- Chen, Y., H. Atashi, S. Franceschini, C. Grelet, C. Nickmilder, P. Lema, K. Wijnrocx, H. Soyeurt, H. Consortium, and N. Gengler (2026). "A Data-Driven Global Health Classification for Dairy Cows Using Milk Mid-Infrared Spectra and Machine Learning". In: *Journal of Dairy Science*. Submitted.
- Czaplicki, S. (2022). "Contribution à l'appréciation du bien-être animal au travers de méthodes non supervisées basées sur les données collectées dans le cadre du contrôle laitier wallon". Unpublished master's thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/15949>.
- Fastré, C. (2022). "Contribution au développement d'un outil d'aide à la décision pour l'élevage laitier à partir de la connaissance du profil en acides gras du lait prédit par la spectrométrie infrarouge". Master's thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/14927>.
- Franceschini, S., S. Czaplicki, C. Nickmilder, J. Leblois, C. Grelet, N. Gengler, and H. Soyeurt (2026). "Assessing the reproducibility of FT-MIR-derived biomarker clusters across lactations and their direct prediction from milk spectra". (In press).
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt (2022). "Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?" In: *Journal of Dairy Science* 105.8, pp. 6760–6772. ISSN: 0022-0302. DOI: 10.3168/jds.2022-21975.

- Franceschini, S., C. Fastré, C. Nickmilder, D. E. Santschi, D. Warner, M. Bahadi, C. Bertozzi, D. Veselko, F. Dehareng, N. Gengler, and H. Soyeurt (2025). “Detection of Dairy Herd Management Issues Using Fatty Acid Profiles Predicted by Mid-Infrared Spectrometry”. In: *Animals* 15.11. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 1575. ISSN: 2076-2615. DOI: 10.3390/ani15111575.
- Hu, H. (2025). “Investigating the relationships between predicted negative energy balance and its biomarkers in early lactation dairy cows using genetic analysis”. PhD thesis. Université de Liège, Gembloux Agro-Bio Tech. URL: <https://hdl.handle.net/2268/338298>.
- Hu, H., H. Atashi, S. Franceschini, P. Lemal, C. Grelet, Y. Chen, K. Wijnrocx, H. Soyeurt, and N. Gengler (2025a). “Using single-step genome-wide association analyses to compare predicted negative energy balance and a novel energy deficiency score in early-lactation Holstein cows”. In: *JDS Communications* 6.6, pp. 792–796. ISSN: 2666-9102. DOI: 10.3168/jdsc.2025-0778.
- Hu, H., S. Franceschini, P. Lemal, H. Atashi, C. Grelet, Y. Chen, K. Wijnrocx, S. Hélène, and N. Gengler (2025b). “Comparing the genetic architecture of energy balance predicted by mid-infrared spectrometry, a novel energy deficiency score, and several biomarkers”. English. In: *Journal of Dairy Science* 108.12, pp. 13499–13511. ISSN: 0022-0302. DOI: 10.3168/jds.2025-26876.
- Hu, H., S. Franceschini, P. Lemal, C. Grelet, Y. Chen, H. Atashi, K. Wijnrocx, H. Soyeurt, and N. Gengler (2025c). “Exploring the relationship between predicted negative energy balance and its biomarkers of Holstein cows in first-parity early lactation”. In: *Journal of Dairy Science* 108.5, pp. 5433–5447. ISSN: 0022-0302. DOI: 10.3168/jds.2024-25932.
- Jacoby, P.-E. (2024). “Méthode d’agrégation d’alertes pour la gestion de troupeaux laitiers : combinaison des données longitudinales issues de l’analyse moyen infrarouge du lait et des connaissances basées sur la littérature”. Unpublished master’s thesis. Liège, Belgique: Université de Liège. URL: <https://matheo.uliege.be/handle/2268.2/20427>.
- Nickmilder, C., H. Soyeurt, I. Alexakis, T. Przybylski, M. Gertz, and S. Franceschini (2026). “Unsupervised machine learning techniques in precision dairy farming”. In: *Advances in Precision Dairy and Beef Farming Technologies*. Invited book chapter, in press. Burleigh Dodds Science Publishing.

