UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODES D'APPRENTISSAGE MACHINE POUR L'ESTIMATION DE L'EFFET DE LA GÉNÉTIQUE SUR LA PRODUCTION LAITIÈRE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE

PAR

AWA SAMAKÉ

JUILLET 2026

REMERCIEMENTS

Ma sincère gratitude va à mon directeur de recherche, le professeur Abdoulaye Baniré Diallo, pour sa supervision rigoureuse, son accompagnement scientifique et ses précieux conseils, qui ont guidé et enrichi cette thèse.

Je tiens à exprimer ma reconnaissance au Laboratoire de Bioinformatique de l'UQAM pour son accueil chaleureux, son encadrement scientifique et les ressources mises à ma disposition. Mes remerciements vont également à la Chaire de recherche-innovation en bien-être animal et intelligence artificielle (Well-e), dont le soutien financier a permis de mener à bien ce projet dans des conditions optimales.

Je souhaite exprimer ma profonde gratitude à mon époux, Zoumana, pour son amour, son soutien indéfectible, son accompagnement sans faille et sa patience qui m'ont permis de mener à bien cette thèse. Mes remerciements s'adressent également à ma fille Ramata, dont la compréhension et le soutien m'ont apporté la force et la motivation tout au long de ce parcours.

Je tiens à remercier mes parents, Siraman et Salimata, mes frères et sœurs, Siaka, Alou et Mariam, ainsi que mes beaux-parents, Moumine et Aramata, et toute ma belle-famille pour leur appui constant, leurs encouragements et leur présence rassurante à chaque étape de ce cheminement. Une pensée particulière à mon beau-père, Moumine, qui nous a quittés à peine un mois après le dépôt de cette thèse, pour ses encouragements qui m'ont aidée à traverser les périodes les plus éprouvantes.

Je souhaite remercier l'ensemble des collaborateurs qui ont contribué, de près ou de loin, à l'avancement de cette thèse.

Je remercie également l'ensemble des professeurs du programme de doctorat pour la qualité de leur enseignement et la richesse de leurs échanges.

À mes ami.e.s et collègues Golrokh, Hayda, Amanda et Amine, j'exprime ma reconnaissance pour l'environnement bienveillant qu'ils ont su créer et pour les échanges intellectuels et personnels qui ont nourri ma réflexion et adouci les périodes les plus exigeantes.

À tous les membres du laboratoire bioinformatique de l'UQAM et de la chaire de recherche et d'innovation

WELL-E, je leur exprime ma reconnaissance pour leur soutien et leur collaboration.

Je souhaite enfin souligner l'accompagnement offert par le comité de soutien aux parents étudiants (CSPE-UQAM) aux parents en études, l'association facultaire AESSUQAM ainsi que le soutien de la communauté uqamienne. Leur engagement a permis de créer un cadre favorable qui m'a aidée à trouver l'équilibre nécessaire pour mener à bien cette recherche.

À toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de cette thèse, j'exprime ma plus profonde gratitude.

Un grand merci à toutes et à tous. Aw ni tchié !

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

ACF Fonctions d'autocorrélation.

ACP Analyse des composantes principales.

ADN Acide désoxyribonucléique.

AIC Critère d'information d'Akaike.

ANN Réseaux de neurones artificiels / artificial neural network.

ARN Acides ribonucléiques.

ARNm ARN messager.

ARIMA Modèles autorégressifs intégrés à moyenne mobile (autoregressive integrated moving average, en anglais).

BIC Critère d'information bayésien.

BLUP Meilleure prédiction linéaire non biaisée (best linear unbiased prediction, en anglais).

CDN Réseau laitier canadien / Canadian dairy network.

CCS Compte de cellules somatiques (somatic cell count SCC, en anglais).

DL Apprentissage profond (deep learning, en anglais).

VEE Valeurs génétiques estimées (estimated breeding values EBV, en anglais).

VGE Valeurs génomiques estimées (genomic estimated breeding values GEBV, en anglais).

IDA Information dropout adaptée.

i.i.d Indépendantes et identiquement distribuées.

LSTM Réseaux à mémoire court et long terme (long short-term memory, en anglais).

LuPI Apprentissage avec les informations privilégiées (learning using privileged information, en anglais).

LuPTS LuPI pour séries chronologiques/LuPI for time series.

OLS Moindres carrés ordinaires (en anglais, ordinary least squares).

PACF Fonctions d'autocorrélation partielle.

SARIMA Modèles autorégressifs intégrés à moyenne mobile saisonnière (Seasonal autoregressive integrated moving average, en anglais).

SCS Score de cellules somatiques.

SNCF Société nationale des chemins de fer.

SNP Polymorphismes nucléotidiques simples.

REML Maximum de vraisemblance restreinte (en anglais, Restricted Maximum Likelihood).

RNN Réseaux neuronaux récurrents / Recurrent neural network.

UQAM Université du Québec à Montréal.

VARMAX Modèle vectoriel autorégressif à moyennes mobiles avec variables exogènes (vector autoregressive moving average with exogenous regressors, en anglais).

RÉSUMÉ

La génétique animale joue un rôle central dans l'amélioration des performances des troupeaux laitiers en permettant la sélection des reproducteurs les plus performants, ainsi que la prévision des maladies. Ces dernières années, l'accessibilité grandissante aux données génétiques, particulièrement celles génomiques, a un impact positif sur l'évaluation génétique au sein de la communauté de recherche académique et industrielle. L'évaluation génétique vise à estimer les valeurs génétiques (VEE, VGE) des animaux à partir de données de production, de généalogie ou de pédigrée, et de séquençage génomique, afin de prédire les performances futures et d'optimiser les décisions de sélection. Elle passe par l'estimation du patrimoine génétique ainsi que par son effet sur les générations futures. L'analyse de cet effet peut se traduire de différentes manières, selon le secteur ; par exemple, la prédisposition à certaines maladies en santé et la production dans l'industrie animalière. Cependant, la complexité des données (grande dimension des SNP, structures de pédigrée profondes, séries temporelles de production) et la variabilité environnementale constituent des défis majeurs pour les méthodes classiques telles que le BLUP. Nous répondons aux questions suivantes : *comment estimer l'effet de la génétique sur le profit dans la production laitière à partir des données de séquençages génomiques ? Comment estimer les futures productions laitières en intégrant les caractères génétiques à moyen/long terme ?*

Cette thèse propose plusieurs contributions méthodologiques reposant sur l'apprentissage automatique et l'apprentissage profond pour améliorer l'estimation et la prévision des performances laitières. Nous utilisons deux paradigmes d'apprentissage : la notion d'informations privilégiées et celle de variables exogènes. À ce jour et à notre connaissance, la notion d'informations privilégiées n'est pas encore appliquée dans le secteur animalier, plus particulièrement dans l'industrie laitière. *LSTMDropout*, un modèle récurrent qui intègre un mécanisme de *dropout hétéroscédastique* et l'information privilégiée, permet de prédire les VEE, VGE, ainsi que les composantes et la valeur économique du lait, avec une robustesse accrue face aux données bruitées. *AgriGen*, une architecture méthodologique intégrée qui permet la fusion des données génétiques, phénotypiques et environnementales en vue de l'évaluation génétique multi-traits et de la prise de décision. *DairyLuPTS* est une architecture optimisée pour la modélisation de séries temporelles de production laitière. *DairyBLUP* est un modèle linéaire mixte basé sur le pédigrée. Enfin, *Information dropout adaptée (IDA)*, une généralisation de l'information dropout, intègre des informations privilégiées pour régulariser les représentations latentes et améliorer les performances de prévision. Les résultats expérimentaux montrent que l'intégration d'informations privilégiées et de données multi-sources permet d'améliorer significativement la précision des prévisions, tout en réduisant le coût computationnel par rapport aux méthodes classiques. *LSTMDropout* et *DairyLuPTS* se distinguent en particulier par leurs performances en prévision multi-sorties et leur capacité à capturer les relations complexes entre les traits génétiques et les rendements laitiers.

Les perspectives de recherche incluent l'extension de *LSTMDropout* à d'autres types de séries temporelles, l'étude des interactions *génotype-environnement*, ainsi que l'analyse de l'effet des variants et des loci (positions d'un gène) sur la santé et la productivité. L'exploration de nouvelles stratégies d'augmentation de données génomiques et de modélisation mathématique des informations privilégiées constitue également un axe prometteur pour renforcer la robustesse des modèles et optimiser la sélection génétique.

INTRODUCTION

L'industrie laitière joue un rôle majeur dans l'économie mondiale, à la fois en termes d'alimentation de la population, d'emplois, de revenus agricoles et de commerce international (Food and Agriculture Organization of the United Nations, 2013; Food and Agriculture Organization of the United Nations *et al.*, 2018). C'est une part importante de l'agroalimentaire mondial. D'un autre côté, elle soutient les communautés rurales. Au Canada, l'industrie laitière représente une part majeure du secteur agricole, se classant au deuxième rang en termes de valeur économique du revenu agricole (Agriculture and Agri-Food Canada, 2025). Le secteur laitier canadien offre une large gamme de produits laitiers et de lait. Pour maintenir cette réputation internationale, les fermes laitières et les entreprises de transformation sont soumises à des normes de qualité rigoureuses, témoignant d'un engagement constant envers les meilleures pratiques en matière de bien-être animal (Agriculture and Agri-Food Canada, 2025; Dairy Farmers of Canada, 2025).

Depuis plusieurs décennies, de nombreux chercheurs s'intéressent au développement d'outils et de stratégies d'aide à la décision visant à optimiser aussi bien les performances des exploitations laitières que le bien-être animal (Slob *et al.*, 2021; Shine et Murphy, 2022). La *rentabilité ou le profit à vie* d'une vache se définit comme la *différence entre la valeur cumulée du lait produit et le coût associé à la gestion et à l'alimentation de la vache*, dans le secteur agricole. Elle est influencée par plusieurs facteurs comme la quantité et la qualité du lait, la santé et la longévité des vaches, et les coûts de gestion du troupeau (Chesnais, 2007; Frasco. *et al.*, 2020). La figure 0.1 montre les relations directes ou indirectes entre la production laitière et les facteurs comme la santé, la gestion, l'environnement et la *génétique* (Ahlborn-Breier et Hohenboken, 1991; Koerhuis et Thompson, 1997; Frasco. *et al.*, 2020; Špehar *et al.*, 2022; Cole *et al.*, 2023). Ainsi, la rentabilité des exploitations laitières est directement liée à l'optimisation de la productivité, à la maximisation des ressources, ainsi qu'à la limitation des dépenses. L'application des concepts et théories relatifs à l'amélioration génétique des caractères de production ainsi que des traits de conformation ou de mensurations corporelles chez les bovins laitiers occupe aujourd'hui une place centrale dans l'industrie laitière (Shine et Murphy, 2022). Au-delà de son rôle déterminant dans l'optimisation de la productivité animale, la *sélection génétique* constitue également un levier essentiel pour l'amélioration de la santé des animaux. Étant donné la difficulté d'observer directement l'effet précis des marqueurs génétiques sur un trait quantitatif, il est nécessaire de recourir à des approches d'estimation (Lynch et Walsh, 1998; Meuwissen *et al.*, 2001). Celles-ci reposent notamment sur le calcul des *valeurs génétiques*, définies comme des prédictions quantitatives de la contribution additive des gènes d'un individu à la performance attendue de

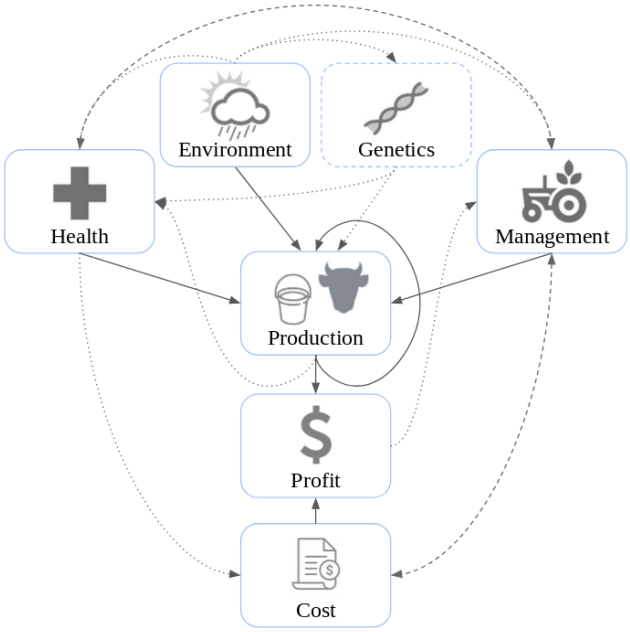


Figure 0.1 – Relations entre les principaux facteurs influençant la production laitière (Frasco. *et al.*, 2020). La production laitière dépend directement de la santé des animaux, de la gestion du troupeau et de l'environnement. La gestion du troupeau et les soins vétérinaires engendrent des coûts qui, combinés aux revenus issus de la production, déterminent la rentabilité de l'exploitation (profit). L'environnement influence également la santé, la gestion du troupeau et les performances génétiques. La génétique agit directement sur les caractéristiques de l'animal et contribue ainsi à sa santé ainsi qu'à son potentiel de production.

sa descendance. Ces estimations fournissent ainsi une approximation de l'effet génétique réel, autrement inobservable. L'ensemble de ces procédures constitue ce que l'on désigne sous le terme d'*évaluation génétique* (Lynch et Walsh, 1998). L'évaluation génétique vise principalement à fournir des outils permettant le classement des reproducteurs en fonction de leur valeur génétique estimée, facilitant ainsi la prise de décision au sein des programmes d'amélioration génétique. Par ailleurs, ces estimations peuvent être exploitées pour orienter les stratégies d'accouplement, que ce soit pour rechercher des combinaisons génétiques complémentaires ou, au contraire, pour éviter des croisements entre individus présentant une divergence génétique marquée. Elles contribuent ainsi à une gestion raisonnée de la reproduction, conciliant progrès et diversité génétique (Lynch et Walsh, 1998; Meuwissen *et al.*, 2001).

À chaque génération, selon un critère défini, le progrès génétique correspond à la sélection des individus présentant le plus grand intérêt pour engendrer la génération suivante (Ducos, 1995). Sa réalisation repose sur l'estimation de la part héréditaire des caractères d'intérêt et de l'impact sur les générations futures. Plusieurs études examinent la relation entre la rentabilité et le mérite génétique (VEE ou VGE). Par exemple,

Weersink *et al.* évaluent cette relation à partir d'une moyenne pondérée du mérite génétique des taureaux utilisés dans les programmes d'insémination artificielle en élevage (Weersink *et al.*, 1991). Lund *et al.* estiment les héritabilités ainsi que les corrélations génétiques et phénotypiques associées à divers caractères linéaires, au nombre de cellules somatiques (CCS), ainsi qu'à diverses maladies, dont la mammite, une inflammation de la glande mammaire caractérisée par une réponse immunitaire à une infection et entraînant des changements dans la composition du lait. Les cellules somatiques sont des cellules présentes dans le lait dont la concentration constitue un indicateur de la santé mammaire. Leur analyse repose sur les données de première lactation issues du programme danois d'évaluation des jeunes taureaux (Lund *et al.*, 1994). Deux jeux de données ont été analysés au moyen d'une procédure REML, c'est-à-dire une estimation du maximum de vraisemblance restreinte mise en œuvre dans un cadre linéaire multicaractères, appliquée à un modèle animal (Meyer et Smith, 1996; Ducrocq, 1990; Da et Grossman, 1991). Ce dernier, également appelé modèle individuel, repose sur une régression linéaire multiple prenant en compte simultanément la valeur génétique de l'animal et celle de ses apparentés (Fournet *et al.*, 1997; Leclère *et al.*, 2014; Ducos et Bidanel, 1993). Les résultats ont montré que les corrélations phénotypiques étaient globalement faibles ; toutefois, une sélection visant à améliorer les caractères du système mammaire permettrait de limiter l'augmentation du CCS et de réduire l'incidence de la mammite. La forte corrélation (0,96) observée entre le CCS et la mammite confirme d'ailleurs que le CCS pourrait constituer un indicateur pertinent de la résistance à la mammite. Par ailleurs, Pryce *et al.* développent le Dairy Information System (DAISY), destiné à l'enregistrement systématique des caractères de santé et de fertilité, à la fois pour la recherche et pour l'appui à la gestion des troupeaux (Pryce *et al.*, 1998). Les données, collectées sur trente-trois troupeaux pendant six ans, ont permis d'estimer les paramètres génétiques des caractères de santé (mammite, boiterie, score cellulaire somatique, SCS), de fertilité (intervalle vêlage–vêlage, jours jusqu'au premier service, conception au premier service) et de production (rendement laitier à 305 jours, matières grasses et protéines). Les composantes de (co)variance sont estimées par REML linéaire multicaractères avec un modèle animal. Les auteurs évaluent également le potentiel des caractères linéaires de conformation comme prédicteurs de caractères de santé, en régressant ces derniers sur les valeurs génétiques estimées des taureaux pour la conformation. Les résultats suggèrent que certaines mensurations corporelles, également appelées traits de conformation, pourraient constituer l'avenir des critères de sélection pertinents pour améliorer la santé des animaux.

Lawson *et al.* mettent en évidence l'effet négatif d'une sélection génétique unilatérale sur le rendement du lait (la quantité moyenne de lait produite par animal et par jour), notamment par l'augmentation de

l'incidence des maladies de production (Lawson *et al.*, 2004). En ce qui concerne Rivest *et al.* , ils estiment le potentiel économique de nouveaux caractères génétiques dans le secteur porcin québécois et ont développé des outils pour calculer les valeurs financières des indices paternels et maternels (Rivest *et al.*, 2008). Leur approche repose sur une régression multiple du revenu net et des caractères économiquement importants, incluant la race et la génétique. Le coefficient de régression associé à une variable donnée représente ainsi l'augmentation attendue du revenu pour une unité d'augmentation de cette variable, les autres variables étant fixées. De leur côté, Roibas *et al.* analysent l'impact du progrès génétique sur la rentabilité des exploitations laitières (Roibas et Alvarez, 2010).

Par ailleurs, plusieurs travaux se sont concentrés sur l'estimation de l'héritabilité des caractères (Lembeye *et al.*, 2016; Lee *et al.*, 2020; Xue *et al.*, 2023). L'étude de Xue *et al.* montre que les caractères de conformation des vaches laitières sont étroitement liés à leur production et à leur reproduction, deux critères majeurs en élevage laitier. Les auteurs estiment l'héritabilité de vingt-trois caractères de conformation et de cinq caractères de production laitière, ainsi que leurs corrélations génétiques, chez des vaches Holstein chinoises. Les analyses, réalisées à l'aide du logiciel DMU et de modèles animaux monocaractères et multicaractères, révèlent des corrélations génétiques positives entre la majorité des caractères de conformation et les caractères de production laitière (Xue *et al.*, 2023; Madsen *et al.*, 2014).

Dans le domaine de la santé, de nombreuses études ont été menées pour analyser les liens entre marqueurs génétiques et maladies rares et/ou chroniques. Ces travaux, appelés études d'association à l'échelle du génome (*Genome-Wide Association Studies*, GWAS), visent à identifier les régions du génome statistiquement associées à un caractère ou à une pathologie spécifique. Les GWAS reposent sur le test simultané de centaines de milliers de variantes génétiques à partir de données issues d'un grand nombre d'individus. Cette approche a permis de mettre en évidence un large éventail d'associations robustes pour divers phénotypes et maladies, et le nombre de variants identifiés devrait continuer à croître à mesure que la taille des échantillons augmente. Les résultats obtenus présentent de nombreuses applications, notamment la compréhension de la biologie sous-jacente d'un phénotype, l'estimation de son héritabilité, le calcul de corrélations génétiques, la prédiction de risques cliniques, l'orientation des programmes de développement de médicaments, ainsi que l'inférence de relations causales potentielles entre facteurs de risque et indicateurs de santé (Uffelmann *et al.*, 2021). Ce schéma (Figure 0.2) illustre les principales étapes d'une étude GWAS, depuis la collecte des données génétiques jusqu'aux analyses post-GWAS, soulignant la rigueur méthodo-

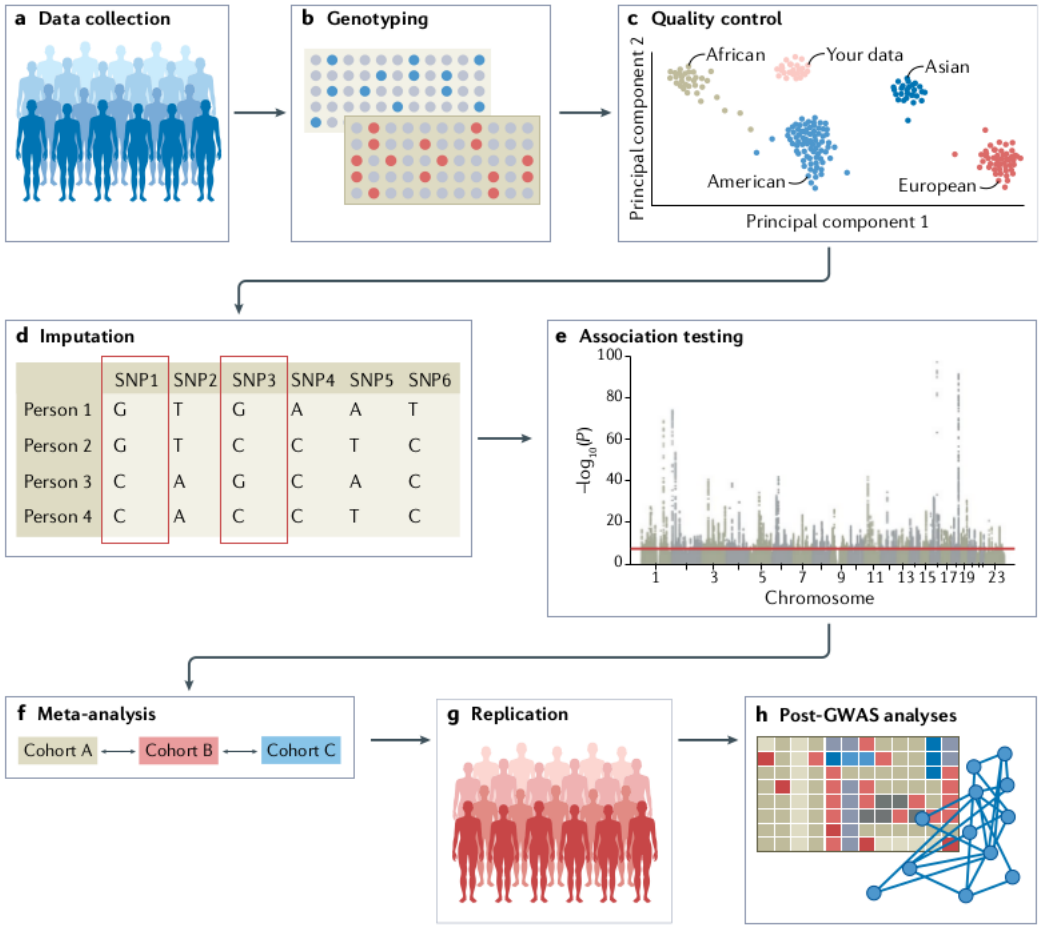


Figure 0.2 – Schéma général d'une étude d'association à l'échelle du génome (GWAS). Huit étapes clés : la collecte de données, le génotypage, le contrôle de qualité, l'imputation, les analyses d'association, la **méta-analyse**, la réplique (reproduction) et les analyses post-GWAS (Uffelmann *et al.*, 2021).

logique nécessaire à l'identification de loci associés à des caractères complexes.

Dans le secteur laitier, l'*évaluation génétique* conventionnelle, basée sur les performances phénotypiques et les liens de parenté, contribue de manière significative à l'amélioration des populations bovines laitières. Néanmoins, cette méthode demeure limitée par la nécessité de collecter des données complètes sur un grand nombre d'animaux et par le délai nécessaire pour évaluer la descendance avant d'estimer avec précision la valeur génétique des reproducteurs. Les progrès du génotypage à haut débit et la disponibilité croissante de marqueurs moléculaires (comme les séquençages génomiques) répartis sur l'ensemble du génome ont permis l'émergence de la *prédiction génomique*. En intégrant directement l'information génétique dans les modèles d'estimation, cette approche améliore la précision des évaluations dès les premiers

stades de la vie, réduit l'intervalle de génération et accélère le gain génétique annuel. On distingue généralement trois approches de prédiction génomique :

- Basée sur le pedigree (*pedigree-based*), utilisée uniquement pour les individus non génotypés ;
- Basée sur les données génomiques (*genomic-based*), appliquée exclusivement aux individus génotypés ;
- Combinée (*single-step genomic*), intégrant simultanément les informations des individus génotypés et non génotypés.

Malgré ses nombreux avantages, l'évaluation génétique reste confrontée à plusieurs défis. Le Tableau 0.1 en présente quelques-uns ainsi que leurs impacts. Parmi ceux-ci figurent notamment la disponibilité de données de qualité, le coût élevé du génotypage ainsi que la complexité biologique de certains caractères.

Dans le cadre de cette étude, nous nous intéressons à l'analyse de l'effet de la génétique sur des traits ou caractères de production. Plus précisément, nous cherchons à déterminer l'influence exercée par différents caractères génétiques, tels que les traits de conformation et les données issues du séquençage génomique, notamment les polymorphismes nucléotidiques simples (SNP), sur les performances de production laitière et la santé. Nous proposons des solutions pour résoudre certains défis précédemment mentionnés (Tableau 0.1). Pour ce faire, il est important de se poser les bonnes questions à savoir :

- *Comment estimer l'effet de la génétique sur le profit dans la production laitière ?*
- *Comment déterminer ou choisir les caractères pertinents pour une meilleure production ?*
- *Comment estimer les futures rentabilités à vie ainsi que les composantes du lait (lait, protéines et gras), en intégrant les caractères génétiques, à moyen ou à long terme ?*

La structure de cette thèse s'organise de manière progressive afin d'accompagner le lecteur dans la compréhension des enjeux scientifiques abordés. Le *chapitre* 1 introduit d'abord un ensemble de notions fondamentales nécessaires à la contextualisation du travail. Le *chapitre* 2 est ensuite consacré à la définition de la problématique ainsi qu'à la formulation des objectifs de la thèse. Le *chapitre* 3 propose une revue de littérature détaillée, permettant de situer notre contribution dans l'état de l'art. Le *chapitre* 4 présente les méthodologies retenues et décrit les données utilisées pour les expérimentations. Les *chapitres* 5 et 6 exposent les résultats obtenus à partir des approches développées dans le chapitre méthodologique, en lien direct avec les objectifs fixés. Enfin, la thèse se conclut par une synthèse générale des travaux réalisés, accompagnée de perspectives de recherche.

Table 0.1 – Principaux défis de l'évaluation génétique et leur impact.

Défi	**Impact**	**Exemple concret**
Qualité et disponibilité des données	Réduction de la précision des estimations si les phénotypes sont incomplets ou bruités	Données de production laitière mal enregistrées en ferme
Coûts du génotypage et du séquençage	Limite l'accès aux technologies modernes de sélection, surtout dans les pays en développement	Génotypage de masse difficilement accessible pour les petits éleveurs
Représentativité des populations de référence	Baisse de la précision prédictive lorsque la population de référence est trop réduite ou éloignée génétiquement	Faible précision pour une nouvelle race bovine non représentée dans la base de référence
Complexité biologique des caractères	Difficultés à modéliser les caractères faiblement héritables et fortement influencés par l'environnement	Fertilité ou résistance aux maladies dans les troupeaux
Modélisation statistique et informatique	Limitations des modèles classiques pour des relations non linéaires ; besoins en calcul intensif	Utilisation de modèles bayésiens
Biais de présélection	Introduction de biais statistiques lorsque les individus sont présélectionnés avant évaluation	Sélection de taureaux sur VGE avant intégration dans les évaluations
Gestion de la consanguinité et de la diversité	Risque de perte de variabilité génétique et d'augmentation des maladies récessives	Sélection intensive dans les races laitières très spécialisées

Les contributions externes de cette thèse s'articulent autour d'un ensemble de productions et de participations scientifiques qui témoignent de la diffusion et de la reconnaissance du travail de recherche au sein de la communauté. Elles comprennent notamment la publication d'un article dans les Proceedings de la conférence *PAMDAS* (voir *Annexe*), accompagnée d'une présentation orale. Elles se prolongent également par une présentation par affiches à la conférence *Deep Learning Indaba (DLI)* et à l'atelier *Women in Machine Learning (WiML@NeurIPS)*. Enfin, la thèse a donné lieu à l'animation d'ateliers scientifiques au sein de la communauté *Women in Machine Learning and Data Science (WiMLDS)*, contribuant ainsi à la vulgarisa-

tion, au partage de compétences et au renforcement de la diversité dans les domaines du machine learning, de la prédiction génomique et de la science des données.

CHAPITRE 1
NOTIONS DE BASE

Dans ce premier chapitre, nous présentons certaines notions fondamentales nécessaires à une meilleure compréhension des travaux développés dans cette thèse. Nous rappelons la terminologie propre à l'industrie laitière et à la génétique. Nous abordons également la notion de séries temporelles, l'intelligence artificielle ainsi que des concepts qui y sont liés, particulièrement l'apprentissage automatique et l'apprentissage profond.

1.1 Industrie laitière

L'industrie laitière constitue un secteur spécifique de l'industrie agroalimentaire, centré sur la production, la collecte, la transformation et la commercialisation du lait ainsi que des produits laitiers dérivés comme le fromage, le yaourt, le beurre et le lait en poudre.

Quant à la production laitière, elle correspond à l'activité agricole qui consiste à élever des animaux producteurs de lait (principalement des vaches, des chèvres et des brebis) et à obtenir le lait brut. Dans le cadre de cette étude, l'attention est portée principalement sur les vaches laitières. Ces dernières engendrent, au-delà des bénéfices, des coûts de gestion. Ce qui nous amène à parler de profit ou de rentabilité.

Définition 1.1 *Le profit à vie ou la rentabilité d'une vache est défini comme la différence entre le revenu total du lait et le coût engagé pour la gestion de cette dernière.*

Plusieurs indicateurs permettent d'évaluer la production laitière et le profit, dont la *production totale* (la quantité totale de lait produite par un animal ou un troupeau sur une période donnée en kg), le *rendement de lait* (la quantité moyenne de lait produite par animal et par jour, en kg), la *santé de la mamelle*, mesurée par le nombre absolu (CCS) ou relatif (SCS) de cellules somatiques (cellules/mL), et la *durée de la lactation* qui désigne la période pendant laquelle l'animal produit du lait après la mise bas.

Définition 1.2 *Après l'insémination et la période de gestation, la lactation commence au moment de la mise bas. Ce cycle, appelé lactation, se poursuit jusqu'au tarissement, c'est-à-dire l'arrêt de la production de lait*

afin de préparer la prochaine mise bas.

En moyenne, la lactation dure environ dix (10) mois, suivie d'une période de tarissement de deux (2) mois. Il faut noter que la forme et la durée des courbes de lactation peuvent varier d'un cycle à l'autre pour un même animal, sous l'effet de facteurs tels que l'environnement, la gestion de l'élevage, l'état de santé et bien d'autres. Les figures 1.1 et 1.2 illustrent les courbes de lactation de différentes vaches. Pour chacune d'elles, on observe la variabilité de la production laitière au cours des cycles de lactation, ainsi que les différences de niveau et de dynamique de production pour un même individu.

Outre les facteurs mentionnés précédemment, la production des vaches est également influencée par les *traits de conformation*.

Définition 1.3 *Ces traits de conformation, exprimés en centimètres (cm), désignent l'ensemble des caractéristiques morphologiques et anatomiques observables chez un animal, en particulier sa structure corporelle et ses proportions.*

Chez les bovins laitiers, ces traits incluent, par exemple, la stature, la forme de la mamelle, l'angle des membres, la largeur du bassin ou encore la profondeur du thorax. La figure 1.3 illustre les différentes régions mesurables. Ces traits, en grande partie héréditaires, sont utilisés comme critères d'évaluation car ils influencent la production laitière, la longévité, la santé et la facilité de traite des animaux. Ces traits de conformation constituent des phénotypes et font partie des caractères génétiques. Dans la section suivante, nous aborderons certaines notions et terminologies propres à la génétique.

1.2 Génétique

La génétique classique a fait ses débuts avec les travaux du moine autrichien, *Gregor Mendel* (1822-1884), qui mena, pendant neuf (9) années, des expériences sur le *pois* (Pisum sativum) afin d'établir et de confirmer ses lois de l'hérédité.

Quant à l'invention du terme *génétique*, elle revient au biologiste anglais *William Bateson* (1861-1926). *Il l'emploie pour la première fois en* 1905 *pour désigner la science de l'hérédité et de la variation des organismes*. Elle étudie les caractères héréditaires des individus d'une espèce, leur transmission au fil des

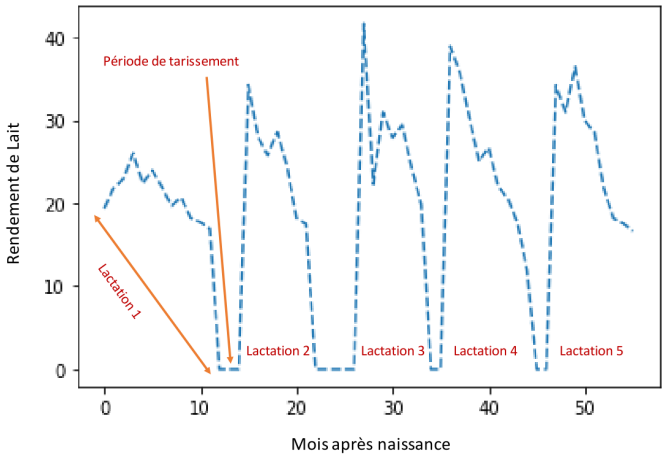
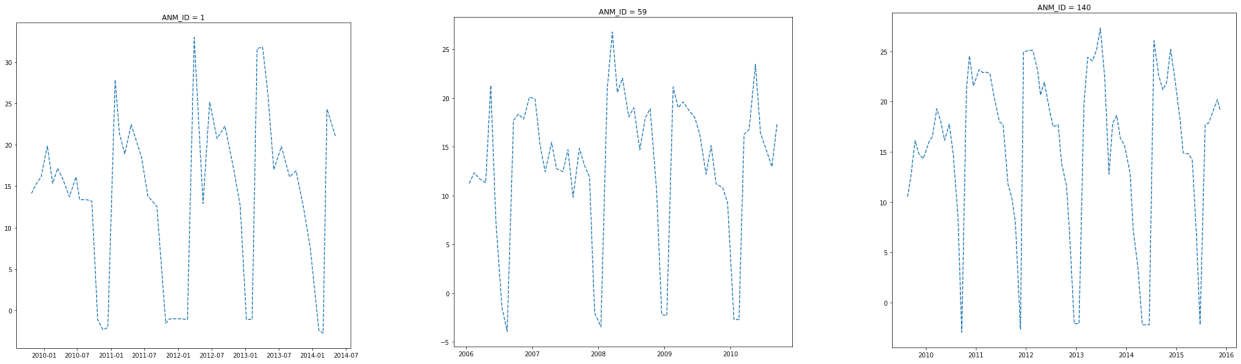


Figure 1.1 – Cycle de lactation de l'animal 140. Cet animal a cinq (5) courbes de lactation. L'unité de mesure du rendement du lait est le kilogramme.



(a) Vache 1. (b) Vache 59. (c) Vache 120.

Figure 1.2 – Courbes de lactation de trois vaches. Chacune représente le rendement de lait (axe des y, en kilogrammes) par rapport au temps (axe des x, nombre de mois après la naissance), c'est-à-dire les cycles de lactation consécutifs d'un même individu. Le nombre de lactations et la forme des courbes sont propres à chaque individu. Anm_id correspond à l'identité de l'animal.

générations et leurs variations. C'est l'étude de cette transmission héréditaire qui a permis l'établissement des lois de Mendel.

La génétique tente de comprendre à quoi sert chaque gène d'un être vivant et de quelle façon ceux-ci se transmettent entre générations. Elle peut se diviser pour des raisons de commodité en trois domaines d'études : la génétique mendélienne ou formelle, la génétique moléculaire et la génétique des populations, la génétique quantitative (Bourgeois et Robinson, 2021). Chez les organismes vivants, on distingue deux

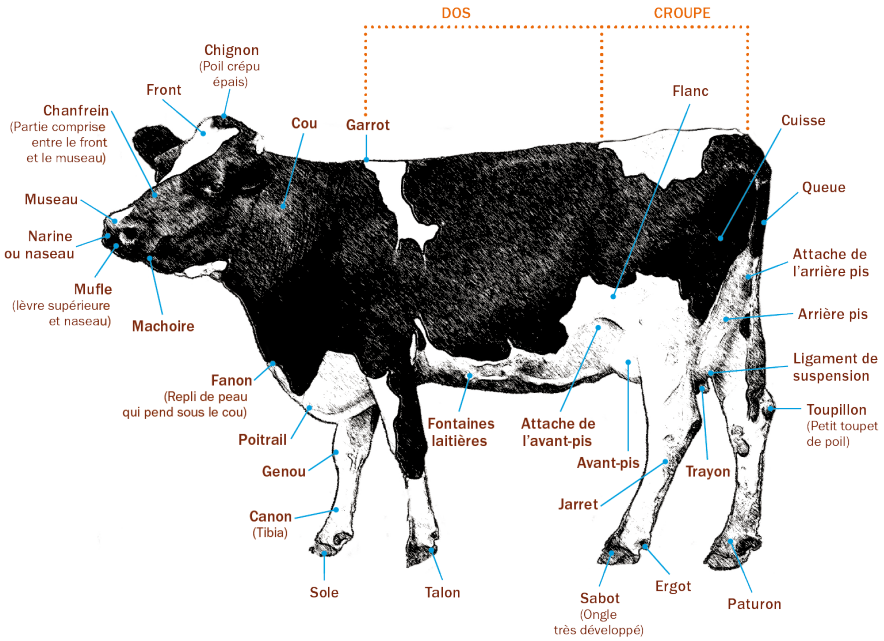


Figure 1.3 – Traits de conformation d'une vache. Ces traits sont des mensurations corporelles de l'animal, en centimètres (cm) (CNIEL, nd).

grands types de systèmes cellulaires :

— Les *eucaryotes*, dont les cellules possèdent un noyau et des organites délimités par des membranes. Les animaux, y compris les humains, les poissons et les oiseaux, les plantes, ainsi que les champignons appartiennent au groupe des eucaryotes.

— Les *procaryotes* (les bactéries), dépourvus de noyau, où l'ADN est librement localisé dans le *cytoplasme*, c'est-à-dire la portion de la cellule comprise entre la membrane plasmique et le noyau, contenant divers composants nécessaires aux fonctions cellulaires.

Chez les *eucaryotes*, le matériel génétique est contenu dans le noyau sous forme de plusieurs brins linéaires qui se condensent, lors des divisions cellulaires, en structures appelées chromosomes. À l'inverse, les bactéries (*procaryotes*) possèdent généralement un seul chromosome circulaire formant une boucle. Un chromosome est constitué d'ADN (acide désoxyribonucléique) associé à des protéines. L'ADN constitue le support moléculaire de l'information génétique héréditaire.

Définition 1.4 *Chez les eucaryotes, un gène est une séquence spécifique d'ADN située dans le noyau et codant pour un produit fonctionnel (souvent une protéine ou un ARN non codant). Il comprend des régions codantes (exons), séparées par des introns, ainsi que des séquences régulatrices qui contrôlent son expression.*

L'ADN subit une réplication au cours du cycle cellulaire, garantissant la transmission fidèle de l'information génétique aux cellules filles. Lorsqu'un gène eucaryote est exprimé, il est transcrit en ARN messager (ARNm) par l'ARN polymérase II. L'ARNm subit ensuite un processus de maturation avant d'être exporté vers le cytoplasme pour être traduit en protéine par les ribosomes.

Chez les *procaryotes*, la structure du gène est généralement plus simple : absence d'introns, organisation en opérons, et couplage transcription-traduction (l'ARNm est traduit alors même qu'il est encore en cours de synthèse). On appelle *opéron* un ensemble de gènes réunis sur un chromosome et soumis à une même régulation.

Pour détecter les variations génétiques, qu'il s'agisse de gènes eucaryotes ou procaryotes, on procède à l'extraction et au séquençage de l'ADN. La comparaison de séquences entre individus ou souches permet d'identifier des variations ponctuelles appelées polymorphismes mononucléotidiques (single nucleotide polymorphisms, SNP), correspondant au changement d'un seul nucléotide.

Sur la figure 1.4, nous avons trois séquences différentes dont l'une est la séquence de référence et les deux dernières. Les SNP peuvent se situer dans les régions codantes, modifiant éventuellement la séquence et la structure des protéines, ou dans les régions non codantes, influençant la régulation de l'expression génique. Leur détection repose sur des technologies de séquençage à haut débit et sur l'analyse bioinformatique, qui permettent l'alignement des séquences et l'identification de ces substitutions.

Ainsi, du gène à l'identification des SNP, qu'il s'agisse d'un eucaryote ou d'un procaryote, le cheminement implique la connaissance de l'organisation du matériel génétique, des mécanismes d'expression et des outils de détection des variations moléculaires. L'ensemble du matériel génétique d'un organisme, y compris tous ses gènes et ses séquences non codantes, constitue son génome.

L'étude des gènes relève de la génétique, discipline indissociable de la *génomique*, qui se concentre sur l'analyse de l'ensemble des gènes d'un organisme, par exemple le génome humain. Le terme *génomique* a été introduit en 1986.

Définition 1.5 *La génomique est une discipline de la biologie moderne. Elle étudie le fonctionnement d'un organisme, d'un organe, d'un cancer, à l'échelle du génome, et non pas limitée à celle d'un ou de plusieurs*

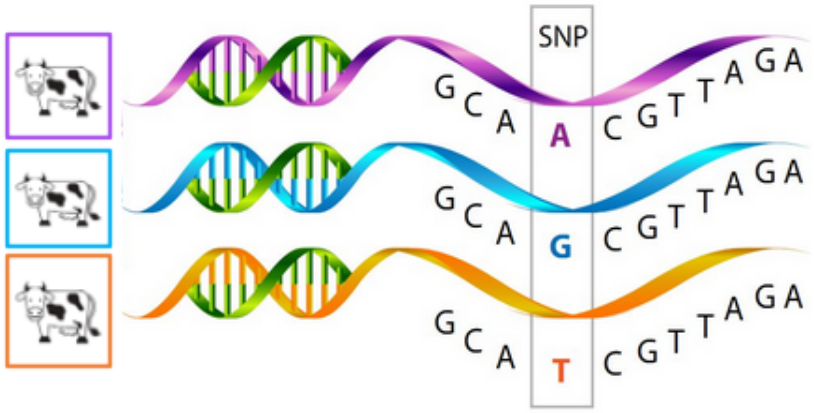


Figure 1.4 – Exemple de polymorphismes nucléotidiques (SNP). La première séquence (en violet) est la référence. Les deux dernières sont les séquences des individus à séquences. Par exemple : SNP 1 = AG (en bleu), SNP 2 : AT (en rouge).

gènes. La génomique se divise en trois branches :

- *La génomique comparative analyse le génome en le comparant à d'autres.*
- *La génomique structurale étudie les génomes en fonction des structures secondaires, tertiaires et quaternaires associées à la conformation du génome ;*
- *La génomique fonctionnelle vise à déterminer la fonction et l'expression des gènes séquencés en caractérisant le transcriptome, le protéome et le métabolome.*

En conclusion, la *génétique* est la discipline qui étudie les variations génétiques, telles que les SNP et autres variants présents dans les gènes, tandis que la *génomique* se concentre sur l'analyse des *génomes* dans leur ensemble.

Pour comprendre l'impact de ces disciplines sur la biologie des organismes, il est nécessaire de distinguer deux concepts fondamentaux : le *génotype* et le *phénotype*. Le *génotype* correspond à l'ensemble de l'information génétique portée par le génome d'un organisme, présente dans chacune de ses cellules sous forme d'ADN (acide désoxyribonucléique).

Définition 1.6 *L'expression d'un génotype, perceptible à travers des caractères morphologiques, physiologiques ou comportementaux, est appelée phénotype. En d'autres termes, le phénotype regroupe l'ensemble des caractéristiques observables d'un individu ou d'un organisme, issues de l'interaction entre son patri-*

moine génétique (génotype) et les facteurs environnementaux qui l'influencent.

La figure 1.5 l'illustre clairement. Depuis plusieurs décennies, l'*évaluation génétique* suscite un intérêt croissant chez les chercheurs, qu'il s'agisse des domaines de la santé, de l'industrie laitière, de l'écologie ou encore de la microbiologie. Elle peut être définie comme l'estimation de sa valeur génétique, réalisée à partir des performances observées sur lui-même et/ou sur ses apparentés (par exemple, son père, sa mère ou d'autres membres de sa lignée). L'objectif principal est de déterminer la *valeur génétique* transmissible à la descendance. Cette valeur est dite *additive* (Ahlborn-Breier et Hohenboken, 1991), car elle correspond à la somme des effets des gènes hérités et peut être transmise de manière stable de génération en génération. Sur les figures 1.6a et 1.6b, on observe comment les individus héritent génétiquement de leurs ancêtres. En

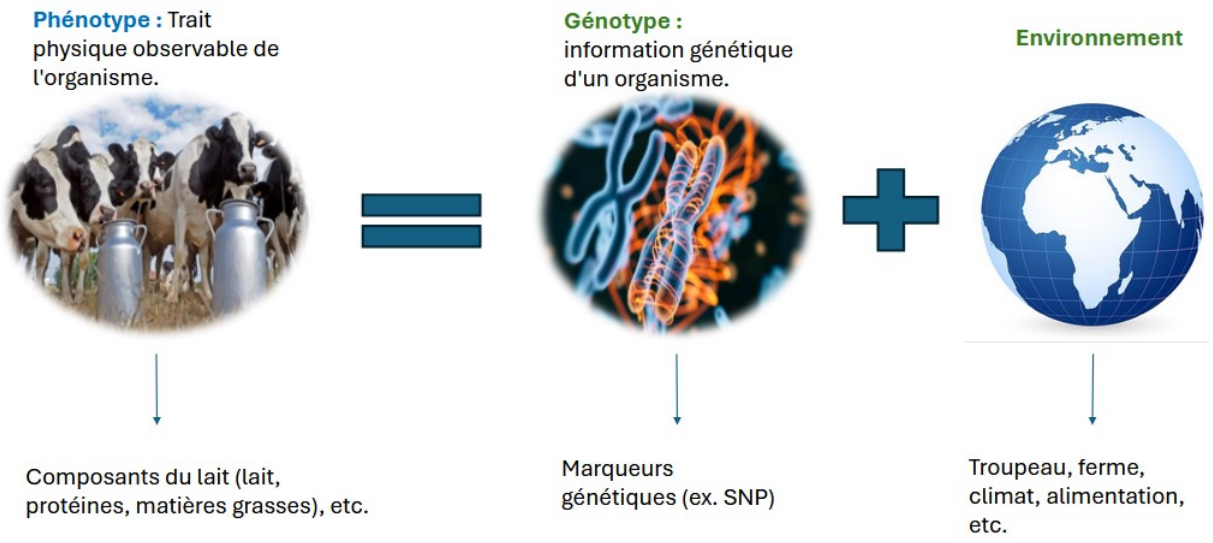


Figure 1.5 – Décomposition classique du phénotype en génétique quantitative. Le phénotype résulte de l'interaction entre le génotype et l'environnement.

pratique, cette estimation repose sur l'exploitation conjointe des informations phénotypiques (mesures de performance) et généalogiques (liens de parenté). Par exemple :

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

avec

- $\mathbf{y}$ le vecteur d'observations, le phénotype (la quantité journalière de lait en kilogrammes) ;
- $\mathbf{X}$ la matrice d'informations (génotype, pédigrée, troupeau, interaction) ;

— β les effets à estimer, les valeurs génétiques estimées (VEE ou VGE). Ils ont la même unité que le phénotype $\mathbf{y}$.

— ϵ l'erreur résiduelle.

À l'inverse, la partie du patrimoine génétique qui résulte de combinaisons nouvelles et aléatoires à chaque génération est qualifiée de *non additive* (Ahlborn-Breier et Hohenboken, 1991). Elle regroupe notamment les *effets de dominance* (interaction entre les deux allèles d'un même gène) et d'*épistasie* (interaction entre des gènes situés sur des loci différents). On appelle *locus*, ou *loci* au pluriel, la localisation précise d'un gène sur le chromosome qui le porte.

$$\text{Phénotype (P)} = \text{Génotype (G)} + \text{Environnement (E)} \tag{1.1}$$

avec

$$G = A + D + I$$

où :

— A = effets additifs (somme des effets des allèles pris individuellement),

— D = effets de dominance (interactions entre allèles au même locus),

— I= effets d'épistasie (interactions entre gènes situés sur des loci différents).

En somme :

$$P = (A + D + I) + E$$

C'est à partir de cette décomposition que l'on introduit la notion d'héritabilité. Avant d'en proposer une définition formelle, il convient de préciser ce que recouvre le concept d'héritabilité.

Définition 1.7 *L'héritabilité désigne le degré de correspondance entre le phénotype et la valeur génétique additive d'un individu pour un caractère donné (Carena* et al., *1981).*

L'estimation de cette valeur génétique repose le plus souvent sur l'étude de la descendance, en cultivant et en évaluant la progéniture de l'individu considéré. Une forte corrélation entre le phénotype et la valeur génétique traduit une héritabilité élevée, généralement associée à des caractères qualitatifs, simples et peu influencés par l'environnement. À l'inverse, lorsque les effets de dominance, d'épistasie ou de l'environnement sont prédominants, l'héritabilité tend à être faible, ce qui est typiquement le cas des caractères quantitatifs. Il importe de souligner que toute estimation est strictement dépendante de la population sur laquelle elle est réalisée (Carena *et al.*, 1981).

Les estimations de l'héritabilité peuvent être établies *en sens large* ou *en sens étroit*, selon la composante de la variance génétique considérée. De plus, elles peuvent être calculées à l'échelle individuelle ou sur la base des moyennes de descendances, en fonction de la génération étudiée. L'*héritabilité au sens étroit* se définit comme le rapport de la variance génétique additive à la variance phénotypique totale.

Revenons à l'équation 1.1. Nous pouvons l'écrire en termes de variance :

$$V_P = V_G + V_E = (V_A + V_D + V_I) + V_E$$

où :

- V_P = variance phénotypique (variabilité observée dans la population),
- V_G = variance génétique totale,
- V_A = variance additive (héritable, base de la sélection),
- V_D = variance due à la dominance,
- V_I = variance due à l'épistasie,
- V_E= variance environnementale.

- *Héritabilité au sens large* :

$$h^2 = \frac{V_G}{V_P}$$

- *Héritabilité au sens étroit* (celle largement utilisée en sélection génomique) :

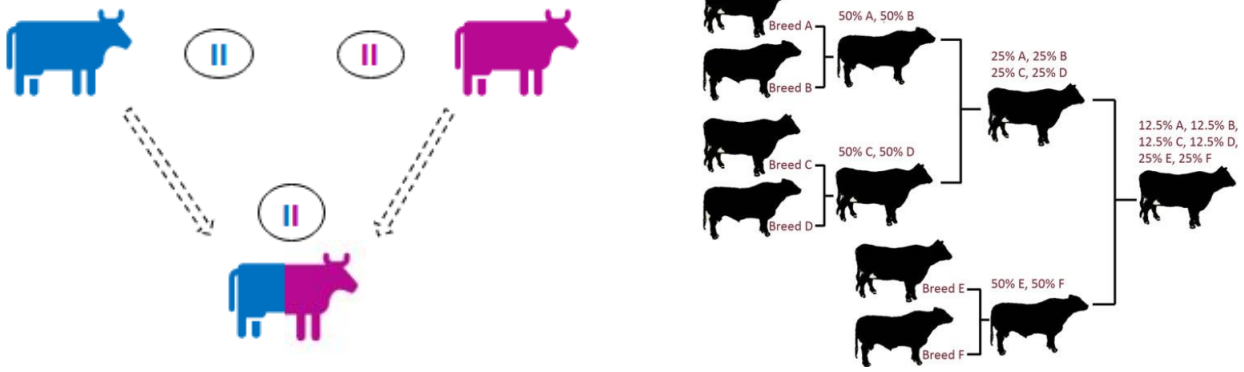
$$h^2 = \frac{V_A}{V_P}$$

1.3 Séries temporelles ou chronologiques

Définition 1.8 *Une série temporelle ou chronologique est une séquence de données mesurées à des instants de temps successifs, généralement espacés de manière régulière (par exemple : quotidiennement, mensuellement ou annuellement).*

Les données de séries temporelles peuvent prendre différentes formes selon le contexte d'étude :

- *Quantitatives* : mesures numériques continues ou discrètes, comme la température quotidienne, la pression sanguine mesurée chaque heure, ou la production laitière journalière.



(a) Exemple 1.

(b) Exemple 2 (Oreobike Blog, 2023).

Figure 1.6 – Hérédité. Les images (a) et (b) illustrent le schéma de transmission du patrimoine génétique d'une génération à l'autre. En effet, chaque individu hérite en moyenne de 50% de son matériel génétique de chacun de ses deux parents.

- *Qualitatives* : observations catégorielles enregistrées dans le temps, par exemple l'état de santé d'un animal (sain, malade, en traitement) ou les conditions météorologiques (ensoleillé, nuageux, pluvieux).
- *Images* : séquences temporelles d'images, utilisées en imagerie médicale (IRM ou échographies successives) ou en vision par ordinateur (vidéosurveillance).
- *Textuelles* : flux d'informations textuelles datées, comme les articles de presse publiés quotidiennement ou les publications sur les réseaux sociaux.
- Sonores : enregistrements acoustiques, par exemple l'analyse des chants d'oiseaux au cours des saisons ou la surveillance des bruits industriels.
- *Vidéos* : séries temporelles d'images animées, comme l'étude du comportement animal filmé en continu ou le suivi de phénomènes naturels (fonte des glaciers, croissance des plantes).

On distingue également plusieurs catégories, parmi lesquelles :

- Séries *univariées* versus *multivariées* :
  - Une série univariée ne contient qu'une seule variable évoluant dans le temps (exemple : prix du lait par jour).
  - Une série multivariée implique plusieurs variables observées simultanément (exemple : la quantité journalière de lait, de protéine et de gras, et la consommation alimentaire d'une vache suivies ensemble, i.e une série $X \in \mathbb{R}^4$).
- Séries *multiples* : composées de différentes séries temporelles correspondant à différents individus

(exemple : $X = [X_1, X_2]$ où $X_1, X_2 \in \mathbb{R}^n$ des séries pour l'individu 1 et 2 respectivement, et $n \in \mathbb{N}$ un entier).

- Séries *stationnaires* versus *non stationnaires* :
  - Une série stationnaire présente des propriétés statistiques (moyenne, variance, autocorrélation) qui restent constantes au cours du temps.
  - Une série non stationnaire évolue avec des tendances ou des variations saisonnières, nécessitant des méthodes de transformation (différenciation, décomposition).

L'analyse de séries temporelles vise à décrire, modéliser et prévoir l'évolution d'un phénomène dans le temps. Elle est largement utilisée dans des domaines variés tels que l'économie, la météorologie, l'écologie, la génétique, ou encore l'industrie laitière (suivi de la production laitière au cours du temps). Les méthodes d'analyse peuvent être classées en deux types : la *régression* et la *prévision* (en anglais, forecasting). La *régression* est une approche statistique qui cherche à établir une relation entre une variable dépendante et une ou plusieurs variables explicatives. Les modèles de régression temporelle (linéaire, polynomiale, régression multiple) permettent d'expliquer l'évolution d'une série. Or, la *prévision* désigne l'ensemble des méthodes permettant de prédire l'avenir d'un phénomène à partir de ses observations passées. Elle dépend du temps.

1.4 Intelligence artificielle

L'intelligence artificielle (IA, ou artificial intelligence AI en anglais) désigne l'ensemble des techniques qui permettent aux machines d'imiter une forme d'intelligence humaine ou des capacités cognitives humaines, notamment la perception, le raisonnement, l'apprentissage, la prise de décision ou encore le langage (McCarthy, 2007; Russell et Norvig, 2021). Ses principaux sous-domaines sont :

- *Apprentissage automatique* (Machine Learning, ML) : permet à des modèles d'apprendre à partir de données (numeriques) pour réaliser des prédictions ou des classifications.
- *Apprentissage profond* (Deep Learning, DL) : sous-domaine du ML qui repose sur les réseaux de neurones capables de modéliser des représentations complexes (images).
- *Vision par ordinateur* (Computer Vision, CV) : vise à analyser, interpréter et comprendre des images et des vidéos.
- *Traitement automatique du langage naturel* (Natural Language Processing, NLP) : permet la compréhension et la génération du langage humain (données textuelles).

— *Apprentissage par renforcement* (Reinforcement Learning, RL) : consiste à entraîner un agent à la prise de décisions par tentatives-erreurs dans un environnement, en maximisant une fonction de récompense cumulative.

Ces sous-domaines reposent sur des concepts communs, parmi lesquels :

1. les *étapes de développement du modèle*
   — Entraînement (training) : apprentissage du modèle par optimisation des paramètres à partir de données (ML/DL) ou d'interactions avec l'environnement (RL).
   — Validation (validation) : étape intermédiaire qui permet d'ajuster les hyperparamètres, de sélectionner le meilleur modèle et de prévenir le surapprentissage.
   — Test (testing) : évaluation finale sur des données ou situations inédites afin de mesurer la capacité de généralisation.
2. la *gestion des données* (dataset split) : dans les approches supervisées (ML/DL, NLP, CV), les données sont en général divisées en trois ensembles :
   — ensemble d'entraînement : pour la construction du modèle.
   — ensemble de validation : permet l'ajustement des hyperparamètres et prévient le surapprentissage.
   — ensemble de test : utilise pour estimer les performances réelles.

   La logique est différente en RL : l'agent génère lui-même ses données en interagissant avec l'environnement, mais des ensembles distincts peuvent être utilisés pour tester la robustesse dans de nouveaux contextes.
3. les métriques : qui quantifient la performance d'un modèle. Elles dépendent du domaine et du type de tâche. On peut citer :
   — Classification (ML/DL, CV, NLP) : précision (accuracy), rappel (recall), F1-score.
   — Régression (ML/DL) : erreur quadratique moyenne (RMSE), erreur absolue moyenne (MAE), coefficient de détermination R^2.

Le Tableau 1.1, (page 21) présente un récapitulatif de certains sous-domaines de l'IA.

L'IA rencontre également deux principaux problèmes :

— *Surapprentissage* (overfitting) qui survient lorsque le modèle s'adapte trop spécifiquement aux données (y compris le bruit) ou aux situations rencontrées, au détriment de sa capacité de généralisa-

Table 1.1 – Tableau comparatif des sous-domaines de l'IA appliqués à la production laitière, l'élevage, l'agriculture et la santé.

Sous-domaine	*Données utilisées*	*Applications typiques (lait, élevage, agriculture, santé)*
Machine Learning (ML)	Données tabulaires (production, alimentation, traits de conformation), séries temporelles (trajectoires de lactation, capteurs IoT)	Prédiction de la production laitière, détection de maladies animales, optimisation des rations alimentaires, systèmes d'alerte en agriculture
Deep Learning (DL)	Images (IRM, échographies, photos d'animaux ou de plantes), signaux capteurs, données textuelles	Détection de pathologies animales ou humaines, classification de maladies des plantes, reconnaissance visuelle pour suivi des animaux, diagnostic assisté par image
Computer Vision (CV)	Images, vidéos (drones, caméras de surveillance d'élevage ou de cultures)	Suivi automatique du comportement animal, détection de boiterie, estimation du poids, détection de stress thermique, surveillance des cultures par drone
Reinforcement Learning (RL)	États et actions générés par interactions avec l'environnement (capteurs, robots, fermes connectées)	Optimisation de la gestion des troupeaux, contrôle intelligent des robots de traite, planification d'irrigation, gestion des ressources agricoles
Natural Language Processing (NLP)	Textes (rapports vétérinaires, littérature scientifique, dossiers médicaux)	Analyse automatique de rapports vétérinaires, extraction d'informations de dossiers cliniques, assistants virtuels pour éleveurs et vétérinaires
Traitement du son (Speech & Audio Processing)	Enregistrements vocaux, signaux acoustiques (sons animaux, bruits d'environnement)	Détection de maladies respiratoires par analyse des sons animaux, suivi du bien-être par vocalisations, interfaces vocales pour éleveurs, détection de bruits anormaux en élevage ou en salle de traite

tion. *Résultats* : très bonnes performances en entraînement, mais mauvaises sur de nouvelles données. *Solutions possibles* : régularisation, dropout, plus de données, early stopping.

— *Sous-apprentissage* (underfitting) qui se produit quand le modèle est trop simple ou mal entraîné pour capturer la complexité des données ou de l'environnement. *Résultats* : mauvaises performances en train et en test. *Solutions possibles* : modèle plus complexe, meilleures caractéristiques (features), plus d'entraînement.

Au cours des dernières décennies, l'*éthique de l'IA* a suscité un intérêt croissant en raison du développement rapide et de l'adoption croissante des technologies basées sur l'intelligence artificielle.

— *Biais et discrimination* : étudient le risque que les modèles reproduisent ou amplifient les inégalités.

— *Transparence et explicabilité* : ont pour but de rendre des décisions prises par l'IA compréhensibles.

— *Vie privée et données personnelles* : respect de la confidentialité et des exigences des lois sur la protection des renseignements personnels (Office of the Privacy Commissioner of Canada, 2026; Government of Canada, 2000; Gouvernement du Québec, 2021; European Union, 2016).

— *Responsabilité* : permet d'identifier et d'attribuer la responsabilité en cas d'erreurs ou de dysfonctionnements d'une technologie d'IA.

— *Durabilité* : analyse l'impact énergétique des modèles massifs (ex. un grand modèle de langage, en anglais Large Language Model LLM).

Pour les études menées au cours de cette thèse, nous n'aurons besoin que du ML et du DL. Avant de définir formellement l'apprentissage automatique (ML) et l'apprentissage profond (DL), nous allons introduire quelques notions d'apprentissage importantes pour cette thèse.

1.4.1 Quelques notions d'apprentissage

En plus des grands paradigmes classiques (supervisé, non supervisé, semi-supervisé, par renforcement), certaines approches spécifiques enrichissent et permettent d'améliorer la qualité des modèles prédictifs : les *variables exogènes* et l'*apprentissage utilisant les informations privilégiées* (en anglais, learning using privileged information, *LuPI*).

1.4.1.1 Variables exogènes

On distingue deux types de variables *exogènes* et *endogènes*.

Définition 1.9 *L'adjectif exogène qualifie ce qui provient d'une source extérieure à un modèle ou un système, ou encore qui se trouve en dehors du modèle. La variable exogène est explicative, mais non expliquée par le modèle. C'est un élément déterminant impactant l'effet de la relation définie par le modèle.*

Définition 1.10 *A l'inverse d'une variable exogène, une variable endogène, dite dans le modèle, est une des variables composant la relation définie par le modèle.*

Définition 1.11 *Une donnée hétérogène est un ensemble de données formé : soit à partir de différentes sources ; soit de différents types de données (quantitative, qualitative, séquences, images, etc).*

1.4.1.2 Apprentissage utilisant les informations privilégiées (LuPI)

Définition 1.12 *Le paradigme d'apprentissage en utilisant des informations privilégiées s'inspire de l'apprentissage enseignant-apprenant. En plus d'un accès aux contenus d'un cours (informations de base), l'apprenant a également accès aux explications de l'enseignant (informations privilégiées) et ce jusqu'au jour de l'examen final. Lors de l'examen final, l'apprenant ne peut compter que sur ses connaissances acquises (informations de base) et n'a pas accès aux explications de l'enseignant. L'apprenant opère sans les informations privilégiées (explications) de l'enseignant. Ce paradigme a été introduit en 2009 par (Vapnik et Vashist, 2009).*

Plus formellement, dans le cadre de l'apprentissage en utilisant des informations privilégiées (LUPI), l'apprentissage est formulé comme un problème supervisé dans lequel, en plus des données d'entraînement habituelles $(\mathbf{x}_i, \mathbf{y}_i)$, on dispose d'informations supplémentaires $(\mathbf{x}_i^*, \mathbf{y}_i^*$, dites privilégiées, accessibles uniquement durant la phase d'apprentissage, mais indisponibles lors de la phase de test et de validation. L'objectif est alors d'estimer une fonction de prédiction $f : \mathcal{X} \rightarrow \mathcal{Y}$ en exploitant ces informations privilégiées afin d'améliorer la généralisation du modèle (Vapnik, 1995; Vapnik et Vashist, 2009). La figure 1.7 illustre le fonctionnement général du paradigme LuPI.

Une approche générale pour exploiter les informations privilégiées est la distillation généralisée, dans laquelle un modèle étudiant est entraîné à minimiser son erreur non seulement vis-à-vis des vraies étiquettes, mais aussi vis-à-vis des cibles générées par un modèle enseignant préalablement entraîné à partir des in-

formations privilégiées (Vapnik et al., *2009; Lopez-Paz* et al., *2015; Hayashi* et al., *2019; Tang* et al., *2019)*.

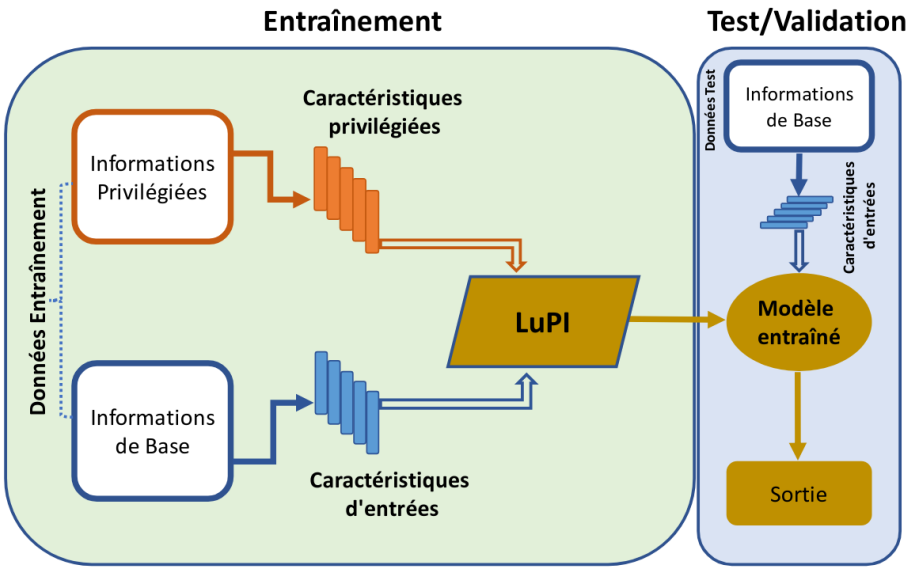


Figure 1.7 – Fonctionnement général de l'apprentissage utilisant les informations privilégiées.

1.4.2 Apprentissage automatique

Le *Machine Learning* (ML) vise à apprendre, à partir de données, une fonction de prévision sur des exemples futurs. On note l'espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$, l'espace de sortie $\mathcal{Y}$ (par exemple $\mathbb{R}$ en régression, $\{1, \ldots, K\}$ en classification avec K le nombre total de classes) et une distribution inconnue P sur $\mathcal{X} \times \mathcal{Y}$. On observe un échantillon d'entraînement indépendant et identiquement distribué (i.i.d).

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \sim P^n.$$

1.4.3 Réseaux de neurones

Un réseau de neurones artificiel, ou réseau de neurones, est un modèle qui s'inspire du fonctionnement du cerveau humain et se compose d'un ensemble de neurones artificiels interconnectés et organisés en couches (Goodfellow *et al.*, 2016). Le neurone artificiel constitue une modélisation mathématique simplifiée du neurone biologique. Il reçoit des signaux en entrée, les combine à l'aide d'une transformation pondérée, puis transmet le signal résultant aux neurones des couches suivantes. Un réseau de neurones comprend généralement trois types de couches :

— Couche d'entrée : reçoit les variables explicatives et les données d'entrée.

— Couches cachées : servent aux transformations intermédiaires.

— Couche de sortie : permet d'émettre une décision finale (la sortie finale).

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{1.2}$$

où x_i sont les entrées, w_i les poids, b le biais, et $f(\cdot)$ une fonction d'activation (par exemple ReLU, sigmoïde, $\tanh$).

1.4.3.1 Perceptron multicouche (MLP)

Un MLP est une composition de couches entièrement connectées. Pour la couche cachée l :

$$\boldsymbol{h}^{(l)} = f\Big(\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}\Big), \tag{1.3}$$

où $\boldsymbol{h}^{(0)} = \boldsymbol{x}$ est l'entrée, $\boldsymbol{W}^{(l)}$ la matrice de poids, $\boldsymbol{b}^{(l)}$ le biais et f l'activation. La couche de sortie prend souvent la forme

$$\hat{\boldsymbol{y}} = g\Big(\boldsymbol{W}^{(L)}\boldsymbol{h}^{(L-1)} + \boldsymbol{b}^{(L)}\Big), \tag{1.4}$$

avec g adapté à la tâche.

1.4.3.2 Réseaux de neurones récurrents (RNN)

Un réseau de neurones récurrent est une architecture conçue pour la modélisation des dépendances séquentielles, dans laquelle la sortie à un instant donné dépend non seulement de l'entrée courante, mais également d'un état interne (état caché) mémorisant l'information issue des étapes précédentes de la séquence (Rumelhart *et al.*, 1986; Goodfellow *et al.*, 2016). Pour une séquence $\{\boldsymbol{x}_t\}_{t=1}^{T}$, on définit :

$$\boldsymbol{h}_t = f(\boldsymbol{W}_{xh}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h), \tag{1.5}$$

$$\boldsymbol{y}_t = g\big(\boldsymbol{W}_{hy}\boldsymbol{h}_t + \boldsymbol{b}_y\big), \tag{1.6}$$

où $\boldsymbol{h}_t$ est l'état caché au temps t. Les RNN classiques peuvent souffrir de gradients explosifs.

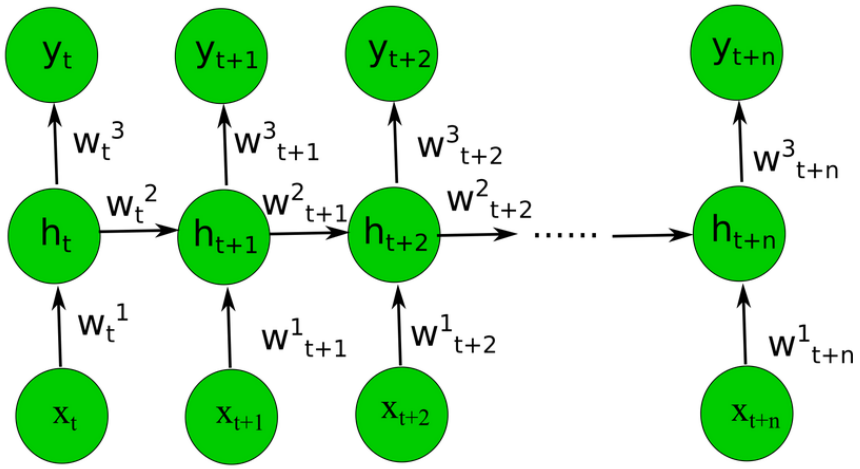


Figure 1.8 – Réseau de neurones récurrent (RNN). Le schéma illustre le fonctionnement d'un réseau de neurones récurrent, de l'entrée des données à la prise de décision finale. Il met en évidence les différentes transformations intermédiaires ainsi que le mécanisme de propagation de l'information au cours de la séquence (Kongakura, 2026).

1.4.3.3 Long Short Term Memory LSTM

L'architecture neuronale LSTM [1] introduit des *portes* pour contrôler le flux d'information et maintenir une mémoire à long terme. Il est souvent utilisé pour modéliser des séries temporelles. La cellule mémoire d'un LSTM (Figure 1.9) est composée de plusieurs portes : une porte d'entrée, une porte de sortie et une porte d'oubli. Ces portes régulent le flux d'informations à l'intérieur de la cellule mémoire, permettant ainsi de contrôler les informations à retenir et celles à oublier. Cela donne aux LSTM la capacité de mémoriser des informations importantes sur de longues séquences et d'ignorer les éléments moins pertinents. Pour les réseaux RNN, l'état caché h_t représente également la mémoire courte du neurone pour les réseaux LSTM. En plus, nous ajoutons un deuxième état pour les réseaux LSTM, appelé c_t qui représente la mémoire à long terme. À chaque pas t, nous avons :

$$\boldsymbol{f}_t = \sigma\big(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f\big) \quad \text{(porte d'oubli)},$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \quad \text{(porte d'entrée)},$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \quad \text{(état candidat)},$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \quad \text{(mise à jour de la mémoire)},$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \quad \text{(porte de sortie)},$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \quad \text{(état caché et sortie)}.$$

1. https://datascientest.com/long-short-term-memory-tout-savoir

avec $\boldsymbol{x}_t$ l'entrée au temps t ; $\boldsymbol{W}_*$, $\mathbf{U}_*$ et $\boldsymbol{b}_*$ sont les paramètres du modèle ; $\sigma(\cdot)$ et $\tanh(\cdot)$ les fonctions d'activation ; et $\odot$ le produit de Hadamard (élément par élément).

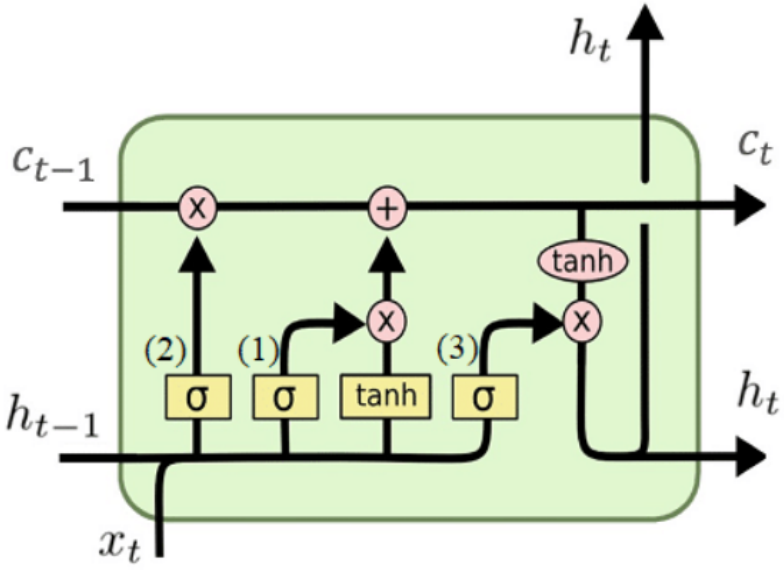


Figure 1.9 – Schéma d'une cellule mémoire d'un LSTM (Ramisarijaona, 2026).

CHAPITRE 2
PROBLÉMATIQUE ET OBJECTIFS

Ce chapitre est consacré à la présentation de la problématique étudiée ainsi qu'à la définition des principaux objectifs associés à sa résolution.

2.1 Définition du problème

La rentabilité constitue un enjeu permanent pour les entreprises de production laitière. Elle repose sur la mise en place de stratégies visant à accroître les revenus tout en réduisant les coûts, afin d'améliorer durablement la performance économique. Plusieurs facteurs influencent directement cette optimisation, parmi lesquels la santé des animaux, leur environnement, la gestion de l'exploitation, la production et la *génétique*. L'estimation de l'effet des caractères génétiques sur la production est généralement réalisée à l'aide de modèles statistiques classiques, tels que la régression. En revanche, les approches plus récentes reposant sur l'apprentissage automatique ou l'apprentissage profond demeurent encore peu exploitées dans ce domaine.

Toutefois, les modèles statistiques traditionnels, bien qu'efficaces pour estimer des effets linéaires et relativement simples, présentent des limites lorsqu'il s'agit de capturer la complexité de certaines interactions entre facteurs génétiques, environnementaux et de gestion. C'est dans cette perspective que l'apprentissage automatique et l'apprentissage profond offrent un potentiel considérable : ils permettent de modéliser des relations non linéaires, de prendre en compte un grand nombre de variables simultanément et de *détecter des interactions complexes difficilement accessibles par les méthodes classiques*. Leur intégration dans l'analyse de la production laitière pourrait ainsi améliorer la précision prédictive et fournir de nouveaux outils d'aide à la décision pour les éleveurs et les sélectionneurs.

Dans le cadre de ce projet de recherche, la problématique centrale réside dans la conception de méthodes pour l'estimation de l'effet des marqueurs génétiques sur la profitabilité des animaux.

- *Comment estimer l'effet de la génétique sur le profit dans la production laitière ?*
- *Comment déterminer ou choisir les caractères pertinents pour une meilleure production ?*
- *Comment estimer la rentabilité future ainsi que les composantes du lait (lait, protéines et gras), en intégrant les caractères génétiques, sur le moyen/long terme ?*

Dans cette étude, les marqueurs génétiques considérés regroupent, d'une part, les traits de conformation ou mensurations corporelles (voir Définition 1.3) mesurés chez des vaches en station debout, et d'autre part, des données de séquençage génomique simulées.

L'association de ces deux sources d'information présente un intérêt particulier. Les traits morphologiques, facilement observables et héritables, fournissent des indicateurs indirects mais pertinents de la productivité, de la longévité et de la santé des animaux. En parallèle, les marqueurs génomiques issus du séquençage permettent d'explorer directement la variabilité génétique au niveau moléculaire et d'identifier des polymorphismes (tels que les SNP) associés à des performances économiques (Atkins, 2018). La combinaison de ces approches phénotypiques et génomiques constitue ainsi une stratégie complémentaire pour affiner l'estimation de la valeur génétique et mieux comprendre leur impact sur la profitabilité des animaux.

Les données sont hétérogènes (Définition 1.11) et proviennent de *Lactanet*, qui regroupe Valacta et le Réseau Laitier Canadien (Canadian Dairy Network, CDN).

La **première partie** des données contient des informations sur l'identité de l'animal (numéro d'identification, troupeau, date de naissance, sexe, date d'entrée et de sortie du troupeau, informations sur ses parents) et sur la production laitière (numéro de lactation, rendement de lait, de protéine et de gras, les coûts, la date du test). Ces données sont recueillies de façon périodique, à des dates de test entre 2006 et 2017. Habituellement dans l'industrie laitière et du fait de l'appartenance des exploitations au programme d'amélioration du cheptel laitier *DHI* (Dairy Herd Improvement), les données sont généralement collectées *dix* fois par an au cours d'une lactation de longueur moyenne 305 jours. Ainsi, elles peuvent être considérées comme des séries temporelles (Définition 1.8). Nous faisons face à des données hétérogènes en raison de la présence de variables quantitatives et qualitatives en même temps.

La **deuxième partie** concerne les caractéristiques génétiques des animaux (Définition 1.2) et des traits de conformation. Contrairement aux données de la première partie, cette dernière est collectée une seule fois avant la première lactation des animaux et dépend de plusieurs facteurs comme l'environnement, la santé et le troupeau d'appartenance de l'animal (pour les EBV). Elles ne peuvent donc pas être considérées comme des séries temporelles. Elles sont également hétérogènes. Certains renseignements (mensurations corporelles) sont directement prélevés et d'autres estimés (EBV) en fonction de la moyenne du troupeau d'appartenance. Nous avons 42 caractéristiques et environ 1 million d'animaux. La **dernière partie** repose

sur des *données de séquençage génomique* de type Illumina, largement utilisées pour la caractérisation des variations génétiques à haut débit.

Sachant les données *multi-dimensionnelles*, génétiques et de production, des premières dates de test (premières lactations et/ou secondes lactations) d'une vache, les *composantes du lait* (*lait*, *protéine* et *gras*) et le *profit* à un *horizon futur* h sont des valeurs prévues. Par conséquent, l'entrée de notre problème se compose d'un *certain horizon futur* (en termes de jours) et de *données exogènes* formant une suite de séquences multi-dimensionnelles. Chaque séquence multi-dimensionnelle contient des facteurs liés à l'identification des animaux, à la qualité et quantité de la production laitière, ainsi qu'aux caractéristiques génétiques.

Une **première hypothèse** de résolution est de traiter comme un problème de prévision en ayant comme :

- *Valeurs d'entrées* : un certain *horizon* (temps en jours et non fixé) ainsi que des données exogènes définies comme des *séries temporelles multivariées* et *multiples* :
  - multivariées : les données contiennent différents facteurs ;
  - multiples : chaque animal (vache) a un certain nombre d'enregistrements (nombre variable d'une vache à une autre). Ce qui nous amène à une série pour chaque animal.
- *Valeurs de sorties* : une séquence de rendements des composantes du lait (*lait*, *protéine* et *gras*) et de profits laitiers à un certain horizon futur.

Soit une matrice X de dimension $T \times N$ où T (formule (2.1)) est la taille (le nombre total d'enregistrements disponibles) de la matrice et N le nombre de variables.

$$T = t_1 + t_2 + \cdots + t_k + \cdots + t_K \tag{2.1}$$

avec k l'individu (animal), K le nombre total d'individus, et t_k le nombre d'enregistrements correspondant à l'individu k.

La matrice X est composée d'autres matrices $X_{t_k}^k$ de dimensions $t_k \times N$. En vue de la formulation sous forme matricielle, nous allons fixer une taille t_u telle que $t_k <= t_u, \forall k \in \{1, 2, \cdots, K\}$. Nous ajouterons des valeurs manquantes aux matrices pour tout $t_k < t_u$.

Nous définissons X comme suit :

$$X = \begin{bmatrix} X_{t_1}^1 & Y_{t_1}^1 \\ X_{t_2}^2 & Y_{t_2}^2 \\ \vdots & \vdots \\ X_{t_k}^k & Y_{t_k}^k \\ \vdots & \vdots \\ X_{t_K}^K & Y_{t_K}^K \end{bmatrix} \tag{2.2}$$

où $Y_{t_k}^k$ une matrice $t_k \times M$-dimensionnelle des variables à prédire (des premières dates de test) pour un individu (vache) k ; et une matrice $X_{t_k}^k$ $t_k \times N$-dimensionnelle pour l'individu k définie par :

$$X_{t_k}^k = \begin{bmatrix} x_{1,1}^k & x_{1,2}^k & \cdots & x_{1,N}^k \\ x_{2,1}^k & x_{2,2}^k & \cdots & x_{2,N}^k \\ \vdots & \ddots & \cdots & \vdots \\ x_{t_k,1}^k & x_{t_k,2}^k & \cdots & x_{t_k,N}^k \end{bmatrix}$$

Les sorties attendues sont représentées par $\hat{Y}$ un vecteur de dimension K qui est défini comme suit :

$$\hat{Y} = \begin{bmatrix} \hat{Y}_{t_1+h_1}^1 \\ \hat{Y}_{t_2+h_2}^2 \\ \vdots \\ \hat{Y}_{t_k+h_k}^k \\ \vdots \\ \hat{Y}_{t_K+h_K}^K \end{bmatrix} \tag{2.3}$$

avec

$$\hat{Y}_{t_k+h_k}^k = f(t_k + h_k | X_{t_k}^k, Y_{t_k}^k) \tag{2.4}$$

où $f(.|.)$ est une fonction.

Une **seconde représentation** serait de considérer que tous les animaux ont le même nombre t de données historiques, tel que $t_k = t, \forall k$.

Une **troisième hypothèse** serait de considérer certaines informations comme des *informations privilégiées* (Définition. 1.12), (Lambert *et al.*, 2018; Tang *et al.*, 2019; Karlsson *et al.*, 2021).

- *Valeurs d'entrée (apprentissage)* : une matrice X_1^p contenant les données de bases, ici les données de la première (1^{ere}) lactation (plus les caractéristiques génétiques).
- *Informations privilégiées* : les données des autres lactations (2^{eme}, 3^{eme}, ainsi de suite) avant un certain horizon temporel cible, i.e. $X_2^p, \cdots, X_L^p$ avec L le nombre total de lactation avant un certain horizon temporel.

  On pourrait considérer également des informations privilégiées *partielles*, car les autres lactations ne sont toujours pas disponibles pour tous les individus (vaches).
- *Sortie* : $\hat{Y}$, les rendements futurs des composantes du lait et la rentabilité future à un certain horizon temporel.

A noter que les données contiennent les attributs de la production laitière et les caractéristiques dont on souhaiterait estimer l'effet. Dans notre cas, ce sont les caractéristiques génétiques. On pourrait faire une combinaison de ces caractéristiques génétiques afin de mieux étudier leur effet. Par exemple : mensurations corporelles, EBV, mensurations corporelles $+$ EBV.

La matrice $X^{p,\ell}$, avec ℓ le numéro de lactation, est définie comme suit :

$$X^{p,\ell} = \begin{bmatrix} X_{t_1}^{p,\ell} & Y_{t_1}^{p,l} \\ X_{t_2}^{p,\ell} & Y_{t_2}^{p,\ell} \\ \vdots & \vdots \\ X_{t_k}^{p,\ell} & Y_{t_k}^{p,\ell} \\ \vdots & \vdots \\ X_{t_K}^{p,\ell} & Y_{t_K}^{p,\ell} \end{bmatrix} \tag{2.5}$$

ou $X_{t_k}^{p,\ell}$ est presque définie de la même façon que $X_{t_k}^k$ de la première hypothèse. La seule différence est que $X_{t_k}^{p,\ell}$ ne contient que les données de la ℓ^{ieme} lactation. La même chose pour les $Y_{t_k}^{p,\ell}$ qui est similaire à celle de $Y_{t_k}^k$.

Une **quatrième représentation** serait de faire une combinaison des *variables exogènes* (les caractéristiques génétiques) et des *informations privilégiées*.

2.2 Objectifs

Afin d'apporter des solutions à notre problème d'évaluation génétique, notre étude aborde quatre (4) objectifs principaux définis comme :

2.2.1 Objectif 1 : Méthode d'estimation de l'effet des marqueurs génétiques sur la rentabilité

Concevoir des modèles d'évaluation de l'effet des marqueurs génétiques sur la rentabilité à vie et les rendements du lait (ainsi que ses composantes protéines et gras).

L'impact de la génétique sur la production laitière ainsi que sa rentabilité est indéniable. Des modèles statistiques (comme BLUP, modèle animal) basés sur les modèles de régression ont été développés (Ducrocq, 1990; Schaeffer, 1991; Tribout *et al.*, 1998; Valergakis *et al.*, 2010a; Upra-Bovine et LAFLEUR, 2017; Špehar *et al.*, 2022)] pour estimer les marqueurs génétiques et étudier leur effet. D'autres recherches menées ont déjà proposé des modèles d'évaluation de l'environnement et de la santé dans le secteur animalier en utilisant des modèles de régression (linéaire, logistique) et d'autres modèles statistiques (comme NARX) (Ruebel *et al.*, 2022; Zhang *et al.*, 2019; Zhang *et al.*, 2020; Ducrocq, 1990). En santé, des modèles de régression linéaire mixte (Uffelmann *et al.*, 2021) ont également été développés. Le but est de proposer des modèles d'apprentissage automatique et profond (comme LSTM, CNN) et VARMAX afin d'estimer l'effet de la génétique sur la rentabilité à vie et les rendements de lait.

2.2.2 Objectif 2 : Méthode d'évaluation génétique à partir des séquençages génomiques (polymorphismes nucléotidiques SNP)

Construire des modèles d'évaluation génétique à partir directement des SNP sans tenir compte de l'effet de la population / cohorte.

Les données génétiques sont collectées *une fois dans la vie d'un individu* et dépendent de la population ou troupeau d'appartenance (et de l'environnement). Cet effet de la population est susceptible à des changements modifiant les données génétiques d'une année à une autre. En d'autres termes, les données géné-

tiques sont souvent issues du *phénotypage* des individus. Ce qui rend toutes comparaisons (et recherche de relations) complexes. La conception de modèles d'évaluation génétique comme dans (Uffelmann *et al.*, 2021) (voir l'étape d'imputation de son architecture Figure 0.2) qui considère directement les SNP (Définition 1.2) est abordée comme notre second objectif à travers une combinaison de CNN et LSTM en utilisant les informations privilégiées comme dans (Lambert *et al.*, 2018; Karlsson *et al.*, 2021).

2.2.3 Objectif 3 : Méthode d'évaluation génétique d'un individu à partir des marqueurs génétiques de ses ancêtres

Concevoir des modèles d'évaluation génétique d'un individu connaissant les caractères génétiques de ses ancêtres.

Parmi les défis des précédents travaux de recherche (Ducrocq, 1990; Uffelmann *et al.*, 2021), l'inaccessibilité aux informations des ancêtres (parents, grands-parents, arrière-grands-parents), à un certain niveau générationnel, est souvent un frein à la conception de certains modèles d'évaluation génétique. A noter que chaque parent (père ou mère) transmet la moitié de son patrimoine génétique à sa progéniture. De ce fait, un individu (enfant) reçoit deux moitiés différentes des patrimoines génétiques parentaux. Étant donné que nous avons à notre disposition des données répondant à ces critères, notre but est de proposer des modèles d'évaluation génétique d'un individu connaissant le patrimoine génétique de ses ancêtres ; par exemple à travers les graphes dynamiques (Boutemine, 2016; Zheng *et al.*, 2019; Wu *et al.*, 2020).

2.2.4 Objectif 4 : Méthode de prédiction du profit et des rendements de production

Construire des modèles de prévision de la rentabilité à vie et les rendements laitiers en intégrant les caractères génétiques et le management.

Beaucoup de modèles de prédiction existent déjà dans le secteur laitier, mais sont limités pour la majorité par manque de généralisation (des hypothèses difficilement satisfaisables dans d'autres secteurs). Notre objectif est de proposer d'autres architectures basées sur des modèles d'apprentissage automatique et/ou profond, en intégrant les caractéristiques génétiques sous forme de variables exogènes (Définition 1.9) ou d'informations privilégiées (Définition 1.12). Le but est de proposer des modèles capables de prédire, à un certain horizon temporel (par exemple un horizon $h = 10$), la rentabilité à vie et les rendements laitiers d'un individu donné, sachant ses caractéristiques génétiques.

Après la formulation de la problématique et la définition des principaux objectifs de cette thèse, le chapitre suivant propose un état de l'art des travaux pertinents dans ce domaine.

CHAPITRE 3
REVUE DE LITTÉRATURE

Dans ce chapitre, nous réalisons une revue de la littérature consacrée aux travaux en lien avec notre problématique. Nous y présentons les contributions majeures ainsi que leurs limites. Cette analyse permet de mettre en évidence les perspectives de recherche afin de positionner précisément l'apport scientifique de cette thèse.

3.1 Modèles d'évaluation génétique

Aujourd'hui, les modèles d'évaluation génétique peuvent être regroupés en trois (3) grandes catégories principales :

- **Modèles basés uniquement sur le pédigrée** : ils utilisent les informations de parenté (ascendance et descendance) pour estimer la *valeur génétique* (VEE), voir la Définition 3.1.
- **Modèles basés uniquement sur les données génomiques** : ils exploitent directement les marqueurs moléculaires (ex. : SNP) pour prédire la valeur génétique.
- **Modèles combinant pédigrée et données génomiques** : ils intègrent simultanément l'information généalogique et les marqueurs moléculaires, offrant une approche plus complète et précise (ex. : *single-step GBLUP*).

Définition 3.1 *Les **valeurs d'élevage estimées** (VEE), ou **estimated breeding values**, sont des estimations de la composante génétique des performances individuelles (Upra-Bovine et LAFLEUR, 2017). Elles sont calculées en comparant les performances d'un animal à la moyenne des performances d'un **lot**, c'est-à-dire un groupe d'animaux de même race, sexe et catégorie d'âge, élevés dans des conditions identiques. Les VEE ne sont donc pas comparables entre races différentes.*

3.1.1 Modèles basés uniquement sur le pédigrée

3.1.1.1 Définition mathématique du problème

L'objectif est d'estimer la valeur génétique de l'élevage, sachant les informations phénotypiques (Définition 1.6) et généalogiques (liens de parenté) de l'individu (l'animal). Mathématiquement, on peut le définir à l'aide d'un modèle linéaire mixte représenté sous forme matricielle, comme suit :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \epsilon, \tag{3.1}$$

avec

$$\begin{bmatrix} \mathbf{u} \\ \epsilon \end{bmatrix} = \mathbb{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \Big),$$

- $\mathbf{y}$ un vecteur connu d'observations phénotypiques (ex. : productions laitières des vaches), de moyenne $\mathbf{Xb}$;
- $\mathbf{b}$ un vecteur inconnu des effets fixes (ex. : année, saison de vêlage, troupeau) ;
- $\mathbf{u}$ un vecteur inconnu d'effets aléatoires (effets génétiques additifs aléatoires comme les valeurs génétiques d'élevage des animaux), de moyenne 0 et ayant pour matrice de variance-covariance $var(\mathbf{u}) = \mathbf{G}$;
- $\mathbf{X}$ et $\mathbf{Z}$ des matrices d'incidence reliant respectivement $\mathbf{b}$ et $\mathbf{u}$ aux observations $\mathbf{y}$;
- ϵ un vecteur inconnu d'erreurs aléatoires ou de résidus aléatoires, de moyenne 0 et de variance $var(\epsilon) = \mathbf{R}$.

On suppose que :

$$\mathbf{u} \sim \mathbb{N}(0, \mathbf{A}\sigma^2_{\mathbf{u}}), \quad \epsilon \sim \mathbb{N}(0, \mathbf{I}\sigma^2_{\epsilon})$$

avec :

- $\mathbf{A}$: matrice de parenté (basée sur le pédigrée),
- $\sigma^2_{\mathbf{u}}$: variance génétique additive,

— σ^2_ϵ : variance résiduelle,

— $\mathbf{G} = \mathbf{A}\sigma^2_\mathbf{u}$ et $\mathbf{R} = \mathbf{I}\sigma^2_\epsilon$.

À noter que le modèle linéaire mixte est également un outil flexible pour l'adaptation d'autres modèles.

Voici un exemple appliqué aux vaches laitières, nous allons considérer l'évaluation de la production laitière :

— $\mathbf{y}$: productions de lait observées (en kg) pour chaque vache,

— $\mathbf{b}$: effets fixes = année, saison de vêlage, troupeau,

— $\mathbf{u}$: valeurs génétiques additives des vaches (VEE),

— $\mathbf{A}$: matrice de parenté (ex. : une vache et sa mère partagent 50% de leurs gènes).

— Les éléments de $\mathbf{Z}$ (matrice d'incidence des effets aléatoires) indiquent le lien entre une observation et l'animal auquel elle appartient :

$$z_{ij} = \begin{cases} 1 & \text{si l'observation } i \text{ provient de l'animal } j \\ 0 & \text{sinon} \end{cases}$$

Supposons trois vaches V_1, V_2, V_3 dont on observe chacune en lactation :

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \mathbf{Xb} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \mathbf{e}$$

Ici :

— la matrice $\mathbf{Z}$ est une matrice identité 3×3, car chaque observation correspond directement à une vache,

— u_1, u_2, u_3 sont les valeurs génétiques additives (VEE) des trois vaches.

Si une vache a plusieurs mesures (par exemple plusieurs lactations), alors plusieurs lignes de $\mathbf{Z}$ contiendront un *1* dans la même colonne correspondant à cette vache.

Nous aborderons à présent les principales approches méthodologiques développées pour l'estimation des valeurs génétiques additives (VEE).

3.1.1.2 Modèles existants

(Ducrocq, 1990) a fait une étude comparative des différentes techniques d'évaluation génétique des bovins laitiers qui existaient jusqu'en 1990. (Ducos et Bidanel, 1993) ont fait recourt au modèle animal multi-caractères pour estimer l'évolution génétique de six caractères de croissance, carcasse et qualité de la viande entre 1977 et 1990, dans les races porcines *large white* et *landrace français*.

Des recherches se sont concentrées sur : des modèles d'estimation des effets maternels sur le poids corporel juvénile des poulets de chair (Koerhuis et Thompson, 1997) ; la rentabilité d'un programme d'amélioration génétique de moutons laitiers utilisant l'insémination artificielle l'estimation (Valergakis *et al.*, 2010b) ; des effets génétiques sociaux sur le comportement alimentaire et les traits de production chez les porcs (Kavlak *et al.*, 2021) ; l'effet de la consanguinité sur les caractères laitiers chez les moutons d'Istrie (Špehar *et al.*, 2022) ; etc.

Henderson *et al.* ont analysé la survie des veaux à l'aide d'une analyse de survie avec un modèle de risques proportionnels de Weibull et une analyse linéaire de caractères multiples (modèle de taureaux) (Henderson *et al.*, 2011).

Quelques modèles d'évaluation génétique sont rencontrés plus souvent, comme le modèle du meilleur prédicateur linéaire non biaisé (best linear unbiased predictor,**BLUP**), le **modèle animal**, et le **modèle des taureaux** (Henderson, 1950; Ducrocq, 1990). Il existe plusieurs dérivations de ces modèles de base. C'est en 1950 que l'évaluation génétique par le BLUP a été introduite par *Henderson* (Henderson, 1950; Dempfle, 1977; Schaeffer, 1991; Robinson, 1991). Le BLUP permet de prendre en compte l'information fournie par la mesure des performances sur tous les apparentés connus afin d'estimer la valeur génétique d'un animal donné. Dans le secteur animalier, le BLUP est une technique d'estimation des mérites génétiques. Il s'agit généralement d'une méthode d'estimation des effets aléatoires. Le BLUP suppose qu'on connaît déjà les composantes de variance ($\mathbf{G}$, $\mathbf{R}$). Mais en pratique, on ne les connaît pas et il faut les estimer. D'où l'intervention de la méthode d'estimation des moindres carrés généralisés (en anglais, generalized least squares GLS). La méthode d'estimation du *maximum de vraisemblance restreinte* (REML, en anglais Restricted Maximum Likelihood) peut également être utilisée pour estimer ces composantes de variance (Meyer et Smith, 1996; Dempfle, 1977).

Le but est de trouver une meilleure estimation $\hat{\mathbf{u}}$, les valeurs génétiques d'élevage $\widehat{\text{VEE}}$, pour le vecteur

inconnu $\mathbf{u}$ étant donné les données $\mathbf{y}$. Les estimations BLUP des valeurs des variables aléatoires $\mathbf{u}$ doivent être :

- **linéaires** : ce sont des fonctions linéaires des données $\mathbf{y}$;
- **non biaisées** : dans le sens où la valeur moyenne de l'estimation est égale à la valeur moyenne de la quantité estimée, c'est-à-dire $E(\hat{\mathbf{u}} - \mathbf{u}) = 0$;
- **meilleures** : elles ont une erreur quadratique moyenne minimale dans la classe des estimateurs linéaires non biaisés. Autrement dit, $Var(\hat{\mathbf{u}} - \mathbf{u})$ ne doit pas être plus *grand* que le $Var(\mathbf{v} - \mathbf{u})$, avec $\mathbf{v}$ comme un autre prédicteur linéaire non biaisé.
- des **prédicteurs** : afin les distinguer des estimateurs d'*effets fixes*.

Le **modèle animal** (ou individuel) est un modèle qui s'intéresse aussi bien à la valeur génétique de l'animal auteur de la performance mesurée, qu'à celle d'individus apparentés (le père, la mère...). D'autres types de modèle animal plus complexe existent comme : modèle animal avec effet de groupes génétiques, modèle animal avec répétabilité (Ducrocq, 1990). Le modèle animal a été utilisé dans plusieurs travaux de recherche (Fournet *et al.*, 1997; Leclère *et al.*, 2014; Ducos et Bidanel, 1993). Il existe également d'autres modèles (Ducrocq, 1990) tels que : BLUP avec modèle animal, le modèle de taureaux (de pères), modèles père et mère, etc.

L'évaluation génétique est réalisée également à l'aide d'autres modèles statistiques et génétiques de plus en plus complexes (Bidanel, 1998) comme :

- **Modèles animaux multivariés** : tiennent compte correctement du biais de sélection lorsque plusieurs caractères sont sélectionnés ; utilisent des informations supplémentaires sur les caractères corrélés ou des informations directes sur des individus apparentés lorsqu'un ou plusieurs caractères de l'objectif de sélection sont difficile (ou même impossible) à mesurer sur les candidats à la sélection. L'amélioration de la précision dépend fortement des paramètres génétiques des caractères considérés (une élévation de la corrélation génétique entre les caractères, différence notable entre les héritabilités ou/et les covariances génétiques et résiduelles, ...).
- **Modèles non linéaires** : Même si les techniques BLUP sont plutôt robustes aux écarts par rapport à la normalité, les modèles non linéaires décrivent plus adéquatement les données observées, lorsqu'on est face à des caractères qui ne sont pas normalement distribués. L'inférence dans les modèles non linéaires est souvent basée sur la **méthodologie bayésienne**, qui offre un cadre global pour l'estima-

tion des paramètres de localisation et de dispersion. Jusqu'à récemment, les algorithmes disponibles étaient basés sur des approximations asymptotiques des distributions postérieures (Ducrocq, 1990). Cependant, les choses ont changé au cours des dernières années avec le développement de nouveaux algorithmes basés sur les méthodes de **Monte Carlo à chaîne de Markov**, qui peuvent calculer des distributions postérieures entières.

Alors que l'évaluation classique des valeurs génétiques s'appuie principalement sur les pédigrées et les performances phénotypiques, l'essor des technologies de génotypage à haut débit a ouvert la voie à la *prédiction génomique*, qui intègre directement l'information moléculaire (SNP) dans le processus d'évaluation. En effet, l'information moléculaire (données issues de l'ADN) rend possible le suivi de l'héritage de segments chromosomiques spécifiques et offre ainsi une mesure de l'identité par descendance plus précise que celle obtenue uniquement à partir des pédigrées (Henderson, 1950; VanRaden, 2008; Herrera *et al.*, 2021; Mota *et al.*, 2023). Dans la section suivante, nous présenterons les approches fondées exclusivement sur les données génomiques.

3.1.2 Modèles basés uniquement sur les données génomiques

Avec la *prédiction génomique*, l'objectif est d'estimer la valeur génétique génomique (VGE), ou en anglais genomic estimated breeding value, d'un individu à partir de ses marqueurs génétiques (SNP).

Le modèle de base utilisé est l'équation classique de la génétique quantitative : *Phénotype = Génotype + Environnement*. Il utilise également la régression linéaire mixte, comme les modèles basés sur le pédigrée (Meuwissen *et al.*, 2001). La différence est la définition des composantes de l'équation.

Le modèle génomique s'écrit :

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \epsilon, \tag{3.2}$$

où :

- $\mathbf{y}$ est le vecteur des phénotypes observés,
- $\mathbf{b}$ représente les effets fixes,
- $\mathbf{g}$ est le vecteur des valeurs génomiques d'élevage (VGEs),
- $\mathbf{X}$ et $\mathbf{Z}$ sont les matrices d'incidence,
- ϵ est le vecteur des résidus aléatoires.

On suppose que :

$$\mathbf{g} \sim \mathbb{N}(0, \mathbf{G}\sigma^2_{\mathbf{g}}), \quad \mathbf{e} \sim \mathbb{N}(0, \mathbf{I}\sigma^2_{\epsilon}),$$

avec $\mathbf{G}$ la matrice de relation (parenté) génomique calculée à partir des génotypes SNP. Ici, $\mathbf{Xb}$ représente la moyenne générale. Plusieurs travaux préfèrent le définir, en ignorant les effets fixes, par $\mathbf{I}\mu$, i.e :

$$\mathbf{y} = \mathbf{I}\mu + \mathbf{Zg} + \epsilon. \tag{3.3}$$

Pour estimer les **valeurs génomiques** ($\widehat{\text{VGE}}$) des individus, Meuwissen *et al.* (Meuwissen *et al.*, 2001) ont proposé une extension du modèle BLUP, appelé BLUP génomique (Genomic Best Linear Unbiased Prediction, **gBLUP**), dans lequel la matrice de parenté issue du pédigrée ($\mathbf{A}$) est remplacée par une matrice de relation (similarité) génomique $\mathbf{G}$ construite à partir des marqueurs SNP (l'information moléculaire). Pour calculer cette dernière, une approche a été proposée par (VanRaden, 2008). Soit $\mathbf{M}$ une matrice des génotypes de dimension $n \times m$ (n individus, m SNP), codée en $\{0, 1, 2\}$; et p_j la fréquence allélique du SNP j. On définit la matrice centrée :

$$\mathbf{G} = \frac{\mathbf{WW}^\intercal}{2\sum_{j=1}^{m} p_j(1-p_j)}, \quad \text{avec} \quad W_{ij} = M_{ij} - 2p_j$$

où :

- $\mathbf{W}$ = matrice centrée des génotypes ($n \times m$),
- le dénominateur est un facteur de *normalisation* qui correspond à la somme des variances attendues sur tous les SNP.
- Deux autres versions existent :
  - Méthode 1 : centrage simple + *division par le nombre* de SNP

$$\mathbf{G} = \frac{\mathbf{WW}^\intercal}{m}.$$

,

  - Méthode 2 : centrage + *standardisation* SNP par SNP

$$\mathbf{G} = \frac{\mathbf{WW}^\intercal}{m}, \quad \text{avec} \quad W_{ij} = \frac{M_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}}.$$

D'autres travaux (Meuwissen *et al.*, 2001; Hayashi, 2013) se sont intéressés à l'utilisation de la méthode bayésienne pour estimer les valeurs génomiques d'élevage. Réécrivons l'équation 3.2 s'écrit :

$$y_i = \mu + \sum_{j=1}^{m} z_{ij} g_j + \epsilon_i,$$

où :

— y_i : phénotype de l'individu i,

— μ : moyenne générale,

— z_{ij} : génotype de l'individu i au SNP j,

— g_j : effet du SNP j,

— m : nombre total de SNP,

— $\epsilon_i \sim \mathbb{N}(0, \sigma_\epsilon^2)$: résidu aléatoire.

Les modèles bayésiens les plus connus sont :

— **Bayes A** Chaque SNP est supposé avoir un effet non nul, avec une variance propre :

$$g_j \sim \mathbb{N}(0, \sigma_j^2), \quad \sigma_j^2 \sim \text{Scaled-Inv-}\chi^2(\nu, S),$$

c'est-à-dire une distribution normale centrée avec une variance spécifique tirée d'une loi *scaled inverse-chi carré*.

Bayes B

$$g_j \sim \begin{cases} 0 & \text{avec probabilité } \pi, \\ \mathbb{N}(0, \sigma_j^2) & \text{avec probabilité } 1-\pi, \end{cases}$$

où π est la proportion de SNP supposés sans effet, et σ_j^2 une variance spécifique à chaque SNP.

— **Bayes C**

$$g_j \sim \begin{cases} 0 & \text{avec probabilité } \pi, \\ \mathbb{N}(0, \sigma_{\mathbf{g}}^2) & \text{avec probabilité } 1-\pi, \end{cases}$$

où tous les SNP non nuls partagent une même variance commune $\sigma_{\mathbf{g}}^2$.

— **Bayes π (Bayes Pi)**

$$g_j \sim \begin{cases} 0 & \text{avec probabilité } \pi, \\ \mathbb{N}(0, \sigma_j^2) & \text{avec probabilité } 1-\pi, \end{cases}$$

avec la différence que π n'est pas fixé a priori mais estimé à partir des données, généralement via une loi a priori de type $\pi \sim \text{Beta}(a, b)$.

Hayashi *et al.* (Hayashi, 2013) ont travaillé sur un modèle de régression bayésien intégrant la sélection de variables pour prédire conjointement les valeurs génétiques d'élevage (VGEs) de plusieurs traits. Ils ont proposé des procédures bayésiennes avec une itération de *Monte-Carlo par chaîne de Markov* (Markov chain Monte Carlo, MCMC) et une approximation variationnelle pour l'estimation bayésienne des paramètres

dans ce modèle multi-traits, appelées respectivement *MCBayes* et *varBayes*. En utilisant des jeux de données simulés de génotypes SNP et de phénotypes pour trois caractères présentant des héritabilités élevées ou faibles, les auteurs ont montré que l'analyse multi-traits améliore nettement la précision des prédictions des valeurs génomiques (VGEs) pour les caractères faiblement héritables lorsqu'ils sont corrélés à des caractères fortement héritables, en exploitant la structure de corrélation entre traits. En revanche, pour les caractères faiblement héritables mais non corrélés, la précision obtenue avec l'approche multi-traits était comparable, voire inférieure, à celle de l'analyse à trait unique, selon le paramètre de probabilité a priori qu'un SNP ait un effet nul. Par ailleurs, bien que varBayes fournisse généralement une précision de prédiction légèrement inférieure à celle de MCBayes, cette perte reste modeste, tandis que le temps de calcul est considérablement réduit avec varBayes.

Les méthodes statistiques et bayésiennes mentionnées jusque-là supposaient la linéarité aussi bien entre les traits eux-mêmes qu'entre les SNP et les traits. Or, dans la réalité, cette hypothèse est difficile à satisfaire. D'où l'exploration de méthodes pouvant identifier les relations génétiques complexes, i.e. non linéaires, entre des traits corrélés. Lee *et al.* (Lee *et al.*, 2023) ont développé **deepGBLUP**, un algorithme de prédiction génomique combinant réseaux profonds (deep learning DL en anglais) (Goodfellow *et al.*, 2016) et gBLUP pour améliorer la précision des prédictions génomiques et la sélection. deepGBLUP extrait les effets locaux des SNP via une couche localement connectée (LCL) — similaire à un réseau de neurones convolutionnel (CNN) mais sans partage de poids. Ensuite, il intègre les relations génétiques avec un GBLUP modifié estimant les composantes additive, dominante et épistatique (Définition 1.5). L'idée est de garder la robustesse statistique du BLUP tout en exploitant la capacité des méthodes d'apprentissage profond à capturer des relations non linéaires et complexes entre SNP et phénotypes. Les auteurs de (Shokor *et al.*, 2025) ont également proposé DLGBLUP, un modèle basé sur le DL. DLGBLUP utilise les valeurs génétiques prédites (VGE) issues du GBLUP, puis les affine en capturant les relations génétiques non linéaires entre les traits à l'aide d'un *perceptron multicouche* (MLP).

Quant à Pérez-Rodríguez *et al.* (Pérez-Rodríguez *et al.*, 2020), ils ont introduit un réseau neuronal régularisé bayésien (BRNNO) destiné à la modélisation de données ordinales. Les paramètres du modèle sont obtenus par une combinaison de l'algorithme de *Gibbs Maximum a Posteriori* et de l'algorithme *Expectation-Maximization* (EM) généralisé.

3.1.3 Modèles combinant pédigrée et données génomiques

Certains de ces modèles se basent sur le meilleur prédicteur linéaire non biaisé BLUP qui satisfait à certaines conditions :

— sans biais : la moyenne de la différence entre la valeur réelle et la valeur prédite est nulle ;

— et a la plus petite variance de l'erreur de prédiction.

Le meilleur prédicteur linéaire non biaisé génomique en une seule étape (ssGBLUP, single-step genomic best linear unbiased predictor en anglais) (Hanrahan *et al.*, 2017; Buaban *et al.*, 2021) combine en une unique étape :

— l'information pédigrée (classique BLUP avec la matrice de parenté $\mathbf{A}$),

— et l'information génomique (GBLUP avec la matrice de relations génomiques $\mathbf{G}$).

$$\text{ssGBLUP} : \mathbf{g} \sim \mathbb{N}(0, \mathbf{H}\sigma^2_{\mathbf{g}}), \quad \text{avec} \quad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}^{-1}_{22} \end{bmatrix}$$

La Table 3.1 regroupe les principales méthodologies définies précédemment.

Méthode	Matrice de relations	Informations utilisées
BLUP	$\mathbf{A}$	Pédigrée uniquement
GBLUP	$\mathbf{G}$	Génotypage (SNP) uniquement
ssGBLUP	$\mathbf{H}$	Pédigrée + Génotypage

Table 3.1 – Récapitulatif des différentes catégories : BLUP, GBLUP et ssGBLUP.

3.2 Modèles mathématiques d'estimation des effets d'un facteur

L'étude de (Liseune *et al.*, 2021) propose un modèle pour prédire l'ensemble de la courbe de lactation des vaches laitières en tirant parti des informations historiques sur la production laitière observées au cours du cycle précédent. En outre, les prévisions de la productivité totale d'une vache ne peuvent être obtenues qu'à partir du moment où la dernière entrée de rendement laitier requise par le modèle est observée. Au cours des quatre dernières décennies, l'influence des facteurs météorologiques sur la production laitière a été explorée dans plusieurs études. Ainsi, la relation entre les données météorologiques et la production laitière présente un intérêt particulier pour les systèmes laitiers basés sur les pâturages, tout comme son impact sur la précision des prévisions de la production laitière. Zhang *et al.* ont eu à étudier les effets de l'intégration des facteurs météorologiques : précipitations, heures d'ensoleillement et température du sol

aux modèles de prévision de la production laitière ont été testés. Malgré des résultats variables entre huit scénarios météorologiques différents, le modèle NARX s'est avéré fournir une plus grande précision de prédiction que le modèle MLR pour prévoir le rendement laitier au niveau de chaque vache (Zhang *et al.*, 2020). Dans une autre étude (Smith, 1968), un modèle de régression linéaire a été utilisé pour prédire la production laitière annuelle au niveau national pendant 13 ans en utilisant des facteurs basés sur la production laitière moyenne d'environ un million de vaches. L'impact de l'ajout de ces paramètres météorologiques n'a pas été comparé aux prévisions basées uniquement sur les données de production laitière en raison de la variation des horizons de prévision et des paramètres météorologiques non dynamiques ; par conséquent, l'effet de l'application des paramètres météorologiques n'a pas été quantifié. Il n'est pas certain que le niveau d'amélioration ait été atteint en ajoutant ces paramètres météorologiques au modèle de prévision de la production laitière. L'étude de Smith est le seul ensemble de travaux qui s'est concentré sur l'introduction de paramètres météorologiques pour améliorer la précision des prévisions de production de lait. Cependant, l'étude s'est limitée à des chiffres annuels moyens au niveau de la production laitière nationale, ainsi qu'à utiliser uniquement le modèle de réseau de neurones multi-couches (MLP) sans comparer de modèles supplémentaires. De plus, pendant la période de l'étude sélectionnée, les systèmes de pâturage étaient beaucoup plus sensibles aux effets des conditions météorologiques, car de nombreuses techniques et technologies de gestion des prairies utilisées aujourd'hui n'étaient pas encore développées.

3.3 Modèles avec variables exogènes

Une vaste collection de modèles de prévision de séries temporelles existe, allant des modèles statistiques (tels que ARMA, VARMA, SARIMA (Peixeiro, 2022)) aux modèles d'apprentissage profond (tels que N-BEATS) (Oreshkin *et al.*, 2020; Dallago *et al.*, 2019; Frasco. *et al.*, 2020). Les modèles de prévision sont utilisés, en général, lorsqu'on fait face à des données temporelles. Ces données temporelles sont appelées **séries temporelles** (Définition 1.8). Pour le traitement de ces types de séries temporelles, certains l'abordent comme des séries temporelles hiérarchiques et/ou groupées. D'autres approches se basent sur les modèles d'apprentissage profond (plus particulièrement les réseaux de neurones **récurrents** RNN tels que les modèles DeepAR et MQRNN d'Amazon (Salinas *et al.*, 2017; Wen *et al.*, 2017) ; ainsi que les réseaux neuronaux multicouches complètement connectés comme N-BEATS (Oreshkin *et al.*, 2020)). L'intégration de certains paramétrages à ces modèles ne date pas d'aujourd'hui, plus particulièrement la notion de variables exogènes. Cette notion n'est pas nouvelle pour les modèles de séries temporelles. On peut trouver les modèles de prévision utilisant les modèles statistiques qui intègrent la notion de variables exogènes comme : VARMAX (Peixeiro,

2022), SARIMAX (Peixeiro, 2022), NARX (Xiu et Zhang, 1704; Boussaada *et al.*, 2018; Cheng *et al.*, 2019; Zhang *et al.*, 2020; Yin *et al.*, 2020), etc.

3.4 Apprentissage utilisant les informations privilégiées

Le paradigme d'apprentissage en utilisant des informations privilégiées (Définition 1.12, learning under privileged information (LuPI) en anglais) a été introduit par Vapnik et Vashist (Vapnik et Vashist, 2009). Les auteurs de (Vapnik et Vashist, 2009; Vapnik et Izmailov, 2015a) ont proposé un algorithme LuPI pour seulement les machines à vecteurs de support (SVM).

En effet, de nombreux auteurs ont montré que l'information privilégiée peut être introduite dans la fonction de perte dans le cadre d'une perte multi-tâche ou d'une perte de distillation de manière agnostique par rapport à l'algorithme. Cependant, la question soulevée par (Lambert *et al.*, 2018) est la suivante : *pourrait-on et devrait-on introduire les informations privilégiées en tant qu'entrée au lieu d'une tâche supplémentaire ?* Lambert *et al.* (Lambert *et al.*, 2018) ont utilisé le CNN et RNN (LSTM) en considérant deux hypothèses : l'apprentissage en utilisant les informations privilégiées & celle utilisant les informations privilégiées *partielle*. Leur expérimentation portait sur la classification d'images et la traduction automatique multimodale. Étant donné que l'apprentissage en utilisant des informations privilégiées (LuPI) permet d'inclure des informations supplémentaires (privilégiées) "seulement" lors de l'entraînement de modèles d'apprentissage automatique ; Gauraha *et al.* (Gauraha *et al.*, 2019) se sont inspirés de l'approche de la validation croisée afin de partitionner les données d'apprentissage en K-folds (parties). Ensuite, ils ont utilisé chaque partie pour l'apprentissage d'une fonction de transfert et les parties restantes ont servi pour les approximations des caractéristiques privilégiées en utilisant des modèles de régression. Ils ont appelé leur modèle : un transfert de connaissances robuste (RKT-LuPI). Ils se sont limités aux classifications binaires pour leur expérimentation.

Karlson *et al.* (Karlsson *et al.*, 2021) se sont intéressés à la prédiction à un certain horizon temporel. Ils ont proposé des modèles d'apprentissage en utilisant des séries temporelles dans les systèmes dynamiques linéaires (LuPTS). Ils ont utilisé la stratégie récursive dans un système de Markov. C'est une stratégie naturelle pour prédire Y dans un système de Markov qui consiste à prédire successivement des séries temporelles comme X_2 à partir de X_1, puis X_3 à partir de la prédiction X_2, et ainsi de suite. Leur expérimentation a porté sur la prédiction de la qualité de l'air et de l'évolution des maladies chroniques (comme l'Alzheimer). Leurs travaux présentent certaines limitations. Ils ne se sont intéressés qu'aux séries temporelles provenant de systèmes dynamiques linéaires à temps discret avec un bruit gaussien isotrope où les matrices de

transition pour chaque pas de temps sont estimées séparément. La structure de leurs séries temporelles est markovienne. Afin d'élargir la compréhension et l'applicabilité de l'apprentissage en utilisant des séries temporelles privilégiées, une extension de leur théorie et de leur algorithme LuPTS, pour inclure des transitions et des estimateurs non linéaires ou exploiter des séries stationnaires, pourrait être intéressante. Les travaux de (Tang *et al.*, 2019) ont porté sur l'usage des informations privilégiées pour l'apprentissage multi-tâches. Ils se sont intéressés aux cas de maladies rares dont certaines informations (notes détaillées du médecin) ne sont disponibles, en général, qu'après les diagnostics. Leurs modèles se basent sur les réseaux de neurones récurrents (RNN) et le multi-perceptron (MLP). Ils ont fait leurs expériences et modèles en introduisant trois paramétrages : seulement les caractéristiques de base (CB), seulement les informations privilégiées (PI) et une combinaison des deux CB + PI.

CHAPITRE 4
MÉTHODOLOGIES ET DONNÉES

Ce chapitre présente les cinq méthodologies développées au cours de cette thèse. Il introduit également les données utilisées pour l'expérimentation de ces approches.

4.1 Méthode 1 : AgriGen

4.1.1 Définition du problème

Le problème considéré dans cette section porte sur l'identification d'individus rentables économiquement (en termes de profit) et la mise en place de stratégies personnalisées fondées sur leurs caractères génétiques.

Il peut être formulé de la manière suivante :

Entrée :

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2] \quad \text{avec} \quad \mathbf{X}_i \in \mathbb{R}^{p_i}, \quad i \in \{1, 2\},$$

où p_i représente le nombre de variables du jeu de données i.

Sortie :

$$\mathbf{y} \in \mathbb{R}^{n \times m},$$

où n correspond au nombre d'individus et m au nombre de sorties (indicateurs de production ou de rentabilité).

Objectif : Identifier les individus présentant une rentabilité faible et formuler des recommandations personnalisées, en tenant compte de leurs caractères génétiques et d'autres facteurs influençant le profit.

Pour atteindre cet objectif, nous développons *AgriGen*, un algorithme d'apprentissage. Dans la prochaine section, nous présentons en détail son architecture.

4.1.2 Modèle AgriGen

L'architecture présentée à la Figure 4.1 illustre l'architecture de *AgriGen*, conçu pour identifier les vaches à faible rentabilité et proposer des recommandations personnalisées en s'appuyant sur leurs valeurs génétiques et les variables déterminantes pour le profit.

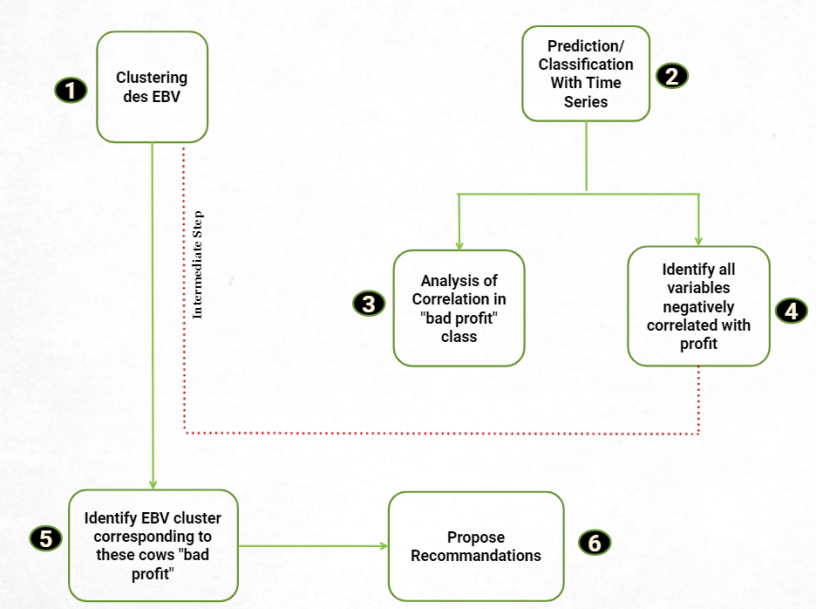


Figure 4.1 – Architecture du modèle *AgriGen*.

Dans ce qui suit, nous détaillons chacune des étapes du protocole méthodologique élaboré et mis en œuvre dans le cadre de ce travail de recherche.

1. *Classification non supervisee (clustering) des EBV (Estimated Breeding Values)* :

   On regroupe les vaches en clusters (clustering, une technique non supervisée qui regroupe des exemples non étiquetés selon leur similarité) selon leurs valeurs génétiques estimées (EBV). Cela permet d'identifier des groupes d'animaux avec des profils génétiques similaires. À cette étape, nous avons fait appel à deux algorithmes de clustering, K-means et Gaussian Mixture Models (GMM), appliqués aux données génétiques (EBV et traits de conformation (Définition 1.3)). Ces clusters nous serviront plus tard dans la partie intermédiaire de notre apprentissage.

2. *Prédiction / Classification avec séries temporelles* :

   On utilise les données de performance (ex. production laitière, coût d'alimentation, santé) sur plusieurs périodes pour prédire ou classer les vaches selon leur rentabilité (profit). Les vaches sont classées dans des groupes comme « *bon profit* » et « *mauvais profit* ».

3. *Analyse de corrélation dans la classe « mauvais profit »* :

   Pour les vaches classées comme peu rentables, on analyse quelles variables (nutrition, santé, reproduction…) sont les plus corrélées négativement avec le profit. Cela permet de cibler les facteurs problématiques.

4. *Identification des variables négativement corrélées au profit* :

   On dresse la liste de toutes les variables ayant un effet défavorable sur la rentabilité. Cette étape peut être combinée avec l'étape 3 pour isoler les leviers d'action.

5. *Identification du cluster EBV des vaches « mauvais profit »* :

   On associe les vaches peu rentables à leur cluster génétique (issu de l'étape 1). Cela permet de voir si certaines caractéristiques génétiques sont liées à de faibles performances économiques.

6. *Proposition de recommandations :*

   En fonction des résultats précédents, on émet des recommandations :
   — Génétiques : sélection des reproducteurs à éviter ou à privilégier.
   — Management : changements dans l'alimentation, la santé ou la reproduction.

7. *Boucles de rétroaction (lignes rouges)* :

   Les informations issues de l'étape 4 peuvent être réutilisées pour affiner le clustering initial (étape 1) et améliorer l'identification des vaches à faible rentabilité.

4.2 Méthode 2 : DairyLuPTS

Introduisons notre deuxième modèle *DairyLuPTS* qui se base sur le paradigme d'apprentissage en utilisant les informations privilégiées (Définition 1.12).

4.2.1 Définition du problème

Dans l'élevage laitier, les données sont collectées au fil du temps. Dans notre problème de prévision, nous avons à faire à des *données temporelles multiples et multivariées* : des *séries temporelles* (Définition 1.8). Dans le domaine de l'élevage laitier, une lactation dure en moyenne 10 mois (Définition 1.2).

Soit $T \in \mathbb{N}$ le nombre d'observations de la série temporelle. Dans notre cas, T représente la taille totale des premières et deuxièmes lactations. Nous supposons que T est fixe pour tous les échantillons d'individus. Soient $P \in \mathbb{N}$ le nombre de caractéristiques d'entrée (Équation 4.2.1) et $K \in \mathbb{N}$ le nombre total d'individus.

$$P = N + M$$

avec un entier $N \geq 1$ qui est le nombre de variables exogènes, et $M \geq 1$ la taille de la sortie.

Entrée :

Définissons une matrice $\mathbf{X} \in \mathbb{R}^{T \times K \times P}$ comme une combinaison de matrices de séries chronologiques d'entrée :

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^k \\ \vdots \\ \mathbf{X}^K \end{bmatrix} \tag{4.1}$$

où $\mathbf{X}^k$ est une matrice de dimension $T \times P$ pour un individu k, $k \in \{1, \cdots, K\}$ définie par :

$$\mathbf{X}^k = \begin{bmatrix} \mathbf{x}_1^k \\ \mathbf{x}_2^k \\ \vdots \\ \mathbf{x}_t^k \\ \vdots \\ \mathbf{x}_T^k \end{bmatrix} = \begin{bmatrix} x_{1,1}^k & x_{1,2}^k & \cdots & x_{1,P}^k \\ x_{2,1}^k & x_{2,2}^k & \cdots & x_{2,P}^k \\ \vdots & \ddots & \cdots & \vdots \\ x_{t,1}^k & x_{t,2}^k & \cdots & x_{t,P}^k \\ \vdots & \ddots & \cdots & \vdots \\ x_{T,1}^k & x_{T,2}^k & \cdots & x_{T,P}^k \end{bmatrix}$$

Sortie :

Les sorties à l'horizon h sont représentées par $\hat{\mathbf{Y}}$, un vecteur à K dimensions, défini par :

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{\mathbf{y}}^1_{T+h} \\ \hat{\mathbf{y}}^2_{T+h} \\ \vdots \\ \hat{\mathbf{y}}^k_{T+h} \\ \vdots \\ \hat{\mathbf{y}}^K_{T+h} \end{bmatrix} \tag{4.2}$$

avec

$$\hat{\mathbf{y}}^k_{T+h} = f(\mathbf{X}^k_{1,\dots,T}) \tag{4.3}$$

où $f(.)$ est une fonction.

Objectif : prédire les performances des individus sur le moyen et le long terme.

4.2.2 Modèle DairyLuPTS

Nous nous inspirons des travaux de *Karlson et al.*, qui proposent d'appliquer le paradigme d'apprentissage avec informations privilégiées aux séries temporelles, en considérant une partie de la série temporelle comme privilégiée (Karlsson *et al.*, 2021). À cette fin, nous commençons par modéliser les données introduites précédemment dans l'équation 4.1. Nous faisons, dans un premier temps, l'hypothèse que *chaque individu dispose du même nombre d'observations* (Algorithme 2 - *DairyLuPTS*). Soit $\mathbf{X}_t, \forall t \in \{1, \cdots, T\}$ une série temporelle de dimension $K \times P$, définie par :

$$\mathbf{X}_t = \begin{bmatrix} x^1_{t,1} & x^1_{t,2} & \cdots & x^1_{t,P} \\ x^2_{t,1} & x^2_{t,2} & \cdots & x^2_{t,P} \\ \vdots & & & \\ x^k_{t,1} & x^k_{t,2} & \cdots & x^k_{t,P} \\ \vdots & & & \\ x^K_{t,1} & x^K_{t,2} & \cdots & x^K_{t,P} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1_t \\ \mathbf{x}^2_t \\ \vdots \\ \mathbf{x}^k_t \\ \vdots \\ \mathbf{x}^K_t \end{bmatrix} \tag{4.4}$$

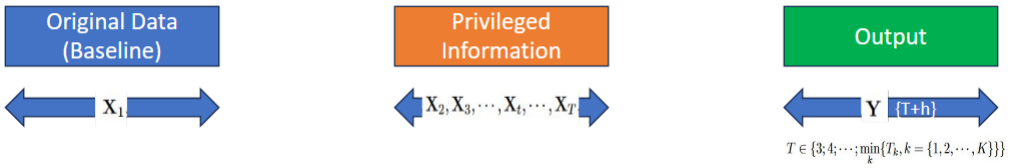


Figure 4.2 – Modélisation des séries temporelles en utilisant les informations privilégiées.

avec $\mathbf{x}_t^k \in \mathbb{R}^P$ un vecteur, pour $k \in \{1, \cdots, K\}$:

$$\mathbf{x}_t^k = \begin{bmatrix} x_{t,1}^k & x_{t,2}^k & \cdots & x_{t,P}^k \end{bmatrix}$$

Ainsi notre série est :

$$\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_t, \cdots, \mathbf{X}_T, \mathbf{Y}$$

Nous cherchons à apprendre des modèles à partir des variables de base $\mathbf{X}_1 \in \mathbb{R}^{K \times P}$, afin de prédire la cible à un horizon $T + h$, avec $h \in \mathbb{N}$. L'objectif est de minimiser la fonction de perte $R(\mathcal{L})$.

Nous faisons l'hypothèse que $\mathbf{X}_1$ correspond aux données de base disponibles à la fois lors de l'apprentissage et de l'évaluation. Les matrices, $\mathbf{X}_2, \mathbf{X}_3, \cdots, \mathbf{X}_t$, $t \leq T$, représentent quant à elles les informations privilégiées, c'est-à-dire des données supplémentaires accessibles uniquement à l'étape d'apprentissage. Enfin, la variable cible est notée Y (Figure 4.2).

Ainsi :

$$\mathbf{X}_{\text{base}} = \mathbf{X}_1, \text{ et } \mathbf{X}_{\text{privilégiées}} = \begin{bmatrix} \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_t \end{bmatrix} \tag{4.5}$$

avec $t \in [\![2, T]\!]$.

4.2.2.1 Stratégie récursive linéaire

Lors de la phase d'apprentissage (Algorithme 2 et Architecture 4.3), nous utilisons une stratégie récursive au sein d'un estimateur linéaire qui exploite des séries temporelles intermédiaires (informations privilégiées). La prédiction à chaque pas temporel t est réalisée au moyen d'un modèle linéaire appris :

$$\widehat{\mathbf{X}}_{t+1} = \widehat{A}_t^\top \mathbf{X}_t, \qquad t = 1, 2, \ldots, T-1, \tag{4.6}$$

où $\widehat{A}_t \in \mathbb{R}^{d \times d}$ est une matrice de paramètres estimée.

La prédiction finale (à l'étape d'apprentissage) est donnée par :

$$\widehat{\mathbf{Y}} = \widehat{\beta}^\top \mathbf{X}_T, \tag{4.7}$$

avec $\widehat{\beta} \in \mathbb{R}^d$.

4.2.2.2 Étape de test

Lors de l'étape de test (Algorithme 2 et Architecture 4.3), seule $\mathbf{X}_1$ est prise en compte. Les modèles sont par la suite combinés de façon récursive pour donner :

$$\widehat{\mathbf{Y}} = (\widehat{A}_1 \widehat{A}_2 \cdots \widehat{A}_{T-1}\, \widehat{\beta})^\top \mathbf{X}_1. \tag{4.8}$$

Ainsi, deux estimateurs peuvent être distingués :

$$\widehat{\mathbf{Y}}_{\text{ols}} = \widehat{\theta}_{\text{ols}}^\top \mathbf{X}_1, \tag{4.9}$$

$$\widehat{\mathbf{Y}}_{\text{lupts}} = \widehat{\theta}_{\text{lupts}}^\top \mathbf{X}_1, \tag{4.10}$$

avec

$$T \in \{2, 3, 4, \ldots, \min_k\{T_k,\ k = 1, 2, \ldots, K\}\}, \tag{4.11}$$

où T_k ($k = 1, 2, \ldots, K$) représente le nombre total d'enregistrements disponibles pour l'individu k.

Les estimateurs s'écrivent :

$$\widehat{\theta}_{\text{ols}} = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{Y}, \tag{4.12}$$

$$\widehat{\theta}_{\text{lupts}} = \widehat{A}_1 \widehat{A}_2 \cdots \widehat{A}_{T-1}\, \widehat{\beta}. \tag{4.13}$$

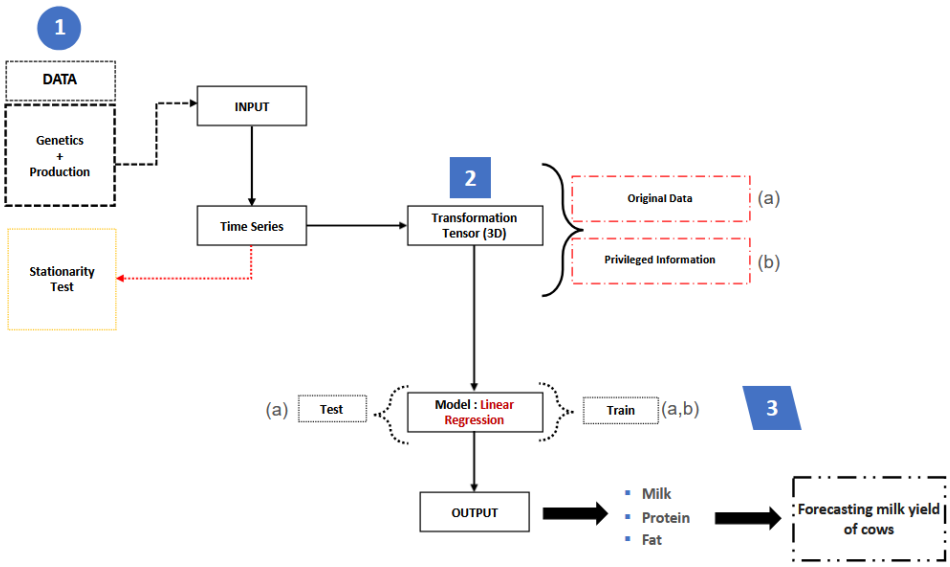


Figure 4.3 – Architecture de *DairyLuPTS*.

Enfin, l'évolution des covariables et la variable cible sont définies par :

$$\mathbf{X}_{t+1} = A_t^\top \mathbf{X}_t, \qquad t = 1, 2, \ldots, T-1, \tag{4.14}$$

$$\mathbf{Y} = \beta^\top \mathbf{X}_T. \tag{4.15}$$

Le modèle de base (Equation 4.5 et Algorithme 1) est la régression linéaire avec la méthode des moindres carrés ordinaires (en anglais, ordinary least squares OLS).

Algorithm 1: Baseline : Moindres carrés ordinaires (OLS) dans les systèmes dynamiques linéaires

Entrées : Données $\mathcal{D} = \{(\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_T^k, \mathbf{y}_k)\}_{k=1}^K$

Préliminaire : Test de stationnarité des séries temporelles

`/* Calculer les derniers paramètres */`

$$\hat{\theta}_{\text{ols}} = \text{argmin}_\beta \frac{1}{K} \sum_{k=1}^{K} \left(\beta^\intercal \mathbf{x}_1^k - \mathbf{y}_k\right)^2$$

Sorties : Estimateurs $\widehat{\theta}_{\text{ols}}$

4.2.2.3 Baseline+

Dans cette partie, nous proposons de modifier les formules

$$\mathbf{X}_{\text{base+}} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{T'} \end{bmatrix} \text{ avec } T' < T, \quad \text{et} \quad \mathbf{X}_{\text{privilegées+}} = \begin{bmatrix} \mathbf{X}_{T'+1} \\ \vdots \\ \mathbf{X}_T \end{bmatrix} \tag{4.16}$$

$$\widehat{\mathbf{Y}}_{\text{ols+}} = \widehat{\theta}_{\text{ols+}}^\top \mathbf{X}_{base+}, \tag{4.17}$$

$$\widehat{\mathbf{Y}}_{\text{lupts+}} = \widehat{\theta}_{\text{lupts+}}^\top \mathbf{X}_{base+}, \tag{4.18}$$

avec

$$T' \in \{2, 3, 4, \cdots < T\}.$$

Algorithm 2: DairyLuPTS : LuPI dans les systèmes dynamiques linéaires

Entrées : Données $\mathcal{D} = \{(\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_T^k, \mathbf{y}_k)\}_{k=1}^K$

Préliminaire : Test of stationnarité des séries temporelles

if *stationnarité* **then**

`/* stationnarité: Calculer une seule fois` $\widetilde{A}$ `pour` $\widehat{A}_t, \forall t = 1, \ldots, T-1$. `*/`

$$\widetilde{A} = \text{argmin}_A \sum_{t=1}^{T-1} \sum_{k=1}^{K} \frac{\|A^\intercal \mathbf{x}_t^k - \mathbf{x}_{t-1}^k\|_2^2}{K(T-1)}$$

$\widehat{A} = \widetilde{A}^{T-1}$

else

`/* absence de stationnarité: Calculer` $\widetilde{A}_t$ `pour chaque pas de temps. */`

for *individu* $t = 1, \ldots, T-1$ **do**

$$\widehat{A}_t = \text{argmin}_A \sum_{k=1}^{K} \frac{\|A^\intercal \mathbf{x}_t^k - \mathbf{x}_{t-1}^k\|_2^2}{K}$$

end

$\widehat{A} = \widehat{A}_1 \widehat{A}_2 \cdots \widehat{A}_{T-1}$

end

`/* Calculer les derniers paramètres */`

$$\widehat{\beta} = \text{argmin}_\beta \frac{1}{K} \sum_{k=1}^{K} \left(\beta^\intercal \mathbf{x}_T^k - \mathbf{y}_k\right)^2$$

Sorties : Estimateurs $\widehat{\theta}_{\text{lupts}} = \widehat{A}\widehat{\beta}$

4.3 Méthode 3 : LSTMDropout

4.3.1 Définition du problème

Entrée : Les données de base sont notées

$$X_{\text{base}} \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}} \times n_{\text{features}_{\text{base}}}},$$

et les données privilégiées

$$X_{\text{priv}} \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}} \times n_{\text{features}_{\text{priv}}}},$$

Algorithm 3: Baseline+ : LuPI dans les systèmes dynamiques linéaires

Entrée : Données $\mathcal{D} = \{(\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_{T'}^k, \mathbf{y}_k)\}_{k=1}^K$

Préliminaire : Test de stationnarité des séries temporelles

if *stationnarité* **then**

`/* stationnarité: Calculer une seule fois` $\widetilde{A}$ `pour` $\widehat{A}_t, \forall t = 1, \ldots, T'-1$. `*/`

$$\widetilde{A} = \text{argmin}_A \sum_{t=1}^{T'-1} \sum_{k=1}^{K} \frac{\|A^\intercal \mathbf{x}_t^k - \mathbf{x}_{t-1}^k\|_2^2}{m(T'-1)}$$

$\hat{A} = \tilde{A}^{T'-1}$

else

`/* absence de stationnarité: Calculer` $\widetilde{A}_t$ `pour chaque pas de temps.` `*/`

for *individu* $t = 1, \ldots, T'-1$ **do**

$$\widehat{A}_t = \text{argmin}_A \sum_{k=1}^{K} \frac{\|A^\intercal \mathbf{x}_t^k - \mathbf{x}_{t-1}^k\|_2^2}{m}$$

end

$\widehat{A} = \widehat{A}_1 \widehat{A}_2 \cdots \widehat{A}_{T'-1}$

end

`/* Calculer les derniers paramètres` `*/`

$$\widehat{\beta} = \text{argmin}_\beta \frac{1}{m} \sum_{k=1}^{K} \left(\beta^\intercal \mathbf{x}_{T'-1}^k - \mathbf{y}_k\right)^2$$

Sorties : Estimateurs $\widehat{\theta}_{\text{OLS+}} = \widehat{A}\widehat{\beta}$

où n_{samples} désigne le nombre d'individus, $n_{\text{seqlength}}$ la longueur des séquences temporelles, et $n_{\text{features}_{\text{base}}}$ et $n_{\text{features}_{\text{priv}}}$ le nombre de caractéristiques de base et privilégiées, respectivement.

Sortie :

$$y = \begin{cases} y \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{outputs}}}, & \text{pour une tâche de régression,} \\ y \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}} \times n_{\text{outputs}}}, & \text{pour une tâche de prévision.} \end{cases}$$

Dans notre cas, $n_{\text{outputs}} \in \{1, 3\}$ représente le nombre de sorties considérées. Une tâche de régression diffère d'une tâche de prévision par la taille des séquences temporelles, qui est de 1 pour la première (ré-

gression) et d'au moins 2 pour la dernière (prévision).

Objectif : Évaluer l'impact de la génétique sur les performances d'un individu. Prédire ses performances futures à partir des données historiques des lactations précédentes, en intégrant des *informations privilégiées (PI)* (voir Définition 1.12), telles que les valeurs génétiques estimées (VEE) et les traits de conformation afin d'améliorer la précision des prédictions.

LSTMDropout offre une approche unique et polyvalente, capable de traiter simultanément les objectifs 1, 2, 3 et 4 définis dans cette thèse. Il sera donc décrit en détail dans la section suivante afin d'en présenter l'architecture.

4.3.2 Modèle LSTMDropout

À la différence des travaux de (Karlsson *et al.*, 2021; Jung et Johansson, 2022; Hayashi *et al.*, 2019), nous définissons les PI en termes de caractéristiques propres à notre étude. Dans le cadre de l'apprentissage de représentations latentes, le problème du goulot d'étranglement (bottleneck problem) constitue un défi, en particulier lors de l'empilement de multiples couches. Le goulot d'étranglement informationnel (information bottleneck, IB) survient lorsque les couches intermédiaires peinent à conserver l'information pertinente tout en éliminant le bruit. La solution la plus couramment adoptée dans la littérature consiste à introduire des mécanismes d'attention, qui se sont révélés efficaces dans divers domaines. En apprentissage sur séries temporelles, lorsque certaines informations intermédiaires font défaut au moment de l'inférence, l'option par défaut consiste à ignorer le champ manquant.

Dans cette thèse, nous proposons une approche alternative fondée sur LuPI avec un dropout hétéroscédastique, tel qu'introduit par Lambert *et al.* (2018) (Lambert *et al.*, 2018). Cette méthode articule le cadre de l'Information Bottleneck (IB) avec des informations privilégiées (PI) afin d'améliorer la précision prédictive, en remplaçant les mécanismes d'attention par une stratégie de dropout spécialisée. Comparée aux modèles à base d'attention tels que le Temporal Fusion Transformer TFT (Lim *et al.*, 2021), notre approche évite le surcoût computationnel lié à l'attention, ce qui la rend particulièrement adaptée aux jeux de données à grande échelle utilisés en élevage laitier. Notre architecture, illustrée à la Figure A.1, s'appuyant sur le paradigme LuPI (Définition 1.12), se compose de trois étapes principales :

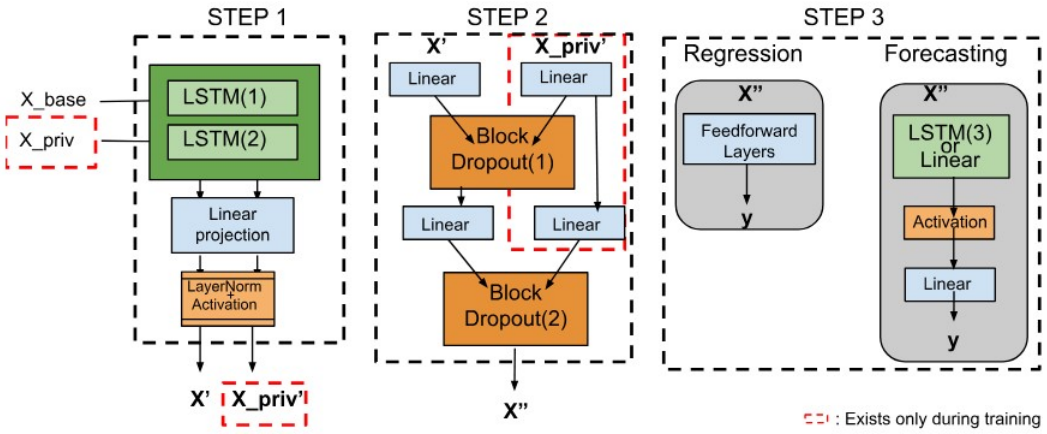


Figure 4.4 – Architecture de notre modèle *LSTMDropout*.

4.3.2.1 Gestion des différences de dimensionnalité

La première étape traite le décalage de dimensionnalité entre les données de base X_{base} de dimension $n_{samples} \times n_{seqlength} \times n_{features_{base}}$ et les données privilégiées X_{priv} de dimension $n_{samples} \times n_{seqlength} \times n_{features_{priv}}$. Des blocs LSTM distincts sont employés pour apprendre des représentations à partir de ces deux entrées. Une fois transformées, ces représentations passent par une couche de projection linéaire, une couche de normalisation, puis par une activation ReLU, de manière à aligner les deux flux pour le traitement ultérieur.

À la différence de Lambert *et al.* (2018) (Lambert *et al.*, 2018), où X_{base} et X_{priv} sont issus de la même source (respectivement une image complète et des régions recadrées), notre approche gère des entrées hétérogènes : des séries temporelles (X_{base}) et des traits phénotypiques (traits de conformation et VEE, X_{priv}). Cette adaptation accroît la complexité, mais garantit que chaque type de données est traité de façon optimale.

4.3.2.2 Estimation de la variance au moyen d'un dropout hétéroscédastique

Cette deuxième étape consiste à estimer la variance du dropout hétéroscédastique, proposé par Lambert *et al.* , à partir des deux représentations apprises précédemment (Lambert *et al.*, 2018). À la différence de l'article de référence (Lambert *et al.*, 2018), où X_{base} et X_{priv} proviennent de la même entrée (respectivement l'image complète pour X_{base} et des recadrages de cette image pour X_{priv}), notre approche traite des entrées hétérogènes : séries temporelles pour X_{base} et traits phénotypiques (traits de conformation et VEE)

pour X_{priv}. Cette hétérogénéité ajoute de la complexité, mais garantit un traitement adapté à chaque type de données.

En modélisant la variance, le dropout hétéroscédastique permet à notre modèle de capturer l'incertitude associée aux informations privilégiées, telles que les traits de conformation ou les valeurs d'élevage estimées (VEE). Il en résulte des prédictions plus fiables que celles obtenues avec un dropout standard ou des mécanismes d'attention, qui ne traitent pas explicitement la variance. Cette augmentation de la fiabilité découle du fait que la variance du dropout est estimée à partir des données observées, plutôt que déterminée par un processus d'optimisation aléatoire. Ainsi, pour calculer la variance du dropout hétéroscédastique pendant l'entraînement, nous appliquons deux couches linéaires distinctes aux deux représentations apprises à l'étape précédente. Ces représentations serviront ensuite d'entrées à la troisième étape. À noter qu'en phase de test, nous n'avons accès qu'aux représentations de X_{base} en entrée de cette deuxième étape ; nous répliquons alors ces représentations avant d'appliquer les couches linéaires et le bloc de dropout.

4.3.2.3 Étape finale

La troisième partie de l'architecture est consacrée aux phases d'entraînement et de test. Selon la tâche considérée, une sortie adaptée est générée.

Pour la *régression*, la sortie est définie comme

$$y \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{outputs}}},$$

comme illustré à la Figure A.1.1. Cette configuration permet de prédire une valeur unique (single output) par individu à l'aide de couches entièrement connectées (feedforward).

Pour la *prévision*, la sortie prend la forme

$$y \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}_{\text{output}}} \times n_{\text{outputs}}},$$

avec par exemple $n_{\text{seqlength}_{\text{output}}} = 10$, comme indiqué à la Figure A.1.2. Dans ce cas, l'architecture applique d'abord soit une couche linéaire (approche feedforward simple), soit une couche LSTM (voir Section 1.4.3) pour capturer les dépendances temporelles, avant de recourir à une couche linéaire finale afin de produire la dimension de sortie souhaitée.

4.3.2.4 Approches pour calculer la variance des dropouts

4.3.2.4.1 Dropout gaussiens

Le *dropout* est une technique de régularisation utilisée pour limiter le surapprentissage dans les modèles d'apprentissage automatique. Son principe consiste à désactiver aléatoirement, lors de l'entraînement, certaines unités d'un réseau de neurones ainsi que leurs connexions. Ce processus stochastique empêche le réseau de dépendre excessivement de certains exemples d'entraînement et favorise ainsi sa capacité de généralisation. Srivastava *et al*. (Srivastava *et al.*, 2014) ont montré que le *dropout gaussien*, qui introduit un bruit multiplicatif de type gaussien, réduit efficacement le surapprentissage par rapport aux méthodes classiques de *dropout* bernoullien.

4.3.2.4.2 Astuce de re-paramétrage

Les auteurs de (Kingma et Welling, 2013; Kingma *et al.*, 2015) ont proposé le *variational dropout* comme une généralisation du *dropout gaussien*. Cette approche permet d'apprendre les taux de *dropout* au cours de l'entraînement, améliorant ainsi les performances du modèle grâce à l'intégration de principes d'apprentissage bayésien. La technique de reparamétrisation consiste à reparamétriser localement le processus de *dropout*, ce qui facilite une optimisation plus efficace par descente de gradient.

4.3.2.4.3 Information Bottleneck

Le *Information Bottleneck* (IB) constitue un cadre théorique puissant pour imposer différentes hypothèses structurelles (Tishby *et al.*, 1999). *Kingma et al*. ont appliqué le cadre IB aux CNN et aux RNN (Définition 1.4) en utilisant la méthode *stochastic gradient variational Bayes* ainsi que l'astuce de reparamétrisation (Kingma et Welling, 2013). Pour les SVM, *Motiian et al*. ont proposé une combinaison entre le paradigme IB et le cadre LUPI (Motiian *et al.*, 2016). L'approche de *Achille et al*. apparaît comme la plus proche de celle proposée par *Lambert et al*. (Achille et Soatto, 2018; Lambert *et al.*, 2018). En effet, *Achille et al*. utilisent le principe de l'IB afin d'apprendre des représentations désentrelacées (*disentangled representations*) lorsqu'un CNN avec *dropout* gaussien est employé (Achille et Soatto, 2018).

4.4 Méthode 4 : Information Dropout Adaptée (IDA)

4.4.1 Définition du problème

Entrée : Les données de base sont notées

$$X_{\text{base}} \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}} \times n_{\text{features}_{\text{base}}}},$$

et les données privilégiées

$$X_{\text{priv}} \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}} \times n_{\text{features}_{\text{priv}}}},$$

où n_{samples} désigne le nombre d'individus, $n_{\text{seqlength}}$ la longueur des séquences temporelles, et $n_{\text{features}_{\text{base}}}$ et $n_{\text{features}_{\text{priv}}}$ le nombre de caractéristiques de base et privilégiées, respectivement.

Sortie :

$$y = \begin{cases} y \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{outputs}}}, & \text{pour une tâche de régression,} \\ y \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{seqlength}} \times n_{\text{outputs}}}, & \text{pour une tâche de prévision (avec } n_{\text{seqlength}} = 10\text{).} \end{cases}$$

Ici, $n_{\text{outputs}} \in \{1, 3\}$ représente le nombre de sorties considérées.

Objectif : Prédire les performances de lactation futures d'un individu à partir des données historiques des lactations précédentes, en intégrant des *informations privilégiées (PI)* (Définition 1.12), telles que les valeurs génétiques estimées (VEE) et les traits de conformation, afin d'améliorer la précision des prédictions.

L'*Information Dropout Adaptée (IDA)* répond spécifiquement à l'objectif 1. Le principe de fonctionnement de ce modèle est exposé de manière détaillée dans la section ci-après.

4.4.2 Modèle Information Dropout Adaptée (IDA)

L'*Information Dropout* (Achille et Soatto, 2018) est une méthode de régularisation basée sur le principe de l'Information Bottleneck (IB). Elle consiste à injecter un bruit multiplicatif gaussien dans les activations d'un réseau de neurones afin de contrôler la quantité d'information transmise par la représentation latente. Formellement, si z désigne l'activation avant *dropout*, la variable perturbée $\tilde{z}$ est définie comme :

$$\tilde{z} = z \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(1, \sigma^2), \tag{4.19}$$

où σ^2 est un paramètre appris, permettant d'adapter le niveau de bruit de manière dépendante des données.

Afin d'adapter cette approche aux tâches de prévision, nous proposons une extension intégrant l'information privilégiée $\mathbf{X}_{priv}$ lors de l'entraînement. Cette information est utilisée pour moduler la variance du bruit injecté de manière hétéroscédastique, ce qui permet un contrôle plus fin de la capacité de la représentation latente.

Cette variante, que nous appelons *Information Dropout Adaptée (IDA)* permet donc d'exploiter les variables supplémentaires disponibles uniquement pendant l'entraînement afin d'améliorer la régularisation et la précision de la prévision.

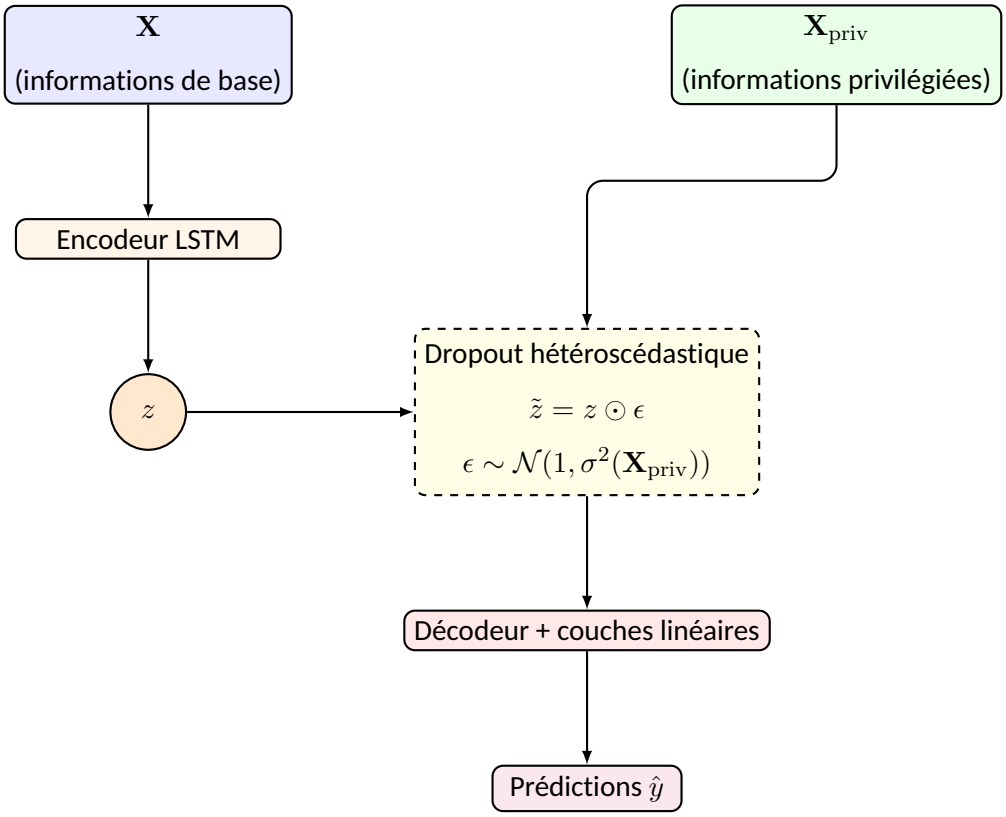


Figure 4.5 – Architecture de *Information Dropout Adaptée (IDA)*.

4.5 Méthode 5 : DairyBLUP

4.5.1 Définition du problème

Entrée : Les données d'entrée comprennent le *pédigrée*, les *performances des individus* sous forme de séries temporelles, ainsi que certaines informations sur les individus (le pays, le nombre de lactations, etc.).

Sortie :

$$\widehat{g} \in \mathbb{R}^{q \times m},$$

où $\widehat{g}$ représente le vecteur des valeurs génétiques estimées (VEE), q le nombre d'individus et m le nombre de caractères génétiques.

Objectif : Estimer les valeurs génétiques de chaque individu en exploitant simultanément les informations de pédigrée, de performance et de traits de conformation.

DairyBLUP représente l'approche méthodologique proposée pour résoudre l'objectif 3.

4.5.2 Modèle DairyBLUP

Les effets fixes considérés incluent les variables suivantes : pays, région, troupeau, nombre de lactations, race et nombre de jours en lait (Days in Milk, DIM). Les effets aléatoires correspondent soit au pédigrée, soit aux données de séquençage génomique.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon}, \tag{4.20}$$

où $\mathbf{Y} \in \mathbb{R}^{n \times m}$ représente la matrice des observations pour m traits, $\mathbf{X} \in \mathbb{R}^{n \times p}$ la matrice des effets fixes regroupant $\mathbf{X}_{\text{pays}}, \mathbf{X}_{\text{region}}, \mathbf{X}_{\text{troupeau}}$, $\mathbf{X}_{\text{lactation}}$, $\mathbf{X}_{\text{race}}$ et $\mathbf{X}_{\text{dim}}$, $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} & \boldsymbol{\nu} & \boldsymbol{\gamma} & \boldsymbol{\beta} & \boldsymbol{\rho} & \boldsymbol{\delta} \end{pmatrix}^\top \in \mathbb{R}^{p \times m}$ la matrice des coefficients associés, $\mathbf{Z} \in \mathbb{R}^{n \times q}$ la matrice d'incidence des effets aléatoires, $\mathbf{g} \in \mathbb{R}^{q \times m}$ la matrice des effets aléatoires, et $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times m}$ la matrice des résidus.

4.5.2.1 Modèle BLUP mono-trait

Soient :

— n : nombre d'observations, n_a : nombre d'animaux.

— $\mathbf{y} \in \mathbb{R}^n$: vecteur des observations.

— $\mathbf{X} \in \mathbb{R}^{n \times p}$: matrice d'incidence des effets fixes (intercept, sexe, région, numéro de lactation, DIM, etc.).

— $\mathbf{Z} \in \mathbb{R}^{n \times n_a}$: matrice d'incidence des effets aléatoires des animaux (1 si l'observation appartient à l'animal, 0 sinon).

— $\mathbf{a} \in \mathbb{R}^{n_a}$: effets génétiques additifs (valeurs d'élevage).

Le modèle linéaire mixte s'écrit :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \epsilon, \qquad \mathbf{a} \sim \mathcal{N}\left(\mathbf{0},\ \sigma_a^2 \mathbf{A}\right), \qquad \epsilon \sim \mathcal{N}\left(\mathbf{0},\ \sigma_\epsilon^2 \mathbf{I}_n\right),$$

où $\mathbf{A}$ est la matrice de parenté additive entre les animaux (Ducrocq, 1990).

Pour la résolution de cette équation, nous pouvons faire appel à la méthode des *moindres carrés ordinaires (OLS)*. Mais Henderson découvre qu'avec une légère modification consistant à ajouter l'inverse de la matrice de variance $Var(a)$, le système d'équations obtenu avec la méthode OLS résout simultanément les solutions des effets fixes $\widehat{\beta}$ et celles du BLUP $\widehat{a}$ (Ducrocq, 1990).

Ces nouvelles équations appelées *équations de Henderson* (MME) sont :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X}/\sigma_\epsilon^2 & \mathbf{X}'\mathbf{Z}/\sigma_\epsilon^2 \\ \mathbf{Z}'\mathbf{X}/\sigma_\epsilon^2 & \mathbf{Z}'\mathbf{Z}/\sigma_\epsilon^2 + \lambda \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y}/\sigma_\epsilon^2 \\ \mathbf{Z}'\mathbf{y}/\sigma_\epsilon^2 \end{bmatrix}, \qquad \lambda = \frac{\sigma_\epsilon^2}{\sigma_a^2}.$$

4.5.2.2 Extension multi-traits : DairyBLUP

Pour m traits, on définit les matrices bloc :

$$\widetilde{\mathbf{X}} = \mathbf{I}_m \otimes \mathbf{X}, \qquad \widetilde{\mathbf{Z}} = \mathbf{I}_m \otimes \mathbf{Z}, \qquad \widetilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix},$$

où $\otimes$ désigne le produit de Kronecker et, pour tout $i = 1, \ldots, m$, $\mathbf{y}_i \in \mathbb{R}^n$ est un vecteur d'observations, de sorte que $\widetilde{\mathbf{y}} \in \mathbb{R}^{mn}$.

Le modèle devient :

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{Z}}\mathbf{a} + \epsilon, \tag{4.21}$$

avec :

$$\mathbf{a} \sim \mathcal{N}(\mathbf{0},\ \mathbf{A} \otimes \mathbf{G})\,, \qquad \epsilon \sim \mathcal{N}(\mathbf{0},\ \mathbf{I}_n \otimes \mathbf{R})\,,$$

où $\mathbf{G}$ est la matrice de covariance génétique entre les traits et $\mathbf{R}$ la matrice de covariance résiduelle.

Les équations mixtes multivariées (MME) s'écrivent alors :

$$\begin{bmatrix} \widetilde{\mathbf{X}}^\top(\mathbf{I}_n \otimes \mathbf{R}^{-1})\widetilde{\mathbf{X}} & \widetilde{\mathbf{X}}^\top(\mathbf{I}_n \otimes \mathbf{R}^{-1})\widetilde{\mathbf{Z}} \\ \widetilde{\mathbf{Z}}^\top(\mathbf{I}_n \otimes \mathbf{R}^{-1})\widetilde{\mathbf{X}} & \widetilde{\mathbf{Z}}^\top(\mathbf{I}_n \otimes \mathbf{R}^{-1})\widetilde{\mathbf{Z}} + \mathbf{A}^{-1} \otimes \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{X}}^\top(\mathbf{I}_n \otimes \mathbf{R}^{-1})\widetilde{\mathbf{y}} \\ \widetilde{\mathbf{Z}}^\top(\mathbf{I}_n \otimes \mathbf{R}^{-1})\widetilde{\mathbf{y}} \end{bmatrix}.$$

4.5.2.3 Matrice de parenté additive

La matrice de parenté $\mathbf{A}$ se construit à partir de la matrice d'incidence de pédigrée $\mathbf{T}$, dans laquelle les coefficients associés aux parents connus prennent la valeur $0,5$, ainsi que de la matrice diagonale $\mathbf{D}$ représentant les variances de ségrégation mendélienne propres à chaque individu. Elle peut être définie comme suit :

$$\mathbf{A} = (\mathbf{I} - \mathbf{T})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{T})^{-\top}.$$

En évaluation génétique à grande échelle, il est rarement nécessaire de calculer explicitement la matrice A. En effet, son inverse A^{-1} peut être obtenu de manière directe, efficace et stable à l'aide de l'algorithme proposé par Henderson. Cet algorithme repose sur une construction récursive exploitant uniquement les informations de filiation (père et mère) pour chaque individu. Ce qui permet d'éviter l'inversion explicite d'une matrice A potentiellement dense (Henderson, 1976; Henderson, 1950).

4.6 Récapitulatif des méthodologies proposées

Les méthodes proposées dans cette étude se focaliseront sur les modèles de prédiction de séries temporelles et d'estimation basés sur des modèles statistiques classiques ou d'apprentissage automatique et profond. Pour rappel, nous avons deux sources de données : Valacta (les données de performances des vaches) et Canadian Dairy Network, CDN (les données génétiques, EBV et traits de conformation). Récemment, les deux se sont fusionnés pour devenir *Lactanet*. Les données de performance sont des données temporelles : séries temporelles (Définition 1.8) recueillies environ dix (10) fois par an à des dates de test. Contrairement aux données de performances, les données génétiques sont collectées une seule fois dans la vie d'un animal.

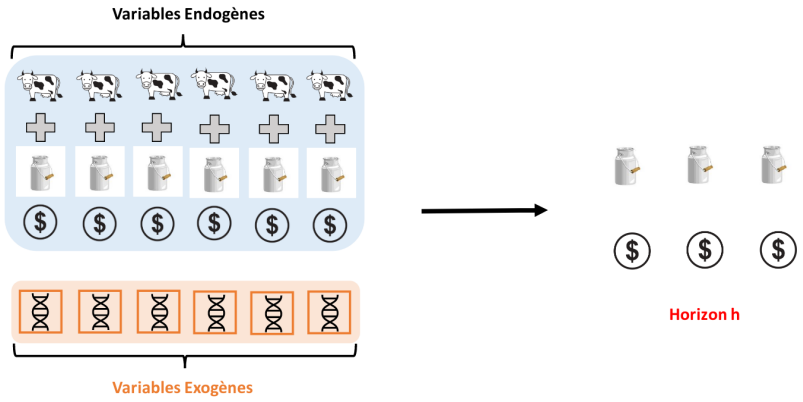


Figure 4.6 – Schéma illustratif du problème.

■ **Méthodologie 1 :** Nous utilisons des modèles statistiques classiques combinés avec la notion de *variables exogènes*.

— **Cheminement des étapes de la méthode :**

La première hypothèse consiste à prendre toutes les variables sans la notion d'exogénéité et la seconde hypothèse sera d'intégrer les caractères génétiques comme des variables exogènes. A la suite comme dernière hypothèse, nous excluons les caractères génétiques pour ne garder que les autres informations d'identification et de production. A noter que nous faisons appel à ces 3 hypothèses à chaque fois que nous introduirons la notion de variables exogènes. Dans la continuité nous ferons appel aux modèles multivariés autorégressifs tels que : VARIMA, VARMA, SARIMA (version multivariée). Par la suite, on intègre le concept des variables exogènes mentionné plus haut dans nos hypothèses. La version multi-output du modèle NARX (modèle autorégressif non linéaire avec variables exogènes) est également utilisé pour résoudre le problème d'évaluation génétique et nous appliquons la stratégie récursive dans un système de Markov. En supposant x_1, x_2 et x_3 des caractères génétiques à estimer ; cette stratégie nous permettra de prédire les $x_i, i \in \{1, 2, 3\}$ successivement x_2 à partir de x_1 , puis x_3 à partir de la prédiction x_2.

■ **Méthodologie 2 :** Dans cette seconde méthode en intégrant la notion de variables exogènes nous effectuons une combinaison de CNN et LSTM (Peixeiro, 2022). A la suite, nous utilisons le NAR avec les réseaux de neurones.

— **Cheminement des étapes de la méthode :**

Pour l'intégration des variables exogènes, on s'inspire de la méthodologie de (Lambert *et al.*, 2018). En d'autres termes, les variables exogènes sont ajoutées au niveau du *dropout* dans toutes les étapes

d'apprentissage (entraînement, test, validation). Ensuite, nous étudions la stabilité de chaque modèle pour leur validation théorique et pratique. Nous faisons une étude comparative avec les modèles de la méthode 1 afin de sélectionner les deux meilleurs modèles en termes de performance et robustesse.

- **Méthodologie 3 :** Pour la troisième approche, nous faisons appel à la notion d'apprentissage avec les informations privilégiées LuPI.
  - **Cheminement des étapes de la méthode :**

  Une première étape consiste à considérer une partie (échantillon) des séries temporelles observées comme des informations privilégiées (Karlsson *et al.*, 2021) et la seconde étape à prendre une partie des variables d'identification (sur les parents et troupeaux) (Tang *et al.*, 2019). Ensuite, nous étudions la mise en place de ses deux approches pour des séries multiples et multi-outputs. Nous utilisons à la suite les modèles de la méthode 2 basés sur les réseaux de neurones (CNN et LSTM) tout en intégrant cette notion d'informations privilégiées.

Table 4.1 – Tableau récapitulatif des méthodes.

Méthodes	**Objectif 1**	**Objectif 2**	**Objectif 3**	**Objectif 4**
Méthode 1 - Agrigen	Oui	Non	Non	Non
Méthode 2 - DairyLuPTS	Oui	Non	Oui	Oui
Méthode 3 - LSTMDropout	Oui	Oui	Oui	Oui
Méthode 4 - Information Dropout Adaptée (IDA)	Oui	Non	Non	Oui
Méthode 5 - DairyBLUP	Non	Oui	Oui	Non

4.7 Métriques d'évaluation

Il existe une variété de critères d'évaluation pour des séries temporelles, mais nous ne nous intéresserons qu'à :

- **Erreur absolue moyenne MAE (mean absolute error)** : En statistique, l'erreur absolue moyenne (EAM) est une mesure des erreurs entre des observations appariées exprimant le même phénomène. Elle est calculée comme la somme des erreurs absolues divisée par la taille de l'échantillon.
- **Erreur quadratique moyenne MSE (mean squared error)** : En statistique, l'erreur quadratique moyenne (ou l'écart quadratique moyen) d'un estimateur (d'une procédure d'estimation d'une quantité non

observée) mesure la moyenne des carrés des erreurs, c'est-à-dire la différence quadratique moyenne entre les valeurs estimées et la valeur réelle.

— **Racine de l'erreur quadratique moyenne RMSE (root mean squared error)** : correspond à la racine carrée de MSE.

— **Pourcentage d'erreur absolu moyen MAPE (mean absolute percentage error)** : Il sert a mesurer la précision (qualité) de la prédiction pour les méthodes de prévision et indépendamment de l'échelle des données. Cela signifie que, le MAPE sera toujours exprimée en pourcentage que l'on travaille avec des valeurs à deux chiffres ou à dix chiffres . Ainsi, MAPE renvoie le pourcentage de l'écart entre les valeurs de prévision et les valeurs observées ou réelles en moyenne, que la prévision soit supérieure ou inférieure aux valeurs observées. Le MAPE est défini par :

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{a_i - p_i}{a_i}|\times 100 \tag{4.22}$$

où a_i est la valeur actuelle (réelle) a un temps i, p_i est la valeur prédite au temps i et n le nombre de prévisions (nombre de valeurs prédites). $|.|$ représente la valeur absolue (Peixeiro, 2022).

— **Q-Q plots** : Le graphique Q-Q est un graphique des quantiles de deux distributions l'un par rapport à l'autre. Dans les prévisions en série temporelle, nous traçons la distribution des résidus sur l'axe des y par rapport à la distribution normale théorique sur l'axe des x. Le graphique Q-Q nous permet d'évaluer la qualité de l'ajustement du modèle. Si la distribution de nos résidus est similaire à une distribution normale, alors une ligne droite apparaît sur $y = x$. Ce qui signifie que notre modèle est bien ajusté, car les résidus sont similaires à un bruit blanc. D'autre part, une ligne courbe apparaît si la distribution de nos résidus est différente d'une distribution normale. Ce qui permet de conclure que notre modèle n'est pas bien ajusté car les résidus ne sont pas similaires à un bruit blanc. En d'autres termes, la distribution des résidus n'est pas proche d'une distribution normale) (Peixeiro, 2022).

— **Corrélogramme** : Dans l'analyse des données, un corrélogramme est un graphique de statistiques de corrélation. Par exemple, dans l'analyse des séries chronologiques, un graphique des auto-corrélations d'un échantillon en fonction des décalages temporels est un autocorrélogramme.

— **Test de Ljung-Box** : Le test de Ljung-Box (appelé aussi test de Box-Pierce modifié, ou simplement test de Box) [(Study, 2023)] est un test statistique qui détermine si les auto-corrélations des erreurs ou des résidus sont non nulles, i.e. s'ils sont similaires au bruit blanc. Le test de Ljung-Box est calcule par :

$$Q(m) = n(n+2)\sum_{j=1}^{n}\frac{r_j^2}{n-j}$$

avec r_j les auto-corrélations accumulées de l'échantillon, et m le décalage temporel.

L'hypothèse nulle H_0 : *les résidus sont indépendamment distribuées et non corrélés* est rejetee si :

$$Q > \chi^2_{1-\alpha,h}$$

avec $\chi^2_{1-\alpha,h}$ la valeur sur le tableau de distribution du Chi-deux χ^2 pour un niveau de signification α et de degrés de liberté h. Le rejet de l'hypothèse nulle signifie que les résidus ne sont pas indépendamment distribués. Par conséquent, il y a auto-corrélation, donc les résidus ne sont pas similaires à un bruit blanc et le modèle ne peut pas être utilisé.

4.8 Données

4.8.1 Données de Lactanet (Valacta & Réseau Laitier Canadien)

Les données utilisées dans ce projet proviennent de *Lactanet*, organisation issue de la fusion de *Valacta* et du *Réseau laitier canadien* (Canadian Dairy Network, *CDN*) (Lactanet Canada, 2026; Frasco. *et al.*, 2020). Elles ont été recueillies entre 2006 et 2017 auprès de 6672 fermes localisées principalement au Québec et en Ontario. Au Canada, plusieurs races de vaches laitières sont présentes, notamment les Holstein, Jersey, Canadienne, Ayrshire et autres (Table 4.5). La race Holstein domine très largement le cheptel national, représentant environ 90% des effectifs laitiers, tandis que les autres races demeurent minoritaires. Notre

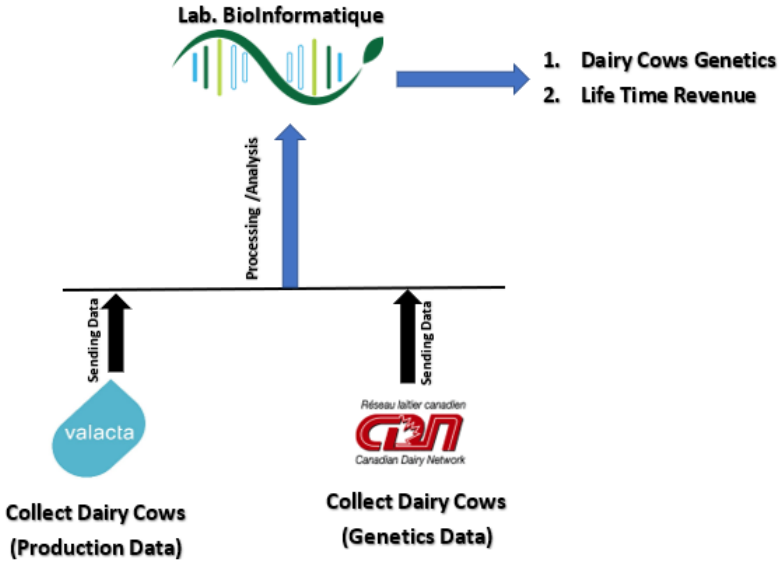


Figure 4.7 – Chaîne d'acquisition des types de données de Lactanet.

base de données comporte *cinq* grands facteurs présents dans la production laitière, à savoir : *la santé*,

l'environnement, *la gestion*, *la production et la génétique* (Frasco. *et al.*, 2020). Pour mener nos travaux de recherche, nous intéressons uniquement à cinq (5) ensembles de données réelles :

- Données **Animaux** : qui nous renseignent sur les informations d'identification des animaux (date de naissance, troupeau, pays, parents, etc.).
- Données **Production/Lactation** : qui contiennent les renseignements sur la production (nombre de lactations, le rendement de lait, le rendement de protéine, le rendement de gras, les différents coûts, nombre de vêlage, etc.).
- Données **Génétique** : qui comporte les valeurs génétiques estimées (EBV) des caractères génétiques des animaux ainsi que des données de séquençages génomiques (SNPs).
- Données **Life Time Revenue** : ce sont des données de suivi (et temporelles) contenant des informations d'identification et de production des animaux.
- Données de **séquençages génomiques** : un fichier de résultats de génotypage SNP Illumina pour des bovins, basé sur la *puce GGP Bovine 100K*. Il contient pour chaque échantillon et chaque SNP : les allèles appelés, leur codage Illumina, un score de qualité (GC Score) et les intensités brutes de fluorescence (X/Y) (Tables 4.12 et 4.13).

Table 4.2 – Résumé des jeux de données utilisés.

Jeu de données	Nombre d'individus	Lignes	Colonnes
Animaux	2 055 532	2 245 954	23
Production (lactations)	1 471 312	35 976 274	24
Génétique (EBV + traits de conformation)	1 035 532	1 (par individu)	42
Revenus à vie (Life Time Revenue)	17 734	100 000	21
Séquençages génomiques	1	16	11

4.8.2 Analyse statistique

4.8.2.1 Données Animaux

L'ensemble de données *Animaux* contient les informations d'identification de $2\,055\,052$ individus (animaux). La taille de *Animaux* est $2\,245\,954$ lignes et 22 colonnes (attributs). Et les données proviennent de plusieurs régions administratives du Québec, incluant 01 – Bas-Saint-Laurent, 02 – Saguenay–Lac-Saint-Jean, 03 – Capitale-Nationale, 04 – Mauricie, 05 – Estrie, 06 – Montréal, 07 – Outaouais, 08 – Abitibi–Témiscamingue, 09 – Côte-Nord, 10 – Nord-du-Québec, 11 – Gaspésie–Îles-de-la-Madeleine, 12 – Chaudière–Appalaches, 13 – Laval, 14 – Lanaudière, 15 – Laurentides, 16 – Montérégie-Est, 17 – Centre-du-Québec et 18 – Montérégie-Ouest.
Les pays de provenance identifiés (Table 4.5) des animaux sont : 840 – États-Unis, AUS – Australie, AUT – Autriche, BEL – Belgique, BGR – Bulgarie, CAN – Canada, CHE – Suisse, CZE – République tchèque, DEU – Allemagne, DNK – Danemark, ESP – Espagne, EST – Estonie, FIN – Finlande, FRA – France, GBR – Royaume-Uni, HUN – Hongrie, IJE – Île de Jersey, IRL – Irlande, ISR – Israël, ITA – Italie, JPN – Japon, NLD – Pays-Bas, NOR – Norvège, NZL – Nouvelle-Zélande, POL – Pologne, ROM – Roumanie, RUS – Russie, SVN – Slovénie, SWE – Suède, USA – États-Unis, YUG – Yougoslavie, ZAF – Afrique du Sud.

La race *Holstein* est la plus représentée dans les données et les animaux proviennent majoritairement du *Canada* (Table 4.5).

4.8.2.2 Données Production/Lactation

Nous avons des renseignements sur les différents cycles de lactation de $1\,471\,312$ animaux provenant de $6\,672$ troupeaux. Ces données sont recueillies à des dates de test, soit environ dix (10) fois par an, ce qui

Table 4.3 – Attributs catégoriels et le pourcentage de données manquantes.

Attributs	**Significations**	**Unique (total)**	**Données Manquantes (%)**
SEX	Sexe	2	0,001291
COU_CD	Pays de Provenance	13	8,078
ANB_CD	Race	26	0
REGION	Région	17	0
DAM_SEX	Sexe Mère	1	14,332
COU_CD_DAM	Pays (Mère)	15	16,579
ANB_CD_DAM	Race Mère	16	16,028
SIRE_SEX	Sexe Père	2	11,849
COU_CD_SIRE	Pays (Père)	20	14,068
ANB_CD_SIRE	Race Père	35	13,541

Table 4.4 – Résumé des attributs et pourcentage de données manquantes.

Attributs	**Significations**	**Données Manquantes (%)**
ANM_ID	Animal ID	0
HRD_ID	Troupeau ID	0
HRD_PRV_CD	Province du Troupeau	0
COMPTR_NO	Numéro de l'ordinateur (d'enregistrement)	0
LHR_CD	Raison de quitter le troupeau	22,96
LHR_CD_2	Raison de quitter le troupeau 2	97,258
REG_ID	Enregistrement ID	13,361
DAM_REG_ID	Enregistrement ID (mère)	16,585
SIRE_REG_ID	Enregistrement ID (père)	14,069

nous fait environ $35\,976\,274$ lignes et 24 variables. La table de Figure 4.6 nous fournit plus d'informations sur l'ensemble de données *Production*.

À noter que la lactation est la production de lait d'une vache. Elle est déclenchée par la naissance d'un veau. Une lactation dure généralement de *dix* mois après la mise bas de la vache, soit en moyenne de 305 jours. Entre deux lactations, il y a une période de *repos* ou de *tarissement* d'environ *deux* avant le prochain vêlage. Dans la table 4.7, seule la variable LEFT_HERD_DATE a $22,96\%$ de valeurs manquantes.

La matrice de corrélation présentée en Figure 4.8 met en évidence plusieurs relations importantes entre les

Table 4.5 – Attributs catégoriels : le nombre de modalités et leurs valeurs uniques (avant pré-traitement).

Attributs	**Significations**	**Unique (total)**	**Valeurs Uniques**
SEX	Sexe	2	['F', 'M']
COU_CD	Pays de Provenance	13	['CAN', 'USA', '840', 'AUS', 'NLD', 'CZE', 'HUN', 'GBR', 'ISR', 'CHN', 'BGR', 'SVN', 'ARG']
ANB_CD	Race	26	['HO', 'JE', 'AY', 'BS', 'CN', 'XX', 'MS', 'DL', 'GU', 'RP', 'NR', 'UU', 'AN', 'MO', 'ZZ', 'SM', 'PS', 'AG', 'BB', 'NM', 'CH', 'SR', 'HP', 'AR', 'FL', 'LM']
REGION	Région	17	[1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18]
DAM_SEX	Sexe Mère	1	['F']
COU_CD_DAM	Pays (Mère)	15	['CAN', 'USA', 'NLD', 'CHE', 'ZAF', 'DEU', 'BEL', 'AUS', 'FRA', 'GBR', '840', 'HUN', 'ITA', 'SVN', 'FIN']
ANB_CD_DAM	Race Mère	16	['HO', nan, 'JE', 'AY', 'BS', 'CN', 'XX', 'GU', 'MS', 'DL', 'ZZ', 'UU', 'AG', 'OH', 'MO', 'SH']
SIRE_SEX	Sexe Père	1	['M']
COU_CD_SIRE	Pays (Père)	20	['CAN', 'USA', 'AUS', 'ITA', 'DEU', 'NLD', 'CZE', 'FRA', 'HUN', 'GBR', 'ESP', 'BEL', '840', 'SWE', 'CHE', 'FIN', 'DNK', 'NOR', 'NZL', 'IRL']
ANB_CD_SIRE	Race Père	35	['HO', 'JE', 'AY', 'BS', 'CN', 'PS', 'GU', 'LM', 'XX', 'BB', 'AN', 'AR', 'MS', 'HP', 'CH', 'SM', 'NR', 'SR', 'ZZ', 'SP', 'MO', 'NM', 'DL', 'HC', 'WW', 'BD', 'PA', 'FL', 'KB', 'UU', 'GV', 'SA', 'IS', 'PI', 'BG']

variables de production, de santé et de performance économique. Les résultats montrent que la production quotidienne de lait (HR_24_MILK) est fortement corrélée à la valeur quotidienne du lait (DAILY_MILK_VALUE, $r = 0.94$) ainsi qu'à la fréquence de traite (MILKING_FQCY, $r = 0.74$). De même, les composantes du lait, à savoir la matière grasse (HR_24_FT) et la protéine (HR_24_PT), présentent une corrélation très élevée ($r = 0.95$), traduisant la stabilité de la composition chimique du lait. L'analyse révèle également que le profit est presque parfaitement corrélé à la valeur quotidienne du lait ($r = 0.99$), ce qui confirme que les revenus laitiers constituent le principal déterminant de la rentabilité. En revanche, certaines corrélations négatives méritent une attention particulière. La production de lait diminue avec les jours en lait (DIM, $r = -0.58$), ce qui se traduit par une baisse progressive du profit au fil de la lactation ($r = -0.49$).

Table 4.6 – Résumé des attributs des données de production : pourcentage de données manquantes et facteur associé

Attributs	**Significations**	**Données Manquantes (%)**	**Facteur**
ANM_ID	Animal ID	0	Animal
HRD_ID	Troupeau ID	0	Animal
BIRTH_DATE	Date de naissance	0	Animal
LACT_NO	Numéro de Lactation	0	Production
MILKNG_FQCY	Milking Frequency	3,133	Production
MILKNG_PTRN	Milking Pattern	0	Production
TEST_DATE	Date au jour du test	0	Production
ANS_CD	Animal status	0	Production
SCC	Somatic cell count / Comptage de cellules somatiques (x 1000)	19,724	Santé
MUN	Milk urea nitrogen / Urée (mg/dL)	58,894	Santé
BHB	Beta hydrobutyrate	81,109	Santé
CCD	Condition Code	0	Santé
ABNRML_STATUS	Abnormal Statut	3,133	Production
LACTOSE	Lactose	39,988	Production
DIM	Days in Milk / Jours en terme de lait	0,00166	Production
HR_24_MILK	Quantité de lait le jour du test	3,133	Production
HR_24_PT	Quantité de protéine le jour du test	3,394	Production
HR_24_FT	Quantité de gras le jour de test	3,394	Production
DAILY_MILK_VALUE	Revenu journalier du lait	3,501	Production
CUMUL_MILK_VALUE	Revenu cumulé du lait	16,718	Production
DAILY_FEED_COST	Coût journalier d'alimentation	61,995	Production
CUMUL_FEED_COST	Coût cumulé de l'alimentation	62,858	Production
MaB	Months after Birth	0	Production
MaBint	Months after Birth (in integer)	0	Production

Par ailleurs, le coût alimentaire cumulé (CUMMIL_FEED_COST) est modérément et négativement corrélé au profit ($r = -0.41$), soulignant son impact sur la marge économique. Enfin, des variables telles que le compte somatique cellulaire (SCC), l'urée du lait (MUN) et les états anormaux (ABNRML_STATUS) présentent de faibles corrélations avec la production et le profit, suggérant qu'elles apportent une information complémentaire susceptible d'améliorer la modélisation de la variabilité individuelle. Ces observations

Table 4.7 – Variables de dates.

Attributs	Significations	Minimum	Maximum	Unique (total)
BIRTH_DATE	Date de Naissance	21/11/1975	06/03/2018	9485
ENTER_HERD_DATE	Date d'entrée dans le troupeau	23/03/1987	31/12/2017	8475
LEFT_HERD_DATE	Date de sortie du troupeau	06/02/2006	06/04/2018	4481

suggèrent que l'inclusion de variables hautement corrélées, telles que HR_24_MILK, DAILY_MILK_VALUE et profit, dans un même modèle pourrait induire des problèmes de multicolinéarité. L'utilisation de méthodes de régularisation (Ridge, Lasso) ou de sélection de variables apparaît donc nécessaire afin d'améliorer la robustesse de l'estimation et l'interprétation des effets.

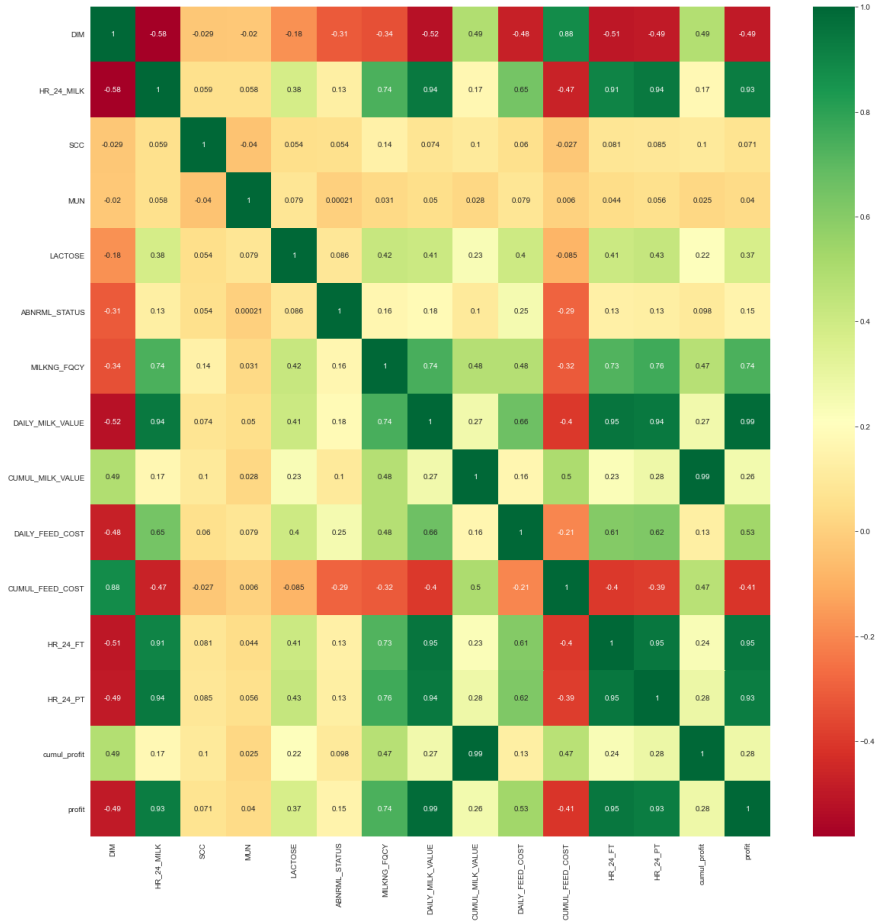
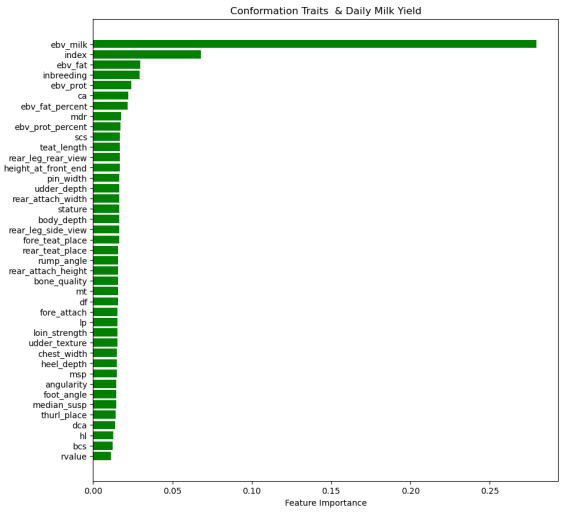
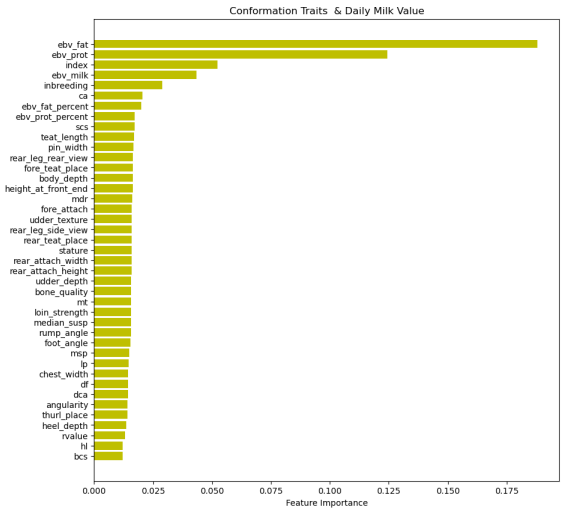


Figure 4.8 – Matrice de corrélation du jeu de données *Production*.

Les histogrammes des quantités quotidiennes de lait, de matière grasse et de protéines (Figure 4.10) montrent



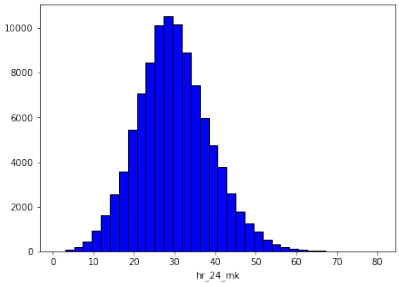
(a) Variable : la quantité journalière du lait.



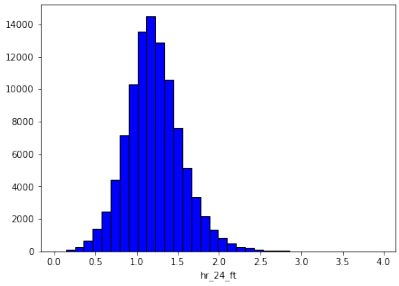
(b) Variable : la valeur journalière du lait.

Figure 4.9 – Analyse comparative des importances de variables génétiques (EBV et traits de conformation) par rapport à la quantité et la valeur journalières du lait en utilisant le modèle de régression de forêt aléatoire (Random Forest regressor).

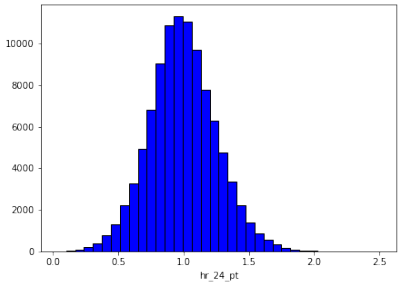
des distributions unimodales et relativement symétriques. Les variables associées au gras et à la protéine (Figures 4.10b et 4.10c) présentent des formes quasi-gaussiennes, tandis que la distribution du lait (Figure 4.10a) présente une légère asymétrie positive due à quelques observations de très forte production (valeurs aberrantes). Dans l'ensemble, ces variables peuvent être approximées par une loi normale, ce qui justifie l'utilisation de modèles linéaires pour leur modélisation.



(a) Lait.



(b) Gras.



(c) Protéine.

Figure 4.10 – Histogrammes de la production journalière du lait, des matières grasses et des protéines.

4.8.2.3 Données Life Time Revenue

Ce jeu de données comprend un total de $100\,000$ enregistrements, correspondant à $17\,734$ individus, et décrit par 21 variables (voir Table 4.8), dont trois variables cibles : lifetime milk, lifetime protein et lifetime fat.

La matrice de corrélation illustrée en Figure 4.11 met en évidence plusieurs relations biologiquement cohérentes entre les variables de production, les caractéristiques individuelles des vaches et leurs performances antérieures. Les productions journalières de lait (hr_24_mk), de matière grasse (hr_24_ft) et de protéines (hr_24_pt) présentent des corrélations positives modérées entre elles. De même, les rendements cumulés sur la lactation (lt_mk, lt_ft, lt_pt) sont positivement associés aux productions journalières correspondantes, confirmant la relation attendue entre performance quotidienne et performance globale. Les jours en lait (dim) sont négativement corrélés avec la production de lait et ses composantes (gras et protéine). Les performances de la lactation précédente (prev_tot_lact_date_yld_*) présentent des corrélations positives avec les performances actuelles, indiquant une certaine régularité inter-lactations, possiblement liée à la génétique ou au management du troupeau. Enfin, le score cellulaire linéaire (scc_linear_score) présente une corrélation légèrement négative avec la production, ce qui reflète l'impact défavorable des troubles de la santé mammaire sur les performances laitières. Comme remarqué précédemment, la fréquence de traite (milking_freq) est positivement associée à la production.

4.8.2.4 Données génétiques

La première partie des données comporte les caractéristiques génétiques de $1\,035\,532$ individus qui sont fournies à travers 42 attributs dont les mensurations corporelles et les EBV (Définition. 3.1).
La seconde partie du jeu de données regroupe les résultats de séquençage génomique de $100\,000$ individus, comprenant pour chacun $92\,259$ SNPs décrits par 11 variables (SNP Name, Sample ID, Allele1/2 - Forward/Top, Allele1/2 - AB, GC Score, X, Y) (Table 4.13). Toutefois, dans le cadre de cette étude, nous ne disposons que d'un sous-échantillon extrêmement restreint, correspondant aux données d'un seul individu.

— **Données de traits de conformation**

Le jeu de données, qui inclut à la fois les traits de conformation et les valeurs d'élevage estimées (EBV), est exploré de manière descriptive. Une analyse statistique présentant la moyenne, le minimum et le maximum

Table 4.8 – Attributs de *Life Time Revenue*.

Attributs	**Significations**
lact_no	Numéro de Lactation
age_at_calving	Âge au vêlage
dim	Days in Milk / Jours en terme de lait
milkng_fqcy	Fréquence des traits
test_cnt	Compte des Tests
scc_linear_score	Pointage linéaire des cellules somatiques
hr_24_mk	Quantité de lait le jour du test
lact_date_mk	-
hr_24_ft	Quantité de gras le jour du test
lact_date_ft	-
hr_24_pt	Quantité de protéine le jour du test
lact_date_pt	-
bhb	Beta hydrobutyrate
prev_tot_lact_date_yld_mk	Rendement total de lait cumulé
prev_tot_lact_date_yld_ft	Rendement total de gras cumulé
prev_tot_lact_date_yld_pt	Rendement total de protéine cumulé
lt_mk	Quantité de lait cumulé
lt_ft	Quantité de gras cumulé
lt_pt	Quantité de protéine cumulé
age_d	Âge en jours
ano_anm	Animal ID

de ces variables est réalisée (Tables 4.9, 4.10 et 4.11). Aucune valeur manquante n'a été identifiée dans cet ensemble de données. Ces informations permettent de caractériser l'intervalle de variation de chaque attribut ainsi que leur tendance centrale.

— **Données de séquençages génomiques**

Les données de génotypage ont été obtenues à l'aide du logiciel GenomeStudio Genotyping Module (GSGT, version 2.0.4, Illumina). Le traitement a été réalisé le 18 novembre 2022 à partir de la puce *GGP Bovine 100K (Illumina)*, qui contient environ $100\,000$ polymorphismes nucléotidiques simples (SNP). Au total, 52 échan-

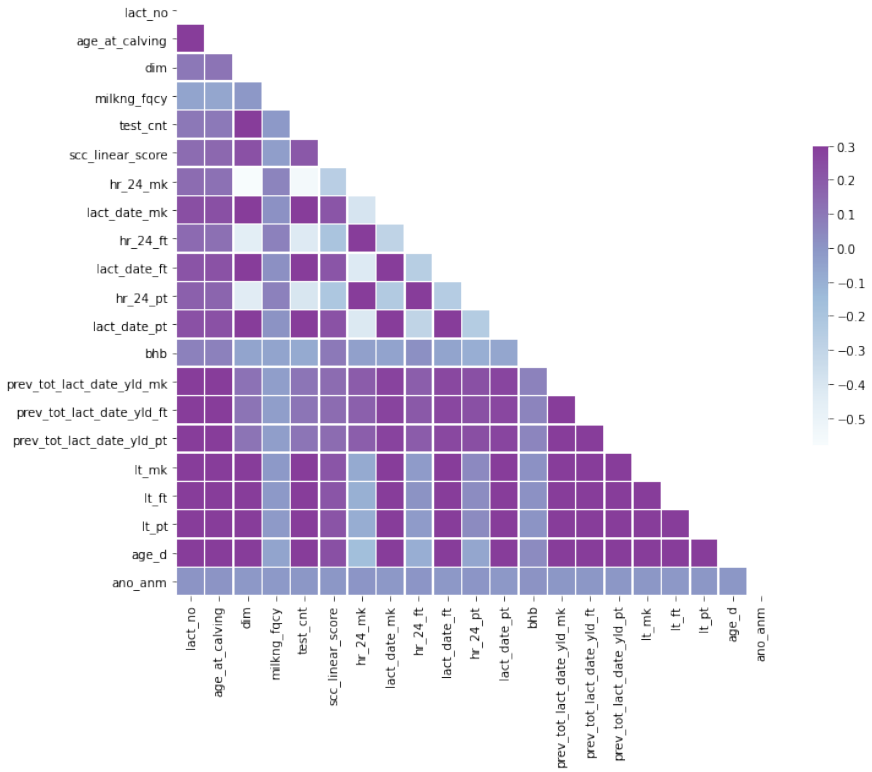


Figure 4.11 – Matrice de corrélation des données Lifetime Revenue.

Table 4.9 – EBV et autres attributs.

Attributes	Mean	Minimum	Maximum
Inbreeding	6.116	0	41.2
Rvalue	13.028	0	26
Somatic Cell Score	99.038	77	118
Estimated Breeding Values EBVs			
EBV Milk	-295.56	-3580	3370
EBV Protein	-10.822	-105	94
EBV Fat	-14.077	-150	137
EBV Protein Percent	-0.007	-0.67	0.63
EBV Fat Percent	-0.0235	-1.4	1.45

tillons bovins ont été analysés. La puce contenait $99\,229$ SNP, dont $95\,256$ ont été effectivement appelés après contrôle qualité. Un échantillon des données ainsi qu'une description détaillée des attributs sont fournis dans les Tables 4.12 et 4.13. Les valeurs des X et Y peuvent servir pour vérifier les liens de parenté

Table 4.10 – Traits de conformation (1/2).

Attributes	Mean	Minimum	Maximum
Front End & Body Capacity			
Stature	-2.375	-30	23
Height at Front End	-0.517	-21	20
Body Condition BCS	101.072	85	121
MR	99.19	85	114
Chest Width	0.357	-17	21
Body Depth	1.257	-17	19
Loin Strength	-0.027	-21	18
Rump			
Rump Angle	0.454	-22	22
Pin Width	-1.289	-23	19
Angularity	-2.807	-29	17
Feet & Legs			
Hock Lesion HL	99.206	87	114
Foot Angle	-1.603	-22	18
Heel Depth	-2.222	-19	20
Bone Quality	-1.091	-23	17
Rear Leg Side View	-0.284	-17	18
Rear Leg Rear View	-1.346	-19	21
Thurl Place Area over the Pelvis	-0.076	-16	16

Table 4.11 – Traits de conformation (2/2).

Attributes	Mean	Minimum	Maximum
Udder (Mammary)			
Udder Depth	-2.487	-21	20
Udder Texture	-3.646	-25	17
Teat Length	1.258	-18	21
Fore Attach	-2.921	-20	16
Fore Teat Place	-2.545	-26	18
Rear Teat Place	-2.173	-22	20
Median Susp	-2.852	-22	16
Rear Attach Height	-2.929	-25	16
Rear Attach Width	-3.179	-25	16
Dairy Strength			
Lactase Persistence LP	98.607	85	113
Minimum Support Price MSP	99.483	83	115
MT	99.495	84	114
Daugther fertility	100.689	85	116
Calving Ability	98.577	78	114
Daily Cow Average	99.474	84	115
Multi-Drug Resistant	98.072	81	113

ou même comme mesure de validation.

Table 4.12 – Échantillon de données SNP.

SNP Name	Sample ID	Allele1 - Forward	Allele2 - Forward	Allele1 - Top	Allele2 - Top	Allele1 - AB	Allele2 - AB	GC Score	X	Y
1_41768691	1428	G	G	C	C	B	B	0,5985	0,040	0,255
1_41997985	1428	T	T	A	A	A	A	0,3945	0,813	0,077
10-104012831-C-G-rs442869917	1428	C	C	C	C	A	A	0,5420	1,428	0,009

Table 4.13 – Description des colonnes du fichier de génotypage SNP.

Variables	**Significations**
SNPs name	Nom du SNP 10-104012831-C-G-rs442869917 : SNP, identifiant du variant, locus (chromosome et position)
Sample ID	Identifiant de l'animal
Allele1 - Forward	Intensité du signal du premier allèle (souvent l'allèle de référence) pour le brin inférieur de l'ADN
Allele1 - Top	Intensité du signal du premier allèle (souvent l'allèle de référence) pour le brin supérieur de l'ADN
Allele2 - Forward	Intensité du signal du deuxième allèle (souvent l'allèle variant) pour le brin inférieur de l'ADN
Allele2 - Top	Intensité du signal du deuxième allèle (souvent l'allèle variant) pour le brin supérieur de l'ADN
Allele1 - AB	Appels de génotypes représentés par deux lettres (ex. A-A, A-C, G-T)
Allele2 - AB	Allèle dominant. Les lettres indiquent les deux allèles présents (ex. A-A homozygote, A-C hétérozygote, G-T hétérozygote)
GenCall Score	Score qualité du génotypage. Mesures supplémentaires : score GC ou fréquence de l'allèle B (BAF)
X	Chromosome X permet de déterminer homo-/hétérozygotie. Intensité du signal pour le chromosome X. Utile chez les femelles (XX).
Y	Chromosome Y. Intensité du signal pour le chromosome Y. Pertinent pour les échantillons masculins.

4.8.3 Prétraitement

À cette étape, nous réalisons une analyse exploratoire des différents jeux de données, comprenant le calcul de statistiques descriptives, l'évaluation des données manquantes ainsi que la visualisation de certains attributs.

4.8.3.1 Prétraitement : Données de Lactanet

Dans cette étude, nous sélectionnons les animaux présentant au moins trois (3) cycles de lactation dans le jeu de données de production laitière. Les informations relatives à ces animaux ont ensuite été identifiées dans les jeux de données *animal* et *génétique*, sans procéder à leur fusion à ce stade. En ce qui concerne le traitement des données manquantes, nous avons supprimé une partie des colonnes présentant plus de $50\ \%$ de données manquantes, tout en tenant compte de l'importance biologique et statistique des variables concernées. Les autres données manquantes ont été traitées comme suit :

- **Valeur laitière quotidienne :**
  1. Les valeurs manquantes ont été imputées par la *moyenne de groupe*, c'est-à-dire la moyenne des individus ayant débuté leur cycle de lactation à une période similaire, en tenant compte du mois de production (afin de respecter l'effet du cycle).
  2. Les valeurs manquantes correspondant à une **période de tarissement** ont été remplacées par zéro (0).
- **Coût alimentaire quotidien :** imputation par la moyenne ou la mode selon la distribution des données.
- **BHB, valeur laitière cumulée (cumul milk value), coût alimentaire cumulé :** variables supprimées en raison d'un taux de données manquantes trop élevé ou d'une redondance avec d'autres mesures.

Ces étapes de préparation visent à *garantir la qualité et la cohérence du jeu de données*, afin de disposer d'une base fiable pour les analyses statistiques ultérieures et l'entraînement de modèles prédictifs (par exemple, GBLUP ou approches d'apprentissage profond).

4.8.3.2 Prétraitement : Données de séquences génomiques

Une première étape consiste à *filtrer les données* en éliminant les lignes ne présentant pas une structure complète *locus–SNP–variant* dans la colonne `SNP Name`. Par la suite, les lignes associées à un *GenCall Score* faible sont également supprimées afin de garantir la qualité des génotypes retenus. Les informations contenues dans `SNP Name` sont ensuite *décomposées* en colonnes distinctes correspondant aux loci, aux SNP et aux variants. Enfin, une nouvelle colonne est créée pour indiquer si chaque SNP est *homozygote* ou *hétérozygote*, sur la base des variables `Allele1-AB`, `Allele2-AB`, `X` et `Y`. Nous synthétisons l'ensemble de ces étapes sous forme de schéma présenté à la Figure 4.13.

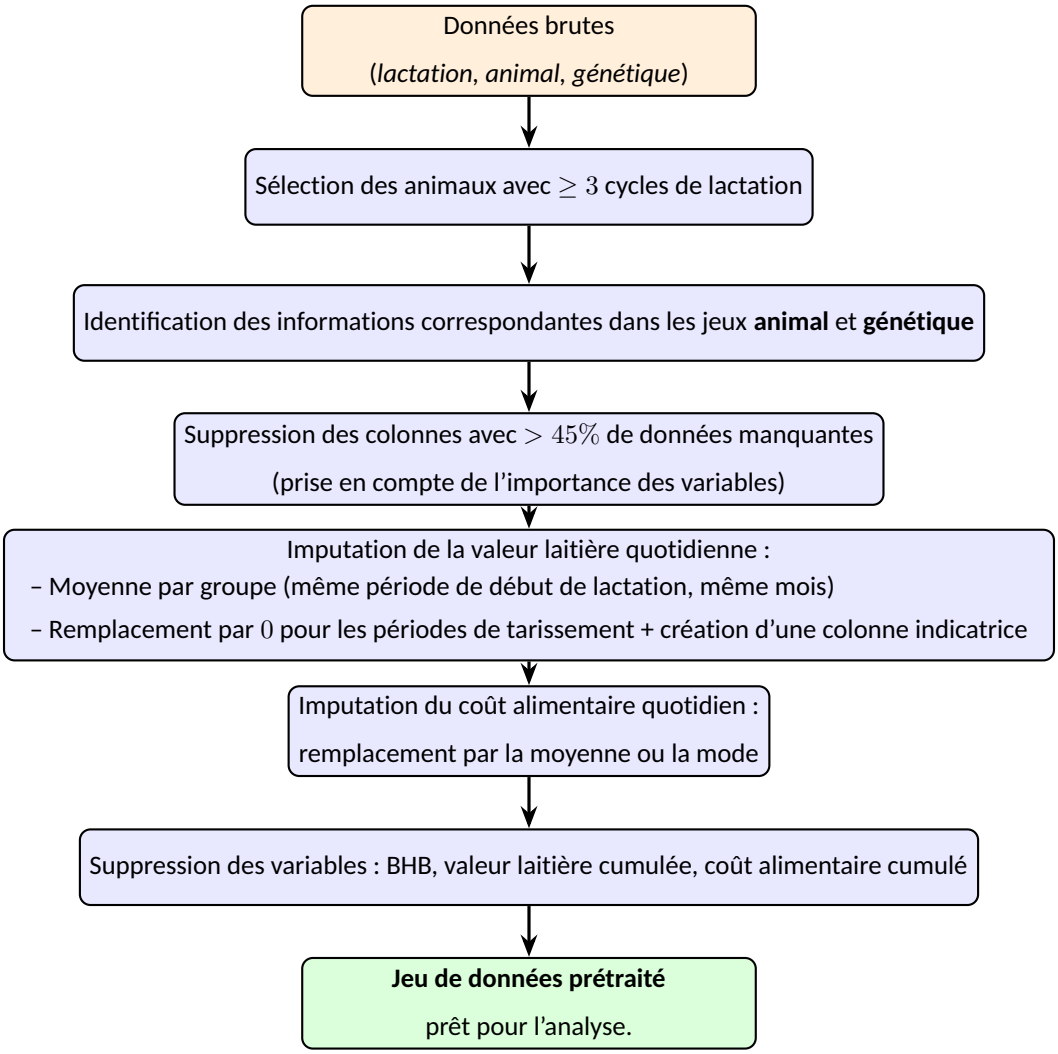


Figure 4.12 – Pipeline de prétraitement des données de lactation : sélection des animaux, filtrage des variables, imputation des données manquantes et production d'un jeu de données propre pour les analyses.

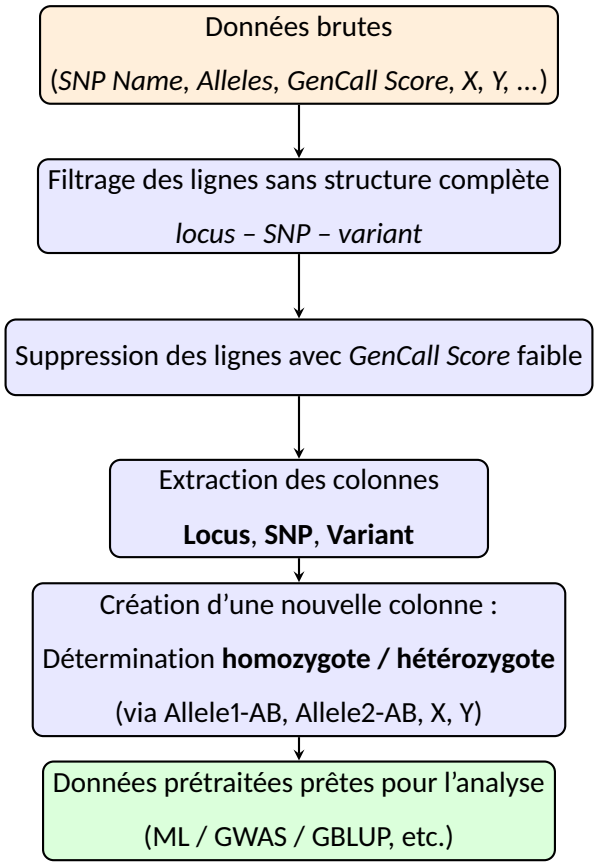


Figure 4.13 – Pipeline de prétraitement des données génomiques : filtrage, extraction des composantes de `SNP Name`, et identification de l'état zygotique.

CHAPITRE 5
ESTIMATION DE L'EFFET DES MARQUEURS GÉNÉTIQUES SUR LA RENTABILITÉ

Dans le chapitre précédent, nous avons présenté une introduction aux méthodes développées, ainsi qu'une description accompagnée d'une analyse statistique des données. Le présent chapitre a pour objectif d'analyser, à travers des expérimentations, l'effet des marqueurs génétiques sur la rentabilité des individus.

5.1 Définition du problème

Notre problème peut être défini comme :

- **Entrée** : les données temporelles contenant les performances de production des vaches et les marqueurs génétiques (traits de conformation et valeurs génétiques estimées VEE).
- **Sortie** : les rendements de lait (DMY), de protéines (DPY) et de gras (DFY), ou les valeurs journalières du lait (DMV) ou le profit (Définition 1.1).
- **Objectif** : étudier l'effet de la génétique sur les composantes du lait (lait, protéines, gras) et la valeur journalière du lait.

Nous allons mener deux expérimentations en utilisant les méthodes *AgriGen* (Section 4.1) et *LSTMDropout* (Section 4.3) pour prédire les rendements de lait, de protéine et de gras en sachant les caractères génétiques, les traits de conformation (Définition 1.3) et les VEE (Définition 3.1).

5.2 Données

Nos expérimentations reposent sur des jeux de données de production laitière et de génétiques (Figure 1.3 et Table 5.1) fournis par Lactanet (Table 4.2) (Frasco. *et al.*, 2020). Notre étude se concentre spécifiquement sur les vaches ayant complété au moins trois cycles de lactation (Définition 1.2 et Figure 5.1). Notre jeu de données final contient $11\,000$ individus, soit à $100\,000$ lignes et 58 variables (Tables 4.6 et 5.1). Ces données sont des *séries temporelles* (Définition 1.8). Nous avons procédé à la fusion des différents jeux de données sur la base des identifiants animaux (ANM_ID et HRD_ID). Les données ont été partitionnées en trois sous-ensembles : 60% pour l'entraînement, 20% pour le test et 20% pour la validation.

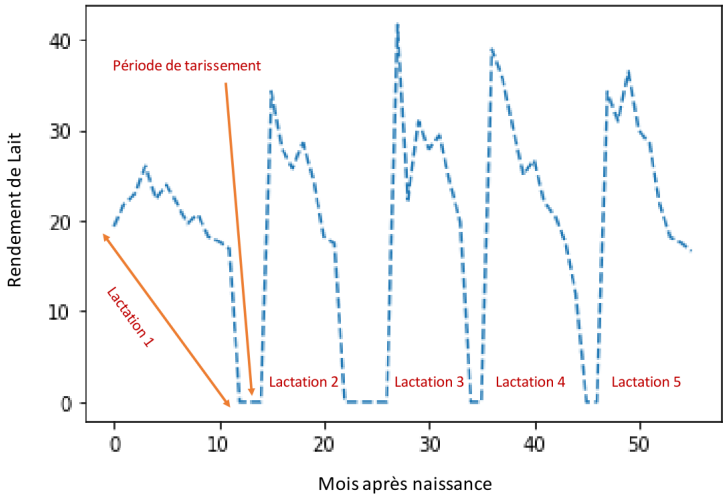


Figure 5.1 – Différents cycles de lactation de l'animal 140.

Table 5.1 – Répartition des traits de conformation.

Names	Conformation Traits
Estimated Breeding Values EBV	EBV Milk ; EBV Protein ; EBV Fat ; EBV Protein Percent ; EBV Fat Percent
Body Capacity	Stature ; Height at Front End ; Body Condition BCS ; MR ; Chest Width ; Body Depth ; Loin Strength
Rump	Rump Angle ; Pin Width ; Angularity
Feet & Legs	Hock Lesion HL ; Foot Angle ; Heel Depth ; Bone Quality ; Rear Leg Side View ; Rear Leg Rear View ; Thurl Place Area over the Pelvis
Udder (Mammary)	Udder Depth ; Udder Texture ; Teat Length ; Fore Attach ; Fore Teat Place ; Rear Teat Place ; Median Susp ; Rear Attach Height ; Rear Attach Width
Dairy Strength	Lactase Persistence LP ; Minimum Support Price MSP ; MT ; Daugther fertility ; Calving Ability ; Daily Cow Average ; Multi-Drug Resistant
Others	Inbreeding ; Rvalue ; Somatic Cell Score (SCS)

5.3 Expérimentation 1 : AgriGen

5.3.1 Implémentation

Pour l'implémentation d'AgriGen (voir Section 4.1 et Figure 4.1), nous mobilisons plusieurs approches et outils statistiques afin de traiter, analyser et modéliser les données :

- Réduction de dimensionnalité : utilisation de l'analyse en composantes principales (PCA) pour extraire les composantes les plus pertinentes.
- Clustering : application de méthodes non supervisées telles que K-Means et GMM pour identifier des groupes homogènes de données.
- Prédiction et classification : recours au modèle VARMAX (Vector Autoregression Moving-Average with Exogenous Regressors) pour modéliser les relations temporelles et effectuer les prédictions.
- Analyse de corrélation : construction d'une matrice de corrélation pour évaluer les dépendances entre les variables.

5.3.2 Résultats

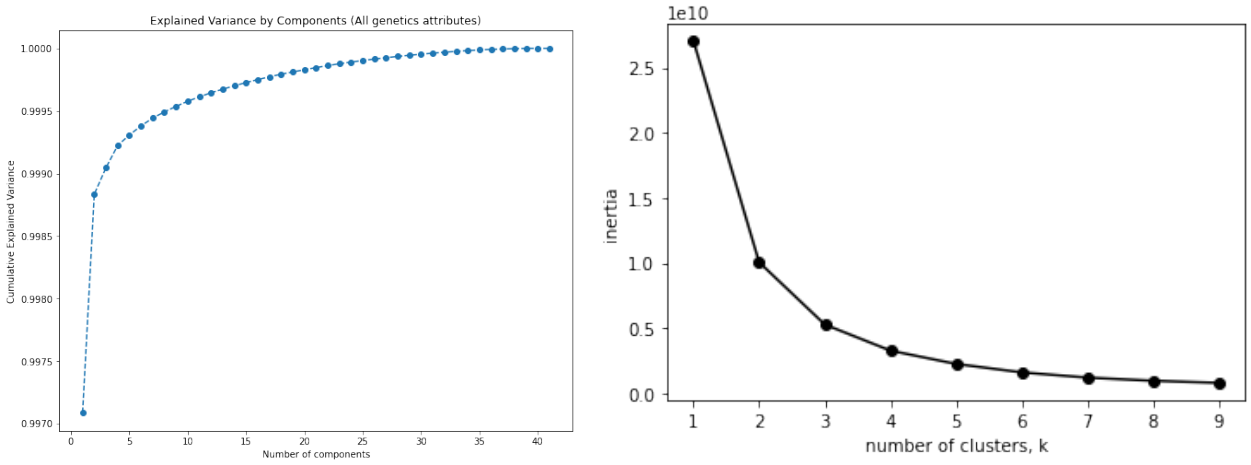
5.3.2.1 Réduction de dimensionnalité

Avant de faire le clustering, nous avons fait une réduction de dimensionnalité à l'aide de l'analyse des composantes principales PCA (Figure 5.2a) sur les données génétiques. Ainsi, nous allons travailler avec deux composantes principales pour cette étape de clustering.

5.3.2.2 Classification non supervisée (Clustering)

Deux méthodes de clustering, K-means et GMM, sont appliquées. Une étape intermédiaire et importante, avant leur application, est la détermination du nombre de clusters. La figure 5.2b ainsi qu'une étude des silhouettes nous ont permis de partir sur deux (2) clusters.

La prochaine étape consiste à construire un modèle de prévision (forecasting) avec comme variables cibles : le profit, le revenu de lait journalier ainsi que le coût journalier.



(a) PCA sur les données génétiques

(b) Choix du nombre de clusters

Figure 5.2 – Analyse exploratoire des données génétiques : (a) projection PCA et (b) détermination du nombre optimal de clusters.

5.3.2.3 Time Series Forecasting

Cette étape nous a permis de construire notre modèle de prédiction qui nous sera utile au moment opportun. Après avoir appliqué plusieurs méthodes de séries chronologiques ARIMA (Autoregressive Integrated Moving Average), VARMA (Vector Autoregressive Moving Average) et VARMAX (Vector Autoregression Moving-Average with Exogenous Regressors), dont chacune représente une généralisation du modèle ARMA (Autoregressive Moving Average), nous avons retenu le modèle VARMAX(p, q) ayant la plus petite erreur RMSE $= 4.242$ (Table 5.2). Sur la Figure 5.3, nous pouvons déduire l'ordre de notre modèle grâce aux fonctions d'auto-corrélation ACF et de la partielle auto-corrélation PACF : $p = 1$ et $q = 1$.

Table 5.2 – Métriques des modèles statistiques de séries temporelles.

Métriques	*AIC*	*BIC*	*RMSE*
ARIMA	91782.714	91797.704	7.694
VARMA	104899.597	105101.958	–
VARMAX	36653.269	37266.511	4.242

En plus des métriques AIC (Akaike information criterion) et BIC (Bayesian information criterion) de la table 5.2, nous avons également calculé le *MAPE* (mean absolute percentage error) de *VARMAX*($p = 1, q = 1$), le mo-

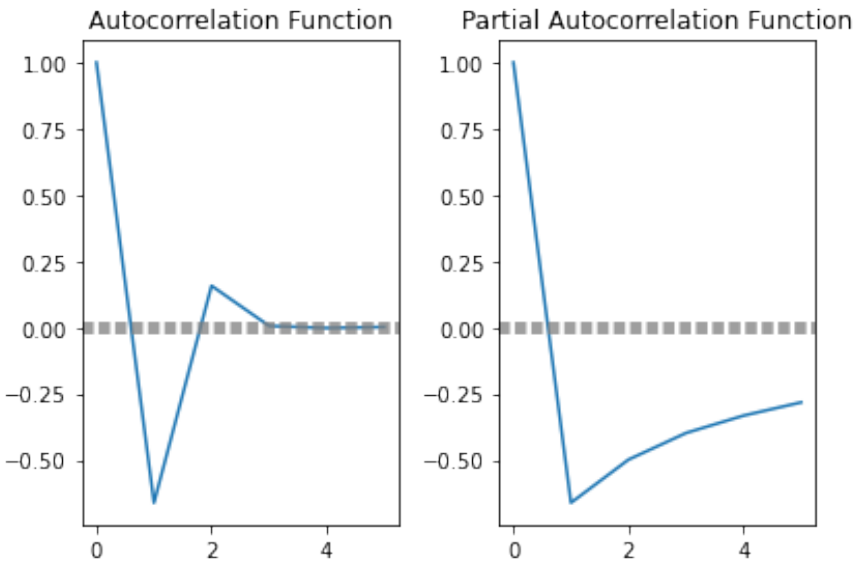


Figure 5.3 – Fonctions d'auto-corrélation ACF et de partielle auto-corrélation PACF.

dèle final : -1.938.

Passons aux dernières étapes de notre architecture qui consistent à l'identification des variables corrélées négativement au profit.

5.3.2.4 Analyse de corrélation

La matrice de corrélation présentée en Figure 5.4 met en évidence plusieurs relations importantes entre les variables de production, de santé et de performance économique. Plusieurs variables ont un impact positif sur le profit (Définition 1.1) telles que : la valeur journalière du lait (≈ 0.99), les rendements journaliers de lait, de protéine et de gras (≈ 0.93), la fréquence de traite (0.74). Les attributs de santé, le compte somatique cellulaire (SCC) et l'urée du lait (MUN), sont faiblement corrélés avec le profit. Certaines corrélations négatives méritent une attention particulière, comme le profit avec les jours en lait (DIM, $r = -0.49$). Ce qui se traduit par une baisse progressive du profit au fil de la lactation. Par ailleurs, le coût alimentaire cumulé (CUMMIL_FEED_COST) est modérément et négativement corrélé au profit ($r = -0.41$), soulignant son impact sur la marge économique.

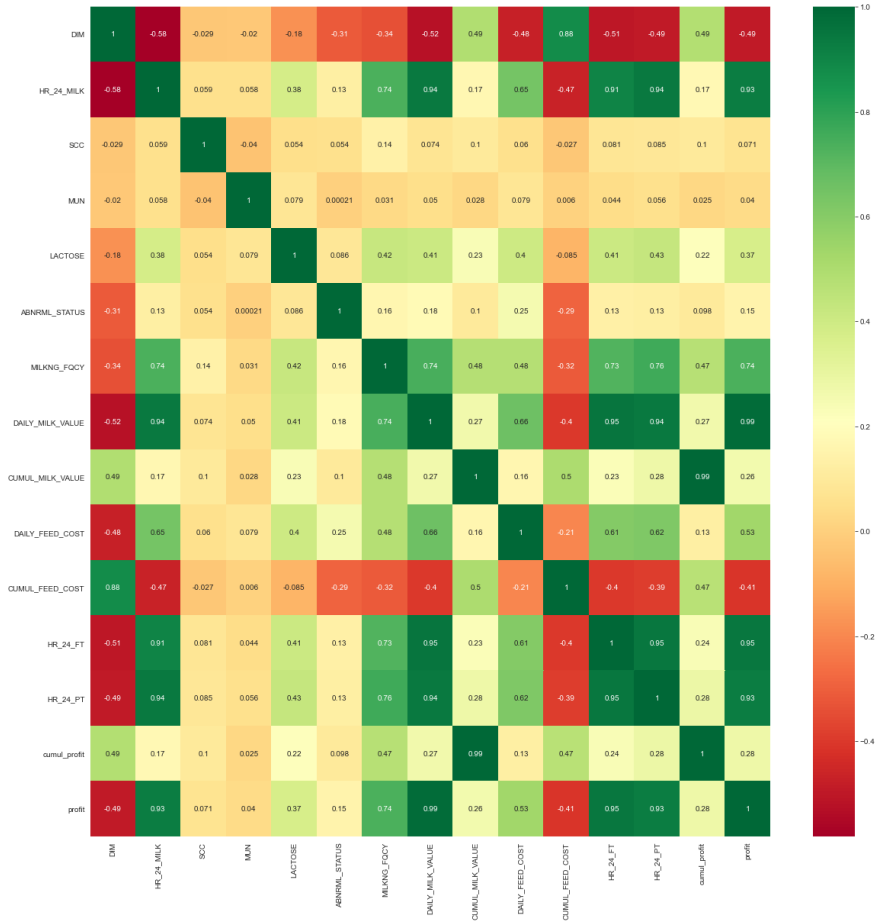


Figure 5.4 – Matrice de corrélation du jeu de données *Production*.

5.3.2.5 Identification des attributs

Après avoir analysé les corrélations entre les variables et le profit, nous avons récupéré celles qui sont négativement corrélées avec le profit. Une classification du profit a été menée en utilisant 2 seuils (Figure 5.5) : 0 (Figure 5.6a) et la moyenne du profit de toutes les vaches (Figure 5.6b).

On observe sur la figure de calibration (Figure 5.7) la robustesse de la corrélation linéaire entre les probabilités prédites et observées du profit.

5.3.3 Discussion

L'étape de modélisation d'*AgriGen* a permis de construire un modèle prédictif robuste, constituant une base solide pour l'évaluation intégrée de la rentabilité laitière et de l'impact génétique. La classification non

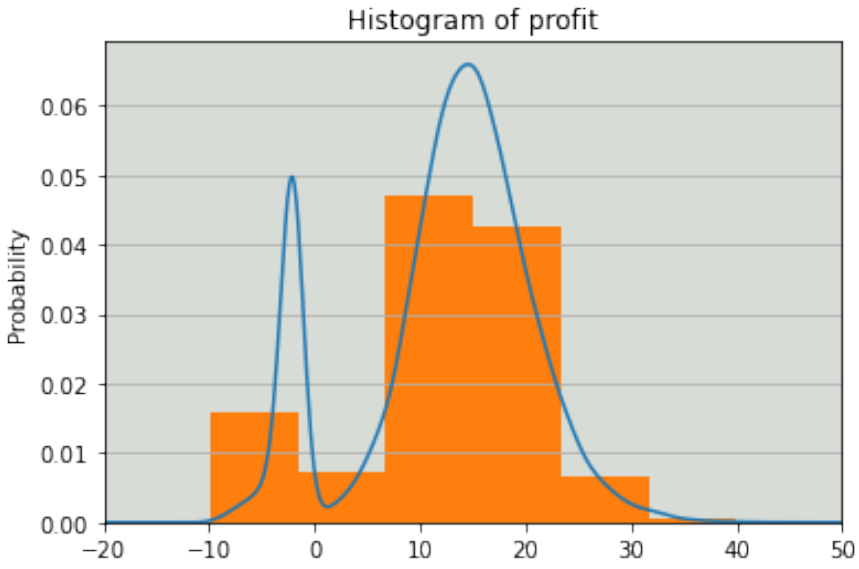
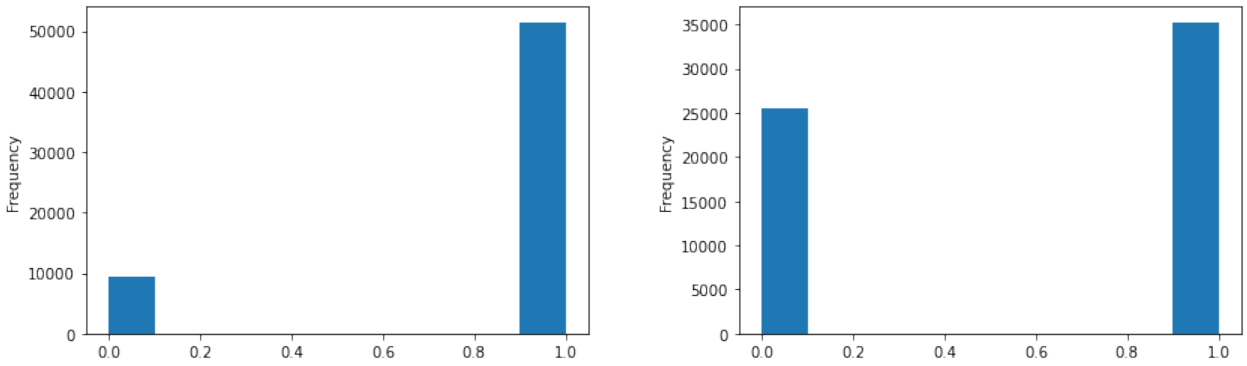


Figure 5.5 – Distribution du profit.



(a) Seuil 0 : 0 = faible profit, 1 = meilleur profit.

(b) Seuil "moyenne" : 0 = faible profit, 1 = meilleur profit.

Figure 5.6 – Distribution des classes de profit selon deux seuils différents.

supervisée avec K-means et GMM nous a permis de regrouper nos individus en deux classes. Après comparaison de plusieurs approches de modélisation des séries temporelles, ARIMA (Autoregressive integrated moving average), VARMA (Vector Autoregressive Moving Average) et VARMAX (Vector autoregression moving average with exogenous regressors), toutes généralisations du modèle ARMA, le modèle VARMAX($p = 1, q = 1$) a été retenu pour ses performances supérieures, avec une erreur minimale (RMSE $= 4.242$, Table 5.2). Enfin, la figure de calibration (Figure 5.7) montre une forte cohérence entre les valeurs prédites et observées, confirmant la robustesse et la validité du modèle VARMAX.

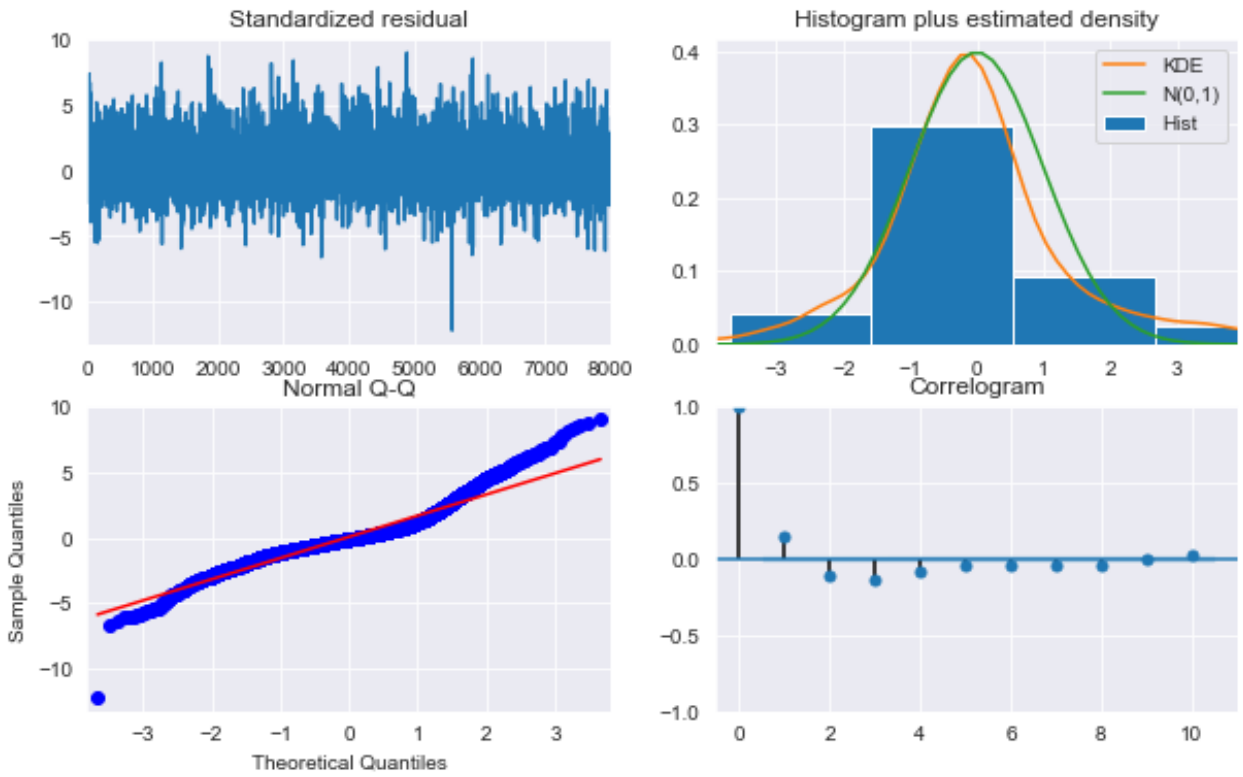


Figure 5.7 – Résultats du VARMAX : Profit.

L'analyse de la matrice de corrélation (Figure 5.4) met en évidence des relations à la fois biologiquement pertinentes et économiquement déterminantes. La valeur journalière du lait présente une corrélation quasi parfaite avec le profit ($r \approx 0.99$), confirmant son rôle central dans la détermination du profit (Définition 1.1). Les rendements journaliers en lait, protéines et matière grasse (environ $r \approx 0.93$) renforcent cette observation et soulignent l'importance d'une optimisation simultanée de l'ensemble des composantes du lait. De plus, la fréquence de traite ($r = 0.74$) est positivement associée au profit, ce qui corrobore l'idée qu'une gestion adéquate de la fréquence de traite peut contribuer à maximiser la production et, par conséquent, la rentabilité.

À l'inverse, certaines variables présentent des corrélations négatives qui méritent une attention particulière. Les jours en lait (DIM) affichent une corrélation négative modérée avec le profit ($r = -0.49$), traduisant la baisse progressive de la production et de l'efficacité économique au fur et à mesure que la lactation progresse (Figure 1.1). Le coût alimentaire cumulé (CUMMIL_FEED_COST) présente également une corrélation négative ($r = -0.41$), mettant en lumière le rôle majeur de l'efficacité alimentaire dans l'optimisation de la marge économique.

Après avoir appliqué deux algorithmes de clustering (K-means et GMM) et regroupé les individus en deux clusters génétiques, nous avons observé que les vaches à *faible profit* étaient présentes dans les deux groupes (clusters). Cette répartition homogène complique la formulation de recommandations directes basées uniquement sur l'appartenance au cluster génétique. Néanmoins, les résultats de la matrice de corrélation confirment que les coûts alimentaires et la durée de lactation ont un impact négatif sur le profit, ce qui est cohérent avec la définition du profit (la différence entre la valeur du lait produit et le coût associé, Définition 1.1). Comme attendu, la production tend à diminuer en fin de lactation, ce qui entraîne une baisse de rentabilité alors que les coûts de gestion restent constants.

Dans la suite de ce travail, il sera pertinent d'analyser, pour chaque cluster génétique identifié, les individus présentant les *meilleurs profits*, afin d'identifier les facteurs distinctifs qui expliquent leurs performances supérieures. Une telle analyse permettra de dégager des recommandations ciblées, adaptées à chaque profil génétique, en vue d'améliorer la productivité et la rentabilité globale du troupeau.

5.3.4 Conclusion

Après comparaison de plusieurs approches de séries temporelles (ARIMA, VARMA et VARMAX), le modèle VARMAX($p = 1, q = 1$) a été retenu pour ses performances supérieures, avec une erreur minimale (RMSE $= 4.242$). La calibration du modèle a confirmé sa capacité à reproduire fidèlement les tendances observées, validant ainsi son utilisation pour la prédiction des indicateurs économiques.

Quant à l'analyse de la matrice de corrélation, elle montre des relations clés entre les variables de production, de santé et de rentabilité. La valeur journalière du lait et les composantes du lait (lait, protéine, gras) se sont révélées fortement corrélées au profit. À l'inverse, les jours en lait et le coût alimentaire cumulé présentent des corrélations négatives, soulignant l'importance de maîtriser la durée de lactation et d'améliorer l'efficacité alimentaire pour maintenir une rentabilité optimale.

La classification non supervisée (K-means et GMM) a permis de regrouper les individus en deux clusters, mais les vaches à *faible profit* sont présentes dans les deux groupes, rendant difficile l'émission de recommandations uniquement sur la base du cluster génétique. Ces résultats mettent néanmoins en lumière l'importance de combiner des informations économiques, de production et génétiques pour une analyse fine de la rentabilité.

Néanmoins, des améliorations demeurent possibles. Par exemple, les *K-médoïdes* constituent une alternative intéressante aux algorithmes K-means et GMM, notamment en raison de leur robustesse face aux valeurs aberrantes, et mériteraient une investigation plus approfondie. Leur mise en œuvre n'a toutefois pas été réalisée dans le cadre de cette thèse, en raison de contraintes de temps ainsi que de limitations pratiques liées aux performances.

5.4 Expérimentation 2 : LSTMDropout et Information Dropout Adaptée (IDA)

5.4.1 Implémentation

Dans cette section, nous présentons les détails pratiques de l'implémentation de notre cadre méthodologique. Les entrées sont X_{base}, un tenseur 3-dimensionnel, et X_{priv} également un tenseur 3-dimensionnel. L'architecture développée repose sur des couches LSTM (Section 1.4) et des couches linéaires. Chaque couche de dropout est remplacée par notre version de dropout hétéroscédastique. Aucun dropout n'est appliqué lors de la phase de test.

Comme indiqué à la section A.4.1, deux ensembles de données sont utilisés *production* et *génétiques*. Les informations de base X_{base}, de dimension $n_{samples} \times n_{seqlength} \times n_{features_{priv}}$, correspondent aux variables du jeu de données *production*, soit $n_{features_{base}} = 19$ variables après prétraitement (Table 5.3). Les informations privilégiées X_{priv} correspondent aux variables de données *génétiques*, soit $n_{features_{priv}} = 39$ variables au total, 8 relatives aux VEE (EBV) et 31 aux traits de conformation (Table 5.1). Les tailles respectives des séquences d'entrée et de sortie sont de $n_{seqlength_{input}} = 22$ (correspondant aux deux premières lactations, chacune comportant 11 enregistrements) et $n_{seqlength_{output}} = \{1; 11\}$, où 1 correspond à la tâche de régression et 11 à la tâche de prévision (forecasting). La taille de la sortie est $n_{outputs} \in \{1,3\}$.

5.4.1.1 Modèles de référence

Nous comparons notre méthode à plusieurs modèles de référence, regroupés en deux catégories : les problèmes à sortie unique et les problèmes à sorties multiples.

Problèmes à sortie unique (single-output) : *ARMA* est un modèle autorégressif à moyenne mobile, adapté aux séries temporelles stationnaires (Contreras *et al.*, 2003). Il combine des composantes autorégressives (AR) et de moyenne mobile (MA). *SARIMA* est un modèle ARIMA saisonnier qui étend l'ARIMA (autoregressive integrated moving average) afin de prendre en compte la composante saisonnière des séries tempo-

Table 5.3 – Ensemble des paramètres nécessaires pour notre problème d'élevage laitier.

Paramètres		**Valeurs**	**Détails**
Taille des entrées	Variables de base	19	Représente le nombre total de variables de production, de santé et de gestion.
	Informations privilégiées	$\{8, 39\}$	Nombre de variables correspondant soit uniquement aux *EBV*, soit à l'ensemble des caractéristiques phénotypiques (*EBV + traits de conformation*).
Taille des séquences	Entrée	22	Correspond aux deux premières lactations.
	Sortie	$\{1, 11\}$	1 pour la tâche de régression et 11 pour la tâche de prévision (troisième lactation).
Répartition Entraînement Test Validation	Entraînement	60%	
	Test	20%	Pourcentages de division du jeu de données entre entraînement, test et validation.
	Validation	20%	
Mode Fast-forward	-	Vrai, Faux	Option permettant d'utiliser ou non l'information privilégiée durant l'entraînement.

relles. *N-BEATS* est un modèle univarié et à sortie unique spécifiquement conçu pour la prévision de séries temporelles (Oreshkin *et al.*, 2020). *MuMu+Attention* : extension du modèle MuMu (Frasco. *et al.*, 2020) intégrant un mécanisme d'attention (Naghashi *et al.*, 2023).

Problèmes à sorties multiples (multi-output) : *Information Dropout* est la méthode proposée par Achille *et al.* (Achille et Soatto, 2018). *Temporal Fusion Transformer (TFT)* est une architecture fondée sur l'attention,

offrant des capacités d'interprétation des dynamiques temporelles (Lim *et al.*, 2021).

5.4.1.2 Métrique d'évaluation

De manière générale, l'évaluation des modèles de régression ou de prévision repose sur des métriques standards telles que l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). *Erreur quadratique moyenne (RMSE)* : mesure l'amplitude moyenne des erreurs en prenant en compte leur carré, ce qui accentue l'impact des grandes erreurs. *Erreur absolue moyenne (MAE)* : fournit la moyenne des différences absolues entre les valeurs prédites et les valeurs réelles. Ces deux métriques sont complémentaires : la MAE est plus robuste aux valeurs aberrantes, tandis que la RMSE y est plus sensible et capture également la variance des erreurs. Ainsi, la RMSE est particulièrement pertinente lorsque les grandes erreurs doivent être pénalisées plus fortement, comme c'est le cas dans nos tâches de régression et de prévision.

5.4.1.3 Ajustement des hyperparamètres

Pour l'optimisation des hyperparamètres (Tables 5.4 et 5.5), un algorithme de recherche en grille (*grid search*) est employé. Le taux d'apprentissage est fixé à 10^{-3}, avec un ordonnanceur permettant jusqu'à cinq (5) époques consécutives sans amélioration de l'erreur de validation avant l'application d'un facteur de décroissance de 5×10^{-3} pour le modèle *LSTMDropout*. Et une patience identique mais un facteur de décroissance de 5×10^{-4} est appliqué pour le modèle *Adaptive Information Dropout*. Afin de limiter le surapprentissage, un mécanisme d'*early stopping* avec une patience de 10 est utilisé, basé sur un critère de perte. La fonction d'activation ReLU et l'optimiseur Adam sont également retenus (Table 5.5). Enfin, lors de l'entraînement, la fonction de perte utilisée (Lambert *et al.*, 2018) intègre un terme de régularisation pondéré par l'hyperparamètre β, et les valeurs ν_1 et ν_2, selon la formule suivante :

$$\mathcal{L} = \mathcal{L} + \beta \times \frac{f(\nu_1, \nu_2)}{\text{batch size}}, \tag{5.1}$$

où $f(\cdot, \cdot)$ désigne une fonction définie sur ces deux variables.

5.4.2 Résultats

Pour évaluer les performances de nos modèles *LSTMDropout* et *Information Dropout Adaptée (IDA)*, des expérimentations ont été menées sur deux types de problèmes : la régression et la prévision (forecasting).

Table 5.4 – Optimisation des hyperparamètres.

Hyperparamètres	**Valeurs**
Nombre d'époques	{ 25 ; 50 ; 100 ; 250 }
Taille du lot (*batch size*)	{ 10 ; 30 ; 50 ; 100 ; 250 }
Taux d'apprentissage	$\{ 10^{-2} ; 10^{-3} ; 10^{-4} ; 10^{-5} \}$
Fonction d'activation	{ReLU, Tanh, SoftSign}
Optimiseur	Adam
Arrêt anticipé (*early stopping*)	patience $\in$ { 5 ; 8 ; 10 }
	critère = perte
Ordonnanceur (*scheduler*)	patience $\in$ { 5 ; 8 ; 10 }
	facteur de décroissance
Taille des couches cachées	{64 ; 128 ; 512 ; 1024 }
Taille du contexte	{16 ; 32 ; 64 ; 128 ; 512 }
Taille des couches fully-connected (FC)	{64 ; 128 ; 512 ; 1024 }
Nombre de couches	{ 5 ; 8 ; 10 }
Beta β	$\{ 3 \times 10^{-1} ; 3 \times 10^{-2} \}$
Taux de *dropout*	{ 0,15 ; 0,2 ; 0,3 }

Chacune d'elles est appliquée en deux scénarios distincts : *single-output* et *multi-output*.

— Cibles $\boldsymbol{y}$ pour les tâches **single-output** : la *valeur quotidienne du lait* (**DMV**). En production laitière, la DMV représente la valeur économique moyenne, en dollars, des composants du lait.
— Cibles $\boldsymbol{y}$ pour les tâches **multi-output** : les rendements quotidiens de *lait* (**DMY**), de *matière grasse* (**DFY**) et de *protéines* (**DPY**).

5.4.2.1 Tâche de régression

Dans la Table 5.7, les résultats obtenus sont présentés et comparés aux modèles N-BEATS, ARMA $(2, 2)$, SARIMA $(1, 0, 1, 11)$, MuMu+Attention, et Information Dropout (Achille et Soatto, 2018), pour les tâches de régression *single-output*. Il convient de noter que, dans ce cas, le modèle de référence MuMu+Attention a accès aux EBV et aux traits de conformation à la fois durant l'entraînement et le test. Les paramètres des modèles ARMA et SARIMA ont été déterminés à l'aide des fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF). La stationnarité des séries temporelles a également été vérifiée et confirmée.

Table 5.5 – Valeurs finales des hyperparamètres retenus pour les deux architectures.

Hyperparamètres	**Valeurs de LSTMDropout**	**Valeurs de Information Dropout Adaptée (IDA)**
Nombre d'époques	150	150
Taille du lot (*batch size*)	100	100
Taux d'apprentissage	10^{-3}	10^{-1}
Fonction d'activation	ReLU	ReLU
Optimiseur	Adam	Adam
Arrêt anticipé (*early stopping*)	patience = 10	patience = 10
Ordonnanceur (*scheduler*)	patience = 5	patience = 5
	facteur de décroissance = 5×10^{-3}	facteur de décroissance = 5×10^{-4}
Taille des couches cachées	64	64
Taille du contexte	64	256
Taille des couches fully-connected (FC)	256	32
Nombre de couches	8	5
Beta β	3×10^{-2}	3×10^{-1}
Taux de *dropout*	0,15	0,15
Fonction de perte	MSE	MSE

Table 5.6 – Performances de prédiction (métrique MSE) avec et sans information privilégiée (PI).

Paramètres	**Valeurs**	**Sans PI**		**Avec PI**	
		Test	Validation	Test	Validation
Activation	ReLU	**0,1529**	**0,1526**	**0,1101**	**0,1109**
	Softsign	0,1595	0,1591	0,1485	0,1488
	Tanh	0,1784	0,1781	0,1383	0,1387
Taille cachée	512	0,1126	0,1131	**0,0856**	**0,0855**
	128	**0,0725**	**0,0726**	0,0934	0,0939
	64	0,2781	0,2780	0,2377	0,2373

Les résultats montrent que notre modèle *LSTMDropout* obtient les plus faibles erreurs RMSE et MAE. Cela indique que l'incorporation d'informations privilégiées, en particulier les EBV et les traits de conformation (X_{priv}), a un effet positif sur les performances de prédiction. Ainsi, leur utilisation exclusive durant l'entraî-

Table 5.7 – Comparaison de nos résultats. **M** = multivarié, **U** = univarié, **RNN** = recurrent, **FF** = feed-forward simple.

Tâche	Modèles	Single		Multi-output	
		RMSE	MAE	RMSE	MAE
Régression	LSTMDropout [M]	**10,5516**	**8,4237**	8,6232	6,9874
	Information Dropout [M]	10,8508	8,6177	—	—
	Information Dropout Adaptée (IDA) [M]	10,8483	8,6234	—	—
	ARMA(2,2) [U]	21,0400	18,0087	—	—
	SARIMA(1,0,1,10) [M]	21,0392	18,0120	—	—
	N-BEATS [U]	21,7107	18,7554	—	—
Forecasting	LSTMDropout [RNN]	**10,5517**	**8,4237**	**5,4873**	**4,4752**
	LSTMDropout (sans PI) [RNN]	11,2615	9,6288	11,5358	10,4416
	LSTMDropout [FF]	10,6870	8,5263	**5,4832**	**4,4626**
	LSTMDropout (sans PI) [FF]	11,2615	9,6288	11,6316	10,4691
	Information Dropout [RNN]	10,8508	8,6177	7,3014	6,4017
	Information Dropout [FF]	10,9178	8,6238	5,8689	4,8780
	Information Dropout Adaptée (IDA) [RNN]	10,8483	8,6234	7,3918	6,3502
	Information Dropout Adaptée (IDA) [FF]	10,9012	8,6179	**5,7492**	**4,7723**
	Temporal Fusion Transformer (TFT)	14,9444	10,3264	12,4981	11,1652

nement améliore les performances du modèle tout en réduisant les ressources d'apprentissage (Table 5.7). Autrement dit, les EBV et les traits de conformation (X_{priv}) influencent la valeur quotidienne du lait, ce qui corrobore les travaux de (Frasco. *et al.*, 2020; Atkins, 2018).

En revanche, le modèle *Information Dropout Adaptée (IDA)* présente des performances comparables à celles du modèle *Information Dropout* (Achille et Soatto, 2018), suggérant que l'intégration de l'information privilégiée dans ce modèle n'apporte pas de gain significatif dans ce contexte.

Dans la suite, nous analysons les performances du modèle *LSTMDropout* dans le cadre des tâches de prévision *single-* et *multi-output*. Les performances des deux variantes de réseaux feedforward, simple et RNN, seront comparées.

5.4.2.2 Tâche de prévision (Forecasting)

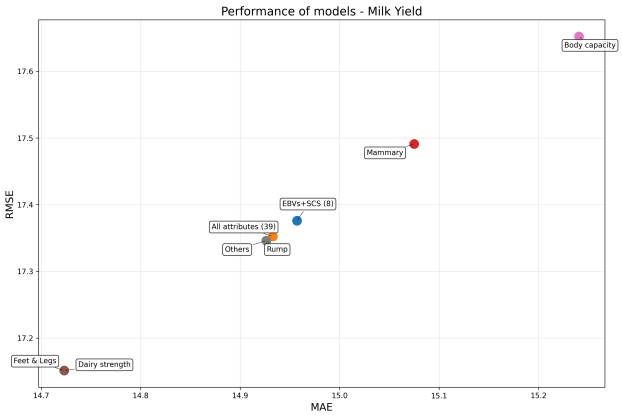
Pour les tâches de prévision *multi-output* (Table 5.7), les résultats indiquent que le modèle *LSTMDropout* surpasse systématiquement les modèles de référence, quel que soit le type de décodeur utilisé (*feedforward RNN* ou *feedforward* simple). Il est également observé que ce modèle obtient de meilleures performances pour les problèmes *multi-output* que pour les problèmes *single-output*. Cela peut s'expliquer par l'importance de certains traits de production, tels que DMY, DMF, DMP et le compte cellulaire somatique, dans la détermination de la valeur quotidienne du lait (DMV) et dans l'évaluation globale de la performance de production d'une vache. La comparaison entre le modèle *LSTMDropout* et sa version *sans information privilégiée* met en évidence que l'utilisation de l'information privilégiée a un impact réel et significatif sur les performances prédictives. La Table 5.7 montre également que notre modèle *LSTMDropout* associé à un décodeur de type *feedforward RNN* surpasse celui utilisant une architecture *feedforward simple* pour les tâches de prévision *single-output*.

Par ailleurs, le modèle *Information Dropout Adaptée (IDA)*, lorsqu'il est implémenté avec une architecture *feedforward simple*, présente de meilleures performances que l'approche *Information Dropout* proposée par (Achille et Soatto, 2018) pour la prévision *multi-output*.

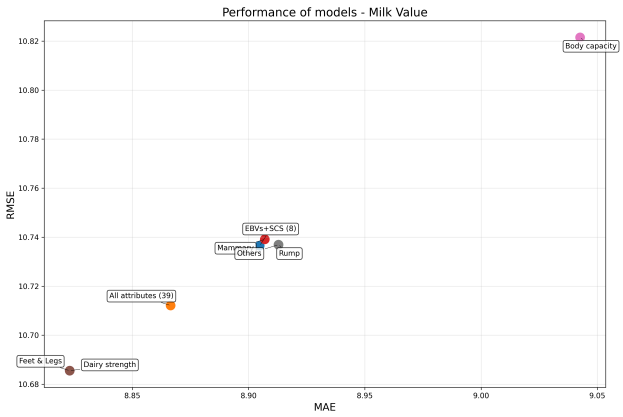
5.4.2.3 Études d'ablation

Compte tenu de l'efficacité démontrée de notre modèle *LSTMDropout* pour les tâches de prévision, nous menons l'étude d'ablation se concentrant exclusivement sur cette architecture. Par étude d'ablation, on entend une démarche expérimentale consistant à évaluer la contribution relative des différents attributs en les retirant afin de mesurer leur impact sur les performances globales. Cette approche permet ainsi d'identifier les attributs à l'origine de son efficacité.

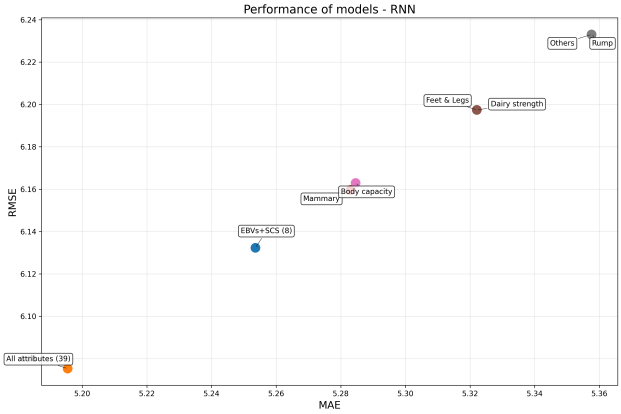
Cette analyse vise à évaluer l'impact des variables d'entrée sur les performances du modèle *LSTMDropout*. La Table 5.1 présente plus de détails concernant les variables utilisées. La Figure 5.8 montre que les attributs *dairy strength* et *feet & legs* améliorent significativement les performances dans les tâches de prévision *single-output*. De plus, l'étude révèle que la sélection de l'attribut *body capacity* et de l'ensemble des attributs (39) en tant qu'informations privilégiées permet d'obtenir de meilleures performances dans les tâches de prévision *multi-output*, notamment avec les architectures *feedforward* simple et *feedforward RNN*, respectivement.



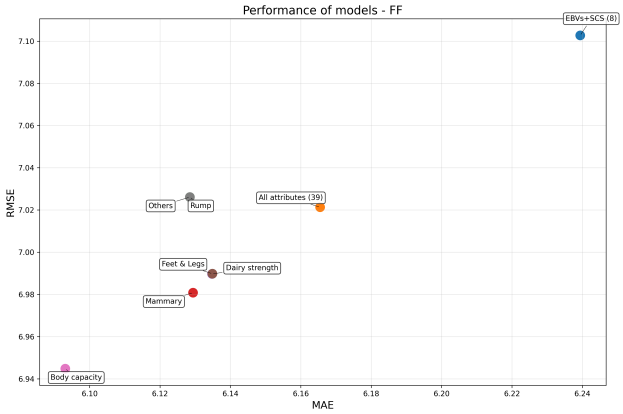
(a) Prevision unique-sortie - LSTMDropout : cible = rendement journalier de lait (DMY).



(b) Prévision unique-sortie - LSTMDropout : cible = valeur journalière de lait (DMV).



(c) Prévision multi-sortie - LSTMDropout [RNN] : cibles = rendements de lait, de protéine et de gras.



(d) Prévision multi-sortie - LSTMDropout [FF] : cibles = rendements de lait, de protéine et de gras.

Figure 5.8 – Étude d'ablation évaluant l'impact de différentes partitions des traits de conformation, utilisés comme informations privilégiées (PI), sur les performances de prévision.

5.4.3 Discussion

Les résultats montrent que les modèles LSTMDropout et Information Dropout Adaptée (IDA), intégrant le paradigme des informations privilégiées (PI), surpassent les modèles de référence ARMA(2,2), SARIMA(1,0,1,10), N-BEATS et TFT, quelle que soit la nature de la tâche, qu'il s'agisse de régression ou de prévision.

En particulier, le modèle LSTMDropout obtient les plus faibles erreurs RMSE et MAE pour les tâches de régression ainsi que pour les prévisions à sortie unique et multi-sorties, avec ou sans PI.

Les études d'ablation menées ont permis d'identifier les variables génétiques les plus déterminantes, tant pour la performance du modèle que pour la production laitière elle-même. Les attributs des groupes *dairy strength* et *feet & legs* contribuent de manière significative à l'amélioration des performances de prévision à sortie unique, que la sortie soit la valeur quotidienne ou le rendement quotidien du lait. Pour la prévision multi-sorties, les attributs les plus influents varient selon l'architecture de propagation choisie (RNN ou réseau feed-forward simple), mettant en évidence l'importance des *39 attributs* retenus ou du groupe *body capacity* respectivement. Ces résultats soulignent l'importance de la sélection de variables et le rôle des informations privilégiées dans l'optimisation des performances des modèles.

Concernant le modèle IDA, aucune amélioration notable n'a été observée par rapport au modèle Information Dropout d'Achille *et al.* (Achille et Soatto, 2018) pour les tâches de régression. En revanche, pour les tâches de prévision, AID présente de meilleures performances que le modèle Information Dropout. Néanmoins, LSTMDropout reste le modèle le plus performant dans l'ensemble des tâches évaluées.

Ces résultats ouvrent des perspectives intéressantes pour l'amélioration génétique des bovins laitiers et la gestion stratégique des élevages. La capacité du modèle à intégrer les traits de conformation et à produire des prévisions fiables sur la production et la rentabilité permettrait aux sélectionneurs et aux producteurs d'optimiser les décisions de reproduction, de réduire les coûts associés aux animaux moins productifs et, in fine, d'améliorer la durabilité économique et environnementale des exploitations.

5.4.4 Conclusion

Dans cette thèse, nous avons proposé et évalué différents modèles d'apprentissage statistique et profond pour estimer l'impact des traits de conformation sur la production laitière et la rentabilité à vie. Les résultats obtenus démontrent la supériorité du modèle *LSTMDropout*, qui surpasse les modèles de référence (ARMA, SARIMA, N-BEATS, TFT) aussi bien pour les tâches de régression que pour celles de prévision, en sortie unique comme en sorties multiples.

Les études d'ablation ont permis d'identifier les groupes de traits de conformation les plus déterminants, notamment *dairy strength*, *feet & legs* et, pour certains scénarios, *body capacity*, confirmant leur rôle central dans la prédiction de la production et de la rentabilité.

Ces résultats apportent une contribution significative à la mise en place de stratégies de sélection génétique

basées sur des données, en offrant un outil robuste, capable de s'adapter à différentes configurations de sortie et d'intégrer des informations privilégiées. En pratique, ces modèles pourraient aider les éleveurs et les sélectionneurs à optimiser les décisions de reproduction, réduire les coûts liés à des animaux moins productifs et améliorer la rentabilité et la durabilité des exploitations.

L'intégration d'informations privilégiées telles que les valeurs génétiques estimées (EBV) et les traits de conformation améliore de manière significative les performances des modèles de prévision et de régression dans les applications en production laitière. L'estimation de la variance à l'aide d'un *dropout* hétéroscédastique fondée sur ces informations privilégiées constitue une approche prometteuse pour accroître la précision des prédictions en analyse de séries temporelles. Ces méthodologies innovantes, LSTMDropout et Information Dropout Adaptée (IDA), permettent non seulement d'améliorer la qualité des modèles prédictifs, mais mettent également en évidence le potentiel de l'exploitation des données pour faire progresser les pratiques d'élevage laitier. Les résultats soulignent la capacité des modèles proposés à généraliser efficacement sur des tâches variées. En effet, ils démontrent des performances élevées tant en régression qu'en prévision (forecasting), et ce, pour des problèmes à sortie unique comme à sorties multiples. Cette robustesse suggère qu'un modèle unique peut constituer une solution polyvalente, réduisant ainsi le besoin de développer et d'optimiser plusieurs architectures spécifiques selon la nature de la tâche.

Les perspectives de recherche incluent l'élaboration d'une analyse théorique rigoureuse des algorithmes proposés ainsi que la généralisation du modèle *LSTMDropout* à un éventail plus large de séries temporelles. Des travaux futurs viseront également à comparer de manière systématique ce modèle aux approches existantes exploitant l'information privilégiée pour des tâches de régression et de prévision.

CHAPITRE 6

ÉVALUATION GÉNÉTIQUE D'UN INDIVIDU À PARTIR DES MARQUEURS GÉNÉTIQUES DE SES ANCÊTRES

À la suite de l'analyse de l'effet des marqueurs génétiques d'un individu sur sa rentabilité, ce chapitre vise à évaluer l'influence des marqueurs génétiques de ses ascendants sur cette même rentabilité.

6.1 Définition du problème

En évaluation génétique, chaque parent (père ou mère) transmet la moitié de son patrimoine génétique à sa progéniture (Figure 1.6).

- **Entrées** : pédigrée, performances des individus, performances et traits de conformation de leurs parents
- **Sortie** : valeurs génétiques estimées, composantes du lait.
- **Objectif** : prédire les valeurs génétiques d'un individu a partir des marqueurs génétiques (traits de conformation) de ses ancêtres.

On peut le définir comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon}, \tag{6.1}$$

où $\mathbf{Y} \in \mathbb{R}^{n \times m}$ représente la matrice des observations pour m traits, $\mathbf{X} \in \mathbb{R}^{n \times p}$ la matrice des effets fixes et $\mathbf{X}_{\text{dim}}$, $\boldsymbol{\theta} \sim \mathbb{N}$ la matrice des coefficients associés, $\mathbf{Z} \in \mathbb{R}^{n \times q}$ la matrice d'incidence des effets aléatoires, $\mathbf{g}$ la matrice des effets aléatoires, et $\boldsymbol{\epsilon}$ la matrice des résidus.

$$\begin{bmatrix} \mathbf{g} \\ \epsilon \end{bmatrix} = \mathbb{N}\Bigg(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \Bigg),$$

où $\mathbf{G} = \mathbf{A}\sigma^2_{\mathbf{u}}$ et $\mathbf{R} = \mathbf{I}\sigma^2_{\epsilon}$.

Les effets fixes considérés incluent les variables suivantes : pays, région, troupeau, nombre de lactations, race et nombre de jours en lait (Days in Milk, DIM). Les effets aléatoires correspondent au pédigrée et traits

de conformation.

6.2 Données

Nous utilisons les données Lactanet, comprenant (i) des mesures de performance individuelles, (ii) les traits de conformation des parents, employés comme informations privilégiées (PI) et (iii) les informations sur les relations de parenté (pédigrée). Nous avons retenu $1\,501$ animaux présentant au moins trois cycles de lactation. Parmi eux, 301 ont des progénitures. Notre sortie est composée des *quantités journalières de lait*, *de protéines et de gras*.

Tout d'abord, nous construisons la matrice des pédigrees qui servira à calculer les matrices d'incidence et les variances des effets aléatoires pour le modèle *DairyBLUP*. Les Figures 6.1 montrent que certains individus présentent plusieurs descendants, répartis sur deux générations.

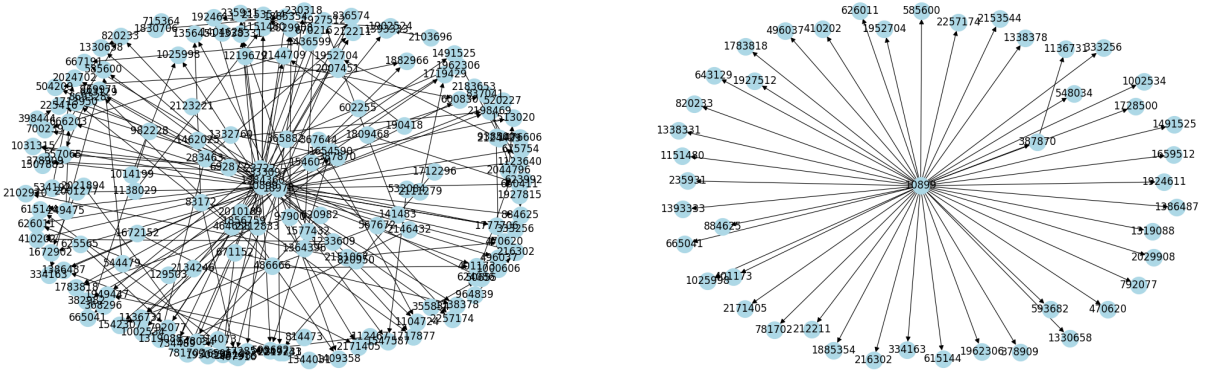


Figure 6.1 – Graphes de pédigrée. Le premier graphe (à gauche) met en évidence la variabilité de la profondeur des relations généalogiques observées dans la base de données, certains individus présentant des lignées particulièrement étendues. Le second graphe (à droite) illustre quant à lui la structure de la descendance en montrant qu'un individu peut engendrer plusieurs descendants directs.

6.3 Expérimentation 1 : DairyBLUP

6.3.1 Implémentation

Faute de données de séquençage génomique pour les animaux et leurs parents, nous appliquons *DairyBLUP*, un *modèle linéaire mixte* basé sur le pédigrée (Section 4.5). L'implémentation de *DairyBLUP* est divisé en

deux grandes parties. En premier, nous allons construire la matrice des pédigrée. Les variances des effets aléatoires ($\mathbf{G} = \mathbf{A}\sigma_u^2$) et des résidus ($\mathbf{R} = \mathbf{I}\sigma_e^2$) sont généralement estimées par la méthode du maximum de vraisemblance restreinte (REML) à partir du modèle mixte (Équation 4.21), la matrice de parenté $\mathbf{A}$ étant dérivée du pedigree (via Henderson), tandis que les matrices d'incidence $\mathbf{X}$ et $\mathbf{Z}$ sont construites à partir des facteurs fixes et des observations en lien avec les effets aléatoires.

Ensuite, nous appliquons le modèle *DairyBLUP*.

6.3.2 Résultats

Les estimations des variances des effets aléatoires $\hat{\mathbf{G}}$ et des résidus $\hat{\mathbf{R}}$ obtenues à l'aide des équations MME 4.21 :

$$\hat{\mathbf{G}} = \begin{pmatrix} 10,8238 & 0,2928 & 0,3507 \\ 0,2928 & 0,0108 & 0,0115 \\ 0,3507 & 0,0115 & 0,0189 \end{pmatrix}$$

$$\hat{\mathbf{R}} = \begin{pmatrix} 96,9482 & 3,1042 & 3,6121 \\ 3,1042 & 0,1164 & 0,1325 \\ 3,6121 & 0,1325 & 0,1762 \end{pmatrix}$$

Table 6.1 – Echantillon. Parents ayant deux descendants ou plus, avec nombre d'enfants par rôle (père ou mère).

ID Parent	# Enfants (Sire)	# Enfants (Dam)	# Total Enfants
10899	44	0	44
14556	30	0	30
18978	28	0	28
1126	23	0	23
24314	23	0	23
24298	23	0	23
18927	20	0	20
22559	20	0	20
21774	18	0	18
34	17	0	17

Les résultats indiquent que le modèle présente des performances globalement satisfaisantes (Figure 6.2a).

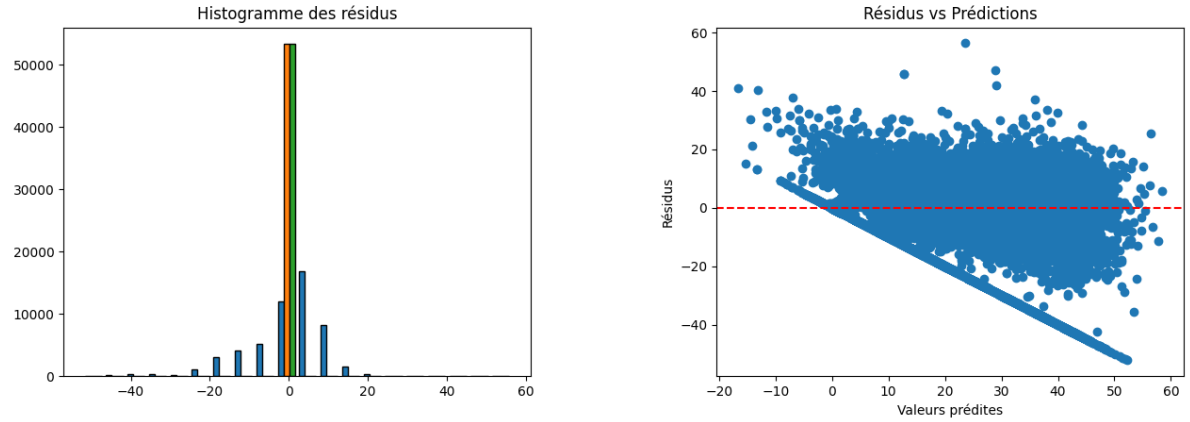
La Figure 6.2b montre que la majorité des résidus est centrée autour de zéro, ce qui suggère une bonne calibration globale du modèle. Néanmoins, la présence de résidus de grande amplitude révèle certains cas atypiques qui méritent une analyse plus approfondie afin d'en identifier les causes possibles.

En termes de métriques quantitatives, l'*erreur quadratique moyenne* (MSE) est de 32.414. Le *coefficient de détermination* R^2 atteint 0.49975 en moyenne pondérée et se répartit comme suit selon les traits considérés :

$$R^2_{\text{hr\_24\_milk}} = 0,49997, \quad R^2_{\text{hr\_24\_pt}} = 0,40267, \quad R^2_{\text{hr\_24\_ft}} = 0,41679.$$

Les résultats issus du modèle *pedigree-BLUP* montrent que les performances globales sont satisfaisantes. L'histogramme des résidus (Figure 6.2a) révèle que ceux-ci sont majoritairement centrés autour de zéro, ce qui indique une bonne absence de biais dans les prédictions des valeurs génétiques. Cependant, la distribution présente des résidus extrêmes, suggérant la présence de quelques individus atypiques ou des observations influentes dans le jeu de données.

La Figure 6.2b, représentant les résidus en fonction des valeurs prédites, confirme que la plupart des résidus sont proches de zéro, mais quelques points présentent des écarts importants, en particulier pour les valeurs prédites élevées.



(a) Histogramme des résidus.

(b) Valeurs réelles vs prédites.

Figure 6.2 – Résultats du modèle DairyBLUP : (a) histogramme des résidus ; (b) valeurs réelles vs prédites.

La Table 6.2 indique que les dix variations observées, les plus élevées, dans les prédictions du modèle *DairyBLUP* (Section 4.5) proviennent principalement de sa difficulté à modéliser les valeurs nulles ainsi que les

valeurs aberrantes.

Afin de pallier ces limitations, le modèle *LSTMDropout* sera appliqué dans la suite.

Table 6.2 – Les 10 résidus multi-traits les plus élevés.

anm_id	resid_norm	hr_24_milk_obs	hr_24_milk_pred	hr_24_milk_resid	hr_24_pt_obs	hr_24_pt_pred	hr_24_pt_resid
137348	56,428933	79,9	23,534775	56,365225	2,789	0,687727	2,101273
32840	52,286329	0,0	52,225946	-52,225946	0,000	1,512277	-1,512277
136381	51,891653	0,0	51,842464	-51,842464	0,000	1,470780	-1,470780
95046	51,428890	0,0	51,368634	-51,368634	0,000	1,517597	-1,517597
116120	50,963247	0,0	50,908092	-50,908092	0,000	1,498603	-1,498603
120127	50,205805	0,0	50,145371	-50,145371	0,000	1,550068	-1,550068
113824	49,518847	0,0	49,456603	-49,456603	0,000	1,529779	-1,529779
43722	49,153969	0,0	49,097488	-49,097488	0,000	1,340865	-1,340865
77802	48,594728	0,0	48,537818	-48,537818	0,000	1,476634	-1,476634
121989	48,504024	0,0	48,446000	-48,446000	0,000	1,444705	-1,444705

6.3.3 Discussion

Les erreurs indiquent que près de 50% de la variabilité de la production laitière quotidienne est expliquée par le modèle, ce qui représente un niveau de performance considérable pour un problème de régression sur des données animales.

Cette tendance pourrait refléter une sous-modélisation de certaines sources de variation ou l'absence de covariables explicatives pertinentes (par exemple, des effets environnementaux ou génétiques non capturés par le pédigrée).

Dans un contexte de BLUP, des résidus centrés et relativement dispersés sont attendus car les solutions BLUP sont, par construction, des estimateurs sans biais des effets aléatoires. Toutefois, l'existence de résidus extrêmes justifiait une investigation complémentaire afin d'identifier les individus ou enregistrements qui influencent le plus l'estimation des valeurs génétiques (VEE). De telles observations pourraient indiquer des erreurs d'enregistrement, des animaux présentant des performances atypiques, ou des interactions non modélisées, comme dans notre cas. La Table 6.2 indique que les variations observées dans les prédictions du modèle *DairyBLUP* (Section 4.5) proviennent principalement de sa difficulté à modéliser les périodes de

tarissement (valeurs nulles) ainsi que les valeurs aberrantes (une erreur de saisie, un pic de lactation).

Le modèle **DairyBLUP**, fondé sur le principe du Best Linear Unbiased Prediction (BLUP), constitue une approche robuste et largement utilisée pour l'estimation des valeurs génétiques. Il exploite efficacement les informations de pédigrée et de performance en utilisant des modèles linéaires mixtes, ce qui en fait un outil fiable pour la sélection génétique. Cependant, DairyBLUP repose sur des hypothèses de linéarité et de normalité qui peuvent limiter sa capacité à capturer des relations complexes et non linéaires, ainsi que les interactions entre génotypes et environnement. De plus, il considère généralement une seule observation par individu, ce qui peut s'avérer insuffisant lorsque les données comportent plusieurs valeurs de la lactation ou des périodes de production irrégulières.

Les modèles d'apprentissage profond, tels que les réseaux de neurones récurrents (LSTM, Définition 1.4), constituent une alternative intéressante, car ils sont capables de modéliser des relations non linéaires et de tirer parti des dépendances temporelles présentes dans les séries de données. Ils permettent ainsi d'exploiter l'ensemble de l'historique de production d'un individu et de mieux gérer les irrégularités et les valeurs aberrantes. Néanmoins, ces modèles nécessitent des volumes de données plus importants et une puissance de calcul plus élevée, et leur interprétation reste plus complexe que celle des modèles linéaires classiques.

6.3.4 Conclusion

En général, les modèles basés sur le pedigree-BLUP considèrent une seule observation par individu. Dans notre cas, chaque individu présente plusieurs cycles de lactation, ce qui complexifie la résolution du problème de régression linéaire mixte, qui constitue la base du BLUP.

Bien que DairyBLUP affiche de bonnes performances globales, il présente des difficultés à modéliser correctement les périodes de tarissement ainsi que les valeurs aberrantes. Une amélioration s'avère donc nécessaire, en recourant à des modèles capables de capturer des structures de données plus complexes et de mieux gérer les irrégularités présentes dans les séries temporelles de production.

Dans cette optique, nous proposons l'utilisation de modèles d'apprentissage profond, tels que LSTMDropout, qui offrent une capacité supérieure à modéliser des relations non linéaires, à exploiter les dépendances temporelles et à s'adapter à des données hétérogènes et bruitées. Cette approche ouvre la voie à une meilleure estimation de l'effet des traits de conformation sur la production et la rentabilité laitière.

6.4 Expérimentation 2 : LSTMDropout

6.4.1 Implémentation

Dans notre protocole expérimental, nous adoptons le paradigme des informations privilégiées (PI) et implémentons principalement le modèle LSTMDropout (Section 4.3). Les variantes du modèle sont appliquées aussi bien aux tâches de régression qu'aux tâches de prévision, y compris pour les problèmes à sorties multiples. Dans ce cadre, les pédigrées et les performances des vaches constituent les variables de base X_{base}, tandis que les traits de conformation sont utilisés comme informations privilégiées X_{priv}, enrichissant ainsi l'apprentissage du modèle.

6.4.2 Résultats

Les résultats présentés dans la Table 6.3 montrent que le modèle de prédiction des valeurs génétiques estimées (VEE) atteint un niveau d'erreur relativement faible, avec un MAE $\approx 6,9479$ et un RMSE $\approx 8,2960$ dans le cadre de la régression. Cela signifie que, en moyenne, l'écart absolu entre les VEE prédites et les valeurs de référence reste modéré, ce qui témoigne d'une bonne précision du modèle. Pour la tâche de prévision, le modèle avec décodeur *feedforward simple* présente des erreurs légèrement plus faibles que la variante *RNN*, avec un RMSE $\approx 9,498$ contre $10,556$.

Table 6.3 – Performances du modèle de prédiction des VEE pour les tâches de régression et de prévision (MAE et RMSE) en fonction des variables de base et des informations privilégiées (PI) utilisées.

Tâche	Base (Features)	PI (Features)	MAE	RMSE
Régression	19 (lactation)	46 (pédigrée + conformation)	**6,9479**	**8,2960**
	26 (lactation + pédigrée)	39 (conformation)	**6,9479**	**8,2960**
	19 (lactation)	39 (conformation)	6,9874	8,6232
Prévision (feed-forward = simple)	26 (lactation + pédigrée)	39 (conformation)	*8,1996*	*9,4975*
Prévision (feed-forward = RNN)	26 (lactation + pédigrée)	39 (conformation)	8,9790	10,5557

6.4.3 Discussion

Pour les tâches de régression, nos résultats (Table 6.3) montrent que l'utilisation conjointe des informations de lactation et de pédigrée comme variables de base, complétées par les traits de conformation en tant qu'informations privilégiées (PI), n'apporte qu'une amélioration marginale des performances par rapport au

scénario où les informations de pédigrée sont considérées comme PI aux côtés des traits de conformation. Cette observation suggère que le pédigrée, qu'il soit intégré en entrée principale ou en tant que PI, explique déjà une part importante de la variance des VEE. Dans une optique d'optimisation computationnelle, le second scénario, qui mobilise moins de ressources tout en offrant des performances comparables, constitue donc une solution préférable.

Concernant les tâches de prévision, le modèle LSTMDropout obtient ses meilleures performances avec l'architecture feed-forward simple. Ce résultat laisse penser qu'un modèle plus simple capte plus efficacement les relations linéaires et non linéaires entre les traits de production et les VEE futures, tandis que la variante récurrente pourrait être sujette à un surapprentissage ou à une accumulation d'erreurs lors de la propagation sur des séquences temporelles longues.

Comparés aux résultats présentés dans la Table 5.7, ces constats confirment que l'intégration du pédigrée en tant qu'information privilégiée améliore la qualité des prédictions pour une tâche de régression. Ils mettent également en évidence l'importance du choix de l'architecture du décodeur, qui doit être adapté à la complexité du problème afin d'optimiser la précision des estimations tout en limitant le risque de surajustement (pour une tâche de prévision).

6.4.4 Conclusion

En résumé, les résultats obtenus démontrent que l'intégration du pédigrée en tant qu'information privilégiée améliore la qualité des prédictions, tout en réduisant le coût computationnel, en particulier pour les tâches de régression. Cela confirme que la régression constitue un cadre plus stable et précis que la prévision pour l'estimation des valeurs génétiques.

Pour les tâches de prévision, l'architecture feed-forward simple de LSTMDropout surpasse systématiquement la variante récurrente, ce qui suggère que des modèles plus simples généralisent mieux et sont moins sensibles au surapprentissage et à la propagation d'erreurs sur les séquences temporelles longues. Ces résultats renforcent l'idée qu'un compromis entre la complexité du modèle et sa performance est essentiel pour optimiser la robustesse des prédictions.

CHAPITRE 7
PRÉDICTION DU PROFIT ET DES RENDEMENTS DE PRODUCTION

Ce chapitre traite de la problématique de la prévision de la rentabilité ainsi que des composantes du lait. Alors que les chapitres précédents étaient consacrés à l'évaluation génétique, il convient à présent de développer des modèles de prévision afin de compléter ces analyses.

7.1 Définition du problème

Étant donné les performances phénotypiques et temporelles d'un individu, ainsi que ses traits de conformation, notre problème peut être défini comme suit :

- Entrée : performances des individus (phénotypes et séries temporelles), traits de conformation.
- Sortie : multiples sorties (composantes du lait).
- Objectif : prédire la rentabilité à vie ou les rendements laitiers d'un individu donné sachant ses caractéristiques génétiques, à un certain horizon temporel (par exemple $profit_{t+h}$, avec $t = 22, \quad h = 10$).

7.2 Données

La base de données *Life Time revenue* (Table 7.1 et Figure 7.1) est utilisée pour mener nos expérimentations. On sélectionne des individus ayant au minimum deux lactations et sept enregistrements. Résultats : 850 individus dont 70% pour l'entraînement et 30% pour le test. Les données sont structurées comme suit : la première partie ($t = 1$), une deuxième partie ($t = \{2, ..., T\}$) et une dernière partie (target $\mathbf{Y}$) (Figure 7.2).

7.3 Expérimentation 1 : DairyLuPTS

7.3.1 Implémentation

Nous avons appliqué notre premier modèle *DairyLuPTS* (Section 4.2). Dans un premier temps, nous avons étudié l'effet de l'horizon de prévision h sur les performances du modèle. Ensuite, nous avons fait varier le pas temporel (*timestep*) entre les observations afin d'analyser son impact sur la qualité des prédictions.

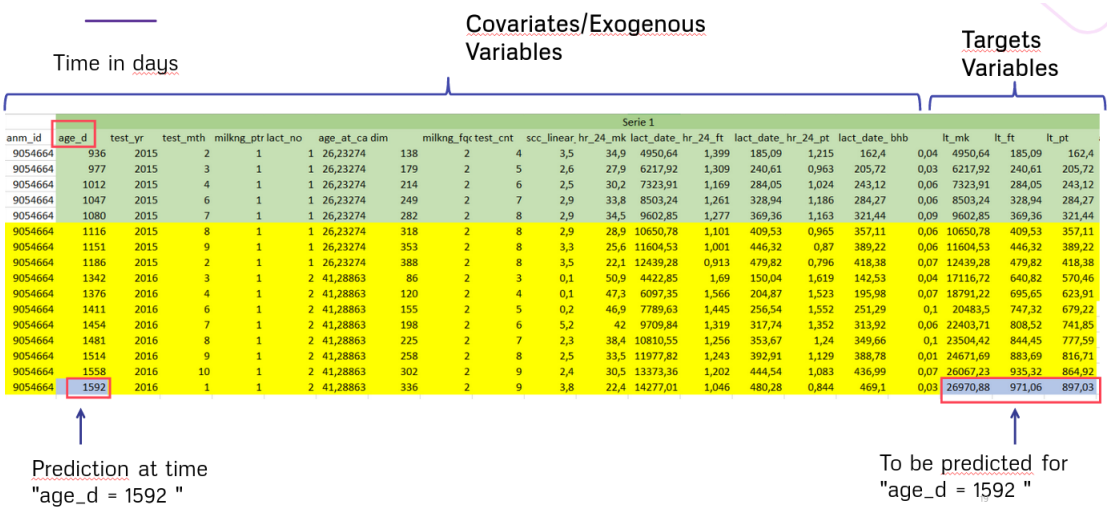


Figure 7.1 – Exemple du problème de prédiction.

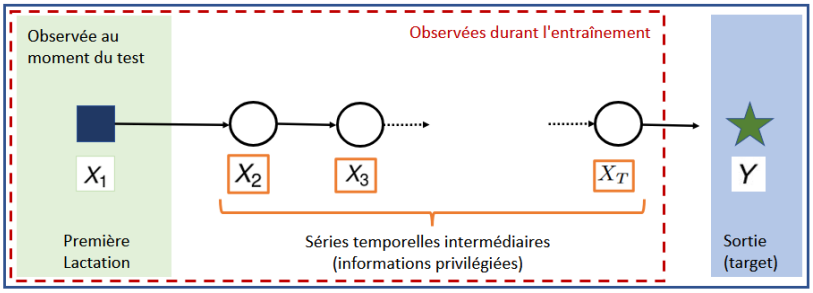


Figure 7.2 – Répartition des données.

- **Entrées** : données temporelles de lactation (*Life Time Revenue*) et données *génétiques*, avec input $\in \mathbb{R}^{(\text{n\_samples,n\_sequences,n\_features})}$.
- **Base** : données de base au temps $t = 1$.
- **Informations privilégiées** : séries temporelles intermédiaires $t \in \{2, \ldots, T\}$.
- **Sortie** : prédiction de la production totale de lait à la seconde lactation.
- **Modèle** : *DairyLuPTS*.

Deux aspects sont étudiés : l'influence de l'horizon de prévision h et celle du pas temporel.

Exemple illustratif : considérons une série temporelle de 10 valeurs $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

- Pour un horizon $h = 12$,
- avec un pas $= 1$, la série reste inchangée : $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

— avec un pas $= 2$, la série devient $\{1, 3, 5, 7, 9\}$.

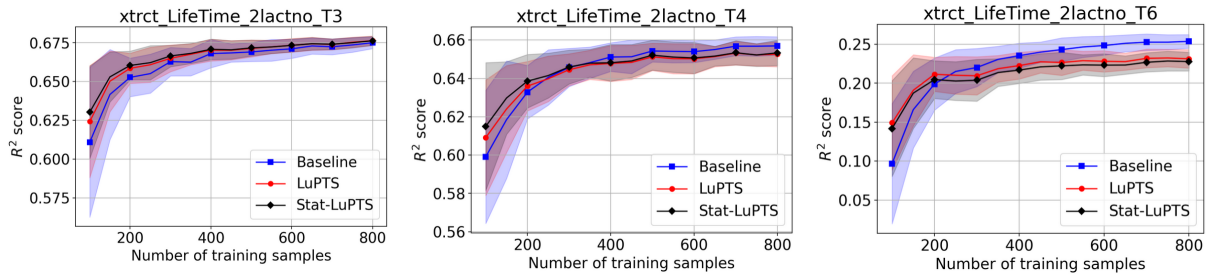
— avec un pas $= 3$, la série devient $\{1, 4, 7, 10\}$.

7.3.2 Résultats

Pour mener à bien ces expérimentations, une analyse approfondie a été menée sur l'influence du pas d'échantillonnage et de l'horizon de prévision sur la performance du modèle.

Pour un pas égal à 1, le modèle *DairyLuPTS* obtient ses meilleures performances avec des horizons de prévision courts, tels que $h \in \{3, 4\}$ (Figure 7.3). Nous observons également que le modèle fonctionne mieux avec un pas temporel de 2 (Figure 7.4).

L'augmentation du pas temporel réduit la corrélation entre observations consécutives, diminuant ainsi la redondance dans les données tout en conservant les tendances globales. Cela peut améliorer la capacité du modèle à généraliser, mais un pas élevé risque de perdre des informations fines et de dégrader la précision de la prévision.

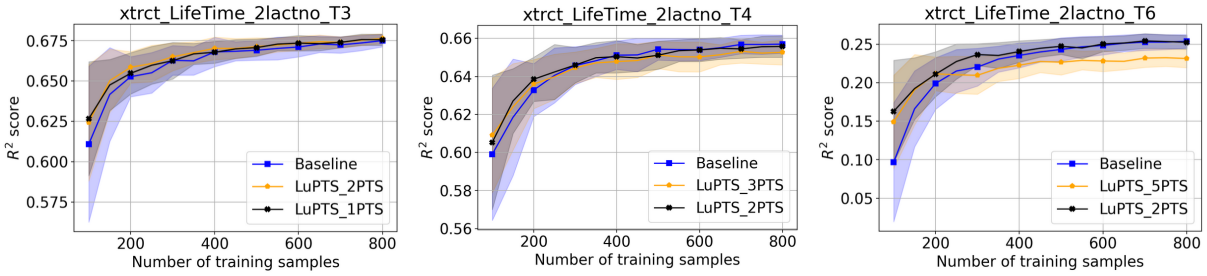


(a) Horizon 3 (b) Horizon 4 (c) Horizon 6.

Figure 7.3 – Performances de prévision pour différents horizons temporels et pas égal à 1. *LuPTS* représente le modèle *DairyLuPTS* appliqué à des séries non stationnaires, tandis que *Stat-LuPTS* correspond à sa variante pour les séries différenciées (stationnaires).

7.3.3 Discussion

Les résultats mettent en évidence deux variantes de DairyLUPTS : Stat-LUPTS, conçue pour les séries temporelles stationnaires, et LUPTS, adaptée aux séries non stationnaires (Figures 7.3 et 7.4).



(a) Pas temporels $\{1, 2\}$.

(b) Pas temporels $\{2, 3\}$.

(c) Pas temporels $\{2, 5\}$.

Figure 7.4 – Variation des pas temporels et horizons.

Les Figures 7.3 et 7.4 montrent que les modèles DairyLUPTS surpassent le modèle de référence (régression linéaire simple, Algorithme 1) pour les petits échantillons (< 400). Ces résultats soulignent le rôle central des données dans la précision des prédictions de séries temporelles.

Nous observons également qu'un pas temporel trop élevé entraîne une perte d'information pertinente pour notre cas d'étude dans le secteur laitier (Figure 7.4). Il convient de rappeler qu'une lactation s'étend sur une durée d'environ dix mois, incluant une période de tarissement (Définition 1.2 et Figure 1.1). Par ailleurs, les Figures 7.3 mettent en évidence que l'intégration des informations privilégiées sous forme de séries temporelles constitue une approche prometteuse lorsque l'horizon de prévision reste modéré.

Ce constat est cohérent avec la dynamique des courbes de lactation, caractérisée par une phase initiale de croissance jusqu'à un pic, suivie d'une décroissance progressive (Figure 1.1, Définition 1.2). La première valeur de production en début de lactation ne contient pas, à elle seule, suffisamment d'informations pour prédire avec précision les productions à un horizon éloigné.

Pour pallier cette limitation, il serait pertinent d'investiguer, d'un point de vue théorique et pratique, un modèle amélioré tel que Baseline+ (Algorithme 3). Une attention particulière devra cependant être portée au coût computationnel associé à ce modèle, afin de garantir sa faisabilité dans un contexte de production.

En perspective, ces observations ouvrent la voie à des recherches futures visant à optimiser DairyLUPTS, notamment en explorant des architectures plus robustes aux pas temporels longs et en intégrant des techniques avancées de sélection de variables ou de régularisation. De telles améliorations pourraient renforcer la précision des prévisions tout en maintenant un compromis acceptable entre performance et coût de cal-

cul. Une autre voie serait l'exploration de modèles d'apprentissage profond, capables de mieux exploiter les dépendances à long terme pour améliorer les prédictions à grand horizon.

7.3.4 Conclusion

Nous avons proposé et évalué le modèle DairyLUPTS (Section 4.2) pour la prédiction de séries temporelles dans le secteur laitier, en le comparant à un modèle de régression linéaire simple (Baseline). Les résultats obtenus montrent que DairyLUPTS offre de meilleures performances, en particulier pour les petits échantillons (< 400), confirmant l'importance d'exploiter la dynamique temporelle des données de lactation.

L'analyse des pas temporels a mis en évidence qu'un pas trop élevé entraîne une perte d'information critique pour la prévision. Par ailleurs, l'intégration des informations privilégiées sous forme de série temporelle s'est révélée particulièrement bénéfique pour les horizons de prévision modérés, en accord avec la forme caractéristique de la courbe de lactation, marquée par une phase de montée, un pic et une décroissance progressive.

Cependant, la première valeur de lactation ne suffit pas à prédire avec précision les productions à un horizon éloigné, ce qui limite la capacité du modèle sur le long terme. Pour remédier à cette limite, l'exploration d'un modèle amélioré, tel que Baseline+ (Algorithme 3), constitue une piste prometteuse. Une attention particulière devra être portée au coût computationnel de ce modèle afin de maintenir un compromis acceptable entre précision et efficacité.

Ces travaux ouvrent ainsi la voie à de futures recherches visant à optimiser DairyLUPTS, notamment en testant des architectures plus robustes aux grands horizons de prévision, en intégrant des techniques de régularisation.

Dans la section suivante, d'autres modèles statistiques seront examinés afin de traiter la problématique étudiée et d'enrichir l'analyse comparative.

7.4 Expérimentation 2 : Autres modèles statistiques

7.4.1 Implémentation

La prédiction des rendements des trio lait, protéine et gras à un certain horizon futur sera abordée en utilisant des modèles d'apprentissage automatique : arbre de décision (Rokach et Maimon, 2005; Charbuty et Abdulazeez, 2021), régression multi-sortie directe et régression multi-sortie chaînée. Ces deux derniers modèles sont des algorithmes de régression multi-sorties enveloppants (Wrapper Multioutput Regression Algorithms) (Borchani *et al.*, 2015; Xu *et al.*, 2019). Les bases de données *Life-Time Revenue* ont été utilisées pour l'expérimentation des modèles précédemment cités. La figure 7.1 nous illustre également le problème abordé ici. Nous avons fait appel également a la validation croisée $K = 10$.

Les données *Life Time Revenue* exploitées ici sont un échantillon complet (pas de données manquantes). Néanmoins, nous avons sélectionné les individus ayant au moins sept (7) enregistrements. Nous avons divisé notre sélection en ensemble d'entraînement, de validation et de test, respectivement 60%, 25% et 15%.

7.4.2 Résultats

7.4.2.1 Arbre de Décision

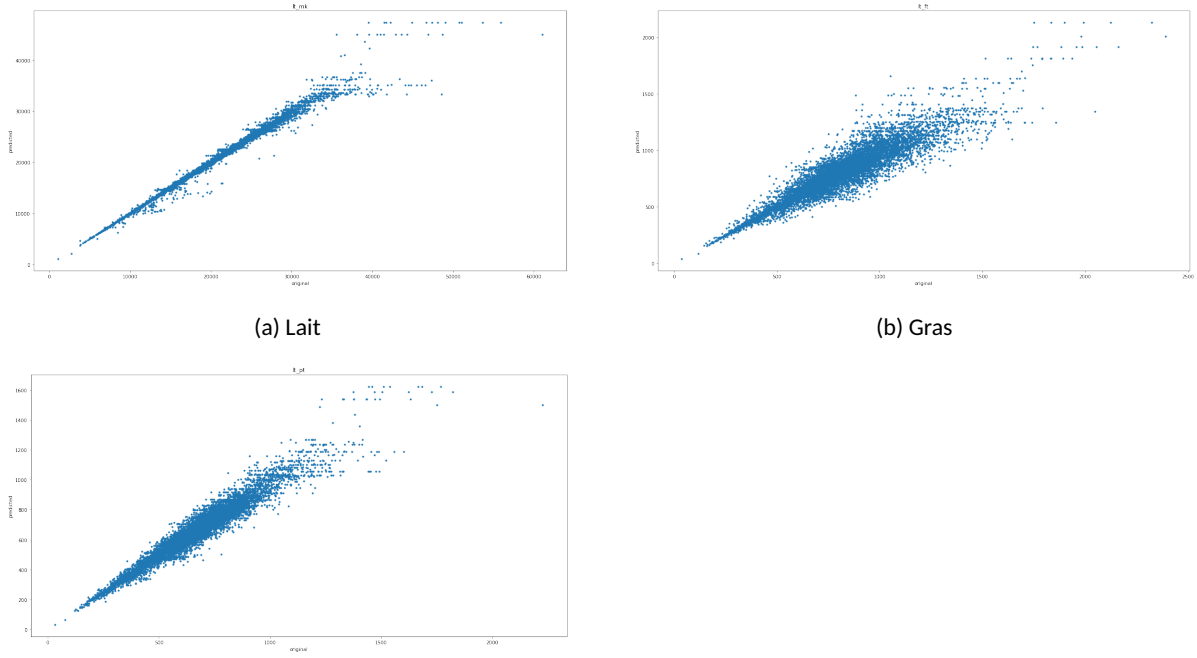
Un arbre de décision (decision tree) est un classificateur exprimé comme une partition récursive de l'espace d'instance. L'arbre de décision est composé de nœuds qui forment un arbre à racine, c'est-à-dire un arbre dirigé avec un nœud appelé "racine" qui n'a pas d'arêtes entrantes. Tous les autres nœuds ont exactement un bord entrant. Un nœud avec des bords sortants est appelé nœud interne ou nœud de test. Tous les autres nœuds sont appelés feuilles (également appelées nœuds terminaux ou de décision). Dans un arbre de décision, chaque nœud interne divise l'espace d'instance en deux ou plusieurs sous-espaces en fonction d'une certaine fonction discrète de la valeur des attributs d'entrée. Les métriques de MAPE présentées dans la Table 7.1 mettent en évidence les limites de l'arbre de décision appliqué à notre problème. Bien que la précision soit relativement satisfaisante pour la composante *lait* (MAPE ≈ 1.31), les erreurs augmentent sensiblement pour les composantes *protéine* (MAPE ≈ 5.40) et surtout *gras* (MAPE ≈ 8.30).

Les diagrammes de dispersion des données réelles et prédites sont l'une des formes les plus riches de visualisation des données. Les diagrammes de dispersion pour les rendements de lait, de gras et de protéine sont illustrés par les graphiques des Figures 7.5a, 7.5b et 7.5c. On constate que nos points tendent plus vers

Table 7.1 – Précision de l'arbre de décision.

Composantes	MAPE
Lait	1,3126
Protéine	5,4033
Gras	8,2990

une ligne diagonale $axe_y = axe_x$ pour la prédiction du lait contrairement aux autres composantes (gras et protéine). Ce qui confirme les métriques du tableau 7.1.



(a) Lait

(b) Gras

(c) Protéine

Figure 7.5 – Graphiques des valeurs prédites versus les valeurs réelles avec l'arbre de décision.

7.4.2.2 Algorithmes de régression multi-outputs enveloppants : Régression multi-output directe

L'approche directe de la régression multi-output implique de diviser le problème de régression en un problème distinct pour chaque variable cible à prédire. Cela suppose que les sorties sont indépendantes les unes des autres, ce qui peut ne pas être une hypothèse correcte. Néanmoins, cette approche peut fournir des prédictions étonnamment efficaces sur une gamme de problèmes et peut valoir la peine d'être essayée,

au moins comme base de performance.

Les résultats présentés dans la Table 7.2 mettent en évidence la haute précision du modèle de régression multi-output directe. Les valeurs de MAPE obtenues sont extrêmement faibles pour l'ensemble des composantes (lait, protéine et gras), indiquant que le modèle reproduit presque parfaitement les valeurs observées.

Table 7.2 – Précision de la régression multi-output directe.

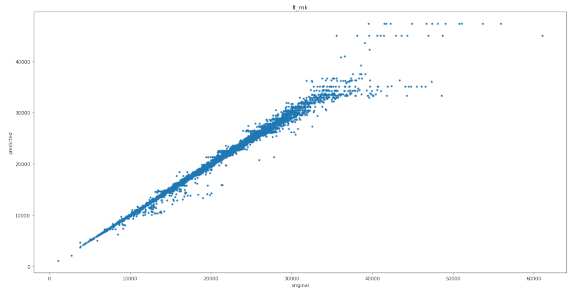
Composantes	**MAPE**
Lait	0,000002
Protéine	0,000382
Gras	0,000477

Les diagrammes de dispersion pour les rendements de lait, de gras et de protéine sont illustrés par les graphiques des figures 7.6a, 7.6b & 7.6c. On constate également que nos points tendent plus vers une ligne diagonale $axe_y = axe_x$ pour la prédiction du lait, contrairement aux autres composantes (gras et protéine).

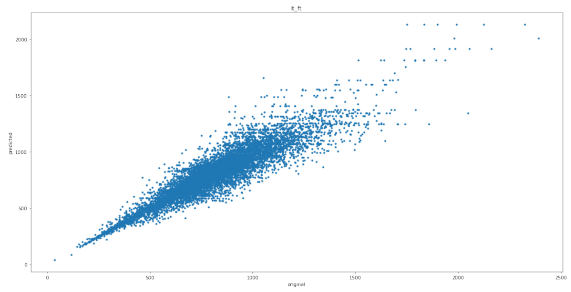
7.4.2.3 Algorithmes de régression multi-output enveloppants : Régression multi-output chaînée

Une autre approche de l'utilisation de modèles de régression à output unique pour la régression à outputs multiples consiste à créer une séquence linéaire de modèles. Le premier modèle de la séquence utilise l'entrée et prédit une sortie ; le deuxième modèle utilise l'entrée et la sortie du premier modèle pour faire une prédiction ; le troisième modèle utilise l'entrée et la sortie des deux premiers modèles pour faire une prédiction, et ainsi de suite. Par exemple, si un problème de régression à sorties multiples nécessite la prédiction de trois valeurs y_1, y_2 et y_3 à partir d'une entrée X, il peut être divisé en trois problèmes de régression à sortie unique dépendants, comme suit :

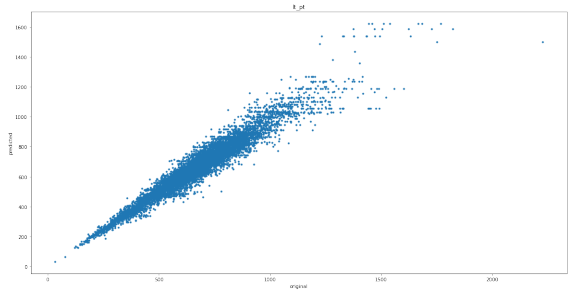
- Problème 1 : Sachant X, prédire y_1.
- Problème 2 : Sachant X et $\hat{y}_1$, prédire y_2.
- Problème 3 : Sachant X, $\hat{y}_1$ et $\hat{y}_2$, prédire y_3.



(a) Lait



(b) Gras



(c) Protéine

Figure 7.6 – Algorithmes de la régression multi-output enveloppants : Régression multi-output directe.

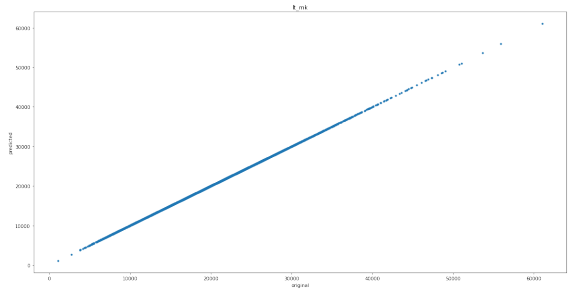
Le tableau 7.3 contient les métriques MAPE du modèle de régression multi-output chaînée. Les diagrammes de dispersion pour les rendements de lait, de gras et de protéine, illustrés respectivement par les graphiques des figures 7.7a, 7.7b & 7.7c, nous montrent que nos points sont très proches d'une ligne diagonale $axe_y = axe_x$ pour toutes nos prédictions (lait, gras et protéine).

Table 7.3 – Précision de la régression multi-output chaînée.

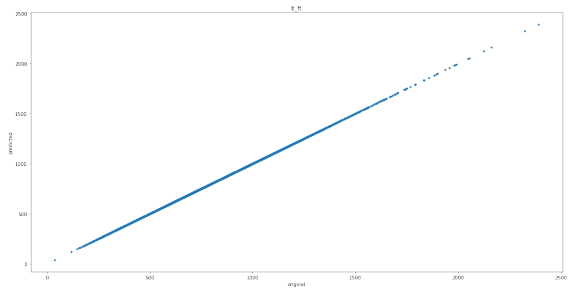
Composantes	**MAPE**
Lait	0,000003
Protéine	0,000346
Gras	0,004985

7.4.3 Discussion

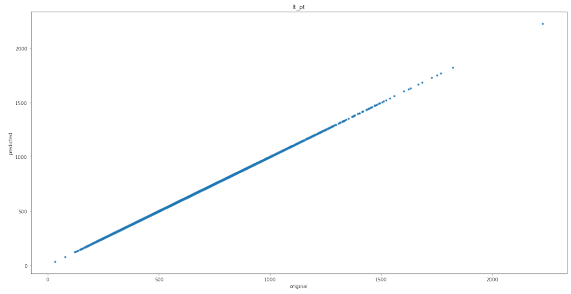
Les résultats de l'arbre de décision indiquent qu'il peine à capturer la variabilité des composantes plus sensibles, probablement en raison de la nature fortement non linéaire et continue des relations entre les traits de production et les variables explicatives. De plus, les arbres de décision ont tendance à segmenter



(a) Graphique de prédiction du lait



(b) Gras



(c) Protéine

Figure 7.7 – Algorithmes de régression multi-output enveloppants : Régression multi-output chaînée.

Table 7.4 – Validation croisée ($k = 10$) pour différents modèles.

Modèle	**MAE moyen**	**Écart-type (MAE)**	**Remarque**
Régression multi-output directe	0.002	0.000	—
Arbre de décision	**30.008**	1.153	Performance nettement inférieure
Régression multi-output chaînée	0.002	0.001	—

l'espace des données en régions disjointes, ce qui peut induire une perte de précision dans les prédictions lorsque la relation entre les variables est plus progressive et lisse. Ainsi, même si l'arbre de décision peut constituer un modèle de base simple et interprétable, il ne semble pas adapté pour atteindre une précision fine sur l'ensemble des composantes de la production laitière. Ces observations confirment la nécessité de recourir à des modèles plus expressifs, capables de mieux représenter les relations complexes entre variables, comme les modèles de régression multi-sorties ou les approches d'apprentissage profond.

Comparées aux résultats de la régression multi-output chaînée (Table 7.3), les performances de régression multi-output directe (Table 7.2) apparaissent similaires, voire légèrement meilleures pour les composantes *lait* et *gras*, bien que l'écart soit très faible. Cela suggère que, dans le contexte de notre jeu de données, l'ap-

proche directe est tout aussi efficace que l'approche chaînée pour capturer les relations entre les variables explicatives et les composantes de la production laitière.

Quant aux Figures 7.6 et 7.7, elles indiquent que la régression multi-output chaînée fournit des performances supérieures à celles de la régression multi-output directe. La raison principale tient au traitement des dépendances entre sorties : la régression multi-output directe considère que chaque variable de sortie est indépendante des autres, ce qui n'est pas le cas dans notre étude où les composantes du lait sont fortement corrélées (Figure 4.8). En revanche, la régression multi-output chaînée modélise explicitement ces dépendances en prédisant chaque sortie successivement en fonction des précédentes, permettant ainsi une meilleure capture des relations entre traits et une amélioration des performances globales de prédiction. Ces résultats soulignent l'importance de tenir compte de la structure de dépendance entre sorties dans les modèles multi-output.

Les résultats de la validation croisée, présentés dans la Table 7.4, permettent de comparer trois approches pour un problème multi-sortie sur notre jeu de données. La régression multi-output directe et la version chaînée présentent toutes deux un MAE moyen extrêmement faible (≈ 0.002) et des écarts-types MAE quasi nuls, indiquant des performances très stables sur l'ensemble des plis de la validation croisée. Ces résultats confirment que les deux approches sont adaptées au problème et généralisent correctement sur les données testées.

En revanche, le modèle basé sur un arbre de décision affiche un MAE moyen nettement plus élevé (30.008) avec un écart-type important (1.153), ce qui indique une forte variabilité des performances et une précision globalement très insuffisante. Cette contre-performance peut s'expliquer par la nature des arbres de décision, qui ont tendance à surajuster les données d'apprentissage lorsque celles-ci sont continues et multidimensionnelles, et à mal généraliser en présence de relations complexes et continues entre les variables.

Ainsi, la comparaison met clairement en évidence la supériorité des approches de régression multi-output, qu'elles soient directes ou chaînées, sur des modèles arborescents pour ce type de tâche. Ce résultat peut s'expliquer par leur capacité à exploiter les relations de dépendance existantes entre les différentes composantes du lait (lait, protéine et gras).

En pratique, la régression directe constitue un choix robuste et peu coûteux computationnellement en raison de sa simplicité de mise en œuvre et de ses performances comparables à celles de la version chaînée,

tout en évitant l'effet cumulatif d'erreurs qui peut se produire lors de la prédiction séquentielle des sorties.

7.4.4 Conclusion

Les analyses comparatives menées montrent clairement que les modèles de régression multi-output, qu'ils soient directs ou chaînés, surpassent largement les arbres de décision pour la prédiction des composantes de la production laitière. Les arbres de décision, bien que simples et interprétables, peinent à capturer la variabilité des traits sensibles en raison de la nature continue et non linéaire des relations entre variables, et présentent une forte variabilité de performances.

La régression multi-output directe et chaînée offre toutes deux des performances stables et précises, avec un MAE moyen extrêmement faible et des écarts-types négligeables lors de la validation croisée. La version chaînée bénéficie d'un avantage théorique en modélisant explicitement les dépendances entre sorties, ce qui améliore la capture des relations corrélées entre les composantes du lait. Cependant, la régression directe constitue un choix robuste et efficace, offrant des performances comparables à la version chaînée tout en limitant la complexité computationnelle et l'accumulation potentielle d'erreurs lors de la prédiction séquentielle.

En pratique, ces résultats soulignent l'importance de choisir des modèles capables de représenter les relations complexes entre variables et de tenir compte des dépendances entre sorties pour optimiser la précision des prédictions multi-sorties dans le contexte de la production laitière. Ils ouvrent également la voie à l'exploration de modèles non linéaires ou d'apprentissage profond, afin de tester leur capacité à capturer d'éventuelles relations plus complexes et à améliorer encore la précision des estimations.

CONCLUSION

Les travaux présentés dans cette thèse ont permis de développer, d'évaluer et de comparer plusieurs modèles statistiques et d'apprentissage profond pour l'estimation des valeurs génétiques et la prévision de la production laitière, en intégrant des données issues de multiples sources (lactation, pedigree, traits de conformation, SNP). L'objectif était de proposer un cadre méthodologique complet pour améliorer la précision des prédictions et fournir des outils d'aide à la décision pour les éleveurs et sélectionneurs.

L'exploitation conjointe de l'*information privilégiée* (PI) et de techniques d'*apprentissage profond* (DL) s'est révélée particulièrement bénéfique. Elle a permis de réduire le coût computationnel, de mieux modéliser les relations complexes et non linéaires entre variables, de capturer les dépendances temporelles et de s'adapter à des données hétérogènes et bruitées, notamment pour les tâches combinées de régression et de prévision.

Plusieurs enseignements clés ont émergé. Le modèle *LSTMDropout* a démontré sa supériorité par rapport aux modèles de référence (ARMA, SARIMA, N-BEATS, TFT) pour des tâches de régression et de prévision, en sortie unique comme en sorties multiples. Les études d'ablation ont mis en évidence l'importance des groupes de traits de conformation, notamment dairy strength, feet & legs et, dans certains cas, body capacity, confirmant leur rôle central dans la prédiction de la production et de la rentabilité. Pour les tâches de prévision, l'architecture feedforward *simple* s'est avérée plus performante que la variante *récurrente*, suggérant qu'un *modèle plus simple favorise une meilleure généralisation et limite le surapprentissage sur des séquences temporelles longues*.

Le modèle *DairyLuPTS* a montré une nette supériorité sur des données strictement temporelles et s'est révélé particulièrement pertinent pour les prédictions sans intégration des traits de conformation.

Le modèle *Information Dropout Adaptée (IDA)*, quant à lui, a produit des performances globalement comparables à celles de l'Information Dropout classique, mais avec un léger gain en prévision lorsque l'information privilégiée est intégrée.

Enfin, le modèle *AgriGen* a permis d'évaluer l'impact des caractères génétiques sur la production laitière, en combinant modélisation VARMAX et classification non supervisée (KMeans et GMM). Bien que les individus

à faible profit aient été répartis dans les deux clusters, rendant les recommandations directes plus difficiles, l'analyse de corrélation a confirmé l'*importance de la maîtrise des coûts alimentaires et de la durée de lactation pour optimiser la rentabilité*.

Ces travaux incluent également le développement du modèle *DairyBLUP*, qui, bien qu'efficace, montre certaines limites pour la modélisation de données multi-lactations et des périodes de tarissement. L'intégration de l'apprentissage profond, et en particulier de LSTMDropout, s'impose comme une solution capable de capturer des relations plus complexes et de mieux gérer les irrégularités et valeurs aberrantes.

Les principales contributions méthodologiques peuvent se résumer comme suit :

Primo, le développement de l'architecture *LSTMDropout*, issu d'un transfert de connaissances des domaines de la vision par ordinateur et du traitement automatique du langage vers la génétique animale, pour la prédiction des VEE, VGE et des composantes de production. *Secundo*, la proposition de l'approche *Information Dropout Adaptée (IDA)*, exploitant l'information privilégiée pour améliorer la précision des prédictions. *Tertio*, la conception du modèle *DairyLuPTS*, tirant parti des séries temporelles comme PI pour optimiser les prévisions. *Quarto*, le développement du modèle *AgriGen*, destiné à évaluer l'impact des caractères génétiques sur la production laitière. Enfin, quinto, la mise en place et l'évaluation de *DairyBLUP* pour l'estimation des valeurs génétiques en contexte multi-lactations.

Ces contributions démontrent l'efficacité de l'intégration de données multi-sources et de l'apprentissage profond pour améliorer la précision et la robustesse des modèles de sélection génétique. Elles constituent un pas important vers des outils plus fiables et exploitables en pratique pour la gestion des élevages et l'optimisation de la rentabilité.

Perspectives et recommandations : Plusieurs axes de recherche méritent d'être explorés pour prolonger ce travail. D'abord, l'*augmentation des données* par des stratégies de génération de SNP synthétiques permettrait d'enrichir les ensembles d'apprentissage, à condition de garantir la cohérence biologique et statistique des génotypes produits. Ensuite, une analyse des interactions (effets génotype × environnement, épistasie, variants et loci) approfondirait la compréhension des déterminants de la production et de la santé animale.

Une *exploration avancée du paradigme de l'information privilégiée* pourrait consister à formaliser l'utilisation des premiers pas temporels comme X_{base} (Baseline+) ou à tester leur intégration comme PI pour diffé-

rents horizons de prévision. Enfin, une évaluation systématique et une généralisation des modèles proposés, en particulier LSTMDropout et DairyLuPTS (avec Baseline+), à d'autres séries temporelles et contextes, constituerait un pas supplémentaire vers des solutions robustes et généralisables.

Ces perspectives visent à approfondir la compréhension des relations entre génotype, phénotype et environnement, tout en renforçant la capacité de généralisation des modèles prédictifs et en soutenant des décisions plus éclairées en matière de sélection génétique et de gestion des troupeaux.

ANNEXE A

MILK PREDICTION USING CONFORMATION TRAITS AS PRIVILEGED INFORMATION

ABSTRACT

Predicting milk production is essential in agriculture, driving advancements in herd management and optimization of animal performance. Traditional approaches include statistical models (autoregressive moving average ARMA, seasonal autoregressive integrated moving average SARIMA) and deep learning (DL) methods such as recurrent networks and multilayer perceptrons. However, these models often address regression or forecasting separately, limiting their adaptability. DL models lack flexibility due to fixed input sizes, while statistical models face scalability issues as population size increases. To overcome these limitations, we propose LSTMDropout, a novel model that leverages training-exclusive features through the learning using privileged information paradigm. Combining long-short-term memory (LSTM) networks with heteroscedastic dropout improves variance estimation and uncertainty modeling, enhancing prediction robustness. Furthermore, we investigate the role of conformation traits as privileged information in predicting single- and multi-output milk yield. In the benchmark dataset, the LSTMDropout model surpasses ARMA, SARIMA, and N-BEATS in regression (errors RMSE : 10.55 and MAE : 8.42) and outperforms the temporal fusion transformer in forecasting, achieving lower errors in both single-output (10.55, 8.42) and multi-output tasks (5.49, 4.48). These results highlight the effectiveness of incorporating privileged information with heteroscedastic dropout, offering accuracy, scalability, and computational efficiency for the dairy industry.
Our code is available at : `https://github.com/AwaSamake/LSTMDropout`

Keywords : Dairy farming ; conformation traits ; time series ; privileged information ; deep learning ; heteroscedastic dropout.

A.1 Introduction

In the domain of dairy farming, prediction of milk production represents a crucial aspect of dairy cattle breeding, facilitating advancements in animal monitoring, herd management, and decision-making for farmers (Frasco. *et al.*, 2020; Vergani *et al.*, 2024). The dairy industry represents a major part of Canada's agricultural sector, ranking second in terms of the economic value of its agricultural income. [1] Several studies have shown that dairy production is influenced by multiple factors, including health, environment, management, and genetics (Sawa *et al.*, 2013; Kadarmideen, 2004; Zhang *et al.*, 2020; Špehar *et al.*, 2022; Cole *et al.*, 2023; Korösi *et al.*, 2024). In order to improve the profitability of dairy production and the living conditions of the animals, models have been developed that help predict the income (value) or yields of dairy farms (Frasco. *et al.*, 2020; Liseune *et al.*, 2021; Salamone *et al.*, 2022; Braginets *et al.*, 2022; Li *et al.*, 2023), determine the impact of the necessary nutritional and energy consumption (Špehar *et al.*, 2022; Cole *et al.*, 2023; Lee *et al.*, 2020; Parra *et al.*, 2021), improve disease detection (King *et al.*, 2018; Franceschini

1. `https://agriculture.canada.ca/en/sector/animal-industry/canadian-dairy-information-centre/dairy-industry`

et al., 2022), optimize the breeding decisions (Zhang *et al.*, 2022; Vergani *et al.*, 2024), and genetic predictions (Short et Lawlor, 1992; Kadarmideen, 2004; Khansefid *et al.*, 2021; Korösi *et al.*, 2024).The models employed in these studies consist of linear regression, traditional statistical techniques (ARMA, ARIMA, VARMA), machine learning models (decision tree, random forest), and deep learning models (ANN, multilayer perceptron, LSTMs). Linear regression-based methods assume a linear relationship between features, which represents a simplistic assumption. Moreover, the complexity of statistical methods increases as the population size grows. Deep learning models, on the other hand, require identical input sizes for both training and testing. In our case, conformation traits are only available at the beginning of the first lactation.

Our study aims to develop a model able to predict future lactation curves using historical lactation data and phenotypic information, such as estimated breeding values (EBVs) and conformation traits (or body measurements) while minimizing computational expenses. More specifically, we predict two sets of target variables depending on the type of task : either a single-output or a multi-output formulation. In the single output case, the target variable corresponds to the daily milk value (**DMV**). In contrast, for the multi-output setting, the target variables include the daily yields of milk (**DMY**), fat (**DFY**) and protein (**DPY**). We also address two problems : regression (Venkatesh *et al.*, 2023) and forecasting (Bojer, 2022).

To achieve this, we propose a new architecture using conformation traits available only during training to counter future model overfitting. We employ the learning paradigm using *privileged information* (**LuPI**) (Vapnik *et al.*, 2009), which leverages information available exclusively during training. The application of this paradigm to temporal data forecasting remains relatively underexplored. With this paradigm, high-performance models can be trained by exploiting data available only at the beginning of the first lactation. The incorporation of this kind of information exclusively during the training phase not only enhances model performance but also contributes to reducing the computational resource requirements in terms of parameters. We extend this framework by integrating heteroscedastic dropout (Lambert *et al.*, 2018) and by using privileged information to estimate variance within Long Short-Term Models (LSTMs) neural networks. Our results demonstrate that the proposed LSTMDropout model significantly outperforms traditional approaches in multi-output forecasting scenarios, compared to single-output models. These findings highlight the potential for enhanced predictive accuracy in dairy farming through the incorporation of privileged information and heteroscedastic dropout in predictive modeling.

The structure of the paper is as follows. Section A.2 provides a review of existing research on regression and

forecasting problems, and introduces the concept of privileged information. In Section A.3, we outline our proposed methodology. The details of the dataset, experimental setup, and the performance evaluation of our model are provided in Sections A.4 and A.5. Finally, Section A.6 conclude the paper by summarizing our findings and discussing their implications.

A.2 Related works

In this section, we review existing studies and categorize them into two main approaches : regression and forecasting methods. Additionally, we explore previous research that applies the LuPI paradigm.

A.2.1 Regression methods

Heinrichs *et al*. (1992) (Heinrichs *et al*., 1992) utilized regression analysis to derive optimal prediction equations by incorporating various body measurements, including heart girth, body length, hip width, and wither height. This approach established a framework for predictive modeling in dairy cattle research. Furthermore, Dallago *et al*. (2019) (Dallago *et al*., 2019) identified key Dairy Herd Improvement (DHI) metrics that significantly impact the first test day milk yield of first lactation dairy cows. They developed and compared three predictive models : multivariate linear regression, random forest, and artificial neural networks. Salamone *et al*. (2022) (Salamone *et al*., 2022) developed a random forest methodology to predict the first test day milk yield after calving using information from the previous lactation.

A.2.2 Forecasting methods

Recent research has emphasized the integration of advanced technologies and data analytics to improve both accuracy and efficiency in predicting milk yields. Traditional lactation models typically forecast milk production based on an initial fixed quantity of milk at the start of the lactation period. Liseune *et al*. (2021) (Liseune *et al*., 2021) introduced a deep learning framework designed to encode the input of the model, predict latent representations of milk yield sequences, and generate the corresponding lactation curves for dairy cows. This approach leverages historical milk yield data from previous lactation cycles to enhance predictive performance. The authors of (Frasco. *et al*., 2020) have proposed a methodology for the design of extensive predictive models that reflect a wider range of factors, centered around the LSTM neural network : univariate and multivariate models. Braginets *et al*. (2022) (Braginets *et al*., 2022) developed a neural network model to predict milk production and fat content by incorporating ancestral lineage data, including

information from mothers, grandmothers, and great-grandmothers. Furthermore, Li *et al.* (2023) (Li *et al.*, 2023) applied convolutional neural network-based models to forecast raw milk prices based on consumer behavior. Franceschini *et al.* (2022) (Franceschini *et al.*, 2022) used Fourier transform mid-infrared (FT-MIR) spectroscopy to explore whether unsupervised learning methods could offer new insights into dairy cow health (Vergani *et al.*, 2024).

A.2.3 Learning using Privileged Information LuPI

The current paradigm in machine learning follows a straightforward approach : given a set of learning examples, the goal is to identify a function from a predefined set that best approximates the unknown decision rule. In human learning, the teacher plays a crucial role by providing additional insights such as explanations, comments, comparisons, and metaphors, collectively referred to as *privileged information* (**PI**). A central aspect of this paradigm is that privileged information is only available during the training phase when the teacher interacts with the learner, and not during the testing phase when the learner operates independently (Vapnik, 1982). It is important to recognize that PI is relevant to nearly all learning problems and can significantly accelerate the learning process. Vapnik *et al.* introduced the *LuPI* paradigm (Vapnik, 1982; Vapnik et Izmailov, 2015b), initially applied to the support vector machine SVM$+$ method. Recent studies (Karlsson *et al.*, 2021; Jung et Johansson, 2022; Hayashi *et al.*, 2019) have demonstrated the utility of LuPI for enhancing early predictions using time-series data. This approach utilizes observed data between the time of prediction and the future outcome as privileged information. Hayashi *et al.* (2019) (Hayashi *et al.*, 2019) applied this method to both synthetic datasets and real-world atmospheric datasets with eight variables. Karlsson *et al.* (2021) (Karlsson *et al.*, 2021) formalized the framework for time series under the term *learning using privileged time-series* (LuPTS), proving that learning efficiency is improved when time series are drawn from a *non-stationary Gaussian-linear dynamical system*. For time series prediction, distillation has emerged as a standard technique for leveraging PI (Hayashi *et al.*, 2019; Karlsson *et al.*, 2021; Jung et Johansson, 2022). Lambert *et al.* (2018) (Lambert *et al.*, 2018) proposed a novel LuPI algorithm tailored for convolutional neural networks (CNNs) and recurrent neural networks (RNNs). They employed pre-trained models (such as VGG networks) with dropout layers, utilizing the *information bottleneck re-parametrization trick*. To manage dropout variance, they suggested computing it as a function of PI, leading to what is termed heteroscedastic dropout. This approach uses PI to control uncertainty in the model output, significantly enhancing sample efficiency and accuracy, even with limited training data. Their method was tested on CNNs and RNNs for *multi-modal machine translation* and *image classification*.

A.3 Methodology

A.3.1 Problem settings

The problem of predicting lifetime milk production, i.e. DMY, DFY, DPY and DMV, in dairy farming involves longitudinal data collection in lactation cycles over time. In this prediction task, we are dealing with multivariate and multiple time series.

Multivariate aspect : This refers to the inclusion of multiple variables beyond the primary output of interest. These variables may have significant effects on the forecasting problem.

Multiple individuals : The dataset includes temporal data for millions of individuals, each represented by a time series. Our dataset contains records collected over multiple lactation cycles for millions of animals. During each lactation cycle, data collection occurs ten (10) times a year.

Our objective is *to predict future lactation performance using historical data from previous lactation while incorporating PI*, such as EBVs and conformation traits (other phenotypic data). In contrast to (Karlsson *et al.*, 2021; Jung et Johansson, 2022; Hayashi *et al.*, 2019), we define our PI in terms of features in our study.

A.3.2 LSTMDropout

In the context of learning hidden representations, the bottleneck problem is a challenge particularly when stacking multiple layers. The information bottleneck (IB) issue arises as intermediate layers struggle to retain relevant information while discarding noise. The most commonly adopted solution in the literature is the introduction of attention mechanisms which have proven effective across various domains. When learning time series with missing intermediate information during inference, discarding the missing field is the default option. In this work, we propose an alternative approach by leveraging LuPI with heteroscedastic dropout, as introduced by Lambert *et al.* (2018) (Lambert *et al.*, 2018). This method combines the IB framework with privileged information (PI) to improve predictive accuracy, replacing attention mechanisms with a customized dropout strategy. Compared to attention-based models such as the Temporal Fusion Transformer (TFT) (Lim *et al.*, 2021), our method avoids the computational overhead associated with attention mechanisms, making it suitable for large-scale datasets such as those used in dairy farming.

Our architecture, illustrated in Figure A.1, builds on the LuPI paradigm (Vapnik *et al.*, 2009; Lambert *et al.*,

2018) and consists of three main steps :

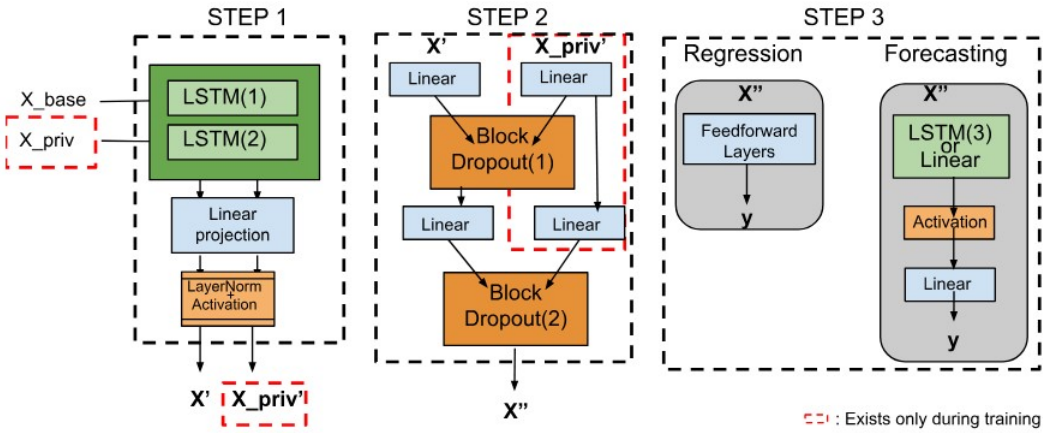


Figure A.1 – General architecture of our model.

A.3.2.0.1 *Handling dimensionality differences*

The first step addresses the dimensional mismatch between the baseline data $X_{base} \in \mathbb{R}^{n_{samples}, n_{seqlength}, n_{features_{base}}}$ and privileged data

$X_{priv} \in \mathbb{R}^{n_{samples}, n_{seqlength}, n_{features_{priv}}}$. Separate LSTM blocks are employed to learn representations from these two inputs. Once transformed, these representations are passed through a linear projection layer, a normalization layer, and a ReLU activation layer, ensuring that both inputs are aligned for subsequent processing. Unlike Lambert *et al.* (2018) (Lambert *et al.*, 2018), where X_{base} and X_{priv} are derived from the same input (respectively a full image and cropped regions), our approach handles heterogeneous inputs, time series data (X_{base}) and phenotypic traits (conformation traits & EBVs, X_{priv}). This adaptation adds complexity but ensures that each type of data is processed optimally.

A.3.2.0.2 *Estimating variance using heteroscedastic dropout*

This second step involves estimating the variance of the heteroscedastic dropout, proposed by Lambert *et al.* (2018) (Lambert *et al.*, 2018), from the two previously learned representations. Note that in the reference paper (Lambert *et al.*, 2018), they extract the baseline and privileged information from the same input, so they do not have to deal with the dimensionality problem encountered in the first step. To calculate the variance of the heteroscedastic dropout during training, we apply two distinct linear layers to the two representations learned in the first step. These learned representations will be the inputs for the third step.

Note that during testing, we will only have access to the X_{base} representations as inputs in the second step. We will duplicate these representations before applying the linear layers and the block dropout.

A.3.2.0.3 *Final step*

The third and final step is common to both training and testing. However, it varies depending on the nature of the task :

• *Regression*. We use feed-forward layers and output a single value.

•*Forecasting*. First we apply either a linear layer (feed-forward simple) or an LSTM layer (feed-forward RNN). At the end, to obtain the final desired output dimension we apply another linear layer.

In the following sections, we present the experimental settings and evaluate the performance of our LSTM-Dropout model using real-world milk production data.

A.4 Experimental settings

A.4.1 Data and preprocessing

Our experiments utilize dairy production datasets and phenotypic dataset (e.g. conformation traits/body measurements) provided by Lactanet (Frasco. *et al.*, 2020). These datasets cover production records for more than a million Holstein cows. Our focus is specifically on cows that have completed at least three lactation cycles, resulting in a dataset of $11,000$ individuals with 100,000 rows and 58 features. Each selected individual has at least three lactation cycles, with one cycle corresponding to 10 data collections per year (10 months per year). The features of the dataset are categorized into several main components.

- **Animal** : Contains information about individual animals including birthdate, entry and exit dates from the herd, breed, country, as well as details about their parents such as breed and country. Number of individuals = $2,055,052$ & Rows = $2,245,954$ & Features = 22.
- **Production** : Provides details on lactation cycles, including the quantity of milk, protein and fat on test day, lactose, number of lactation cycles per cow, test dates, daily milk values, daily feed costs and health characteristics such as somatic cell count, β-hydroxybutyrate (BhB). Number of individuals = $1,471,312$ & Rows = $35,976,274$ & Features = 24. Figure A.2 shows the different lactation cycles of a cow. As can be seen, a cow can reach a very high peak during lactation and then see her production

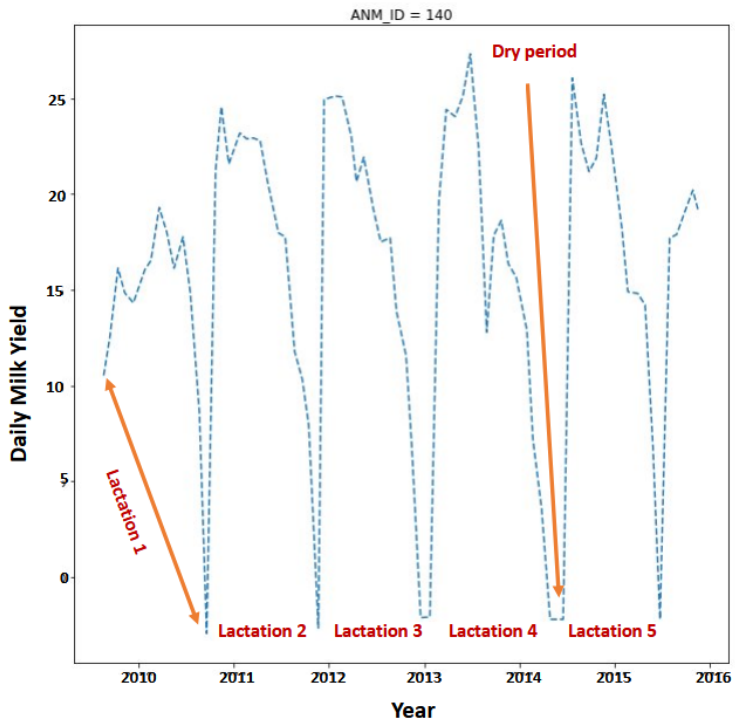


Figure A.2 – Different lactation cycles of cow 140.

levels gradually decline at 305 days in milk (i.e. 10 months). We can also see that the curve differs from one lactation to the next for the same cow.

— **Phenotypes** : Includes information on phenotypic markers, taken before the first lactation cycle, such as conformation traits (or body measurements) (e.g. mammary measures, chest width, body depth, loin strength, rump angle) and estimated breeding values (EBVs). Number of individuals = $1,035,532$ & Rows = $1,035,532$ (1 per individual) & Features = 42. In Table 4.9, we can find more details on our phenotypic features.

The dataset *Production* has been structured as a time series collected between 2006 and 2017, enabling a comprehensive analysis in various dimensions of dairy production and health management. After selecting cows that had completed their first three lactations from the *Production* dataset, we merged the datasets based on animal IDs. The final dataset was split into 60% training, 20% test and 20% validation sets.

Based on their relevance as established by expert knowledge in the dairy production domain, we then deleted the features with more than 50% missing values. Other missing values were filled in for numerical features by global mean (like daily milk value) or local mean (per individual's records, like daily milk yield) ; and for categorical features by the mode. Finally, we proceeded to standardize the data.

Note that *Production* & *Animal* datasets is X_{base}, and *phenotype* one (EBVs and conformation traits) is X_{priv}.

Table A.1 – Repartition of Conformation Traits

Names	Conformation Traits
Estimated Breeding Values EBVs	EBV Milk ; EBV Protein ; EBV Fat ; EBV Protein Percent ; EBV Fat Percent
Body Capacity	Stature ; Height at Front End ; Body Condition BCS ; MR ; Chest Width ; Body Depth ; Loin Strength
Rump	Rump Angle ; Pin Width ; Angularity
Feet & Legs	Hock Lesion HL ; Foot Angle ; Heel Depth ; Bone Quality ; Rear Leg Side View ; Rear Leg Rear View ; Thurl Place Area over the Pelvis
Udder (Mammary)	Udder Depth ; Udder Texture ; Teat Length ; Fore Attach ; Fore Teat Place ; Rear Teat Place ; Median Susp ; Rear Attach Height ; Rear Attach Width
Dairy Strength	Lactase Persistence LP ; Minimum Support Price MSP ; MT ; Daugther fertility ; Calving Ability ; Daily Cow Average ; Multi-Drug Resistant
Others	Inbreeding ; Rvalue ; Somatic Cell Score (SCS)

A.4.2 Baselines models

We benchmark our method against several baseline models, categorized into single-output and multi-output problems :

Single-output problems : ARMA (Contreras *et al.*, 2003) : An autoregressive moving average model that handles stationary time series data. ARMA combines autoregressive (AR) and moving average (MA) components.**SARIMA** : Seasonal ARIMA, which extends ARIMA, autoregressive integrated moving average model, to account for seasonal patterns in time series data. **NBEATS** (Oreshkin *et al.*, 2020) : An univariate and single-output model designed for time series forecasting. **MuMu+Attention** : An extension of the MuMu model (Frasco. *et al.*, 2020) incorporating an attention mechanism (Naghashi *et al.*, 2023).

Multi-Output problems : Information Dropout (Achille et Soatto, 2018) : Originally a regularization method involving multiplicative noise injection in neural network activations. Adapted for forecasting tasks, we modify its architecture to integrate the concept of privileged information $\mathbf{X}_{priv}$ during training, distinguishing it as **Information Dropout (with privileged information)**. This model is referred to as **Information Dropout new** in Table A.3. **Temporal Fusion Transformer** (TFT) (Lim *et al.*, 2021) – an attention-based architecture with interpretable insights into temporal dynamics.

A.4.3 Evaluation metrics and hyperparameter tuning

A.4.3.0.1 *Evaluation metrics*

In general, the evaluation of regression or forecasting models is based on standard metrics such as root mean squared error (RMSE) and mean absolute error (MAE). *Root Mean Squared Error (RMSE)* : Measures the average magnitude of the error. *Mean Absolute Error (MAE)* : Provides the average absolute difference between predicted and actual values. MAE and RMSE are complementary metrics. MAE is more robust to outliers, while RMSE is more sensitive to outliers or large deviations. While MAE takes into account the mean of the errors, RMSE takes into account their variance. Moreover, RMSE is more relevant in situations where larger errors have a greater impact and require minimization, as in our case (regression and forecasting tasks).

Table A.2 – Hyperparameters's tuning.

Hyperparameters	Values
Number of epochs	{ 25 ; 50 ; 100 ; 250 }
Batch size	{ 10 ; 30 ; 50 ; 100 ; 250 }
Learning rate	{ 10^{-2} ; 10^{-3} ; 10^{-4} ; 10^{-5} }
Activation	{ReLU, Tanh, SoftSign }
Optimizer	Adam
Early stopping	patience $\in$ { 5 ; 8 ; 10 }
	criterion = loss
Scheduler	patience $\in$ { 5 ; 8 ; 10 }
	decrease factor
Hidden size	{64 ; 128 ; 512 ; 1024 }
Context size	{16 ; 32 ; 64 ; 128 ; 512 }
FC size	{64 ; 128 ; 512 ; 1024 }
Number of layers	{ 5 ; 8 ; 10 }
Dropout rate	{ 0.15 ; 0.2 ; 0.3 }

A.4.3.0.2 *Hyperparameter tuning*

To optimize the hyperparameters (Table A.2), we employ a grid search algorithm. The learning rate is set to 10^{-3}, with a scheduler that allows up to five (5) epochs of non-decreasing validation error before applying a decay factor of 5×10^{-3}. To mitigate overfitting, we implement *early stopping* with a patience of 10 based on a loss criterion. Additionally, we utilize the ReLU activation function and the Adam optimizer.

A.5 Results

To evaluate the performance of our *LSTMDropout* and *information dropout with privileged information*, we evaluated it across two distinct types of problems : regression and forecasting. Each type further comprises specific tasks categorized into single-output and multi-output scenarios.

— Targets $\mathbf{y}$ of **single-output** task : *daily milk value* (**DMV**). In dairy farming, the DMV represents the average dollar value of the milk components.

— Targets $\mathbf{y}$ of **multi-output** task : daily yields of *milk* (**DMY**), *fat* (**DFY**), and *protein* (**DPY**).

Task	Models	Single		Multi-output	
		RMSE	MAE	RMSE	MAE
Regression	LSTMDropout [M]	**10.5516**	**8.4237**	8.6232	6.9874
	Information Dropout [M]	10.8508	8.6177	—	—
	Information Dropout new [M]	10.8483	8.6234	—	—
	ARMA(2,2) [U]	21.0400	18.0087	—	—
	SARIMA(1,0,1,10) [M]	21.0392	18.0120	—	—
	NBEATS [U]	21.7107	18.7554	—	—
Forecasting	LSTMDropout [RNN]	**10.5517**	**8.4237**	**5.4873**	**4.4752**
	LSTMDropout (without PI) [RNN]	11.2615	9.6288	11.5358	10.4416
	LSTMDropout [FF]	10.6870	8.5263	**5.4832**	**4.4626**
	LSTMDropout (without PI) [FF]	11.2615	9.6288	11.6316	10.4691
	Information Dropout [RNN]	10.8508	8.6177	7.3014	6.4017
	Information Dropout [FF]	10.9178	8.6238	5.8689	4.8780
	Information Dropout new [RNN]	10.8483	8.6234	7.3918	6.3502
	Information Dropout new [FF]	10.9012	8.6179	5.7492	4.7723
	Temporal Fusion Transformer (TFT)	14.9444	10.3264	12.4981	11.1652

Table A.3 – Comparison of our results. Regression versus forecasting tasks. **M** = multivariate models. **U** = univariate models. **RNN** = feed-forward RNN. **FF** = feed-forward simple.

A.5.0.0.1 ■ *Regression task*

In Table A.3, we present our results compared to NBEATS, ARMA (2,2), SARIMA (1,0,1,11), MuMu + Attention, Information Dropout (Achille et Soatto, 2018), for *single output* regression tasks. It is important to note that in this case, the baseline MuMu+Attention model has access to EBVs and conformation traits data during both training and testing. To determine the parameters of the ARMA and SARIMA models, we used the ACF (autocorrelation function) and PACF (partial autocorrelation function). Additionally, we examined the stationarity of our time series, confirming that they are stationary. We observe that our LSTMDropout model has the smallest RMSE and MAE errors. This shows that privileged information, in particular EBVs and conformation traits (X_{priv}), has a positive effect. As a result, their incorporation during training improves the performance of prediction models (see Table A.3). In other words, EBVs and conformation traits (X_{priv}) influence the daily value of milk. This is also confirmed by the work (Frasco. *et al.*, 2020; Atkins, 2018). Integrating EBVs and conformation traits only during training improves both the prediction performance of

our regression model and the reduction of learning resources. On the other hand, the model *information dropout with privileged information* shows performance similar to the original model in (Achille et Soatto, 2018). This suggests that integrating privileged information with this model does not add significant value here.

Let us now analyze the LSTMDropout model for single- and multi-output forecasting tasks. We will compare the performance of our two feedforwards, simple and RNN.

A.5.0.0.2 ■ *Forecasting task*

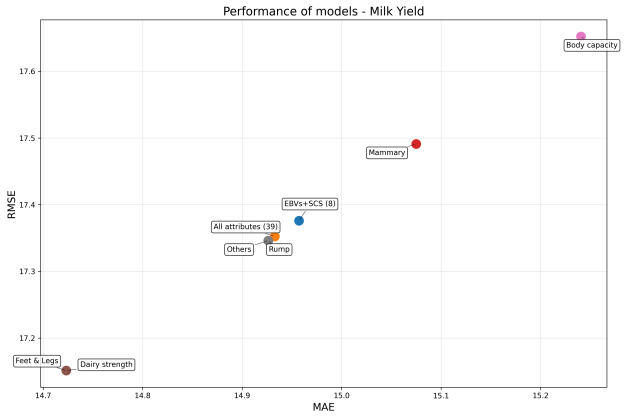
For multi-output forecasting (see Table A.3), the results show that our LSTMDropout model performs better than baselines regardless of the type of decoder chosen, RNN feedforward or simple feedforward. We also observe that our model outperforms for multi-output problems compared with single-output problems. This can be explained by the important role played by certain production traits, such as DMY, DMF, DMP, and somatic cell count, on DMV and in assessing a cow's production performance. The comparaison between LSTMDropout and LSTMDropout (without PI) shows that the use of PI has real impact on the performance.

Table A.3 also demonstrates that our LSTMDropout model with a RNN feedforward outperforms the one with a simple feedforward architecture for single-output forecasting. The model *information dropout with privileged information*, when utilizing a simple feed-forward architecture, outperforms the approach in (Achille et Soatto, 2018) for multi-output forecasting.

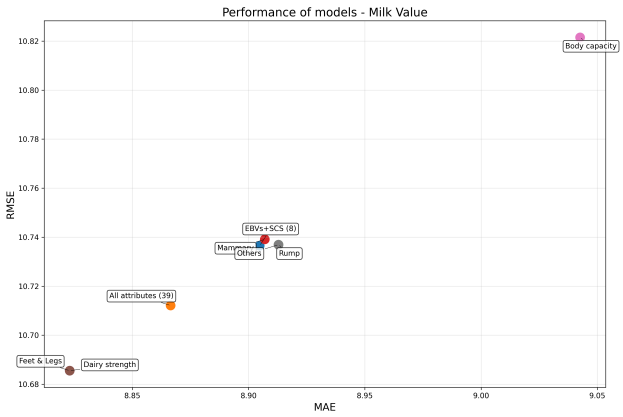
A.5.0.0.3 ■ *Ablation studies*

Given the demonstrated efficiency of our LSTMDropout model in forecasting tasks, we focus our ablation study exclusively on this architecture. This analysis aims to evaluate the impact of input features on the performance of LSTMDropout. In Table A.1, we can find more details on our features.

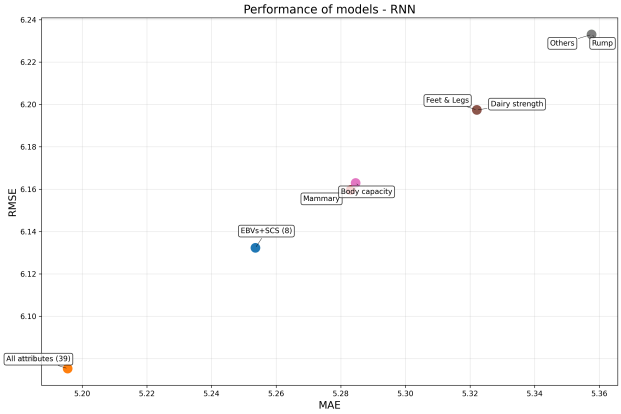
Figure A.3 highlights that the features *dairy strength* and *feet & legs* significantly enhance the performance in single-output forecasting tasks. Furthermore, the study reveals that selecting *body capacity* and *all attributes (39)* as privileged information yields improved performance in multi-output forecasting tasks, particularly with simple and RNN feedforward architectures, respectively. These findings underscore the importance of feature selection and the role of privileged information in optimizing model performance.



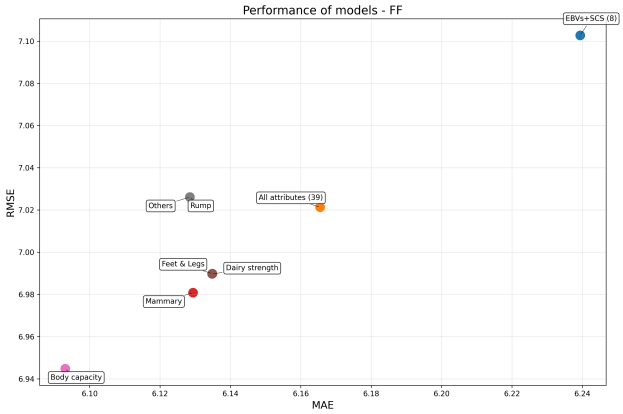
(a) Single-output forecasting - LSTMDropout : target = daily milk yield (DMY).



(b) Single-output forecasting - LSTMDropout : target = daily milk value (DMV).



(c) Multi-output forecasting - LSTMDropout [RNN] : targets = daily yields of milk, fat & protein.



(d) Multi-output forecasting - LSTMDropout [FF] : targets = daily yields of milk, fat & protein.

Figure A.3 – Ablation study evaluating the impact of different partitions of conformation traits as PI.

.

A.6 Conclusion

Integrating privileged information such as estimated breeding values (EBVs) and phenotypic data (conformation traits) significantly improves the performance of forecasting and regression models in dairy farming applications. Estimating the variance using heteroscedastic dropout from privileged information represents a promising approach to improve prediction accuracy in time series analysis. This innovative methodology not only improves predictive models, but also underscores the potential to advance dairy farming practices through data-driven insights. Future research will focus on conducting a rigorous theoretical analysis of the proposed algorithms. Additionally, we plan to generalize the LSTMDropout model for broader applications

across various types of time series data. Our upcoming efforts will include benchmarking the model against existing methods that leverage privileged information for regression & forecasting tasks.

BIBLIOGRAPHIE

Achille, A. et Soatto, S. (2018). Information dropout : Learning optimal representations through noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12).

Agriculture and Agri-Food Canada (2025). *Overview of Canada's Dairy Industry : 2024 Highlights*. Rapport technique, Government of Canada

Ahlborn-Breier, G. et Hohenboken, W. (1991). Additive and nonadditive genetic effects on milk production in dairy cattle : Evidence for major individual heterosis. *Journal of Dairy Science*, *74*(2), 592–602. http://dx.doi.org/https://doi.org/10.3168/jds.S0022-0302(91)78206-4

Atkins, G. (2018). How is functional conformation affecting your herd's longevity and profitability ? *AgProud*.

Bidanel, J. P. (1998). Benefits and limits of increasing sophisticated models for genetic evaluation : the example of pig breeding. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, NSW, Australia, 11-16 January, 1998. Volume 25 : Lactation ; growth and efficiency ; meat quality ; role of exotic breeds in the tropics ; design of village breeding programs ; non-conventional species ; breeding objectives ; design of breeding objectives ; estimation of genetic parameters ; prediction of breeding values*, 577–584.

Bojer, C. S. (2022). Understanding machine learning-based forecasting methods : A decomposition framework and research opportunities. *International Journal of Forecasting*, *38*(4), 1555–1561. Special Issue : M5 competition.

Borchani, H., Varando, G., Bielza, C. et Larranaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, *5*. http://dx.doi.org/10.1002/widm.1157

Bourgeois, P. et Robinson, T. R. (2021). *La Génétique pour les Nuls* (2 éd.). Paris : First Éditions.

Boussaada, Z., Curea, O., Ahmed, R., Camblong, H. et Najiba, m. b. (2018). A nonlinear autoregressive exogenous (narx) neural network model for the prediction of the daily direct solar radiation. *Energies*, *11*, 620. http://dx.doi.org/10.3390/en11030620

Boutemine, O. (2016). Une nouvelle approche de détection de communautés dans les réseaux multidimensionnels.

Braginets, S. A., Galanina, O. V. et Grachev, V. S. (2022). *Neural Networks and Artificial Intelligence in Stockbreeding and Forecasting Dairy Cattle Productivity*, Dans *AgroTech : AI, Big Data, IoT*, (p. 199–207). Springer Nature Singapore : Singapore

Buaban, S., Prempree, S., Sumreddee, P., Duangjinda, M. et Masuda, Y. (2021). Genomic prediction of milk-production traits and somatic cell score using single-step genomic best linear unbiased predictor with random regression test-day model in thai dairy cattle. *Journal of Dairy Science*. http://dx.doi.org/10.3168/jds.2021-20263

Carena, M. J., Hallauer, A. R. et Filho, J. B. M. (1981). Quantitative genetics in maize breeding. Dans *Quantitative Genetics in Maize Breeding*.

Charbuty, B. et Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. http://dx.doi.org/10.38094/jastt20165

Cheng, F., Li, T., Wei, Y.-m. et Fan, T. (2019). The VEC-NAR model for short-term forecasting of oil prices. *Energy Economics*, *78*(C), 656–667. http://dx.doi.org/10.1016/j.eneco.2017.12.0

Chesnais, J. (2007). La génomique, que peut-elle faire pour le producteur laitier ?

CNIEL (n.d.). Dico du lait. https://dico-du-lait.fr/. 2026-04-15.

Cole, J., Makanjuola, B., Rochus, C., van Staaveren, N. et Baes, C. (2023). The effects of breeding and selection on lactation in dairy cattle. *Anim Front*. http://dx.doi.org/10.1093/af/vfad044

Contreras, J., Espinola, R., Nogales, F. et Conejo, A. (2003). Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, *18*(3), 1014–1020.

Da, Y. et Grossman, M. (1991). Multitrait animal model with genetic groups1. *Journal of Dairy Science*, *74*(9), 3183–3195. http://dx.doi.org/10.3168/jds.S0022-0302(91)78504-4

Dairy Farmers of Canada (2025). proaction : On-farm excellence. Récupéré le 2026-06-14 de https://www.dairyfarmers.ca/proaction

Dallago, G. M., de Figueiredo, D. M., de Resende Andrade, P. C., dos Santos, R. A., Lacroix, R., Santschi, D. E. et Lefebvre, D. M. (2019). Predicting first test day milk yield of dairy heifers. *Computers and Electronics in Agriculture*, *166*, 105032. http://dx.doi.org/https://doi.org/10.1016/j.compag.2019.105032

Dempfle, L. (1977). Relation entre blup (best linear unbiased prediction) et estimateur bayésiens. *Annales de Génétique et de Sélection Animale*, *9*. http://dx.doi.org/10.1186/1297-9686-9-1-27

Ducos, A. (1995). Evolutions de l'objectif de sélection dans l'espèce porcine en france. conséquences sur les progrès génétiques attendus. *Revue de Médecine Vétérinaire*, *146*(11), 715–722.

Ducos, A. et Bidanel, J. P. (1993). Use of a multitrait animal model for estimating the genetic trend in six characters of growth, carcasses and meat quality between 1977 and 1990, in the large white and french landrace breeds of pig utilisation d'un modele animal multi-caracteres pour estimer l'evolution genetique de six caracteres de croissance, carcasse et qualite de la viande entre 1977 et 1990, dans les races porcines large white et landrace francais. *Journees de la Recherche Porcine en France*, *25*, 59–64.

Ducrocq, V. (1990). Les techniques d'évaluation génétique des bovins laitiers. *Productions animales*, *3*(1), 3–16.

European Union (2016). General data protection regulation (gdpr), regulation (eu) 2016/679. Récupéré le 2026-04-17 de https://gdpr.eu/tag/gdpr/

Food and Agriculture Organization of the United Nations. (2013). *Milk and Dairy Products in Human Nutrition*. Rome : Food and Agriculture Organization of the United Nations (FAO). Récupéré le 2026-06-12 de https://www.fao.org/3/i3396e/i3396e.pdf

Food and Agriculture Organization of the United Nations, Global Dairy Platform et Global Agenda for Sustainable Livestock. (2018). *Climate Change and the Global Dairy Cattle Sector – The Role of the Dairy Sector in Sustainable Livelihoods*. Rapport technique ISBN 978-92-5-131232-2, Food and Agriculture Organization of the United Nations (FAO) and Global Dairy Platform, Rome

Fournet, F., Elsen, J., Barbieri, M. et Manfredi, E. (1997). Effect of including major gene information in mass selection : A stochastic simulation in a small population. *Genetics Selection Evolution*, *29*. http://dx.doi.org/10.1186/1297-9686-29-1-35

Franceschini, S., Grelet, C., Leblois, J., Gengler, N. et 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. http://dx.doi.org/https://doi.org/10.3168/jds.2022-21975

Frasco., C. G., Radmacher., M., Lacroix., R., Cue., R., Valtchev., P., Robert., C., Boukadoum., M., Sirard., M. et Diallo., A. B. (2020). Towards an effective decision-making system based on cow profitability using deep learning. Dans *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2 : ICAART*, 949–958. INSTICC, SciTePress. http://dx.doi.org/10.5220/0009174809490958

Gauraha, N., Söderdahl, F. et Spjuth, O. (2019). Robust knowledge transfer in learning under privileged information framework. Uppsala University.

Goodfellow, I., Bengio, Y. et Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gouvernement du Québec (2021). Loi modernisant des dispositions législatives en matière de protection des renseignements personnels (loi 25). Récupéré le 2026-04-17 de https://www.quebec.ca/gouvernement/travailler-gouvernement/normes-gouvernance-pratiques-internes/protection-des-renseignements-personnels

Government of Canada (2000). Personal information protection and electronic documents act (pipeda). Récupéré le 2026-04-17 de https://laws-lois.justice.gc.ca/eng/acts/P-8.6/

Hanrahan, L., Geoghegan, A., O'Donovan, M., Griffith, V., Ruelle, E., Wallace, M. et Shalloo, L. (2017). Pasturebase ireland : A grassland decision support system and national database. *Computers and Electronics in Agriculture*, *136*, 193–201. http://dx.doi.org/https://doi.org/10.1016/j.compag.2017.01.029

Hayashi, Takeshi AU Iwata, H. (2013). A bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics*, *14*. http://dx.doi.org/10.1186/1471-2105-14-34

Hayashi, S., Tanimoto, A. et Kashima, H. (2019). Long-term prediction of small time-series data using generalized distillation. *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Heinrichs, A., Rogers, G. et Cooper, J. (1992). Predicting body weight and wither height in holstein heifers using body measurements. *Journal of dairy science*, *75*(12), 3576–3581.

Henderson, C. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics*, *21*, 309–310.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, *32*(1), 69–83.

Henderson, L., Miglior, F., Sewalem, A., Kelton, D., Robinson, A. et Leslie, K. (2011). Estimation of genetic parameters for measures of calf survival in a population of holstein heifer calves from a heifer-raising facility in new york state. *Journal of Dairy Science*, *94*(1), 461–470. http://dx.doi.org/10.3168/jds.2010-3243

Herrera, J. R. V., Flores, E. B., Duijvesteijn, N., Moghaddar, N. et van der Werf, J. H. (2021). Accuracy of genomic prediction for milk production traits in philippine dairy buffaloes. *Frontiers in Genetics*, *12*. http://dx.doi.org/10.3389/fgene.2021.682576

Jung, B. et Johansson, F. D. (2022). Efficient learning of nonlinear prediction models with time-series privileged information. Dans A. H. Oh, A. Agarwal, D. Belgrave, et K. Cho (dir.). *Advances in Neural Information Processing Systems*.

Kadarmideen, H. N. (2004). Genetic correlations among body condition score, somatic cell score, milk production, fertility and conformation traits in dairy cows. *Animal Science*, *79*(2), 191–201.

Karlsson, R., Willbo, M., Hussain, Z. M., Krishnan, R. G., Sontag, D. A. et Johansson, F. D. (2021). Using time-series privileged information for provably efficient learning of prediction models. *CoRR*, *abs/2110.14993*.

Kavlak, A., Strandén, I., Lidauer, M. et Uimari, P. (2021). Estimation of social genetic effects on feeding behaviour and production traits in pigs. *Animal*, *15*(3), 100168. http://dx.doi.org/10.1016/j.animal.2020.100168

Khansefid, M., Haile-Mariam, M. et Pryce, J. (2021). Including milk production, conformation, and functional traits in multivariate models for genetic evaluation of lameness. *Journal of Dairy Science*, *104*(10), 10905–10920.

King, M., LeBlanc, S., Pajor, E., Wright, T. et DeVries, T. (2018). Behavior and productivity of cows milked in automated systems before diagnosis of health disorders in early lactation. *Journal of Dairy Science*, *101*(5), 4343–4356.

Kingma, D. P., Salimans, T. et Welling, M. (2015). Variational dropout and the local reparameterization trick. Dans C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, et R. Garnett (dir.). *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Kingma, D. P. et Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, *abs/1312.6114*.

Koerhuis, A. et Thompson, R. (1997). Model to estimate maternal effects for juvenile body weight in broiler chickens. *Genetics Selection Evolution*, *29*. http://dx.doi.org/10.1186/1297-9686-29-2-225

Kongakura (2026). Les réseaux de neurones récurrents. Récupéré le 2026-04-22 de https://kongakura.fr/article/Les-reseaux-de-neurones-reccurent

Korösi, Z. J., Bognár, L., Bene, S. et Szabó, F. (2024). The role of the conformation of holstein cows in the sustainability of milk production. *Chemical Engineering Transactions*, *114*, 745–750.

Lactanet Canada (2026). Lactanet canada. https://lactanet.ca/. Consulté le 23 juin 2026.

Lambert, J., Sener, O. et Savarese, S. (2018). Deep Learning Under Privileged Information Using Heteroscedastic Dropout . Dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8886–8895., Los Alamitos, CA, USA. IEEE Computer Society. http://dx.doi.org/10.1109/CVPR.2018.00926

Lawson, L., Bruun, J., Coelli, T., Agger, J. et Lund, M. (2004). Relationships of efficiency to reproductive disorders in danish milk production : A stochastic frontier analysis. *Journal of Dairy Science*, *87*(1), 212–224. http://dx.doi.org/10.3168/jds.S0022-0302(04)73160-4

Leclère, F., Kolb, F., Lewbart, G., Casoli, V. et Vögelin, E. (2014). *Un modèle animal simple pour l'apprentissage des techniques de microanastomoses vasculaires de congruences différentes*, volume 22. Spring.

Lee, H.-J., Lee, J. H., Gondro, C., Koh, Y. J. et Lee, S. H. (2023). deepgblup : joint deep learning networks and gblup framework for accurate genomic prediction of complex traits in korean native cattle. *Genetics Selection Evolution*, *55*(56), 1297–9686. http://dx.doi.org/10.1186/s12711-023-00825-y

Lee, Y. M., Dang, C. G., Alam, M. Z., Kim, Y. S., Cho, K. H., Park, K. D. et Kim, J. J. (2020). The effectiveness of genomic selection for milk production traits of holstein dairy cattle. *Asian-Australasian journal of animal sciences*. http://dx.doi.org/10.5713/ajas.19.0546

Lembeye, F., Lopez-Villalobos, N., Burke, J. et Davis, S. (2016). Estimation of genetic parameters for milk traits in cows milked once- or twice-daily in new zealand. *Livestock Science*, *185*, 142–147. http://dx.doi.org/10.1016/j.livsci.2016.01.022

Li, Z., Zuo, A. et Li, C. (2023). Predicting raw milk price based on depth time series features for consumer behavior analysis. *Sustainability*, *15*(8), 6647.

Lim, B., Arık, S. , Loeff, N. et Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748–1764.

Liseune, A., Salamone, M., Van den Poel, D., van Ranst, B. et Hostens, M. (2021). Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning. *Computers and Electronics in Agriculture*, *180*, 105904. http://dx.doi.org/https://doi.org/10.1016/j.compag.2020.105904

Lopez-Paz, D., Bottou, L., Schölkopf, B. et Vapnik, V. N. (2015). Unifying distillation and privileged information. *CoRR*, *abs/1511.03643*.

Lund, T., Miglior, F., Dekkers, J. et Burnside, E. (1994). Genetic relationships between clinical mastitis,

somatic cell count, and udder conformation in danish holsteins. *Livestock Production Science*, *39*(3), 243–251. http://dx.doi.org/10.1016/0301-6226(94)90203-8

Lynch, M. et Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates.

Madsen, P., Jensen, J., Labouriau, R., Christensen, O. F. et Sahana, G. (2014). Dmu – a package for analyzing multivariate mixed models in quantitative genetics and genomics. Dans *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production (WCGALP)*, Vancouver, Canada.

McCarthy, J. (2007). What is artificial intelligence ? Récupéré le 2026-04-17 de https://www-formal.stanford.edu/jmc/whatisai.pdf

Meuwissen, T. H. E., Hayes, B. J. et Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. http://dx.doi.org/10.1093/genetics/157.4.1819

Meyer, K. et Smith, S. (1996). Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genetics Selection Evolution*, *28*, 23–49. http://dx.doi.org/10.1051/gse:19960102

Mota, L. F. M., Giannuzzi, D., Pegolo, S., Trevisi, E., Ajmone-Marsan, P. et Cecchinato, A. (2023). Integrating on-farm and genomic information improves the predictive ability of milk infrared prediction of blood indicators of metabolic disorders in dairy cows. *Genetics Selection Evolution*, *55*(23). http://dx.doi.org/10.1186/s12711-023-00795-1

Motiian, S., Piccirilli, M., Adjeroh, D. A. et Doretto, G. (2016). Information bottleneck learning using privileged information for visual recognition. Dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1496–1505. http://dx.doi.org/10.1109/CVPR.2016.166

Naghashi, V., Dallago, G., Diallo, A. B. et Boukadoum, M. (2023). Univariate and multivariate time-series methods to forecast dairy income. Dans *2nd AAAI Workshop on AI for Agriculture and Food Systems*.

Office of the Privacy Commissioner of Canada (2026). Pipeda in brief. Récupéré le 2026-04-17 de https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/

Oreobike Blog (2023). Disadvantages of highland cattle. Récupéré le 2026-04-17 de https://oreobike.blogg.se/2023/october/disadvantages-of-highland-cattle.html

Oreshkin, B. N., Carpov, D., Chapados, N. et Bengio, Y. (2020). N-BEATS : Neural basis expansion analysis for interpretable time series forecasting. Dans *International Conference on Learning Representations*. http://dx.doi.org/10.48550/ARXIV.1905.10437

Parra, C., Bragatto, J., Piran Filho, F., Silva, S., Tuzzi, B., Jobim, C. et Daniel, J. (2021). Effect of sealing strategy on the feeding value of corn silage for growing dairy heifers. *Journal of Dairy Science*, *104*(6), 6792–6802. http://dx.doi.org/https://doi.org/10.3168/jds.2020-19895

Peixeiro, M. (2022). *Time Series Forecasting in Python*. Manning.

Pryce, J. E., Esslemont, R. J., Thompson, R., Veerkamp, R. F., Kossaibati, M. A. et Simm, G. (1998). Estimation of genetic parameters using health, fertility and production data from a management recording system for dairy cattle. *Animal Science*, *66*(3), 577–584. http://dx.doi.org/10.1017/S1357729800009152

Pérez-Rodríguez, P., Flores-Galarza, S., Vaquera-Huerta, H., del Valle-Paniagua, D. H., Montesinos-López, O. A. et Crossa, J. (2020). Genome-based prediction of bayesian linear and non-linear regression models for ordinal data. *The Plant Genome*, *13*(2), e20021. http://dx.doi.org/https://doi.org/10.1002/tpg2.20021

Ramisarijaona, A. (2026). Long short-term memory (lstm) : de quoi s'agit-il ? Récupéré le 2026-04-22 de https://liora.io/long-short-term-memory-tout-savoir

Rivest, J., Fortin, F., Maignel, L., Riendeau, L., Morin, M., Plourde, N., Richard, Y. et de développement du porc du Québec inc., C. (2008). Estimation du potentiel économique de nouveaux caractères génétiques et développement d'un outil de calcul des valeurs économiques des indices paternel et maternel. *Revue de Médecine Vétérinaire*.

Robinson, G. K. (1991). That BLUP is a Good Thing : The Estimation of Random Effects. *Statistical Science*, *6*(1), 15 – 32. http://dx.doi.org/10.1214/ss/1177011926

Roibas, D. et Alvarez, A. (2010). Impact of genetic progress on the profits of dairy farmers. *Journal of Dairy Science*, *93*(9), 4366–4373. http://dx.doi.org/10.3168/jds.2010-3135

Rokach, L. et Maimon, O. (2005). *Decision Trees*, Dans *Data Mining and Knowledge Discovery Handbook*, (p. 165–192). Springer US : Boston, MA

Ruebel, M. L., Martins, L. R., Schall, P. Z., Pursley, J. R. et Latham, K. E. (2022). Effects of early lactation body condition loss in dairy cows on serum lipid profiles and on oocyte and cumulus cell transcriptomes. *Journal of Dairy Science*, *105*(10), 8470–8484. http://dx.doi.org/10.3168/jds.2022-21919

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Russell, S. et Norvig, P. (2021). *Artificial Intelligence : A Modern Approach* (4 éd.). London : Pearson.

Salamone, M., Adriaens, I., Vervaet, A., Opsomer, G., Atashi, H., Fievez, V., Aernouts, B. et Hostens, M. (2022). Prediction of first test day milk yield using historical records in dairy cows. *animal*, *16*(11), 100658.

Salinas, D., Flunkert, V., Gasthaus, J. et Januschowski, T. (2017). Deepar : Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191. http://dx.doi.org/10.1016/j.ijforecast.2019.07.001

Sawa, A., Bogucki, M., Krężel-Czopek, S. et Neja, W. (2013). Relationship between conformation traits and lifetime production efficiency of cows. *International Scholarly Research Notices*, *2013*(1), 124690.

Schaeffer, L. (1991). C. r. henderson : Contributions to predicting genetic merit. *Journal of Dairy Science*, *74*(11), 4052–4066. http://dx.doi.org/10.3168/jds.S0022-0302(91)78601-3

Shine, P. et Murphy, M. D. (2022). Over 20 years of machine learning applications on dairy farms : A comprehensive mapping study. *Sensors*, *22*(1), 52. http://dx.doi.org/10.3390/s22010052

Shokor, F., Croiseau, P., Gangloff, H., Saintilan, R., Tribout, T., Mary-Huard, T. et Cuyabano, B. (2025). Deep learning and genomic best linear unbiased prediction integration : An approach to identify potential nonlinear genetic relationships between traits. *Journal of Dairy Science*, *108*(6), 6174–6189. http://dx.doi.org/https://doi.org/10.3168/jds.2024-26057

Short, T. et Lawlor, T. (1992). Genetic parameters of conformation traits, milk yield, and herd life in holsteins. *Journal of Dairy Science*, *75*(7), 1987–1998.

Slob, N., Catal, C. et Kassahun, A. (2021). Application of machine learning to improve dairy farm management : A systematic literature review. *Preventive Veterinary Medicine*, *187*, 105237. http://dx.doi.org/https://doi.org/10.1016/j.prevetmed.2020.105237. Récupéré le 2026-04-17 de https://www.sciencedirect.com/science/article/pii/S0167587720309211

Smith, L. (1968). Forecasting annual milk yields. *Agricultural Meteorology*, *5*(3), 209–214. http://dx.doi.org/https://doi.org/10.1016/0002-1571(68)90004-6

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. et Salakhutdinov, R. (2014). Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

Study, P. A. (2023). Box-pierce q-statistic and ljung-box q-statistic.

Tang, F., Xiao, C., Wang, F., Zhou, J. et Lehman, L.-w. H. (2019). Retaining privileged information for multi-task learning. Dans *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & amp ; Data Mining*, KDD '19, p. 1369–1377., New York, NY, USA. Association for Computing Machinery. http://dx.doi.org/10.1145/3292500.3330907

Tishby, N., Pereira, F. C. et Bialek, W. (1999). The information bottleneck method. Dans *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 368–377.

Tribout, T., Bidanel, J. P., Garreau, H., Fleho, J., Gueblez, R., Tiran, M., Ligonesche, B., Lorent, P. et Ducos, A. (1998). Continuous genetic evaluation of on-farm- and station-tested pigs for production and reproduction traits in france. *Journees de la Recherche Porcine en France*, *30*, 95–100.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. et Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *59*. http://dx.doi.org/10.1038/s43586-021-00056-9

Upra-Bovine et LAFLEUR, C. (2017). UPRA NEWS No15 estimated breeding values (ebv). https://www.agripedia.nc/conseils-techniques/productions-animales/conduite-des-elevages-et-qualite/estimated-breeding-values.

Valergakis, G. E., Gelasakis, A. I., Oikonomou, G., Arsenos, G., Fortomaris, P. et Banos, G. (2010a). Profitability of a dairy sheep genetic improvement program using artificial insemination. *Animal*, *4*(10), 1628–1633. http://dx.doi.org/10.1017/S1751731110000832

Valergakis, G. E., Gelasakis, A. I., Oikonomou, G., Arsenos, G., Fortomaris, P. et Banos, G. (2010b). Profitability of a dairy sheep genetic improvement program using artificial insemination. *Animal*, *4*(10), 1628–1633. http://dx.doi.org/10.1017/S1751731110000832

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. http://dx.doi.org/10.3168/jds.2007-0980

Vapnik, V. (1982). *Estimation of dependences based on empirical data : Springer series in statistics*. Berlin, Heidelberg : Springer-Verlag.

Vapnik, V. et Izmailov, R. (2015a). Learning using privileged information : Similarity control and knowledge transfer. *J. Mach. Learn. Res.*, *16*(1), 2023–2049.

Vapnik, V. et Izmailov, R. (2015b). Learning with intelligent teacher : Similarity control and knowledge transfer. Dans A. Gammerman, V. Vovk, et H. Papadopoulos (dir.). *Statistical Learning and Data Sciences*, 3–32., Cham. Springer International Publishing.

Vapnik, V. et Vashist, A. (2009). A new learning paradigm : Learning using privileged information. *Neural Networks*, *22*(5), 544–557. Advances in Neural Networks Research : IJCNN2009.

Vapnik, V., Vashist, A. et Pavlovitch, N. (2009). Learning using hidden information (learning with teacher). Dans *2009 International Joint Conference on Neural Networks*, 3188–3195. IEEE.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

Venkatesh, K., Mohanasundaram, K. et Pothyachi, V. (2023). 8 - regression tasks for machine learning. In T. Goswami et G. Sinha (dir.), *Statistical Modeling in Machine Learning* 133–157. Academic Press.

Vergani, A., Bagnato, A. et Masseroli, M. (2024). Predicting bovine daily milk yield by leveraging genomic breeding values. *Computers and Electronics in Agriculture*, *219*, 108777.

Weersink, A., Howard, W. H., Dekkers, J. C. et Lohuis, M. (1991). Sire selection decisions and farm profitability relationships for ontario dairy fanners. *Journal of Dairy Science*, *74*. http://dx.doi.org/10.3168/jds.S0022-0302(91)78594-9

Wen, R., Torkkola, K., Narayanaswamy, B. et Madeka, D. (2017). A multi-horizon quantile recurrent forecaster. http://dx.doi.org/10.48550/ARXIV.1711.11053

Wu, C., Nikolentzos, G. et Vazirgiannis, M. (2020). Evonet : A neural network for predicting the evolution of dynamic graphs. *CoRR*, *abs/2003.00842*.

Xiu, Y. et Zhang, W. (2017/04). Multivariate chaotic time series prediction based on narx neural networks. Dans *Proceedings of the 2017 2nd International Conference on Electrical, Automation and Mechanical Engineering (EAME 2017)*, 164–167. Atlantis Press. http://dx.doi.org/10.2991/eame-17.2017.40

Xu, D., Shi, Y., Tsang, I. W., Ong, Y., Gong, C. et Shen, X. (2019). A survey on multi-output learning. *CoRR*, *abs/1901.00248*.

Xue, X., Hu, H., Zhang, J., Ma, Y., Han, L., Hao, F., Jiang, Y. et Ma, Y. (2023). Estimation of genetic parameters for conformation traits and milk production traits in chinese holsteins. *Animals*, *13*(1). http://dx.doi.org/10.3390/ani13010100

Yin, X., Han, Y., Sun, H., Xu, Z., Yu, H. et Duan, X. (2020). A multivariate time series prediction schema based on multi-attention in recurrent neural network. Dans *2020 IEEE Symposium on Computers and Communications (ISCC)*, 1–7. http://dx.doi.org/10.1109/ISCC50000.2020.9219721

Zhang, F., Upton, J., Shalloo, L. et Murphy, M. (2019). Effect of parity weighting on milk production forecast models. *Computers and Electronics in Agriculture*, *157*, 589–603. http://dx.doi.org/10.1016/j.compag.2018.12.051

Zhang, F., Upton, J., Shalloo, L., Shine, P. et Murphy, M. D. (2020). Effect of introducing weather parameters on the accuracy of milk production forecast models. *Information Processing in Agriculture*, *7*(1), 120–138. http://dx.doi.org/10.1016/j.inpa.2019.04.004

Zhang, F., Weigel, K. et Cabrera, V. (2022). Predicting daily milk yield for primiparous cows using data of within-herd relatives to capture genotype-by-environment interactions. *Journal of Dairy Science*, *105*(8), 6739–6748.

Zheng, Y., Chen, S., Zhang, X. et Wang, D. (2019). Heterogeneous graph convolutional networks for temporal community detection. *CoRR*, *abs/1909.10248*.

Špehar, M., Ramljak, J. et Kasap, A. (2022). Estimation of genetic parameters and the effect of inbreeding on dairy traits in istrian sheep. *Italian Journal of Animal Science*, *21*(1), 331–342. http://dx.doi.org/10.1080/1828051X.2022.2031320