ROYAUME DU MAROC
Université Mohammed V - RABAT -
جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ÉTUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre : 3995

THÈSE

en vue de l'obtention du : **DOCTORAT**

Centre de Recherche : Biotechnologies Végétales et Microbiennes, Biodiversité et Environnement
Structure de Recherche : Laboratoire de Biodiversité, Écologie, et Génome
Discipline : Biologie
Spécialité : Génomique, Biostatistiques, et Bio-informatique

Présentée et Soutenue le : *19 /10/ 2024*

par :

Narjice CHAFAI

Genetic Evaluation and Genomic Prediction of Zootechnical traits in Moroccan Dairy Cattle Using Bayesian Approaches and Artificial Intelligence Algorithms

Devant le JURY :

Hocein BAZAIRI	PES	Faculté des Sciences, Université Mohammed V de Rabat	Président
Laila SBABOU	PES	Faculté des Sciences, Université Mohammed V de Rabat	Examinatrice/Rapportrice
Mohammed PIRO	PES	Institut Agronomique et Vétérinaire Hassan II, Rabat	Examinateur/Rapporteur
Taniguchi HIRAOKI	PES	Institut de Génétique et de Biotechnologie Animale de l'Académie Polonaise des Sciences, Pologne/ Centre Africain du Génome de l'Université Polytechnique Mohammed VI, Bengrir	Examinateur/Rapporteur
Kaoutar TAHA	PH	Faculté des Sciences, Université Mohammed V de Rabat	Examinatrice/Rapportrice
Bouchra EL AMIRI	Experte	Institut National de la Recherche Scientifique de Settat	Examinatrice
Bouabid BADAOUI	PES	Faculté des Sciences, Université Mohammed V de Rabat	Directeur de thèse

Année Universitaire : 2023 - 24

DEDICATION

To my beloved family, the pillars of my life: my mother, my father, and my dear sister. Your unwavering love, support, and belief in me have been the foundation of every accomplishment, including this work.

To my mother, whose endless love and sacrifices have shaped me into who I am today. Your strength and compassion continue to inspire me every day.

To my father, whose wisdom and encouragement have guided me through every challenge. Your faith in my abilities has given me the courage to pursue my dreams.

To my lovely sister, my confidante and cheerleader. Your laughter, support, and friendship have been my solace throughout this journey.

To all my colleagues in UM5R and UGA, Ichrak, Amanda, Zuleica, Arielly, Marisol, Larissa, Mahsa, Aleja, Jennifer, Mary-Kate, Heegun, Joe, Koushik, Sergio, Evan, and Masum, thank you for your support and encouragement.

To all the post-docs in UGA, Dr. Carrara, Dr. Veroneze, Dr. Bussiman, Dr. Gowane, and Dr. Oliveira, thank you for sharing your knowledge with me. Your support and kindness are highly appreciated.

To all the professors at UGA, Dr. Ignacy Misztal, Dr. Daniela Lourenco, Dr. Jorge Hidalgo. Thank you for teaching me how to think as a geneticist and statistician. Your courses and discussions will be with me for life.

This work is as much yours as it is mine. Thank you for being my constant source of strength, inspiration, and love. I am forever grateful for your presence in my life and for the sacrifices you've made to help me reach this milestone.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the Laboratory of Biodiversity, Ecology, and Genome for the constant support and invaluable opportunities I received throughout my PhD. Working in such a stimulating and collaborative environment has allowed me to develop my research skills and broaden my scientific horizons. I am also thankful to the members of the laboratory for their guidance, insightful advice, and the team spirit that greatly contributed to the completion of my work. It was a privilege to be part of such an exceptional group.

Second, I would like to express my deepest gratitude to my supervisor, Mr. **Bouabid BADAOUI**, Professor of Higher Education (PES), whose guidance and support were instrumental in the completion of this thesis. Prof. BADAOUI not only taught me the fundamentals of computational skills and genomics but also provided unwavering support throughout my thesis work and Fulbright journey. His willingness to introduce me to researchers worldwide has broadened my academic horizons, opening up new opportunities for collaboration and learning. Prof. BADAOUI's dedication to his students' success, combined with his approachable nature, makes him the kind of supervisor any student would be fortunate to have. His impact on my academic and professional development extends far beyond this thesis, and I am profoundly grateful for his mentorship, patience, and encouragement throughout this journey.

Additionally, I would like to express my deepest gratitude to the President of the Jury Mr. **Hocein BAZAIRI**, Professor of Higher Education (PES), for accepting to evaluate this thesis. Your insightful feedback and valuable time have contributed significantly to improving the quality of this work.

Also, I would like to express my gratitude for the members of the Jury and reviewers of the document, Ms. **Laila SBABOU**, Professor of Higher Education (PES), Mr. **Mohammed PIRO**, Professor of Higher Education (PES), Mr. **Taniguchi HIRAOKI**, Professor of Higher Education (PES), and Ms. **Kaoutar TAHA**, Habilitated Professor (PH). Your constructive comments and detailed feedback have been instrumental in enhancing the quality and depth of this research. I deeply appreciate the time and effort you dedicated to reviewing my work, and your contributions have undoubtedly helped shape it into a more rigorous and comprehensive study. Thank you for your invaluable support and guidance.

I would like to extend my sincere thanks to Dr. **Bouchra EL AMIRI** for her invaluable contribution as a jury member during my PhD defense. Her insightful feedback, thoughtful questions, and deep expertise have greatly enriched my work.

I also would like to express my sincere appreciation to my co-supervisor, Professor **Romdhane REKAYA** from the University of Georgia, Athens, for his invaluable guidance and support throughout my thesis journey. Prof. REKAYA's expertise and insights have significantly enriched my research experience and contributed greatly to the depth of this work. His thoughtful feedback, challenging questions, and encouragement to explore new perspectives have been instrumental in shaping both this thesis and my growth as a researcher. Prof. REKAYA's dedication to academic excellence and his willingness to share his knowledge have been truly inspiring. I am grateful for the time he invested in our discussions, his patience in explaining complex concepts, and his continuous support in helping me navigate the challenges of my research. His contribution to my academic development at UGA has been immeasurable, and I feel fortunate to have had the opportunity to work under his co-supervision.

Finally, I would like to express my heartfelt gratitude to my family, whose unwavering support has been a constant source of strength throughout this journey. Your encouragement, love, and belief in me have been the foundation upon which I have built this work. Through every challenge and milestone, you have stood by my side, providing both emotional and practical support. This achievement would not have been possible without your presence, and I am forever grateful for your endless patience and sacrifices. Thank you for always being there for me.

ABSTRACT

This thesis addresses the genetic improvement of dairy cows in Morocco, with a focus on estimating the genetic parameters of production and reproduction traits in Holstein cows. It reveals that the heritability of milk yield is moderate, while that of fertility traits is low, making direct selection for the latter difficult. However, integrating fertility and production traits into a selection index could help mitigate the decline in fertility. For censored data, the threshold model proves effective for imputing phenotypes, enhancing predictive accuracy, which could be a solution for Holstein cow databases in Morocco. The thesis also analyzes the use of artificial intelligence in predicting genetic values, evaluating its effects on accuracy and computational complexity. By comparing several SNP selection methods, including PCA, Gradient Boosting Machines, Ridge regression, and LASSO, GBM demonstrated an advantage in selecting 500 and 1000 SNPs. The accuracy achieved with 1000 SNPs selected by GBM is almost equivalent to that obtained with a full panel of 50,000 SNPs.

Keywords (5) : Genetic parameters, fertility, milk yield, artificial intelligence, SNPs.

RESUMÉ

Cette thèse traite l'amélioration génétique des vaches laitières au Maroc, avec un focus sur l'estimation des paramètres génétiques des traits de production et de reproduction des vaches Holstein. Elle révèle que l'héritabilité du rendement laitier est modérée, tandis que celle des traits de fertilité est faible, rendant la sélection directe pour ces derniers difficile. Toutefois, l'intégration de traits de fertilité et de production dans un indice de sélection pourrait aider à limiter la baisse de la fertilité. Pour les données censurées, le modèle à seuil s'avère efficace pour l'imputation des phénotypes, améliorant la précision prédictive, ce qui pourrait être une solution pour les bases de données des vaches Holstein au Maroc. La thèse analyse également l'utilisation de l'intelligence artificielle dans la prédiction des valeurs génétiques, en évaluant ses effets sur la précision et la complexité computationnelle. En comparant plusieurs méthodes de sélection de SNPs, notamment l'ACP, le Gradient Boosting Machines, la régression Ridge et LASSO, le GBM a démontré un avantage en sélectionnant 500 et 1000 SNPs. La précision obtenue avec 1000 SNPs sélectionnés par GBM est presque équivalente à celle obtenue avec un panel complet de 50 000 SNPs.

Mots-clefs (5) : paramètres génétiques, fertilité, rendement laitier, intelligence artificielle, SNPs.

RESUMÉ DETAILLÉ

Cette thèse traite des défis liés à l'évaluation génétique du rendement laitier et des caractères de fertilité des vaches laitières Holstein, ainsi que des problèmes liés à l'augmentation constante de la dimensionnalité des données de marqueurs génétiques, ce qui accroît la complexité informatique de la sélection génomique. La recherche aborde les défis liés à la sélection génétique de caractères antagonistes, révèle comment le choix des caractères de fertilité à inclure dans les indices de sélection affecte l'efficacité de la sélection génétique, examine la capacité de plusieurs modèles à traiter les données de fertilité censurées sur le terrain, et propose enfin plusieurs modèles d'apprentissage automatique capables de réduire les bases de données de marqueurs sans compromettre la précision de la sélection génomique. Cette dissertation est organisée comme suit:

Introduction

La section d'introduction de cette recherche fournit une vue d'ensemble approfondie qui situe le lecteur dans le contexte actuel de l'industrie laitière marocaine. Elle met en lumière les diverses initiatives lancées par le gouvernement pour stimuler la production laitière et détaille leur impact positif, tout en identifiant les défis actuels qui doivent être relevés pour améliorer davantage le secteur. L'introduction souligne comment l'intégration de la génomique pourrait accélérer le progrès génétique et résoudre les problèmes liés aux évaluations basées sur la généalogie. Cependant, elle reconnaît également l'émergence de nouveaux défis avec la génomique, tels que l'augmentation de la dimensionnalité des jeux de données de marqueurs et le nombre croissant d'animaux génotypés. Enfin, cette section met en avant les avancées remarquables de l'IA dans des domaines tels que la télédétection et le traitement du langage naturel, en soulignant l'importance d'explorer le potentiel de l'IA pour détecter des motifs dans de grands jeux de données génomiques et réduire la charge informatique de la sélection génomique.

Chapitre 1: Revue de la littérature

La revue de littérature de ce chapitre explore l'évolution des pratiques de sélection dans l'élevage animal, retraçant la montée de la génétique quantitative et des méthodologies clés telles que les Moindres Carrés Généralisés (MCG), l'Indice de Sélection et les Équations du Modèle Mixte de Henderson (MME), ainsi que la Meilleure Prédiction Linéaire Sans Biais (BLUP). Elle offre une discussion approfondie sur

l'estimation des composantes de la variance, explorant les méthodes du Maximum de Vraisemblance (ML), du Maximum de Vraisemblance Restreinte (REML) et des méthodes bayésiennes. La revue se tourne ensuite vers une vue d'ensemble de la sélection génomique dans l'élevage animal, mettant en lumière ses succès et les défis futurs. Elle offre également une base théorique pour les algorithmes d'apprentissage automatique (AA), suivie d'une revue extensive des modèles d'AA appliqués à la prédiction génomique en élevage animal. Le chapitre se termine par une discussion sur les applications potentielles de ces modèles en sélection animale et sur les opportunités spécifiques pour les pays en développement de bénéficier des avancées de l'IA dans la sélection génomique.

Chapitre 2: Paramètres génétiques du rendement laitier et des caractères de fertilité chez les Holstein marocains

Dans ce chapitre, nous présentons un de nos travaux scientifiques portant sur l'estimation des paramètres génétiques du rendement laitier et des caractères de fertilité chez les Holstein marocains. Ce chapitre discute de la corrélation antagoniste entre le rendement laitier à 305 jours et la fertilité chez les bovins laitiers, ainsi que de la faible héritabilité des caractères de fertilité qui compliquent la sélection génétique pour ces deux traits. En outre, ce chapitre comprend une explication détaillée d'un cadre bayésien pour estimer conjointement les paramètres du rendement laitier à 305 jours, deux caractères continus de fertilité (les jours ouverts et le nombre d'inséminations par conception), et un caractère binaire (succès de la première insémination). Enfin, il évalue l'effet de la composante génétique sur l'observation d'une première insémination réussie.

Chapitre 3: Analyse génétique de l'intervalle vêlage conception chez les Holstein Marocains en utilisant différents modèles pour traiter les données censurées

Dans ce chapitre, nous discutons des différents défis liés à l'analyse des données de fertilité sur le terrain, notamment les niveaux élevés de censure. Il explique comment la censure peut introduire des biais dans l'évaluation génétique et propose ensuite deux modèles pour traiter ce problème dans une base de données de Holstein Marocains. À des fins de comparaison, nous avons évalué la précision prédictive du modèle linéaire (sans observations censurées), de la méthode de pénalité et du modèle linéaire à seuil avec pénalité. Nous avons également calculé le biais, la dispersion, la corrélation de Spearman entre les valeurs

génétiques estimées des individus de validation, et le pourcentage d'animaux similaires dans le top 20% des individus sélectionnés pour comparer les trois méthodes.

Chapitre 4 : Exploration du potentiel de la sélection de caractéristiques par apprentissage automatique et méthodes conventionnelles pour réduire la dimensionnalité des jeux de données de marqueurs et améliorer la précision de la sélection génomique : une étude de simulation

Ce chapitre cherche à explorer le potentiel de plusieurs modèles d'apprentissage automatique, y compris la régression régularisée L1 et L2, les machines à gradient boosté, les réseaux neuronaux profonds, et l'analyse en composantes principales (ACP), pour la sélection de caractéristiques et leur capacité à améliorer la précision de la sélection génomique. Grâce à une approche par simulation, nous avons évalué la performance de ces modèles à différents niveaux d'héritabilité, en examinant leur efficacité à réduire la dimensionnalité des jeux de données SNP et leur influence sur les résultats de la sélection génomique.

Conclusions

Cette thèse a fourni des informations précieuses sur l'amélioration génétique des vaches Holstein dans l'environnement marocain et propose des alternatives potentielles pour résoudre les défis liés à la complexité informatique de la sélection génomique à l'avenir. Cette thèse peut être résumée en quatre conclusions:

- Démontre le potentiel d'amélioration à la fois de la fertilité et du rendement laitier malgré la corrélation antagoniste entre ces traits, en employant un indice de sélection complet qui prend en compte les traits pertinents.
- Souligne le rôle crucial des caractères de fertilité dans la durabilité économique des troupeaux laitiers. Elle aborde les défis posés par les données de fertilité censurées, montrant l'efficacité des méthodes à seuil pénalisé.
- Met en avant l'efficacité de la sélection de caractéristiques, notamment des machines à gradient boosté (GBM), pour identifier les SNPs pertinents et produire des estimations fiables de la valeur d'élevage.
- Enfin, elle confirme que les algorithmes d'apprentissage automatique ont montré un grand potentiel pour ajuster et extraire des motifs à partir de grands jeux de données bruyants. Cependant,

leur adoption dans l'élevage reste encore à ses débuts. Par conséquent, davantage de recherches doivent être menées pour trouver de nouveaux aperçus pour la sélection génomique.

Orientations futures:

Pour les perspectives futures, plusieurs domaines clés devraient être explorés pour améliorer l'amélioration génétique des vaches laitières au Maroc:

- Élargir les données phénotypiques : incorporer des caractères supplémentaires liés à la qualité du lait, à la fertilité, à la santé et à la longévité.
- Évaluation économique: évaluer les implications économiques de ces caractères pour leur inclusion dans des indices économiques afin d'améliorer les processus de prise de décision.
- Initiatives de génotypage: effectuer le génotypage des vaches laitières marocaines et établir systématiquement une population de référence nationale.
- Collaboration internationale: collaborer avec d'autres pays pour améliorer la précision des prédictions génomiques.
- Implémentation de technologies avancées: intégrer des données phénotypiques à haut débit avec des données génomiques pour découvrir de nouvelles perspectives sur l'application de l'IA dans les prédictions génomiques et phénotypiques.

LIST OF TABLES

CHAPTER 4:

LIST OF FIGURES

LIST OF ABREVIATIONS

AI	Artificial Intelligence
AI-REML	Average Information Restricted Maximum Likelihood
ANN	Artificial Neural Networks
ANOVA	Analysis Of Variance
APY	Algorithm for Proven and Young
AUC	Area Under the Curve
AUROC	Area Under the ROC
BLUE	Best Linear Unbiased Estimates
BLUP	Best Linear Unbiased Predictors
BNN	Bayesian Neural Networks
CART	Classification And Regression Trees
CNN	Convolutional Neural Networks
CNV	Copy Number Variation
DF	Derivative-Free
DNN	Deep Neural Networks
DO	Days Open
DT	Decision Trees
EBV	Estimated breeding values
ECP	Extreme-category Problems
EM-REML	Expectation-Maximization Restricted Maximum Likelihood
FN	False Negatives
FP	False Positives
GA	Genetic Algorithms
GBM	Gradient Boosting Machines
GLS	Generalized Least Squares
GP	Genomic prediction
GPTA	Genomic Predicted Transmitting Ability
GPU	Graphics Processing Unit
GS	Genomic Selection
GWAS	Genome-Wide Association Studies
HD	High Density
IBD	Identical By Descent
ID3	Iterative Dichotomiser 3
KcRR	Cosine Kernel-based Ridge Regression
LASSO	Least Absolute Shrinkage and Selection Operator

LD	Linkage Disequilibrium
LD	Low density
LM	Linear Model
LSE	Least Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MAS	Marker Assisted Selection
MCMC	Markov Chain Monte Carlo
ML	Machine Learning
ML	Maximum Likelihood
MLP	Multilayer Perceptron
MME	Mixed Model Equations
MSE	Mean Square Error
MY	Milk Yield
NIC	Number of Inseminations per conception
NLP	Natural Language Processing
NMSE	Normalized Mean Square Error
NN	Nearest Neighbors
OLS	Ordinary Least Squares
PAC	Prediction Accuracy
PBVs	Predicted Breeding Values
PCA	Principal Component Analysis
PFEMs	Performance Fitness and Error Metrics
PLM	Linear Model with Penalty
PSO	Particle Swarm Optimization
PTM	Linear Threshold Model with Penalty
QTL	Quantitative Trait Loci
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
REML	Restricted Maximum Likelihood
RF	Random Forest
RHKS	Reproducing kernel Hilbert spaces
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
SCS	Somatic cell score
SFI	Success of the First Insemination
SNPs	Single Nucleotide Polymorphism
ssGBLUP	Single-Step GBLUP
SVM	Support Vector Machines

TN	True Negatives
TP	True Positives
USDA	United States Department of Agriculture
VCE	Variance Components Estimation
XgBoost	Extreme Gradient Boosting

LIST OF PUBLICATIONS

1. Submitted and published papers containing the work founding the thesis:

Chafai N, Hayah I, Houaga I and Badaoui B (2023) A review of machine learning models applied to genomic prediction in animal breeding. Front. Genet. 14:1150596. doi: 10.3389/fgene.2023.1150596.

Chafai, N.; Badaoui, B. Genetic Analysis of Days Open in Moroccan Holstein Using Different Models to Account for Censored Data. Animals 2024, 14.

Chafai, N.; Badaoui, B.; Rekaya, R. Genetic parameters of milk yield and fertility traits in Moroccan Holsteins. Front. Anim. Sci.

2. Other publications not included in this work:

Chafai N, Bonizzi L, Botti S, Badaoui B. Emerging applications of machine learning in genomic medicine and healthcare. Crit Rev Clin Lab Sci. 2024 Mar;61(2):140-163. doi: 10.1080/10408363.2023.2259466. Epub 2023 Oct 10. PMID: 37815417.

Hayah I, Talbi C, Chafai N, Houaga I, Botti S and Badaoui B (2023) Genetic diversity and breed-informative SNPs identification in domestic pig populations using coding SNPs. Front. Genet. 14:1229741. doi: 10.3389/fgene.2023.1229741

TABLE OF CONTENTS

INTRODUCTION

The Moroccan dairy industry offers crucial socio-economic assets, it generates over 13 billion MAD and a total of 460.000 permanent jobs (Labriji et al., 2021). Several governmental initiatives have increased national milk production during the last two decades. For example, in the 1970s, the Ministry of Agriculture set up a "National Dairy Plan" plan to improve Moroccan dairy herd genetics by introducing improved dairy cattle, implementing milk recording, and extending artificial insemination (AI). Consequently, since 1996, Morocco has imported around 30,000 purebred heifers per year to improve the genetics of Moroccan dairy cows. Additionally, in 2008, the government reprioritized the agricultural sector by adopting the "Green Morocco Plan" which targeted a milk production of 4.0 million tons by 2020. Despite the great improvement in the dairy sector, milk production stagnated at 2.5 million tons and the dairy sector still struggles with serious issues mostly related to difficult environmental conditions, bad sanitary conditions, bad farm management practices, and inadequate genetics. This results in a limited profit margin for dairy farms due to high expenses and low income, as the price of one litter of milk is around 3.5 MAD.

The introduction of purebred heifers in large numbers has brought in breeds with high milk production potential, such as Holstein, Dutch Red Frisian, Montbelliard, German Fleckveth, and Danish Red Friesian. Of these, the Holstein breed dominates with over 70% of the national milk production. However, the arid Moroccan environment and poor management practices limit the milk production potential of these cows and shorten their productive lives. As a result, heifers seldom reach their third lactation and are often involuntarily culled (Sraïri & Baqasse, 2002). A study conducted by Boujenane (2017), 98% of the reasons for culling were involuntary. The main causes of involuntary culling were disease (accounting for 38% of culling) and reproductive issues (accounting for 36%). This highlights that large-scale importation is an inefficient and unsustainable method for improving the genetics of Moroccan Holsteins. Therefore, it is vital to quantify the environmental impact and suggest an index that includes fertility traits for selecting replacement heifers in the Moroccan environment.

Several problems can impede the development of a national genetic evaluation system. First, 84% of the farms are small farms with less than 4 cows. These cows are generally crossbred, and their milk production is low. However, there are a few progressive farms that are relatively well managed and where farmers

are willing to pay a premium price to have a Holstein heifer. Developing an on-farm evaluation system can help these farms improve the genetics of their herd while accounting for the local environmental conditions. Another challenge in developing a genetic evaluation system for Morocco is the scarcity and subpar quality of available data. In most Moroccan farms, milk quality is not systematically recorded on a per-cow basis, and fertility data often contain numerous missing or censored records and a significant amount of noise. Additionally, pedigree information is frequently inaccurate, which lower notably the reliability of estimated breeding values.

One of the most effective ways to improve Moroccan Holstein genetics and avoid the deterioration of fertility while preserving high milk production is to develop a selection index that includes both yield and fertility traits. Choosing informative fertility traits that can account for veterinary costs, labor, semen costs, and management decisions is crucial for the efficiency of genetic selection. Additionally, addressing challenges related to fertility field data including the noise and the censorship is crucial for improving the predictive ability of genetic models and thus the estimated breeding values reliability. Furthermore, introducing genomic selection for Moroccan herds can be extremely beneficial to accelerate genetic gain. Genotyping dairy cows in Morocco will help address issues related to inaccurate pedigree recording as it checks parentage (Hayes et al., 2009), increases the accuracy of estimating breeding values, accelerates the genetic gain, and allows for further exploring of genetic by environment interactions (Bouquet & Juga, 2013). To achieve high reliability and genetic gain through genomic selection, it is essential to possess a large reference population size. Given that Morocco typically imports semen from elite sires, the potential Moroccan Holstein reference population will comprise exclusively cows. However, the reliability of cows is relatively low, and thus collaboration with other countries and exploring alternative approaches would be advantageous for maximizing the benefits of genomic selection in the Moroccan context.

The advancement of molecular technologies and the significant decrease in genotyping costs have led to an unprecedented increase in the number of genotyped animals. In certain cases, the number of genotyped animals exceeds the number of markers, which poses a primary challenge for the statistical methods utilized in genetic evaluation. The first challenge relates to the necessity of inverting the **G** matrix, which is computationally challenging due to its cubic cost with the number of animals. To overcome this challenge, an approximation of the inverse of **G** using the Algorithm for Proven and Young Animals (APY) has been proposed. However, as this method is data-driven, the results may vary depending on the

animals selected in the core, and critical downward bias in genomic predictions for certain animals may arise.

Additionally, on top of the growing number of genotyped animals and sequence data, the development of high-throughput phenotyping has facilitated the exploration of new insights in genomic prediction. These advancements have been fueled by the remarkable progress in machine learning (ML) models and artificial intelligence across various fields, including remote sensing, computer vision, and natural language processing (NLP). As a result, there has been a growing interest in investigating the potential of applying ML to animal breeding for identifying important markers relevant to certain traits and modeling the non-linear effects involved in the genetic determinism of traits, such as dominance and epistasis.

In this intricate context, the primary objective of this dissertation was to carry out a genetic assessment for productive and reproductive traits of a commercial herd of Holstein cows in the Moroccan environment. The study aimed to calculate the variance components and genetic parameters of these traits using various statistical models. Additionally, the research examined potential statistical methods able to impute censored fertility data and increase the predictive ability of genetic models in the Moroccan context. Finally, this thesis has also investigated the use of machine learning algorithm for feature selection of important SNPs to decrease the computational complexity of genomic evaluations by estimating SNP effects instead of inverting G matrix. This thesis is organized into chapters. A literature review is provided in Chapter 2. Chapter 3 comprises a review of machine learning algorithms used in animal breeding for genomic prediction. Chapter 4 consists of an estimation of genetic parameters of milk yield and fertility traits for Moroccan Holstein and assesses the effect of the genetic components of some traits across different production environments. In Chapter 5, we explore the implementation of three linear models in imputing censored fertility data and we assess the effect of imputing the censored records compared to deleting missing information. Chapter 6 investigates the potential of various machine learning algorithms in identifying relevant SNPs and potentially increasing breeding values estimation accuracy. Lastly, the general conclusions of this dissertation are presented in Chapter 7 and future directions in Chapter 8.

CHAPTER 1- LITERATURE REVIEW

1. Evolution of selection practices in animal breeding: The rise of quantitative genetics

The history of animal breeding began long before the principles of genetics were understood. Early breeders used to select and mate the best individuals based on observable traits and their relatives' performance. The first major transition in animal breeding came in the mid-19th century with Mendel's discovery of genes, the units of inheritance. This was followed by Galton's contributions in the late 19th century, where he introduced the concept of a linear regression model to predict an individual's height based on a relative's height, with the slope varying according to the closeness of their genetic relationship. This statistical regression of offspring traits on those of their parents was later termed "realized heritability" by Falconer. Since then there was an active debate between the biometrician and Mendelian schools of thought until Ronald Fisher came up with a finding that revolutionized the field of genetics and unified the two schools (Nelson et al., 2013).

Fisher's infinitesimal model, presented in 1919 (Fisher, 1919), transformed animal breeding and consisted of a solid framework for following genetic studies. Fisher introduced the assumption that the genetic variance within families is due to a large number of Mendelian factors with small contributions that act additively. He also suggested the decomposition of trait values and their (co)variances into several components such as dominance and epistasis, which became foundational to quantitative genetics, laying the groundwork for the later development of the animal model by Henderson in 1960.

While Fisher introduced the underlying statistical concepts, Jay L. Lush contributed a lot to the understanding and the application of the narrow and broad sense of "heritability" within the field of animal breeding. According to Bell (1977), the exact origin of the word "heritability" is unclear, but its usage has evolved through three distinct stages, each becoming more specific in its meaning. Initially, around 1832 or possibly earlier, "heritability" referred broadly to the hereditary transmission of characteristics, followed by the second stage, around the early 20th century, the term followed Johannsen's classical distinction between genetic (genotypic) differences and environmental (nongenetic) fluctuations, approximating what we now call "broad-sense heritability." Finally, in 1936, the term evolved to "narrow-sense heritability" that we use until now in animal breeding, defined as the ratio of additive genetic variance to the total phenotypic variance within a population (Bell, 1977).

Heritability is a fundamental concept in quantitative genetics, playing a crucial role in understanding and predicting the genetic potential for trait improvement within populations. It quantifies the proportion of phenotypic variation in a trait that is attributable to genetic differences among individuals. Heritability was first estimated using the breeder's equation, $R = h^2S$, where R is the ratio of the change in mean phenotype between generations, and S. is the selection differential that means the difference in mean phenotype between the parents selected for breeding and the overall mean in their generation (Visscher & Goddard, 2019).

Other traditional methods of estimating heritabilities used the concept of decomposition of phenotypic variance. Traditionally, variance estimates were based on simple designs, such as comparing offspring traits to parental traits, looking at correlations among siblings, and comparing correlations between monozygotic and dizygotic twins. However, as research expanded to include more complex family structures and relationships across generations, these traditional methods became less efficient. In such cases, where the data is unbalanced or involves mixed relationships, the linear mixed model is preferred for estimating both genetic and environmental influences on phenotypic traits. Within this approach, the "animal model" has gained popularity, particularly in fields like livestock genetics, evolutionary genetics, and some areas of human genetics.

Henderson introduced in the late 1940s and early 1950s, a set of equations for mixed models that resemble the typical least-squares equations. However, these equations include an important adjustment for random effects called the shrinkage term. This adjustment makes the equations more compact and easier to work with compared to the alternative equations based on maximum likelihood estimation. The mixed model equations provide estimates for both fixed and random effects. The estimates for fixed effects are known as Best Linear Unbiased Estimates (BLUE), while the estimates for random effects are referred to as Best Linear Unbiased Predictors (BLUP). These estimates are considered optimal because they are unbiased and have minimum variance among all linear estimators. The first implementation of BLUP models was focused on sire genetic effects and within sire deviations. These models served the fundamental requirement for evaluating the genetic merit of dairy sires and were computationally viable. However, a comprehensive analysis necessitates the incorporation of the "animal model", a concept first elucidated by Henderson in 1976 although not explicitly named at the time. In this model, the breeding value of each individual is estimated in terms of its effects and the covariance structure among the breeding values of different individuals which is Wright's numerator relationship $\boldsymbol{A}$. The solutions of BLUP required the

inverse of the additive relationship matrix $\mathbf{A}^{-1}$. However, Henderson came up with a simpler method to build the inverse from the pedigree information without the direct inversion of $\mathbf{A}$. Henderson's mixed model equations, BLUP, and variance components estimation are a landmark in the field of statistics and quantitative genetics, they are still used today for estimating breeding values in animal and plant breeding.

The following sections will elucidate the genetic evaluation methodologies employed in animal breeding, arranged chronologically. Commencing with the classical least squares approach, followed by the selection index, subsequently progressing to the single-trait mixed-model incorporating both sire and animal models, and finally, the multiple-trait mixed-model analysis.

2. Generalized Least Squares (GLS)

Consider the linear model below.

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} \qquad [1]$$

Where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{b}$ is the vector of fixed effect, $\boldsymbol{e}$ is the vector of random residuals, and $\boldsymbol{X}$ is the incidence matrix linking the observations to the fixed effects. $\boldsymbol{e}$ follow a normal distribution with mean $\mathrm{E}(\boldsymbol{e}) = 0$ and $\mathrm{Var}(\boldsymbol{e}) = \mathbf{V}$.

To get the best unbiased estimator (BLUE) of the fixed effects, GLS differentiates the expression $(\boldsymbol{y} - \boldsymbol{Xb})'\boldsymbol{V^{-1}}(\boldsymbol{y} - \boldsymbol{Xb})$ with respect to b (the vector of fixed effects) and then equating the derivative to zero. This leads to the formula below.

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}Y \qquad [1]$$

When the residuals in the model are uncorrelated and have constant variance, $\boldsymbol{V}$ equals the identity matrix $\boldsymbol{I}$, and GLS is simplified to ordinary least squares (OLS) and when the residuals are still uncorrelated, but their variance might vary across different levels of independent variables $\boldsymbol{V}$ becomes a diagonal matrix $\boldsymbol{D}$, GLS becomes weighted least squares.

3. Selection Index

Selection indices were introduced to select superior individuals for multiple traits. These indices write each trait based on its economic importance and the genetic correlations among traits. According to Hazel & Lush (1943), selecting using an index that appropriately weights each trait is more efficient than selecting for one trait at a time. The selection index formula proposed by Hazel (1943) is,

$$I = w_1 z_1 + w_2 z_2 + \cdots + w_n z_n \quad [2]$$

Where $\boldsymbol{I}$ is the selection index value, w_i are the weight assigned to trait i, and z_i are the standardized deviation of trait i from the mean.

The weights w_i can be determined based on economic values, heritabilities, and genetic correlations among traits. Economic values represent the relative economic importance of each trait in achieving the breeding goal. Heritabilities indicate the proportion of phenotypic variation in a trait that is due to genetic factors, and genetic correlations quantify the relationship between different traits.

4. Henderson Mixed Model Equations (MME), and Best Linear Unbiased Prediction (BLUP)

The Henderson Mixed Model Equations (MME) stand as a cornerstone in statistical genetics and the analysis of complex traits. Originating from the pioneering work of Charles Roy Henderson, these equations provide a robust framework for estimating genetic parameters in populations with diverse structures, encompassing both fixed and random effects. Their elegance lies in their ability to handle the inherent complexities of hierarchical data structures, accommodating various sources of variation while ensuring efficient and accurate estimation.

The mixed model equations introduced by Henderson (1974) for a univariate animal model are shown below.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \quad [3]$$

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}}\boldsymbol{y} \\ \mathbf{Z'R^{-1}}\boldsymbol{y} \end{bmatrix} \quad [4]$$

Where $\boldsymbol{y}$ is the vector of phenotypes, $\widehat{\boldsymbol{\beta}}$ is the vector of estimated fixed effects, and $\widehat{\boldsymbol{u}}$ is the vector of estimated breeding values. $\mathbf{G^{-1}}$ is the inverse of the covariance matrix of random effects. $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices that connect the observations to the fixed and random effects respectively.

The model can support multiple random effects, for example, permanent environment effect, maternal effect, or interactions. In this case, $\mathbf{G^{-1}}$ can be decomposed into contributions from each random effect.

The covariance matrix of random effects changes according to the model that is used. In the case of a sire model, which is a specific type of mixed model that has been in use for several decades, particularly for dairy cattle, that focuses on the genetic contribution of sires to their offspring, the genetic covariance of sire additive effect can be modeled as:

$$\mathbf{G_i} = \mathbf{A_s}\sigma_i^2 \qquad [3]$$

For the animal additive effect, $\mathbf{G}$ is equal to the additive relationship matrix multiplied by the additive genetic variance.

$$\boldsymbol{G_i} = \boldsymbol{A}\sigma_i^2 \qquad [4]$$

When dealing with multiple traits, the model needs to account for the correlation between traits. The multivariate mixed model can be written as:

$$\begin{bmatrix} \mathbf{X}'(\mathbf{R}^{-1} \otimes \mathbf{I})\mathbf{X} & \mathbf{X}'(\mathbf{R}^{-1} \otimes \mathbf{I})\mathbf{Z} \\ \mathbf{Z}'(\mathbf{R}^{-1} \otimes \mathbf{I})\mathbf{X} & \mathbf{Z}'(\mathbf{R}^{-1} \otimes \mathbf{I})\mathbf{Z} + (\mathbf{G_0^{-1}} \otimes \mathbf{A}^{-1}) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'(\mathbf{R}^{-1} \otimes \mathbf{I})\boldsymbol{y} \\ \mathbf{Z}'(\mathbf{R}^{-1} \otimes \mathbf{I})\boldsymbol{y} \end{bmatrix} \qquad [5]$$

where $\mathbf{G_0^{-1}}$ is the inverse of the genetic (co)variance matrix of the traits, and $\mathbf{G_0^{-1}} \otimes \mathbf{A^{-1}}$ is the chronicle product of the additive relationship matrix and the genetic covariance matrix.

Henderson's mixed models contributed remarkably to all the advances in animal breeding. First, it simplifies the computation of $X\hat{\beta}$ remarkably as the left-hand side of the MME has an order equal to the number of fixed and random effects. Furthermore, calculating the inverse of the residual matrix is not very computationally expensive as it has the dimensions of the number of traits in the model, and Henderson came up with a way to compute the inverse of the additive relationship matrix directly without building the A matrix and inverting it, which makes computing $\mathbf{G^{-1}}$ easier. Second, it enables computing the solutions for fixed effects and random effects simultaneously. This dual estimation is crucial in accurately capturing the genetic contributions to observed traits. However, performing BLUP assumes that the variance components are known, which is not the case for most datasets. Therefore, variance component estimation is a key step for estimating breeding values using mixed models' equations. Variance components estimation (VCE) methods are discussed in the following section.

5. Variance components estimation

Henderson proposed three influential methods to estimate variance components, crucial in the field of animal breeding and quantitative genetics. The first method is Henderson's Method 1, which is based on the analysis of variance (ANOVA) approach, leveraging expected mean squares from a balanced data structure to estimate components. The second, Method 2, employs the Maximum Likelihood (ML) estimation, providing a more flexible approach that can accommodate unbalanced data and complex models, though it can be computationally intensive. The third, Method 3, utilizes Restricted Maximum Likelihood (REML), which refines ML by accounting for the loss of degrees of freedom due to fixed effects, thereby yielding more accurate and unbiased estimates (Searle, 1991). Each method has its strengths and is chosen based on the specific requirements and structure of the data at hand.

5.1.Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML)

The likelihood-based methods are considered a standard method to estimate variance components nowadays. They are adapted to animal breeding as they can be applied to unbalanced data. Maximum likelihood, developed first by Hartley and Rao (1967), assumes that the phenotype y follows a normal distribution $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \mathbf{V})$ and the residuals follow a normal distribution with mean 0 and variance $\mathbf{R}$, $\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{R})$, the likelihood is then defined as :

$$\mathrm{L}(\boldsymbol{\beta}, \mathbf{V}; \boldsymbol{y}) = \frac{1}{\sqrt{2\pi|\mathbf{V}|}} e^{-\frac{-(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})}{2}} \quad [6]$$

Since maximizing the likelihood is the same as maximizing log-likelihood, the formula can be simplified to

$$\ln\mathrm{L}(\boldsymbol{\beta}, \mathbf{V}; \boldsymbol{y}) = -\frac{1}{2}\left[\ln(2\pi) + \ln|\mathbf{V}| + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right] \quad [7]$$

Obtaining estimates of the variance components involves equating the derivatives of the likelihood function with respect to the variance components to zero. Maximum likelihood (ML) estimates are particularly valued for their unbiasedness and desirable properties in large samples (Hartley and Rao, 1967). Among the various ML algorithms, the expectation-maximization (EM) algorithm is frequently

used for variance components estimation. In the context of EM, the estimators are asymptotically unbiased, which means that with increasing sample size, the estimators approach the true parameter values. Furthermore, the variability of these estimators can be described by the asymptotic dispersion matrix, which is the inverse of the Fisher information matrix and is a function of the unknown variance components. By substituting the estimated variance components into this matrix, we obtain an estimated dispersion matrix. Additionally, the dispersion matrix of the estimators achieves the Cramer-Rao lower bound asymptotically. This implies that the estimators are asymptotically efficient, reaching the minimum possible variance for unbiased estimators, thus making the ML method highly effective for variance components estimation (Hofer, 1998).

Maximum Likelihood (ML) is still widely used for variance components estimation, but it can be biased, especially in small samples. Restricted Maximum Likelihood (REML) was developed to address this issue. REML adjusts for the loss of degrees of freedom associated with estimating fixed effects in the model, resulting in less biased estimates of variance components compared to ML (Fleming et al., 2019).

In REML, the goal is to estimate the vector of unknown parameters by maximizing a restricted log-likelihood function while bypassing the estimation of the vector of fixed effects $\boldsymbol{\beta}$. Therefore, rather than using the vector of phenotypes $\boldsymbol{y}$ directly, a linear combination of $\boldsymbol{y}$ is chosen in a way that those combinations do not contain any fixed effects. In matrix notation, the model in equation [3] is multiplied by a vector $\boldsymbol{K}$.

$$\boldsymbol{Ky} = \boldsymbol{K}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{KZu} + \boldsymbol{Ke} \qquad [8]$$

Such that $\boldsymbol{K}'\mathbf{X} = \mathbf{0}$. The vector $\boldsymbol{K}$ is referred to as error contrast (Harville, 1977). The REML model becomes as follows.

$$\boldsymbol{y}^* = \boldsymbol{KZu} + \boldsymbol{Ke} \qquad [9]$$

Where the variance of $\boldsymbol{y}^*$ is equal to $\boldsymbol{KVK}'$. Therefore, the likelihood can be written as:

$$\mathrm{L}(\mathbf{V}|\boldsymbol{y}) \propto -\frac{1}{2}\ln|\boldsymbol{K}'\mathbf{V}\boldsymbol{K}| - \frac{1}{2}(\boldsymbol{Ky})'(\boldsymbol{K}'\mathbf{V}\boldsymbol{K})^{-1}(\boldsymbol{Ky}) \qquad [10]$$

In subsequent derivations, the error contrast cancels out and the restricted maximum likelihood becomes.

$$L(\mathbf{V}|\boldsymbol{y}) \propto -\frac{1}{2}(\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \boldsymbol{y}(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1})\boldsymbol{y}) \quad [11]$$

EML estimates are obtained by iteratively maximizing the likelihood function with respect to the parameters of interest using various maximization algorithms, which may be derivative-free (DF) or employ first or second derivatives (Misztal, 2008). One commonly used algorithm is Expectation Maximization (EM) REML, which utilizes the first derivatives of the likelihood. The EM-REML process involves two main steps. The first step called the Expectation step (E-step), calculates the expected value of the log-likelihood function concerning the parameters to be estimated. The second step, the Maximization step (M-step), updates the parameters based on the expected log-likelihood computed in the E-step. By iterating these two steps until convergence, reliable parameter estimates are achieved.

The most popular maximization algorithm that uses the second derivates is the average information restricted maximum likelihood (AI-REML) proposed by Johnson & Thompson (1995). AI-REML is built upon the efficiency of Newton's methods, which use both the first and second derivatives of the likelihood function and are known for their rapid convergence. AI-REML uses the average of the observed and expected second derivatives to cancel out the trace term resulting in less computational cost compared to traditional Newton's methods.

Overall, derivative-free (DF) algorithm is advantageous for its simplicity and speed in handling straightforward models; however, it becomes prohibitively slow and unreliable with more complex models, such as multiple trait models, leading to its infrequent use today. The EM algorithm once considered the most reliable despite its slow convergence (sometimes requiring hundreds of iterations) fails with random regression models, producing inconsistent estimates based on initial values. Additionally, EM does not directly provide standard errors of estimates and becomes significantly more complex when dealing with missing traits. On the other hand, the Average Information (AI) algorithm, though complex to program, offers direct computation of standard errors and typically achieves convergence in just a few iterations for many models (Misztal, 2008).

5.2. Bayesian methods

The theory behind Bayesian inference is to use probability for expressing uncertainty. In other words, in the frequentist approach, if we want to estimate a parameter given a set of data, we look for the value of

that parameter that can produce the data with the highest probability (maximizes the likelihood). Bayesian inference, instead, provides a probabilistic distribution of the parameter we want to estimate. The only challenge is that we need prior information about the parameter, and we need a method to introduce this prior information in the analysis (Blasco, 2017).

The Bayesian inference is based on 'Bayes theorem':

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad [12]$$

where $P(A|B)$ is the conditional distribution of A given B, $P(B|A)$ is the conditional distribution of B given A, $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively without any condition. In animal breeding, Bayesian inference is used for several objectives including the estimating of the genetic parameters given the data. In case of variance component estimation, $P(A|B)$ is the probability distribution of the parameter of interest, $P(B|A)$ is the distribution of the data for a given trait. This distribution is known or assumed to be known from reasonable hypotheses. $P(A)$ is the prior information that we have a priori about the parameter of interest, and $P(B)$ is a scaling factor, the probability of the sample.

The constant $P(B)$ can be calculated by integrating the probability distribution of the data. However, these integrals are often complicated to compute directly. The introduction of Markov Chain Monte Carlo (MCMC) methods made the implementation of the Bayesian framework simpler by allowing the computation of the marginal posterior distribution without solving these integrals. MCMC methods are numerical tools that enable us to randomly sample from the marginal posterior distributions. We can make inferences using these samples, for example, given a random MCMC sample for the marginal posterior distribution $P(B|A)$, the probability of B being higher than 0 can be calculated by counting how many samples are higher than 0 divided by the total number of samples. However, there's a sampling error related to these methods, and it depends on the sample size and how correlated are the samples. To address some of these challenges, we can improve our sampling, by thinning the MCMC chain—removing some consecutive samples—we reduce the autocorrelation between samples, making them more representative of the posterior distribution. Second, we can increase the total number of samples collected to enhance the precision of our posterior estimates. However, simply increasing the number of samples does not necessarily guarantee the precision of our experiment, as it is also crucial to ensure that the samples are

well-mixed and representative of the entire parameter space. Proper diagnostics and convergence checks are essential to validate the reliability of the MCMC estimates. The most common MCMC method used in animal breeding is Gibbs sampling. Conceptually, the Gibbs sampler is a relatively straightforward algorithmic process. However, a key aspect of its implementation involves efficiently sampling from the full conditional distributions, which can be approached in various ways. Often, for some parameters, the prior distribution will be conjugate to the likelihood, resulting in the full conditional distribution being a posterior update of a standard prior (Gelfand, 2000). When the resulting conditional distribution is not a known function other MCMC techniques can be used to sample from these distributions.

Gibbs sampler algorithm can be explained as follows. Consider the random variables $X_1, X_2,$ and X_3. At iteration 0 we set their values to some initial values $x_1^{(0)}, x_2^{(0)}$, and $x_3^{(0)}$. At iteration i, we sample $x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}$, $x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}$, and $x_3^{(i)} \sim p(X_3 = x_3 | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}$. We repeat the process for many rounds to aim for convergence (when the values have the same distribution as if they were sampled from the true posterior distribution).

The theory of MCMC guarantees that after a large number of iterations, we can generate samples from the target distribution. Additionally, we can clearly tell from the aforementioned algorithm that the first rounds are not sampled from the target distribution as they are highly influenced by the starting values. These samples are discarded and referred to as burn-in (Yildirim, 2012). One can also try different chains with different starting points to make sure that the chains converge to the same target distribution. Trace plots are usually used to check the convergence and behavior of the Gibbs sampler.

Overall, the Bayesian framework has been widely used for estimating genetic parameters in animal breeding and has demonstrated significant potential in addressing various challenges. First of all, Bayesian inference allows us to make inferences about the unknowns using probabilities and marginalization. Second, it enables us to incorporate prior information about unknown parameters, which is particularly useful when the available data is insufficient, but we have enough confidence in at least the range of the estimates. However, integrating the prior information in the inference is still difficult and most modern Bayesians do not use prior information except to allow them to work with probabilities (Blasco, 2017). Third, all the variables are random, so we do not decide which variable is fixed or random. Finally, we have a systematic method for making inferences and can express uncertainty in various ways using these distributions.

6. Overview of genomic selection in animal breeding

As outlined in earlier sections, estimating genetic variance using pedigree information has considerably improved genetic gain and the precision of estimated breeding values. However, the main focus has always been on utilizing inherited genes to recognize those that govern the phenotypes of interest. Since specific mutations have been associated with abnormalities and polymorphisms that significantly influence quantitative traits, attempts have been made to identify these polymorphisms and integrate them into breeding strategies. Unfortunately, research found that these known causal polymorphisms account for only a small fraction of the genetic variance. Therefore, this approach was limited by the inability to link most causal polymorphisms to objective traits.

In the 90's, new genetic markers were discovered and tested for associations with quantitative traits. These associations proved too weak and unreliable for practical use in livestock selection. Many of these markers influenced traits indirectly by being in linkage disequilibrium (LD) with the Quantitative Trait Loci (QTLs) controlling the traits (Goddard et al., 2010). Microsatellites were the first genetic markers discovered that tracked QTLs. Typically, 100 to 200 markers were used to detect QTLs by LD within full- or half-sib families. This method, called Marker Assisted Selection (MAS), has also been unsuccessful, as it tracks the QTLs imprecisely. The linkage disequilibrium between the markers and QTLs varied across families, making this approach impractical for improving livestock breeding strategies (Blasco & Toro, 2014). In 2001, Meuwissen et al. proposed the concept of Genomic Selection (GS), which is a form of Marker Assisted Selection (MAS) that is based on two fundamental assumptions. Firstly, it necessitates the development of extensive marker panels comprising tens of thousands of genetic markers. Secondly, it requires achieving sufficient marker density to ensure that all genes that influence a particular trait are in linkage disequilibrium with their neighboring markers. As the density of marker panels is increasing, and the cost of genotyping is significantly decreasing genomic selection turned into a practical reality in livestock breeding. Genomic selection can be implemented in two main steps. First, estimate the effects of markers, usually Single Nucleotide Polymorphism (SNPs), using a reference population that has been phenotyped and genotyped. Second, use the estimated SNP effects to predict the breeding value of candidates for selection in a testing population that has been genotyped only. A naïve model using 50,000 SNPs can be written as follows.

$$y_i = \mu + X_{1i} \times b_1 + X_{2i} \times b_2 + \cdots + X_{50000i} \times b_{50000} + e_i \qquad [13]$$

Where y_i is the phenotype of animal i, μ is an overall mean, X_{ji} is the genotype of animal i for marker j, usually coded 0,1,2 for homozygote for allele 1, heterozygote, and homozygote for allele 2, and e_i is the residual (Meuwissen et al., 2016).

The changes in the genetic models were coupled with the development of various statistical methods and techniques to estimate breeding values using SNP effects. One of the challenges these approaches tackled is the "large p and small n problem" (Blasco & Toro, 2014). As the number of variables in these models (SNPs) exceeds the number of observations, we cannot estimate the SNP effects using classical statistical approaches if they were treated as fixed. Therefore, SNP effects were assumed random, and Bayesian methods were used to make all effects estimable by including prior information about their distribution. Given the type of prior used, the method changed. BayesA assumes that all markers have a non-zero effect and are normally distributed with a mean of zero and a locus-specific variance. In BayesB, the effects of SNPs can be partitioned into two components: one with zero effects, which occurs with probability π, and the other with non-zero effects, which occurs with a probability of 1- π. The distribution of the latter component is assumed to be normal, with a mean of 0 and a locus-specific variance following a scaled inverse chi-square distribution (Gianola, 2013). These two methods have been subject to criticism from both genetic and statistical perspectives. The notion that all single nucleotide polymorphisms (SNPs) have small effects that are normally distributed does not accurately reflect the biological determinism of the traits. Furthermore, the fact that the mixture proportion π is not inferred from the data but instead chosen based on assumptions or prior beliefs makes the assignment of π arbitrary. BayesC proposed by Habier et al., (2011) modifies the Bayesian approaches A and B by introducing a common variance for all SNP effects, rather than locus-specific variances. This common variance is assigned a scaled inverse chi-square prior similarly to BayesA and BayesB. As a result, the distribution of a SNP effect, when it is not zero (with probability 1−π), comes from a mixture of multivariate Student's t-distributions. For example, in an analysis involving three SNPs, there are four possible models where the effect of any given SNP is not zero. Each model corresponds to a different configuration of non-zero effects among the SNPs, resulting in the effect of a SNP being drawn from a mixture of multivariate t-distributions. This approach considers the combined uncertainty and variability across different possible models, reducing the influence of individual priors on SNP variances and providing a more robust and flexible model for genetic effect estimation. Several Bayesian alphabet methods were introduced assuming different regularization terms to shrink the effect of some SNPs or set these effects to 0, such as Bayesian Lasso (BayesL) or Bayesian Ridge Regression (BayesR) with different prior information assumptions about the regularization term.

The BLUP have also been used to estimate the marker effects. It is similar to a least square estimate with a shrinkage towards zero. For the case where the vector of marker effects $\boldsymbol{b}$ is normally distributed, the model in matrix notation can be presented as follows:

$$\boldsymbol{y} = \mathbf{M}\boldsymbol{b} + \boldsymbol{e} \qquad [14]$$

Where **y** is the vector of observations, **M** is the matrix of genotypes of dimension number of observations by number of SNPs, $\boldsymbol{b}$ is the vector of SNP effect, and **e** is the vector of residuals. The breeding values are estimated by summing the effects of SNPs for each animal using the formula $\boldsymbol{bv}$ = **M**$\boldsymbol{b}$, which means that $\boldsymbol{y} = \boldsymbol{bv} + \boldsymbol{e}$ with $\boldsymbol{bv} \sim \boldsymbol{N}(0, \mathbf{MM}'\sigma_g^2)$**.** This model, also called SNPBLUP, is equivalent to an animal model where the relationship among animals is computed using the matrix of genotypes **MM**' instead of A (pedigree-based relationship matrix) (Goddard et al., 2010). This method is referred to as GBLUP. Several modifications were developed for the matrix **MM**' to increase the liability of estimated breeding values. VanRaden (2007) proposed a method to calculate the genomic relationship matrix based on genotype information. This method modifies the marker genotype matrix M by subtracting a matrix P that contains the frequencies of the second allele at each locus p_i. This step gives more weight to rare alleles when computing genomic relationships. The allele frequencies in **P** should be those of the base population before the selection occurs. The resulting matrix **Z** is multiplied by the transpose and divided by $2\sum p_i(1-p_i)$ in order to be analogous to **A**. The matrix $\mathbf{G} = \mathbf{ZZ}'/[\,2\sum p_i(1-p_i)]$ is positive semi-definite and can be singular if the number of markers is too small so individuals can have the same genotypes, if the data contains identical twins, or if the number of alleles is less than the total number of genotyped animals. Contrarily to the expected relationship matrix A, G computes the similarities between animals' genotypes, consequently, it does not only track blood relationships, also called identical by descent (IBD) but also marker genotypes that are identical by state (IBS). Matrix G can be used in mixed model equations to estimate breeding values of a simple model $\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e}$ using the following formula:

$$\widehat{\boldsymbol{a}} = \mathbf{G}\left(\mathbf{G} + \frac{\mathbf{R}\sigma_e^2}{\sigma_a^2}\right)^{-1}(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{b}}) \qquad [15]$$

Where $\widehat{\boldsymbol{a}}$ is the vector of estimated breeding values, **G** is the genomic relationship matrix**,** $\boldsymbol{y}$ is the vector of phenotypes**,** $\widehat{\boldsymbol{b}}$ is the vector of estimated fixed effects, and **X** is the incidence matrix that related fixed effects to observations (VanRaden, 2007).

Generally, in genomic selection, some animals are not genotyped but as they are related to genotyped animals in the population, we are interested in including their phenotypic information to maximize the amount of information included to estimate the breeding values. The traditional GBLUP required the genotypes for all animals. Two methods were proposed to address this challenge. First is the multiple-step GS method which consists of 3 steps. The first step is to generate estimated phenotypes for genotyped animals. These pseudo-phenotypes incorporate information from the animal's ungenotyped relatives. For example, if a bull is genotyped but has no direct phenotype measurements, its pseudo-phenotype might be based on the average production of its daughters. Second step consists of using the pseudo-phenotypes along with the genotype data of the animals, to perform genomic predictions. Finally, we merge the traditional estimated breeding values (EBV) with the genomic estimated breeding values (GEBV) to produce a total EBV. This combined value provides a more comprehensive assessment of an animal's genetic merit (Meuwissen et al., 2016). The second method is more commonly used in animal breeding, it's called single-step GBLUP and it was proposed in parallel by Legarra et al. (2009) and Christensen & Lund, (2010). The idea behind this approach is to integrate genotyped and non-genotyped animals in one single analysis by using the genomic relationship matrix for genotyped animal and pedigree information for non-genotyped animals. Basically, it's an extension of a linear mixed model where an animal's breeding value is composed of two components: a) A genomic genetic effect based on marker information, and b) A polygenic genetic effect based on pedigree information. After tedious mathematical derivations, the authors found that the augmented genomic relationship matrix **H** can be written as:

$$\mathbf{H} = \begin{pmatrix} \mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{12}} + \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{G}\mathbf{A_{22}^{-1}}\mathbf{A_{12}} & \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{G} \\ \mathbf{G}\mathbf{A_{22}^{-1}}\mathbf{A_{12}} & \mathbf{G} \end{pmatrix} \quad [16]$$

Where 1 stands for non-genotyped animals and 2 stands for genotyped animals. In other words, A_{11} is the pedigree-based relationship matrix between non-genotyped animals, A_{22} is the pedigree-based relationship matrix between genotyped animals, and A_{12} is the pedigree-based relationship matrix between non-genotyped and genotyped animals. G is the genomic relationship matrix between genotyped animals.

Moreover, the inverse of the matrix **H** was found to possess a notably straightforward structure as follows.

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{pmatrix} \quad [17]$$

The matrix $\mathbf{A}^{-1}$ is known using Henderson's method, and thus we can easily obtain $\mathbf{A}_{\mathbf{22}}^{-1}$. The inverse of matrix **G** can be obtained by using blending or biding methods. Henderson mixed model equations can support the H matrix which gave the single-step method an immediate application for genomic evaluations (Legarra et al., 2014). For a single trait model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \mathbf{W}\boldsymbol{u} + \boldsymbol{e} \quad [18]$$

Where $var(\boldsymbol{u}) = \mathbf{H}\sigma_u^2$, and $var(\boldsymbol{e}) = \mathbf{I}\sigma_e^2$, the MME can be written as:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\lambda \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{y} \\ \mathbf{W}'\boldsymbol{y} \end{pmatrix} \quad [19]$$

Genomic selection has gathered significant attention since the early 2000s, particularly following the initiation of the Bovine Genome Project in 2003. Various companies and research laboratories have committed resources to developing genotyping panels with a higher number of single nucleotide polymorphisms (SNPs). In 2005, ParAllele BioScience (Affymetrix, Santa Clara, CA) released the MegAllele Genotyping Bovine 10,000-SNP panel, which included 10,410 SNP markers. However, the distribution of markers across the genome was suboptimal, prompting further research to create a more comprehensive SNP assay with markers spanning each block of linkage disequilibrium (LD) across the entire genome. In 2010, genomic selection experienced a significant breakthrough with the introduction of two commercial genotyping chips from Illumina: a low-density 3K chip with 2,900 SNPs and a high-density chip with 777,962 SNPs. The following year, the release of the BovineLD chip, which contained 6,909 SNPs, marked a notable improvement and facilitated the practical application of genomic selection.

In recent years, the implementation of genomic selection (GS) in various countries has been facilitated by the development of commercial genotyping chips. New Zealand took the lead by introducing the BovineSNP50 chip in 2008, while Australia followed suit in 2011 with its initial genomic evaluations. Canada also joined in, collaborating with the USDA to release its genomic evaluations in 2009. To enhance the accuracy of evaluations, EuroGenomics was established in 2009 as a collaboration between breeder-owned companies from Belgium, Denmark, Finland, France, Germany, the Netherlands, and Sweden. This collaboration led to the development of a data-exchange system for the BovineSNP50 BeadChip and a customized Netherlands chip, resulting in a reference population of 18,500 bulls by March 2010. The USA has genotyped over 6 million dairy cattle for genomic purposes (Guinan et al., 2022; Wiggans et al., 2011, 2017). Consequently, the implementation of GS in livestock has remarkably

enhanced genetic gain as it increases the reliability of estimated breeding values and substantially reduces the generation interval. For example, the application of genomic selection to dairy herds has achieved more than two times the rate of genetic gain reached using conventional breeding programs (Pryce & Hayes, 2012). In fact, the accuracy of genomic prediction can reach 0.8 for production traits and 0.7 for fertility and other traits, if the reference population is large enough and if many animals in the reference population are progeny-tested bulls with a high number of daughters. Guinan et al. (2022) reported that the application of GS in US dairy cattle has changed the genetic trends of various traits. The change was noticeable particularly for Holstein breed as it has the largest reference population, and for milk production trait, as it has been, historically, emphasized in selection indexes, constituting 52% in 1971 index.

Genotyping young bulls results in the highest genetic gain. In their study, Buch et al., (2012) indicated that selecting sires to become active based on their genomic estimated breeding values (GEBV) as soon as they reach sexual maturity, referred to as a "turbo scheme," leads to the highest annual gain in the aggregate genotype. This finding aligns with de Roos et al. (2001), who also observed that a turbo scheme yields the highest annual gain when genetic markers account for 20% or more of the genetic variation. However, in the context where the price of genotyping has extremely decreased and become affordable, additionally to the development of low-density panels with even lower prices together with the increase of imputation software accuracy, genotyping females can be beneficial for the selection scheme. For example, in dairy cattle, the population of dairy cows is way larger than bulls, thus in order to increase the size of the reference population, genotyping females with good phenotypic records can help increase the accuracy of GS. Kemper et al. (2015) mentioned that adding 10,000 and 5000 cows to a reference population of Holstein and Jersey cattle respectively resulted in an increase of 5-8% in the accuracy depending on the trait.

Additionally, genotyping females can help identify replacement heifers. The farm's profit from selecting replacement heifers can be assessed using an adaptation of the breeder's equation (Falconer and Mckay, 1996), as follows:

$$SI \times rel \times SD_A PR \qquad [20]$$

Where *SI* is the selection intensity, *rel* is the individual reliability of bulls, $SD_A PR$ is the standard deviation of APR which is the dollar advantage of the heifers. For instance, the standard deviation of APR within the Australian Holstein breed is AU$80.4 (Pryce & Hayes, 2012). Furthermore, genomics enhances the

reliability of heifer breeding values. Based on the trait, the reliability of estimated breeding values can attain 60% at birth through genomics, which is comparable to a cow that has undergone three to four lactations without genomics (Pryce & Hayes, 2012). As a result, genotyping heifers can accurately identify elite female to become dams of the next generation. These selected heifers can also be sold for higher prices providing additional income for dairy farms. Moreover, the development of reproductive technologies, including sexed semen and embryo transfer, presents opportunities to select cattle along the breeding pathway from cow to cow (Pryce et al., 2012).

Genotyping female dairy cattle serves a crucial role in the development of strategic mating plans. From a technical perspective, these plans are designed to optimize the genetic quality of female progeny while concurrently minimizing inbreeding coefficients and the probability of producing offspring homozygous for deleterious recessive alleles (Bérodier et al., 2021). The incorporation of female genotypic data can facilitate the correction of parentage, particularly in herds with imprecise pedigree information or those experiencing high-volume calving events within compressed timeframes. Moreover, this approach has demonstrated the capacity to mitigate inbreeding while maintaining genetic gains associated with profitability. Although pedigree-based methods have proven their efficacity, research has indicated that genomically controlled inbreeding strategies are approximately twice as effective when evaluated on a genomic scale (Pryce et al., 2012).

In conclusion, genomic selection has been implemented successfully across diverse species. In the context of dairy cattle, this approach has facilitated significant alterations in genetic trends for various traits, particularly enhancing the improvement of yield-related characteristics. The utilization of genomic relationship matrices has enabled researchers to not only track blood relationships but also elucidate marker similarities (identity by state) among unrelated animals. This methodological advancement has resulted in a substantial increase in the reliability of estimated breeding values. Consequently, young bulls have become viable for breeding at an earlier age, leading to a notable reduction in generation intervals and a concomitant increase in genetic gain. The decrease in genotyping cost, coupled with advancements in molecular technologies, have significantly enhanced the practicality of genomic selection (GS). Numerous countries have already implemented national genetic evaluation systems based on genomic data. Furthermore, international collaborations have accelerated improvements in genetic trends on a global scale. Despite the current efficacy of genomic selection (GS), its long-term advantages may attenuate over time. Several challenges must be addressed to maintain the benefits of GS while mitigating

potential drawbacks. The subsequent section will elucidate the diverse challenges confronting GS and explore future directions for its development and application.

7. Future challenges of genomic selection

Genomic selection has revolutionized the dairy industry, offering unprecedented opportunities to accelerate genetic progress and improve the efficiency of dairy production. However, the implementation of GS has also brought about several challenges that must be addressed to fully realize its potential. One of the primary challenges in implementing genomic selection nowadays is the ever-increasing number of genotyped animals. This directly affects the computational cost of estimating breeding values, calculating reliabilities, and making predictions. Unlike the pedigree-based relationship matrix (**A**), for which a sparse inverse can be directly constructed using Henderson's method, the genomic relationship matrix (**G**) necessitates a computationally expensive matrix inversion. The cost of a direct inversion of G is cubic with the number of animals and thus prohibitively expensive for large datasets. Currently, the majority of software imposes a practical limit of approximately 150,000 individuals (Misztal, 2016). However, this threshold has been significantly surpassed in many breeding programs. For instance, the United States dairy cattle industry now encompasses over 6 million genotyped animals, and more than 950,000 are Holstein cows. This exponential growth in genomic data volume raises critical questions. How can we efficiently invert the **G** matrix when dealing with a large number of genotyped animals? and is it feasible to develop methods that circumvent the need for **G** matrix inversion altogether? Misztal et al. (2014) proposed a recursive method to compute the inverse of the genomic relationship matrix for large datasets. According to Misztal et al. (2014), the genomic relationship matrix comprises lot of redundancies, therefore there's no need to use the observed genomic relationships between all the animals in the population.

The proposed algorithm Proven and Young (APY) is based on recursions of "proven" animals with their phenotypes or their progeny phenotypes. The algorithm splits the population into two subpopulations, core, and noncore animals, and uses only the genomic relationships between the core animals. APY was tested on a population using proven bulls as core animals, the correlation of genomic estimated breeding values (GEBVs) with APY G-1 was greater than 0.99. additionally, when only cows were included in the recursion, the correlation remained high at >0.99. finally, when using random subsets of 5000, 10,000, and 15,000 animals as core animals, the correlations were 0.97, 0.98, and 0.99, respectively, with minimal

variability among replicates. Furthermore, the convergence rates with random subsets were superior, indicating better numerical conditioning (Misztal, 2016).

Thus, the inversion of APY-G has a cubic cost for core animals but a linear cost and memory for noncore animals. It is suggested to randomly select a subset of the population because it is more likely to represent individuals from all generations. However, determining the number of core animals to include is crucial. The number of animals to include can be calculated based on the number of eigenvalues that explain 98 to 99% of the variation in the matrix G, which typically equals the number of independent chromosome segments (ICSs) of a population with limited effective population size (Ne). The number of ICSs (or Me) is limited to 4NeL, where Ne is the effective population size and L is the genome length in Morgans (Misztal et al., 2015). Utilizing the algorithm for proven and young (APY) to invert the genomic relationship matrix within the framework of single-step genomic evaluations can lead to a significant decrease in memory and computational cost. However, the algorithm has also some side effects on genomic EBV. According to Edel et al. (2022), excluding parent animals from the APY-core can result in the omission of their Mendelian sampling deviations in the EBV of their offspring. Also, under specific but not uncommon circumstances, a critical downward bias in genomic predictions for certain animals may arise.

In addition to computational complexity, intensive genomic selection results in the reduction of the genetic variance and changes the genetic correlations between traits. Hidalgo et al. (2020) demonstrated how intensive selection can reduce genetic variation and increase unfavorable genetic correlations. Genetic covariances can deviate from the direction preferred by selection due to the induced correlation between pairs of loci, complicating the achievement of genetic changes in the desired direction for each trait and reducing the genetic gain. Moreover, by reducing genetic variance, genomic selection decreases the effective population size, and thus increases inbreeding rates. There is a consensus that inbreeding depression affects negatively the fitness and viability of dairy cows (Makanjuola et al., 2020). This reduction can be of remarkable economic loss to dairy farmers. Due to intensive artificial selection, Holstein cows are mainly produced from a very small number of bulls (Sieklicki et al., 2020). Therefore, the breed has experienced a rise in inbreeding by 1.4% per generation over the past decade (Misztal et al., 2021). This has been manifested in the performance of dairy cows. According to Smith et al. (1998), each 1% increase in inbreeding results in an increase of 0.55 days for age at first calving and decreases of 6 days in productive life and 4.8 days in milk.

Furthermore, in order to increase genetic gain through genomic selection, breeding programs necessitate the availability of a large phenotyped and genotyped reference population. Meeting these requirements is challenging for small breeds, traits that are difficult to measure, and countries with a limited number of animals or financial constraints that preclude genotyping a large number of animals. One way to address this issue is through collaboration between different countries such as EuroGenomics consortia (Lund et al., 2010) or the collaboration between the United States, Canada, Italy, and the United Kingdom (Schenkel et al., 2009). In fact, in a small population, the accuracy of genomic prediction (GP) depends strongly on the level of relationships between the reference population (IBD). Instead, in a larger reference population, similarities between markers become more important and the effect of family relationships decreases (Schöpke & Swalve, 2016). For instance, VanRaden et al. (2012) proved that there was an increase in the accuracy of GP by 2% when including bulls from Canada, Italy, and the United Kingdom which enlarged the population by 24%. Similarly, Lund et al. (2011) reported that there has been an increase in GEBVs reliabilities by 2% and 13% for protein yield when enlarging the European reference population using Norwegian, German, French, and Dutch Holstein bulls. A combined reference population of cross breed can also be an option to increase the size of the reference population. Although it has been demonstrated that the marker-QTL association is breed-specific, the utilization of a joint reference population can be advantageous for small populations. Furthermore, including female information, as previously mentioned, can enhance the accuracy of GP. This increase in reliability can be even more pronounced when phenotypic data is hard to measure (Buch et al., 2012). In their study, Calus et al. (2013) observed that using a reference population of 1,609 cows increased the accuracy by 4% to 9% compared to using 296 bulls. Additionally, combining both bulls and cows resulted in an accuracy increase of 1% to 5%. In conclusion, cows are indeed less informative compared to males due to the low reliability of their EBV, Boichard et al. (2015) reports that depending on the trait's heritability, one bull can be replaced by 3 to 10 cows to achieve the same accuracy. However, including cows in the reference population can be very advantageous for small populations.

In the current era of expanding precision agriculture, there is a significant increase in the availability of high-throughput phenotyping data (Brito et al., 2020). The advancements in sensor technology have made it possible to automatically record phenotypes of animals such as activity, rumination, milk records, etc. It also allows monitoring of environmental factors such as climate conditions, and temperature and humidity indices (Persa et al., 2021). Additionally, it is more accurate to assess the welfare and health issues of the animals which gained a lot of interest lately. However, the growth of high-throughput

phenotypes raises questions about the efficacy of conventional statistical models in handling such large datasets. Consequently, there is ongoing debate in the scientific community regarding whether non-parametric and non-linear models should be explored as alternatives for incorporating these extensive phenotypic datasets into genetic evaluations. In this context, a plethora of artificial intelligence algorithms were developed (and are currently in expansion) to include high-throughput phenotypes and large genotype data to increase genetic gain. The biggest challenge is that sensor data including machine vision, activity sensors, or acoustic sensing contains high levels of noise to which both conventional and machine learning algorithms are sensitive. The following section will discuss the main theory behind machine learning algorithms, the opportunities, and the challenges of their application in GP.

8. Theoretical background of machine learning algorithms

Machine learning (ML), a branch of artificial intelligence (AI), involves creating algorithms and statistical models that enable computers to learn and make decisions based on data. These ML algorithms are designed to handle vast amounts of data, often surpassing human cognitive capabilities. In animal breeding. the main goal of applying machine learning is to develop intelligent algorithms able to predict either breeding values, and phenotypes, track associations between regions of DNA and traits of interest or reduce the dimensionality of marker datasets.

Machine learning algorithms are typically categorized into supervised learning, unsupervised learning, and semi-supervised learning based on the structure and availability of the input data. In supervised learning, models are trained to predict a target variable using a labeled dataset that fits a tabular representation with multiple columns. Depending on the nature of this variable, supervised learning algorithms perform either regression or classification tasks. Regression involves predicting continuous variables and quantifying them with real values, such as using SNPs to predict milk yield. Classification involves predicting a discrete target variable, such as predicting mastitis and thus classifying cows into healthy or non-healthy. In unsupervised learning, the training data is unlabeled, and the learning process involves extracting patterns and clusters from the data. These algorithms are generally used for clustering or dimensionality reduction tasks. Semi-supervised learning falls between supervised and unsupervised learning, using a dataset containing both labeled and unlabeled data. For example, we can impute the genotypes of dams in a population that have phenotypes and progeny but lack genotypes, using the genotypes of their progeny and parents.

The application of machine learning (ML) models to genomic datasets typically involves a multi-step workflow. Initially, diverse data types are collected, including phenotypic records, environmental data, genomic information, and sometimes omics datasets. This heterogeneous data presents analytical challenges due to the difficulty of uniform representation as numerical vectors. Consequently, preprocessing is crucial and encompasses formatting, cleaning, scaling, and normalizing the data to make it suitable for ML algorithms. The resulting dataset is then partitioned into training and validation sets. The training process requires careful consideration of several factors. Feature selection is paramount, involving the reduction of data dimensionality through the identification of relevant features and elimination of noise, thereby mitigating overfitting, and enhancing model performance. This step is particularly critical in bioinformatics due to the high-dimensional nature of omics data. Genomic feature selection methods include statistical models (e.g., correlation analysis, mutual information, hypothesis testing) and ML-based techniques (filters, wrappers, embedded methods). Regularization is another essential procedure in ML model training, addressing the balance between model complexity and generalizability. Overly complex models may overfit the training data and perform poorly on new datasets, while excessively simplistic models may underfit and fail to capture underlying patterns. Regularization techniques such as L1 and L2 for regression and Bayesian models and dropout for neural networks help mitigate these issues by penalizing model complexity. Hyperparameter tuning is central to optimizing ML algorithms and improving their accuracy. Various optimization methods are employed, including gradient descent, stochastic gradient descent, random search, grid search, Bayesian optimization, and genetic algorithms. Following training, model validation is necessary to assess accuracy. While cross-validation using the original dataset is common, external validation with an independent dataset is highly recommended for final model approval.

Machine learning has demonstrated remarkable potential across a variety of fields, thanks to advancements in algorithms, computational power, and the use of GPUs. These technological advancements have enabled significant progress in areas such as image analysis, computer vision, and natural language processing (NLP). However, the development of machine learning applications in the domain of genomic prediction, particularly for livestock, has been comparatively limited due to several constraints, including the complexity of biological data, the need for extensive and diverse datasets, and challenges in model interpretability. In Chapter One, a comprehensive explanation of the various models applied to animal breeding will be provided, along with a thorough state-of-the-art review of machine

learning applications in animal breeding, highlighting current achievements and identifying areas for future research.

9. A review of Machine Learning models applied to genomic prediction in animal breeding

This section provides a comprehensive review of machine learning algorithms applied to animal breeding. It begins by defining artificial intelligence (AI) and machine learning (ML), outlining the different types of learning—supervised, unsupervised, and reinforcement learning. The review then delves into a range of ML models used in genomic prediction within animal breeding, explaining their theoretical foundations and core concepts. Following this, it discusses numerous studies that have applied ML models in various areas, including genomic prediction, feature selection, and genotype imputation. Finally, the review explores the potential applications of ML in animal breeding in developing countries, highlighting how these technologies could address specific challenges and enhance genetic improvement programs in resource-limited environments.

A review of Machine Learning models applied to genomic prediction in animal breeding

Narjice Chafai[1], Ichrak Hayah[1], Isidore Houaga[2,3], Bouabid Badaoui[1,4]. Published in Frontiers in genetics.

1. Abstract

The advent of modern genotyping technologies has revolutionized genomic selection in animal breeding. Large marker datasets have shown several drawbacks for traditional genomic prediction methods in terms of flexibility, accuracy, and computational power. Recently, the application of machine learning models in animal breeding has gained a lot of interest due to their tremendous flexibility and their ability to capture patterns in large noisy datasets. Here, we present a general overview of a handful of machine learning algorithms and their application in genomic prediction to provide a meta-picture of their performance in genomic estimated breeding values estimation, genotype imputation, and feature selection. Finally, we discuss a potential adoption of machine learning models in genomic prediction in developing countries. The results of the reviewed studies showed that machine learning models have indeed performed well in fitting large noisy data sets and modeling minor nonadditive effects in some of the studies. However, sometimes conventional methods outperformed machine learning models, which confirms that there's no universal method for genomic prediction. In summary, machine learning models have great potential for extracting patterns from single nucleotide polymorphism datasets. Nonetheless, the level of their adoption in animal breeding is still low due to data limitations, complex genetic interactions, a lack of standardization and reproducibility, and the lack of interpretability of machine learning models when trained with biological data. Consequently, there is no remarkable outperformance of machine learning methods compared to traditional methods in genomic prediction. Therefore, more research should be conducted to discover new insights that could enhance livestock breeding programs.

2. Introduction

Farmers and animal breeders have long used artificial selection to produce offspring with specific desired traits. Assessing the performance of animals was based solely on phenotypes for centuries; it was not until the 20th century that pedigree records and performance data became the keys to genetic selection programs (Boichard et al., 2016). Several statistical methods were developed to predict the breeding values of individuals, such as selection index and Mixed Model Equations (MME), which allowed, due to advances in computational power, the Best Linear Unbiased Prediction (BLUP) (Henderson, 1984) to become the most sophisticated approach for breeding value estimation and thus enable accurate selection decisions (Meuwissen et al., 2016). Nevertheless, traditional genetic evaluation techniques are generally more reliable in estimating breeding values for phenotypic traits that can be easily measured and have moderate to high heritability (Boichard et al., 2016). Conversely, traits with low heritability necessitate a substantial quantity of pedigree and phenotype data, which increases the generation interval and subsequently

diminishes the overall genetic improvement accomplished through the breeding program. The emergence of molecular genetics has prompted researchers to delve into a comprehensive investigation of how traits are determined at the DNA level. Numerous studies have been carried out with the aim of pinpointing particular segments within the genome that play a crucial role in accounting for variations in genetic characteristics known as Quantitative Trait Loci. Later in the 1980s to the 2000s, several methods were proposed for marker-assisted selection (MAS) research that incorporate information about QTL in the MME as fixed effects, and thus breeding value estimation is performed by summing the estimated effects for every QTL (Weigel et al., 2017). Nevertheless, the effectiveness of incorporating Quantitative Trait Loci into estimating breeding values was constrained by the sparse distribution of markers that were in linkage disequilibrium with QTL across the entire population. Furthermore, it was discovered that quantitative traits are influenced by a multitude of QTL with relatively minor individual contributions. Meuwissen et al. (2001) proposed a multiple QTL methodology named genomic selection, that estimates breeding values using a dense marker map. Genomic selection assumes that estimating the effects of a large number of single nucleotide polymorphism (SNP) across the genome will enable breeding value estimation without prior knowledge of the location of specific genes on the genome (Eggen, 2012).

In 2007, progress in molecular technology allowed the first assembly of the bovine genome. The Illumina Company and an international consortium introduced a chip to genotype simultaneously over 54,000 SNPs, which revolutionized dairy cattle breeding (Boichard et al., 2016), and consequently, various methods were developed for whole-genome selection in plants and other domestic animal species. Recently, the availability of highthroughput genotyping and the decrease in genotyping costs have made genomic selection a standard method in animal breeding schemes in many countries (Meuwissen et al., 2016). The underlying concept is based on predicting markers effects using phenotypic information and the genomic relationship between individuals of a reference population previously genotyped and phenotyped to forecast the breeding values of a certain trait for a population of genotyped selection candidates (Goddard et al., 2010). Various statistical methods, such as Genomic Best Linear Unbiased Prediction (GBLUP) or Bayesian methods with different prior assumptions, have been developed to predict markers' effects and thus the genomic breeding values of individuals. Nevertheless, these conventional methods were unable to consider non-additive effects such as epistasis and interactions between genotypes (Bayer et al., 2021) which can have a large effect on phenotypes in animal species. Furthermore, genotyping provides ever-increasing marker datasets, which exacerbates the "curse of dimensionality" also known as the "large P, small N" paradigm (Nayeri et al., 2019). Consequently,

traditional linear models became inadequate for capturing patterns and explaining the complex relationships hidden in this mass of large noisy data.

Recently, the development of machine learning (ML) algorithms and the concomitant boost in computational processing power have generated buzz in the scientific community. ML models are known for their tremendous flexibility and their ability to extract hidden patterns in large noisy datasets, such as image-based data (Xiao et al., 2015), massive datasets of heterogeneous records (Li et al., 2018b), or digital data, which is increasing remarkably due to advancements in computer vision, natural language processing (NLP), internet of things (IoT), or computer hardware (David et al., 2019). Genomics, due to the advent of sequencing technologies, became a field where researchers deal with massive, heterogeneous, redundant, and complex omics datasets. Thus, the application of machine learning models in genomics has been investigated in several studies. In this paper, we review the application of ML algorithms to genomic prediction (GP) in livestock breeding. This work is organized as follows: First, we discuss machine learning fundamentals and provide a brief description of common algorithms used in genomic prediction. Second, we outline the different evaluation methods used to assess the performance of ML models. Afterwards, we review some of the published studies concerning the application of ML models in genomic prediction to provide a meta-picture of their potential in terms of prediction accuracy and computational time. Finally, we discuss the potential of applying ML to animal breeding in low- and middle-income countries.

3. Machine learning fundamentals

Machine learning can be defined as a branch of artificial intelligence that empowers computer systems to learn without being voraciously programmed (Sharma and Kumar, 2017). In other words, a learning computer system can be described as a computer whose performance P on task T improves as its experience E increases (Kang and Jameson, 2018). Based on the learning process, machine learning algorithms can be classified into supervised learning, unsupervised learning and reinforcement learning.

3.1. Supervised learning

In supervised learning, the learning process consists of conceiving a meaning from labeled data. Mainly, supervised learning algorithms tend to estimate or predict a response variable y, based on a set of explicative variables x, through a function called predictor $f(x, \beta)$ where β is a vector of model parameters. The performance criterion we use to define the best predictor is called a loss function L, we thus define the best predictor as the predictor who minimizes the loss function L (Crisci et al., 2012;

Pereira and Borysov, 2019). Depending on the nature of the response variable y (continuous or discrete), supervised learning algorithms are applied to either regression or classification problems. If the main task of an algorithm is to predict a numeric value of a continuous target variable, the ML algorithm performs a regression problem. Alternatively, a classification problem consists of training the algorithm using a set of labeled features (discrete variable), to learn how to successfully classify new features accordingly (Kang and Jameson, 2018). Sometimes the training data involves labeled and unlabeled data. This type of learning is called semi-unsupervised learning, and it is considered a class of supervised learning tasks. Anomaly detection is a typical application of semi-supervised learning algorithms (Kang and Jameson, 2018).

3.2. Unsupervised learning

Unsupervised learning consists of finding patterns or clusters in the training data where the target variable is not present. Algorithms learn on their way to discovering interesting structures in the training data (Mahesh, 2018). Since the features fed to the algorithms are unlabeled, there is no way of assessing the accuracy of these algorithms, unlike supervised learning and reinforcement learning. These models are mainly used for clustering and feature reduction (Sharma and Kumar, 2017).

3.3. Reinforcement learning

In reinforcement learning, software agents perceive and interpret their environment, perform actions and get rewards or penalties in return. Explicitly, a reinforcement learning algorithm enables an agent connected to its environment, to choose an action a1 and generate an output y, given an input i and an environment s1. The action changes the environment, and a value is attributed to the transition of the environment's state through a scalar reinforcement signal r. Consequently, the agent chooses actions that increase the sum of values of the reinforcement signal (Kaelbling et al., 1996). Similar to biological systems, animals living in specific environments face fundamental challenges such as locating sustenance, avoiding harm, and reproducing. These environmental conditions are subject to dynamic changes and sudden variations. Consequently, animals must continuously acquire knowledge from their surroundings and adapt their behaviors accordingly (Neftci and Averbeck, 2019) . Similarly, when a robot is assigned the task of navigating a maze in reinforcement learning scenarios, it functions as an agent within this process. In its interactions with the maze environment, the robot seeks to identify optimal paths by taking successive actions (i.e., moving) while simultaneously receiving feedback through rewards for proximity to the exit or penalties for deviating further away or finding no escape route. By integrating these multiple-

step feedback signals into its decision-making processes over time, the robot gradually enhances its navigation capabilities.

In the field of genomic prediction, supervised learning stands out as the most widely employed technique. This approach leverages labeled data to develop and assess models, thereby allowing for more direct predictions based on established patterns. In contrast, less prominence is given to unsupervised learning and reinforcement learning in relation to genomic prediction.

4. Common ML models used for genomic prediction

In the sections below, we present a short description of some widely used machine learning algorithms for genomic prediction.

4.1. Linear regression

Linear regression is a model usually used to forecast the value of a continuous variable y also called label or target variable using ML terminology, through a vector of explanatory variables also called independent variables or features X, and a linear function. If the model involves a single independent variable x, simple linear regression defines the relationship between the variables using the model:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad [1]$$

where β_0 is the intercept term and β_1 is a regression coefficient that represents the variation in the outcome for a 1-unit increase in the value of the independent variable x, and ε represents the error term also called noise. The dependent variable y can be explained with more than one explanatory variable. In that case, we are talking about Multivariate Linear Regression (MLR). The basic model for MLR is (Maulud and Abdulazeez, 2020):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \qquad [2]$$

Linear regression is considered a supervised learning algorithm because we feed the model with a data set containing features x_i and the corresponding values of the target variable y_i, and we expect an accurate prediction of y_j for another set of features x_j. In order to reach sufficient accuracy, the model minimizes the value of a chosen loss function (Nasteski, 2017). The most commonly used loss function for linear regression is Least Squared Error (LSE) (Maulud and Abdulazeez, 2020).

4.2. Logistic regression

Logistic regression is a classification model regularly applied for the analysis of dichotomous or binary outcomes (LaValley, 2008). In other words, logistic regression is used to study the effects of predictor variables on binary or categorical outcomes, such as the presence or absence of an event (Nick and Campbell, 2007). Training data is fed to a model that uses a logistic function in order to predict the probability of the event. Unlike linear regression, logistic regression doesn't require a linear relationship between dependent and independent variables, the model uses a log transformation to the odds ratio defined as the ratio of the probability of the event happening divided by the probability of the event not happening (LaValley, 2008). The logistic regression hypothesis is defined as (Nasteski, 2017):

$$h_\theta(x) = g(\theta^T x) \quad [3]$$

Where the function g is a sigmoid function defined as the following:

$$g(z) = \frac{1}{1 + e^{-z}} \quad [4]$$

Logistic regression uses a Maximum Likelihood Estimation (MLE) loss function, which is a conditional probability. The algorithm assigns each observation to class 0 or class 1 based on whether the probability is greater or smaller than a given threshold, 0.5 for example (Belyadi and Haghighat, 2021).

4.3. Decision trees

Decision Trees (DT), also known as Classification And Regression Trees (CART) is one of the most popular supervised learning algorithms based on recursive partitioning (Jiang et al., 2020). This approach was first introduced by Breiman et al. (1984), and it relies on dividing a heterogeneous large dataset into multiple smaller homogeneous subsets, which leads to a branching structure. This structure (Figure 1) consists of nodes connected through branches. If a node does not represent an incoming edge, it is called a root. Generally, all nodes have one incoming edge and two or more outgoing edges. The nodes with no outgoing edges are called leaves. In decision trees, splitting the training data is performed by answering several questions incrementally from the topmost node to a leaf. A good question can split a heterogeneous dataset into several homogenous subsamples. Decision trees can deal with both classification and regression problems. For continuous variables, the split is performed using a threshold, the rule takes the form $x < s$where s is a threshold over the variable x. Contrary, when the variable is discrete, the split has the form $x \in$ L where L is a subset of possible levels of x. When the target variable

is continuous, which means we are dealing with regression, the predicted value of each subgroup is the average value of y for all observations in the training set assigned to that subgroup (Crisci et al., 2012). In contrast, when *y* is discrete and DT algorithm is dealing with classification problems, the most frequent level of *y* over the leaf observation is assigned to the target value. The basic algorithm used to build decision trees for regression matters is the Iterative Dichotomiser 3 (ID3) which uses the standard deviation reduction (SDR) to generate the decision tree. In classification situations, the ID3 algorithm uses entropy, defined as a measure of the homogeneity of subsamples, and information gain (Choudhary and Gianey, 2017). This method is widely used because of its flexibility and ease of interpretability.

4.4. Ensemble learning

4.4.1. Bagging

Bagging, also called Bootstrap aggregating, is an ensemble method used for assembling multiple versions of a predictor to get an aggregated strong predictor (Breiman, 1996). Given a labeled training set

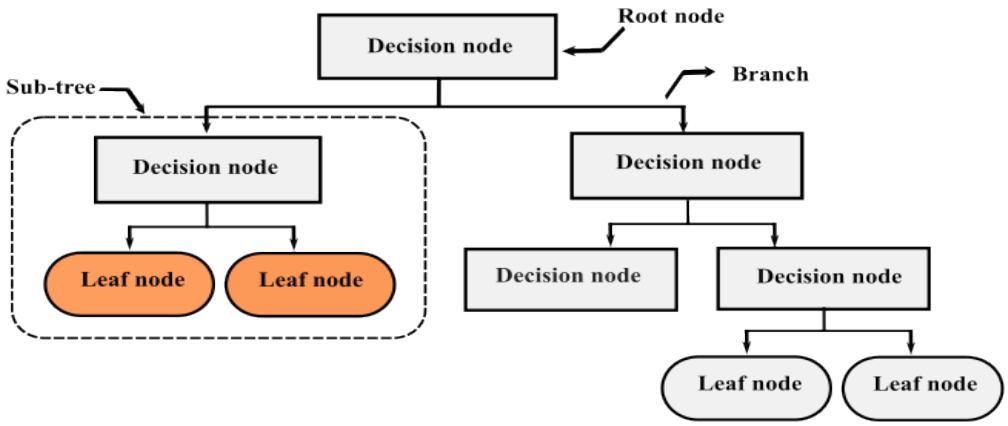

Figure 1: Decision trees structure

(X_1, Y_1) ... (X_n, Y_n), bagging algorithm constructs a bootstrap replicate (X_1^*, Y_1^*) ... (X_n^*, Y_n^*), by randomly selecting samples n times with replacement from the original dataset, and then using them as new learning sets for the CART model. The final model is obtained by repeating these steps *M* times during the learning process. When predicting a numerical outcome, the aggregation algorithm averages the outcome of all predictors. If the target variable is a class label, the bagging predictor is then defined as the majority vote over the M models (Bühlmann, 2012). Bagging algorithms outperformed simple CART models, showing substantial gains in accuracy and significant optimization for weak learners who exhibit unstable behavior. However, bagging algorithms are sensitive to changes in training sets and can slightly reduce the

performance of stable procedures (Breiman, 1996; Freund and Schapire, 1996; Bühlmann, 2012; Crisci et al., 2012).

1.1.1. Random Forest

Random Forest consists of a combination of tree predictors that operates as an ensemble (Breiman, 2001). These decision trees are generated by a randomized tree-building algorithm. The algorithm builds several trees using different random samples of the same size as the original training set by including certain items more than once. Additionally, at each node of the decision trees, the split considers a small random subset of features. As a result, the predictions of these trees can be different. The target value is then assigned to a certain class based on the majority vote over the prediction given by the trees (Kingsford and Salzberg, 2008). Random forests can also be used for regression, in which case the estimated value of the output variable is the average of the predictions of the trees in the forest (Choudhary and Gianey, 2017).

1.1.2. Boosting

Boosting is a strategy used to enhance the accuracy of prediction models. It works by merging multiple simple models, known as weak learners, into one comprehensive and more accurate model. These weak learners, such as basic decision trees, do not have high predictive power on their own. However, when many of them are combined using a boosting algorithm, their collective accuracy significantly improves (Freund and Schapire, 1996).

The Adaboost is one of the most widely used practical boosting algorithms. The learning procedure of this algorithm starts by taking m labeled training examples $S = ((X_1, Y_1) \ldots (X_m, Y_m))$, where x_i belongs to some space *X* and it is represented as a vector of input values, and $y_i \in Y$ is the labeled output associated with x_i. Boosting algorithm runs repeatedly in a series of rounds t = 1, . . ., T, and every weak learner who's given a distribution *Dt*, which refers to the distribution of weights assigned to the examples in the training set *S* at each iteration, finds a weak hypothesis ht: $X \rightarrow Y$. The overall aim of the weak learning algorithm is to find a hypothesis, called weak hypothesis, that minimizes the weighted error t associated to *Dt*. The final outcome of the boosting algorithm is a combination of all the weak hypotheses, where each one is assigned a weight (αt) according to its importance. The more accurate a weak hypothesis is, the higher its weight. This final combination is a kind of "majority vote" of all the weak hypotheses, and it is much more accurate than any of the individual weak learners. Mathematically, the final hypothesis *H* is represented as a weighted majority vote of the weak hypotheses, where every hypothesis *ht* is

multiplied by a weight *αt* (Freund and Schapire, 1996). Boosting is effective at reducing both random variability (variance) and systematic error (bias) in the predictions. It also has a unique feature where it focuses more on the more challenging examples, based on the performance of the previous weak learners. This makes boosting algorithms perform better than other methods like bagging and makes them less sensitive to changes in the training data (Freund and Schapire, 1996).

1.2.Kernel-based algorithms

1.2.1. Reproducing kernel Hilbert spaces (RHKS)

Reproducing kernel Hilbert (RKHS) is a semi-parametric regression model applied for the first time on marker genotypes by Gianola et al. (2011). This method has shown great computational potential, especially when p >> n. RKHS is a Hilbert space (H) of functions where every function can be thought of as a point in Euclidean space, and is assumed to be bounded and linear. In other words, if two functions f and g have close norms f(8) − g(x) → 0 , they also have close values f(x) − g(x) → 0|. The learning task of RKHS can be described as follows: Let xi be a vector of marker genotypes (input), yi a vector of genetic values (output), and g(x) an unknown function of genetic effects.

To infer g, RKHS proceeds by defining a space of functions from which an element ^g will be chosen if it minimizes the loss function bellow:

$$l(g|\lambda) = \|y - g\|^2 + \lambda\|g\|_H^2 \qquad [5]$$

Where λ is a regularization parameter that controls tradeoffs between goodness of fit and model complexity, H represents a Hilbert space, and $\|g\|_H^2$ is the square of the norm of g on H. The square of the norm measures the model complexity. According to Manton and Amblard (2014), RKHS theory can be used to solve three types of problems:

(i) when the problem is defined over a subspace that happens to be RKHS. This suggests that mapping the problem space into a higher dimensional space makes the problem easier. Genomic selection poses a high-dimensional challenge as the number of genotypes (p) typically exceeds the number of individuals (n). By leveraging an RKHS framework, it becomes possible to mitigate this dimensionality and facilitate solving such problems. Introducing a Gaussian kernel allows for transforming the genotypic data into an appropriate RKHS representation, whereby subsequent linear regression models can be effectively used for predicting genetic values within this reduced-dimensional space.

(ii) when a problem has a positive semi-definite function: In the field of genomic selection, a critical component is the genetic relationship matrix (also referred to as the kinship matrix), which quantifies the genetic similarity between individuals. This function serves an important purpose in correcting for confounding factors such as population structure and familial relatedness in association studies. Utilizing a reproducing kernel Hilbert space is one solution to the problem that high-dimensional genotypes present. By applying this approach, we can leverage the kernel trick to effectively handle and make more manageable this complex problem.

(iii) When the data points can be embedded into a RKHS with the kernel function capturing the characteristics of the distance function, given all the data points and a function determining the distance between them Nayeri et al. (2019). One common task in genomic selection is to group individuals based on their genotypes. This is typically done for purposes such as identifying subpopulations or accounting for population structure. To achieve this, the genotypes can be embedded into a reproducible Kernel Hilbert Space using an appropriate kernel function, such as a Gaussian or linear kernel. By doing so, we are able to capture the genetic similarity among individuals. The clustering algorithm operates within this RKHS and aims to find clusters that are well-separated in the RKHS even if they may not appear well-separated in the original genotype space.

1.2.2. Support vector machines

Support vector machines (SVM) is a non-parametric algorithm proposed by Cortes and Vapnik (1995). It was first conceived for two-group classification problems; however, it is widely used nowadays for both regression and classification. When dealing with clustering, the aim of SVM algorithm is to identify an optimal hyperplane defined as a boundary that maximally separates classes (Jiang et al., 2020). When data points are linearly separable, the SVM algorithm performs a linear classification and the optimal hyperplane is found using numerical optimization (Crisci et al., 2012). Otherwise, SVM can perform a non-linear classification using the Kernel function. Gaussian kernel function is used to map the data points from a data space to a high-dimensional feature space. In the feature space, small spheres appear to enclose the image of data, these spheres are mapped back to the data space and form cluster boundaries that enclose data points of the same cluster (Ben-Hur et al., 2001). The boundaries should maximize the margin between them and the classes to minimize the classification error (Mahesh, 2020). When the SVM algorithm is applied to regression problems, the loss function should include a distance

measure. The possible loss functions are the quadratic, Laplacian loss function, Huber and the insensitive loss function (Gunn, 1998). SVM algorithms can result in highly accurate predictions due to their flexibility. However, they're described as a black box because no metrics are provided for how predictors optimize the hyperplane, which makes the predictions hard to interpret (Jiang et al., 2020).

1.3.Nearest neighbors

Nearest neighbors model is one of the most simple and intuitive machine learning algorithms. The idea of this approach is to forecast the value of a target variable yi associated with an input variable xi based on the distance between xi and other data points. Generally, Euclidean distance is used, but there are other methods to calculate this distance, such as Manhattan distance (Zhang, 2016). In classification, yi is assigned to the class label of the majority of the nearest data points in the space. Alternatively, when dealing with regression, the predictor is the average of the output over the nearest neighbors (Crisci et al., 2012). The K-nearest neighbors (KNN) is the most popular algorithm in this category. It is based on the same idea that the nearest patterns to a datapoint xi deliver useful label information. The unknown parameter K decides how many neighbors will be considered in the learning process (Kramer, 2013). The number of neighbors K has a significant impact on the performance of the algorithm. An optimal K is the one that strikes a balance between overfitting (low bias but high variance) and underfitting (low variance but high bias). Some authors suggest K to the square root of the number of observations in the training set (Zhang, 2016).

1.4.Deep neural networks

Deep learning is a family of powerful learning methods capable of recognizing complex patterns in raw data (Vieira et al., 2020). The well-known Rosenblatt "perceptron" proposed in the 1950s was the first attempt to conceive a model closely analogous to the perceptual processes of the human brain (Rosenblatt, 1957). Deep neural networks' (DNN) structure (Figure 2) consists of stacked layers of connected neurons. In other words, the DNN model comprises a certain number of layers, each layer contains several neurons. Each neuron is connected to the neurons in adjacent layers through weights that reflect the strength and direction of the connection (excitatory or inhibitory) (Montesinos-López et al., 2021). DNN models are characterized by their depth, size, and width. The number of layers that a DNN contains, excluding the input layer, is called depth. The total number of neurons in the model is referred to as the size. Finally, the width of the DNN is the layer that comprises the largest number of neurons.

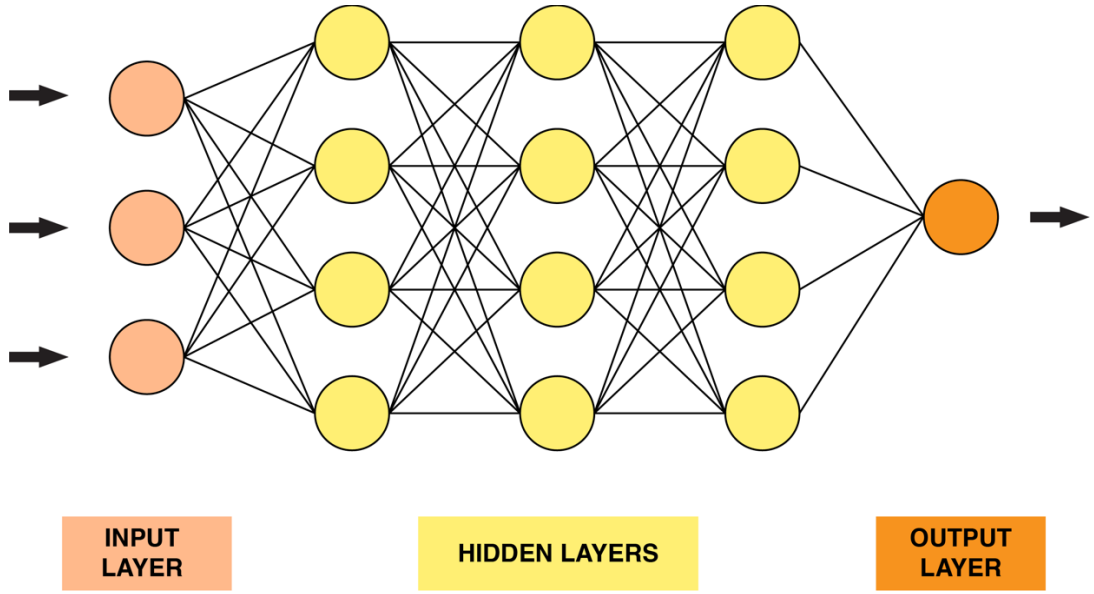

Figure 2: A graphical representation of a simple neural networks.

When running DNN, a set of observations X enter the model through the input layer. The observations x_i are the input and the output of this layer. In the hidden layers of the DNN, every neuron of a given layer receives from the layer of lower hierarchical level, the weighted sum of its neurons' output, and then passes it through an activation function to drive it as an output for that neuron. In the hidden layers, the most widely used activation functions are the rectified linear unit, hyperbolic tangent activation and the sigmoid function. In the output layer, the DNN is meant to perform either a classification or a regression based on the nature of the target variable. When dealing with classification, the number of neurons in the output layer is equal to the number of classes. Additionally, different activation functions could be used according to the type of the target variable. Softmax is used for categorical variables, the exponential function for count data and the sigmoid function for binary outcomes (Vieira et al., 2020; Montesinos-López et al., 2021). In regression problems, the output layer represents the estimated values of the target variables and linear activation functions are applied. The most successful activation function when dealing with a continuous variable is the rectified linear unit (ReLU) (Bircanoğlu and Arıca, 2018). The tanh activation function is used in DNN to introduce non-linearity in the model and to allow the model to learn from both positive and negative weights since it is centered around zero (unlike the sigmoid function). It is typically used in the hidden layers.

Like other ML models, training DNN consists of choosing optimal weights that minimize the differences between real and estimated values of the target variable. The gradient descent is used to

minimize the loss function. These parameters need to be updated during the learning process. When first training the DNN model, the weights are randomly initialized. Once an observation has entered the model, the information is forward propagated through the network until it predicts a certain output value. The gradients of the loss function are then computed using a hyperparameter called the learning rate η, which indicates how big the steps of gradient descent should be, and then used to update the function parameters (weights and biases). Backpropagation is another efficient method of computing gradients. The concept of this method is based on the fact that the contribution of each neuron to the loss function is proportional to the weight of its connection with the neurons of the following layer. Therefore, these contributions could be calculated starting from the output layer and backpropagated through the network using the weights and the derivative of the activation function (Pereira and Borysov, 2019; Vieira et al., 2020; Montesinos-López et al., 2021).

Deep learning comprises a wide variety of architectures. The most popular ones are the feedforward networks, also called the multilayer perceptron (MLP), recurrent neural networks (RNN) and the convolutional neural networks (CNN).

1.4.1. Multiplayer perceptron (MLP)

The multilayer perceptron (MLP) is a layered feedforward network where all layers are fully connected. Every neuron of a given layer is connected to neurons of the adjacent layer, the information flows in a single direction. In other words, there are no intralayer or supralayer connections. MLPs are found to be powerful and simple to train. However, these networks are not suitable to deal with spatial or temporal datasets and they're prone to overfitting (Montesinos-López et al., 2021).

1.4.2. Recurrent neural networks (RNN)

In Recurrent Neural Networks (RNN), information flows in both directions. Every neuron has three types of connections: incoming connections from the previous layer, ongoing connections toward the subsequent layer, and recurrent connections between neurons of the same layer (Montesinos- López et al., 2021). This recursive structure allows this network to have some notion of memory since the output of a layer depends on both current and previous inputs. RNN are frequently used to model space-temporal structures. It is also used in the fields of natural language processing and speech recognition (Pereira and Borysov, 2019; Zingaretti et al., 2020).

1.4.3. Convolutional neural networks (CNN)

Convolutional Neural Networks (CNN) are designed to accommodate situations where data is represented in the form of multiple arrays. The input variable can have one-dimension such as SNPs, two dimensions such as color images, or three dimensions for videos or volumetric images (LeCun et al., 2015). The architecture of CNNs is made up of convolutional and pooling layers followed by fully connected neural networks (Pereira and Borysov, 2019). When training CNNs, the first two types of layers, namely, convolutional, and pooling layers, perform feature extraction. The fully connected neural network is meant to perform the classification or the regression task. In the convolutional layer, a mathematical operation is performed to generate one filtered version of the original matrices of the input data. This convolutional operation is called “kernel” or “filter”. A non-linear activation function, generally ReLU, is applied after every convolution to produce the output, which is organized as feature maps. The pooling operation comes after to smooth out the results, its role is to merge semantically similar features into one. In other words, pooling reduces the number of parameters and makes the network less computationally expensive. Max pooling is a typical pooling operation that proceeds by extracting patches from the feature maps, determining the maximum value in each patch, and then eliminating all the other values. Finally, after turning the input matrices into a one-dimensional vector, the features are mapped by a network of fully connected layers similar to the aforementioned feedforward deep network to obtain the final output, the probabilities of a given feature belonging to a given class for example. The output of the fully connected neural network is fed to another different activation function to perform classification or regression based on the output variable (Yamashita et al., 2018). CNNs have been successfully applied in visual and speech recognition, natural language processing, and various classification tasks (LeCun et al., 2015; Yamashita et al., 2018; Pereira and Borysov, 2019).

2. Performance fitness and error metrics

Machine learning algorithms need to be rigorously evaluated in order to confirm their validity in understanding complex datasets and hence extend the use of this model in different datasets. Generally, the performance of ML models is assessed using Performance Fitness and Error Metrics (PFEMs), defined as mathematical constructs used to measure how close the predicted and real observed values of a given variable are. Choosing the right metric for assessing the performance of a predictor is very delicate because a limited understanding of the behavior of algorithms can lead to misinterpretations of results and thus

false assumptions. In addition, PFEMs are used differently when dealing with regression and classification problems.

In regression, performance metrics are based on calculating the distance between predicted and real values using subtraction or division operations, sometimes supplemented with absoluteness or squareness. Moreover, PFEMs in regression also investigate the distribution of residuals, whether it is random or regular, which indicates that the regression model does not explain all the regularity in the dataset. The most common PFEMs used in regression are (Table 1): mean square error (MSE) or root mean square error (RMSE), normalized mean squared error (NMSE), correlation coefficient (R), r squared (R2), mean absolute error (MAE), and mean absolute percentage error (MAPE). They are easy to interpret, straightforward, and they indicate the magnitude of the difference between measured and predicted values (Naser and Alavi, 2021). The interpretation of these metrics can be found elsewhere (Botchkarev, 2018).

Table 1: Common performance metrics used for the evaluation of regression models.

Metric abbreviation	Metric Name	Metric Formula
MSE	Mean Squared Error	$MSE = \frac{1}{N}\sum_{n=1}^{N}[y(n) - \hat{y}(n)]^2$
RMSE	Root Mean Squared Error	$RMSE = \sqrt{MSE}$
NMSE	Normalized Mean Squared Error	$NMSE = \frac{\sum_{n=1}^{N}[y(n) - \hat{y}(n)]^2}{\sum_{n=1}^{N}[y(n) - y]^2}$
MAE	Mean Absolute Error	$MAE = \frac{1}{N}\sum_{n=1}^{N}\lvert[y(n) - \hat{y}(n)]^2\rvert$
MAPE	Mean Absolute Percentage Error	$MAPE = \frac{100}{N}\sum_{n=1}^{N}\left\lvert\frac{[y(n)-\hat{y}(n)]^2}{y(n)}\right\rvert$
R2	Coefficient of determination	$R^2 = 1 - NMSE$

Where N(1 , ..., n) is the number of observations, yn refers to observed values, and yn refers to the estimated values.

Classification models are meant to categorize data into distinct classes. Therefore, assessing the performance of classifiers relies on a confusion matrix where columns represent the predicted values, while rows represent the actual values as described in Figure 3, where TP refers to true positives, TN

denotes true negatives, FP denotes false positives, and FN refers to false negatives. The performance of classifiers is often evaluated using prediction accuracy (PAC), sensitivity or recall, specificity, and precision. Based on the confusion matrix, these metrics are defined as below:

$$PAC = \frac{TP + TN}{TP + FP + TN + FN}, \quad precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}, \quad [6]$$

$$specificity = \frac{TN}{TN + FP}$$

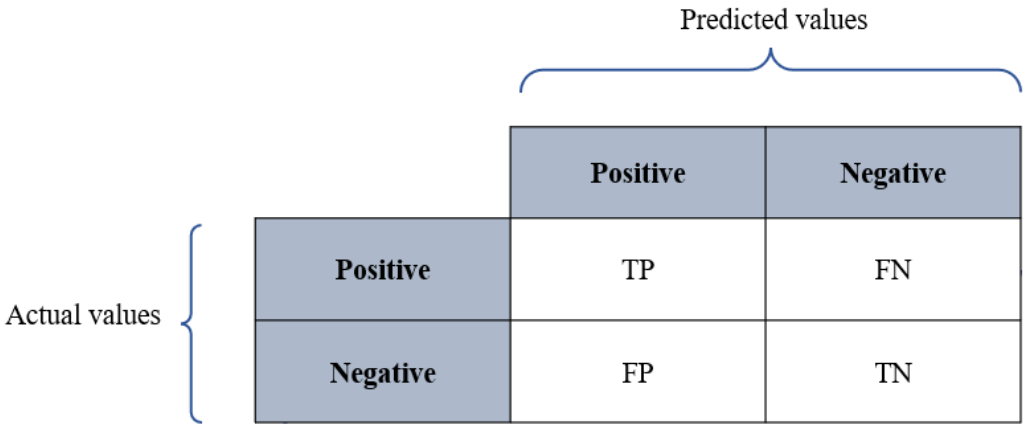

Figure 3: Confusion matrix

Other methods based on the aforementioned metrics have also been broadly used in assessing the performance of classifiers. The F1 score that combines both precision and recall in a harmonic mean in the following formula:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \quad [7]$$

Moreover, Matthews (1975) introduced a coefficient used to measure the performance of binary classifiers, called the Matthews correlation coefficient (MCC). This coefficient combines all four

measures in the confusion matrix, and thus it is qualified as the most informative metric especially when a significant imbalance in class sizes is noticed (Nayeri et al., 2019). MCC formula is represented below:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad [8]$$

Another criterion widely used to measure the performance of classifiers is the Area Under the Receiver Operating Characteristic (ROC) curve (AUC). The ROC curve visualizes the tradeoff between sensitivity and specificity. In other words, the curve captures the ratio of false to true positive rates under variation of the decision threshold (Hoffmann et al., 2019). Generally, good performance is detected when the curve is high and close to the left in the ROC space. In contrast, an inaccurate method has a curve close to the main diagonal (figure 4). Thus, when comparing several ML models, the one with the highest AUC value is the most accurate (Metz, 1978).

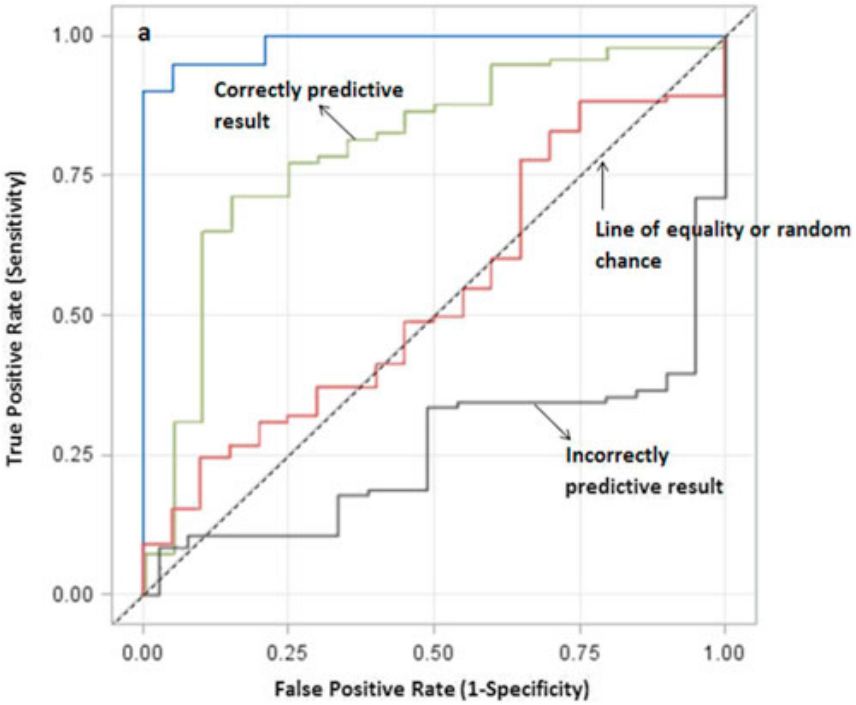

Figure 0-1 Interpretation of ROC curves of varying sensitivity and specificity. The sensitivity and the specificity of the test increases as the curve approaches the point a (x = 0, y = 1). The closer the curves are to the diagonal line the less precise they are.

3. Machine learning models applied to genomic prediction in animal breeding

Machine learning algorithms have been widely used in various fields. Their ability to discover patterns in large, messy datasets has driven researchers to investigate their performance in dealing with complex models and nonlinearities in large datasets. Animal breeding in the post-genomic era is a domain that deals with high-dimensional marker datasets such as genomics, epigenomics, transcriptomics, proteomics and metabolomics. The most commonly used marker data sets in animal breeding are single nucleotide polymorphism (SNPs) data sets that represent the genetic variation in a genome. SNP markers data sets are very large, for example, the data set resulting from genotyping 2,000 individuals for 10,000 SNP markers, contains 20 million data points. Furthermore, they can be complex and noisy due to genotyping errors, missing data, batch effects, and biological variability. Copy number variation (CNV) is another valuable form of genetic variation that complements SNPs analysis. CNV datasets are used to investigate diversity within populations (Yang et al., 2018). They can serve as informative markers for marker-assisted selection by identifying CNVs associated with desirable traits (Ma et al., 2018), and genomic prediction to enhance the accuracy of predicting breeding values (Hay et al., 2018), etc. In addition, microarray data provide valuable information concerning gene expression, by measuring the mRNA expression levels of tens of thousands of genes. Gene expression datasets are known to be massive (large number of genes) and redundant, and thus, their manipulation requires a lot of preprocessing and dimensionality reduction (Liu and Motoda, 2007). Applying machine learning models is hence becoming attractive in genomics, due to their potential in dealing with large, noisy data and modeling minor nonadditive effects as well as interactions between phenotypes and genotypes.

Machine learning models have several important applications in genomics. Through the introduction of sophisticated algorithms and computational models, ML can be trained using large datasets of genotypes and phenotypes to predict animals' breeding values for certain traits. This would enable an accurate selection of animals with the highest genetic merit and allow for more informed breeding decisions. ML models have successfully been implemented to predict genomic breeding values across various animal species, including dairy cattle (Beskorovajni et al., 2022), beef cattle (Srivastava et al., 2021), pigs (Zhao et al., 2020), and broilers (González-Recio et al., 2008). The estimated GEBVs provide an accurate prediction of animals' genetic potential and thus identify animals with high genetic potential that surpass the population average. Therefore, ML models can have a valuable role in allowing breeders to make more precise breeding decisions, leading to faster genetic progress. In addition, machine learning algorithms can also be deployed to predict disease occurrence based on integrated information of

genotypes and health records. For example, Ehret et al. (2015) applied ML to encounter a serious health problem in the intensive dairy industry, which is subclinical ketosis risk. The authors proposed an ANN to investigate the utility of combining metabolic, genomic and milk performance in predicting milk levels of β-hydroxybutyrate. Data comprised SNP markers, and weekly records of the concentrations of glycerophosphocholine, phosphocholine, and milk composition data (milk yield, fat and protein percentage). The deep learning model deployed provided an average correlation between real and predicted values up to 0.643 when incorporating information about metabolite concentration, milk yield, and genomic information.

Moreover, ML models can be coupled with GWAS and population genomics to identify genetic variants and biological pathways linked to specific phenotypic traits. A deep learning framework was proposed by Zeng et al. (2021) to predict quantitative phenotypes of interest and discover genomic markers considering the zygosity of SNP information from plants and animals as input. Furthermore, ML models can be used to impute moderate-density genotypes when genotyping large populations can be expensive and time-consuming. ML models can accurately infer missing genotypes and fill the gaps to create moderate density genotypes. This has already been implemented in the beef cattle genomic dataset (Sun et al., 2012).

Taken together, ML models appear to be a powerful tool for enabling more accurate predictions, targeted selection, and an improved understanding of genetic mechanisms. However, when training ML models on biological data, several challenges can occur. For example, when using markers data, environmental data, and phenotypic records all together to predict a certain variable, the large heterogeneity of the input data can be a hurdle. Therefore, it is indispensable to perform a pre-processing step that includes formatting, cleaning, scaling, and normalizing the data. This step ensures that the data is prepared to optimize the performance and accuracy of the machine learning model. Markers data sets are usually massive and comprise a lot of noise. Using the raw data can lead to a low performance and overfitting. Thus, performing feature selection is vital when manipulating omics data in order to reduce the dimensionality of the data by selecting relevant features while eliminating noise from the model. Multiple methods can be used to perform feature selection including statistical methods, correlations, or hypothesis testing. Recently, ML models were proved to be very powerful in feature selection. The most broadly used machine learning-based methods for feature selection are filters, wrappers, and embedded methods that combine filter and wrapper methods (Tadist et al., 2019). Machine learning-based feature

selection is widely used when manipulating animal species marker data sets. Finally, when training ML models on biological data, several steps should be performed to ensure the quality of the data fed to the model. In addition, adjusting the hyperparameters and generalizing the model through regularization techniques are also central to optimizing the performance of the model. There are multiple techniques to optimize ML models, such as gradient descent, stochastic gradient descent, random search, grid search, Bayesian optimization, and genetic algorithms. Now that we have discussed the overall applications of ML models in genomic prediction and the multiple issues encountered while implementing those models on markers data, we will review, in this section, some of the published studies on the application of different ML models for genomic prediction in animal breeding, feature selection, and genotype imputation separately, to provide a meta-picture of their potential in terms of prediction accuracy and computational time. Data sets and different machine learning models applied to genomic prediction in a handful of the reviewed papers are summarized in Table 2. In Supplementary Materials; Table 1 contains the full summary of the reviewed papers, and Table 2 presents the programming languages and packages used to train the models in the aforementioned studies.

Table 2 Machine learning models applied to genomic prediction in animal breeding.

Year	Authors	Species	Breed	No. of individuals	No. of markers	Response variable	ML algorithms	Aim of the study
2016	Naderi et al.	Dairy cattle (simulated)	-	20000 females and 400 males	50025 and 10005 SNPs	Subclinical Ketosis	ANN (MLP)	Building an ANN for an earlier prediction of subclinical Ketosis in lactation.
2016	Yao et al.	Dairy cattle	Holstein	3000 genotyped 792 genotyped and phenotyped	57491 SNPs	RFI	SVM (semi-supervised learning)	Describing a SVM-based semi-supervised learning model, and applying it for genomic prediction of residual feed intake.
2018	Li et al.	Beef cattle	Brahman	2093	40184	BW	RF, GBM, XGBoost	Assessing the efficiency of three ML methods in identifying the top-ranked SNPs and using the subsets of SNPs to

								construct genomic relationship matrices for estimating genomic breeding values.
2020	Liang et al.	Beef cattle	Simmental	1217	671900 SNPs	CW, LW, EMA	Adaboost.RT (integrated SVR), KRR, RF	Applying ensemble learning models to predict genomic breeding values of three economic traits.
2020	Abdollahi-Arpanahi et al.	Dairy cattle	Holstein	1170	57749 SNPs	SCR	MLP, CNN, RF, GB	Comparing the predictive performance of two deep learning methods, two ensemble learning methods, gradient boosting and two parametric methods (GBLUP and Bayes B).
		Simulated data	-	-	100 and 1000 QTNs	A quantitative trait		
2021	Chen et al.	Beef cattle	Nellore	18	16,423 genes	FE	RF, XGBoost, RX, SVM	Applying Rf, XGBoost and RX to identify small subsets of biologically important genes to classify animals into High Feed Efficiency and Low Feed Efficiency.
2021	Srivastava et al.	Beef cattle	Hanwoo	7324	53866 SNPs	CWT, MS, BFT EMA	RF, XGB, SVM	Comparing the predictive ability of three ML models in predicting phenotypes from genotypes.
2021	Wang et al.	Pig	Yorkshire	2566	44922 SNPs	TNB, NBA	SVR, KRR, RF, Adaboost.R2	Exploring and comparing the prediction ability of four ML models to GBLUP, ssGBLUP

								and bayesian methods in genomic prediction of reproductive traits.
2021	Beskorovajni et al.	Dairy cattle	Holstein	92	-	MFP, MPP, CM, FM, LIV, SCE, HCR, CCR, DSB, SSB, GL	MLP	Predicting yield and fertility traits using an MLP model based on the Broyden-Fletcher-Goldfarb-Shanno iterative optimization algorithm for genomic selection.
2021	An et al.	Beef cattle	Simmental	1301	671990 SNPs	Cosine Kernel based KRR (KcRR),SVR	LW, CW, EMA	Assessing the prediction accuracies of 12 traits with various heritabilities and genetic architectures using parametric methods (GBLUP and Bayes B), and two machine learning models (KcRR and SVR)
		Dairy cattle	Holstein	5024	42551 SNPs		MY, MFP, SCS	
		Pig	-	3534	43494, 43407, and 43412 SNPs for each trait		T1, T2,T3	
		Simulated data	-	4000	50 SNPs for each trait (3 traits)		T1, T2,T3	

3.1. Genomic prediction

The wide majority of traits of interest in animal breeding are presumed to be influenced by many genomic regions with complex interactions. Kernel-based methods are gaining consideration over conventional regression models due to their capacity to capture nonadditive effects. A more succinct description of kernel-based methods applied to GP can be found in Morota and Gianola (2014). González-Recio et al. (2008) used the F-metric model, kernel regression, reproducing kernel Hilbert spaces (RKHS) regression, and Bayesian regression to predict mortality in broilers and see how well they did compared to the standard genetic evaluation (E-BLUP), which is only based on pedigree information. The dataset

contained records for mortality rates for 12167 progeny of 200 sires with a total of 5523 SNPs. The authors concluded that kernel regression and RKHS regression had a low residual sum of squares and increased the accuracy from 25% to 150% relative to other methods, and thus the authors recommended their utility in the genomic prediction of early mortality in broilers. An et al. (2021) developed another kernel-based algorithm named Cosine Kernel-based Ridge Regression (KcRR) to perform genomic prediction using simulated and real datasets. The simulated dataset included 4000 individuals and concerned three quantitative traits with various heritabilities (0.36, 0.35, and 0.52). Meanwhile, the real data concerned three species: a Chinese Simmental beef cattle dataset contained 1,301 bulls, with a total of 671990 SNPs and concerned three traits of interest: live weight (LW, kg), cold carcass weight (CW, kg), and eye muscle area (EMA, cm2). The pig dataset included 3,534 animals, and finally, the German Holstein cattle dataset included 5,024 bulls with a total of 42551 SNPs that concerned three phenotype traits, milk yield (MY, kg), milk fat percentage (MFP,%), and somatic cell score (SCS). The designed model consisted of a kernel-based ridge regression, which is a ridge regression built in a higher dimensional feature space that uses a Cosine similarity matrix (CS matrix) instead of the genomic relationship matrix (G matrix). The difference between these two matrices is that the CS matrix measures the cosine of the angle between two projected vectors, and the G matrix in an m-dimensional feature space where m is the number of SNP markers. For comparison purposes, a 20-fold cross-validation approach was used to evaluate the prediction accuracy of KcRR to that of GBLUP, BayesB, and SVR. The authors have also simulated for the quantitative traits different heritabilities, and genetic architectures, including one major gene and a large number of genes with minor effects, a number of genes with moderate effects and many genes with small effects, and finally a large number of genes with small effects, in order to assess the performance and consistency of these methods. Overall, KcRR had the best prediction accuracy among the methods, in addition, it performed stably for all traits and genetic architectures, which confirms its reliability and robustness. Therefore, An et al. (2021) suggested the use of KcRR and the CS matrix as a potential alternative in future GP. Zhao et al. (2020) investigated the performance of SVM in a pig dataset containing 3,534 samples with a different number of SNPs for each trait respectively 45,025, 45,441, 44,190, 44,151, and 44,037 SNPs for T1, T2, T3, T4, and T5. For training the SVM model, a suitable kernel function was selected. The authors tested the prediction ability of four commonly used kernel functions namely, the Radial Basis Function (RBF), the Polynomial Kernel Function, the Linear Kernel Function, and the Sigmoid Kernel Function in previously published pig and maize datasets. The findings demonstrated that SVM-RBF had the best performance, the SVM-sigmoid and the SVM-poly models had

similar accuracies, and the SVM-linear had the lowest accuracy. As a result, the authors chose using the SVM-RBF model to adjust the hyperparameters of the final SVM model. Afterwards, the authors evaluated the performance of SVM-RBF, GBLUP and BayesR in fitting the five pig datasets, using a 10-fold cross validation approach. Overall, the performance of the trained models was similar. However, the SVM model performed better than BayesR but worse than GBLUP in terms of time, and better than GBLUP but worse than BayesR in terms of memory.

Ensemble learning has been broadly used in the genomic prediction of animal breeding values. Naderi et al. (2016) studied the use of RF for genomic prediction of binary disease traits using simulated data from 20,000 cows with different disease incidence scenarios, different heritability (h2 = 0.30 and h2 = 0.10), and different genomic architecture (725 and 290 QTL, populations with high and low levels of linkage disequilibrium). The training set contained 16,000 healthy cows, and the testing data contained the remaining 4,000 sick cows. Afterwards, the number of sick cows was increased progressively by moving 10% of the sick individuals to the training data, ensuring that the size of both the training and testing data remained constant. This study compared the performance of RF and GBLUP using the correlations between estimated genomic breeding values and true breeding values, and the area under the curve (AUROC). The results confirmed that RF had a great advantage in the binary classification for scenarios with a larger marker density. In addition, the best prediction accuracies of RF (0.53) and GBLUP (0.51), and the highest values of AUROC for RF (0.66) and for GBLUP (0.64), were achieved using 50,025 SNPs, a heritability of 0.30, 725 QTL, and a disease incidence similar to the population disease incidence (0.20). The authors also noted that the genetic makeup of the population had an impact on the performance of RF and GBLUP. However, the variability was more pronounced for RF than for GBLUP.

A boosting algorithm called L2-Boosting was suggested by González-Recio et al. (2010) to forecast the progeny test predicted transmitting abilities for the length of productive life (PL) in a dairy cattle dataset, and the average food conversion rate records in a broiler dataset. The dairy cattle data set consisted of 4702 Holstein sires with a total of 32611 SNPs, and the broiler dataset comprised 394 sires of a commercial broiler line with 3,481 SNPs. The L2-Boosting algorithm proceeds by combining two weak learners, namely, ordinary least squares (OLS) and nonparametric (NP) regression. The performance of OLS-Boosting and NP-Boosting was compared to Bayesian LASSO (BL) and Bayes A regression. The results showed that OLS-Boosting had the lowest bias and mean-squared errors (MSEs) in both the dairy cattle (0.08 and 1.08, respectively) and the broiler (0.011 and 0.006, respectively) data sets. The authors

concluded that L2-Boosting with a suitable learner represents a good alternative for genomic prediction, providing high accuracy and low bias in a short computational time.

In another study, a bagging approach using GBLUP (BGBLUP) was performed to predict the genomic predicted transmitting ability (GPTA) of young Holstein bulls for three traits: protein yield (PY), somatic cell score (SCS), and daughter pregnancy rate (DPR) (Mikshowsky et al., 2017). The dataset consisted of 17,276 Holstein bulls with a total of 57,169 SNP markers, and it was split into a reference population set used to train the model and a testing set for the evaluation. The aim of the proposed bagging approach was to create 50 bootstraps containing bulls selected randomly, with replacement, from the reference population, until each bootstrap reaches the same number of individuals as the original reference population. GBLUP was applied to predict the GEBVs of individuals for each trait. According to the results, GBLUP outperformed BGBLUP in the genomic prediction for PY, SCS, and DPR, the correlations between the real and predicted values of each trait for GBLUP were 0.690, 0.609, and 0.557, and 0.665, 0.584, and 0.499 for BGBLUP. In summary, the authors found no advantage to using BGBLUP over GBLUP for genomic prediction.

For comparison purposes, several studies have deployed various machine learning methods to forecast and compare their predictive accuracies when trained using genomic data. For example, Ogutu et al. (2011) compared the performance of three machine learning models, namely RF, stochastic gradient boosting, and SVMs, in estimating genomic breeding values. A simulated dataset of 2,326 genotyped and phenotyped individuals and 900 individuals who lacked phenotypic records was used. As a performance metric, Pearson correlations were used between the simulated values and the predicted values from the validation set, as well as between the predicted and real breeding values for non-phenotyped individuals. The results showed that stochastic gradient boosting and SVM had better correlations between the simulated values and predicted values compared to RF. However, RF provided reasonable rankings of the SNPs, which can be useful for identifying markers for further testing. In conclusion, stochastic gradient boosting and SVM are found to be able to accommodate complex relationships and interactions in marker data such as epistasis. They have also outperformed RF in the genomic prediction of the quantitative trait, however, SVM was computationally intensive due to the grid search for tuning the hyper-parameters. In contrast, Srivastava et al. (2021) found different conclusions when evaluating the performance of RF, XGB, and SVM in predicting four traits namely, carcass weight (CWT), marbling score (MS), backfat thickness (BFT) and eye muscle area (EMA) of 7234 Hanwoo cattle. According to this study, XGB yielded

higher correlations for CWT, MS, (0.43, 0.44, respectively) compared to GBLUP (0.41, 0.42), and lower (0.23, and 0.31) than GBLUP (0.35, and 0.38) for BFT, and EMA. Meanwhile, GBLUP delivered the lowest MSE for all traits. Among the ML methods, XGB had the lowest MSE for CWT and MS, and SVM provided the lowest MSE for BFT and EMA. Despite the good performance of XGB and SVM, the authors still concluded that there was no advantage to using ML methods over GBLUP.

Liang et al. (2021), compared the performance of Adaboost.RT, SVR, KRR, RF to the conventional GBLUP in predicting breeding values for cattle growth traits in Chinese Simmental cattle (carcass weight, live weight, and eye muscle area), using a dataset of 1,217 young bulls with a total of 671990 SNPs. Contrary to the previous study, the authors recommended using ML methods over GBLUP. Indeed, the predictive accuracies of SVR, KRR, RF, Adaboost.RT and GBLUP were 0.346, 0.349, 0.315, 0.349, and 0.290 respectively. In other words, ML methods improved the predictive accuracy by 12.8%, 14.9%, 5.4%, and 14.4%, respectively, over GBLUP. In summary, Liang et al. (2021) found a great advantage in using ML algorithms for GP in Simmental beef cattle, especially Adaboost.RT due to its reliability. However, the authors pointed out that ML models were sensitive to data, which means that two different datasets may have significant differences in predictive accuracy. Wang et al. (2022) used a pig dataset of 2,566 Chinese Yorkshire pigs to compare the same models. The study concentrated on estimating the genomic breeding values of these individuals for two reproductive traits: the total number of piglets born (TNB) and the number of piglets born alive (NBA). The GEBVs were also estimated using classical methods [GBLUP, ssGBLUP, and Bayesian Horseshoe (BayesHE)]. Overall, ML methods outperformed conventional ones, and the degree of improvement over GBLUP, ssGBLUP, and BayesHE was 19.3%, 15.0% and 20.8% respectively. Furthermore, results showed that ML methods had the lowest MSE and MAE in all case scenarios. SVR and KRR provided the most consistent prediction abilities including higher accuracies and lower MSE and MAE. The findings of this study showed that ML methods are more efficient and had better performance in predicting GEBVs for reproductive traits, which can provide new insights for future GP. In another report, Sahebalam et al. (2019) evaluated the predictive ability of RF, SVM, the semiparametric model reproducing kernel Hilbert spaces (RKHS), and two parametric methods, namely, ridge regression and Bayes A. The ability of the above methods to predict was tested by estimating genomic breeding values for traits with different combinations of QTL effects, QTL numbers, three scenarios of heritability, and two training sets with 1,000 and 2,000 individuals. A genome of four chromosomes was simulated, and four generations were considered in the study. In the various simulation scenarios, the parametric methods outperformed semi-parametric (RKHS) and

nonparametric ones (RF and SVM). However, the superiority of parametric models compared to semi-parametric ones was not statistically significant. In summary, Bayes A had the best prediction accuracy among all tested models.

Deep learning algorithms are found to be powerful in discovering intricate patterns and nonlinearity in large, messy datasets. Their application in genomic prediction has been investigated, however, the number of reports on DL application in animal breeding is small, and thus their potential should be further investigated. Gianola et al. (2011) evaluated the predictive ability of an artificial neural network to predict three quantitative traits, namely, milk, fat, and protein yield. In Jersey dairy cows. The dataset contained records of the milk yield of 297 Jersey dairy cows with a total of 35,798 SNPS. The authors conceived different Bayesian neural networks (BNN) with various architectures that differed in terms of the number of neurons, the type of activation function, and the source of the input variables, whether they were derived from pedigree or molecular markers. According to the results, BNNs with at least two neurons in the hidden layer had better performance. Moreover, results also showed that Bayesian regularization helped reduce the number of weights, which helped prevent overfitting. However, an overfitting problem still occurred in the Jersey training set, where large correlations between observed and predicted data were observed in the training set (0.90–0.95) and much lower correlations in the testing set. In another study, Beskorovajni et al. (2022) developed a multi-layer perceptron for predicting yield and fertility traits of 92 genotyped Holstein heifers using several “Key traits” as input variables. These traits consist of Milk Yield, Fat Yield, Protein Yield, Somatic Cell Score (SCS), Productive Life (PL), Daughter Pregnancy Rate (DPR), Daughter Calving Ease (DCE), Final Type (PTA Type) and Genomic Future Inbreeding (GFI). An iterative method called the Broyden-Fletcher- Goldfarb-Shanno algorithm, which proceeds by minimizing the validation error, was used for optimization while training the ANN model. The authors obtained one optimal ANN for each target variable. The obtained ANN contained three layers, 11 neurons in the hidden layer and 276 weights and biases due to the high nonlinearity of the observed system. These hyperparameters led to the highest values of r2 (0.951, 0.947, 0.989, 0.985, 0.902, 0.887, 0.676, 0.953, 0.590, 0.647, and 0.444) for these traits respectively; fat percentage, protein percentage, cheese merit, fluid merit, cow livability, sire calving ease, sire calving ease, heifer conception rate, cow conception rate, daughter stillbirth, sire stillbirth, and gestation length. In the end, Beskorovajni et al. (2022) found that the ANN (network MLP 9-11-11) based on the Broyden-Fletcher-Goldfarb-Shanno optimization algorithm did a good job of fitting the data and predicting yield and fertility traits. Waldmann et al. (2020) combined a one-dimensional CNN model with l1-norm regularization, Bayesian optimization

and ensemble prediction within Genome Wide Prediction framework (CNNGWP) using simulated data with additive and dominance genetic effects and real pig data of 808 Australian Large White and Landrace sows with a total of 50,174 SNPs. In comparison to findings achieved with GBLUP and the LASSO, the results demonstrate that CNNGWP does indeed reduce prediction error by more than 25% on simulated data and by about 3% on real pig data. In summary, Waldmann et al. (2020) pointed out that CNNGWP appears to offer a promising approach for GWP, however the degree of improvement depends on the genetic architecture and the heritability. A detailed guide about the implementation of DL for GP may be found in (Zingaretti et al., 2020).

In order to compare the performance of ensemble learning methods and deep learning algorithms, Abdollahi-Arpanahi et al. (2020), compared the performances of RF and GB with MLP and CNN, and two conventional tools, namely, GBLUP and Bayes B, in predicting quantitative traits using both simulated and real Holstein datasets. The simulated dataset was used to assess the performance of ML methods in different scenarios of genetic architectures. A quantitative trait was simulated and two scenarios of QTN number were considered: [small (100) and large (1,000)]. QTNs were located across the genome in two different ways: clustered or randomly, and gene action were either purely additive or a combination of additive, dominance and epistasis effects. On the other hand, real data from 11,790 US Holstein bulls with a total of 57749 SNPs were used to test how well ML approaches can predict complex phenotypes like SCR, which is affected by both additive and non-additive effects. Abdollahi-Arpanahi et al. (2020) found that results differed depending on the genetic architecture of the trait. When pure additive actions controlled the trait, classic statistical models had better predictive accuracies compared to ML methods. However, the number of loci controlling the trait of interest appears to be an important factor in how well the models predicted outcomes when non-additive genetic effects occurred. The performance of ML algorithms, and in particular, GB, surpassed that of traditional statistical methods when the traits were controlled by a small number of QTN. The researchers finally came to the conclusion that, since Waldmann (2018) had already shown that loci are clustered, ML approaches work well for predicting traits with complex gene action and a small number of QTN (Abdollahi- Arpanahi et al., 2020). Genomic prediction in animal breeding usually involves small reference population issues, especially when it concerns a novel trait, which can be costly and labor-intensive to measure. Machine learning models can be deployed to tackle these challenges. For example, Yao et al. (2016) developed a self-training model, which is a semi-supervised algorithm wrapped around SVM to encounter the challenge of genomic prediction of residual feed intake (RFI). The model uses 792 animals with both genotypes and phenotypes

to train a base predictor, which is used to estimate the "self-trained phenotype" of 3,000 animals with genotypes only. To train a new predictor that is utilized to generate the final genomic predictions, both of these datasets are integrated. A total of 57491 SNPs were used for the analysis. The results showed that indeed, the self training algorithm increased the accuracy of genomic prediction, however, this improvement was small when the dataset already contained more individuals with measured phenotypes. Additionally, the correlation between predicted and measured phenotypes increased by adding more self-trained phenotypes, however, it reached a plateau at a certain level. In summary, Yao et al. (2016) concluded that semi-supervised learning is a powerful tool for enhancing the accuracy of genomic prediction for novel traits and for small reference populations. However, choosing an adequate sample size and an adequate ML algorithm are necessary to prevent poor predictions. As an example, the predictive ability of RF models with a set-up similar to this study was assessed, and the authors found no improvement in accuracy from using self-training models (Yao et al., 2016).

3.2.Feature selection

Feature selection techniques are vital in genomic prediction They allow us to identify the most informative genetic markers mostly SNPs, that contribute to the traits of interest. In genomics the massive amount of markers data poses a challenge in terms o computational efficiency and interpretability. By eliminatin irrelevant markers, feature selection methods reduce noise and dimensionality, and increase the accuracy and performance of ML models. In addition, feature selection procedures enable the identification of key genetic variants, providing valuable insights into the biological mechanisms underlying traits of interest. Therefore, several studies have investigated the potential of ML models in performing feature selection using SNPs datasets of multiple animal species. Li et al. (2018a) applied three machine learning methods, namely, RF, GBM and XgBoost, for ranking the top 400, 1,000, and 3,000 SNPs directly related to the body weight of Brahman cattle to generate genomic relationship matrices (GRMs) for estimating genomic breeding values (GEBVs). The database used consisted of the body weight records of 2093 animals with a total of 38082 SNP markers. According to the results, RF and GBM outperformed XgBoost in identifying a subset of SNPs related to the growth trait. Furthermore, the top 3,000 SNPs identified by RF and GBM provided similar GEBV values to those of the whole SNP panel. In summary, the authors highly recommend the use of RF and GBM for identifying subsets of potential SNPs related to traits of interest. Besides, this approach could be very useful in animal breeding since the vast majority of research suffers from small reference population issues, whether it is due to genotyping cost constraints or to the nature of the target variable, which could be costly and labor-

intensive to measure, such as feed efficiency. In this sense, Chen et al. (2021) compared the performance of two conventional methods, t-test and edgeR and three ensemble learning models, namely, RF, XGBoost, and a combination of both RF and XGBoost (RX) in identifying subsets of potential predictor genes in different tissues related to feed efficiency in Nellore Bulls. The dataset contained RNA sequences of five tissues (adrenal gland, hypothalamus, liver, skeletal muscle, and pituitary) from nine high-feed efficiency (HFE) and nine low-feed efficiency (LFE) bulls. Using the SVM model, the predictor genes that had been found using the above methods were used to divide the animals in the testing set into HFE and LFE. The performance of the classifier was evaluated using four metrics: overall accuracy, precision, recall and F1-score. The results showed that RX provided the best prediction accuracy yet with the smallest subset of genes (117). RF, in contrast, had the worst performance despite the fact that it had identified the largest number of candidate genes, contrary to what has been found in Naderi et al. (2016). The authors emphasize the idea that ML methods demonstrate great potential in identifying biologically relevant genes that can be used in classifying individuals accurately. In another study, Piles et al. (2021) implemented three types of feature selection methods: filter methods (tree-based methods), embedded methods (elastic net and LASSO regression), and a combination of both. Ridge regression, SVM, and GB were used after the pre-selection of relevant SNPs with filter methods. The results showed that using small subsets (50-250 SNPs), the feature selection method had a significant impact on prediction accuracy. In addition, filter methods demonstrated good performance and stability, indicating their potential for designing low density SNP chips for evaluating feed efficiency based on genomic information (Piles et al., 2021).

3.3. Genotype imputation

Genotype imputation plays a crucial role in animal genomics by inferring genotypes at specific positions in a genome by leveraging patterns and correlations within the data. Machine learning can be deployed to perform genotype imputation. For example, Sun et al. (2012) investigated the performance of Adaboost in imputing moderate-density genotypes from low density panels in order to reduce genotyping costs. The proposed model works, in fact, by combining the imputation results of preexisting software packages. The database included 3059 registered genotyped Angus cattle and 51911 SNPs across the whole genome. The missing genotypes were first imputed by previously available packages, of which three were family-based and the others were population-based. Consequently, the possible combinations of the six packages

resulted in 720 unique ensemble systems. The proposed Adaboost-based systems attribute a weight to each imputation method as a weak classifier. During the iterative training, the weights of classifiers that provided good predictions remained constant, whereas the weights of the misclassified samples were increased, which emphasized the focus on difficult samples. Finally, the final imputation of the genotype is the one with the majority of votes from all classifiers in the ensemble system. The results showed that indeed the ensemble method improved the accuracy of imputation in the data, however, the degree of improvement was limited by the fact that the packages used as weak classifiers had already provided highly accurate imputation results. Nevertheless, the authors highlighted the potential of ensemble learning to provide robust systems to address inconsistencies among different imputations of the preexisting methods.

4. Potential for ML applications to genomic prediction in animal breeding in developing countries

The majority of developing countries are grappling with satisfying the nutritional demands of an increasing human population. Meeting the demand for animal protein in a context of difficult environmental conditions and the predominance of smallholder systems in a sustainable manner is a challenging task. In addition, the introduction of highly productive dairy cows and the use of elite AI bulls' semen to inseminate national dairy herds resulted in low productivity due to unfavorable genotypes by environment interaction. Moreover, it is delicate for developing countries to implement a consistent conventional genomic selection breeding scheme due to the lack of reliable phenotypes and pedigree data recording (Mrode et al., 2019). Therefore, in order to improve national livestock systems productivity, developing countries should find alternatives to the aforementioned bottlenecks. The development of genomic technologies and the remarkable decrease in genotyping costs can be valuable for low- and middle-income countries, as they can tackle pedigree error problems by using the genomic relationship matrix (G) instead of the relationship matrix (A) or combining both information in a matrix H. However, the size and structure of the reference population is the biggest struggle for adopting GS in developing countries, the number of genotyped animals is limited, usually between 500 and 3,000 animals, predominated by females due to the non-existence of AI bulls (Mrode et al., 2019). Collaborations with developed nations, as Li et al. (2016) describe, could therefore be advantageous for implementing GS in these nations. Also, the use of a mixture of high-density (HD) and low density (LD) chips followed by imputation to the HD could be an alternative for reducing even more the genotyping costs in order to increase the size of the reference population (Lashmar et al., 2019).

Considering indigenous breeds in breeding programs is indispensable in developing countries. First of all, the majority of smallholder systems' dairy cows are either indigenous dairy cattle or crossbreds. Second, the conservation of genetic resources of local breeds that are adapted to specific agro-ecologies is crucial for the sustainability of the breed and biodiversity (Bulcha et al., 2022). Several countries, such as Kenya, Senegal, East Africa, Ethiopia, etc., have already implemented genomic technologies for indigenous breeds in Africa. Some studies used SNP data to determine the most adequate breed-type for different production environments. Others used genomic technologies to enhance breeding programs by increasing the accuracy of relationships among individuals. In other words, they have adopted genomic procedures to tackle the lack of pedigree recording. Finally, researchers investigated the potential of genomics for creating new breed-types that combine the adaptation and resilience of local breeds with the high productivity of exotic breeds. Genomic procedures and technologies have also been shown to be useful in discovering valuable genes in indigenous breed genomes, with significant effects due to the high levels of genome diversity of local breeds compared to exotic ones (Marshall et al., 2019).

Adopting GS in developing countries could benefit from the implementation of machine learning algorithms. First of all, given that indigenous breeds always have small reference populations, machine learning has shown great advantage in increasing the accuracy of breeding values estimation in small populations, as previously seen in Yao et al. (2016). In addition, ML models increased the accuracy of SNP imputation from low-density (LD) panels to high density (HD) chips, as investigated by Sun et al. (2012). This could result in reducing genotyping costs and increasing the size of genotyped animals (if the reference population is small due to genotyping costs). Overall, the potential of applying machine learning models for animal breeding in low- and medium-income countries is remarkable, as it could provide insightful findings. However, one of the biggest challenges would be the lack of data. Machine learning models typically require a massive amount of data in order to achieve high accuracy, while low- and middle-income countries often struggle with limited access to reliable data. Nonetheless, efforts should be directed toward exploring alternative techniques to enhance genomic prediction accuracy using a small reference population and promoting data sharing through collaborations among institutes and countries. As far as we know, the combination of machine learning models and genomic prediction in developing countries has not been used in any of the published studies, and thus their potential in enhancing breeding programs in low- and middle- income countries should be investigated in future experiments.

5. Conclusion

Machine learning algorithms have proven their high flexibility and ability to extract patterns in large, messy datasets in various fields such as natural language processing, robotics, speech recognition, image processing, etc. Genomic prediction is indeed a field of study where the main challenge is dealing with an ever-increasing marker dataset and capturing interactions and non-additive effects between genotypes. Consequently, investigating the potentiality of ML algorithms in GP is gaining a lot of buzz in the animal breeding community. Here, we reviewed studies that applied ML models to GP, whether they concerned estimating the GEBVs for production traits, health traits, or novel traits. In addition, several studies used ML algorithms for feature selection (FS) and moderate-density genotype imputation from low-density panels. It can be observed that ML algorithms outperformed conventional methods in some studies but were less accurate in others, which indicates that there's no universal method that can be applied to enhance the accuracy of prediction regardless of the domain of application. As a prerequisite, one should pay attention to several factors in order to successfully apply ML algorithms. For instance, the nature of the task, whether it consists of classification, clustering, regression, or dimensionality reduction, the type of the target variable (continuous or discrete), and the quality of the data (redundant, noisy, existence of outliers, missing values). ML models are indeed flexible and powerful, but they also have several drawbacks. One of the most common problems encountered in ML is overfitting. Additionally, finding the optimal hyperparameters can be challenging, and the size of the training data needs to be very large, especially for training deep learning algorithms. It is indeed true that incorporating ML algorithms and biological knowledge provides valuable results. However, marker datasets tend to be very heterogeneous and redundant, which can lower the predictive ability of these models. Moreover, the interpretability of non-parametric ML models is also questionable. Even though the algorithm's prediction for a particular target variable is accurate, the relationship between the input and output variables is not simple to understand. In fact, DL models are broadly known for their "Black Box" nature, which means that their interpretation cannot extract relevant information about variables in the dataset. In summary, ML algorithms showed great potential for fitting and extracting patterns from large, noisy datasets. However, their adoption in livestock breeding is still in its infancy, and hence more research must be done in order to find new insights for GP. The limited number of applications of ML in animal breeding did not allow researchers to clarify the huge potential for these models to improve the genomic prediction of important traits. Therefore, more iterative experimentation needs to be conducted.

References

Bell, A. E. (1977). Heritability in restrospect. The Journal of Heridity, 68, 297–300.

Belyadi, H., & Haghighat, A. (2021). Machine Learning Guide for Oil and Gas Using Python: a step-by-step breakdown with data, algorithms, codes, and applications. Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-821929-4.00004-4.

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. Journal of machine learning research, 2(Dec), 125-137. https://www.jmlr.org/papers/v2/horn01a.

Bérodier, M., Berg, P., Meuwissen, T., Boichard, D., Brochard, M., & Ducrocq, V. (2021). Improved dairy cattle mating plans at herd level using genomic information. Animal, 15(1), 100016. https://doi.org/10.1016/j.animal.2020.100016

Beskorovajni, R., Jovanović, R., Pezo, L., Popović, N., Tolimir, N., Mihajlović, L., & Šurlan-Momirović, G. (2022). Mathematical modeling for genomic selection in Serbian dairy cattle. Genetika, 53(3), 1105-1115. https://doi.org/10.2298/GENSR2103105B.

Bircanoğlu, C., & Arıca, N. (2018, May). A comparison of activation functions in artificial neural networks. In 2018 26th signal processing and communications applications conference (SIU) (pp. 1-4). IEEE. https://doi.org/10.1109/SIU.2018.8404724.

Blasco, A. (2017). Bayesian Data Analysis for Animal Scientists: The Basics. Springer International Publishing. https://doi.org/10.1007/978-3-319-54274-4

Blasco, A., & Toro, M. A. (2014). A short critical history of the application of genomics to animal breeding. Livestock Science, 166, 4–9. https://doi.org/10.1016/j.livsci.2014.03.015

Boichard, D., Ducrocq, V., & Fritz, S. (2015). Sustainable dairy cattle selection in the genomic era. Journal of Animal Breeding and Genetics, 132(2), 135–143. https://doi.org/10.1111/jbg.12150

Boichard, D., Ducrocq, V., Croiseau, P., & Fritz, S. (2016). Genomic selection in domestic animals: principles, applications and perspectives. Comptes rendus biologiques, 339(7-8), 274-277. https://doi.org/10.1016/j.crvi.2016.04.007.

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv preprint arXiv:1809.03006.

Boujenane, I. (2017). Reasons and risk factors for culling of Holstein dairy cows in Morocco. Journal of Livestock Science and Technologies, 1(5), 25–31.

Bouquet, A., & Juga, J. (2013). Integrating genomic selection into dairy cattle breeding programmes: A review. Animal, 7(5), 705–713. https://doi.org/10.1017/S1751731112002248

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification algorithms and regression trees. Classification and regression trees, 15(2), 246.

Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Brito, L. F., Oliveira, H. R., McConn, B. R., Schinckel, A. P., Arrazola, A., Marchant-Forde, J. N., & Johnson, J. S. (2020). Large-Scale Phenotyping of Livestock Welfare in Commercial Production Systems: A New Frontier in Animal Breeding. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.00793

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. Handbook of computational statistics: Concepts and methods, 985-1022. https://doi.org/10.1007/978-3-642-21551-3_33.

Buch, L. H., Kargo, M., Berg, P., Lassen, J., & Sørensen, A. C. (2012). The value of cows in reference populations for genomic selection of new functional traits. Animal, 6(6), 880–886. https://doi.org/10.1017/S1751731111002205

Buch, L. h., Sørensen, M. k., Berg, P., Pedersen, L. d., & Sørensen, A. c. (2012). Genomic selection strategies in dairy cattle: Strong positive interaction between use of genotypic information and intensive use of young bulls on genetic gain. Journal of Animal Breeding and Genetics, 129(2), 138–151. https://doi.org/10.1111/j.1439-0388.2011.00947.x

Bulcha, G. G., Dewo, O. G., Desta, M. A., & Nwogwugwu, C. P. (2022). Indigenous knowledge of farmers in breeding practice and selection criteria of dairy cows at Chora and Gechi Districts of Ethiopia: An Implication for Genetic Improvements. Veterinary Medicine International, 2022.

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery, 159(6), 1638-1645.

Chen, T., He, T., and Benesty, M. (2016). Xgboost: Extreme Gradient Boosting. Available online at: https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf (accessed January 5, 2021).

Chen T, Li M, Li Y, Lin M, Wang N, Wang M. (2017). MXNet: a flexible and efficient library for deep learning. https ://mxnet .incubator.apach e.org/.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z. (2015). MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems (No. arXiv:1512.01274). arXiv. https://doi.org/10.48550/arXiv.1512.01274

Chen, W., Alexandre, P. A., Ribeiro, G., Fukumasu, H., Sun, W., Reverter, A., & Li, Y. (2021). Identification of predictor genes for feed efficiency in beef cattle by applying machine learning methods to multi-tissue transcriptome data. Frontiers in Genetics, 12, 619857. https://doi.org/10.3389/fgene.2021.619857.

Choudhary, R., & Gianey, H. K. (2017, December). Comprehensive review on supervised machine learning algorithms. In 2017 International Conference on Machine Learning and Data Science (MLDS) (pp. 37-43). IEEE. https://doi.org/10.1109/MLDS.2017.11.

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. Genetics Selection Evolution, 42(1), 2. https://doi.org/10.1186/1297-9686-422

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297. https://doi.org/10.1007/BF00994018.

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modeling, 240, 113-122. https://doi.org/10.1016/j.ecolmodel.2012.03.001.

David, L., Arús-Pous, J., Karlsson, J., Engkvist, O., Bjerrum, E. J., Kogej, T., ... & Chen, H. (2019). Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. Frontiers in pharmacology, 10, 1303. https://www.frontiersin.org/articles/10.3389/fphar.2019.01303.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2008). Misc functions of the Department of Statistics (e1071), TU Wien. R package, 1, 5-24. http://CRAN.R-project.org/.

Edel, C., Pimentel, E. c. g., Emmerling, R., & Götz, K.-U. (2022). 338. A critical aspect when using APY inversion with Single-Step GBLUP. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP) (pp. 1416–1419). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-940-4_338

Eggen, A. (2012). The development and application of genomic selection as a new breeding paradigm. Animal frontiers, 2(1), 10-15. https://doi.org/10.2527/af.2011-0027.

Ehret, A., D. Hochstuhl, N. Krattenmacher, J. Tetens, M.S. Klein, W. Gronwald, and G. Thaller. (2015). Short Communication: Use of Genomic and Metabolic Information as Well as Milk Performance Records for Prediction of Subclinical Ketosis Risk via Artificial Neural Networks. Journal of Dairy Science 98 (1): 322–29. https://doi.org/10.3168/jds.2014-8602.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh, 52(2), 399–433. https://doi.org/10.1017/S0080456800012163

Fleming, C. H., Noonan, M. J., Medici, E. P., & Calabrese, J. M. (2019). Overcoming the challenge of small effective sample sizes in home-range estimation. Methods in Ecology and Evolution, 10(10), 1679–1689. https://doi.org/10.1111/2041-210X.13270

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

Gelfand, A. E. (2000). Gibbs Sampling. Journal of the American Statistical Association, 95(452), 1300–1304. https://doi.org/10.1080/01621459.2000.10474335

Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics, 194(3), 573–596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., Okut, H., Weigel, K. A., & Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC genetics, 12, 1-14.

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. (2010). Genomic selection in livestock populations. Genetics research, 92(5-6), 413-421.https://doi.org/10.1017/S0016672310000613.

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2010). Genomic selection in livestock populations. Genetics Research, 92(5–6), 413–421. https://doi.org/10.1017/S0016672310000613

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J., & Avendano, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics, 178(4), 2305-2313.https://doi.org/10.1534/genetics.107.084293.

González-Recio, O., Weigel, K. A., Gianola, D., Naya, H., & Rosa, G. J. (2010). L2-Boosting algorithm applied to high-dimensional problems in genomic selection. Genetics research, 92(3), 227-237. https://doi.org/10.1017/S0016672310000261.

Greenwell, B., Boehmke, B., & Cunningham, J. (2020). Developers (https://github. com/gbm-developers). GBM gbm: Generalized Boosted Regression Models.. https://CRAN.R-project.org/package=gbm.

Guinan, F. L., Wiggans, G. R., Norman, H. D., Dürr, J. W., Cole, J. B., Van Tassell, C. P., Misztal, I., & Lourenco, D. (2022). Changes in genetic trends in US dairy cattle since the implementation of genomic selection. Journal of Dairy Science. https://doi.org/10.3168/jds.2022-22205

Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS technical report, 14(1), 5-16.

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics, 12(1), 186. https://doi.org/10.1186/1471-2105-12-186

HARTLEY, H. O., & RAO, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika, 54(1–2), 93–108. https://doi.org/10.1093/biomet/54.1-2.93

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 72(358), 320–338. https://doi.org/10.1080/01621459.1977.10480998

Hay, El Hamidi A., Yuri T. Utsunomiya, Lingyang Xu, Yang Zhou, Haroldo H. R. Neves, Roberto Carvalheiro, Derek M. Bickhart, Li Ma, Jose Fernando Garcia, and George E. Liu. (2018). "Genomic Predictions Combining SNP Markers and Copy Number Variations in Nellore Cattle." BMC Genomics 19 (1): 441. https://doi.org/10.1186/s12864-018-4787-6.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science, 92(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Hazel, L. N. (1943). THE GENETIC BASIS FOR CONSTRUCTING SELECTION INDEXES. Genetics, 28(6), 476–490. https://doi.org/10.1093/genetics/28.6.476

Hazel, L. N., & Lush, J. L. (1943). The efficiency of three methods of selection. Journal of Heredity, 33, 393–939.

Henderson, C. R. (1973). Sire evaluation and genetic trends. Journal of Animal Science, 1973(Symposium), 10-41. https://doi.org/10.1093/ansci/1973.Symposium.10.

Henderson, C. R. (1974). General Flexibility of Linear Model Techniques for Sire Evaluation. Journal of Dairy Science, 57(8), 963–972. https://doi.org/10.3168/jds.S0022-0302(74)84993-3

Henderson, C. (1984). Applicatıons of lınear models ın animal breedıng. University of Guelph Press, Guelph, 11, 652-653.

Hofer, A. (1998). Variance component estimation in animal breeding: A review†. Journal of Animal Breeding and Genetics, 115(1–6), 247–265. https://doi.org/10.1111/j.1439-0388.1998.tb00347.x

Hoffmann, F., Bertram, T., Mikut, R., Reischl, M., & Nelles, O. (2019). Benchmarking in classification and regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(5), e1318. https://doi.org/10.1002/widm.1318.

Johnson, D. L., & Thompson, R. (1995). Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. Journal of Dairy Science, 78(2), 449–456. https://doi.org/10.3168/jds.S0022-0302(95)76654-1

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. Behavior Therapy, 51(5), 675-687.https://doi.org/10.1016/j.beth.2020.05.002.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285. https://doi.org/10.1613/jair.301.

Kang, M., & Jameson, N. J. (2018). Machine Learning: Fundamentals. In Prognostics and Health Management of Electronics, edited by Michael G. Pecht and Myeongsu Kang, 85–109. Chichester, UK: John Wiley and Sons Ltd. https://doi.org/10.1002/9781119515326.ch4.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees?. Nature biotechnology, 26(9), 1011-1013 https://doi.org/10.1038/nbt0908-1011.

Kramer, O. (2013). Dimensionality reduction with unsupervised nearest neighbors (Vol. 51, pp. 13-23). Berlin: Springer. https://doi.org/10.1007/978-3-642-38652-7_2.

Labriji, A., El Foutayeni, Y., & Rachik, M. (2021). A Subsidy Strategy to Boost the Activity of Small Milk Producers in Morocco. Journal of Applied Mathematics, 2021, e9094551. https://doi.org/10.1155/2021/9094551

Lashmar, S. F., F. C. Muchadeyi, & C. Visser. (2019). Genotype Imputation as a Cost-Saving Genomic Strategy for South African Sanga Cattle: A Review. South African Journal of Animal Science 49 (2): 262–80. https://doi.org/10.4314/sajas.v49i2.7.

LaValley, M. P. (2008). Logistic Regression. Circulation 117 (18): 2395–99. https://doi.org/10.1161/CIRCULATIONAHA.106.682658.

LeCun, Y., Bengio, Y., & Hinton. G. (2015). Deep Learning. Nature 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. Journal of Dairy Science, 92(9), 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. Livestock Science, 166, 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Li, B., Zhang, N., Wang, Y. G., George, A. W., Reverter, A., & Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Frontiers in genetics, 9, 237. https://doi.org/10.3389/fgene.2018.00237.

Li, X., Lund, M. S., Zhang, Q., Costa, C. N., Ducrocq, V., & Su, G. (2016). Improving accuracy of predicting breeding values in Brazilian Holstein population by adding data from Nordic and French Holstein populations. Journal of Dairy Science, 99(6), 4574-4579. https://doi.org/10.3168/jds.2015-10609.

Li, Y., Wu, F. X., & Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. Briefings in bioinformatics, 19(2), 325-340. https://doi.org/10.1093/bib/bbw113.

Liang, M., Miao, J., Wang, X., Chang, T., An, B., Duan, X., Xu, L., Gao, X., Zhang, L., Li, J., & Gao, H. (2021). Application of ensemble learning to genomic selection in chinese simmental beef cattle. Journal of Animal Breeding and Genetics, 138(3), 291-299. https://doi.org/10.1111/jbg.12514.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22. https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.

Liu, H., & Motoda, H. (Eds.). (2007). Computational methods of feature selection. CRC press.

Lund, M. S., de Roos, A. P., de Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., & Su, G. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution, 43(1), 43. https://doi.org/10.1186/1297-9686-43-43

Lund, M. S., de Roos, A. P. W., de Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, M., & Su, G. (2010). Improving genomic prediction by EuroGenomics collaboration. 9. World Congress on Genetics Applied to Livestock Production, Communication 880. https://hal.science/hal-01193765

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., & Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta, 248, 1307-1318.https://doi.org/10.1007/s00425-018-2976-9

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

Makanjuola, B. O., Miglior, F., Abdalla, E. A., Maltecca, C., Schenkel, F. S., & Baes, C. F. (2020). Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. Journal of Dairy Science, 103(6), 5183–5199. https://doi.org/10.3168/jds.2019-18013

Marshall, K., Gibson, J. P., Mwai, O., Mwacharo, J. M., Haile, A., Getachew, T., Mrode, R., & Kemp, S. J. (2019). Livestock genomics for developing countries–African examples in practice. Frontiers in genetics, 10, 297. https://www.frontiersin.org/articles/10.3389/fgene.2019.00297.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451. https://doi.org/10.1016/0005-2795(75)90109-9.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147. https://doi.org/10.38094/jastt1457.

Metz, C. E. (1978, October). Basic principles of ROC analysis. In Seminars in nuclear medicine (Vol. 8, No. 4, pp. 283-298). WB Saunders. https://doi.org/10.1016/S0001-2998(78)80014-2.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. genetics, 157(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819.

Meuwissen, T., Hayes, B., & Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. Animal Frontiers, 6(1), 6–14. https://doi.org/10.2527/af.2016-0002

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. (2015). Misc functions of the Department of Statistics. Probability Theory Group (Formerly: E1071), TU Wien. https://cran.microsoft.com/snapshot/2016-08-05/web/packages/e1071/index.html.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. C. (2019). e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version, 1(2).

Misztal, I. (2008). Reliable computing in estimation of variance components. Journal of Animal Breeding and Genetics, 125(6), 363–370. https://doi.org/10.1111/j.1439-0388.2008.00774.x

Misztal, I. (2016). Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics, 202(2), 401–409. https://doi.org/10.1534/genetics.115.182089

Misztal, I., Aguilar, I., Lourenco, D., Ma, L., Steibel, J. P., & Toro, M. (2021). Emerging issues in genomic selection. Journal of Animal Science, 99(6), skab092. https://doi.org/10.1093/jas/skab092

Misztal, I., Fragomeni, B. O., Lourenco, D. A. L., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., & Lawlor, T. (2015). Efficient inversion of genomic relationship matrix by the algorithm for proven and young (APY). Interbull Bulletin, 49, Article 49. https://journal.interbull.org/index.php/ib/article/view/1602

Misztal, I., Legarra, A., & Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. Journal of Dairy Science, 97(6), 3943–3952. https://doi.org/10.3168/jds.2013-7752

Mikshowsky, A. A., Gianola, D., & Weigel, K. A. (2017). Assessing genomic prediction accuracy for Holstein sires using bootstrap aggregation sampling and leave-one-out cross validation. Journal of Dairy Science, 100(1), 453-464. https://doi.org/10.3168/jds.2016-11496.

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., Gaytan-Lugo, L. S., Santana-Mancilla, P. C., & Crossa, J. (2021). A review of deep learning applications for genomic selection. BMC genomics, 22, 1-23.https://doi.org/10.1186/s12864-020-07319-x.

Morota, G., & Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. Frontiers in genetics, 5, 363. https://doi.org/10.3389/fgene.2014.00363.

Mrode, R., Ojango, J. M. K., Okeyo, A. M., & Mwacharo, J. M. (2019). Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: Current status and future prospects. Frontiers in genetics, 9, 694. https://www.frontiersin.org/articles/10.3389/fgene.2018.00694.

Naderi, S., Yin, T., & König, S. (2016). Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. Journal of Dairy Science, 99(9), 7261-7273. https://doi.org/10.3168/jds.2016-10887.

Naser, M. Z., & Alavi, A. H. (2021). Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. Architecture, Structures and Construction, 1-19. https://doi.org/10.1007/s44150-021-00015-8.

Nasteski, V. (2017). An Overview of the Supervised Machine Learning Methods. HORIZONS.B 4 (December): 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05.

Nayeri, S., Sargolzaei, M., & Tulpan, D. (2019). A review of traditional and machine learning methods applied to animal breeding. Animal health research reviews, 20(1), 31-46. https://doi.org/10.1017/S1466252319000148.

Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. Nature Machine Intelligence, 1(3), 133-143. https://doi.org/10.1038/s42256-019-0025-4.

Nelson, R. M., Pettersson, M. E., & Carlborg, Ö. (2013). A century after Fisher: Time for a new paradigm in quantitative genetics. Trends in Genetics, 29(12), 669–676. https://doi.org/10.1016/j.tig.2013.09.006

Nick, T. G., & Campbell, K. M. (2007). Logistic regression. Topics in biostatistics, 273-301. https://doi.org/10.1007/978-1-59745-530-5_14.

Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011, December). A comparison of random forests, boosting and support vector machines for genomic selection. In BMC proceedings (Vol. 5, No. 3, pp. 1-5). BioMed Central. https://doi.org/10.1186/1753-6561-5-S3-S11.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Louppe, G. (2012). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12.

Pereira, F. C., & Borysov, S. S. (2019). Machine learning fundamentals. In Mobility Patterns, Big Data and Transport Analytics (pp. 9-29). Elsevier. https://doi.org/10.1016/B978-0-12-812970-8.00002-6.

Persa, R., Ribeiro, P. C. de O., & Jarquin, D. (2021). The use of high-throughput phenotyping in genomic selection context. Crop Breeding and Applied Biotechnology, 21, e385921S6. https://doi.org/10.1590/1984-70332021v21Sa19

Piles, M., Bergsma, R., Gianola, D., Gilbert, H., & Tusell, L. (2021). Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. Frontiers in genetics, 12, 611506. https://doi.org/10.3389/fgene.2021.611506.

Pryce, J., & Hayes, B. (2012). A review of how dairy farmers can use and profit from genomic technologies. Animal Production Science, 52(3), 180. https:/ /doi.org/10.1071/AN11172

Ridgeway G: Gbm: Generalized boosted regression models. R package, version 1.6-3.1. Available at http://cran.r-project.org/web/packages/gbm/.

Rosenblatt, F. (1957). The Perceptron-a Perceiving and Recognizing Automaton. Cornell University, Ithaca, NY, Project PARA, Cornell Aeronautical Laboratory, Rep, 85-460.

Sahebalam, H., Gholizadeh, M., Hafezian, H., & Farhadi, A. (2019). Comparison of parametric, semiparametric and nonparametric methods in genomic evaluation. Journal of Genetics, 98, 1-8. https://doi.org/10.1007/s12041-019-1149-3.

Schapire, R. E. (2013). Explaining adaboost. In Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik (pp. 37-52). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5.

Schapire, R. E. (1999, July). A Brief Introduction to Boosting. 99 (999): 1401–6. In Ijcai (Vol. 99, No. 999, pp. 1401-1406).

Schöpke, K., & Swalve, H. H. (2016). Review: Opportunities and challenges for small populations of dairy cattle in the era of genomics. Animal, 10(6), 1050–1060. https://doi.org/10.1017/S1751731116000410

Searle, S. R. (1991). C. R. Henderson, the Statistician; And His Contributions to Variance Components Estimation1. Journal of Dairy Science, 74(11), 4035–4044. https://doi.org/10.3168/jds.S0022-0302(91)78599-8

Sharma, D., & Kumar, N. (2017). A review on machine learning algorithms, tasks and applications. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 6(10), 2278-1323.

Sieklicki, M. de F., Mulim, H. A., Pinto, L. F. B., Valloto, A. A., & Pedrosa, V. B. (2020). Population structure and inbreeding of Holstein cattle in southern Brazil. Revista Brasileira de Zootecnia, 49, e20190052. https://doi.org/10.37496/rbz4920190052

Smith, L. A., Cassell, B. G., & Pearson, R. E. (1998). The Effects of Inbreeding on the Lifetime Performance of Dairy Cattle. Journal of Dairy Science, 81(10), 2729–2737. https://doi.org/10.3168/jds.S0022-0302(98)75830-8

Sraïri, M. T., & Baqasse, M. (2002). Devenir et performances de génisses laitières importées au Maroc. In Prospects for a sustainable dairy sector in the Mediterranean (pp. 331–336). Wageningen Academic. https://doi.org/10.3920/9789086865093_043

Srivastava, S., Lopez, B. I., Kumar, H., Jang, M., Chai, H. H., Park, W, Park, J. E.,& Lim, D. (2021). Prediction of Hanwoo cattle phenotypes from genotypes using machine learning methods. Animals, 11(7), 2066. 2066. https://doi.org/10.3390/ani11072066.

Sun, C., Wu, X. L., Weigel, K. A., Rosa, G. J., Bauck, S., Woodward, B., Schnabel R. D., Taylor, J. F., & Gianola, D. (2012). An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genetics research, 94(3), 133-150.. https://doi.org/10.1017/S001667231200033X.

VanRaden, P. M., Olson, K. M., Null, D. J., Sargolzaei, M., Winters, M., & Kaam, J. B. C. H. M. van. (2012). Reliability Increases from Combining 50,000- and 777,000-Marker Genotypes from Four Countries. Interbull Bulletin, 46, Article 46. https://journal.interbull.org/index.php/ib/article/view/1745

Vieira, S., Pinaya, W. H. L., Garcia-Dias, R., & Mechelli, A. (2020). Deep neural networks. In Machine Learning (pp. 157-172). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00009-2.

Visscher, P. M., & Goddard, M. E. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. Genetics, 211(4), 1125–1130. https://doi.org/10.1534/genetics.118.301594

Waldmann, P. (2018). Approximate Bayesian neural networks in genomic prediction. Genetics Selection Evolution, 50, 1-9. https://doi.org/10.1186/s12711-018-0439-1.

Waldmann, P., Pfeiffer, C., & Mészáros, G. (2020). Sparse convolutional neural networks for genome-wide prediction. Frontiers in Genetics, 11, 25. https://doi.org/10.3389/fgene.2020.00025.

Wang, X., Shi, S., Wang, G., Luo, W., Wei, X., Qiu, A., Luo, F., & Ding, X. (2022). Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. Journal of Animal Science and Biotechnology, 13(1), 1-12. https://doi.org/10.21203/rs.3.rs-1083849/v1.

Weigel, K. A., VanRaden, P. M., Norman, H. D., & Grosu, H. (2017). A 100-Year Review: Methods and impact of genetic selection in dairy cattle—From daughter–dam comparisons to deep learning algorithms. Journal of dairy science, 100(12), 10234-10250. https://doi.org/10.3168/jds.2017-12954.

Wiggans, G. R., Cole, J. B., Hubbard, S. M., & Sonstegard, T. S. (2017). Genomic Selection in Dairy Cattle: The USDA Experience. Annual Review of Animal Biosciences, 5(1), 309–327. https://doi.org/10.1146/annurev-animal-021815-111422

Wiggans, G. R., VanRaden, P. M., & Cooper, T. A. (2011). The genomic evaluation system in the United States: Past, present, future. Journal of Dairy Science, 94(6), 3202–3211. https://doi.org/10.3168/jds.2010-3866

Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2691-2699).https://doi.org/10.1109/CVPR.2015.7298885.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9, 611-629. https://doi.org/10.1007/s13244-018-0639-9.

Yang, L., Xu, L., Zhou, Y., Liu, M., Wang, L., Kijas, J. W., ... & Liu, G. E. (2018). Diversity of copy number variation in a worldwide population of sheep. Genomics, 110(3), 143-148. https://doi.org/10.1016/j.ygeno.2017.09.005.

Yao, C., Zhu, X., & Weigel, K. A. (2016). Semi-supervised learning for genomic prediction of novel traits with small reference populations: an application to residual feed intake in dairy cattle. Genetics Selection Evolution, 48, 1-9. https://doi.org/10.1186/s12711-016-0262-5.

CHAPTER 2 - GENETIC PARAMETERS OF MILK YIELD AND FERTILITY TRAITS IN MOROCCAN HOLSTEINS

Narjice Chafai[1,2], Bouabid Badaoui[1], Romdhane Rekaya[2]. Submitted to Frontiers in livestock.

Abstract

The dairy industry in Morocco is going through a modernization process. Large and medium size farms are rapidly increasing their presence in the Moroccan dairy industry both in number of milking cows and milk output. Import of purebred dairy heifers and semen of elite sires has led to a substantial increase in milk production. However, less than optimum farm management procedures and environmental factors are becoming a major challenge to more productive Moroccan dairy cows. This is especially the case for reproduction related traits. The negative relationship between milk yield and fertility traits is being exacerbated by the severe heat stress conditions. As the country attempts to locally select their replacement heifers, improvement or at least the avoidance of further deterioration of reproduction performance is a priority. Two data sets consisting of 4,186 first parity and 5,511 first and multi-parity cows were used to assess the genetic correlations between 305-d milk yield and 3 reproduction traits (number of inseminations per conception, success of first insemination, and days open). The pedigree files for both data sets consisted of 8,758 and 9,935 animals, respectively. A threshold-linear model was used for the analysis. For first parity, estimates of heritability for 305-day milk yield (MY), days open (DO), number of inseminations per conception (NIC), and success of first insemination (SFI) were 0.26±0.04, 0.17±0.04, 0.10±0.03, and 0.10±0.04, respectively. For multi-parity data, these estimates were 0.19±0.03, 0.12±0.02, 0.10±0.02, and 0.09±0.02 for MY, DO, NIC, and SFI, respectively. The genetic correlations between MY and reproduction traits were 0.15±0.11, 0.38±0.12, and -0.43±0.11 for DO, NIC, and SFI respectively. Overall, the heritability of fertility traits was low, and their genetic correlations with MY were moderately negative allowing for the possibility for further selection for milk production without at least additional deterioration in reproduction performance. The relative impact of using high fertility bulls compared to low fertility bulls on the success of first insemination ranged between 1.2 and 6.3% depending on the production environment. Collectively, these results point towards the possibility of implementing a viable selection program based on an appropriate weighted selection index.

Background

The Moroccan dairy industry is an important socio-economic sector with significant contribution to the country economic growth. Additionally, it provides livelihoods and employment to a substantial

portion of the population either directly or indirectly (Labriji et al., 2021). Despite its pivotal role, the industry faces persistent challenges that hinder national milk production, notably climatic factors. The latter are characterized by irregular rainfall and a prolonged hot season, leading to pronounced heat stress and fodder shortages. Poor farm management practices further exacerbate the situation resulting in a low productivity of the dairy enterprise. Compounding these issues is the heavy reliance on imported purebred dairy heifers and elite sires for genetic improvement, a strategy that often fails due to the inability of these animals to fully realize their genetic potential under Moroccan conditions leading to early involuntary culling and significant economic losses to farmers (Sraïri and Baqasse, 2002; Sraïri, 2011).

The poor fertility performance together with the low productivity are major challenges to the sustainability and the survival of the Moroccan dairy industry. Attempts to increase milk production has resulted in marked deterioration of reproduction performance. This is expected as high milk producing cows often struggle with reduced fertility. Continuing to rely on imported germplasm, under the current environmental and management conditions, does not seem to be the best option. To mitigate these challenges, two interventions are needed: 1) produce local replacement heifers, and 2) include reproduction traits into the genetic selection index. In this study, only the second challenge will be addressed. Specifically, the main aim of this study is to estimate the genetic parameters of milk yield and three fertility related traits (305d-milk yield, days open, number of inseminations per conception, and success of first insemination) using the appropriate statistical model for the first parity and multi parity Holstein cows.

Material and methods

1. Data

It consisted of 13,255 records collected on 7,928 cows in four large dairy herd located in the northern region of Morocco between 2015 and 2023. The number of parities per cow ranged between 1 and 5. The four locations are characterized by a Mediterranean climate with a mild and humid winters and a long hot and dry summer season. Cows were housed in open stalls within an intensive feeding and production systems. Cows are fed a total mixed ration twice a day. Artificial insemination (AI) was the sole breeding method utilized and it is administered by specialized AI technicians using frozen semen imported from Europe and North America. Estrus detection was conducted through visual inspection. Daily records encompassing calving events, veterinary interventions, health issues, abortions, and dry-off dates were meticulously logged using a herd management software.

In order to reduce inconsistencies and non-random missingness, only cows with the first 3 lactations were retained. 305-d milk yield (MY), days open (DO), number of inseminations per conception (NIC) as continuous traits, and success of first insemination (SFI) as a discrete response were considered in this study. Parities with matching calving and abortion dates were removed. DO and NIC were considered missing when the conception date is missing, and no subsequent calving was reported. NIC and DO ranges respectively between 1 and 17, and 20 and 574 days. Cows with first calving before 530 days of age were excluded. After edit, the final multi-parity data consisted of 7,600 records collected on 5,511 cows. The first parity dataset included 4,186 records.

2. Statistical model

A Bayesian multivariate linear-threshold model was used to jointly analyze the four productive and reproductive traits for the first parity and in multiparity cows. The model included for the continuous traits and the liability for the binary response the fixed effects of herd (4 levels), calving month (12 classes), health status (2 levels), abortion status (3 levels), and two covariates (age at calving and days in milk at first insemination). Herd and calving month were separated and not included as a contemporary group to avoid extreme-category problems (ECP) or the binary response (all the observations in one fixed effect class are either 0 or 1). For the health status, a cow was considered healthy if no health issues, excluding abortion, occurred during the lactation otherwise it was considered non-healthy independently of the number of health problem episodes. Abortion status was coded 1 if no abortion occurred, 2 if one or more abortions occurred during milk production period, and 3 if one or more abortions occurred during the dry period. In addition, for the repeatability model, the parity number was included as a fixed effect.

At the liability scale, the following mixed linear model was used to analyze the data of the first three parities

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Wpe} + \boldsymbol{e} \qquad [21]$$

where $\boldsymbol{y} = (\boldsymbol{y_{MY}}, \boldsymbol{y_{DO}}, \boldsymbol{y_{NIC}}, \boldsymbol{l})'$ is the vector of phenotypes for milk yield ($\boldsymbol{y_{MY}}$), days open ($\boldsymbol{y_{DO}}$), number of inseminations to conception ($\boldsymbol{y_{NIC}}$), and the liabilities $\boldsymbol{l}$ for the binary response (success of first insemination). $\boldsymbol{\beta, u, pe}$ and **e** are the vectors of fixed, additive effects, random permanent environmental effects, and error terms, respectively. $\boldsymbol{X, Z, W}$ are known incidence matrices with the appropriate dimensions.

Conditionally on the model parameters, the joint distribution of the three continuous traits and the liability of the binary response was assumed to be normal

$$\begin{pmatrix} \boldsymbol{y_{MY}} \\ \boldsymbol{y_{DO}} \\ \boldsymbol{y_{NIC}} \\ \boldsymbol{l} \end{pmatrix} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Wpe}, \quad \boldsymbol{R_0} \otimes \boldsymbol{I})$$

where $\boldsymbol{R_0}$ is a 4x4 residual covariance matrix between the three continuous traits and the liability for the success of first insemination and $\boldsymbol{I}$ is the identity matrix with dimensions equal to the number of animals with data. Note that the last diagonal element of the residual covariance matrix is fixed to 1 to make the model identifiable (Gianola, 1982).

When using only the first parity data, the permanent effect and the parity number were dropped from the model presented in equation [1]. To finalize the Bayesian formulation, the following priors ere assumed for the position parameters

$$p(\boldsymbol{\beta}) \sim \mathrm{U}[-10^6, 10^6\,]$$

$$\boldsymbol{u}|\boldsymbol{A}, \boldsymbol{G} \sim \mathrm{N}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A})$$

$$\boldsymbol{p}|\boldsymbol{P} \sim \mathrm{N}(\mathbf{0}, \mathbf{P} \otimes \mathbf{I})$$

where $\boldsymbol{G}$ and $\boldsymbol{P}$ are 4x4 genetic and permanent environmental covariance matrices, respectively. **A** is the additive relationship matrix.

For the genetic and permanent environment covariance matrices, scaled inverse Wishart distribution priors were assigned

$$\boldsymbol{G}\,|\boldsymbol{S_g}, \vartheta_g \sim IW(\boldsymbol{S_g}, \vartheta_g)$$

$$\boldsymbol{P}|\boldsymbol{S_p}, \vartheta_p \sim IW(\boldsymbol{S_p}, \vartheta_p)$$

where $\boldsymbol{S_g}$ and $\boldsymbol{S_p}$ are 4x4 covariance matrices and ϑ_g and ϑ_p are the degrees of belief a priori for the genetic permanent environmental covariances, respectively.

The residual covariance matrix $\boldsymbol{R_0}$, is not completely random due to the fixation of the fourth diagonal element (corresponding to the binary traits) to 1. Consequently, the direct sampling of $\boldsymbol{R_0}$ is not

feasible. To overcome this problem of sampling of the residual (co) variance, the methods described by Rekaya et al., (2013) were used.

Briefly, the model in [1] is multiplied by a matrix $\boldsymbol{D} = \boldsymbol{D_0} \otimes \boldsymbol{I}$ where $\boldsymbol{D_0}$ is a 4x4 diagonal matrix, and $\boldsymbol{I}$ is an identity matrix with the appropriate dimensions, yields an equivalent model:

$$\boldsymbol{y}^* = \boldsymbol{X\beta}^* + \boldsymbol{Zu}^* + \boldsymbol{Wp}^* + \boldsymbol{e}^* \qquad [22]$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*)'$, $\boldsymbol{u}^* = (\boldsymbol{u}_1^*, \boldsymbol{u}_2^*, \boldsymbol{u}_3^*, \boldsymbol{u}_4^*)'$, $\boldsymbol{p}^* = (\boldsymbol{p}_1^*, \boldsymbol{p}_2^*, \boldsymbol{p}_3^*, \boldsymbol{p}_4^*)'$ with $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_i d_{ii}$, $\boldsymbol{u}_i^* = \boldsymbol{u}_i d_{ii}$, and $\boldsymbol{p}_i^* = \boldsymbol{p}_i d_{ii}$ are the vectors of fixed, additive, and permanent environmental effects for the trait i in the identifiable model in [1], respectively, and d_{ii} is the diagonal element i of the matrix $\boldsymbol{D_0}$. The resulting model in [2] is not identifiable because of $\boldsymbol{D}$ is not known. The residual (co)variance matrix of the non-restricted model in [2] given by:

$$var(\boldsymbol{e}^*) = \boldsymbol{DRD}' = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \qquad [23]$$

$$var(\boldsymbol{e}^*) = \boldsymbol{DRD}' = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}$$

where $\boldsymbol{R}$ is the original restricted residual (co)variance matrix in [1], and $\boldsymbol{\Sigma}_0$ is a 4x4 residual (co)variance matrix of the non-restricted model in [2]. Thus, estimates of the restricted residual covariance matrix, $\boldsymbol{R}$, could be easily obtained using equation [3] and the non-restricted matrix $\boldsymbol{\Sigma}_0$. The lack of restrictions in $\boldsymbol{\Sigma}$ facilitate enormously the Bayesian implementation vi Markov Chain Monte Carlo (MCMC) methods. However, in order to obtain the parameters of the identifiable model in [1] from the draws of the non-identifiable parameters, the matrix $\boldsymbol{D}$ needs to be defined. The identifiable parameters, based on expressions in [2] and [3], can be retrieved as:

$$\boldsymbol{\beta}_i = \frac{1}{d_{ii}}\boldsymbol{\beta}_i^*; \boldsymbol{u}_i = \frac{1}{d_{ii}}\boldsymbol{u}_i^*; \text{ and } \boldsymbol{p}_i = \frac{1}{d_{ii}}\boldsymbol{p}_i^* \qquad [24]$$

$$\boldsymbol{R} = \boldsymbol{D^{-1}\Sigma D^{-1}}; \textbf{ and } \boldsymbol{R_0} = \boldsymbol{D_0^{-1}\Sigma_0 D_0^{-1}} \qquad [25]$$

Given that the diagonal element of R_0 corresponding to the binary response is fixed to 1, the last diagonal element of the matrix D_0 must be equal to the square root of their corresponding element in $\mathbf{\Sigma}_0$, and the first 3 diagonal elements of D_0 corresponding to the continuous traits are set equal to one as indicated in Rekaya et al., (2013) and Chang et al., (2017). A flat bounded prior was assumed $\mathbf{\Sigma}_0$.

3. Effects of the genetic component on the probability of success of first insemination

Success of first insemination was scored as a binary response. Under the model assumed at the liability scale, the liability for an animal *i*, conditionally on the model parameter is given by:

$$l_i|\mu_i \sim N(\mu_i, 1) \qquad [26]$$

where l_i is the liability for animal *i,* $\mu_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i + p_i$, and $\boldsymbol{\beta}, u_i,$ and p_i are the vector of solutions for the fixed effect, the animal breeding value, and the environmental permanent environmental effect, respectively. $\boldsymbol{x}_i$is a known incidence vector relating the fixed effects to the liability.

Under the assumed threshold model, the relationship between the observed ordered categorical response and the liabilities is given by:

$$c_j = \begin{cases} 1 \text{ if } l_i > T = 0 \\ 0 \text{ otherwise} \end{cases} \qquad [27]$$

where $c_i = j$ is the binary response for animal *i* is equal to j ($j = 0, 1$), and the threshold value (T) was assumed equal to zero.

Given the distribution in equation [6] and the relationship in equation [7], the probability of observing class $j = 1$ for animal *i* is given by:

$$pr(c_i = 1|\boldsymbol{\beta}, u_i, p_i, \sigma_e^2 = 1, T = 0) = 1 - \Phi(T = 0|\eta_i, \sigma_e^2 = 1) \qquad [28]$$

where $\Phi(T = 0|\eta_i, \sigma_e^2 = 1)$ is the cumulative distribution function for the normal distribution with mean equal to η_i and variance equal to σ_e^2 evaluated at $T = 0$.

The impact of the using the best (highest breeding value for SFI) versus the worst bull on the probability of success of first insemination was assessed across three production environments (bad: μ_i=-1; average=μ_i; and good: μ_i=+1)

Results and discussion

1. Descriptive statistics for the traits

Table 1 presents a summary description of the four traits. Around 43% of the records were removed during the data editing process due to several reasons. Observation loss increases with parity number. This trend can be attributed to the elevated culling rate and the dependence on imported heifers resulting in a substantial portion of cows failing to progress beyond their initial lactation due to fertility and health issues. This trend underscores the unsustainability of the current production system. The 305-d milk production increases with the parity and is comparable to the results reported by Fahim et al. (2021). However, it was slightly higher than what have been reported in Ouarfli & Chehma (2021). Overall, the cows' reproductive performance across all parities were notably subpar. Across the three first parities, DO ranged between 153 to 161 days (Table 1). These results align closely with findings from Tunisia where M'Hamdi et al. (2009) reported a similar average DO interval of 151 days. Boujenane & Draga (2021), reported a shorter interval of 125 days, while El-Sherief et al. (2022) reported a much longer DO interval (173.17 days).

The NIC was relatively high across all three lactations, averaging around 4 inseminations per conception. Similar results were reported by Fahim et al. (2021) for Egyptian Holstein cows, Smaller estimates were reported by El-Sherief et al. (2022) and Boujenane & Draga, (2021). It's worth noting that the number of inseminations per conception is directly linked to Days Open (DO), as both metrics reflect similar aspects of the cow reproductive efficiency. Nevertheless, DO fails to account for the time elapsed between calving and first insemination, a gap addressed by the number of inseminations per conception (NIC) as suggested by Abdollahi-Arpanahi et al., (2013). However, DO is recommended in population that lacks a good recording scheme (González-Recio et al., 2006). Therefore, NIC emerges as a more comprehensive indicator of fertility. This comparison underscores the importance of considering multiple metrics to gain a holistic understanding of dairy cow reproductive health and efficiency.

Table 1: Least square means (LSM) and standard deviations for 305d-milk yield (305-MILK), days open (DO), number of inseminations per conception (NIC), and success of first insemination (SFI) across the first three parities

Parity	Trait	Number of observations	LSM	Standard deviation
First	305-MILK, kg	4186	8759.0	1489.1
	DO, d		152.5	91.3
	NIC		3.7	2.9
	SFI		0.2	
Second	305-MILK, kg	2380	9838.0	1620.3
	DO, d		157.5	94.4
	NIC		4.0	3.6
	SFI		0.2	
Third	305-MILK, kg	1034	9965.0	1696.4
	DO, d		161.4	98.6
	NIC		4.3	3.9
	SFI		0.2	
Total		7600		

The success of the first insemination, a reflection of pregnancy rate and closely linked to the non-return rate, was remarkably low across all parties, particularly in the first parity. This finding is unexpected, considering that first insemination success rate typically tends to be higher for the first parity cows compared to older ones. One possible explanation for this discrepancy, on top of the limited number of observations for older cows, is the strong culling rate for cows failing the first 2 or 3 inseminations if they calved in fall or early winter. Thus, older cows are heavily selected for their ability of conceiving early. The observed conception rate in this study appears to be lower than the 0.46 and 0.41 reported by Boujenane & Draga, (2021) and El-Sherief et al. (2022), respectively.

2. Estimates of variance and genetic parameters

Table 2 presents estimates of the posterior mean (PM) and posterior standard deviation (PSD) of the heritabilities, the genetic and the residual correlations for the four traits in the first parity. For fertility traits, heritability estimates were 0.17, 0.10, and 0.10 for DO, NIC, and SFI, respectively. To be noted are the negative correlations of DO and NIC with milk production, and the small correlations between SFI and the other traits However, these results have to be interpreted with caution given the wide high posterior density intervals associated with these estimates spanning from negative to positive values due the limited size of the dataset and the sparseness of the pedigree information.

Table 2: Heritability estimates (diagonal), genetic correlations (above), and residual correlations (below) for 305d-milk yield (305-MY), days open (DO), number of inseminations per conception (NIC), and success of first insemination (SFI) in first parity

	305-MILK, kg	**DO, d**	**NIC**	**SFI**
305-MILK, kg	0.26 ±0.04	-0.11±0.14	-0.23±0.14	0.04±0.10
DO, d	0.21±0.04	0.17±0.04	0.96± 0.019	-0.73±0.08
NIC	0.21±0.04	0.89±0.006	0.10±0.03	-0.80±0.05
SFI	-0.32±0.04	-0.67±0.02	-0.81±0.01	0.10±0.04

Estimates of the heritability and repeatability using the first three parity data are presented in Table 3. The 305-day milk yield heritability estimate (0.19 ± 0.03) was comparable to the findings reported by Bakri et al., (2022). However, it was slightly lower than those reported in several studies conducted in arid and Mediterranean regions (Fahim et al., 2021; M'Hamdi et al., 2009; Ojango & Pollott, 2001; Sadek et al., 2021; Wahinya et al., 2020; Windig et al., 2006). Heritability estimates for fertility traits were low and well within the range of reported estimates in the literature (Biffani et al., 2005; González-Recio & Alenda, 2005; Inchaisri et al., 2011; Jamrozik et al., 2005; M'Hamdi et al., 2009; Ojango & Pollott, 2001; Sadek et al., 2021; Wahinya et al., 2020), with posterior mean of 0.12, 0.07, and 0.08 for DO, NIC, and SFI, respectively.

Table 3 : Heritability and repeatability estimates for deviations) for 305d-milk yield (305-MY), days open (DO), number of inseminations per conception (NIC), and success of first insemination (SFI) using the first three parities

Trait	Heritability	Repeatability
305-MILK, kg	0.19±0.03	0.29 ± 0.02
DO, d	0.12±0.02	0.14 ± 0.02
NIC	0.07±0.02	0.08 ± 0.01
SFI	0.08±0.02	0.09 ± 0.02

Repeatability estimates (Table 3) ranged between moderate for milk yield (0.29) to low for fertility traits (0.08 to 0.014). Genetic correlations estimated via the repeatability threshold-linear model indicated positive associations between 305-day milk yield, days open, and number of inseminations per conception, suggesting that high milk-producing cows often require more inseminations to conceive and exhibit prolonged days open. Conversely, a negative correlation was observed between success of first insemination and all other traits, indicating that high milk yield cows are more likely not to conceive soon after calving. This is likely due high milk production around the peak of the lactation leading to a negative energy balance. The genetic correlations between SFI and DO and NIC ranged -0.97 and -0.89 (Table 4). The correlation between NIC and DO was high and positive (0.83). The residual correlations between the four traits were similar in trend and magnitude to those observed for the genetic correlations (Table 4) with the exception of the correlation between milk yield and SFI which was substantially smaller (in absolute value) than the genetic correlation (-0.12 vs -0.43). It is imperative to note that the standard deviations associated with the estimates of genetic correlations between milk yield and the fertility traits are relatively large indicating potential heterogeneity in managing cows based on their milk yield or/and reproductive performance.

Table 4 : Genetic (above the diagonal) and residual correlations (below diagonal) between 305d-milk yield (305-MILK), days open (DO), number of inseminations per conception (NIC), and success of first insemination (SFI)

	MILK, kg	**DO, d**	**NIC**	**SFI**
MILK, kg		0.15±0.11	0.38±0.12	-0.43±0.11
DO, d	0.16±0.03		0.83±0.06	-0.89±0.04
NIC	0.12±0.03	0.81±0.007		-0.97±0.01
SFI	-0.12±0.03	-0.86±0.02	-0.99±0.004	

Figure 1 presents the impact of using the best sire for success of first insemination (based on estimated breeding values) compared to using the worst sire across three generic production environments (bad, average, and good). Using the best sire resulted in a 6.3, 3.3, and 1.2% increase in SFI in the bad, average, and good production environment, respectively. Although the relative increase in the SFI was in the bad environment, the marginal gain (the difference in the success of first insemination using the best and worst bull) was 0.01, 0.015 and 0.01 in the bad, average, and good environment, respectively.

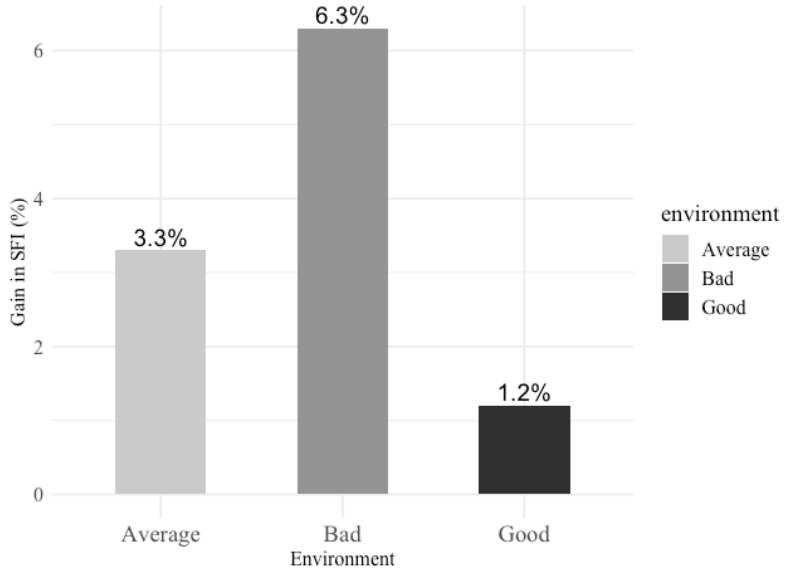

Figure 1 The relative gain in the success of first insemination using the best and the worst bull across different production environments.

In general, the results of this study align with previous research indicating the relatively low heritability of fertility traits in Holstein cows (Fahim et al., 2021; Inchaisri et al., 2011; M'Hamdi et al., 2009; Ojango & Pollott, 2001; Sadek et al., 2021; Wahinya et al., 2020). Our analysis underscores the challenge of accurately quantifying the genetic component of traits related to reproductive performance in Moroccan Holstein cows. Moreover, our investigation reveals a strong positive correlation between days open and the number of inseminations per conception. This correlation suggests that cows experiencing prolonged intervals between calving and conception tend to require a greater number of inseminations to achieve conception. Notably, the repeatability of days open is observed to be higher than that of NIC. This can be attributed to the fact that days open interval is more susceptible to management interventions, such as delaying the first insemination for high-yielding cows. In fact, it was argued against using days open for fertility evaluation, particularly in settings with comprehensive recording schemes and management practices (González-Recio et al., 2006). However, in contexts where such schemes are lacking, days open may still offer valuable insights into reproductive performance and management strategies (González-Recio et al., 2006). The success of first insemination is an important economic determinant, encompassing the cost of semen and the labor involved in monitoring heat cycles and conducting inseminations. Integrating SFI into the fertility index of a herd holds considerable merit, as it allows for a more comprehensive evaluation of reproductive performance and cost-effectiveness. Furthermore, the outcome of each insemination can be assessed as a longitudinal binary variable, as demonstrated by (Averill et al., 2006). This modeling approach enables the incorporation of all breeding information in a certain period, enabling more precise estimations.

Conclusion

Female fertility represents a multifaceted set of traits critical for the profitability of dairy herds, particularly evident for Holstein cows. However, this crucial relationship is often marred by an antagonistic correlation with milk yield. The moderate genetic correlations between milk yield and the three fertility traits used in this study clearly support the possibility for improving fertility or at least to attenuate its further deterioration without scarifying the selection for higher milk yield. This could be achieved through a selection index that includes all relevant traits with their relative weights. Defining appropriate traits to include in selection indices is paramount. A particular emphasis should be put on traits directly impacting profitability, such as the number of inseminations per conception and the success rate at first insemination, which include information on veterinary, breeding, and labor costs. A

fundamental hurdle in fertility trait analysis lies in the non-normal distribution of some traits, notably the discrete nature of the success rate of first insemination.

References

Abdollahi-Arpanahi, R., Peñagaricano, F., Aliloo, H., Ghiasi, H., & Urioste, J. I. (2013). Comparison of Poisson, probit and linear models for genetic analysis of number of inseminations to conception and success at first insemination in Iranian Holstein cows. *Livestock Science*, *153*(1), 20–26. https://doi.org/10.1016/j.livsci.2013.01.009

Averill, T., Rekaya, R., & Weigel, K. (2006). Random Regression Models for Male and Female Fertility Evaluation Using Longitudinal Binary Data. *Journal of Dairy Science*, *89*(9), 3681–3689. https://doi.org/10.3168/jds.S0022-0302(06)72408-0

Biffani, S., Canavesi, F., & Samoré, A. (2005). Estimates of genetic parameters for fertility traits of Italian Holstein-Friesian cattle. *Stockbreeding (Hrvatsko-Agronomsko-Drustvo@zg.t-Com.Hr); Vol.59 No.2*.

Boujenane, I., & Draga, B. (2021). Non-genetic factors affecting reproductive performance of Holstein dairy cows. *Livestock Research for Rural Development*, *33*, Article # 10.

Chang, L.-Y., Toghiani, S., Ling, A., Hay, E. H., Aggrey, S. E., & Rekaya, R. (2017). Analysis of Multiple Binary Responses Using a Threshold Model. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(4), 640–651. https://doi.org/10.1007/s13253-017-0305-6

El-Sherief, A. A., El-Komy, S. M., Rashad, A., & El-Hedainy, D. K. (2022). Reproductive Performance of Lactating Holstein Cows as Influenced by Season of Calving and Parity Under Subtropical Conditions. *Journal of Advanced Veterinary Research*, *12*(1), Article 1.

Fahim, N. H., Mohamed Ibrahim, M. A. A., Amin, A. H., & Sadek, R. R. (2021). Milk Production and Reproductive Performance of Retained and Culled Cows in a Large Holstein Herd in Egypt. *World's Veterinary Journal*, *11*(3), Article 3.

Gianola, D. (1982). Theory and Analysis of Threshold Characters. *Journal of Animal Science*, *54*(5), 1079–1096. https://doi.org/10.2527/jas1982.5451079x

González-Recio, O., & Alenda, R. (2005). Genetic Parameters for Female Fertility Traits and a Fertility Index in Spanish Dairy Cattle. *Journal of Dairy Science*, *88*(9), 3282–3289. https://doi.org/10.3168/jds.S0022-0302(05)73011-3

González-Recio, O., Alenda, R., Chang, Y. M., Weigel, K. A., & Gianola, D. (2006). Selection for Female Fertility Using Censored Fertility Traits and Investigation of the Relationship with Milk

Production. *Journal of Dairy Science*, *89*(11), 4438–4444. https://doi.org/10.3168/jds.S0022-0302(06)72492-4

Inchaisri, C., Jorritsma, R., Vernooij, J., Vos, P., van der Weijden, G., & Hogeveen, H. (2011). Cow Effects and Estimation of Success of First and Following Inseminations in Dutch Dairy Cows. *Reproduction in Domestic Animals*, *46*(6), 1043–1049. https://doi.org/10.1111/j.1439-0531.2011.01782.x

Jamrozik, J., Fatehi, J., Kistemaker, G. J., & Schaeffer, L. R. (2005). Estimates of Genetic Parameters for Canadian Holstein Female Reproduction Traits. *Journal of Dairy Science*, *88*(6), 2199–2208. https://doi.org/10.3168/jds.S0022-0302(05)72895-2

Labriji, A., El Foutayeni, Y., & Rachik, M. (2021). A Subsidy Strategy to Boost the Activity of Small Milk Producers in Morocco. *Journal of Applied Mathematics*, *2021*, e9094551. https://doi.org/10.1155/2021/9094551

M'Hamdi, N., Aloulou, R., Brar, S., Mahdi, B., & Ben Hamouda, M. (2009). Phenotypic and genetic parameters of reproductive traits in Tunisian Holstein cows. *Livestock Research for Rural Development*, *11*. https://doi.org/10.2298/BAH1006297M

Ojango, J. M. K., & Pollott, G. E. (2001). Genetics of milk yield and fertility traits in Holstein-Friesian cattle on large-scale Kenyan farms1. *Journal of Animal Science*, *79*(7), 1742–1750. https://doi.org/10.2527/2001.7971742x

Ouarfli, L., & Chehma, A. (2021). Effect of Temperature-Humidity-Index on Milk Performances of Local Born Holstein Dairy Cows Under Saharan Climate. *Archiva Zootechnica*, *24*(2), 24–36.

Rekaya, R., Sapp, R. L., Wing, T., & Aggrey, S. E. (2013). Genetic evaluation for growth, body composition, feed efficiency, and leg soundness. *Poultry Science*, *92*(4), 923–929. https://doi.org/10.3382/ps.2012-02649

Sadek, R. R., Abou-Bakr, S., Nigm, A. A., Abd El-Aziz Mohamed Ibrahim, M., Badr, M. M., & Awad, M. A. A. (2021). Evaluation of Milk Yield and Reproductive Performance of Pure Holstein and Its F1 Crossbreds with Montbeliarde in Egypt. *World's Veterinary Journal*, *11*(3), Article 3.

Sraïri, M. T., & Baqasse, M. (2002). Devenir et performances de génisses laitières importées au Maroc. In *Prospects for a sustainable dairy sector in the Mediterranean* (pp. 331–336). Wageningen Academic. https://doi.org/10.3920/9789086865093_043

Wahinya, P. K., Jeyaruban, G., Swan, A., & Magothe, T. (2020). Estimation of genetic parameters for milk and fertility traits within and between low, medium and high dairy production systems in

Kenya to account for genotype-by-environment interaction. *Journal of Animal Breeding and Genetics*, *137*(5), 495–509. https://doi.org/10.1111/jbg.12473

Windig, J. J., Calus, M. P. L., Beerda, B., & Veerkamp, R. F. (2006). Genetic Correlations Between Milk Production and Health and Fertility Depending on Herd Environment. *Journal of Dairy Science*, *89*(5), 1765–1775. https://doi.org/10.3168/jds.S0022-0302(06)72245-7

CHAPTER 3- GENETIC ANALYSIS OF DAYS OPEN IN MOROCCAN HOLSTEIN USING DIFFERENT MODELS TO ACCOUNT FOR CENSORED DATA

Narjice Chafai[1], Bouabid Badaoui[1]. Published in Animals.

Simple summary

Reproductive performance is a critical factor for the economic success and long-term viability of dairy herds. Intense selection for milk production has resulted in a decline in fertility traits in Holstein cows, leading to negative consequences for the industry. To combat this, incorporating fertility traits into the genetic evaluation is a potential solution. However, fertility data are often in-complete due to a variety of reasons; in this study, which has a limited dataset, it is essential to address this issue. The study utilized three methods to handle censorship: a linear model, a penalty method, and a threshold linear model with a penalty. The findings revealed that the penalized threshold model showed a slightly higher heritability compared to linear models. Moreover, both the penalty method and the threshold method exhibited comparable predictive abilities and substantial overlap in common animals, suggesting that both methods can be employed to impute days open censored data in this population.

Abstract

Reproductive efficiency is a key element of profitability in dairy herds. However, the genetic evaluation of fertility traits is often challenged by the presence of high censorship rates due to various reasons. An easy approach to address this challenge is to remove the censored data from the dataset. However, removing data might bias the genetic evaluation. Therefore, addressing this issue is crucial, particularly for small populations and populations with limited size. This study uses a Moroccan Holstein dataset to compare two Gaussian linear models and a threshold linear model to handle censored records of days open (DO). Data contained 8646 records of days open across the first three parities of 6337 Holstein cows. The pedigree file comprised 11,555 animals and 14.51% of the dataset was censored. The genetic parameters and breeding values of DO were computed using three different methods: a linear model where all censored records were omitted (LM), a penalty method in which a constant equal to one estrus cycle in cattle was added to the maximum value of DO in each contemporary group to impute the censored records (PLM), and a bivariate threshold model with a penalty (PTM). The heritability estimates were equal to 0.021 ± 0.01 (PLM), 0.029 ± 0.01 (LM), and 0.033 ± 0.01 (PTM). The lack of changes in the ranking of animals between PLM and PTM suggests that both methods can be used to address censored data in this population.

Introduction

Fertility traits are critical in the genetic evaluation of dairy cattle as they directly influence reproductive efficiency, which is essential for the sustainability and profitability of dairy operations. Effective female fertility performance is generally defined as a female capable of displaying estrus and achieving pregnancy with a minimal number of in-seminations per pregnancy [1]. A reduced number of inseminations translates to lower veterinary and hormonal treatment costs, reduced semen expenses, and diminished labor associated with artificial insemination. Due to intensive selection for milk production, particularly in the Holstein breed, and the negative correlation between milk yield and fertility traits, the reproductive performance of these cows has significantly declined [2]. Implementing genetic evaluations incorporating milk and fertility traits enables breeders to select animals with superior productive performance without remarkably compromising reproduction [3]. In the past decades, most breeding companies shifted their emphasis from milk production to including functional traits through selection indices. This approach could potentially improve herd fertility, enhance genetic progress, and yield greater economic returns. For instance, Scandinavian countries included health, fertility, and longevity traits in their selection indices for over two decades. Consequently, the decline in mean performance for such traits was stabilized [4]. Similarly, in 2003, the US introduced genetic evaluations for daughter pregnancy rates to improve fertility in dairy cows. This evaluation used DO data and transformed it into 21-day pregnancy rates as a measure of reproductive efficiency. As a result, significant improvements were observed with notable differences in fertility between the sires [4].

Fertility trait field data contain high noise levels due to several factors [5], including preferential treatments for different individuals and varying management practices. For example, avoiding breeding cows during hot seasons, when the success of insemination is very low due to heat stress, can introduce variability. Additionally, missed heat detections unfairly penalize cows for not conceiving, further contributing to the noise in fertility data. On top of that, fertility data generally comprise censored records as a result of retrieving the datasets before some cows conceive or give birth. Consequently, there are generally a lot of missing records due to censoring for days open (DO), defined as the days from calving to the following conception. This issue can be addressed by simply removing all the censored records. This approach offers a straightforward solution to this issue and computation simplicity, but it may also compromise the efficiency and accuracy of the estimation process and bias the genetic evaluation [6]. An alternative approach to tackle this problem is by imputing these censored records. Several methods have been proposed to address this challenge. The penalized Gaussian linear model (PLM) is one approach that

assigns a penalty to each censored record. In this method, the highest value within each contemporary group is added to censored records [7]. The constant value of 21 days is included to account for the duration of an estrus cycle in cattle. This approach assumes that any female who failed to conceive would have conceived if given an extra cycle [8]. The penalty method has been considered suitable to treat censored records in the genetic evaluations of fertility-related traits [9]. Data augmentation techniques can also be utilized for censored data imputation. One such model is the linear-threshold approach [7]. This method proposes a mixed effects model that incorporates a latent variable, liability (l), for each observation. If a record is censored and the pregnancy status of the cow with the censored record was non-pregnant at the last insemination, then its corresponding liability must be greater than a specified threshold T. The penalized threshold model is similar to the linear threshold model but assigns a penalty for censored records. As previously discussed, [6], the highest value in each contemporary group was identified, and a constant of 21 days was added to each censored record. Additionally, a latent variable indicating the censorship status was incorporated. The Bayesian alternative provides an appealing solution for censored data. However, the complex models and data structures typically found in animal breeding records can present computational and numerical challenges that are not yet fully understood [10]. Overall, the potential advantage of threshold-linear models over linear models with censoring in assessing fertility traits has not linear models over linear models with censoring in assessing fertility traits has not been quantified [11].

Datasets on reproductive traits in Moroccan Holstein suffer from high rates of censorship. Furthermore, the size of fertility trait datasets is generally limited. To perform genetic evaluations to choose elite dams' progeny for replacement, imputing censored records is crucial. The primary objective of this study is to apply various censoring models and evaluate their predictive ability to potentially identify the best model to address the issue of censored data of DO traits within a dataset of Holstein cows in Morocco.

Materials and methods

1. Data

Data were provided by the 'Les Domaines Agricoles' company. The dataset comprises records of Holstein cows raised in four herds located in two regions of northern Morocco. These regions are characterized by a Mediterranean climate, with mild winters and long, hot, and dry summers. The animals were managed under an intensive pro-duction system, where they were fed a total mixed ration and milked twice a day. Artificial insemination with frozen imported semen was the only method used by specialized

technicians. Pregnant cows were kept in free stalls and cleaned twice a day. The farm employees kept a constant watch on the pregnant cows and groups in maternity pens, monitoring for any visual signs of parturition. Once parturition occurred, the calves were promptly removed from the maternity pens and relocated to a heated calf pen. Records of milk yield, calving dates, artificial insemination dates, conception dates, dry-off dates, and heat detection dates were recorded using herd management software.

The raw dataset included 13,501 observations collected from 2017 to 2023. These observations were distributed across four herds: Herd 1 consisted of 3812 observations from 2149 cows, Herd 2 comprised 3170 records from 2303 cows, Herd 3 included 4092 records from 2098 cows, and Herd 4 encompassed 2427 records from 1464 cows. The age at each calving in the raw data was 781 days for the first calving, 1231 days for the second calving, and 1666 days for the third calving. The dataset included parities up to the ninth parity. The number of observations decreased at the fourth parity to 855 and at the ninth parity to 9 observations.

To address non-random missingness, only the first three lactations were retained. Data processing involved removing cows with incorrect identification, cows that calved for the first time before 530 days of age, those with identical calving and abortion dates, and those with missing calving dates. For the linear covariates age at calving and days in milk at first insemination, records exceeding the mean plus three standard deviations were excluded. The trait of interest, days open (DO), was defined as the number of days from calving to subsequent conception. Observations of DO were retained only if they fell within the range of the mean plus or minus three standard deviations within parity. Additionally, records of DO smaller than 20 days were deleted. Contemporary groups were created by combining herd, year of calving, and season of calving (HYS). The seasons were defined as winter (December to February), spring (March to May), summer (June to August), and fall (September to November). Contemporary groups with fewer than three observations were removed. The maximum number of observations in a contemporary group was 248. Due to the size of the dataset, 287 sire had one progeny. Therefore, no restriction was applied to the number of daughters per sire. Records were assumed censored if no pregnancy was confirmed by a subsequent conception or calving.

After data processing, the dataset consisted of 8646 records from 6337 cows and 14.51% of the DO records were censored. The pedigree comprised 11,555 animals and 931 sires. The dataset with no censored records consisted of 7550 observations of 5468 cows and the pedigree file consisted of 10,375 animals.

2. Statistical Models

Two Gaussian linear models were used in this analysis, a classical linear model with no censored data and a linear model with penalty. In the first linear model (LM), the censored data were completely removed from the dataset. In matrix notation, the animal model can be expressed as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Wpe} + \boldsymbol{e} \quad (1)$$

where y is the vector of observations of DO; β is the vector of fixed effects, including the contemporary groups her-year-season with 122 classes (HYS), parity (3 classes), and two linear covariates: days in milk at first insemination and age at calving; u is the vector of additive genetic effect; pe is the vector of permanent environment effect; and e is the vector of residuals. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known incidence matrices. The random genetic effect u was assumed to be normally distributed $\boldsymbol{u} \sim N(0, \boldsymbol{A}\sigma_u^2)$ where A is the numerator relationship matrix and σ_u^2 is the additive genetic variance. The permanent environment effect is also assumed to follow a normal distribution $\boldsymbol{pe} \sim N(0, \boldsymbol{I}\sigma_{pe}^2)$ where I is an identity matrix and σ_{pe}^2 is the permanent environment variance. The residuals are assumed to be independent and follow a normal distribution $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ where σ_e^2 is the residual variance and I is an identity matrix.

The penalty method (PLM) proposed by Johnston and Bunter [8] assigns a penalty for each censored record. Specifically, it adds the highest value of DO within each contemporary group, which is equivalent to the length of the estrus cycle in cattle at 21 days, to each censored record [7,12]. This method assumes that if these animals were given the opportunity for an additional estrus cycle, they would likely conceive. The analysis for this approach utilized the previously described animal model (1).

The threshold linear model with a penalty (PTM) is a Bayesian bivariate model in the scale of a latent variable, liability (l). In this approach, data were augmented using an additional binary trait corresponding to censorship status. The binary trait was equal to 1 if the record was censored and 0 otherwise. The underlying continuous liability was associated with the binary trait by the following formula:

$$y_{binary,i} = \begin{cases} 1 \text{ if } l_i > T = 0 \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

where $y_{binary,i}$ is the binary response for animal i, and the threshold T was assumed equal to 0. The liability values were updated at each iteration of the Gibbs sampler.

$$\begin{bmatrix} \boldsymbol{y_{DO}} \\ \boldsymbol{l} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_{DO}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_l} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_{DO}} \\ \boldsymbol{\beta_l} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_{DO}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z_l} \end{bmatrix} \begin{bmatrix} \boldsymbol{u_{DO}} \\ \boldsymbol{u_l} \end{bmatrix} + \begin{bmatrix} \boldsymbol{W_{DO}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W_l} \end{bmatrix} \begin{bmatrix} \boldsymbol{pe_{DO}} \\ \boldsymbol{pe_l} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_{DO}} \\ \boldsymbol{e_l} \end{bmatrix} \quad (3)$$

The threshold-linear animal model on a liability scale is given by

where $\boldsymbol{y_{DO}} = [\boldsymbol{y_o}, \boldsymbol{y_c}]$ consists of the vector of observed records of DO ($\boldsymbol{y_o}$) and the vector of augmented records with a penalty ($\boldsymbol{y_c}$), calculated as the maximum value of DO within each contemporary group (HYS) plus a constant equal to 21 days, as described by Costa et al. [13]; l is the censoring liability; β is the vector of fixed effects including contemporary groups, parity, age at calving, and days in milk at first insemination; u is the vector of additive genetic effect; pe is the vector of permanent environment effects; and e is the vector of residuals. $X_{DO}, Z_{DO}, W_{DO}, X_l, Z_l$, and W_l are the incidence matrices related to the fixed effects, the additive genetic effect, and the permanent environment effect, for both DO records (observed and augmented with a penalty) and the correlated binary variable corresponding to censorship status, respectively.

For all the models, a Gibbs sampler was run for a single chain of 500,000 iterations with a burn-in of 100,000 iterations and a thinning interval of 50. The genetic parameters and breeding values were inferred using BLUPF90+ programs [14]. A burn-in period was determined using visual inspection and the thinning interval was chosen using the autocorrelations computed by the Postgibbsf90 program [14]. Convergence was assessed using the Geweke diagnostic [15] and visual inspection of trace plots.

3. Accuracy of genetic predictions using the three models

The prediction accuracy for the three models LM, PLM, and PTM was evaluated using the LR method proposed by Legarra and Reverter [16]. The validation cohort comprised animals with phenotypes born in 2020 or later. As a result, the validation set consisted of 30% of the younger animals in the complete dataset. The breeding values were estimated using the whole dataset ($\hat{u}_w$) and a reduced dataset from which the phenotypes of the younger animals (the validation set) were removed ($\hat{u}_p$). The accuracies of the estimated breeding values using the LR method $\widehat{ACC}_{LR}$ can be obtained using the following formula:

$$\widehat{ACC}_{LR} = \sqrt{\frac{cov(\hat{u}_w, \hat{u}_p)}{(1-\bar{F})\sigma_a^2}}, \quad (4)$$

where $cov(\hat{u}_w, \hat{u}_p)$ is the covariance between estimated breeding values obtained using the whole dataset and the partial dataset, respectively. $\bar{F}$ is the average inbreeding coefficient in the validation population and σ_a^2 is the estimated genetic variance using each method. The average inbreeding coefficient of the

validation set was computed using RENUMF90 of BLUPF90 software version 2.53 [14] and was equal to 0.009.

Furthermore, we conducted a comparison of the models' bias and dispersion using the LR method [15]. The bias is calculated as the difference between the average estimated breeding values (EBVs) of individuals in the validation set based on partial data and the whole data. The following formula was applied to determine the bias [17]:

$$Bias_{LR} = \overline{\hat{u}_p} - \overline{\hat{u}_w} \quad (5)$$

The dispersion was measured through the slope of the regression of $\hat{u}_w$ on $\hat{u}_p$ for the animals in the validation dataset [17]. The dispersion formula is given by

$$Dispersion_{LR} = \frac{cov(\hat{u}_w, \hat{u}_p)}{var(\hat{u}_p)} \quad (6)$$

Additionally, the Spearman correlation between the estimated breeding values of the animals in the validation obtained by the three methods (LM, PLM, and PTM) was computed using R software [18]. Finally, the percentage of similar animals in the top 20% was identified to evaluate potential changes in animals' ranking.

Results and discussion

1. Genetic parameters

A descriptive summary of the edited data with no censored record used in this study is presented in Table1. Data cleaning and processing reduced the original dataset to 64%. The first parity contains more observations than other parities because of high culling rates due to fertility-related problems. The distribution of DO (Figure 1) is asymmetric and has a long tail to the right. The average DO was around 155.46 ± 96.50 for the first parity, 156.55 ± 94.10 for the second parity, and, finally, 159.71 ± 96.81. These numbers align closely with previous studies concerning Holstein cattle in the Mediterranean and arid areas [19,20]. The mean age at calving was 771.75 ± 77.81 (25 months), 1201.27 ± 116.79 (40 months), and 1628.29 ± 149.93 (54 months) for the first, second, and third calving, respectively. Similar results were found for Holstein cows in similar climates [21,22].

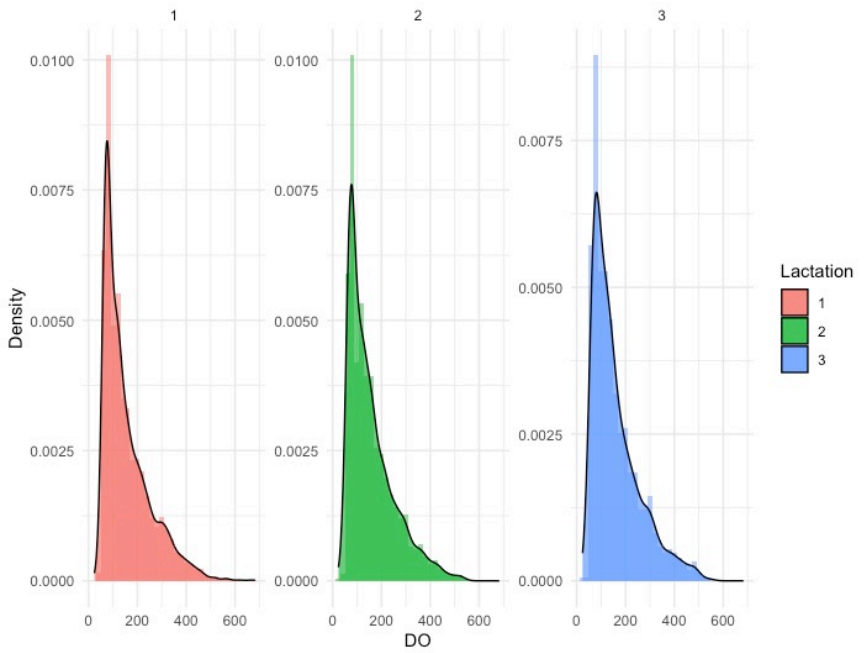

Figure 1 The distribution of days open (DO) across the three first parities

Days in milk at first insemination ranged between 74.91 and 76.91 across the first three lactations. This can be explained by the influence of the management of the farms. The third parity presents a somewhat larger standard deviation than the first two lactations. This might be explained by both a smaller number of observations and a possible larger rate of occurrence of clinical mastitis, levels of somatic cell count, and other health issues [23]. These health issues are shown to increase by parity and affect the variability of DO [24].

Table 1 Descriptive statistics of age at calving, days in milk at first insemination, and days open across the first three lactations: number of records, mean, standard deviation (SD), and minimum and maximum values in the uncensored dataset.

Parity	Number of Records	Age at Calving				Days in Milk at First Insemination				DO			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
1	4183	771.75	77.81	553	993	75.89	27.46	22	499	155.46	96.50	37	683
2	2334	1201.27	116.79	880	1528	74.91	20.15	24	509	156.55	94.10	24	546
3	1033	1628.29	149.93	1196	2080	76.15	51.49	30	1556	159.71	96.81	30	550

The heritabilities of DO (Table 2) ranged from 0.021 to 0.033 across the three different methods. These values are low but somewhat still in the range of what has been published in the literature [19,20,25]. The low heritability can be explained by the strong influence of the environment on the DO [26] on top of the shallow pedigree information [27]. PLM and PTM provided similar heritability results for DO, while LM with no censored records provided a slightly smaller value of 0.02. Garcia et al. [28] used six models to impute censored data and found out that the Gaussian linear model with penalty censored the Gaussian linear model, and the penalized threshold-linear model provided similar heritabilities for age at first calving and calving interval.

Table 2 Posterior means and standard deviation for heritability and variance components for days open (DO) provided by different models. ($\boldsymbol{\sigma_a^2}$is the additive genetic variance; $\boldsymbol{\sigma_{pe}^2}$ is the permanent environment variance; $\boldsymbol{\sigma_e^2}$ is the residual variance; and $\boldsymbol{h^2}$ is the heritability).

Method	$\boldsymbol{\sigma_u^2}$	$\boldsymbol{\sigma_{pe}^2}$	$\boldsymbol{\sigma_e^2}$	$\boldsymbol{h^2}$
LM	160.16 ± 77.04	363.87 ± 229.86	6766.4 ± 258.39	0.021 ± 0.010
PLM	389.22 ± 193.85	490.61 ± 376.20	12166.0 ± 389.59	0.029 ± 0.014
PTM	437.64 ± 217.97	705.76 ± 373.61	11932.0 ± 433.08	0.033 ± 0.016

Similar results were reported by Malhado et al. [29] and Costa et al. [13] for age at first calving in Nellore cattle. Despite similar estimates of heritability, the variance components have changed. The most no-ticeable difference was in the magnitude of all variance components between LM and both PLM and PTM. Posterior means of genetic, permanent environment, and residual var-iances for LM were the smallest. One potential explanation suggested by Hou et al. [12] is the overestimation of residual variances in models that account for censored records due to values beyond the upper limits. In both PLM and PTM, values were assigned equal to the maximum values of DO within each contemporary group plus 21 days. PTM provided higher additive genetic variance compared to that estimated with standard linear models, with or without a penalty, and a smaller residual variance compared to PLM. Similar results were found for Spanish Holstein cows by Gonzàlez-Recio et al. [30]. In fact, the authors reported that a censored Bayesian linear model is theoretically more appropriate than standard linear models for addressing censored data. However, it is important to construct contemporary groups carefully to avoid extreme-class problems (ECPs) where all observations in the same group are either censored or non-censored.

Additionally, the genetic correlation between DO records and the binary trait indicating censorship status could influence the results, providing higher additive genetic variance in the PTM model.

2. The prediction accuracy of the models

We assessed the prediction accuracy using the LR method [16] (Table 3). This approach can provide more accurate estimates of prediction accuracies than the predictive ability method as it does not require the adjusted phenotypes [31]. The overall accuracy rates obtained through the LR method were relatively low and no substantial differences were spotted among the methods. The predictive accuracy of LM (0.16) was somewhat smaller than that of PLM (0.21) and PTM (0.20). Different results were reported by Lázaro et al. [32] where LM exhibited a higher correlation between observed and predicted phenotypes (0.30), followed by PLM with an accuracy of 0.25 for age at first calving. Instead, Urioste et al. [11] reported that the threshold linear model had the highest Pearson correlations between sire breeding value predictions (0.67, 0.68, and 0.67 in the first, second, and third parity, respectively). The low accuracy found in this study can be attributed to the trait's low heritability [12,33] and missing pedigree information in the studied population. According to Latifi and Nadiri [27], remarkably, these two factors influence the prediction accuracy. In fact, in simulated data of sex-limited traits with heritabilities 0.05, 0.1, and 0.2 and different rates of missingness of pedigree information, the lowest accuracy of prediction (0.018 ± 0.006) was for heritability 0.05 when 30% of paternal pedigree was missing and 10% of maternal pedigree was missing. Despite the low predictive accuracy ($\widehat{ACC}_{LR}$) of LM, the model yielded the lowest bias (-0.06) compared to PLM (-0.10) and PTM (-0.14) (Table 3).

The negative bias of the latter two models suggests an underestimation of the true breeding values [34]. This finding was previously reported by Donoghue et al. [7] who claimed that the threshold linear model, when applied to high levels of censoring, slightly underestimates the records. The models' slopes ($Dispersion_{LR}$) were all found to be less than 1, with PTM having a marginally higher slope of 0.73. While a deviation from one may indicate a potential bias in the genetic evaluation, it could also be attributed to the small size of the validation cohort (2052 animals). Legarra and Revereter [16] noted that a dispersion lower than 1 is not always indicative of model quality; it may also depend on the size and relatedness of the animals in the validation set. Furthermore, the sources of biases related to statistical models in animal breeding may stem from various factors, such as the use of incorrect heritability, inaccurate modeling of the age effect, or improper definition of contemporary groups [34]

Table 3 Estimates of $\boldsymbol{Bias_{LR}}$, $\boldsymbol{Dispersion_{LR}}$, *and Spearman correlations between the estimated breeding values of the validation dataset using the three models LM, PLM, and PTM (above diagonal), and the percentage of animals in common in the top 20% of selected individuals (below diagonal).*

Method	LM	PLM	PTM
$\widehat{\boldsymbol{ACC}}_{\boldsymbol{LR}}$	0.16	0.21	0.20
$\boldsymbol{Bias_{LR}}$	−0.06	−0.10	−0.14
$\boldsymbol{Dispersion_{LR}}$	0.53	0.57	0.73
LM		0.80	0.80
PLM	52.67%		0.99
PTM	52.77%	96.34%	

. The results of our study showed that PLM and PTM methods produced similar prediction accuracy. However, when it came to variance components, using PTM resulted in a larger genetic variance and a smaller environmental variance, which led to a slightly higher estimated heritability. This difference may be due to the correlation between the censorship status variable and the DO trait. Additionally, both methods had similar bias and PTM had a slightly higher dispersion. Overall, PTM appears to be a promising method to address censored fertility data in this population. On the other hand, the LM approach had the lowest heritability, genetic variance, predictive accuracy, and dispersion. While the LM approach had a bias closer to 0, the low slope also suggests that there is a sort of bias related to removing all censored data. The Spearman correlation between the predicted breeding values of the validation dataset was 0.99 between PTM and PLM, 0.80 between PLM and LM, and 0.80 between PTM and LM. In the lower diagonal of Table 3, the ranking of animals varies at a higher magnitude between the penalty method, the penalized threshold model, and the Gaussian linear models. Similarly, Lázaro et al. [32] stated that the percentage of animals in common between the methods at 10% of selected individuals was 82.96% between LM and PM, 55.65% between LM and PTM, and 51.48% between PM and PTM, indicating that the largest changes in the ranking were spotted when using the threshold linear model. Different results

were reported by Garcia et al. [28] where there were no significant changes in the ranking of the top 10% of Nellore bulls across the three methods.

Conclusions

Including fertility traits in conducting genetic evaluations is crucial for sustaining the profitability of dairy farms by selecting replacement heifers. Fertility field datasets often contain censored data for numerous reasons. In the context of this study where the dataset is limited, imputing censored records is important. Results for estimating genetic parameters for days open using LM, PLM, and PTM showed that the penalized threshold model marginally increased the trait's heritability compared to linear models. However, the heritability estimates for all methods indicated a reduced genetic gain by selection. Both PLM and PTM yielded better prediction accuracy when using the LR method. Spearman correlations between the estimated breeding values of the validation dataset were high between PLM and PTM, explaining the large proportion of common animals in the top 20%. The lack of changes in the ranking of animals between PLM and PTM suggests that both methods can be used to address censored data in this population.

Acknowledgements

The authors acknowledge the "Les Domaines Agricoles" company for providing data.

References

González-Recio, O.; Alenda, R. Genetic Parameters for Female Fertility Traits and a Fertility Index in Spanish Dairy Cattle. J. Dairy Sci. 2005, 88, 3282–3289. https://doi.org/10.3168/jds.S0022-0302(05)73011-3.

Berry, D.P.; Friggens, N.C.; Lucy, M.; Roche, J.R. Milk Production and Fertility in Cattle. Annu. Rev. Anim. Biosci. 2016, 4, 269–290. https://doi.org/10.1146/annurev-animal-021815-111406.

Ma, L.; Cole, J.B.; Da, Y.; VanRaden, P.M. Symposium Review: Genetics, Genome-Wide Association Study, and Genetic Improvement of Dairy Fertility Traits*. J. Dairy Sci. 2019, 102, 3735–3743. https://doi.org/10.3168/jds.2018-15269.

Weigel, K.A. Prospects for Improving Reproductive Performance through Genetic Selection. Anim. Reprod. Sci. 2006, 96, 323–330. https://doi.org/10.1016/j.anireprosci.2006.08.010.

Haile-Mariam, M.; Pryce, J. Advances in Dairy Cattle Breeding to Improve Fertility/Reproductive Efficiency. In Advances in Breeding of Dairy Cattle; Burleigh Dodds Science Publishing: Cambridge, UK , 2019. ISBN 978-0-429-27560-9.

Turkson, A.J.; Ayiah-Mensah, F.; Nimoh, V. Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review. Int. J. Math. Math. Sci. 2021, 2021, 9307475. https://doi.org/10.1155/2021/9307475.

Donoghue, K.A.; Rekaya, R.; Bertrand, J.K. Comparison of Methods for Handling Censored Records in Beef Fertility Data: Simulation Study1. J. Anim. Sci. 2004, 82, 351–356. https://doi.org/10.2527/2004.822351x.

Johnston, D.J.; Bunter, K.L. Days to Calving in Angus Cattle: Genetic and Environmental Effects, and Covariances with Other Traits. Livest. Prod. Sci. 1996, 45, 13–22. https://doi.org/10.1016/0301-6226(95)00088-7.

Oliveira, H.R.; Miller, S.P.; Brito, L.F.; Schenkel, F.S. Impact of Censored or Penalized Data in the Genetic Evaluation of Two Longevity Indicator Traits Using Random Regression Models in North American Angus Cattle. Animals 2021, 11, 800. https://doi.org/10.3390/ani11030800.

Sorensen, D.A.; Gianola, D.; Korsgaard, I.R. Bayesian Mixed-Effects Model Analysis of a Censored Normal Distribution with Animal Breeding Applications. Acta Agric. Scand. Sect.—Anim. Sci. 1998, 48, 222–229. https://doi.org/10.1080/09064709809362424.

Urioste, J.I.; Misztal, I.; Bertrand, J.K. Fertility Traits in Spring-Calving Aberdeen Angus Cattle. 2. Model Comparison . J. Anim. Sci. 2007, 85, 2861–2865. https://doi.org/10.2527/jas.2006-550.

Hou, Y.; Madsen, P.; Labouriau, R.; Zhang, Y.; Lund, M.S.; Su, G. Genetic Analysis of Days from Calving to First Insemination and Days Open in Danish Holsteins Using Different Models and Censoring Scenarios1. J. Dairy Sci. 2009, 92, 1229–1239. https://doi.org/10.3168/jds.2008-1556.

Costa, E.V.; Ventura, H.T.; Veroneze, R.; Silva, F.F.; Pereira, M.A.; Lopes, P.S. Bayesian Linear-Threshold Censored Models for Genetic Evaluation of Age at First Calving and Stayability in Nellore Cattle. Livest. Sci. 2019, 230, 103833. https://doi.org/10.1016/j.livsci.2019.103833.

Misztal, I.; Lourenco, D.; Aguilar, I.; Legarra, A. Manual for BLUPF90 Family of Programs; University of Georgia: Athens, GA, USA, 2018.

Geweke, J. Bayesian Inference in Econometric Models Using Monte Carlo Integration. Econometrica 1989, 57, 1317–1339. https://doi.org/10.2307/1913710.

Legarra, A.; Reverter, A. Semi-Parametric Estimates of Population Accuracy and Bias of Predictions of Breeding Values and Future Phenotypes Using the LR Method. Genet. Sel. Evol. 2018, 50, 53. https://doi.org/10.1186/s12711-018-0426-6.

Alexandre, P.A.; Li, Y.; Hine, B.C.; Duff, C.J.; Ingham, A.B.; Porto-Neto, L.R.; Reverter, A. Bias, Dispersion, and Accuracy of Genomic Predictions for Feedlot and Carcase Traits in Australian Angus Steers. Genet. Sel. Evol. 2021, 53, 77. https://doi.org/10.1186/s12711-021-00673-8.

R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2020 .

Boujenane, I.; Draga, B. Non-Genetic Factors Affecting Reproductive Performance of Holstein Dairy Cows. Livest. Res. Rural Dev. 2021, 33, 10.

El-Sherief, A.A.; El-Komy, S.M.; Rashad, A.; El-Hedainy, D.K. Reproductive Performance of Lactating Holstein Cows as Influenced by Season of Calving and Parity Under Subtropical Conditions. J. Adv. Vet. Res. 2022, 12, 11–17.

Boujenane, I.; Hilal, B. Genetic and Non Genetic Effects for Lactation Curve Traits in Holstein-Friesian Cows. Arch. Anim. Breed. 2012, 55, 450–457. https://doi.org/10.5194/aab-55-450-2012.

M'hamdi, N.; Bouallegue, M.; Frouja, S.; Ressaissi, Y.; Brar, S.K.; Hamouda, M.B.; M'hamdi, N.; Bouallegue, M.; Frouja, S.; Ressaissi, Y.; et al. Effects of Environmental Factors on Milk Yield, Lactation Length and Dry Period in Tunisian Holstein Cows. In Milk Production—An Up-to-Date Overview of Animal Nutrition, Management and Health; IntechOpen: London, UK, 2012. ISBN 978-953-51-0765-1.

Yusuf, M.; Nakao, T.; Yoshida, C.; Long, S.T.; Gautam, G.; Ranasinghe, R.B.K.; Koike, K.; Hayashi, A. Days in Milk at First AI in Dairy Cows; Its Effect on Subsequent Reproductive Performance and Some Factors Influencing It. J. Reprod. Dev. 2011, 57, 643–649. https://doi.org/10.1262/jrd.10-097T.

Carlén, E.; Strandberg, E.; Roth, A. Genetic Parameters for Clinical Mastitis, Somatic Cell Score, and Production in the First Three Lactations of Swedish Holstein Cows. J. Dairy Sci. 2004, 87, 3062–3070. https://doi.org/10.3168/jds.S0022-0302(04)73439-6.

M'Hamdi, N.; Aloulou, R.; Brar, S.; Mahdi, B.; Ben Hamouda, M. Phenotypic and Genetic Parameters of Reproductive Traits in Tunisian Holstein Cows. Livest. Res. Rural Dev. 2010, 26, 297–307 . https://doi.org/10.2298/BAH1006297M.

Brookfield, J.F.Y. Heritability. Curr. Biol. 2012, 22, R217–R219. https://doi.org/10.1016/j.cub.2012.02.035.

Latifi, M.; Naderi, Y. The Effect of Pedigree Error on Heritability and Accuracy of Prediction of Breeding Value in Threshold Traits. Res. Anim. Prod. 2023, 14, 139–144. https://doi.org/10.61186/rap.14.39.139.

Garcia, D.A.; Rosa, G.J.M.; Valente, B.D.; Carvalheiro, R.; Albuquerque, L.G. Comparison of Models for the Genetic Evaluation of Reproductive Traits with Censored Data in Nellore Cattle1. J. Anim. Sci. 2016, 94, 2297–2306. https://doi.org/10.2527/jas.2016-0273.

Malhado, C.H.M.; Malhado, A.C.M.; Martins Filho, R.; Carneiro, P.L.S.; Pala, A.; Adrián Carrillo, J. Age at First Calving of Nellore Cattle in the Semi-Arid Region of Northeastern Brazil Using Linear, Threshold, Censored and Penalty Models. Livest. Sci. 2013, 154, 28–33. https://doi.org/10.1016/j.livsci.2013.02.021.

González-Recio, O.; Chang, Y.M.; Gianola, D.; Weigel, K.A. Comparison of Models Using Different Censoring Scenarios for Days Open in Spanish Holstein Cows. Anim. Sci. 2006, 82, 233–239. https://doi.org/10.1079/ASC200519.

Hidalgo, J.; Lourenco, D.; Tsuruta, S.; Masuda, Y.; Breen, V.; Hawken, R.; Bermann, M.; Misztal, I. Investigating the Persistence of Accuracy of Genomic Predictions over Time in Broilers. J. Anim. Sci. 2021, 99, skab239. https://doi.org/10.1093/jas/skab239.

Lázaro, S.F.; Varona, L.; Fonseca e Silva, F.; Ventura, H.T.; Veroneze, R.; Brito, L.C.; Costa, E.V.; Lopes, P.S. Censored Bayesian Models for Genetic Evaluation of Age at First Calving in Brazilian Brahman Cattle. Livest. Sci. 2019, 221, 177–180. https://doi.org/10.1016/j.livsci.2018.11.014.

Strandberg, E.; Danell, B. Genetic and Phenotypic Parameters for Production and Days Open in the First Three Lactations of Swedish Dairy Cattle. Acta Agric. Scand. 1989, 39, 203–215. https://doi.org/10.1080/00015128909438513.

Macedo, F.L.; Reverter, A.; Legarra, A. Behavior of the Linear Regression Method to Estimate Bias and Accuracies with Correct and Incorrect Genetic Evaluation Models. J. Dairy Sci. 2020, 103, 529–544. https://doi.org/10.3168/jds.2019-16603.

CHAPTER 4 - EXPLORING THE POTENTIAL OF FEATURE SELECTION THROUGH MACHINE LEARNING AND CONVENTIONAL MODELS TO REDUCE MARKER DATASET DIMENSIONALITY AND ENHANCE GENOMIC PREDICTION ACCURACY: A SIMULATION STUDY

Narjice Chafai[1], Bouabid Badaoui[1]. Unpublished.

Introduction

The widespread use of large-scale genotyping for single nucleotide polymorphisms (SNPs) has created an invaluable opportunity to explore the connections between genomic variation and key traits. In both livestock and plant breeding, leveraging genomic data goes beyond merely identifying complex traits and discovering relevant genetic markers. It significantly improves the accuracy of breeding value predictions, which in turn accelerates genetic progress through a method known as genomic selection (Boichard et al., 2012). This approach has become a preferred tool in genetic evaluation for livestock and plants because it not only boosts precision but also shortens the generation intervals considerably. Numerous studies have shown that genomic selection outperforms traditional methods of genetic evaluation (Doublet, 2019.; Ruiz-López et al., 2018; Thomasen et al., 2014; Xu et al., 2020). In dairy cattle, for instance, genomically estimated breeding values (GEBVs) show an accuracy improvement of 12% over those obtained through the conventional BLUP method (Lee et al., 2023). Moreover, genomic selection enables the evaluation of young animals at birth or earlier, eliminating the long wait for phenotypic data to be collected, which traditionally delays reliable genetic assessments. When implementing genomic selection through variance component methods like GBLUP and ssGBLUP, the pedigree-based relationship matrix (A) is replaced by the genomic relationship matrix (G), which is derived from SNP genotype data. This genomic approach offers distinct advantages over pedigree-based methods by correcting possible pedigree errors, revealing previously unknown relationships, and providing a more accurate representation of Mendelian sampling effects (Goddard et al., 2010). The application of genomic selection (GS) through methods such as GBLUP and ssGBLUP necessitates the inversion of the genomic relationship matrix (G). However, the direct inversion of G has become computationally unfeasible for many livestock populations due to the rapid growth in the number of genotyped animals. For instance, in the U.S. dairy cattle population, over 6 million animals have been genotyped, including more than 950,000 Holstein cows (Guinan et al., 2022). Moreover, the computational burden of inverting G grows at a cubic rate in relation to the number of genotyped individuals, posing a significant challenge even for datasets of moderate size. A potential solution would be to reduce the redundancy in the G matrix by reducing the number of genotyped animals as described by Misztal et al. (2015) to approximate G matrix in the context of the Algorithm of Proven and Young. Another approach would be to select relevant SNPs that are tracking QTLs and use Bayesian alphabets to directly estimate the effects of these functional SNPs and eventually estimate the breeding values by summing these effects. Additionally, prioritizing sets of SNPs and using them to build a genomic

relationship Matrix has also been investigated by Chang et al. (2019) and were found to increase the genomic accuracy while reducing remarkably the statistical and computational power of the genetic models.

Additionally, the rapid growth of genomic data in livestock breeding and other biological domains has led to an increased need for effective methods to manage and analyze high-dimensional datasets. One of the key challenges in genomic studies is handling the sheer volume of genetic markers, particularly single nucleotide polymorphisms (SNPs), which can reach tens or hundreds of thousands in size. This high dimensionality can hinder the accuracy of genomic predictions, making it essential to develop methods that can reduce dataset size without compromising prediction performance. Feature selection, the process of identifying the most informative variables in a dataset, has emerged as a promising approach to tackle this issue. Feature selection methods can be broadly categorized into three types: filtering methods, wrapping methods, and embedded methods. (i) Filtering methods rely on ranking features based on specific criteria, assessing the importance of each variable before any modeling occurs. These methods evaluate each feature independently, assigning scores to rank them based on their relevance to the target variable. Features that fall below a certain threshold are removed, making this approach efficient and easy to implement. Due to its simplicity, filtering is widely used in practical applications, where the focus is on identifying variables that provide meaningful distinctions between different classes. (ii) Wrapping methods, on the other hand, use a more dynamic approach by involving the model itself in the feature selection process. These methods treat the model as a "black box," evaluating subsets of features based on how well they enhance the model's performance. Since testing all possible combinations of features can be computationally expensive, especially with larger datasets, heuristic search algorithms like Genetic Algorithms (GA) or Particle Swarm Optimization (PSO) are employed to find suboptimal but effective feature subsets. Although this approach often leads to better performance, it is computationally intensive due to the repeated training and testing involved. (iii) Embedded methods aim to integrate feature selection directly into the model's learning process, reducing the computational burden seen in wrapping methods. Instead of evaluating different subsets of features after training, embedded methods perform feature selection as part of the model's training itself. This integration leads to more efficient feature selection, as it eliminates the need to repeatedly reclassify subsets. Commonly used in machine learning algorithms like decision trees and regularization techniques (e.g., Lasso), embedded methods strike a balance between computational efficiency and predictive accuracy (Chandrashekar & Sahin, 2014). Each of these methods offers distinct advantages depending on the complexity of the dataset and the computational resources

available, making them valuable tools for improving model performance and reducing dimensionality. Machine learning (ML) techniques are known for their ability to manage large datasets and identify patterns. These models can be employed to select a subset of relevant markers, thereby reducing computational complexity while maintaining or even enhancing prediction quality. Additionally, conventional methods such as principal component analysis (PCA) have also been used for selecting SNPs (Song et al., 2010). However, feature selection methods that eliminate unnecessary or irrelevant features from a dataset, use a specific criterion i to assess how important each feature is in predicting the output or target variables. In machine learning, if a model uses irrelevant features, it can negatively impact the system's ability to generalize to new data, resulting in poor performance. Nonetheless, removing irrelevant features is different from dimension reduction methods like Principal Component Analysis (PCA), because good features can still be important on their own, even if they are independent of other features in the dataset. Therefore, feature selection and PCA serve different purposes in data analysis (Chandrashekar & Sahin, 2014).

This study aims to investigate the potential of various machine learning models, including regression with L1 and L2 regularizations, gradient boosting machines, deep neural networks, and PCA to perform feature selection and explore potential improvement in genomic prediction accuracy. By employing a simulation study, we evaluate the performance of these methods across different heritability levels, assessing their effectiveness in reducing the dimensionality of SNP datasets and their impact on genomic prediction outcomes.

Materials and methods

1. Simulation

Data used in this study was simulated using QMSim software (Sargolzaei & Schenkel, 2009). The simulation process involved two steps as illustrated in Figure 1. Initially, a historical population of 4000 animals underwent random mating for 1000 generations, followed by 1100 additional generations with population growth from 250 to 25000 animals. This phase established linkage disequilibrium (LD) and mutation-drift equilibrium (Toghiani et al., 2017). The founder population (G0) was formed by randomly selecting 100 males and 2,000 females from the final historical generation. Subsequently, five generations were created and used for the analysis. The final pedigree file comprised 30100 animals with records (Figure 1).

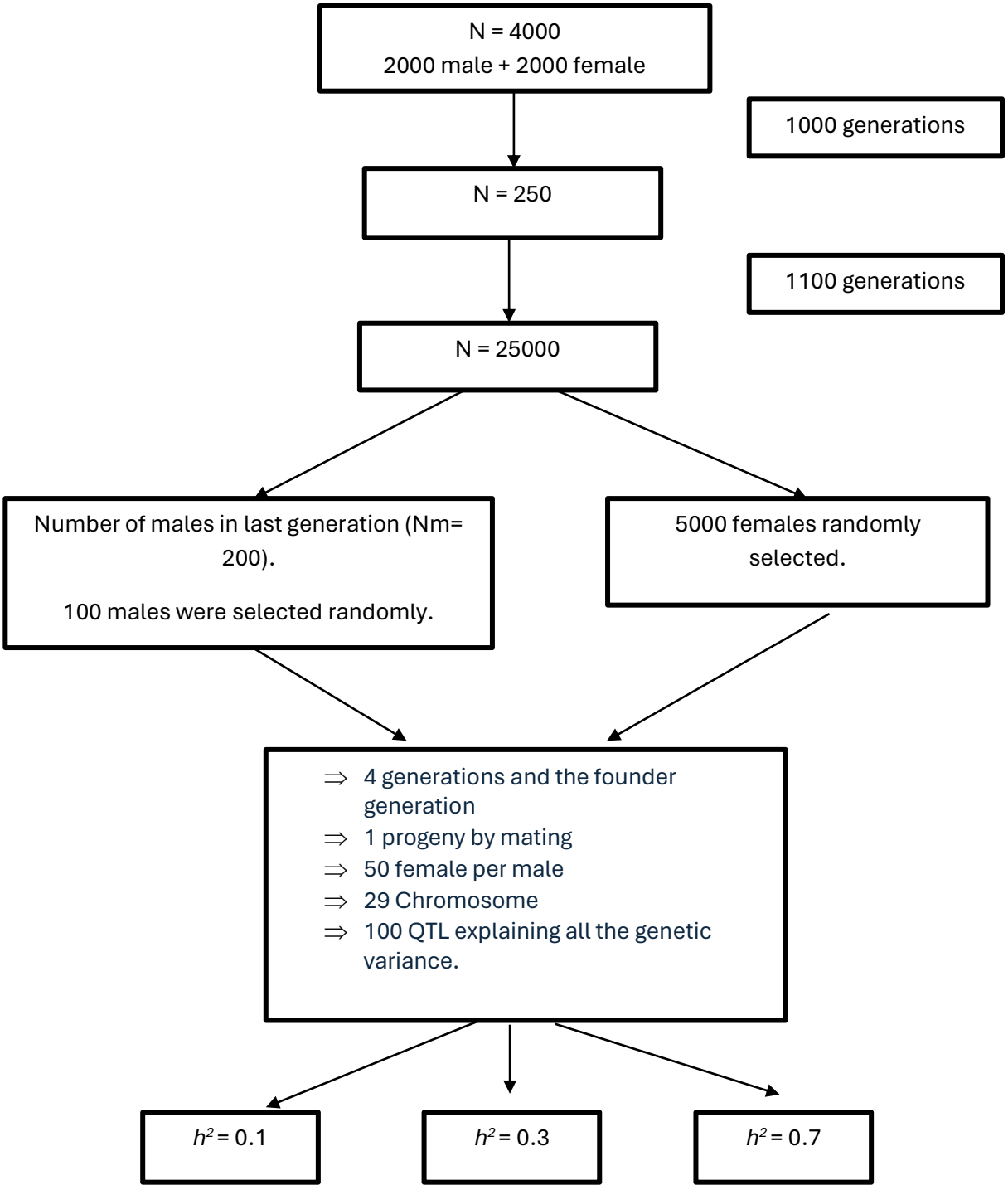

Figure 1 Simulation parameters for the three scenarios.

For generations G0 to G4, animal selection was based on estimated breeding values (EBVs), with 50% male and 30% female replacement rates to maintain offspring sex ratio balance. Each dam produced one offspring. A trait with heritability of 0.7, 0.3 or 0.1 was simulated, assuming all genetic variance was explained by the simulated QTL. Phenotypic variance was standardized to one, with residual variance adjusted to achieve the desired heritability. True breeding values were calculated as the sum of QTL additive effects. Trait phenotypes were generated by adding random errors from a normal distribution with zero mean and standard deviation equal to residual variance.

The simulated genome mimicked the cattle genome with 29 autosomal chromosomes. Fifty thousand SNP markers were uniformly distributed across the genome. In all scenarios, 100 QTLs accounted for 100% of the genetic variance. Both SNPs and markers were biallelic, with no overlap between marker loci and QTL. Three replicates were generated for each simulation. The genotyped animals (2000) were randomly selected from the 4^{th} and 5^{th} generations.

Prior to analysis, genomic data underwent quality control using PreGSf90 of BLUPF90 programs (Misztal et al., 2014). This process removed SNPs and animals with call rates below 0.90, minor allele frequencies (MAF) below 0.05, and parent-progeny conflicts, while also checking for Hardy-Weinberg equilibrium (HWE).

2. Feature selection methods

This study used 4 machine learning feature selection methods and the conventional Principal Component Analysis (PCA) to select the most relevant 100, 500, and 1000 SNPs from a 50k SNP panel.

2.1.Principal Component Analysis (PCA)

Principal component analysis is a powerful statistical technique used to distill the essential features, known as principal components, from a data table consisting of cases and variables. These principal components are comprised of a small number of linear combinations of the original variables that maximize the explanation of variance across all variables. This method also provides a simplified representation of the original data table by utilizing only these key components. In essence, it offers a way to condense complex datasets while preserving their most significant characteristics (Greenacre et al., 2022). Song et al. (2010) used for the first time PCA for feature selection and described the algorithm as follows: (i) Computing the covariance matrix using the training data and calculating the Eigenvalues and Eigenvectors of the covariance matrix. Each eigenvector represents a direction in feature space and the

corresponding eigenvalues quantifies the variance in that direction. (ii) Selecting the eigenvectors with m largest eigenvalues $V_1, V_2, \ldots, V_m$. (iii) Calculating the feature contribution using the formula $c_j = \sum_{p=1}^{m} |V_{pj}|$ where V_{pj} refers to the jth entry of the *pth* eigenvector. (iiii) Finally, ranking and selecting features by first sorting c_j in descending order. The largest contributions correspond to the most important features. PCA was performed in this study using "*prcomp*" package of base R version 4.3.1 (R Core Team, 2023).

2.2. Ridge regression

Ridge regression is a variant of linear regression that incorporates a penalty term to mitigate overfitting and enhance model generalization, particularly in scenarios where the data exhibits multicollinearity or when the system in underdetermined (number of predictors exceeds the number of observations) (Imani & Ghassemian, 2015). In ordinary least squares (OLS) regression, the model aims to minimize the sum of squared residuals. However, ridge regression introduces a regularization term to the cost function, which comprises the sum of the squared coefficients (the L2 norm). This regularization term functions to shrink the coefficients toward zero, thereby discouraging large values and regulating model complexity. Ridge regression aims to minimize the following equation:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad [29]$$

where y_i are the observed values; $\hat{y}_i$ are the predicted values; β_j are the regression coefficients, and λ is the regularization term. A larger λ results in more regularization and hence more shrinkage of the coefficients. Ridge regression was performed in this study using glm function of H2o framework in R (H2O.ai., 2022).

2.3.Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is also a type of linear regression that, analogous to ridge regression, incorporates regularization to enhance model performance by mitigating overfitting. However, in contrast to ridge regression, LASSO employs an L1 regularization penalty, which is defined as the sum of the absolute values of the regression coefficients. This regularization methodology tends to generate sparse models by reducing some coefficients precisely to zero, thereby effectively conducting variable selection and

diminishing the number of predictors. LASSO models aim to minimize the following equation (Muthukrishnan & Rohini, 2016):

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad [30]$$

where y_i represents the observed values; $\hat{y}_i$ the predicted values; β_j the regression coefficients, and λ the regularization term. A larger λ forces more coefficients to zero. LASSO was performed in this study using glm function of H2o framework in R (H2O.ai., 2022).

2.4.Gradient Boosting Machines

The fundamental principle underlying GBMs is the incremental construction of a model through the sequential combination of predictions from multiple decision trees (weak predictors), wherein each subsequent model aims to rectify the errors made by its predecessors. The algorithm initiates by fitting an initial model, generally a simple one such as a decision tree with limited depth, to the data. Subsequently, it computes the residuals. Another weak model is trained to predict these residuals. This process is iterated for a predetermined number of iterations or until a stopping criterion is met. At each stage, the new model focuses on minimizing the residual errors from the previous model. This process is referred to as "boosting". In our study, GBM was performed using gbm function of H2O framework in R (H2O.ai., 2022). 100 decision trees with a maximum depth of 4. The learning process is governed by mean squared error loss function. The learning rate that regulates the contribution of each weak model was chosen through grid search. The final learning rate used was 0.1.

2.5.Deep Neural Networks (DNN)

Deep neural networks model consists of a number of stacked layers containing several neurons. Each neuron is linked to the neurons in neighboring layers through weights, which indicate the strength and direction of the connection, either enhancing or suppressing the signal. In a Deep Neural Network (DNN), a set of observations X is fed to the model through the input layer, where each observation x_i serves as both the input and output for this layer. In the hidden layers, each neuron in a given layer receives a weighted sum of outputs from the neurons in the previous layer, processes it through an activation function, and produces an output. Common activation functions in the hidden layers include the Rectified Linear Unit (ReLU), hyperbolic tangent, and sigmoid functions. In the output layer, the DNN performs either classification or regression, depending on the type of target variable. In our study, a DNN model

was constructed using the h2o.deeplearning function in H2O framework (H2O.ai., 2022). The model comprised 3 hidden layers, each consisting of 32 neurons was run using a Rectified Linear Unit with Dropout. The input layer consisted of 50k SNPs, and the target variable was the phenotypes of the 2000 genotyped animals. The input dropout ratio was set to 0.2 to reduce overfitting. An L1 regularization was applied with a value of 10^{-5} to induce sparsity in the model. Finally, the model was trained for 10 epochs to balance training time and performance and the variable importance scores were analyzed to identify key predictors.

3. Single-step GBLUP (ss-GBLUP)

Single-step GBLUP was used to estimate the breeding values of genotyped and non-genotyped animals using the full 50K SNP panel, 100, 500, and 1000 selected SNPs using the different feature selection methods. The univariate animal model used in this analysis is:

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{u} + \boldsymbol{e} \quad [3]$$

where $var(\boldsymbol{u}) = \mathbf{H}\sigma_u^2$, and $var(\boldsymbol{e}) = \mathbf{I}\sigma_e^2$, and the MME can be written as:

$$(\mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\boldsymbol{\lambda})\hat{\boldsymbol{u}} = \mathbf{W}'\boldsymbol{y} \quad [4]$$

where $\hat{u}$ is the estimated random genetic effect; and $\boldsymbol{W}$ is the incidence matrix with appropriate dimensions; $\boldsymbol{\lambda}$ is the ratio between the genetic and the residual variance, and $\mathbf{H^{-1}}$ is the inverse of the augmented genomic relationship matrix. $\mathbf{H^{-1}}$ can be written as:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{\mathbf{22}}^{-1} \end{pmatrix} \quad [5]$$

where $\mathbf{A}^{-1}$ is the inverse of the average relationship matrix; $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix between genotyped animals and $\mathbf{A}_{\mathbf{22}}^{-1}$ is the inverse of average relationship matrix between genotyped animals.

ssGBLUP was performed using BLUPF90 suite (Misztal et al., 2014).

4. Performance metrics

The performance of the feature selection models was evaluated using 5-fold cross-validation. To assess the reliability and goodness of fit of these models, the mean squared error (MSE), root mean squared

error (RMSE), and mean absolute error (MAE) were computed using the h2o library in R (H2O.ai., 2022). The formulae for these metrics are presented in Table 1.

The MSE and RMSE quantify the average squared difference between predicted and actual values. These metrics assign greater weight to larger errors, rendering them particularly sensitive to outliers. The RMSE, being expressed in the same units as the target variable, offers enhanced interpretability. Conversely, the MAE calculates the average absolute difference between predicted and actual values, treating all errors equally regardless of magnitude. This characteristic makes the MAE less susceptible to outliers compared to MSE and RMSE, and it is often preferred when equal importance is to be given to all prediction errors. For all three metrics, lower values indicate superior model performance, signifying smaller prediction errors.

Table 1 Performance metrics used to assess Feature selection methods' performance.

Metric abbreviation	Metric Name	Metric Formula
MSE	Mean Squared Error	$MSE = \frac{1}{N}\sum_{n=1}^{N}[y(n) - \hat{y}(n)]^2$
RMSE	Root Mean Squared Error	$RMSE = \sqrt{MSE}$
MAE	Mean Absolute Error	$MAE = \frac{1}{N}\sum_{n=1}^{N}\lvert[y(n) - \hat{y}(n)]^2\rvert$

To compare the efficacy of PCA, LASSO, ridge regression, DNN, and GBM in feature selection, and to evaluate the relevance of the selected 100, 500, and 1000 SNPs, we calculated the correlation between true breeding values (computed as the sum of QTL effects provided by QMSim software) and the breeding values obtained using the full 50k SNP panel, as well as the 100, 500, and 1000 SNPs selected by each of the aforementioned methods.

Results and discussion

Table 2 and Table 3 presents the performance metrics including MSE, RMSE, and MAE of the 4 machine learning models Ridge regression, LASSO, GBM, and DNN in the training phase and validation phase, respectively. The tables consist of the mean output of the three replicates for each heritability. The comparative analysis of model performance in the training phase shows that RIDGE regression

demonstrated consistently low error rates (MSE: 0.005707327, RMSE: 0.07554685, MAE: 0.05770006) across all heritability levels. However, this exceptional performance did not translate to the validation set, where RIDGE consistently underperformed, suggesting potential overfitting. Conversely, GBM, which showed moderate performance in training, emerged as the top performer in the validation set, exhibiting the lowest error rates across all heritability levels. This indicates robust generalization capabilities for GBM. LASSO displayed variable performance in the training set, with effectiveness decreasing at higher heritability levels, but showed more consistent performance in the validation set. Notably, DNN, which underperformed in the training set, demonstrated competitive performance in the validation set, particularly at higher heritability levels, suggesting potential underfitting during training but good generalization. This finding can be supported by the small size of training data, as only 2000 animals were randomly selected to be genotyped, while DNN models require large and good quality training dataset (Taylor & Nitschke, 2017).

Table 2 the mean of MSE, RMSE, and MAE across replicates of the four feature selection methods for the three different scenarios (heritabilities) using the training set.

Heritability	Method	MSE	RMSE	MAE
0.1	LASSO	0.0431	0.2077	0.1649
	RIDGE	0.0057	0.0755	0.0577
	GBM	0.0895	0.2992	0.2391
	DNN	1.0477	1.0236	0.8154
0.3	LASSO	1.1309	1.0634	0.8460
	RIDGE	0.0057	0.0755	0.0577
	GBM	0.0895	0.2992	0.2391
	DNN	1.0516	1.022	0.8165
0.7	LASSO	1.0477	1.0235	0.8154
	RIDGE	0.0057	0.0755	0.0577
	GBM	0.0895	1.0254	0.2391
	DNN	1.0477	1.0235	0.8154

The discrepancies observed between training and validation performance underscore the critical importance of cross-validation in model selection and evaluation, particularly in the context of genetic feature selection where model generalization is crucial. These results highlight that while RIDGE may excel in capturing training data patterns, GBM offers superior generalization, making it potentially more suitable for practical applications in genetic feature selection tasks. Similar results were found for soy bean, where Gill et al. (2022) compared the performance of XGBoost, random forest, convolutional neural networks, and deep neural networks in selecting SNPs that are relevant to seven agronomic traits. The authors reported that the boosting algorithm (XGBoost) outperformed the other methods particularly for categorical traits, while for continuous traits RF was the best performer.

Table 3 the mean of MSE, RMSE, and MAE across replicates of the four feature selection methods for the three different scenarios (heritabilities) using the 5 fold validation sets.

Heritability	Method	MSE	RMSE	MAE
0.1	LASSO	1.1309	1.0634	0.8460
	RIDGE	1.2179	1.1035	0.8846
	GBM	1.0397	1.0197	0.8167
	DNN	1.0571	1.0282	0.8175
0.3	LASSO	1.2179	1.1035	0.8846
	RIDGE	1.2902	1.1358	0.9197
	GBM	1.0199	1.0099	0.8018
	DNN	1.0451	1.0223	0.8136
0.7	LASSO	1.0505	1.0249	0.8170
	RIDGE	1.2179	1.1035	0.8846
	GBM	1.0240	1.0119	0.8016
	DNN	1.0505	1.0249	0.8170

Concerning the correlation between the true breeding values (TBVs) and the estimated breeding values (EBVs) using the full 50k SNP and the sets of SNPs selected by ML methods and PCA (Figure 2, 3 , and 4), GBM outperformed the other methods by providing close or even higher correlation coefficients than using the full 50K SNP panel. The mean correlation coefficient across the three replicates of every scenario provided good correlation with TBVs across all heritabilities. Using 500 SNPs selected by GBM provided a higher correlation than using the full 50k SNPs. However, this scenario was not consistent across all heritabilities. Therefore, we can include that GBM provided similar results of using the full panel. PCA provided the least reliable estimated breeding values, particularly when the heritability of the trait was low to moderate. LASSO and Ridge regression provided similar results. However, Ridge tended to overfit the training data which lowers the generality of using this method for feature selection. Overall, selecting 1000 SNPs using GBM provided good results and similar EBVs to using the full 50k SNPs. This finding indicates that GBM identified successfully functional SNPs that are in LD with QTLs. Research has indeed shown that selecting specific genetic markers can reduce the complexity of genomic prediction models or even improve their accuracy (Bermingham et al., 2015; Jeong et al., 2020; Li et al., 2018). This approach can maintain predictive accuracy while using fewer single nucleotide polymorphisms (SNPs), and in some cases, even enhance model performance (Heinrich et al., 2023). This improvement is particularly notable when the number of subjects is significantly less than the genetic features analyzed. By reducing the feature set, models may become less prone to overfitting and more adept at generalizing to new data. Even when the performance merely matches that of more complex models, the ability to achieve similar results with fewer SNPs is valuable. It paves the way for developing trait-specific, low-density SNP arrays. These streamlined tools can significantly reduce genotyping expenses, making genomic prediction a more economically viable option for breeding programs. In addition, in the context where the number of genotyped animals is increasing, and the computational cost of inverting G matrix is cubic to the number of genotyped animals, selecting relevant SNPs that are tracking QTLs and estimating their effect using Bayesian alphabets would be a potential solution. The results of our study show that Gradient boosting machines tend to identify relevant SNPs which lead to similar or even higher correlation with TBVs compared to using the full 50k SNP panel. GBMs strengths lie in their capacity to capture complex patterns in data, offer high predictive accuracy, and provide inherent feature importance metrics. However, they may be susceptible to overfitting if not properly tuned, and the training process can be computationally intensive compared to simpler models such as linear regression (Breiman et al., 2017)

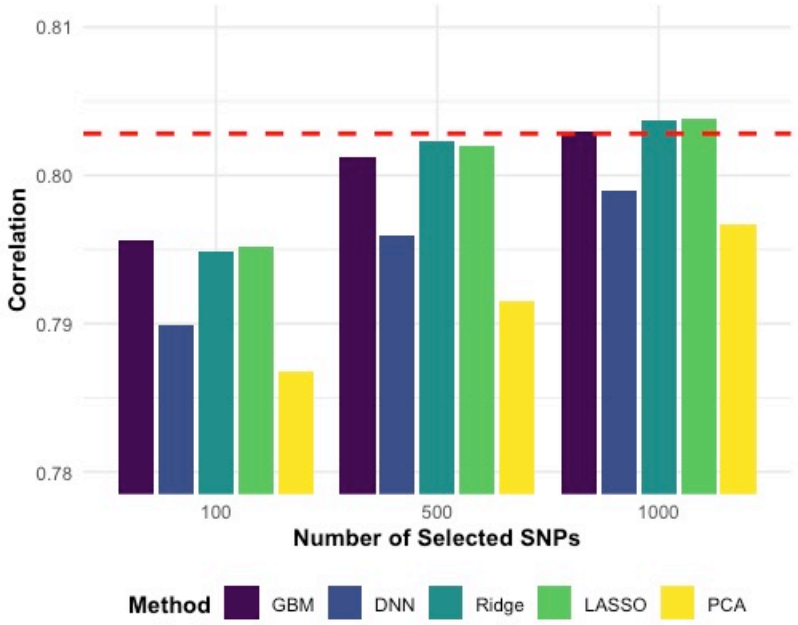

Figure 2 The correlations between EBVs and TBVs using the full 50k SNP panel, 100, 500, and 1000 SNPs selected using Ridge regression, LASSO, GBM, DNN, and PCA for the low heritability trait

**The dashed red line refers to the correlation between TBVs and EBVs using the whole 50k SNP panel.*

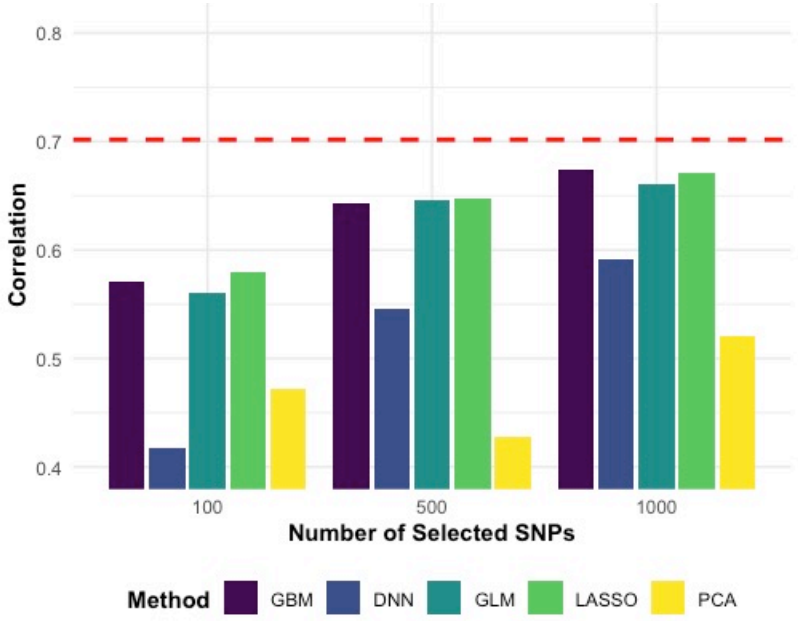

Figure 3 The correlations between EBVs and TBVs using the full 50k SNP panel, 100, 500, and 1000 SNPs selected using Ridge regression, LASSO, GBM, DNN, and PCA for the moderate heritability trait

**The dashed red line refers to the correlation between TBVs and EBVs using the whole 50k SNP panel.*

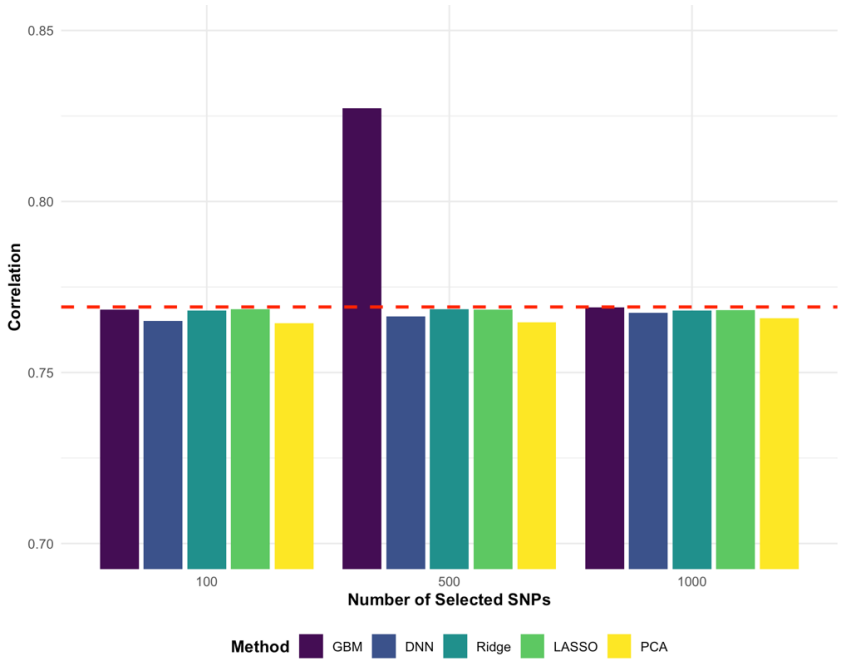

Figure 4 The correlations between EBVs and TBVs using the full 50k SNP panel, 100, 500, and 1000 SNPs selected using Ridge regression, LASSO, GBM, DNN, and PCA for the high heritability trait

**The dashed red line refers to the correlation between TBVs and EBVs using the whole 50k SNP panel.*

Conclusion

The advancements in genotyping technology have resulted in increasingly larger genotype and marker datasets, significantly raising the computational cost and complexity of genomic selection. One effective approach to address this challenge is feature selection, which helps reduce the dimensionality of data while maintaining predictive accuracy. In this study, various feature selection methods were used to identify 100, 500, and 1000 relevant SNPs for genomic prediction. The results demonstrated that using Gradient Boosting Machine (GBM) to select relevant SNPs produced comparable or even superior correlations with true breeding values (TBVs) compared to the full 50k SNP panel. GBM showed strong generalization ability, particularly when selecting 1000 SNPs, leading to similar estimated breeding values (EBVs) as using the entire SNP set. Ridge regression, although effective in training, suffered from overfitting, while PCA was the least reliable, especially at lower heritability levels. Ultimately, GBM's capacity to capture complex data patterns and provide important feature selection metrics makes it a promising tool for reducing dataset complexity in genomic selection, paving the way for more efficient and cost-effective breeding programs.

References

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., & Haley, C. S. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, *5*, 10312. https://doi.org/10.1038/srep10312

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., Druet, T., Genestout, L., Colleau, J. J., Journaux, L., Ducrocq, V., & Fritz, S. (2012). Genomic selection in French dairy cattle. *Animal Production Science*, *52*(3), 115. https://doi.org/10.1071/AN11119

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification And Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, L.-Y., Toghiani, S., Hay, E. H., Aggrey, S. E., & Rekaya, R. (2019). A Weighted Genomic Relationship Matrix Based on Fixation Index (FST) Prioritized SNPs for Genomic Selection. *Genes*, *10*(11), Article 11. https://doi.org/10.3390/genes10110922

Doublet, A.-C. (n.d.). *La diversité génétique à l'ère de la génomique: Évolution de la consanguinité et ses conséquences dans trois races bovines laitières françaises*. 188.

Gill, M., Anderson, R., Hu, H., Bennamoun, M., Petereit, J., Valliyodan, B., Nguyen, H. T., Batley, J., Bayer, P. E., & Edwards, D. (2022). Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. *BMC Plant Biology*, *22*(1), 180. https://doi.org/10.1186/s12870-022-03559-z

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2010). Genomic selection in livestock populations. *Genetics Research*, *92*(5–6), 413–421. https://doi.org/10.1017/S0016672310000613

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 1–21. https://doi.org/10.1038/s43586-022-00184-w

Guinan, F. L., Wiggans, G. R., Norman, H. D., Dürr, J. W., Cole, J. B., Van Tassell, C. P., Misztal, I., & Lourenco, D. (2022). Changes in genetic trends in US dairy cattle since the implementation of genomic selection. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2022-22205

H2O.ai. (2022) *h2o: R Interface for H2O*. R package version 3.42.0.2. https://github.com/h2oai/h2o-3.

Heinrich, F., Lange, T. M., Kircher, M., Ramzan, F., Schmitt, A. O., & Gültas, M. (2023). Exploring the potential of incremental feature selection to improve genomic prediction accuracy. *Genetics Selection Evolution*, *55*(1), 78. https://doi.org/10.1186/s12711-023-00853-8

Imani, M., & Ghassemian, H. (2015). Ridge regression-based feature extraction for hyperspectral data. *International Journal of Remote Sensing*, *36*(6), 1728–1742. https://doi.org/10.1080/01431161.2015.1024894

Jeong, S., Kim, J.-Y., & Kim, N. (2020). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Scientific Reports*, *10*(1), 19653. https://doi.org/10.1038/s41598-020-76759-y

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., & Li, Y. (2018). Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. *Frontiers in Genetics*, *9*, 237. https://doi.org/10.3389/fgene.2018.00237

Misztal, I., Fragomeni, B. O., Lourenco, D. A. L., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., & Lawlor, T. (2015). Efficient inversion of genomic relationship matrix by the algorithm for proven and young (APY). *Interbull Bulletin*, *49*, Article 49. https://journal.interbull.org/index.php/ib/article/view/1602

Misztal, I., Lourenco, D., Aguilar, I., Legarra, A., & Vitezica, Z. (2014). *Manual for BLUPF90 family of programs*. 149.

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 18–20. https://doi.org/10.1109/ICACA.2016.7887916

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ruiz-López, F. J., García-Ruiz, A., Wiggans, G. R., & Van Tassell, C. P. (2018). Impact of Genomic Selection on Genetic Gain of Net Merit of US Dairy Cattle. *Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.(710)*, 11–16.

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*(5), 680–681. https://doi.org/10.1093/bioinformatics/btp045

Song, F., Guo, Z., & Mei, D. (2010). Feature Selection Using Principal Component Analysis. *Engineering Design and Manufacturing Informatization 2010 International Conference on System Science*, *1*, 27–30. https://doi.org/10.1109/ICSEM.2010.14

Taylor, L., & Nitschke, G. (2017). *Improving Deep Learning using Generic Data Augmentation* (arXiv:1708.06020). arXiv. https://doi.org/10.48550/arXiv.1708.06020

Thomasen, J. R., Egger-Danner, C., Willam, A., Guldbrandtsen, B., Lund, M. S., & Sørensen, A. C. (2014). Genomic selection strategies in a small dairy cattle population evaluated for genetic gain and profit. *Journal of Dairy Science*, *97*(1), 458–470. https://doi.org/10.3168/jds.2013-6599

Toghiani, S., Chang, L.-Y., Ling, A., Aggrey, S. E., & Rekaya, R. (2017). Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock. *Livestock Science*, *205*, 24–30. https://doi.org/10.1016/j.livsci.2017.09.007

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B. M., Olsen, M. S., Wang, G., & Zhang, A. (2020). Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Communications*, *1*(1).

CONCLUSIONS

In conclusion, this thesis provides valuable insights into the genetic improvement of Moroccan Holstein cows, focusing on both milk yield and fertility traits. The complex relationship between these traits, particularly the antagonistic correlation between high milk yield and reduced fertility, presents a significant challenge for breeders. However, the moderate genetic correlations identified in this study suggest that it is possible to improve fertility without sacrificing milk production, by employing a comprehensive selection index that includes all relevant traits.

Additionally, our thesis sheds light on fertility traits, such as the number of inseminations per conception and success rate at first insemination, are critical to the economic sustainability of dairy herds. These traits, however, pose analytical challenges due to their non-normal distribution and censored nature. To address the issue of incomplete fertility data, this research applied a range of models, with the penalized threshold model demonstrating the highest heritability. Both the penalty and threshold methods were effective in imputing censored data, proving to be valuable tools in genetic evaluation.

Finally, the application of machine learning (ML) in genomic prediction was another focal point of this thesis. While ML models hold great promise for handling large, complex datasets, their adoption in animal breeding remains in its early stages. Continued research is necessary to fully explore the potential of these models in improving genomic predictions. The simulation study further demonstrated the effectiveness of feature selection methods, particularly Gradient Boosting Machine (GBM), which outperformed other techniques by identifying relevant SNPs and producing reliable breeding value estimates.

Overall, this thesis emphasizes the importance of integrating fertility traits into genetic evaluations, utilizing advanced statistical and machine learning methods to optimize breeding strategies. The findings contribute to a more efficient and sustainable approach to dairy cow breeding in Morocco.

FUTURE DIRECTIONS

For future research, several areas warrant further exploration to enhance the genetic improvement of dairy cows in Morocco. First, expanding the dataset to include more comprehensive fertility records would improve the accuracy of genetic parameter estimates. Given the complexity of fertility traits, especially their non-normal distribution and the presence of censored data. Threshold linear model with a penalty and the penalty method have indeed shown promise in handling incomplete data in our case, but their broader applicability across diverse populations needs further validation. Moreover, the antagonistic relationship between milk yield and fertility highlights the need to refine selection indices that balance these traits effectively. Future studies should focus on including to the genetic evaluation, new fertility traits that are highly linked to farms' profitability and traits related to health issues and longevity. Estimating the genetic components of these traits and assessing their economic influence can help integrating them into multi-trait selection indices with economic weights. This would help mitigate the negative effects of intense selection for milk production in the Moroccan environment.

Another crucial step for the genetic improvement of dairy cows in Morocco is the genotyping of Holstein cows and the establishment of a national reference population. Currently, the reliance on genetic evaluations from foreign populations may not fully capture the unique environmental and management conditions specific to Morocco. By genotyping a substantial number of Holstein cows, Morocco can develop its own reference population, which would allow for more accurate and relevant genomic predictions tailored to local breeding objectives. A national reference population would also facilitate the identification of genetic markers associated with traits important to Moroccan dairy systems, such as heat tolerance and disease resistance, alongside milk yield and fertility traits. Building this population would provide a foundation for implementing genomic selection on a large scale, significantly accelerating genetic progress. Moreover, this initiative could reduce dependency on imported genetics, making the breeding program more self-sufficient and sustainable. Over time, a well-established Moroccan reference population could contribute to the development of breeding strategies that address the specific challenges faced by the dairy industry in the region, ultimately boosting productivity and profitability.

Furthermore, in order to increase genetic gain through genomic selection for Moroccan Holstein where the number of genotyped animals is limited and these animals are mostly cows with low reliability, seeking for potential collaboration between Morocco and other countries, especially those from where

Holstein heifers are mostly imported can help address this challenge. A future study incorporating the genotypes of Moroccan Holstein cows and other animals (bulls and cows) from other countries can increase the reliability of genomic selection, the estimation of breeding values, and the prediction of future phenotypes.

Additionally, the application of machine learning (ML) for genomic prediction holds significant potential, but its use in livestock breeding is still in its early stages. Future research should investigate the scalability of ML models, such as GBM, for larger datasets, and explore other advanced techniques, like DNN and CNN for feature selection and phenotype prediction using large genotype dataset as our study was limited to a simulated dataset with a small number of genotyped animals. Additionally, more iterative experiments are needed to assess the potential of these models in improving accuracy and reducing computational complexity, especially in challenging environments with limited data, as is often the case in Morocco.

Finally, the implementation of advanced sensor technologies for the automatic recording of phenotypes, particularly those related to milk quality, activity, rumination, and environmental indices such as temperature and humidity on medium- and large-scale farms in Morocco holds significant potential to improve animal welfare and farm management practices. This approach not only provides high-throughput phenotypic data, which can enhance the accuracy of genetic evaluations by ensuring high-quality phenotypic records and precise animal identification, but also facilitates the comprehensive assessment of environmental effects on the performance of Holstein cows. Additionally, leveraging artificial intelligence algorithms to analyze these data can further assist farm managers in making informed decisions, thus contributing to more rational and economically sound management strategies.

Finally, the integration of genomic selection with reproductive management strategies, such as optimizing insemination practices, could further enhance the genetic progress in fertility traits. Combining cutting-edge genetic tools with practical herd management could pave the way for more resilient and productive dairy herds in Morocco, contributing to long-term industry sustainability.

Résumé (200 mots max.)

Cette thèse traite l'amélioration génétique des vaches laitières au Maroc, avec un focus sur l'estimation des paramètres génétiques des traits de production et de reproduction des vaches Holstein. Elle révèle que l'héritabilité du rendement laitier est modérée, tandis que celle des traits de fertilité est faible, rendant la sélection directe pour ces derniers difficile. Toutefois, l'intégration de traits de fertilité et de production dans un indice de sélection pourrait aider à limiter la baisse de la fertilité. Pour les données censurées, le modèle à seuil s'avère efficace pour l'imputation des phénotypes, améliorant la précision prédictive, ce qui pourrait être une solution pour les bases de données des vaches Holstein au Maroc. La thèse analyse également l'utilisation de l'intelligence artificielle dans la prédiction des valeurs génétiques, en évaluant ses effets sur la précision et la complexité computationnelle. En comparant plusieurs méthodes de sélection de SNPs, notamment l'ACP, le Gradient Boosting Machines, la régression Ridge et LASSO, le GBM a démontré un avantage en sélectionnant 500 et 1000 SNPs. La précision obtenue avec 1000 SNPs sélectionnés par GBM est presque équivalente à celle obtenue avec un panel complet de 50 000 SNPs.

Mots-clefs (5) : paramètres génétiques, fertilité, rendement laitier, intelligence artificielle, SNPs.

Abstract

This thesis addresses the genetic improvement of dairy cows in Morocco, with a focus on estimating the genetic parameters of production and reproduction traits in Holstein cows. It reveals that the heritability of milk yield is moderate, while that of fertility traits is low, making direct selection for the latter difficult. However, integrating fertility and production traits into a selection index could help mitigate the decline in fertility. For censored data, the threshold model proves effective for imputing phenotypes, enhancing predictive accuracy, which could be a solution for Holstein cow databases in Morocco. The thesis also analyzes the use of artificial intelligence in predicting genetic values, evaluating its effects on accuracy and computational complexity. By comparing several SNP selection methods, including PCA, Gradient Boosting Machines, Ridge regression, and LASSO, GBM demonstrated an advantage in selecting 500 and 1000 SNPs. The accuracy achieved with 1000 SNPs selected by GBM is almost equivalent to that obtained with a full panel of 50,000 SNPs.

Keywords (5) : Genetic parameters, fertility, milk yield, artificial intelligence, SNPs.