**Application**

To predict dairy cows with high residual feed intake using machine learning models to increase the precision of genomic selection in Holstein dairy cattle.

**Introduction**

The economic viability of livestock production systems, particularly dairy operations, is significantly constrained by feed costs, which account for the majority of production expenses (Alqaisi & Schlecht, 2021). Profitability can be improved by identifying and cultivating herds with cattle that exhibit high feed efficiency (FE), allowing feed costs to be minimized without reductions in milk yield. The use of genomic selection combined with precision livestock farming (PLF) has been identified as a method to enable the detection of cattle with desirable genetic traits for FE, supporting the development of breeding programs aimed at improving FE across entire herds (Madilindi et al., 2022). However, FE is recognized as a complex trait, and its measurement within genomic selection programs is hindered by the limited availability of direct observational data. Machine learning (ML) approaches have been proposed as a means to address these challenges by facilitating the integration of large and complex genomic datasets. These techniques can enhance the identification of genetic markers associated with FE and improve the predictive accuracy of genomic selection models (Mota et al., 2024). In this study, ML algorithms were applied to genomic data from Holstein dairy cattle to predict residual feed intake (RFI), which was estimated using feed intake and milk production data

**Materials and Methods**

Genotypic data from 195 Holstein-Friesian dairy cattle at the Agri-Food and Biosciences Institute (AFBI) were collected using the 50K Illumina Bovine SNP panel. Over the period of 2014 to 2023, a total of 9984 weekly records of individual cattle milk production, feed consumption and weight data were used to calculate the target phenotype, RFI. RFI was regressed using a random effect model and were classified as high or low RFI cattle. The predictions of RFI were then modeled using Genomic Best Linear Unbiased Prediction (GBLUP) in addition to various ML algorithms, including logistic regression models (LogReg), support vector machines (SVM), stochastic gradient boosting (SGB), and random forests (RF) using genotype data. Feature selection was conducted to select the top 1% of SNPs for predicting RFI using random forest. Training and validation datasets were created by randomly splitting the data, with 70% used for training and 30% reserved for validation. Fisher’s Least Significant difference (LSD) test of the accuracy of machine learning (ML) algorithms was performed to evaluate the performance metric between the models.

**Results**

The logistic regression model achieved the highest accuracy values demonstrating significantly better performance than other ML models using the LSD test. The mean accuracy scores for LogReg is 0.79, RF is 0.76, SVM is 0.74, DT is 0.58 and GBLUP is 0.51 where GBLUP achieved the lowest accuracy scores as compared to the other ML models (Figure 1). The findings demonstrated that ML-driven approaches can effectively predict complex traits such as RFI, highlighting the potential of these methodologies in genomic selection. Genetic markers identified as highly associated with RFI will be further analyzed using gene ontology to explore the biological pathways involved, offering deeper insights into the genetic mechanisms underlying FE.

A chart with blue squares and black text

Description automatically generated

Figure 1: Boxplot of the distribution of test accuracy valyes for random forest (RF), support vector machine (SVM), logistic regression (LogReg Model), decision tree (DT) and Genomic best linear unbiased prediction (GBLUP)

**Conclusions**

This study demonstrates the benefit of using machine learning to enhance genomic selection for complex traits like feed efficiency in dairy cattle. The machine learning models achieved significantly higher accuracy values than traditional GBLUP model. The performance of ML algorithms, particularly logistic regression had the highest mean accuracy scores (0.79) and GBLUP (0.51). This highlights a more effective method in identifying genetic markers for breeding optimization in dairy production systems.

**References**

Alqaisi, O., & Schlecht, E. (2021). Feeding Models to Optimize Dairy Feed Rations in View of Feed Availability, Feed Prices and Milk Production Scenarios. *Sustainability*, *13*(1), 215. <https://www.mdpi.com/2071-1050/13/1/215>

Madilindi, M. A., Zishiri, O. T., Dube, B., & Banga, C. B. (2022). Technological advances in genetic improvement of feed efficiency in dairy cattle: A review. *Livestock Science*, *258*, 104871. <https://doi.org/https://doi.org/10.1016/j.livsci.2022.104871>

Mota, L. F., Arikawa, L. M., Santos, S. W., Fernandes Júnior, G. A., Alves, A. A., Rosa, G. J., Mercadante, M. E., Cyrillo, J. N., Carvalheiro, R., & Albuquerque, L. G. (2024). Benchmarking machine learning and parametric methods for genomic prediction of feed efficiency-related traits in Nellore cattle. *Scientific Reports*, *14*(1), 6404.