*Livestock Studies, 2025;65(2), 85-89*
http://doi.org/10.46897/livestockstudies.1848872

LIVESTOCK STUDIES

RESEARCH ARTICLE

# Comparative Evaluation of Machine Learning Models—LASSO and Elastic Net—for Genetic Association Mapping Using Simulated Phenotype Data

**Semih YAZICI[1]*, Yalçın YAMAN[1]**

[1] Siirt University Veterinary Medicine Faculty, Department of Genetics, Siirt, Türkiye.

**Article History**
Received: 15 Dec 25
Accepted: 22 Dec 25

**Corresponding Author**
E-mail: vet.semihyazici@gmail.com

**Keywords**
Phenotype prediction
SNP selection
Regularized regression
High-dimensional genomics
Simulation study

## Abstract

Traditional Genome-Wide Association Studies (GWAS), have several limitations that have prompted the search for more advanced analytical methods. Machine learning (ML) models have emerged as promising alternatives. This study evaluates two regularized regression models, LASSO (Least Absolute Shrinkage and Selection Operator) and Elastic Net, implemented via the GLMNET package, for phenotype prediction and SNP selection. Genotype data from the Sheep HapMap consortium (Sheep 50K) were combined with phenotypes simulated using Genome-wide Complex Trait Analysis (GCTA) under three heritability scenarios ($h^2$ = 0.1, 0.3, 0.56). After quality control, imputation, and LD pruning, 38,448 SNPs and 2,819 individuals were retained. Model performance increased with heritability. At low heritability ($h^2$ = 0.1), both models showed limited predictive power (Elastic Net: $R^2$ = 0.079; LASSO: $R^2$ = 0.091). Performance improved at moderate heritability ($h^2$ = 0.3), with Elastic Net achieving $R^2$ = 0.415 and LASSO $R^2$ = 0.385. At high heritability ($h^2$ = 0.6), both models achieved moderate-to-strong predictive accuracy (Elastic Net: $R^2$ = 0.672; LASSO: $R^2$ = 0.683). Concordance between the top 50 SNPs identified by both models was high across scenarios (84%, 90%, and 100%). In conclusion, the utility of ML-based regularization methods for association mapping in high-dimensional genomic studies.

## Introduction

Genetic association studies aim to identify genomic regions that contribute to phenotypic variation in complex traits. Traditional genome-wide association studies (GWAS) predominantly rely on single-marker statistical tests, which are often insufficient to fully capture the polygenic architecture underlying complex traits and face substantial challenges when applied to high-dimensional genomic datasets. In addition, classical GWAS approaches are sensitive to confounding factors such as linkage disequilibrium, population structure, and multiple testing burden, often leading to reduced statistical power and inflated false-positive rates (Waldmann et al., 2013; Hong et al., 2014; Bush and Moore, 2012). As high-throughput genotyping technologies continue to advance, modern datasets routinely include tens or even hundreds of thousands of single nucleotide polymorphisms (SNPs), further exacerbating these limitations and underscoring the need for more scalable and robust analytical frameworks.

To address these challenges, multivariate and machine learning–based approaches have gained increasing attention in genomic research. In particular, regularized regression models such as Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net provide an effective means of jointly modeling large numbers of correlated genetic markers while controlling model complexity (Tibshirani, 1996; Zou and Hastie, 2005). By incorporating penalty terms, these methods mitigate overfitting, handle multicollinearity among SNPs, and enable simultaneous phenotype prediction and variable selection, making them especially suitable for high-dimensional genomic applications where the number of predictors far exceeds the number of observations. Recent studies have demonstrated the potential of regularized machine learning models in genomic

prediction and association mapping across both human and animal populations (Ogutu et al., 2012; Lourenço et al., 2014). However, despite their growing use, several important aspects of their performance remain insufficiently explored. In particular, the influence of underlying genetic architecture and heritability on model behavior, predictive accuracy, and SNP selection consistency has not been systematically evaluated. Many existing studies rely on empirical phenotypes with unknown genetic properties, making it difficult to disentangle methodological performance from biological complexity.

In this context, the use of simulated phenotypes offers a controlled and reproducible framework for benchmarking machine learning–based regularization methods. Simulation-based approaches allow for explicit control over heritability levels and genetic architecture, thereby enabling a more rigorous and interpretable comparison of model performance across different scenarios.

Accordingly, the present study aims to comparatively evaluate LASSO and Elastic Net models using simulated phenotype data under multiple heritability settings. By assessing both predictive performance and SNP selection concordance, this work provides novel insights into the strengths and limitations of regularized regression methods for high-dimensional genomic association analyses.

## Material and Methods

### Genotype and Phenotype Data

Genotype data used in this study were obtained from the Sheep HapMap Consortium and consisted of 2,819 individuals genotyped with the Sheep 50K SNP array, containing 46,925 markers. Phenotypes were generated using the Genome-wide Complex Trait Analysis (GCTA) software under three different heritability scenarios ($h^2$ = 0.1, 0.3, and 0.56). Each simulated phenotype dataset was produced using the same genotype matrix, enabling direct comparison of model performance across varying heritability levels.

### Population Structure Analysis

The potential for confounding effects arising from population stratification necessitates careful evaluation in genomic studies. To address this crucial factor, Principal Component Analysis (PCA) was performed on the quality-controlled genotype data. This analysis served to identify and quantify the genetic structure within the studied populations.

### SNP Filtering

Quality control procedures were performed in PLINK 1.9 using commonly applied thresholds. SNPs with a call rate below 0.95, minor allele frequency lower than 0.01, or Hardy-Weinberg equilibrium p-values below $1 \times 10^{-6}$ were excluded. Individuals with a call rate below 0.95 were also removed. After these filtering steps, a total of 38,448 SNPs and the complete set of 2,819 individuals remained. Missing genotypes were then imputed using Beagle 5.0 with default parameter settings, and the resulting dataset was used for subsequent analyses. In order to minimize multicollinearity caused by linkage disequilibrium (LD), LD pruning was carried out in PLINK using a 50-SNP sliding window, a 5-SNP step, and an $r^2$ threshold of 0.2; the pruned dataset was used for all machine learning analyses.

### Machine Learning Analyses

Phenotypes were simulated in GCTA by specifying the desired heritability values, where genetic and environmental variances were defined proportionally to the target $h^2$ for each scenario. This approach provided a controlled analytical framework

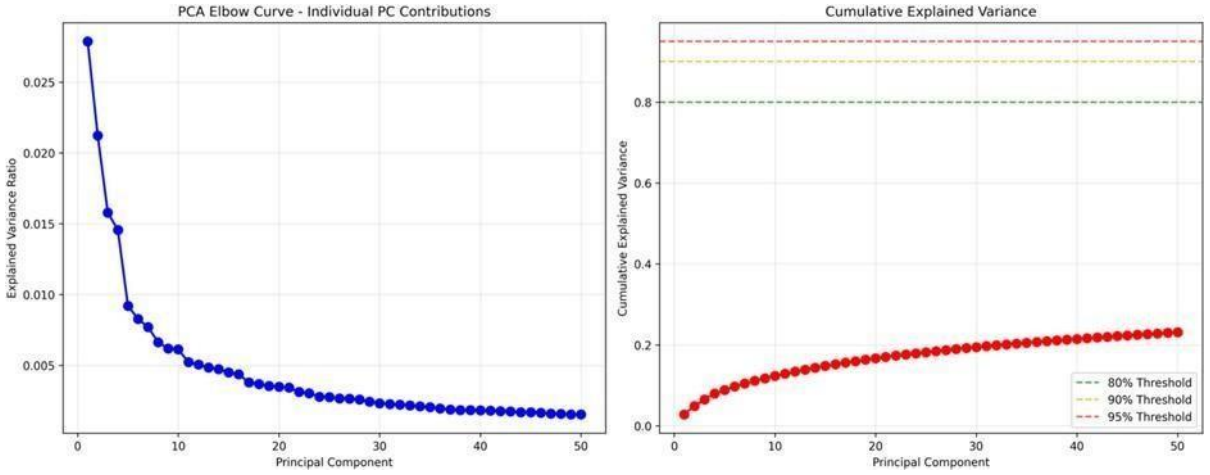


**Figure 1.** The elbow method was applied to determine the optimal number of principal components. Ten principal components were identified, explaining more variance than the others.

for evaluating the accuracy, stability, and SNP detection capability of machine learning models under different levels of genetic determinism.

Machine learning analyses were performed in R using the GLMNET package, applying both LASSO regression and Elastic Net regression. For LASSO, an L1 regularization penalty was used to achieve simultaneous prediction and variable selection. For Elastic Net, a combined L1 and L2 penalty was applied with α set to 0.5, giving equal weight to both penalty types. In both methods, the regularization parameter (λ) was determined using 10-fold cross-validation. Model performance was evaluated based on the coefficient of determination ($R^2$), the correlation between observed and predicted phenotypes, and the number of SNPs selected by each method. Additionally, the overlap of top-ranked SNPs between LASSO and Elastic Net was calculated for each heritability scenario to assess the consistency of variable selection. All data processing, quality control, phenotype simulation, and statistical analyses were conducted using PLINK v1.9, Beagle v5.0, GCTA, and R (version 4.x), with GLMNET serving as the primary machine learning implementation.

## Results and Discussion

### Population Structure

The Principal Component Analysis (PCA) scatter plot reveals significant population structure among the analyzed sheep breeds. PC1 (2.79% variance) and PC2 (2.12% variance) capture the main axes of genetic differentiation.

After quality control and linkage disequilibrium (LD) pruning, a total of 38,448 SNPs and 2,819 individuals were retained for downstream analyses. Phenotypes simulated under three heritability scenarios ($h^2$ = 0.1, 0.3, 0.6) yielded realised heritability estimates of 0.11, 0.26, and 0.60, respectively, consistent with the simulation design.

### Model Evaluation

Both LASSO and Elastic Net models were trained using 10-fold cross-validation. Model performance varied substantially across heritability levels. At $h^2$ = 0.1, predictive accuracy was low for both methods: Elastic Net yielded $R^2$ = 0.079 and correlation = 0.399, whereas LASSO produced $R^2$ = 0.091 and correlation = 0.421. At $h^2$ = 0.3, predictive performance increased markedly. Elastic Net achieved $R^2$ = 0.415 and correlation = 0.710, while LASSO showed $R^2$ = 0.385 and correlation = 0.686. At $h^2$ = 0.6, both models reached moderate-to-high performance. Elastic Net resulted in $R^2$ = 0.672 and correlation = 0.862, and LASSO performed similarly with $R^2$ = 0.683 and correlation = 0.866. The number of non-zero SNP coefficients selected by the models decreased as heritability increased and model sparsity improved. At $h^2$ = 0.1, Elastic Net selected 91 SNPs and LASSO 108 SNPs. At $h^2$ = 0.3, the models selected 292 and 318 SNPs, respectively; and at $h^2$ = 0.6, the selected SNP counts further increased due to stronger genetic signal.

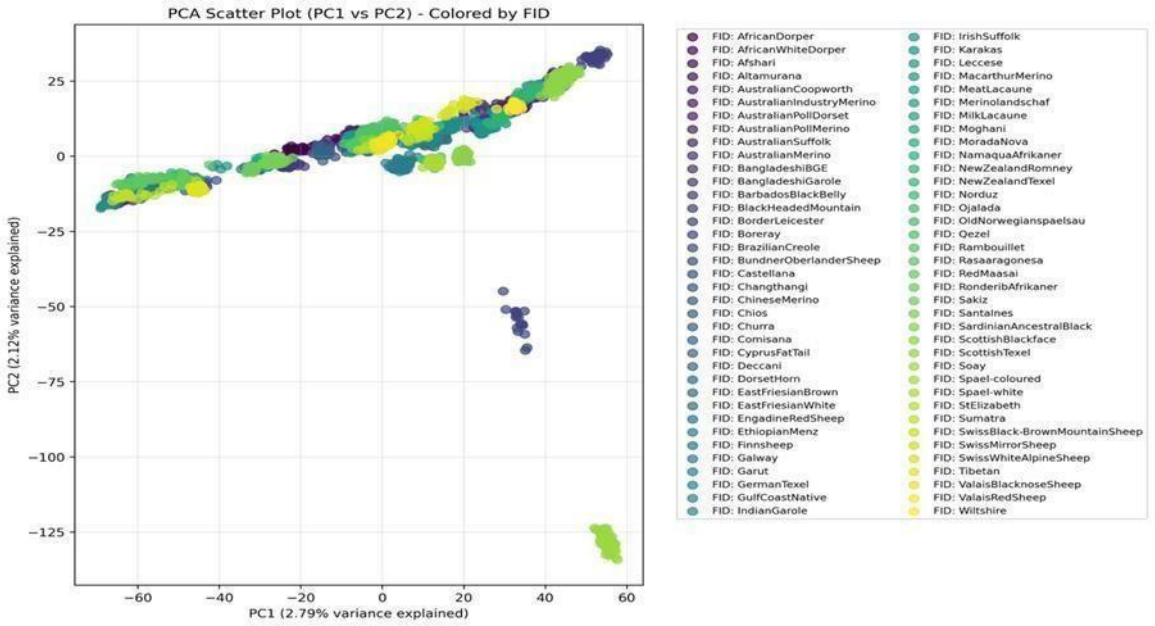


**Figure 2.**The analysis indicates that groups such as Blackheadedmountain, Border Leicester, Australian breeds, and Scottish Texel exhibit clear divergence from the main genetic pool, while the remaining breeds are concentrated within a single principal plane, reflecting close genetic relationships.

Across all scenarios, the overlap among the top 50 highest-weighted SNPs identified by both models was substantial, with concordance rates of 84%, 90%, and 100% for $h^2$ = 0.1, 0.3, and 0.6, respectively. These results indicate strong agreement between LASSO and Elastic Net in SNP ranking, especially at higher heritability levels. Traditional Genome-Wide Association Studies, despite their advantages in identifying genotype–phenotype relationships, face several limitations that have prompted the search for more advanced analytical methods. ML models have emerged as promising alternatives, offering new opportunities for improved prediction and variable selection (Miao et al., 2024). Therefore, ML approaches offer flexible and powerful alternatives to classical statistical methods for genetic association studies, especially when dealing with high-dimensional genomic datasets.

Simulation-based approaches have been extensively used to benchmark genome-wide association study methodologies under controlled genetic architectures. For instance, Cebeci et al. (2023) compared traditional and advanced GWAS methods using simulated quantitative traits with varying heritability levels in domesticated goats, demonstrating that multi-locus approaches such as BLINK and FarmCPU outperform single-locus models, particularly in controlling false positives and improving detection power under different genetic architectures. Previous studies primarily focused on evaluating classical and multi-locus GWAS models, such as mixed linear models, FarmCPU, and BLINK, using simulated phenotypes to assess statistical power and false-positive control (Porter & O'Reilly, 2017; Meyer & Birney, 2018; Tang&Liu, 2019). While these studies provided valuable insights into the performance of traditional GWAS frameworks, they largely emphasized hypothesis-testing paradigms and single-marker or multi-locus statistical inference.

In contrast, relatively limited attention has been given to machine learning–based regularized regression models, such as LASSO and Elastic Net, within a controlled simulation framework explicitly designed for genetic association mapping. Existing simulation tools have rarely been used to systematically evaluate the dual capability of these models in simultaneous phenotype prediction and SNP selection across varying heritability levels. Moreover, direct comparisons of LASSO and Elastic Net in terms of model stability, predictive accuracy, and consistency of SNP detection under different genetic architectures remain scarce.

The present study addresses this gap by conducting a comprehensive simulation-based evaluation of LASSO and Elastic Net models for genetic association mapping. By analyzing simulated phenotypes with low, medium, and high heritability, this work provides a structured assessment of how regularized machine learning models perform under different genetic signal strengths. The findings contribute to the growing body of evidence supporting the use of machine learning approaches as complementary tools to traditional GWAS methods, particularly in high-dimensional genomic settings where multicollinearity and variable selection pose significant analytical challenges. Nevertheless, validation of these findings using real-world genotype–phenotype datasets is essential, and future studies should extend this framework to empirical data to confirm the robustness and practical applicability of machine learning–based association mapping approaches.

**Limitations and Future Directions**

Despite the advantages of simulation-based analyses in providing a controlled and reproducible framework, several limitations of the present study should be acknowledged. First, the use of simulated phenotypes, while allowing precise control over genetic architecture and heritability, may not fully capture the complexity of real-world genotype–phenotype relationships. Factors such as gene–gene interactions, gene–environment effects, and unobserved population stratification are inherently simplified in simulation settings.

Second, although this study focuses on evaluating LASSO and Elastic Net under varying heritability scenarios, the results are contingent upon the specific simulation assumptions, including effect size distributions and the proportion of causal variants. Different genetic architectures may influence model performance differently, and therefore the findings should be interpreted within the context of the defined simulation framework.

Future research should extend this comparative framework to empirical genotype–phenotype datasets to validate the robustness and generalizability of the observed results. Applying machine learning–based regularization methods to real-world populations will be essential for assessing their practical utility in the presence of biological noise, complex linkage disequilibrium patterns, and heterogeneous population structures. Additionally, integrating non-linear machine learning models and exploring hybrid approaches that combine classical GWAS and machine learning techniques may further enhance the detection of biologically meaningful genetic associations.

## Conclusion

In conclusion, the findings indicate that while regularized ML models perform poorly under low-heritability conditions, they exhibit substantial predictive ability and high SNP-selection

consistency when $h^2 \geq 0.3$. Although empirical validation is needed, these results support the utility of ML-based regularization methods for association mapping in high-dimensional genomic studies.

## Author Contributions

All authors contributed to manuscript drafting, critical review.

## Conflict of Interest

The author(s) declare that they have no known competing financial or non-financial, professional, or personal conflicts that could have appeared to influence the work reported in this paper.

## Ethical Statement

This study was conducted based on advanced analytical methods without using any animal material.

## References

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology, 8(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Cebeci, Z., Bayraktar, M., & Gökçe, G. (2023). Comparison of the statistical methods for genome-wide association studies on simulated quantitative traits of domesticated goats (Capra hircus L.). Small Ruminant Research, 227, 107053. https://doi.org/10.1016/j.smallrumres.2023.107053

Cho, S., Kim, H., Oh, S., Kim, K., & Park, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proceedings, 3(S7), S25. https://doi.org/10.1186/1753-6561-3-S7-S25

Hong, S., Kim, Y., & Park, T. (2014). Practical issues in screening and variable selection in genome-wide association analysis. Cancer Informatics, 13(Suppl. 7),CIN.S16350. https://doi.org/10.4137/CIN.S16350

Li, S., Yu, J., Kang, H., & Liu, J. (2022). Genomic selection in Chinese Holsteins using regularized regression models for feature selection of whole genome sequencing data. Animals, 12(18), 2419. https://doi.org/10.3390/ani12182419

Lourenço, V. M., Ogutu, J. O., Rodrigues, R. A., Posekany, A., & Piepho, H. (2024). Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. BMC Genomics, 25(1), 152. https://doi.org/10.1186/s12864-023-09933-x

Miao, J., Wu, Y., Sun, Z., Miao, X., Lu, T., Zhao, J., & Lu, Q. (2024). Valid inference for machine learning-assisted GWAS. medRxiv. https://doi.org/10.1101/2024.01.03.24300779

Meyer, H. V., & Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. Bioinformatics, 34(17), 2951–2956. https://doi.org/10.1093/bioinformatics/bty197

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings,6(S2). https://doi.org/10.1186/1753-6561-6-s2-s10

Porter, H. F., & O'Reilly, P. F. (2017). Multivariate simulation framework reveals performance of multi-trait GWAS methods. Scientific Reports, 7(1), 38837. https://doi.org/10.1038/srep38837

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Statistical Methodology), 58(1), 267–288. https://doi.org/10.1111/j.2517- 161.1996.tb02080.x

Tang, Y., & Liu, X. (2019). G2P: a Genome-Wide- Association-Study simulation tool for genotype simulation, phenotype simulation and power evaluation. Bioinformatics, 35(19), 3852–3854. https://doi.org/10.1093/bioinformatics/btz126

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the LASSO and the elastic net in genome-wide association studies. Frontiers in Genetics, 4, 270. https://doi.org/10.3389/fgene.2013.00270

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B (Statistical Methodology), 67(2),301–320. https://doi.org/10.1111/j.1467- 9868.2005.00503.x