Meta-analysis on the Possible Application of Machine Learning (ML) in Leguminous Plant Production: Faba bean and Common vetch

Abstract - This meta-analysis demonstrates that machine learning (ML) combined with meta-analysis (MA) holds revolutionary promise for improving understudied legumes like common vetch and faba bean. By synthesizing 115 studies, it reveals ensemble methods excel in disease resistance (AUC 0.88–0.91) and yield prediction (R^2 up to 0.92). ML enhances genomic selection (85–95% accuracy) and root phenotyping. Key challenges include data standardization and computational demands for large genomes. The integration of ML and MA accelerates trait discovery for precision breeding. Precision breeding for nutritional quality and climatic resistance is made possible by the faster trait discovery made possible by the combination of ML and MA.Future work will focus on explainable AI, multi-omics integration, and cloud-based pipelines to bridge the gap between model and neglected crops.

Authors

Ebube Oliver Chukwunyere[1], Valery Anatolievich, Burlutsky [1], and Meisam Zargar[1].

Affiliations

[1]Department of Agro-Biotechnology, Peoples' Friendship University of Russia, Moscow, Russian Federation

Introduction

Legumes, such as faba beans (*Vicia faba*) and common vetch (*Vicia sativa*), serve as sustainable, plant-based protein sources crucial for food security in developing nations facing population growth and climate change (Ferrari et al., 2022; Feddern et al., 2024). Advances in interdisciplinary research integrating genetics, molecular biology, and machine learning (ML) are revolutionizing plant breeding. ML enables precision breeding through two pathways: (1) **knowledge-driven approaches** for polygenic traits (e.g., disease resistance) by identifying agronomically valuable genes, and (2) **data-driven modeling** for complex quantitative traits (e.g., yield) using predictive algorithms (Yan and Wang, 2023). With global demand for animal products projected to rise 60–70% by 2050, legumes are vital for feed and food, yet production faces challenges like land degradation and climate variability. ML optimizes breeding by analyzing omics data, but gaps remain in its application to legume value chains (e.g., feed processing). A **meta-analysis (MA)** of ML in legume production is needed to synthesize global data, neutralize study heterogeneity, and ensure reproducibility (Lee, 2023). ML models (e.g., random forests, SVMs) excel at pattern recognition in high-dimensional data, offering insights for trait prediction and stress resilience (Lima et al., 2021).

Objective

The synergy of MA and ML can transform plant genetics by uncovering hidden data trends (Tahmasebi et al., 2023). While ML algorithms (e.g., XGBoost, CNNs) enhance predictive accuracy, their success depends on data quality and biological complexity highlighting challenges like noisy physiological data in *Vicia villosa* (hairy vetch) versus robust genomic resources in *Phaseolus vulgaris* (common bean). This study addresses the research gap by meta-analyzing ML's role in legume production, focusing on *Vicia* species, to guide sustainable breeding and precision agriculture.

Methodology

This study followed PRISMA 2020 guidelines to conduct a systematic review and meta-analysis on machine learning (ML) applications in legume production (2015–2025). Searches in the Dimension database used terms like *"Machine learning AND Legumes production"* and focused on metrics (RMSE, F1-score, AUC, etc.). Eligible studies covered faba beans, common vetch, and other legumes, with 549 initial papers screened. After removing duplicates/incomplete data, 115 articles were reviewed. Data was categorized into sub-themes (yield prediction, disease detection, etc.). The unregistered protocol lacked a structured search plan but ensured standardized methodology for transparency. Figure 1 details the selection process.

Results

This meta-analysis reviews machine learning (ML) applications across 30 legume species from 82 studies. Herbaceous forage legumes dominate (15 species), including *Lens culinaris*, *Vicia faba*, and *Phaseolus vulgaris*. Tropical neglected legumes (7 species, e.g., *Lablab purpureus*) and multipurpose fodder trees/shrubs (6 species, e.g., *Vicia villosa*) were also analyzed. Forage vetches (*Vicia* spp.) and *P. vulgaris* exhibited high diversity in ML evaluations. Using stability parameters, predictive modeling, and clustering techniques, the analysis identifies species-specific strengths and challenges in breeding and agronomic management

Analysis

The study addressed heterogeneity in legume production data by categorizing species into herbaceous forage legumes, tropical neglected legumes, and multipurpose fodder trees/shrubs. This stratification minimized publication bias and enabled focused analysis. Machine learning (ML) metrics were grouped into four tasks: (1) **Classification** (F1-score, AUC, accuracy), (2) **Multi-class classification** (Cohen's kappa, confusion matrix), (3) **Regression** (RMSE, R^2, Pearson's R), and (4) **Clustering/Feature Importance** (Silhouette Score, SHAP values). These metrics evaluated genomic and phenotypic trait predictions, with applications across legume crops like *Vicia sativa* and *Faba sativa*. The approach standardized ML comparisons despite diverse agroecological conditions, growth patterns, and data limitations (Ajay et al., 2020; and Ipekesen et al., 2024).

Stability and Genotype × Environment Interaction (GEI)	Machine Learning Performance	Nutritional and Biochemical Traits	Clustering and Genomic Insights
Lentil (*Vicia lens*) exhibits high stability (Eberhart & Russell: 0.8–1.2; low Shukla's σ^2_i) with minimal GEI, making it reliable for low-input systems.	**Disease Resistance** determination using ML models a high F1-scores for disease traits was achieved in (lentil: 0.75–0.90; and common bean: 0.80–0.95) due to well-characterized genetic markers. Faba bean scores lower (0.65–0.80) due to *Botrytis fabae* strain diversity.	**Faba bean** pods show 37× higher L-DOPA AUC than seeds, suggesting nutraceutical potential.	**Common bean** seed traits achieve the best clustering (Silhouette Score: 0.86; DBI: 0.41), reflecting advanced breeding.
Faba bean (*Vicia faba*) shows high GEI sensitivity (Eberhart & Russell: 0.7–1.5), requiring environment-specific breeding.	**Stress Tolerance** determination as drought prediction is robust for common vetch (F1: 0.70–0.85) but poor for hairy vetch (F1: 0.60–0.75) due to phenotypic plasticity.	**Common bean** anthocyanins (AUC: 9.41 mg C3GE/g) correlate with antioxidant activity, preserved post-cooking.	**Hairy vetch** winter hardiness clusters well (Silhouette: 0.75), but nitrogen fixation traits are noisy (DBI: 0.77).
Common vetch (*Vicia sativa*) demonstrates moderate stability (CV: 15–35%), suitable for forage but variable under drought.	**Yield Prediction** was observed in Common bean with an accuracy (R^2: 0.81–0.86; RMSE: 290 kg/ha), while forage vetches (*V. sativa/villosa*) lag (R^2: 0.51–0.64) due to grazing variability.	**Common vetch** protein digestibility improves with heat treatment (AUC +85.55%), reducing anti-nutrients.	**Faba bean** phenotypic traits (e.g., flower color) cluster poorly (DBI: 1.12), highlighting unresolved genetic complexity.
Hairy vetch (*Vicia villosa*) is highly adaptable (Eberhart & Russell: 0.5–1.8) but erratic (CV: 25–50%), needing localized testing.	ML excels in disease resistance (common bean: 95% accuracy) and yield prediction (XGBoost RMSE: 290 kg/ha).		Hairy vetch's phenotypic variability and faba bean's data scarcity limit model accuracy.
Common bean (*Phaseolus vulgaris*) balances stability and yield (CV: 8–20%), excelling in tropical climates.			

Table 1: Analysis identifying species-specific strengths and challenges in breeding and agronomic management

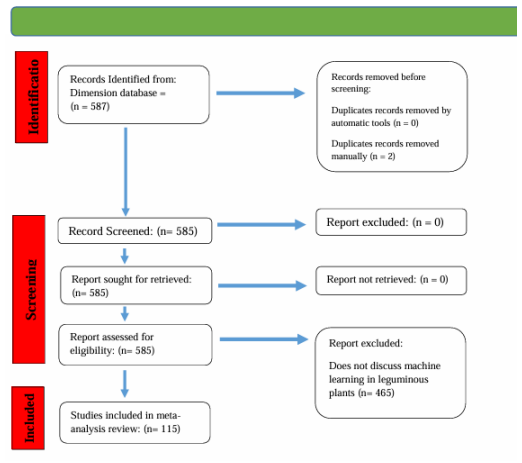

Figure 1: Search and selection process flow diagram. PRISMA 2020 official diagram (Page et al., 2021)

Conclusion

This meta-analysis demonstrates ML's potential to optimize legume breeding, by addressing species-specific challenges with **common bean** as the most model-ready species due to robust genomic resources and *Vicia villosa*'s undomesticated traits limiting accuracy. Making the later a benchmark for translational research**.** Faba bean and hairy vetch require targeted data collection and enhanced phenotyping, while **lentil and common vetch** benefit from stability-focused ML applications. ML-driven insights from stability metrics to nutraceutical AUC guide targeted breeding, though species-specific complexities (e.g., *V. villosa*'s plasticity) demand advanced algorithms like CNNs. The findings underscore the need for species-tailored approaches in precision agriculture. Combining ML and meta-analysis transforms plant breeding into a predictive science, uncovering hidden trends and accelerating crop improvement especially for neglected species like *Vicia*. Transfer learning bridges data gaps, as seen in predicting sorghum protein content and *Vicia faba* diversity using SCoT markers. Challenges include computational limits (e.g., *V. faba*'s 13Gb genome), inconsistent data annotation, and the need for explainable AI. Precision agriculture tools (UAVs, hyperspectral imaging) and advanced algorithms (XGBoost, CNNs) are critical for climate-resilient legume production. Future work should expand datasets for understudied species (*V. sativa*) and integrate multi-omics to bridge accuracy gaps as well as integrate multi-omics (e.g., FT-ICR MS for polyphenols), expanding datasets to improve abiotic stress prediction and integration of artificial general intelligence (AGI) could autonomously design breeding strategies, revolutionizing legume research.

Related Literature

Ipekesen, S., Rufaioğlu, S. B., Tunç, M., and Bicer, B. T. (2024). Machine Learning in Legume Breeding: Modelling Genotype and Environment Interactions. EJONS International Journal on Mathematic, Engineering and Natural 8 (4) 493-503. https://doi.org/10.5281/zenodo.14253789

Lee, H. (2023). Meta-Analysis and Machine Learning: Advancement of Analytic Methodology. Korean J Neurotrauma. 19(4):407-408. https://doi.org/10.13004/kjnt.2023.19.e63. pISSN 2234-8999·eISSN 2288-2243

Lima, E., Hyde, R., and Green, M. (2021). Model selection for inferential models with high dimensional data: synthesis and graphical representation of multiple techniques. *Sci Rep* **11**, 412. https://doi.org/10.1038/s41598-020-79317-8

Tahmasebi, A., Niazi, A., and Akrami, S. (2023). Integration of meta-analysis, machine learning and systems biology approach for investigating the transcriptomic response to drought stress in *Populus* species. *Sci Rep* **13**, 847. https://doi.org/10.1038/s41598-023-27746-6