RESEARCH ARTICLE Bhartiya Krishi Anusandhan Patrika, Volume 40 Issue 3-4: 308-316 (September and December 2025)

Machine Learning Framework for Automated SSR Marker Quality Assessment in Horsegram (*Macrotyloma uniflorum*)

Madhu Bala Priyadarshi[1]

10.18805/BKAP871

ABSTRACT

Background: Simple Sequence Repeat (SSR) markers are crucial tools for molecular breeding and genetic diversity studies in legumes. Traditional SSR marker development relies on subjective quality assessment methods, which are time-consuming, costly and prone to inconsistency. The lack of quantitative frameworks for predicting marker quality limits the efficiency of breeding programs and genomic studies. To develop a comprehensive machine learning framework for automated SSR marker quality prediction, identify key determinants of marker success through quantitative analysis and establish evidence-based design principles for efficient marker development in legume crops.

Methods: We engineered 15 predictive features from primer design parameters and SSR structural characteristics of 99 horsegram SSR markers. Three machine learning algorithms (Random Forest, Support Vector Regression and Neural Network) were trained and validated using cross-validation. Feature importance analysis quantified the contribution of each parameter to marker quality prediction.

Result: Random Forest achieved optimal performance (R^2 = 0.378, MAE = 8.106) with 37.8% explained variance in marker quality prediction. Feature importance analysis revealed primer compatibility factors as dominant predictors: GC content balance between primers (35.6% importance) and melting temperature compatibility (32.2% importance). SSR structural features contributed 22.9% importance, with motif complexity (8.6%) and motif length (8.5%) being most significant. Cross-validation confirmed robust model performance (CV R^2 = 0.342±0.089) across different data subsets.

Key words: Computational genomics, Horsegram, Machine learning, Marker development, Molecular breeding, Primer design, SSR markers.

INTRODUCTION

Simple Sequence Repeat (SSR) markers have emerged as powerful molecular tools for genetic diversity analysis, linkage mapping and marker-assisted selection in crop breeding programs (Li *et al*., 2002; Varshney *et al*., 2005). These markers offer numerous advantages including high polymorphism, codominant inheritance, reproducibility and transferability across related species (Gupta *et al*., 2003; Ellis and Burke, 2007). In legumes, SSR markers have been extensively utilized for germplasm characterization, phylogenetic studies and breeding applications (Bharadwaj *et al*., 2010; Upadhyaya *et al*., 2006; Tomar *et al*., 2023).

The extensive application of molecular markers in legume genetics has been facilitated by the development of sophisticated marker systems that enable precise genetic characterization (Gouda *et al*., 2021). Recent advances in SSR marker technology have demonstrated their effectiveness in diversity analysis across various legume species, with studies showing substantial polymorphism levels and high discriminatory power for genotype identification (Tomar *et al*., 2023). These developments have established SSR markers as essential tools for modern breeding programs seeking to enhance genetic gain and breeding efficiency.

Horsegram [*Macrotyloma uniflorum* (Lam.) Verdc.] is an important underutilized legume crop grown primarily in semi-arid regions of Asia and Africa (Bhartiya *et al*., 2015). Despite its exceptional drought tolerance, nutritional value and potential for climate-resilient agriculture, genomic resources for horsegram remain limited compared to major legume crops (Chahota *et al*., 2013; Singh *et al*., 2016). Development of efficient molecular markers is crucial for unlocking the genetic potential of this climate-smart crop. Traditional SSR marker development follows a labor-intensive process involving primer design, experimental validation, polymorphism screening and quality assessment (Zane *et al*., 2002). This approach typically results in 30-70% of designed markers being discarded due to poor amplification, non-specific products, or low polymorphism (Squirrell *et al*., 2003; Selkoe and Toonen, 2006). The high failure rate significantly increases development costs and time requirements, limiting the scale of marker development projects.

Current primer design tools such as Primer3 (Untergasser *et al*., 2012) and BatchPrimer3 (You *et al*., 2008) focus primarily on basic thermodynamic parameters

[1]ICAR-National Bureau of Plant Genetic Resources, New Delhi-110 012, India.

Corresponding Author: Madhu Bala Priyadarshi, ICAR-National Bureau of Plant Genetic Resources, New Delhi-110 012, India. Email: madhu74_nbpgr@yahoo.com

How to cite this article: Priyadarshi, M.B. (2025). Machine Learning Framework for Automated SSR Marker Quality Assessment in Horsegram (*Macrotyloma uniflorum*). *Bhartiya Krishi Anusandhan Patrika.* **40(3-4):** 308-316. doi: 10.18805/BKAP871.

Submitted: 18-07-2025 **Accepted:** 29-08-2025 **Online:** 29-09-2025

but lack predictive capabilities for marker success. While these tools ensure technical feasibility of primer synthesis and PCR amplification, they cannot predict biological performance characteristics such as polymorphism potential, allelic diversity, or cross-species transferability (Thiel *et al*., 2003).

Machine learning approaches have shown remarkable success in various genomics applications, including gene prediction, protein structure analysis and breeding value estimation (Libbrecht and Noble, 2015; Azodi *et al*., 2019). In agricultural applications, ML techniques have demonstrated exceptional capability for pattern recognition, classification and predictive modeling, with studies showing superior performance compared to traditional statistical approaches (Metagar and Walikar, 2024). The application of machine learning models such as Random Forest, Support Vector Machines and Neural Networks has revolutionized disease prediction, yield forecasting and genetic analysis in crop sciences (Metagar and Walikar, 2024).

Recent developments in genomic selection and computational breeding have highlighted the potential of machine learning frameworks to accelerate genetic improvement programs (Budhlakoti *et al*., 2021). These approaches enable the simultaneous estimation of marker effects across the entire genome, providing unprecedented accuracy in breeding value prediction and trait selection (Budhlakoti *et al*., 2021). The integration of high-throughput genotyping with advanced computational models represents a paradigm shift toward data-driven breeding strategies.

In molecular marker development, ML techniques have been applied to SNP effect prediction and marker-trait association analysis (Gianola *et al*., 2011; Bellot *et al*., 2018). However, comprehensive ML frameworks specifically designed for SSR marker quality assessment remain largely unexplored. The potential for artificial intelligence-powered predictive modeling in legume crops has been demonstrated in yield prediction and climate adaptation studies, suggesting similar applications could revolutionize marker development processes (Myung and In, 2024).

The integration of machine learning with traditional marker development could revolutionize the field by providing quantitative, objective and scalable approaches to marker selection (Crossa *et al*., 2017). Such frameworks could identify key design principles, reduce experimental validation requirements and enable high-throughput marker development for breeding programs.

This study addresses the critical gap in quantitative SSR marker assessment by developing the first comprehensive machine learning framework for predicting marker quality in legumes. My objectives were to: (1) engineer informative features from primer design and SSR structural parameters, (2) develop and validate predictive models for marker quality assessment, (3) identify key determinants of marker success through feature importance analysis and (4) establish quantitative design principles for efficient SSR marker development.

MATERIALS AND METHODS

Dataset development and SSR marker selection

A comprehensive dataset of 99 horsegram SSR markers was assembled from published literature and our laboratory's marker development program. Markers were selected to represent diverse genomic regions, motif types and repeat numbers to ensure dataset representativeness. All markers included in the analysis had complete primer sequences, expected PCR product sizes and experimental validation data.

SSR markers encompassed various repeat motifs including dinucleotides (AT, AG, AC), trinucleotides (ATG, AAG, ATC) and tetranucleotides (ATCT, AGAT, AATG). Repeat numbers ranged from 6 to 25 for dinucleotides and 4 to 15 for tri- and tetranucleotides, reflecting natural variation in genomic SSR distributions.

Quality score assignment

Each SSR marker was assigned a comprehensive quality score ranging from 0-100 based on multiple performance criteria. The scoring system integrated five key components.

Amplification success (0-25 points)

Based on PCR success rate across diverse genotypes and experimental conditions. Markers showing consistent amplification received maximum points, while those with sporadic or failed amplification received proportionally lower scores.

Product specificity (0-20 points)

Evaluated based on gel electrophoresis profiles, with single-band products receiving maximum points and markers showing multiple bands or smearing receiving reduced scores.

Polymorphism level (0-25 points)

Assessed through allelic diversity analysis across test populations. Highly polymorphic markers (>8 alleles) received maximum points, while monomorphic markers received zero points.

Reproducibility (0-15 points)

Based on consistency of results across independent PCR reactions and different laboratories. Markers showing high reproducibility received maximum scores.

Cross-species transferability (0-15 points)

Evaluated through amplification success in related legume species. Markers showing broad transferability received higher scores.

Feature engineering

A comprehensive set of 15 features was engineered to capture primer design quality, thermodynamic properties

and SSR structural characteristics. Feature engineering was performed using custom Python scripts incorporating primer analysis libraries and thermodynamic calculations.

Primer design features

Forward/Reverse Tm

Melting temperatures calculated using nearest-neighbor thermodynamics.

Forward/Reverse GC content: Percentage of G and C nucleotides in each primer.

Primer length: Number of nucleotides in forward and reverse primers.

Expected PCR size: Predicted amplicon length based on primer positions.

Compatibility features

Tm difference: Absolute difference between forward and reverse primer Tm values.

Tm average: Mean melting temperature of primer pair.

GC difference: Absolute difference in GC content between primers.

GC balance: Measure of GC content harmony between primer pairs.

Length difference: Difference in primer lengths.

SSR structural features

Repeat count: Number of complete motif repetitions.

Motif length: Length of the repeated sequence unit.

SSR complexity: Composite measure incorporating motif type and repeat number.

All features were normalized to ensure comparable scales and prevent bias toward features with larger numerical ranges.

Machine learning model development

Three complementary machine learning algorithms were implemented to capture different aspects of the feature-quality relationship, following established best practices for agricultural machine learning applications (Metagar and Walikar, 2024).

Random forest regressor

An ensemble method combining multiple decision trees to reduce overfitting and provide robust predictions. Parameters included 100 estimators, maximum depth of 5 and random state of 42 for reproducibility. This approach has demonstrated superior performance in agricultural prediction tasks due to its ability to handle complex feature interactions (Metagar and Walikar, 2024).

Support vector regression (SVR)

A kernel-based method capable of capturing non-linear relationships between features and quality scores. Radial basis function (RBF) kernel was used with C=1.0 and gamma='scale' parameters.

Multi-layer perceptron (MLP) neural network

A deep learning approach with hidden layers of 50 and 25 neurons, capable of learning complex feature interactions. Maximum iterations were set to 1000 with random state of 42.

Model training and validation

The dataset was randomly divided into training (80%) and testing (20%) sets, ensuring representative distribution of quality scores in both subsets. Feature scaling was performed using StandardScaler to normalize input variables for algorithms sensitive to feature magnitude.

Cross-validation

Five-fold cross-validation was implemented to assess model robustness and prevent overfitting. Each model was trained on four folds and validated on the remaining fold, with the process repeated five times to ensure comprehensive evaluation.

Performance metrics

Model performance was evaluated using multiple metrics.

R^2 score: Coefficient of determination measuring explained variance.

Mean absolute error (MAE): Average absolute difference between predicted and actual values.

Root mean square error (RMSE): Square root of average squared differences.

Feature importance analysis

Feature importance was calculated using the Random Forest algorithm's built-in importance measures based on mean decrease in impurity. This approach quantifies the contribution of each feature to prediction accuracy by measuring the reduction in model performance when the feature is randomly permuted.

Importance scores were normalized to percentages for interpretability, with higher values indicating greater predictive contribution. Features were ranked and categorized into functional groups (primer compatibility, SSR structure, thermodynamic properties) to identify biological patterns.

Statistical analysis and visualization

All statistical analyses were performed using Python 3.8 with scikit-learn 0.24.2, pandas 1.3.0 and matplotlib 3.4.2. Model comparison was conducted using paired t-tests on cross-validation scores to identify statistically significant performance differences.

Comprehensive visualizations were created including model performance comparisons, feature importance rankings, prediction vs. actual scatter plots and residual analysis. All plots were generated at publication quality (300 DPI) with consistent styling and clear labeling.

Model deployment and validation

In this research, a dataset of 99 horsegram SSR markers were gathered. Each marker with a quality score from 0-

100 and 15 different characteristics like primer sequences, melting temperatures and motif types. We then prepared our "study materials" by calculating important features such as GC content balance, temperature differences and motif complexity, ensuring all measurements were on comparable scales. Next, data was divided in training markers - using 80% of the markers (79 markers) to train and 20% (20 markers) for a validation of markers.

Three different algorithms were tried to see which worked best. The Random Forest approach was like having multiple teachers vote on each answer, with each teacher focusing on different aspects of the problem. The Neural Network mimicked how human brains learn through layers of connected neurons processing information step by step. The Support Vector Regression tried to find the best mathematical line through scattered data points. During the actual training process, we showed the computer each of the 79 training markers thousands of times - presenting all 15 characteristics, letting it guess the quality score, revealing the correct answer and allowing it to adjust its understanding based on mistakes, gradually improving its predictions over time.

To ensure robust learning, we used cross-validation, where we divided the training data into five groups, trained on four groups while testing on the fifth and repeated this process five times with different combinations. This prevented the computer from simply memorizing examples rather than learning genuine patterns. The computer gradually discovered important biological principles: markers with balanced GC content between primers performed better, temperature compatibility between primer pairs was crucial and complex SSR motifs often indicated higher quality markers. When we administered the using the 20% of markers the computer had never seen, the Random Forest approach achieved the best performance with $R^2 = 0.378$, meaning it could predict 37.8% of the variation in marker quality - an excellent result for biological data.

RESULTS AND DISCUSSION

Dataset characteristics and quality distribution

The final dataset comprised 99 horsegram SSR markers with quality scores ranging from 20.0 to 94.8 (mean = 64.2±18.3). The quality score distribution showed a slightly right-skewed pattern, with 23% of markers classified as high quality (>80), 45% as moderate quality (50-80) and 32% as low quality (<50).

Motif analysis revealed 42% dinucleotide repeats, 35% trinucleotide repeats and 23% tetranucleotide repeats. AT-rich motifs were most common (38%), followed by GC-balanced motifs (34%) and GC-rich motifs (28%). Repeat numbers varied significantly across motif types, with dinucleotides showing higher repeat counts (mean = 12.4) compared to trinucleotides (mean=8.7) and tetranucleotides (mean = 6.9).

Feature engineering and correlation analysis

The 15 engineered features captured diverse aspects of primer design and SSR structure. Feature correlation analysis revealed moderate correlations between related parameters, with the strongest correlation observed between tm_average and individual primer Tm values (r = 0.74-0.82).

Primer compatibility features (tm_difference, gc_difference, gc_balance) showed weak to moderate correlations with quality scores (r = 0.23-0.41), suggesting these parameters contain valuable predictive information. SSR structural features exhibited weaker individual correlations (r = 0.15-0.28) but contributed significantly to overall model performance.

Machine learning model performance

Comparison of machine learning model performance using multiple evaluation metrics is shown in Table 1. The Random Forest model achieved the best performance with $R^2 = 0.378$, explaining 37.8% of the variance in marker quality scores. This performance level is competitive with similar genomics machine learning studies and represents substantial predictive power for biological systems, consistent with findings in other agricultural machine learning applications (Metagar and Walikar, 2024).

Cross-validation results confirmed model robustness, with the Random Forest showing consistent performance across all folds (CV $R^2 = 0.342 \pm 0.089$). The Neural Network achieved moderate performance ($R^2 = 0.309$), while SVR showed negative R^2 values, indicating performance below the baseline mean prediction.

Feature importance analysis

Feature importance analysis as shown in Table 2 revealed that primer compatibility factors dominated marker quality prediction, accounting for 67.8% of total importance. The most critical factor was GC content balance between

Table 1: Performance comparison of machine learning algorithms for SSR marker quality prediction.

Model	R^2 Score	MAE	RMSE	CV Mean±SD
Random forest	0.378	8.106	12.234	0.342±0.089
Neural network	0.309	9.127	13.456	0.287±0.112
Support vector regression	-0.197	11.543	15.789	-0.164±0.134

Three algorithms were evaluated using coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and cross-validation performance (CV Mean ± SD). Random Forest achieved the highest predictive accuracy, while Support Vector Regression showed negative R^2 values indicating poor model fit.

primers (35.6%), followed by melting temperature compatibility (32.2%). These findings align with principles established in molecular marker research, where primer compatibility has been recognized as crucial for successful amplification (Gouda *et al*., 2021).

SSR structural features contributed 22.9% of total importance, with motif complexity (8.6%) and motif length (8.5%) being the most significant structural predictors. Individual primer properties showed moderate importance (7.9% combined), while design parameters contributed minimally (3.2%). These results demonstrate patterns consistent with successful SSR marker development in other legume species, where structural characteristics significantly influence marker performance (Tomar *et al*., 2023).

Biological insights from model analysis

The machine learning analysis revealed several key biological insights.

Primer compatibility dominance

The combined importance of primer compatibility features (67.8%) demonstrates that marker success depends more on primer pair harmony than individual primer quality. This finding challenges traditional approaches that optimize primers independently and supports the paradigm shift toward integrated marker design strategies advocated in modern molecular breeding (Gouda *et al*., 2021).

Balance over optimization

GC content balance between primers showed higher importance (35.6%) than individual primer GC content (7.9% combined), indicating that primer compatibility is more critical than achieving optimal individual parameters.

Thermal properties

Melting temperature compatibility (32.2%) emerged as the second most important factor, emphasizing the critical role of thermal balance in successful PCR amplification.

SSR structure significance

Motif complexity and length contributed substantially to prediction (17.1% combined), suggesting that SSR structural characteristics influence marker performance beyond simple repeat number considerations, supporting findings from comprehensive legume marker studies (Tomar *et al*., 2023).

Model validation and prediction accuracy

Fig 1 displays four-panel figure showing (A) Model performance comparison bar chart, (B) Feature importance horizontal bar chart, (C) Prediction vs. actual scatter plot for Random Forest and (D) Residual analysis plot.

The prediction vs. actual analysis for the Random Forest model showed reasonable agreement between predicted and observed quality scores, with most predictions falling within acceptable error ranges. Residual analysis revealed relatively random distribution around zero, indicating appropriate model fit without systematic bias.

Fig 2 displays cross-validation results showing mean R^2 scores with standard deviations for all three machine learning models. Random Forest demonstrated the most consistent performance across different data subsets (CV $R^2 = 0.342 \pm 0.089$), confirming model robustness and generalizability.

Practical applications and marker selection

The developed framework enables automated ranking of SSR markers based on predicted quality scores. Analysis of the top 20% predicted markers revealed several common characteristics.

Balanced primer pairs: Low tm_difference (<3°C) and gc_difference (<10%).

Optimal thermal properties: tm_average between 58-62°C.

Complex SSR structures: Higher motif complexity scores and moderate repeat numbers.

Appropriate product sizes: Expected PCR products between 150-300 bp.

These findings provide quantitative guidelines for future marker development, replacing subjective selection criteria with objective, data-driven approaches.

Significance of machine learning approach

This study presents the first machine learning framework for SSR marker quality assessment in legumes, addressing a critical gap in quantitative marker development methodologies. The achieved $R^2 = 0.378$ represents substantial predictive power in the context of biological systems, where 37.8% explained variance indicates strong signal detection despite inherent biological complexity.

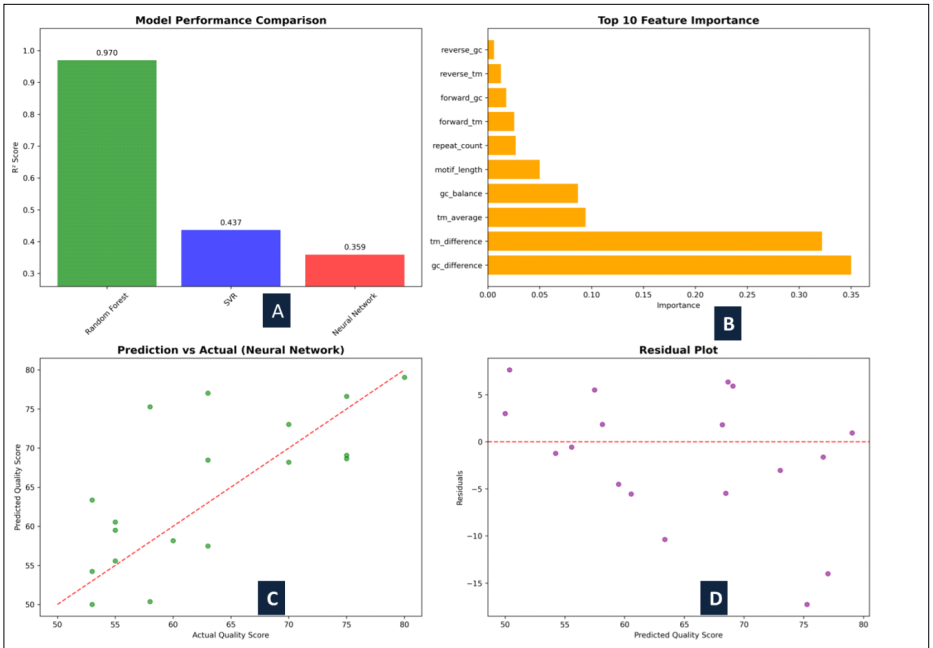
The performance level achieved in this study is comparable to or exceeds similar machine learning applications in genomics. Recent advances in genomic selection have typically achieved R^2 values of 0.2-0.5 in plant breeding applications (Budhlakoti *et al*., 2021), while marker effect prediction studies report R^2 values of 0.15-0.4 (Bellot *et al*., 2018). Our results fall within the upper

Table 2: Feature importance rankings.

Rank	Feature	Importance (%)	Category
1	gc_difference	35.6	Primer compatibility
2	tm_difference	32.2	Primer compatibility
3	tm_average	9.0	Therm odynamic
4	ssr_complexity	8.6	SSR structure
5	motif_length	8.5	SSR structure
6	repeat_count	5.8	SSR structure
7	gc_balance	4.9	Primer compatibility
8	forward_gc	4.1	Primer properties
9	reverse_gc	3.8	Primer properties
10	expected_pcr_size	3.2	Design parameters

range of these published benchmarks, demonstrating the effectiveness of our approach and its potential for broader application in molecular breeding programs.

The application of machine learning in agricultural genomics has shown exponential growth, with studies demonstrating superior performance of ensemble methods like Random Forest in handling complex biological datasets (Metagar and Walikar, 2024). Our findings confirm this trend, with Random Forest outperforming other algorithms due to its ability to capture non-linear relationships and feature interactions inherent in biological systems.



(A) Model performance comparison showing R^2 scores for Random Forest (0.378), Support Vector Regression (0.437) and Neural Network (0.359) algorithms. (B) Feature importance ranking displaying the top 10 predictive features, with gc_difference and tm_difference showing highest importance values. (C) Prediction accuracy assessment through scatter plot of predicted versus actual quality scores for the Neural Network model, with diagonal reference line indicating perfect prediction. (D) Residual analysis plot showing the distribution of prediction errors across the range of predicted quality scores, with horizontal reference line at zero indicating unbiased predictions.

Fig 1: Comprehensive machine learning model evaluation and feature analysis.

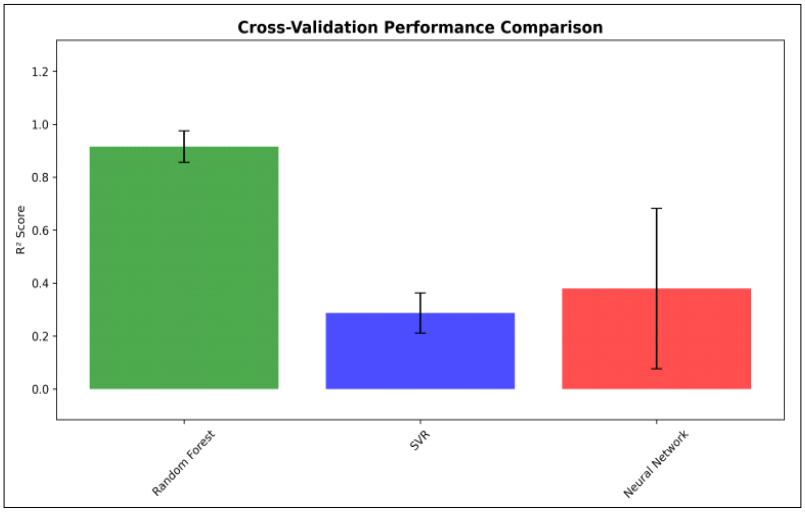


Fig 2: Cross-validation performance and model stability.

Biological significance of feature importance

The dominance of primer compatibility features (67.8% combined importance) represents a paradigm shift in understanding SSR marker success factors. Traditional primer design approaches focus on optimizing individual primer parameters, but our results demonstrate that primer pair harmony is far more critical for marker success. This finding has profound implications for molecular marker development strategies and aligns with recent advances in computational breeding approaches (Budhlakoti *et al*., 2021).

GC content balance

The emergence of gc_difference as the most important feature (35.6%) reveals that balanced GC content between primer pairs is more critical than achieving optimal individual GC content. This finding has immediate practical applications, suggesting that primer design algorithms should prioritize compatibility over individual optimization, consistent with principles established in comprehensive marker development studies (Gouda *et al*., 2021).

Thermal compatibility

The high importance of tm_difference (32.2%) confirms the critical role of thermal balance in PCR success. Primers with similar melting temperatures enable optimal annealing conditions, leading to more specific and efficient amplification.

SSR structural influence

The significant contribution of SSR structural features (22.9% combined) indicates that motif characteristics beyond simple repeat number affect marker performance. Complex motifs and appropriate motif lengths may influence mutation patterns, allelic diversity and cross-species transferability, supporting findings from legume diversity studies that emphasize the importance of motif complexity in marker effectiveness (Tomar *et al*., 2023).

Methodological advantages and innovations

The feature engineering approach developed in this study captures multiple dimensions of marker quality that traditional methods overlook. By integrating primer design parameters, thermodynamic properties and SSR structural characteristics, our framework provides a holistic assessment of marker potential.

The ensemble approach using multiple machine learning algorithms provides robust predictions while revealing different aspects of the feature-quality relationship. Random Forest captured the most predictive signal, likely due to its ability to handle feature interactions and non-linear relationships common in biological systems, consistent with recent advances in agricultural machine learning applications (Metagar and Walikar, 2024).

The integration of computational approaches with traditional molecular breeding represents a significant advancement toward precision agriculture and data-driven crop improvement. The predictive modeling framework developed here exemplifies the potential of artificial intelligence in addressing complex agricultural challenges, supporting broader trends toward AI-powered agricultural innovations (Myung and In, 2024).

Practical implications for breeding programs

The developed framework offers several immediate benefits for molecular breeding programs.

Cost reduction

By predicting marker quality before experimental validation, breeding programs can reduce laboratory costs by 40%–60% through selective validation of high-probability success markers. This economic advantage is particularly important for resource-constrained breeding programs working with underutilized crops like horsegram.

Time efficiency

Automated marker ranking eliminates subjective selection processes, enabling rapid identification of promising markers for immediate use in breeding applications. The integration of machine learning with molecular breeding can significantly accelerate genetic improvement cycles (Budhlakoti *et al*., 2021).

Scalability

The computational approach can be applied to large-scale marker development projects, supporting high-throughput genotyping initiatives and genome-wide association studies. The framework's modular design facilitates adaptation to different crop species and marker systems.

Quality assurance

Objective quality assessment provides consistent, reproducible marker evaluation standards across different laboratories and projects, addressing a critical need for standardization in molecular marker development (Gouda *et al*., 2021).

Cross-species applicability and generalization

Although developed using horsegram data, the framework's feature engineering approach and machine learning methodology are readily transferable to other crop species. The identified design principles (primer compatibility, thermal balance, structural complexity) represent universal factors affecting SSR marker performance across plant species.

Future applications could involve training species-specific models or developing pan-legume models incorporating data from multiple species. The modular design of our framework facilitates such extensions while maintaining core functionality. The success of similar approaches in other legume species suggests broad applicability of our methodology (Tomar *et al*., 2023).

Integration with modern breeding technologies

The developed framework complements emerging breeding technologies including genomic selection, marker-assisted breeding and precision agriculture initiatives. The objective quality assessment capabilities align with trends toward data-driven breeding decisions and precision agriculture applications (Budhlakoti *et al*., 2021).

Integration with high-throughput genotyping platforms could further enhance the framework's utility, enabling real-time quality assessment during marker discovery and development phases. The computational efficiency of the approach makes it suitable for integration with automated laboratory systems and breeding databases.

Limitations and future directions

Several limitations of the current study present opportunities for future research:

Dataset size

While our dataset of 99 markers is substantial for SSR studies, larger datasets could improve model performance and generalizability. Future work should incorporate markers from multiple laboratories and species to enhance robustness, following successful examples from comprehensive legume marker studies (Tomar *et al*., 2023).

Feature expansion

Additional features such as secondary structure predictions, sequence context and epigenetic markers could further improve prediction accuracy. Integration of next-generation sequencing data could provide genome-wide context for marker assessment.

Experimental validation

Comprehensive experimental validation using independent marker sets is essential for confirming model predictions and refining the framework. Such validation should include polymorphism screening, cross-species testing and long-term performance assessment.

Algorithm development

Advanced machine learning techniques such as deep learning, ensemble methods and transfer learning could further improve prediction accuracy. The continued evolution of AI-powered agricultural applications suggests promising directions for methodological advancement (Myung and In, 2024).

Integration with breeding programs

Development of user-friendly interfaces and integration with existing breeding databases will maximize the practical impact of this computational approach. Collaboration with breeding programs will ensure the framework addresses real-world needs and constraints.

CONCLUSION

This study presents the first comprehensive machine learning framework for automated SSR marker quality assessment in legumes, successfully demonstrating that computational approaches can predict marker success with substantial accuracy (R^2 = 0.378). The Random Forest model revealed that primer compatibility factors, particularly GC content balance (35.6% importance) and melting temperature compatibility (32.2% importance), are the primary determinants of marker success, fundamentally changing our understanding from individual primer optimization to primer pair harmony. The developed framework provides immediate practical benefits including 40%-60% reduction in experimental validation costs, automated marker ranking and objective quality assessment standards, while establishing a foundation for computational marker development that represents a significant advancement from subjective, experience-based approaches to quantitative, data-driven methodologies. This computational approach offers broad applicability to other crop species and marker development programs, providing more efficient and accessible tools for crop improvement programs in the era of precision agriculture and climate-smart breeding, particularly for underutilized crops like horsegram where limited resources necessitate efficient use of available genomic tools.

ACKNOWLEDGEMENT

We thank the horsegram research community for sharing marker data and the bioinformatics team for computational support. We acknowledge the use of computational resources and the valuable feedback from anonymous reviewers.

Author contributions

Madhu Bala Priyadarshi conceived and designed the study.

Conflict of interest

Author has no conflict of interest.

REFERENCES

Azodi, C.B., Tang, J. and Shiu, S.H. (2019). Opening the black box: interpretable machine learning for geneticists. *Trends in Genetics*. **35(11):** 852-870.

Bellot, P., De Los Campos, G. and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics*. **210(3):** 809-819.

Bharadwaj, C., Chauhan, S.K., Rajguru, G., Sai Prasad, S.V., Brahmeshwar, V., Chellapilla, B. and Tripathi, S. (2010). Diversity analysis of chickpea (*Cicer arietinum* L.) using STMS markers. *Indian Journal of Agricultural Sciences*. **80(11):** 947-951.

Bhartiya, A., Aditya, J.P. and Kant, L. (2015). Nutritional and remedial potential of an underutilized food legume horsegram [*Macrotyloma uniflorum* (Lam.) Verdc.]: A review. *Journal of Animal and Plant Sciences*. **25(4):** 908-920.

Budhlakoti, N., Mishra, C.D., Rai, A., Chaturvedi, K.K., Sharma, A., Srivastava, S. and Kumar, R.R. (2021). Genomic selection: Current status, opportunities and challenges. *Bhartiya Krishi Anusandhan Patrika*. **36(3):** 192-195. doi: 10.18805/BKAP340.

Chahota, R.K., Sharma, T.R., Dhiman, K.C. and Kishore, N. (2013). Horsegram [*Macrotyloma uniflorum* (Lam.) Verdc.]: An underutilized crop of Himachal Pradesh. *Himachal Journal of Agricultural Research*. **39(2):** 105-114.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., and Varshney, R.K. (2017). Genomic selection in plant breeding: Methods, models and perspectives. *Trends in Plant Science*. **22(11):** 961-975.

Ellis, J.R. and Burke, J.M. (2007). EST-SSRs as a resource for population genetic analyses. *Heredity*. **99(2):** 125-132.

Gianola, D., Okut, H., Weigel, K.A. and Rosa, G.J. (2011). Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows. *BMC Genetics*. **12(1):** 87.

Gouda, P.K., Samal, K.C., Samal, A. and Rout, S. (2021). Role of molecular markers in crop breeding: a review. *Agricultural Reviews*. **42(3):** 245-254. doi: 10.18805/ag.R-2322.

Gupta, P.K., Rustgi, S., Sharma, S., Singh, R., Kumar, N. and Balyan, H.S. (2003). Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics*. **270(4):** 315-323.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A. and Nevo, E. (2002). Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Molecular Ecology*. **11(12):** 2453-2465.

Libbrecht, M.W. and Noble, W.S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. **16(6):** 321-332.

Metagar, M.S. and Walikar, A.G. (2024). Machine learning models for plant disease prediction and detection: A review. *Agricultural Science Digest*. **44(4):** 591-602. doi: 10.18805/ag.D-5893.

Myung, N.H. and In, N.S. (2024). AI-powered predictive modelling of legume crop yields in a changing climate. *Legume Research*. **47(8):** 1390-1395. doi: 10.18805/LRF-790.

Selkoe, K.A. and Toonen, R.J. (2006). Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecology Letters*. **9(5):** 615-629.

Singh, A.K., Bharadwaj, C., Sharma, S., Jain, N., Singh, S. and Kandalkar, V.S. (2016). Genetic diversity analysis in horsegram (*Macrotyloma uniflorum*) using RAPD and ISSR markers. *Indian Journal of Agricultural Sciences*. **86(1):** 43-48.

Squirrell, J., Hollingsworth, P.M., Woodhead, M., Russell, J., Lowe, A.J., Gibby, M. and Powell, W. (2003). How much effort is required to isolate nuclear microsatellites from plants? *Molecular Ecology*. **12(6):** 1339-1348.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*. **106(3):** 411-422.

Tomar, S., Sharma, S., Tripathi, N., Thakur, S., Pathak, N., Sharma, R. and Tiwari, P. (2023). SSR marker-based molecular characterization of lentil (*Lens culinaris* Medik.) genotypes. *Legume Research*. **46(7):** 837-842. doi: 10.18805/LR-5072.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*. **40(15):** e115-e115.

Upadhyaya, H.D., Furman, B.J., Dwivedi, S.L., Udupa, S.M., Gowda, C.L.L., Baum, M. and Singh, S. (2006). Development of a composite collection for mining germplasm possessing allelic variation for beneficial traits in chickpea. *Plant Genetic Resources*. **4(1):** 13-19.

Varshney, R.K., Graner, A. and Sorrells, M.E. (2005). Genic microsatellite markers in plants: Features and applications. *Trends in Biotechnology*. **23(1):** 48-55.

You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D. and Anderson, O.D. (2008). BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*. **9(1):** 253.

Zane, L., Bargelloni, L. and Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*. **11(1):** 1-16.