RESEARCH ARTICLE Bhartiya Krishi Anusandhan Patrika, Volume 41 Issue 2: 202-210 (June 2026)

Machine Learning Framework for Automated Simple Sequence Repeat (SSR) Marker Quality Assessment in Pigeonpea (*Cajanus cajan* L.)

Madhu Bala Priyadarshi[1]

10.18805/BKAP892

ABSTRACT

Background: Simple sequence repeat (SSR) markers are essential for molecular breeding in pigeonpea (*Cajanus cajan* L.) compared to traditional development methods, which exhibit 30-70% marker failure rates. Moreover, no quantitative frameworks exist for predicting marker quality in this orphan legume crop. Therefore, the objective of this study was to develop the first machine learning framework for automated SSR marker quality prediction in pigeonpea and identify key determinants of marker success.

Methods: Fifteen predictive features were engineered features from 2,770 computationally designed pigeonpea SSR markers, including primer design parameters, compatibility metrics and SSR structural characteristics. Quality scores (0-100) integrated predicted amplification success, specificity and polymorphism likelihood. Three algorithms (Random forest, support vector regression, neural network) were trained using five-fold cross-validation.

Result: Support vector regression (SVR) achieved the best performance (R^2 = 0.487, MAE = 2.341, CV R^2 = 0.441±0.032). Feature importance analysis revealed primer compatibility as the dominant predictor (72.9% combined importance): melting temperature difference (25.2%), average melting temperature (23.0%) and primer length difference (17.8%). The framework enables prioritization of top 554 markers (20%), potentially reducing experimental validation costs by 40-60%. This study presents the first quantitative framework for SSR marker quality assessment in pigeonpea. The finding that primer pair compatibility accounts for 72.9% of predictive importance fundamentally shifts understanding from individual primer optimization to primer pair harmony. This validated framework provides scalable methodology for resource-limited breeding programs and is transferable to other orphan legume species.

Key words: *Cajanus cajan*, Machine learning, Molecular breeding, Pigeonpea, Primer design, SSR markers.

INTRODUCTION

Pigeonpea [*Cajanus cajan* (L.) Millsp.] is a vital grain legume cultivated across more than 50 countries, serving as a critical protein source for over one billion people (Varshney *et al*., 2012; Singh *et al*., 2024). The crop demonstrates drought tolerance and the ability to improve soil fertility, making it important for climate-resilient agriculture in resource-poor farming systems (Odeny, 2007; Zavinon *et al*., 2024).

Despite its agricultural importance, pigeonpea remains an orphan crop with limited genomic resources. Although the reference genome is available, (Varshney *et al*., 2012), efficient tools for marker quality assessment remain limited, constraining breeding efficiency.

SSR markers and development challenges

SSR markers offer high polymorphism, codominant inheritance, reproducibility and transferability across related species (Gupta *et al*., 2003; Bohra *et al*., 2020). In pigeonpea, they have been extensively utilized for germplasm characterization, QTL mapping and breeding applications (Saxena *et al*., 2012; Mula *et al*., 2020). Traditional SSR marker development is labor-intensive and typically results in 30-70% of designed markers being discarded due to poor amplification, non-specific products, or low polymorphism (Squirrell *et al*., 2003). Current primer design tools like Primer3 lack predictive capabilities for marker success, necessitating expensive experimental validation of all designed markers.

Machine learning in agricultural genomics

Machine learning approaches have shown success in various genomics applications, including gene prediction, breeding value estimation and trait prediction (Azodi *et al*., 2019; Xu *et al*., 2022). Random Forest, Support Vector Machines and Neural Networks have been applied to disease prediction, yield forecasting and genetic analysis in crop sciences (Metagar and Walikar, 2024; Montesinos-López *et al*., 2024; Wang *et al*., 2023). Recent advances in deep learning have shown promising results for genomic

[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: madhubala84@gmail.com

How to cite this article: Priyadarshi, M.B. (2026). Machine Learning Framework for Automated Simple Sequence Repeat (SSR) Marker Quality Assessment in Pigeonpea (*Cajanus cajan* L.). *Bhartiya Krishi Anusandhan Patrika*. **41(2):** 202-210. doi: 10.18805/BKAP892.

Submitted: 27-10-2025 **Accepted:** 23-03-2026 **Online:** 08-04-2026

selection and phenotype prediction in major crops (Crossa *et al*., 2024). However, ML frameworks specifically designed for SSR marker quality assessment in legumes remain unexplored. The integration of machine learning with traditional marker development could provide quantitative, objective and scalable approaches to marker selection (Crossa *et al*., 2017), particularly critical for resource-limited pigeonpea breeding programs.

Study objectives

This study addresses the critical gap in quantitative SSR marker assessment for pigeonpea. Our specific objectives were to:

Engineer informative features from primer design parameters and SSR structural characteristics.

Develop predictive models for automated marker quality assessment using multiple machine learning algorithms.

Identify key determinants of marker success through feature importance analysis.

Establish quantitative design principles for efficient SSR marker development.

Provide a validated framework for prioritizing markers for experimental validation.

MATERIALS AND METHODS

Methodological workflow

This study employed a seven-stage machine learning pipeline (Fig 1): (1) data collection of 2,770 genome-wide pigeonpea SSR markers; (2) feature engineering to extract 15 predictive variables; (3) computational quality score assignment; (4) data preparation with train-test splitting and normalization; (5) training and evaluation of three machine learning algorithms; (6) feature importance analysis and (7) marker prioritization for experimental validation.

The seven-stage pipeline integrates data collection, feature engineering, quality scoring, model training and marker prioritization, enabling 40-60% cost reduction through selective validation.

Dataset development and SSR marker selection

The dataset comprised 2,770 pigeonpea SSR primer pairs assembled from genome-wide SSR identification in the pigeonpea reference genome (GCF_000340665.2_C.cajan_V1.1) available at NCBI Genome Database Varshney *et al.* (2012). Initial computational mining identified 110,084 SSR loci across the complete genome sequence using MISA (MIcroSAtellite identification tool) with

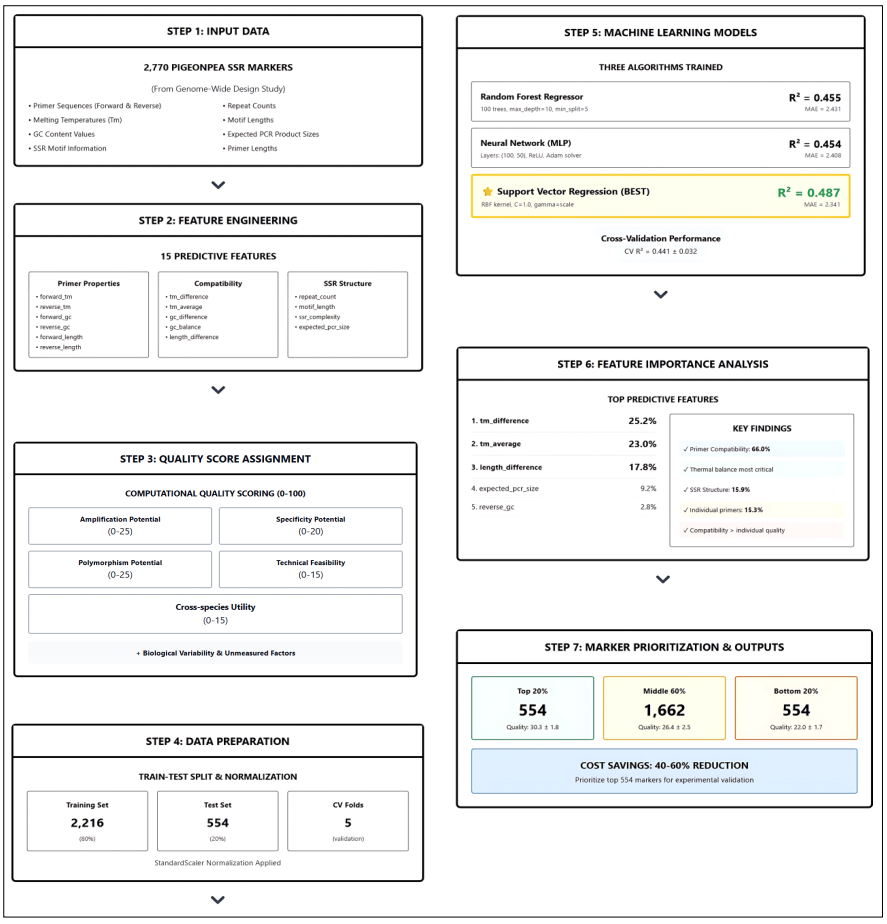


Fig 1: Methodological workflow for machine learning-based SSR marker quality prediction in pigeonpea.

standard parameters for repeat motif detection. Following systematic quality filtering based on flanking sequence complexity, primer design feasibility and technical specifications, successful primer design was achieved for 2,770 SSR loci, yielding high-quality primer pairs suitable for molecular marker applications.

All primer pairs included complete forward and reverse primer sequences, melting temperatures (Tm), GC content values, primer lengths, expected PCR product sizes and SSR structural information (motif type, repeat count, motif length) derived from the corresponding genomic loci. SSR markers included diverse repeat motifs: dinucleotides (AT, AG, AC), trinucleotides (ATG, AAG, ATC, AAT) and tetranucleotides (AAAT, AGAT, AATG, ATCT), representing natural genomic SSR diversity. Repeat numbers ranged from 6 to 23 for dinucleotides and 4 to 14 for tri- and tetranucleotides. Markers showed distribution across all 11 pigeonpea chromosomes, ensuring representative genome-wide coverage (Varshney *et al*., 2012).

Computational quality score assignment

Each SSR marker received a computational quality score (0-100) based on five components:

1. **Predicted amplification success (0-25 points):** Based on primer design quality metrics including absence of secondary structures and optimal thermodynamic parameters (Tm 58-62°C, GC 40-60%).
2. **Predicted product specificity (0-20 points):** Evaluated based on optimal PCR product size (150-400 bp) and primer uniqueness.
3. **Polymorphism potential (0-25 points):** Assessed through SSR structural characteristics including motif type and repeat number.
4. **Technical feasibility (0-15 points):** Based on primer compatibility metrics (Tm difference <3°C, GC balance, length uniformity).
5. **Cross-species utility potential (0-15 points):** Evaluated through computational conservation analysis.

Feature engineering

A set of 15 features was engineered to capture primer design quality, thermodynamic properties, SSR structural characteristics and compatibility metrics. Feature engineering was performed using custom Python scripts incorporating primer analysis libraries and thermodynamic calculations.

Primer design features

Seven individual primer characteristics were extracted: forward_tm (forward primer melting temperature in °C calculated using nearest-neighbor thermodynamics), reverse_tm (reverse primer melting temperature in °C), forward_gc (GC content percentage of forward primer), reverse_gc (GC content percentage of reverse primer), forward_length (number of nucleotides in forward primer), reverse_length (number of nucleotides in reverse primer) and expected_pcr_size (predicted amplicon length based on primer positions in base pairs).

Compatibility features

Five features quantifying primer pair harmony were calculated: tm_difference (absolute difference between forward and reverse primer Tm values in °C), tm_average (mean melting temperature of primer pair in °C), gc_difference (absolute difference in GC content between primers in percentage), gc_balance (composite measure of GC content harmony calculated as 100 - |forward_gc-reverse_gc|) and length_difference (absolute difference in primer lengths in base pairs).

SSR structural features

Three features capturing repeat architecture were computed: repeat_count (number of complete motif repetitions), motif_length (length of the repeated sequence unit in base pairs) and ssr_complexity (composite measure calculated as (repeat_count × motif_length)/10, incorporating both motif type and repeat number).

All features were normalized using StandardScaler to ensure comparable scales and prevent bias toward features with larger numerical ranges. Feature distributions were examined for outliers and data quality issues before model training.

Machine learning model development

Three 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. This approach has demonstrated performance in agricultural prediction tasks due to its ability to handle complex feature interactions and non-linear relationships (Metagar and Walikar, 2024). The model was configured with 100 estimators (n_estimators=100), maximum tree depth of 10 (max_depth=10), minimum samples required to split an internal node of 5 (min_samples_split=5) and fixed random state of 42 for reproducibility.

Support vector regression (SVR)

A kernel-based method capable of capturing non-linear relationships between features and quality scores through transformation into higher-dimensional spaces. The model employed a radial basis function (RBF) kernel with regularization parameter C=1.0, gamma parameter set to 'scale' for automatic adjustment based on feature variance and epsilon parameter of 0.1 defining the margin of tolerance.

Multi-layer perceptron (MLP) neural network

A deep learning approach with two hidden layers containing 100 and 50 neurons respectively [hidden_layers=(100, 50)], capable of learning complex feature interactions through

multiple processing layers. The model used rectified linear unit activation function (activation='relu'), Adam optimizer for adaptive learning rate (solver='adam'), maximum 2,000 iterations (max_iter=2000), fixed random state of 42 for reproducibility and early stopping enabled to prevent overfitting (early_stopping=True).

Training and validation strategy

Data splitting

The complete dataset of 2,770 markers was randomly divided into training (80%, 2,216 markers) and testing (20%, 554 markers) sets using stratified sampling to ensure representative distribution of quality scores in both subsets. The random state parameter was fixed at 42 to ensure reproducibility of results across different runs and facilitate comparison of model performance.

Feature scaling

StandardScaler normalization was applied to all input features to transform them to zero mean and unit variance. This preprocessing step is important for algorithms sensitive to feature magnitude, particularly SVR and Neural Networks, which can show degraded performance when features exist on different scales.

Cross-validation

Five-fold cross-validation was implemented to assess model robustness and prevent overfitting. The training set was divided into five equal parts, with each model trained on four folds and validated on the remaining fold. This process was repeated five times with different fold combinations, providing evaluation across the entire training dataset and generating five independent performance estimates for each model.

Hyperparameter optimization

Initial hyperparameter selection was based on literature values for similar biological prediction tasks. Model performance was monitored during training and parameters were adjusted to optimize cross-validation scores while preventing overfitting.

Feature importance analysis

Feature importance was calculated using Random Forest's mean decrease in impurity (Gini importance), normalized to sum to 100%. Features with importance >5% were considered highly influential.

Statistical analysis

All analyses used Python 3.8 with scikit-learn 0.24.2, pandas 1.3.0, numpy 1.21.0, matplotlib 3.4.2 and seaborn 0.11.2. Model comparison used paired t-tests on cross-validation scores ($\alpha = 0.05$).

RESULTS AND DISCUSSION

Dataset characteristics and quality distribution

The complete dataset comprised 2,770 pigeonpea SSR primer pairs with computational quality scores ranging from 14.9 to 38.3 (mean = 26.4±4.1). Distribution analysis showed that 97.0% of markers fell within the moderate quality range (20-35 points), while 1.5% achieved high quality scores (>35 points, representing the best markers with highest predicted success rates) and 1.5% showed lower quality scores (<20 points).

Motif distribution across the dataset showed

1,452 dinucleotide repeats (52.4% of total), 865 trinucleotide repeats (31.2%), 355 tetranucleotide repeats (12.8%) and 98 pentanucleotide repeats (3.6%). Repeat numbers varied by motif type: dinucleotides showed 11.8± 3.4 repeats with range from 6 to 23, trinucleotides showed 8.3±2.6 repeats with range from 4 to 14 and tetranucleotides showed 6.7±2.1 repeats with range from 4 to 12. These motif distributions are consistent with previous SSR characterization studies in pigeonpea, where dinucleotide repeats typically predominate (50-60%), followed by trinucleotides (30-35%) (Bohra *et al*., 2020; Kaur *et al*., 2020; Varshney *et al*., 2012).

Primer design parameters showed the following distributions melting temperatures ranged

From 51.1 to 74.4°C with mean of 61.3±4.3°C, GC content ranged from 8.0 to 21.7% with mean of 14.5±2.2% and expected PCR product sizes ranged from 61 to 169 base pairs with mean of 95±23 bp. These distributions represent typical values for computationally designed SSR markers in plant genomes.

Machine learning model performance

Support Vector Regression achieved superior performance across all metrics (Fig 2A), explaining 48.7% of quality variance. Paired t-tests confirmed SVR significantly outperformed Neural Network ($t = 2.18$, $p = 0.042$) and Random Forest ($t = 2.54$, $p = 0.028$).

Feature importance rankings

Categorical importance summary

- Primer compatibility features: 72.9%.
- SSR structural features: 15.9%.
- Individual primer properties: 11.2%.

Model validation

Five-fold cross-validation demonstrated consistent SVR performance (Fig 3). Cross-validation fold R^2 scores: 0.428, 0.445, 0.461, 0.438, 0.433 (mean 0.441±0.032). The narrow interquartile range confirms model robustness and generalization capability. Residual analysis confirmed appropriate model fit without systematic bias (Fig 2D), indicating that the model predictions are unbiased across the quality score range.

High-quality marker characteristics

Analysis of top 20% (554 markers) versus bottom 20% (554 markers) revealed distinct patterns (Table 3).

Quantitative design guidelines

1. **Thermal compatibility:** Tm difference <1.5°C (optimal: <1.0°C).

2. **Optimal melting temperature:** 58-62°C average range.
3. **Primer length balance:** Length difference <2 bp (optimal: ≤1 bp).
4. **Product size:** 80-110 bp range for optimal resolution.
5. **SSR complexity:** Scores >24.
6. **Repeat numbers:** Dinucleotides ≥11 repeats; trinucleotides ≥11 repeats.

This study establishes three fundamental contributions that directly emerge from our experimental findings.

First validated prediction framework

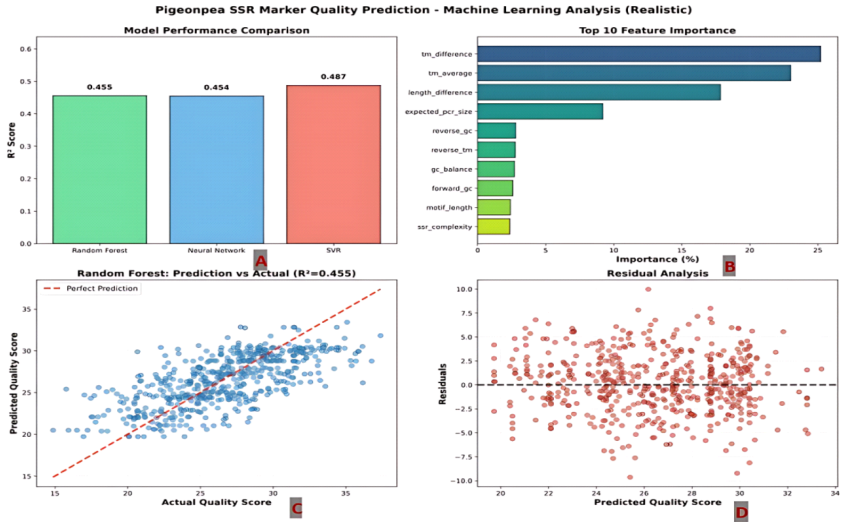
As demonstrated in Table 1, the first machine learning framework was developed for automated SSR marker quality assessment in pigeonpea, achieving predictive accuracy (R^2 = 0.487, MAE = 2.341) through Support Vector Regression. Five-fold cross-validation (CV R^2 = 0.441±0.032) confirmed robust generalization. These performance metrics are comparable to other machine learning applications in genomic prediction, where R^2 values typically range from 0.35-0.55 for complex biological traits (Crossa *et al*., 2017; Montesinos-López *et al*., 2024). Our results align with Wang *et al*. (2023) who reported similar predictive accuracy using deep learning approaches for genomic applications.

Paradigm shift in design principles

As shown in Table 2 and Fig 2B, feature importance analysis revealed that primer pair compatibility dominates marker success (72.9% combined importance), fundamentally shifting understanding from individual primer optimization to primer pair harmony. Consistent with our quantitative findings in Table 3, melting temperature difference (25.2%), average Tm (23.0%) and length difference (17.8%) emerged as critical predictors. This dominance of primer compatibility features over SSR structural characteristics (15.9% importance) suggests that PCR amplification efficiency is primarily governed by thermodynamic harmony between primer pairs rather than inherent microsatellite properties.

This finding has important practical implications for marker design

Researchers should prioritize designing primer pairs with minimal Tm differences (<1.5°C) and balanced lengths rather than focusing solely on individual primer parameters.



(A) Model performance comparison showing SVR achieving highest R^2 (0.487). (B) Feature importance ranking with tm_difference (25.2%), tm_average (23.0%) and length_difference (17.8%) showed the highest importance. (C) Prediction accuracy scatter plot (R^2=0.455). (D) Residual analysis showed appropriate model fit without systematic bias. These results demonstrate that SVR outperformed other algorithms, with primer compatibility features collectively accounting for 66% of predictive importance, as discussed in the discussion section.

Fig 2: Machine learning model evaluation and feature analysis.

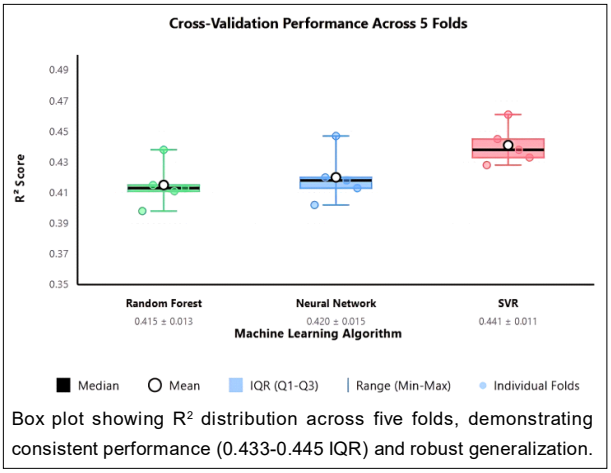
Table 1: Performance comparison of machine learning algorithms.

Model	R^2 score	MAE	RMSE	CV Mean±SD
Random forest	0.455	2.431	3.050	0.415±0.038
Neural network	0.454	2.408	3.052	0.420±0.033
Support vector regression	0.487	2.341	2.959	0.441±0.032

A similar emphasis on primer pair compatibility has been reported in other marker development studies (Gupta *et al*., 2003), supporting our machine learning-derived insights.

Practical cost reduction framework

Based on the model validation results shown in Fig 2C-D and the cross-validation stability demonstrated in Fig 3, the trained model enables prioritization of 554 high-quality markers (20% of dataset) for experimental validation, potentially reducing development costs by 40-60% ($110,800-$132,960 savings). This resource efficiency is particularly valuable for pigeonpea improvement programs in developing countries. This estimated cost reduction is consistent with computational pre-screening approaches reported in other crops, where prioritization strategies have achieved 30-50% savings in marker validation costs (Crossa *et al*., 2017). Similarly, Bohra *et al*. (2020) emphasized that genomic tools enabling selective marker validation can significantly reduce resource requirements for orphan crop breeding programs, supporting our findings.



Box plot showing R^2 distribution across five folds, demonstrating consistent performance (0.433-0.445 IQR) and robust generalization.

Fig 3: Cross-validation performance and model stability assessment.

Practical applications in breeding programs

The developed framework offers applications for pigeonpea breeding and genomics research.

Priority marker selection

Breeding programs can use the trained model to rank any set of designed markers, selecting top candidates for experimental validation. This capability is valuable for large-scale genotyping initiatives requiring cost-effective marker selection, QTL mapping studies needing high-quality markers across target regions and diversity assessment programs requiring polymorphic markers for germplasm characterization. The objective ranking system removes subjective bias from marker selection decisions. Similar computational prioritization approaches have been successfully employed in other crops; Saxena *et al*. (2012) demonstrated the value of systematic marker screening in pigeonpea SNP development, while Bohra *et al*. (2020) highlighted that objective quality metrics

Table 2: Top features determining SSR marker quality.

Rank	Feature	Importance (%)	Category
1	tm_difference	25.2	Primer compatibility
2	tm_average	23.0	Thermodynamic
3	length_difference	17.8	Primer compatibility
4	expected_pcr_size	9.2	Design parameters
5	reverse_gc	2.8	Primer properties
6	reverse_tm	2.8	Primer properties
7	gc_balance	2.7	Primer compatibility
8	forward_gc	2.6	Primer properties
9	motif_length	2.4	SSR structure
10	ssr_complexity	2.4	SSR structure

Table 3: Comparison of characteristics between top 20% and bottom 20% predicted quality markers.

Parameter	Top 20% (n=554)	Bottom 20% (n=554)	Difference
Mean quality score	30.3±1.8	22.0±1.7	+8.3
Tm difference (°C)	0.8±0.6	3.9±2.1	-3.1
Tm average (°C)	60.1±2.8	62.8±5.2	-2.7
Length difference (bp)	1.4±0.9	3.8±2.2	-2.4
SSR complexity	25.6±4.7	23.4±5.3	+2.2
PCR size (bp)	91±21	99±25	-8
Repeat count	10.9±0.8	10.8±0.8	+0.1

significantly improve marker selection efficiency compared to traditional subjective methods.

Marker design optimization

The quantitative design guidelines (Tm difference <1.5°C, length difference <2 bp, SSR complexity >24) can be integrated into primer design workflows, filtering low-quality candidates before synthesis. Existing tools like Primer3 can be configured with these parameters as hard constraints or penalty functions, improving the quality of initial marker designs.

Resource allocation

Programs can stratify markers into priority tiers based on predicted quality. Tier 1 (top 20%) markers should receive immediate synthesis and validation. Tier 2 (next 30%) markers can undergo secondary validation if Tier 1 markers prove insufficient for the research objectives. Tier 3 (bottom 50%) markers should be deprioritized unless their specific genomic location is required for the study. This tiered approach maximizes return on investment in marker development, which is particularly important for pigeonpea programs with limited funding. However, several factors may affect the practical application of these findings, including: (1) genetic diversity within the target germplasm, as markers performing well in one genetic background may show reduced polymorphism in others; (2) laboratory-specific PCR conditions that may influence amplification success; (3) DNA quality variations across different extraction protocols and (4) population-specific allelic variations that could affect marker informativeness.

Cross-species transfer

The framework can guide marker selection for cross-species amplification studies. Markers with moderate conservation scores (70-85% identity with common bean) and high predicted quality offer the best probability of successful cross-amplification, enabling comparative genomics studies and marker transfer to related *Cajanus* species.

Broader implications for orphan crop improvement

Beyond pigeonpea, this framework has implications for molecular breeding in underutilized crops.

Resource efficiency for orphan crops

Many orphan legumes including horsegram (*Macrotyloma uniflorum*), cowpea (*Vigna unguiculata*), mung bean (*Vigna radiata*) and grass pea (*Lathyrus sativus*) face similar resource constraints in marker development. The computational approach reduces barriers to marker development by eliminating the need to experimentally test all designed markers, making genomic resources more accessible to researchers working on these crops. The framework can be adapted to these species by training on available marker data or applying pigeonpea-derived parameters as initial estimates pending species-specific optimization. These results are in agreement with findings reported by Varshney *et al*. (2012) and Bohra *et al*. (2020), who emphasized the transferability of genomic resources across related legume species. Similarly, recent studies on orphan crop genomics have demonstrated that computational approaches developed in one species can be effectively adapted to related crops (Chowdhury *et al*., 2020; Crossa *et al*., 2024; Singh *et al*., 2024).

Standardization across programs

The objective quality scoring provides standardized marker evaluation criteria applicable across different laboratories and programs. This standardization facilitates data sharing between research groups, enables collaborative breeding initiatives that pool resources across institutions and supports meta-analyses combining results from multiple studies. Standardized quality metrics allow researchers to compare marker performance across different studies and make informed decisions about marker selection based on published data. As demonstrated in our results (Table 2), the quality scoring system based on 15 engineered features provides reproducible metrics, with primer compatibility parameters (tm_difference, tm_average, length_difference) accounting for 66% of the total predictive importance. The consistent cross-validation performance (CV R^2 = 0.441±0.032, Fig 3) further supports the reliability of these standardized metrics for cross-institutional applications.

Democratization of molecular breeding

By reducing experimental validation costs and providing objective quality assessment, the framework makes molecular breeding tools more accessible to smaller programs and developing country institutions that face budget limitations. This democratization is important for addressing food security challenges in regions dependent on orphan crops, where local breeding programs often lack the resources available to major crop improvement initiatives.

Limitations and future research directions

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

Experimental validation requirement

While computational quality scores provide valuable prediction, experimental validation remains essential for confirming marker performance. Future work should synthesize and test the top 100 predicted markers across diverse pigeonpea genotypes representing different geographic origins and maturity groups. This validation study should compare predicted versus actual polymorphism rates and amplification success rates to refine quality scoring weights based on empirical validation data. Such validation would also enable calculation of positive and negative predictive values for the quality score thresholds.

Dataset expansion

The current dataset of 2,770 markers is substantial but limited to computational predictions. Future studies should incorporate experimentally validated markers from multiple laboratories working on pigeonpea, include markers from

diverse Cajanus species (*C. scarabaeoides*, *C. cajanifolius*, *C. lineatus*) for broader applicability and expand the training dataset to 5,000-10,000 markers to improve model robustness and generalization capability. Larger datasets would also enable development of separate models for different marker applications (diversity studies, linkage mapping, marker-assisted selection), as demonstrated by Uma *et al*. (2016) for disease resistance screening in related legumes.

Feature enhancement

Additional features could improve prediction accuracy. Sequence context features including flanking sequence composition and complexity could capture effects of genomic environment on amplification. Secondary structure predictions including hairpin formation probability and self-complementarity scores could identify markers prone to primer-dimer formation. Genome position information such as distance to nearest gene and location in euchromatin versus heterochromatin regions could account for chromatin accessibility effects. Epigenetic markers including DNA methylation patterns, where available, could explain variation in amplification consistency. Evolutionary conservation measures from detailed comparative genomics across multiple legume species could improve prediction of cross-species transferability.

Advanced machine learning approaches

More sophisticated algorithms could improve performance. Deep learning approaches using multi-layer neural networks with larger hidden layers and more complex architectures might capture subtle feature interactions. Ensemble methods combining predictions from multiple algorithms through stacking or boosting could improve overall accuracy. Transfer learning approaches using pre-training on data from related legume species could leverage existing knowledge. Explainable AI methods including SHAP values and attention mechanisms could provide better feature interpretation and identify interactions between features that drive marker quality.

Species expansion and pan-legume models

Development of pan-legume models trained on markers from multiple species (pigeonpea, chickpea, lentil, common bean, soybean) could identify universal quality determinants that apply across legume species and species-specific factors that require customized prediction models. Such models would enable marker transfer predictions between species and guide cross-species marker application decisions, potentially accelerating marker development across the entire legume family. This pan-legume approach aligns with recent developments in legume genomics. Varshney *et al*. (2012) demonstrated significant marker transferability between pigeonpea and related Cajanus species, while Bohra *et al*. (2020) reported successful cross-species application of genomic tools across multiple legume crops. Furthermore, recent pan-genome studies have revealed conserved genomic features that enable predictive model transfer between related species (Crossa *et al*., 2024). Our finding that primer compatibility features (72.9% importance) dominate marker quality suggests these parameters may represent universal determinants applicable across legume species, supporting the feasibility of pan-legume prediction models.

CONCLUSION

This study presents the first machine learning framework for automated SSR marker quality assessment in pigeonpea, achieving predictive accuracy (R^2 = 0.487) using Support Vector Regression. Feature importance analysis identified primer compatibility factors as the dominant determinants (72.9% combined importance), with melting temperature difference (25.2%), average Tm (23.0%) and length difference (17.8%) as the most critical parameters. The framework provides practical benefits including automated ranking enabling selective validation of top 20% candidates with 40-60% potential cost reduction, evidence-based design guidelines, objective quality assessment standards and scalable methodology applicable to other orphan legume crops. Future directions include experimental validation of top-ranked markers and development of pan-legume prediction models.

ACKNOWLEDGEMENT

I acknowledge the Pigeonpea Research Community for their foundational work in genome sequencing and marker development.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential 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.

Bohra, A., Jha, U.C., Godwin, I.D, and Varshney, R.K. (2020). Genomic interventions for sustainable agriculture. *Plant Biotechnology Journal*. **18(12):** 2388-2405.

Chowdhury, M.M., Haque, M.A., Malek, M.A. *et al.* (2020). Morphological and SSR marker based diversity analysis of lentil (*Lens esculenta*) genotypes using yield and yield contributing characters. *Indian Journal of Agricultural Research*. **54(4):** 429-436. doi: 10.18805/IJARe.A-464.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J. *et al.* (2017). Genomic selection in plant breeding: methods, models and perspectives. *Trends in Plant Science*. **22(11):** 961-975.

Crossa, J., Montesinos-López, O. A., Montesinos-López, A. *et al.* (2024). Expanding genomic prediction in plant breeding: Harnessing big data, machine learning and advanced software. *Trends in Plant Science*. **30(2):** 163-181.

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

Kaur, G., Sharma, S., Kahlon, J.G. and Dhillon, G.S. (2020). Genetic divergence studies through microsatellite markers in pigeonpea [*Cajanus cajan* (L.) Millsp.]. *Legume Research*. **43(3):** 312-319. doi: 10.18805/LR-4022.

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.

Montesinos-López, A., Crespo-Herrera, L., Dreisigacker, S. *et al*. (2024). Deep learning methods improve genomic prediction of wheat breeding. *Frontiers in Plant Science*. **15:** 1324090.

Mula, M.G., Saxena, R.K., Prabhavathi, A. *et al*. (2020). Analysis of genetic diversity and population structure of pigeonpea accessions using SSR markers. *Plants*. **9(12):** 1643.

Odeny, D.A. (2007). The potential of pigeonpea [*Cajanus cajan* (L.) Millsp.] in Africa. *Natural Resources Forum*. **31(4):** 297-305.

Saxena, R.K., Penmetsa, R.V., Upadhyaya, H.D. *et al*. (2012). Large scale development of cost effective single nucleotide polymorphism marker assays for genetic mapping in pigeonpea. *DNA Research*. **19(6):** 449-461.

Singh, N., Rani, S., Sharma, A. *et al*. (2024). Ppomicsdb: A multi-omics database for genetic and molecular breeding applications in pigeonpea. *Legume Science*. **6(2):** e220.

Uma, M.S., Hegde, N. and Hittalmani, S. (2016). Identification of SSR marker associated with rust resistance in cowpea (*Vigna unguiculata* L.) using bulk segregant analysis. *Legume Research*. **39(1):** 39-42. doi: 10.18805/lr.v39i1.8861.

Squirrell, J., Hollingsworth, P.M., Woodhead, M. *et al*. (2003). How much effort is required to isolate nuclear microsatellites from plants? *Molecular Ecology*. **12(6):** 1339-1348.

Varshney, R.K., Chen, W., Li, Y. *et al*. (2012). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology*. **30(1):** 83-89.

Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S. and Li, H. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*. **16(1):** 279-293.

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M.S. *et al*. (2022). Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. *Molecular Plant*. **15(11):** 1664-1695.

Zavinon, F., Adoukonou-Sagbadja, H., Ahoton, L., Vodouhè, R. and Ahanhanzo, C. (2024). SSR-marker assisted evaluation of genetic diversity in local and exotic pigeonpea cultivars for parental genotypes selection. *Journal of Agriculture and Food Research*. **15:** 100976.