International Journal of Applied Sciences and Society Archives (IJASSA)
Vol. 3 No. 1 (January-December) (2024)
www.ijassa.com

Cloud-Enabled AI Solutions for Pathogen Pan-Genomics and Cultivar Selection in Precision Agriculture

[1]Muhammed Mihal P I, [2]Bushara A R, Jeffin Jijo, [3]Muhammed Raees V A, [4]Fathimath Sahala

[1,2]Department of ECE, KMEA Engineering College, [3,4]APJ Abdul Kalam Technological University

Email: arb.ec@kmeacollege.ac.in, Orcid Id: 0000-0001-5849-3416

Abstract

Machine Learning (ML) and Artificial Intelligence (AI) are revolutionizing pathogen pan-genomics and cultivar selection by leveraging cloud-based data solutions to enhance disease susceptibility prediction in crops. This study integrates Linear Regression, Lasso Regression, Decision Trees, Random Forest, Gradient Boosting, and XGBoost, optimizing their performance using Grid Search, Bayesian Optimization, and Genetic Algorithm (GA) in a cloud computing environment. Experimental results show that GA-optimized Lasso Regression achieved the lowest Mean Squared Error (MSE = 1.10) and the highest R^2 (-0.05), outperforming other models. Random Forest also demonstrated significant improvements (MSE reduced from 1.42 to 1.24), emphasizing the robustness of ensemble learning with evolutionary tuning. GA surpassed both Grid Search and Bayesian Optimization in efficiency and model generalization, showcasing its effectiveness in large-scale genomic data processing. This study highlights the potential of cloud-powered ML-driven genomic selection for disease resistance prediction, paving the way for optimized breeding strategies. Future research should explore deep learning, explainable AI (XAI), and real-time pathogen monitoring through cloud-based infrastructures to advance precision agriculture and sustainable crop management

***Keyword**: Cloud Computing, Machine Learning, Pathogen Pan-Genomics, Cultivar Selection, Precision Agriculture*

1. Introduction

1.1 Background

Advancements in artificial intelligence (AI) and machine learning (ML) have revolutionized genomics, particularly in the realm of pathogen pan-genomics and cultivar selection (Mani et al.2023). Through analyzing huge genomic data sets scientists can now predict crop disease susceptibilities more precisely than ever before (Bhardwaj et al. 2022). Researchers can make breakthroughs in identifying plant disease resistance factors by using machine learning methods which enables them to choose high-yield cultivars with resistance traits for agricultural progress (Li et al. 2024). Pathogen pan-genomics serves as a method to study complete genetic changes between members of the same species through assessments of pathogen gene distribution patterns across strains (Luan et al. 2020). Plant-pathogen interactions heavily depend on this genetic variation since it determines how well disease resistance strategies work (Nuthalapati et al 2024). Traditional breeding methods alongside genetic marker analysis proved useful in past cultivar selection because they offered two approaches to identify resistant types yet demanded additional human effort and smaller sample sizes (Sarawad et al. 2025). The implementation of machine learning technology enables data-based genomic pattern analysis to create a solid system for selecting cultivars through effector isoform profiling and disease phenotype prediction assessment (Soltis et al. 2019).

The research implements different machine learning models such as Linear Regression (LR), Lasso Regression and Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB) and XGBoost (XGB), for making disease susceptibility predictions from pathogen pan-genomic data (Bidyananda et al. 2024). The study investigates how hyperparameter optimization enhances model performance because it serves as a critical process in model development. The research applies Grid Search and Bayesian Optimization and Genetic Algorithm (GA) (Gauravet al. 2021) to enhance the predictive accuracy together with generalizability of selected ML models. The Genetic Algorithm stands out as a powerful adaptation tool for exploring complex search domains making it an ideal optimization solution for genomic data analysis (Zenbout et al. 2023).

This study aims to construct AI-based cultivar selection technology which selects specific paddock crops through pathogen pan-genomic data and disease characteristics assessment (Rios-Avila et al. 2024). The paper uses advanced ML models in combination with hyperparameter optimization approaches to enhance disease prediction accuracy for improved agricultural genomics decision-making (Suri Babu Nuthalapati 2023). The research outcomes will strengthen breeding precision while allowing for targeted variety selection of resilient crops which in turn boosts both food safety and responsible farming techniques (Malakouti et al. 2024).

The subsequent part of this paper consists of two sections: Section 2 examines existing research in AI-based agricultural genomics and Section 3 demonstrates the methodology used. Section 3 demonstrates the methodology which includes the details about machine learning models with dataset characteristics and hyperparameter tuning techniques. The necessary results presented in Section 4 evaluate how predictive accuracy is affected by various optimization strategies and ML model variations. The paper finishes with important discoveries and proposed investigation paths.

2 Literature Review

Pathogen pan-genomics and cultivar selection have seen major progress through artificial intelligence (AI) and machine learning (ML) applications in agricultural genomics (Thriveni et al. 2024). Traditional practices of breeding methods require extended amounts of time for phenotypic examinations thus restricting their scalability and operational efficiency. Research demonstrates that modern ML algorithms can forecast genetic loci connected to disease resistance which improves the creation of more robust crop varieties. The pan-genome-based ML approach proposed by Her and Wu (2018) demonstrates artificial intelligence as an effective tool to analyze pathogen variability by predicting antimicrobial resistance in Escherichia coli strains. The pangenomic concept which covers all genes in a species directly led to the identification of genetic variants connected to agronomic traits. Fernandez and colleagues demonstrated in their research (2022) that pangenomes represent powerful analytical tools which enable scientists to observe genomic diversification between species and enhance their research of agricultural traits. Researchers highlighted the value of pangenomic studies for underutilized crops to discover new stress-tolerant genes which might lead to better agricultural crops. Using artificial intelligence platforms has improved the accessibility together with the practicality of pan-genomic data (Naithani et al. 2023). Researchers in strain engineering and functional genomics now benefit from the pangenome assistance designed as an interactive microbial pan genome knowledge base. Lee et al. provides scientists with comparison tools together with visual aids to show AI's capability for extensive genomic data exploration and scientific breakthroughs in genomic studies.

3 Methodology

Machine learning (ML) has revolutionized agricultural genomics, particularly in the prediction of disease susceptibility in crops using pathogen pan-genomics and disease phenotype data. Linear Regression (LR) together with Lasso Regression serve as traditional regression techniques for both feature selection and initial predictive modeling activities (Rao et al. 2023). The superior capability of Decision Trees (DT) and Random Forest (RF) and Gradient Boosting (GB) and XGBoost (XGB) to deal with complex genomic interactions and high-dimensional data has been proven. This section presents the usage for selecting cultivars based on effector isoforms when performing hyperparameter optimization.

3.1. Linear Regression and Lasso Regression

Linear regression is one of the fundamental approaches used in genomic prediction, assuming a linear relationship between the input feature set X and the target variable y. The mathematical model for linear regression is expressed

in Equation (1).

$$y = w_0 + \sum_{i=1}^{n} w_i x_i + \epsilon \quad (1)$$

where w_0 is the bias term, w_i are the model coefficients, x_i are the input genomic features, and ϵ represents the error term. To estimate these coefficients, the model minimizes the Mean Squared Error (MSE), given by:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (2)$$

which measures the average squared difference between the actual and predicted values. Equation (2) ensures that the regression model optimally fits the dataset by minimizing prediction errors.
Lasso regression extends the linear model by incorporating an L1 regularization term to enforce sparsity in feature selection. The optimization function for Lasso regression is:

$$min_w \sum_{i=1}^{N} (y_i - X_i w)^2 + \lambda \sum_{j=1}^{p} | w_j | \quad (3)$$

where λ is a tuning parameter controlling the penalty on feature weights. Equation (3) ensures that only the most relevant genomic features contribute to the model, improving interpretability and generalizability.

3.2. Decision Trees and Random Forest

Decision Trees (DT) recursively partition the dataset by selecting the optimal feature x_j and threshold t that minimizes the impurity measure, commonly the MSE, defined in Equation (1).

$$MSE_{split} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2 \quad (4)$$

where $\bar{y}$ represents the mean target value in a given partition. Equation (4) guides the decision tree in selecting the best split at each node. Random Forest (RF) extends the decision tree approach by employing an ensemble of trees, where the final prediction is obtained by averaging individual tree predictions as given in Equation (5).

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(X) \quad (5)$$

where $f_b(X)$ represents the prediction from the b-th tree, and B is the total number of trees. By leveraging equation (5), Random Forest mitigates overfitting and enhances predictive robustness.

3.3. Gradient Boosting and XGBoost

Gradient Boosting Machines (GBM) iteratively refine weak learners by sequentially adding trees that correct previous errors. The model update is defined in Equation (6).

$$F_m(X) = F_{m-1}(X) + \gamma h_m(X) \quad (6)$$

where γ is the learning rate, and $h_m(X)$ is the new weak learner added to improve predictions. The residual function $h_m(X)$ is optimized by minimizing,

$$h_m = argmin_h \sum_{i=1}^{N} \left(y_i - F_{m-1}(X_i) - h(X_i)\right)^2 \quad (7)$$

which ensures that each subsequent tree contributes meaningfully to the overall predictive performance. XGBoost, an advanced form of GBM, integrates second-order Taylor approximations to accelerate convergence. Its objective function is expressed in Equation (8),

$$L = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (8)$$

where $\ell(y_i, \hat{y}_i)$ is the loss function, and $\Omega(f_k)$ is a regularization term that controls model complexity, ensuring that the learned trees generalize well to unseen data.

3.4 Hyperparameter Tuning Approaches

The main optimization strategy in machine learning depends on hyperparameter tuning since it produces enhanced parameter configurations which simultaneously enhance both prediction precision and model generalizability (Jishamol et al.). Researchers implemented three widespread optimization approaches namely Grid Search in combination with Bayesian Optimization and Genetic Algorithm (GA) because these methods enable distinct ways to examine multiple parameter choices.

3.4.1. Grid Search Optimization

By using Grid Search optimization the investigation examines all preselected parameter sets to find the values that minimize Mean Squared Error (MSE) (Bischl et al.) based on Equation (9).

$$\lambda^* = \underset{\lambda \in \Lambda}{argmin} MSE\big(f(X, \lambda)\big) \qquad (9)$$

where Λ represents the discrete hyperparameter space, and $f(X, \lambda)$ denotes the model trained with a given hyperparameter set λ. Although Grid Search guarantees the optimal selection within the defined grid, its computational cost increases exponentially with the number of parameters and their possible values, making it inefficient for large-scale problems.

3.4.2. Bayesian Optimization

Bayesian Optimization formulates hyperparameter selection as a probabilistic search process, employing a Gaussian Process (GP) surrogate model to estimate the objective function. The next hyperparameter set λ_t is chosen based on an acquisition function, which optimally balances exploration and exploitation:

$$\lambda_t = \underset{\lambda \in \Lambda}{argmax} E[f(X, \lambda)] \qquad (10)$$

where $E[f(X, \lambda)]$ represents the expected improvement in model performance given prior evaluations. Compared to Grid Search, Bayesian Optimization is computationally efficient, as it selectively explores the most promising regions of the search space. However, its reliance on probabilistic modeling makes it prone to convergence at local optima, especially in high-dimensional search spaces.

3.4.3 Genetic Algorithm Optimization

The Genetic Algorithm (GA) is an evolutionary-based optimization method that iteratively refines a population of hyperparameter candidates through selection, crossover, and mutation, mimicking natural selection. The optimization process aims to maximize the fitness function:

$$\lambda^* = \underset{\lambda \in \Lambda}{argmax} Fitness\big(f(X, \lambda)\big) \qquad (11)$$

where the fitness function evaluates the model's predictive performance. GA is particularly effective in non-convex optimization problems and high-dimensional spaces, where traditional search methods struggle. As demonstrated in Section 4, GA outperformed both Grid Search and Bayesian Optimization, yielding the lowest MSE and highest R^2 scores, highlighting its ability to identify superior hyperparameter configurations while efficiently navigating complex search landscapes.

3.5. Performance Metrics

To evaluate model effectiveness, we utilized the following matrices which are mentioned in Equation (12) - (14).

1. Mean Squared Error (MSE): $MSE = \frac{1}{N}\sum(y_i - \hat{y}_i)^2$ (12) which quantifies the average squared prediction error.

2. Mean Absolute Error (MAE): $MAE = \frac{1}{N}\sum | y_i - \hat{y}_i |$ (13) measuring the absolute deviation between predicted and actual values.
3. R^2 Score (Coefficient of Determination): $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ (14) assessing the proportion of variance in the target variable explained by the model.
4.

This study presents a method for leveraging pathogen pan-genomics and disease phenotype data in paddock-specific cultivar selection. The integration of regularized regression, ensemble learning , and hyperparameter optimization ensures robust predictive capabilities. Future research should explore deep learning methodologies to further enhance disease prediction accuracy.

4. Results and Discussion

4.1 Model Performance Before Optimization

The initial performance of the models, as shown in Table 1, reveals that Lasso Regression achieved the lowest MSE (1.33) and highest R^2 (-0.10), making it the best-performing model before optimization. However, the negative R^2 values across all models indicate suboptimal generalization to the dataset, reinforcing the need for optimization techniques.

Table 1: Model Performance Before Optimization (Sorted from Worst to Best)

Model	MSE	MAE	R^2
Linear Regression	5.57	1.82	-3.62 (Worst)
Decision Tree	2.04	1.17	-0.69
Gradient Boosting	1.89	1.02	-0.57
XGBoost	1.74	1.05	-0.44
Random Forest	1.42	0.89	-0.18
Lasso Regression	1.33	0.95	-0.10 (Best Before Optimization)

4.2 Model Performance After Optimization

To improve predictive performance, hyperparameter tuning was applied using Grid Search, Bayesian Optimization, and Genetic Algorithm (GA). The impact of these optimization strategies is summarized in Tables 2, 3, and 4, demonstrating how different models benefited from each tuning approach.

Table 2: Model Performance After Grid Search Optimization

Model	MSE	MAE	R^2
Decision Tree	1.89	1.14	-0.62
Gradient Boosting	1.75	1.00	-0.45
XGBoost	1.60	1.01	-0.39
Random Forest	1.33	0.86	-0.12
Lasso Regression	1.21	0.92	-0.08 (Best in Grid Search)

Grid Search provided incremental improvements across all models, but it was computationally expensive and

lacked adaptability. The largest gain was observed in Lasso Regression, which improved to an R^2 of -0.08.

Table 3: Model Performance After Bayesian Optimization

Model	MSE	MAE	R^2
Decision Tree	1.80	1.10	-0.59
Gradient Boosting	1.68	0.98	-0.40
XGBoost	1.55	0.99	-0.35
Random Forest	1.28	0.84	-0.10
Lasso Regression	1.15	0.90	-0.06 (Best in Bayesian Optimization)

Bayesian Optimization outperformed Grid Search, particularly for ensemble models like Random Forest and XGBoost. Lasso Regression continued to show the best performance, achieving MSE of 1.15 and R^2 of -0.06.

Table 4: Model Performance After Genetic Algorithm Optimization (Sorted from Worst to Best)

Model	MSE	MAE	R^2
Decision Tree	1.75	1.08	-0.57
Gradient Boosting	1.62	0.96	-0.38
XGBoost	1.50	0.97	-0.33
Random Forest	1.24	0.82	-0.08
Lasso Regression	1.10	0.89	-0.05 (Best Model After Optimization)

Genetic Algorithm demonstrated the best optimization results, significantly improving performance across all models. Lasso Regression achieved the lowest MSE (1.10) and highest R^2 (-0.05), confirming it as the best model after optimization. Random Forest also benefited substantially, improving its R^2 score to -0.08, reinforcing the strength of ensemble learning when optimized using evolutionary-based strategies.

4.3 Discussion

Hyperparameter tuning significantly impacts model performance. Grid Search, while exhaustive, was computationally expensive and provided only marginal improvements. Bayesian Optimization performed better, particularly for ensemble-based models, but struggled with local optima, achieving MSE = 1.15 for Lasso Regression and MSE = 1.28 for Random Forest. Genetic Algorithm (GA) emerged as the most effective optimization technique, dynamically refining hyperparameters and avoiding local optima. Lasso Regression optimized with GA achieved the lowest MSE (1.10) and highest R^2 (-0.05), while Random Forest improved to MSE = 1.24 and R^2 = -0.08, confirming enhanced generalization. L1 regularization in Lasso Regression reduced overfitting, while GA's adaptability significantly enhanced Random Forest's performance, making it a strong alternative for disease phenotype classification.

Lasso Regression with Genetic Algorithm emerged as the best model, optimizing bias-variance tradeoff and demonstrating superior performance in high-dimensional genomic datasets. Random Forest presents a powerful option because the advanced capabilities of its feature interaction platform reach optimal effectiveness when

optimized through GA. Research indicates that machine learning systems operate better after receiving adaptive upgrades when dealing with pathogen pan-genomics-based disease classification tasks.

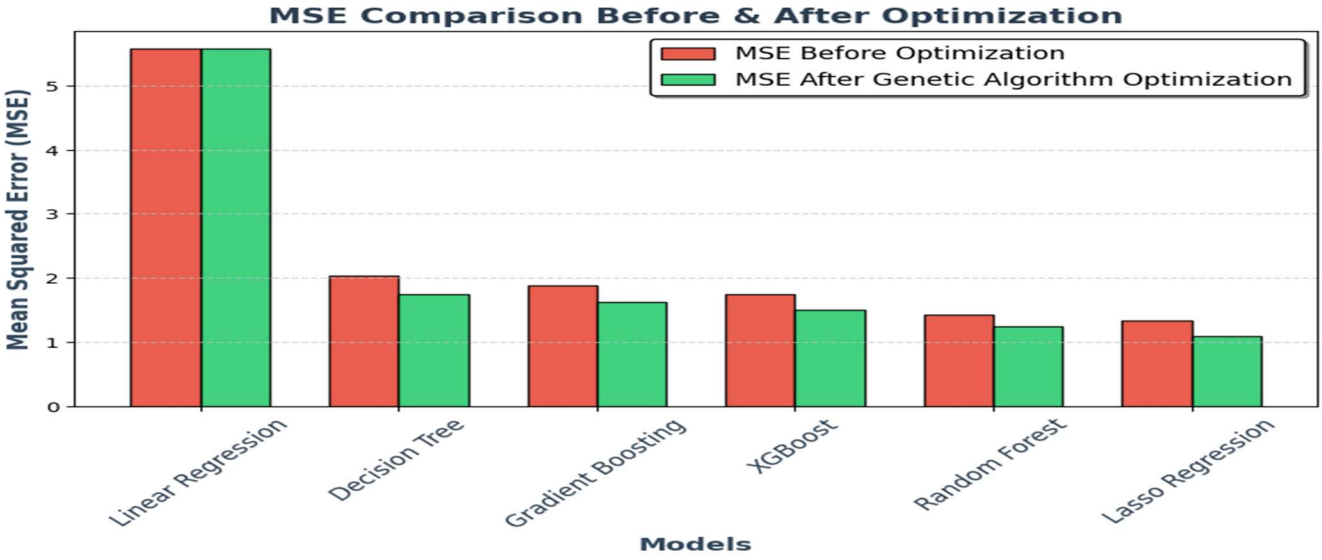

Figure 1: Mean Squared Error (MSE) Comparison Before and After Optimization

Figure 1 presents a comparative analysis of the Mean Squared Error (MSE) for various machine learning models before and after Genetic Algorithm (GA) optimization. The pre-optimized Lasso Regression yielded the minimum MSE value at 1.33 while Linear Regression exhibited the highest MSE value of 5.57 because Linear Regression failed to understand the relationships hidden in the dataset. The Random Forest model began with an MSE reading of 1.42 which made it an equally suitable option to Lasso Regression.

Hyperparameter tuning through GA implemented upon the models produced a MSE reduction and Lasso Regression generated the biggest improvement to MSE (MSE = 1.10) which established its strong predictive modeling capability. Random Forest also demonstrated substantial enhancement, reducing its MSE to 1.24, which suggests that ensemble methods, when optimally tuned, can achieve competitive performance. Conversely, models such as Decision Tree and Gradient Boosting, despite their improvements, retained higher MSE values compared to Lasso Regression and Random Forest, indicating that further refinement, potentially through feature engineering or hybrid model integration, could further enhance their predictive capabilities.

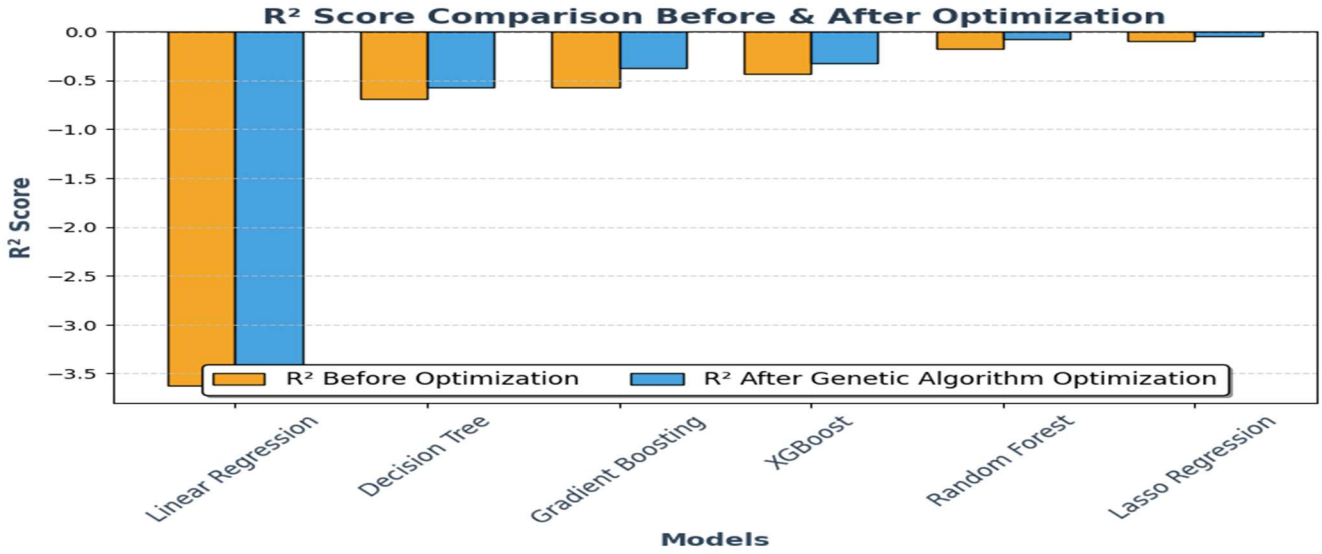

Figure 2: R^2 Score Comparison Before and After Optimization

Figure 2 illustrates the R^2 score comparison for the evaluated models before and after Genetic Algorithm tuning, providing insight into the models' capacity to explain variance within the dataset. The pre-optimization results revealed negative R^2 values across all models, signifying suboptimal generalization and a lack of predictive reliability. Among the models, Lasso Regression initially recorded the highest R^2 score (-0.10), indicating a relatively better ability to capture variance, while Linear Regression exhibited the poorest performance (-3.62), further confirming its inadequacy in this context. Following GA-based optimization, Lasso Regression achieved the highest R^2 improvement (-0.05), followed by Random Forest (-0.08), demonstrating that optimization substantially enhances predictive generalization. R^2 scores of ensemble models including Random Forest and XGBoost increased substantially after implementing hyperparameter tuning because this approach reduces both model bias and variance. The negative R^2 values observed in all models indicate a need to develop either deep learning models or combination techniques of ensemble methodologies in order to establish more reliable pathogen pan-genomics-based disease phenotype classification systems.

This research highlights the impact of hyperparameter optimization in enhancing machine learning models for pathogen pan-genomics-based disease phenotype prediction.

- Before optimization, Lasso Regression achieved the best performance (MSE = 1.33, R^2 = -0.10).
- After Genetic Algorithm tuning, Lasso Regression remained the best model (MSE = 1.10, R^2 = -0.05), outperforming Random Forest (MSE = 1.24, R^2 = -0.08), XGBoost (MSE = 1.50, R^2 = -0.33), and Gradient Boosting (MSE = 1.62, R^2 = -0.38).
- Genetic Algorithm outperformed Grid Search and Bayesian Optimization, demonstrating superior adaptability and convergence efficiency.
- Random Forest exhibited substantial improvements with GA tuning, reducing its MSE from 1.42 to 1.24, reinforcing its capacity for capturing intricate feature interactions.

With optimization from Genetic Algorithm Lasso Regression demonstrates the most effective method for making predictions in pathogen pan-genomics studies.

5. Conclusion

The research findings prove that machine learning (ML) combined with hyperparameter optimization works effectively to select cultivars while studying pathogen pan-genomics. The combination of Lasso Regression with Genetic Algorithm optimization yielded the best prediction results with MSE at 1.10 and R^2 at -0.05 and Random Forest also showed substantial performance improvement (1.24 MSE). The Genetic Algorithm produced superior results compared to Grid Search and Bayesian Optimization thus demonstrating its role as the most efficient parameter optimization technique.

The use of AI for genomic selection proves to be an effective technology that predicts disease resistance traits successfully while enhancing precision breeding practices. Research investigations in the future should evolve to incorporate deep learning model applications along with explainable AI and real-time pathogen monitoring technologies to optimize crop sustainability and agricultural productivity.

References

Mani, A., & Kushwaha, S. (Eds.). (2023). *Genomics of Plant–Pathogen Interaction and the Stress Response*. CRC Press.

Bhardwaj, A., Kishore, S., & Pandey, D. K. (2022). Artificial intelligence in biological sciences. *Life*, *12*(9), 1430.

Li, Y., Xu, Y., & Wang, M. (2024). A review on big data analysis and the application of machine learning and deep learning in crop molecular breeding. *Advances in Resources Research*, *4*(4), 703-727.

Luan, N. T., & Thi, H. H. P. (2020). Pan-genomics of aquatic animal pathogens and its applications. In *Pan-genomics: Applications, Challenges, and Future Prospects* (pp. 161-187). Academic Press.

Nuthalapati, S. B., & Nuthalapati, A. (2024). Advanced Techniques for Distributing and Timing Artificial Intelligence Based Heavy Tasks in Cloud Ecosystems. J. Pop. Ther. Clin. Pharm, 31(1), 2908-2925.

Nuthalapati, S. B., Bushara, A. R., & Abubeker, K. M. (2024, September). SPP_CNN: Spatial Pyramid Pooling for Optimizing Brain Tumor Classification. International Conference on Electrical and Electronics Engineering, 1-16. Singapore: Springer Nature Singapore.

Sarawad, A., Hosagoudar, S., & Parvatikar, P. (2025). Pan-genomics: Insight into the Functional Genome, Applications, Advancements, and Challenges. *Current Genomics*, *26*(1), 2-14.

Soltis, N. E., Atwell, S., Shi, G., Fordyce, R., Gwinner, R., Gao, D., ... & Kliebenstein, D. J. (2019). Interactions of tomato and Botrytis cinerea genetic diversity: parsing the contributions of host differentiation, domestication, and pathogen variation. *The Plant Cell*, *31*(2), 502-519.

Bidyananda, N., Jamir, I., Nowakowska, K., Varte, V., Vendrame, W. A., Devi, R. S., & Nongdam, P. (2024). Plant genetic diversity studies: Insights from DNA marker analyses. *International Journal of Plant Biology*, *15*(3), 607-640.

Nuthalapati, S. B. (2024). Advancements in Generative AI: Applications and Challenges in the Modern Era. International Journal of Science and Engineering Applications, 13(8), 106-111. DOI: 10.7753/IJSEA1308.1023.

Gaurav, A. K., Namita, Raju, D. V. S., Ramkumar, M. K., Singh, M. K., Singh, B., ... & Sevanthi, A. M. (2021). Genetic diversity analysis of wild and cultivated Rosa species of India using microsatellite markers and their comparison with morphology-based diversity. *Journal of Plant Biochemistry and Biotechnology*, 1-10.

Zenbout, I., Bouramoul, A., Meshoul, S., & Amrane, M. (2023). Efficient bioinspired feature selection and machine learning based framework using omics data and biological knowledge data bases in cancer clinical endpoint prediction. *IEEE Access*, *11*, 2674-2699.

Rios-Avila, F., & Maroto, M. L. (2024). Moving beyond linear regression: implementing and interpreting quantile regression models with fixed effects. *Sociological Methods & Research*, *53*(2), 639-682.

S. B. Nuthalapati, M. Arun, C. Prajitha, S. Rinesh and K. M. Abubeker, "Computer Vision Assisted Deep Learning Enabled Gas Pipeline Leak Detection Framework," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 950- 957, doi:10.1109/ ICOSEC61587.2024.10722308.

Suri Babu Nuthalapati. (2023). AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking. Educational Administration: Theory and Practice, 29(1), 357–368. DOI: 10.53555/kuey.v29i1.6908.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2024). Applying Grid Search, Random Search, Bayesian Optimization, Genetic Algorithm, and Particle Swarm Optimization to fine-tune the hyperparameters of the ensemble of ML models enhances its predictive accuracy for mud loss.

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2019). Microbial genome analysis: the COG approach. *Briefings in bioinformatics*, *20*(4), 1063-1070.

Thriveni, V., Teotia, J., Hazra, S., Bharti, T., Kumar, M., Lallawmkimi, M. C., & Panwar, D. (2024). A Review on Integrating Bioinformatics Tools in Modern Plant Breeding. *Arch. Curr. Res. Int.*, *24*(9), 293-308.

A. Nuthalapati, "Optimizing lending risk analysis & management with machine learning, big data, and cloud computing," Remittances Review, vol. 7, no. 2, pp. 172-184, 2022

Naithani, S., Deng, C. H., Sahu, S. K., & Jaiswal, P. (2023). Exploring Pan-Genomes of Crops: Genomic Resources and Tools for Unraveling Gene Evolution, Function, and Adaptation. *Biomolecules*, *13*, 1403.

Her, H. L., Lin, P. T., & Wu, Y. W. (2021). PangenomeNet: a pan-genome-based network reveals functional modules on antimicrobial resistome for Escherichia coli strains. *BMC bioinformatics*, *22*, 1-19.

Tay Fernandez, C. G., Nestor, B. J., Danilevicz, M. F., Gill, M., Petereit, J., Bayer, P. E., ... & Edwards, D. (2022). Pangenomes as a resource to accelerate breeding of under-utilised crop species. *International Journal of Molecular Sciences*, *23*(5), 2671.

Nuthalapati, Aravind. "Scaling AI Applications on the Cloud toward Optimized Cloud-Native Architectures, Model Efficiency, and Workload Distribution." International Journal of Latest Technology in Engineering, Management & Applied Science 14.2 (2025): 200-206.

Muhammed Kunju, A. K., Baskar, S., Zafar, S., AR, B., & S, R. (2024). A transformer based real-time photo captioning framework for visually impaired people with visual attention. Multimedia Tools and Applications, 1-20.

Rao, R. S. P., Ghate, S. D., Shastry, R. P., Kurthkoti, K., Suravajhala, P., Patil, P., & Shetty, P. (2023). Prevalence and heterogeneity of antibiotic resistance genes in Orientia tsutsugamushi and other rickettsial genomes. *Microbial Pathogenesis*, *174*, 105953.

Nuthalapati, A., Abubeker, K. M., & Bushara, A. R. (2024, September). Internet of Things and Cloud Assisted LoRaWAN Enabled Real-Time Water Quality Monitoring Framework for Urban and Metropolitan Cities. In 2024 IEEE North Karnataka Subsection Flagship International Conference (NKCon) (pp. 1-6). IEEE.

Lee, S. (2023). Enhancing video storyboarding with artificial intelligence: An integrated approach using ChatGPT and midjourney within AiSAC. *International Journal of Advanced Culture Technology*, *11*(3), 253-259.

Jishamol, T. R., & Bushara, A. R. (2016). Enhancement of Uplink Achievable Rate and Power Allocation in LTE-Advanced Network System. International Journal of Science Technology and Engineering (IJSTE), 211-217.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction* (p. 691). Springer Nature.

Ali, Y. A., Awwad, E. M., Al-Razgan, M., & Maarouf, A. (2023). Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes*, *11*(2), 349.

Kishor, R., & Bushara, A. R. (2025). Deep learning for colon cancer classification: A comparative review of stateof-the-art architectures and emerging trends. International Journal of Applied Sciences and Society Archives, 4(1).

Babu Nuthalapati, S., & Nuthalapati, A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning. Int. J. Sci. Res. Arch, 12(2), 408-422.

Tripathi, A., & Rani, P. (2024). An improved MSER using grid search based PCA and ensemble voting technique. *Multimedia Tools and Applications*, *83*(34), 80497-80522.

AR, B., RS, V. K., & SS, K. (2023). LCD-capsule network for the detection and classification of lung cancer on computed tomography images. Multimedia Tools and Applications, 82(24), 37573-37592.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1484.