# RiceAML: An Auto Machine Learning Model to Identify Abiotic Stress-Associated Single Nucleotide Polymorphisms in Rice

Hasnaa Imad Al-Shaikhli[1], Suhad A. Yousif[1], Tiba Zaki Abdulhameed[1], and Suzan Jabbar Obaiys[2(✉)]

[1] Al-Nahrain University, College of Sciences, Computer Science Department, Baghdad, Iraq
hasna.imad@nahrainuniv.edu.iq

[2] Department of Computer System and Technology, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
suzan@um.edu.my

**Abstract.** Identifying core single nucleotide polymorphisms (SNPs) associated with rice abiotic stress traits computationally helps researchers to deepen their understanding of genetic diversity, trait inheritance, yield improvements, and marker-assisted selection in rice breeding. The advancement of recent Artificial Intelligence (AI) can facilitate the identification and prediction of essential SNPs. In this study, a RiceAML model is proposed via utilizing Auto Machine Learning (AML) approach like (TPOT) to classify SNPs of abiotic stress traits in rice. Eight datasets (six individual datasets and two combined datasets) with two DNA sequence lengths 40 and 100 nucleotides, are used. Three encoding approaches are applied to convert DNA sequences into numeric values such as One-Hot, DNA2Vec, and concatenation of both feature vectors. The results showed that the prediction accuracy of the model trained on merged or combined dataset, especially Combined_100, outperformed the prediction accuracy of models trained on individual datasets, particularly when using the One-Hot Encoding, achieving an accuracy of 93.7%. The proposed RiceAML model with related datasets can be accessed through https://github.com/HasnaaImad/RiceAML.

**Keywords:** Rice · SNP · AI · TPOT AutoML · DNA mining · One-Hot Encoding · DNA2Vec Encoding · Bioinformatics

## 1 Introduction

Rice (Oryza Sative) is a fundamental staple crop that has played a substantial role in agriculture and human nutrition for over half the world [9,20]. This improves rice genetic research, which is important in global food security.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 81–96, 2025.
https://doi.org/10.1007/978-3-031-96997-3_6

Researchers and breeders face challenges in optimizing rice quality and quantity through its sophisticated genome architecture and many traits. Specifying core single nucleotide polymorphisms (SNP) related to targeted traits computationally from large genomic data accelerates breeding by reducing irrelevant mutation noise [20].

In the rice genome, Single Nucleotide Polymorphisms (SNPs) are the most popular genetic variation, consisting of altering a single nucleotide base. These variations are important in understanding genetic diversity, trait inheritance, and marker-assisted selection in rice breeding. Identifying SNPs related to popular traits such as yield, disease resistance, and stress tolerance can help breeders develop improved rice varieties more efficiently [6,13,25]. However, the rice genome has an extremely high number of SNP markers, ranging from hundreds of thousands to millions. It is difficult to use all the SNPs for analysis because of the dimensionality and comparably small sample sizes found in breeding data sets. Redundant or irrelevant SNPs add noise, decreasing prediction accuracy. Therefore, effective computational techniques must be employed to predict the most relevant SNPs [20].

Recent developments of machine learning (ML) and deep learning (DL) have remarkably boosted the analysis and employment of SNP data in rice research to address these challenges by enabling deeper analysis of complex and high dimensional SNP data. Farooq et al. [5] comprehensively review how artificial intelligence (AI) is revolutionizing modern plant breeding. The authors highlight the importance of AI in handling large-scale genomic and phenomic data to allow faster genetic gains. Through deep learning, machine learning, and natural language processing algorithms, AI improves trait prediction, variant detection, gene discovery, and genomic selection. Integrating multi-omic data with high-throughput phenotyping allows breeders to reveal complex relationships among traits, leading to increased crop resilience and productivity. The study emphasizes the ability of AI to bridge the genotype-phenotype gap and optimize genome editing tools to create cultivars ready for future challenges.

Another study [10] by Eugene Koh et al. explains the application of AI in the analysis of omics data on plant stress. The authors review more than 4,000 BioProjects and identify widespread drought, heat, and pathogen stress with species such as Arabidopsis thaliana, rice, and maize. It shows how tools such as LLM and AlphaFold can extract methods, build knowledge graphs, and predict gene functions from raw DNA. Despite problems such as inconsistent metadata and explainability, it emphasizes the importance of AI in boosting the study of plant resilience through existing tools and collaboration.

How big data and AI revolutionize plant stress research by interpreting plant reaction to environmental stresses is discussed in [11]. It reports on genomic and transcriptomic data from SRA and GEO on drought, salt, heat, cold, and pathogen stress in Arabidopsis thaliana, Oryza sativa, and Zea mays. AI techniques such as deep learning and large language models identify omics data patterns, make gene function predictions, and support literature mining. Tools such as AlphaFold, Enformer, FloraBERT, and AgroNT optimize stress tolerance predictions. Poor data labeling, interpretability, and unequal access to data are some of the hindrances to its application. The article appeals to inter-

disciplinary collaboration and more advanced data infrastructure to take AI to greater heights in sustainable agriculture under climate change.

Early crop genomic prediction utilized linear methods such as genomic best linear unbiased prediction (GBLUP), assuming small additive effects for every SNP. ML and DL can capture nonlinear effects (epistasis, dominance) without assuming prior strong requirements. Genomic rice researchers have used various models, from classical ML algorithms (random forests, support vector machines) to more recent DL architectures (convolutional neural networks, recurrent networks, transformers) to predict phenotypes of SNP genotypes or identify SNPs linked to specific traits [26]. What follows is a summary of significant papers from the last years using these methods, categorized by modeling method.

Traditional ML methods have been widely used to handle rice SNP datasets, specifically for feature selection and genomic prediction tasks. Random forests (RF) and gradient boosting (e.g. XGBoost) have worked effectively in the scenario of high-dimension SNP inputs and have feature importance metrics incorporated. For example, a 2024 research study introduced the PlantMine platform [20], which incorporates feature selection into ML algorithms to identify core SNPs (a reduced set of informative markers most predictive of target traits). Three feature selection methods (ANOVA, MIC, F-score) screened approximately 7202 potential SNPs from the 3000 Rice Genomes dataset in PlantMine, and various ML models (XGBoost, SVM, k-NN, Random Forest) were compared to predict flowering time. The best feature selection method, along with the ML model, produced nearly 98% accuracy in discriminating early and late flowering groups using a random forest based on the corresponding SNP subset. Feature selection ML algorithms can effectively manage genome-wide SNP data and also prevent overfitting. The random forest is particularly notable for the detection of SNPs and the increase in predictability. Feature selection improves model interpretability and performance by removing duplicate SNP features. PlantMine SNP discovery enables rice breeders to identify key genetic markers for traits such as heading date to prioritize for breeding.

In a crop genomics experiment in 2022 [7], the authors found that classical ML algorithms could outperform deep learning in genotype-to-phenotype predictions when data were sparse. Random forest and XGBoost models performed better than neural networks on 13 of 14 trait prediction tasks, and the top SNPs determined using XGBoost's feature importance were similar to established GWAS trait-associated loci. Properly tuned ML models yield strong performance and improve interpretability. Significant SNP markers can be identified by examining characteristic importance scores such as the XGBoost F score or RF Gini, some of which map to previously identified genes. A model could be reduced to 10% of the SNPs without loss of accuracy, increasing interpretability and speed.

During the study [18], machine learning methodologies are used to discriminate between various stress conditions by analyzing stress-responsive genes in the rice plant. Based on comparative transcriptomic data analysis, the models identify putative genes important for high-level stress tolerance, thus allowing

better insight into the genetic underpinnings of stress responses and allowing the generation of stress-tolerant rice cultivars.

Deep learning (DL) has transformed bioinformatics by allowing pattern recognition in large datasets without manual feature engineering. In rice genomics, DL models identify the SNP-associated trait as a classification problem and the quantitative prediction of traits from SNP inputs as a regression problem.

Tan et al. (2022) [4] presented NucleoNet, a novel deep learning framework used to predict rice yields and detect important single nucleotide polymorphisms (SNP) that control yields. NucleoNet outperformed traditional methods such as ordinary least squares (OLS) by leveraging SNP positional data and attention capabilities to identify important genetic markers related to yield. It used 1,536 genotyped SNP markers genotyped in 467 Indonesian rice varieties and tested on 4 environments. In particular, NucleoNet detected several potentially beneficial SNPs in rice breeding programs and highlighted the power of artificial intelligence in the genomic selection of complicated traits.

In 2024, Xu et al. introduced RiceSNP-BST [23], a deep learning framework for predicting SNPs associated with biotic stress traits employing convolutional neural networks. The input to the model consisted of DNA sequences surrounding putative SNP loci in several formats: one-hot representation, DNA2vec embeddings, DNABERT embeddings, and DNAShape features. These features allowed CNN to discover SNPs associated with the trait. RiceSNP-BST was unique in that its 35,749 pairs of SNPs from the GWAS Atlas were used in its training set, with the known pairs being positive and the unrelated SNPs negative. In an independent validation set, the model outperformed baselines such as gradient boosting with improved precision and recall. RiceSNP-BST resulted in an F1-score and AUC of 0.65 and 0.61, respectively, compared to an F1 of around 0.58 by XGBoost. Its generalization performance across species revealed strong learning from unseen data. Genomic task automation eliminates the requirement of tuning CNNs by hand.

A similar approach is followed in early 2025 when the RiceSNP-ABST model or framework [14] was introduced to predict SNPs associated with abiotic stress traits in rice against salinity and drought. It employed CNNs with residual connections to predict SNPs from DNA sequences. Six sets were used to train RiceSNP-ABST because there is poor high-quality data on rice abiotic stress QTLs. Four sequence representation methods (such as RiceSNP-BST: One-Hot, k-mer embeddings), with initial feature selection before CNN modeling, were applied. RiceSNP-ABST's CNN performs better than current machine learning models and past benchmarks. RiceSNP-ABST generalizes well to independent rice data and to other organisms and demonstrates that genomic patterns of stress-responsive variants are captured through learning representations. The model outperforms traditional machine learning methods and multiple state-of-the-art methods in the benchmark data set with a maximum accuracy of 61.38%.

Vourlaki et al. [21] examined the potential improvement in rice trait predictions by combining structural variants (SVs) with single nucleotide polymorphisms (SNPs) using deep learning (DL) methods. Their comparative study of DL models such as multilayer perceptrons (MLP) and convolutional neural networks (CNN) against Bayesian methods showed that DL outperformed conventional modeling approaches in many cases, especially in the case of binary traits like leaf senescence and culm diameter. In particular, adding SVs greatly enhanced predictive power, and SNPs linked to SVs showed high effectiveness. This study emphasizes the great potential that DL and diverse genomic inputs offer to revolutionize rice breeding methods.

Cropformer [22] is an advanced deep learning model with convolutional and attention layers designed in 2024 for predicting crop phenotypes, including rice-specific traits. It is not just applied to rice, but upon testing on five crops (maize, millet, rice, tomato, wheat) on 20+ characteristics, it showcases state-of-the-art deep learning in genomics. Cropformer's architecture integrates CNN layers to capture local haplotype signals and self-attention layers to extract long-range interactions focusing on influential SNPs. Rice genotypes are passed through 1D convolution to identify local linkage disequilibrium signals and then differential weighting by multi-head attention blocks. The architecture greatly improved predictive performance compared to models like rrBLUP and XGBoost. The predictive power was 7.5% better in some characteristics compared to the next-best model. In the case of rice, the cropformer improves predictions of yields and characteristics by capturing epistatic interactions. Crucially, it was designed with interpretability in mind. Once trained on the data, the authors performed haplotype analysis, attention visualizations, SHAP value computation, and clustering to interpret predictions. They selected SNPs with notable weights that could be causal to trait differences. Attention-weighted SNPs in maize identified significant biology associated with height and flowering genes. Likewise, the approach can identify significant SNPs for grain yield or rice disease resistance. The model integrates genomic data with environmental factors to enhance the precision of the predictions. The model's interpretability allows researchers to identify influential genetic markers and environmental factors influencing specific traits, thus facilitating the application of targeted breeding programs.

In this research, a classifier called RiceAML, based on TPOT auto machine learning approach, is proposed to predict the Aboitis Stress-Associated Single Nucleotide Polymorphisms (ABST SNPs) in rice. The first contribution of this work is to check whether auto machine learning will give a machine learning algorithm performance better than the current ones applied individually. To achieve this aim, six benchmark datasets, recently created in [14] with two DNA sequence lengths (40 and 100 bp), are utilized. The second contribution focuses on the study of the effects of combining datasets of the same length on the performance of the proposed model.

This manuscript is organized as follows: Sect. 2 describes the datasets used with the proposed model methodology in detail. Section 3 presents the results of the implementation of ML models with optimized selected features and parame-

**Table 1.** Summary of datasets with total number of samples for both training and testing stages

| Dataset Name (Total Samples) | Training (80%) | | | Testing (20%) | | |
|---|---|---|---|---|---|---|
| | +ve Samples | −ve Samples | Total | +ve Samples | −ve Samples | Total |
| Random_40 (5118) | 2,047 | 2,047 | 4,094 | 512 | 512 | 1,024 |
| Common_40 (5118) | 2,047 | 2,047 | 4,094 | 512 | 512 | 1,024 |
| Rearrange_40 (5604) | 2,241 | 2,241 | 4,482 | 561 | 561 | 1,122 |
| Random_100 (5140) | 2,056 | 2,056 | 4,112 | 514 | 514 | 1,028 |
| Common_100 (5140) | 2,056 | 2,056 | 4,112 | 514 | 514 | 1,028 |
| Rearrange_100 (5604) | 2,241 | 2,241 | 4,482 | 561 | 561 | 1,122 |
| Combined_40 (15,840) | 6,336 | 6,336 | 12,672 | 1,584 | 1,584 | 3,168 |
| Combined_100 (15,884) | 6,353 | 6,353 | 12,706 | 1,589 | 1,589 | 3,178 |

ters using TPOT. Finally, Sect. 4 summarizes the key contributions of this study and discusses the broader implications of using one-hot and DNA2vec encoding to predict (ABST SNP) in rice.

## 2 Materials and Methodology

The proposed model consists of four main phases: datasets preparation, sequence encoding and data splitting, constructing the model or classifier, and finally, classifier's performance evaluation as shown in Fig. 1. The description of each phase is explained in detail below.

### 2.1 Dataset Preparation Phase

The proposed RiceAML model is trained and tested on six recent benchmark datasets, which were created and applied for RiceABST model [14] with two DNA sequence lengths (40 and 100 bp). The RiceAML classifier is further trained and tested on two combined datasets that are created from merging all three datasets with the same DNA sequence length. For instance, a dataset named "Combined_40" is generated from combining datasets "Random_40", "Common_40", and "Rearrange_40" where each sequence has a length of 40 bp as listed in Table 1. It is good to mention that the original datasets' sequence

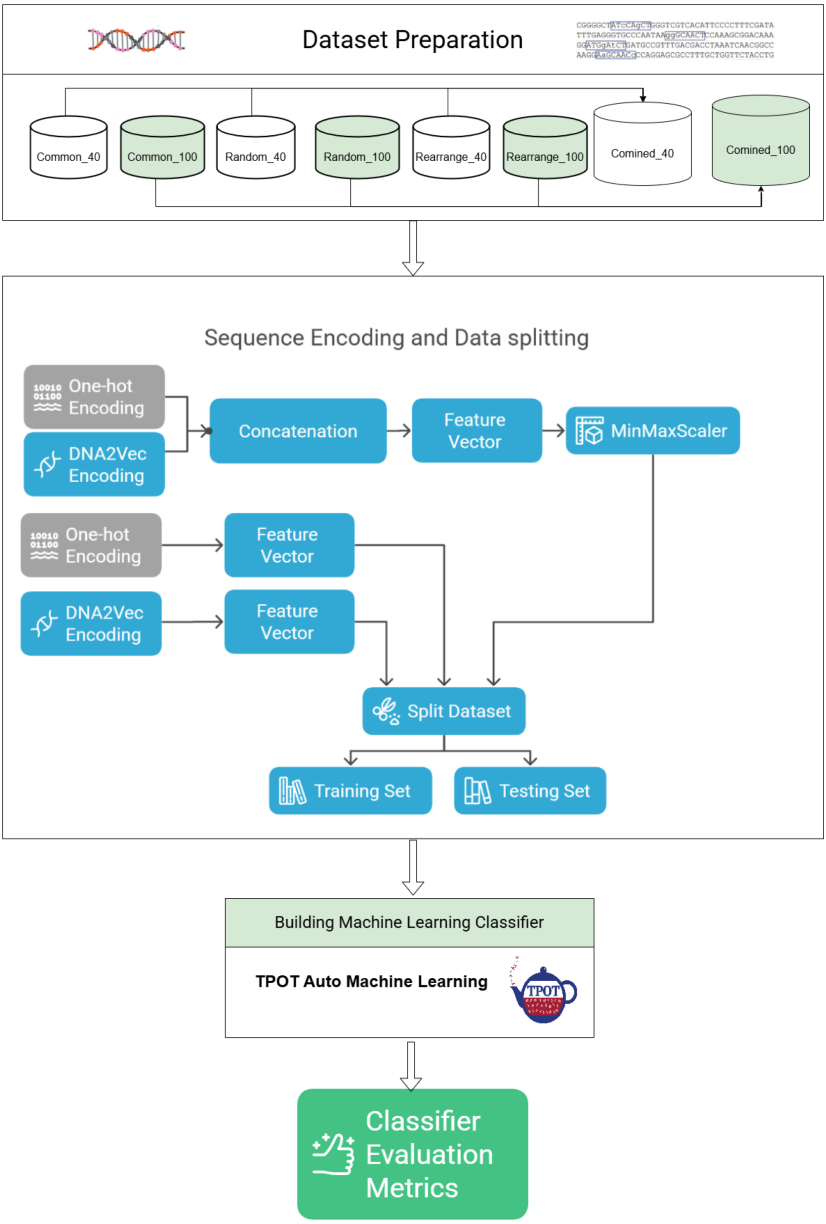


**Fig. 1.** Proposed model phases.

lengths are 40 or 100 nucleotides, not 41 or 101 nucleotides. For this reason, the names of the datasets in this study are different.

### 2.2 Sequence Encoding and Data Splitting Phase

To make the dataset instances (sequences) ready for machine learning classifiers, they must be encoded into numeric feature vectors. Thus, three encoding schemes are tested as listed.

- One-Hot-Encoding: The vocabulary set of the DNA sequence consists of the four letters C,G, T, and B. Thus, 4-bit One-Hot Encoding is used [3]. This technique embeds each nucleotide (A, C, G, and T)in a vector. As an example A=(1,0,0,0), G=(0,1,0,0), C=(0,0,1,0), and T=(0,0,0,1). This encoding is easy to generate and simple, but the vectors produced do not carry any contextual information that can help understand the motifs' pattern. The One-Hot-Encoding provides local information, where each DNA sequence nucleotide is considered a feature.
- DNA2Vec Encoding [17] is a DNA sequence embedding technique that uses a shallow neural network (NN) similar to word2vec [8]. The input to the NN are variable k-mers, which are equivalent to the language's vocabulary when word2vec embedding is applied in Natural Language Processing (NLP) applications. The main challenges of DNA2Vec are deciding the value of k, and the range of the surrounding context of the k-mers.
- Concatenation of One-Hot-Encoding and DNA2Vec Encoding into one vector. This is to test whether combining local and contextual relationships can provide more information about the sequence for the classifier. All features were normalized using MinMaxScaler to ensure scale consistency across dimensions.

After generating the feature vectors, the dataset is split into 80% for training and 20% for testing.

### 2.3 Building Machine Learning Classifier Phase

AutoML automates the process of model development, such as feature selection, hyperparameter tuning, and model evaluation. Some popular AutoML tools are AutoKeras [16], TPOT [15], and H2O.ai [24]. This research uses the Tree-based Pipeline Optimization Tool (TPOT) Auto ML. TPOT applies multiple iterations for feature selection until optimum selected features are passed to the next stage [12,19].

### 2.4 Classifier Evaluation Metrics Phase

This stage evaluates each classifier generated from the previous stage. Evaluation is based on the Accuracy and Confusion Matrix.

Accuracy is computed by the Eq. 1

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \tag{1}$$

Where TP, TN, FN, and FP stands for the total number of True Positive, True Negative, False Negative, and False Positive, respectively. A positive in the Rice trait prediction problem refers to a trait subsequence predicted from the DNA sequence. A TP is the number of Trait subsequences that are correctly classified, and a TN is the number of nontrait subsequences that are also correctly classified. On the other hand, FP is the number of non-trait subsequences incorrectly classified as traits, and FN is the number of trait subsequences incorrectly classified too.

Confusion Matrix is a heat map generated based on the TP, TN, FP, and FN. It illustrates a visualization of the classifier's performance, where higher values of TP and TN indicate better performance.

## 3 Results and Discussion

In this section, we will present the outcomes of our experiments, starting by loading the dataset and transforming it into an ideal shape to train the model on top of it. Then, we process three different encoding strategies to convert categorical or sequential data to numerical data for machine learning model training. In transforming categorical data to represent digitally, the most prevalent method is the One Hot Encoding method; however, in DNA sequence-based data the DNA2Vec Encoding captures a feature-wise representation of DNA sequences, thus working excellently with biological datasets. Finally, we concatenated the two encoding methods to get more features.

Using TPOT, we automatically optimize and select the best pipeline for every encoding method. The models are assessed based on their key performance metrics: classification accuracy and confusion matrix. Lastly, we show the efficacy of our approach by comparing the results with those of previous studies.

### 3.1 Results of Different Encoding Methods and Datasets

Based on the horizontal bar chart shown in Fig. 2., higher accuracy results are represented by a dark bar (combined datasets) or a light bar (individual datasets) for each encoding type, so the analysis is broken down into that structure below:

### 3.2 One-Hot Encoding Results

The combined_100 dataset produced an accuracy result of 0.937, which is higher than all other results by a significant margin, thus demonstrating very high performance. This high accuracy is due to the One-Hot Encoding ability to extract exact positioning information in large aggregated datasets. Individual runs on these datasets achieve an accuracy of 0.583 (rearrange_40), which reflects a significant performance drop from the combined dataset. This large gap shows that single datasets may be underrepresented or non-diverse and that the encoding method can affect accuracy even though a strong encoder is used.

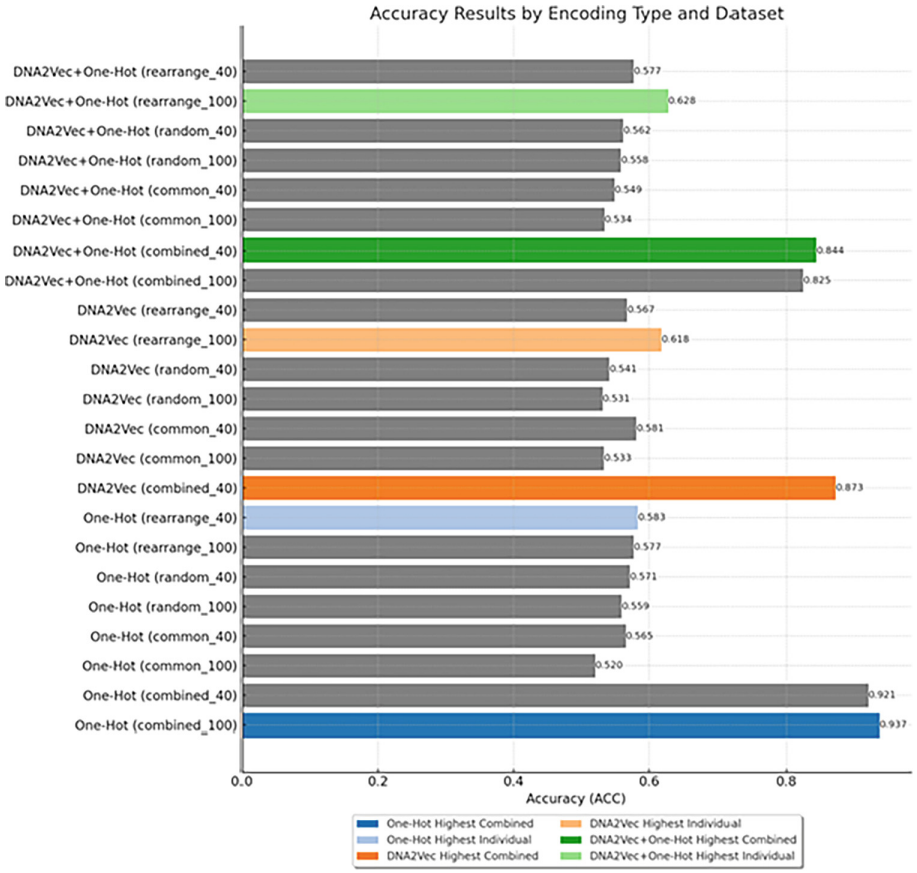


**Fig. 2.** Accuracy results by applying different encoding type on different datasets.

### 3.3 DNA2Vec Encoding Results

The best predictor of accuracy was (0.873) in combined_40 dataset with the DNA2Vec Encoding method, which is moderately lower than the One-Hot results, but remains strong. Although the DNA2Vec model can capture semantic similarities between DNA sequences, it may not encode position-specific information as effectively as One-Hot encoding. The best single result for DNA2Vec was in rearrange_100 with an accuracy of 0.618, which was lower than the combined dataset. It illustrates the drop in accuracy that is based on one specific embedding-based approach, which underwrites its trend to increase the data used for better accuracy.

### 3.4 Concatenated Encoding (One-Hot and DNA2Vec)

While this method did not outperform One-Hot method alone, the combined accuracy of 0.844 was still good. These results suggest that concatenated

encoding efficiently combines the advantages of the two approaches, the positional specificity of One-Hot and embedding-based semantic representations of DNA2Vec, to achieve a more balanced and robust prediction performance. The best accuracy seen from individual datasets was produced by rearrange_100 at 0.628, with slight advancements upon DNA2Vec only (0.618) and notably higher than individual One-Hot results (0.583). This result confirms that the advantage of concatenated encoding, as mentioned, is valid for all datasets; as a result, the previous gap between DNA2Vec and other models about positional representation was effectively compensated for.

These results show that the aggregation of datasets always beats the standalone dataset for any encoding strategy, signifying how critical the magnitude of the data is for any predictive storytelling. Of all encoding methods for the combined datasets, one-hole encoding emerges as the single most powerful method. It captures the fine-grained positional features essential for exploring biological sequences. The concatenation approach works effectively for individual datasets, combining the advantages of two encoding methods DNA2Vec scores moderate accuracy, indicating that it may require larger datasets or hybrid approaches for better performance. The score does not appear to improve much with increased training. This visual and analytical comparison method is structured to quickly identify the encoding strategies and their relative strengths. This will ultimately drive decisions about the analysis of biological data and the modeling methods applied in the future.

Finally, we summarize the highest accuracy values among different encoding methods and datasets, showing that the One-Hot Encoding (combined_100) and DNA2Vec concatenated with the One-Hot Encoding (rearrange_100) performed better overall, as shown in Fig. 3. Their confusion matrices are also given to support these results and provide a visualization, as listed in Fig. 4.

Performance analysis of the top results with the TPOT Auto ML classifier based on varying encoding methods, One Hot Encoding, DNA2Vec and One-Hot Concatenated DNA2Vec, for both combined and individual data sets shows that for the combined dataset, the best predictive performance was achieved based on One Hot Encoding (combined_100) with an accuracy (ACC) of 0.937362, an AUC of 0.937365, and an AUPR of 0.906. This impressive result was achieved by the TPOT AutoML pipeline using a well-planned series of preprocessing methods, comprising StandardScaler for feature scaling, and SelectPercentile for keeping about 18% of the most statistically relevant features, and combined parallel Feature Union layers coupling of SkipTransformer (to transfer critical raw feature signals) and Passthrough (to maintain intact numerically encoded raw feature signals). The bagging classifier built on top of this process now only uses 67 trees, no bootstrapping, 61% features, and 66.7% samples to ensure the model diversity and prediction performance are stable.

The highest single result achieved on the individual dataset was using the One-Hot concatenated DNA2Vec Encoding features from the rearrange_100 dataset, which scored an accuracy of 0.628011, an AUC of 0.628002, and an AUPR of 0.580297. The TPOT pipeline executing this individual dataset's best

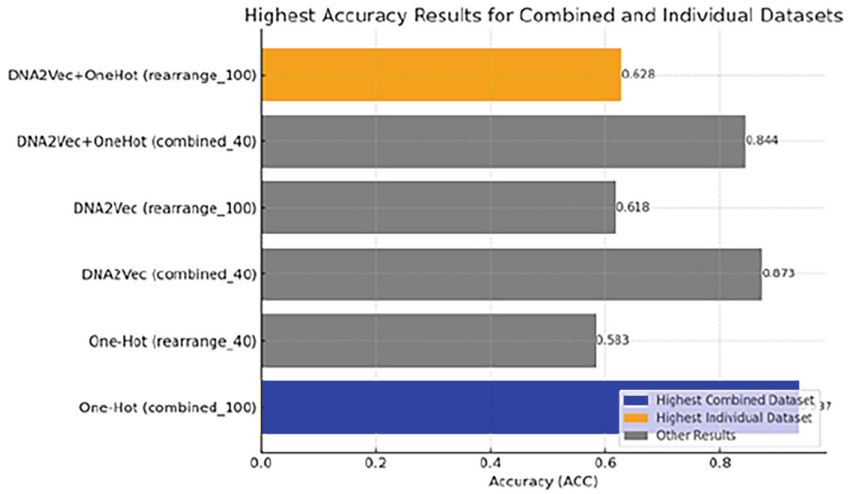


**Fig. 3.** Highest accuracy results for combined and individual datasets.

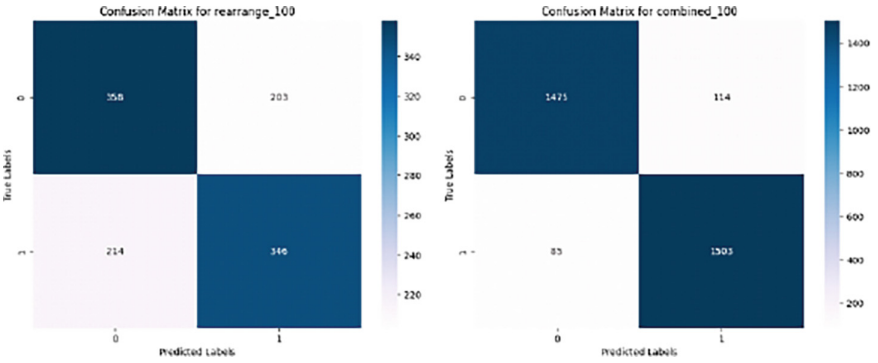


**Fig. 4.** Confusion matrices of TPOT pipeline models.

result performed extensive preprocessing, with initial scaling of feature values to a global range using MinMaxScaler and multiple nested transformations in twin FeatureUnion layers. A binarizer (threshold = 0.856) and PCA (retaining approximately 51.3% variance) were combined alongside the preserved raw features using a passthrough for optimal representation of the features. At its end, the pipeline arrived at an optimized XGBoost Classifier with regularized hyperparameters (learning_rate = 0.0354, max_depth = 10, subsample = 0.7435, gamma = 0.0154, reg_alpha = 0.00545, reg_lambda = 0.0499, and n_estimators = 100) that mostly finds a compromise between diversity of features, control over complexity, and generalization of prediction.

The findings confirm the great gain of working with combined datasets, especially when using One-Hot Encoding, in obtaining better model performance.

Using TPOT pipelines, selective feature selection and dimensionality reduction in the context of raw signal preservation and ensemble modeling were combined to effectively optimize the final predictive accuracy.

To showcase the state of our new work, we compare our experimental setup and results with the RiceSNP-ABST study [14], one of the latest works for RT-

**Table 2.** Conceptual and performance comparison between TPOT ML pipeline and CNN-based RiceSNP-ABST models.

| Aspect | RiceSNP CNN-Based Model | TPOT ML Pipeline (Proposed Experiment) | Notes |
|---|---|---|---|
| Model Architecture | Automated CNN with residual connections optimized via ENAS | Automated TPOT ML pipelines (XGBoost, Bagging) | CNN vs ML pipeline; CNN captures spatial patterns explicitly. |
| Datasets | Six training datasets: Random_41, Common_41, Rearrange_41, Random_101, Common_101, Rearrange_101 with two testing datasets (Independent and cross-species). | Eight datasets: Random_40, Common_40, Rearrange_40, Random_100, Common_100, Rearrange_100, Combined_40, Combined_100 (80% for training and 20% for testing). | Multiple datasets used; your method explicitly analyzes encoding strategies. |
| Preprocessing | Custom preprocessing | Automated within pipeline | Proposed approach enables full automation. |
| Feature Handling | Learns features from raw data | Selects and transforms features via TPOT | Manual vs automated feature processing. |
| Best Achieved Accuracy | ACC = 0.6133, AUC = 0.6633 (trained on rearrange_101nt) | ACC = 0.937, AUC = 0.937 (Combined_100, One-Hot encoding) | Proposed model significantly outperforms RiceSNP-ABST. |
| Individual Dataset Performance | ACC = 0.613 (rearrange_101) | ACC = 0.628 (rearrange_100, One-Hot and DNA2Vec feature vector) | Slightly better individual performance as well. |

based trait prediction in rice. In contrast to RiceSNP-ABST, which emphasizes the use of SNP characteristics in attention-based approaches to classification, our experimental design features alternative encoding mechanisms and model architectures designed for improved interpretability and performance. The comparison is made on various factors such as the dataset used, the characteristics represented, the models chosen, the evaluation metrics calculated, and the performance achieved, as shown in Table 2.

## 4 Conclusions and Future Work

This research presents a comparative evaluation of DNA sequence encoding when used to detect the ABST SNPs in rice. The prediction models were built using TPOT auto machine learning approach, reaching 93.7% with XGBoost model trained on the combined dataset with a sequence length 100 represented with One-Hot encoding.

In general, combined datasets with the two DNA sequence lengths achieved better results. In other words, encoding methods and machine learning algorithms achieve higher accuracy with larger datasets. This is promising situation when analyzing the big genomic data.

Future research should introduce more encoding techniques and larger datasets to uncover the generality of these findings across various crop traits. In particular, integrating these bioinformatics strategies establishes a strong foundation for further genetic work and application for trait prediction, breeding program design, and trait-based data-driven agricultural recommendations in rice and other key crops.

Locating SNPs in DNA rice sequences is similar to the regular motif finding problem in the bioinformatics field. As for future work, motif finding methods or algorithms can be applied not only for predicting SNPs but also for finding their exact locations in the DNA sequences, such as SMF [1] and qSMF [2] algorithms. In our opinion, this can help researchers or breeders to understand which motif(s) (transcription factor binding site(s)) bind to which gene(s) (which affect associated traits).

## References

1. Al-Shaikhli, H., de Doncker, E.: SMF: approximate algorithm for the planted (l, d) motif finding problem in DNA sequences. In: Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), pp. 123–129. The Steering Committee of The World Congress in Computer Science, Computer ... (2018)
2. Al-Shaikhli, H., de Doncker, E.: qSMF: an approximate algorithm for Quotum planted motif search on ChIP-Seq data. In: 2019 IEEE International Conference on Electro Information Technology (EIT), pp. 434–440. IEEE (2019)
3. Choong, A.C.H., Lee, N.K.: Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method. In: 2017 International Conference on Computer and Drone Applications (IConDA), pp. 60–65 (2017)

4. Dominic, N., Cenggoro, T.W., Budiarto, A., Pardamean, B.: Deep polygenic neural network for predicting and identifying yield-associated genes in Indonesian rice accessions. Sci. Rep. **12**(1), 13823 (2022)
5. Farooq, M.A., et al.: Artificial intelligence in plant breeding. Trends in Genetics (2024)
6. Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N., Paterson, A.H.: An SNP resource for rice genetics and breeding based on subspecies Indica and japonica genome alignments. Genome Res. **14**(9), 1812–1819 (2004)
7. Gill, M., et al.: Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. BMC Plant Biol. **22**(1), 180 (2022)
8. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
9. Huang, J., Li, Z., Zhang, J.: Research on plant genomics and breeding (2023)
10. Koh, E., Sunil, R.S., Lam, H.Y.I., Mutwil, M.: Confronting the data deluge: How artificial intelligence can be used in the study of plant stress. Computational and Structural Biotechnology Journal (2024)
11. Koh, E., Sunil, R.S., Lam, H.Y.I., Mutwil, M.: Harnessing big data and artificial intelligence to study plant stress. arXiv preprint arXiv:2404.15776 (2024)
12. Le, T.T., Fu, W., Moore, J.H.: Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics **36**(1), 250–256 (2020)
13. Liu, C.G., Zhang, G.Q.: Single nucleotide polymorphism (SNP) and its application in rice. Yi Chuan= Hereditas **28**(6), 737–744 (2006)
14. Lu, Q., et al.: RiceSNP-ABST: a deep learning approach to identify abiotic stress-associated single nucleotide polymorphisms in rice. Briefings Bioinform. **26**(1), bbae702 (2025)
15. Maher, N., Yousif, S.A.: An automated machine learning model for diagnosing COVID-19 infection. IAES Int. J. Artif. Intell. (IJ-AI) **12**(3), 1360–1369 (2023)
16. Matti, R., Yousif, S.A.: AutoKeras for fake news identification in Arabic: leveraging deep learning with an extensive dataset. Al-Nahrain J. Sci. **26**(3), 60–66 (2023)
17. Ng, P.: dna2vec: Consistent vector representations of variable-length k-mers. arXiv preprint arXiv:1701.06279 (2017)
18. Panahi, B., Hamid, R., Jalaly, H.M.Z.: Deciphering plant transcriptomes: leveraging machine learning for deeper insights. Curr. Plant Biol. 100432 (2024)
19. Ribeiro, P., et al.: Tpot2: A new graph-based implementation of the tree-based pipeline optimization tool for automated machine learning. In: Genetic Programming Theory and Practice XX, pp. 1–17. Springer (2024)
20. Tong, K., et al.: PlantMine: a machine-learning framework to detect core SNPs in rice genomics. Genes **15**(5), 603 (2024)
21. Vourlaki, I.T., Ramos-Onsins, S.E., Pérez-Enciso, M., Castanera, R.: Evaluation of deep learning for predicting rice traits using structural and single-nucleotide genomic variants. Plant Methods **20**(1), 121 (2024)
22. Wang, H., et al.: Cropformer: an interpretable deep learning framework for crop genomic prediction. Plant Communications (2024)
23. Xu, J., et al.: RiceSNP-BST: a deep learning framework for predicting biotic stress–associated SNPs in rice. Briefings Bioinform. **25**(6), bbae599 (2024)
24. Yousif, S.A., Samawi, V.W.: The effectiveness of using AutoML in electricity theft detection: the impact of data preprocessing and balancing techniques. In: International Conference on Computational Science and Its Applications, pp. 68–82. Springer (2024)

25. Zhao, H., et al.: RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res. **43**(D1), D1018–D1022 (2015)
26. Zhao, L., Walkowiak, S., Fernando, W.: Artificial intelligence: a promising tool in exploring the phytomicrobiome in managing disease and promoting plant health. Plants **12**(9), 1852 (2023)