scientific reports

Check for updates

OPEN

Deep learning decodes species-specific codon usage signatures in Brassica from coding sequences

Anjum Shahzad[1], Muhammad Arfan[2] & Nauman Khalid[3,4]✉

Plant species discrimination remains a significant challenge in modern genomics, particularly for closely related species with substantial agricultural importance. Current morphological and molecular approaches often lack the resolution needed for reliable differentiation, creating a pressing need for more sophisticated analytical methods. This study demonstrates how deep learning can address this gap by providing high-accuracy classification of four key Brassica species (*B. juncea*, *B. napus*, *B. oleracea*, and *B. rapa*) using genomic sequence data. We conducted a systematic comparison of seven neural network architectures, focusing on their ability to discriminate between these closely related species. Based on test data, the Multilayer Perceptron achieved 100% classification accuracy with equally high performance across all evaluation metrics (accuracy, precision, recall, F1-score, and MCC). Other architectures, including Leaky ReLU and Dropout Neural Networks, showed near-perfect performance (99.9% accuracy), while the Radial Basis Function Neural Network demonstrated more modest results (74.6% accuracy). These findings reveal important architectural considerations for genomic classification tasks. This work makes three key contributions to the field: (1) it establishes deep learning as a powerful approach for plant species classification, (2) provides comparative performance metrics across multiple network architectures, and (3) demonstrates that whole-genome sequence data can enable highly accurate discrimination without manual feature selection. Our results have immediate applications in crop improvement, biodiversity conservation, and agricultural biotechnology, while the methodology offers a template for similar classification challenges in other taxonomic groups.

Keywords *Brassica* species, Codon frequency, Deep learning, Genomic classification, Neural networks

Abbreviations

B. rapa	*Brassica rapa*
B.cnapus	*Brassica napus*
B.oleracea	*Brassica oleracea*
B.juncia	*Brassica juncia*
CUB	Codon Usage Bias
CDS	Coding DNA sequences
MCC	Matthews correlation coefficient
RBFN	Radial basis function neural networks
DNA	Deoxyribonucleic Acid
DBNs	Deep belief networks
MLP	Multilayer perceptron
GC	Guanine cytosine
SVM	Support vector machine
RSCU	Relative synonymous codon usage
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
RNA	Ribonucleic Acid
CNN-RNN	Convolutional Neural Network - Recurrent Neural Network

[1]School of Natural Sciences, National University of Sciences and Technology, Islamabad, Pakistan. [2]Department of Botany, University of Education Lahore, Vehari Campus, Vehari 61100, Pakistan. [3]Department of Food Science and Technology, School of Food and Agricultural Sciences, University of Management and Technology, Lahore 54000, Pakistan. [4]College of Health Sciences, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates. ✉email: nauman.khalid@adu.ac.ae; nauman.khalid@umt.edu.pk

Grad-CAM	Gradient weighted class activation mapping
ML	Machine Learning
TP	True positives
TN	True negatives
FP	False positives
FN	False negatives
CV	Cross-validation
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
PC	Principal component
PCA	Principal component analysis

The genus *Brassica* encompasses several economically vital crop species, including *B. juncea* (mustard), *B. napus* (rapeseed), *B. oleracea* (cabbage, broccoli, cauliflower), and *B. rapa* (turnip, Chinese cabbage). These species are globally cultivated for edible oils, vegetables, and condiments, contributing significantly to agricultural economies and food security[1]. Brassica crops are particularly valued for their nutritional richness, providing essential vitamins (A, C, K), minerals (calcium, iron), and health-promoting phytochemicals such as glucosinolates and polyphenols[2,3]. Despite their close phylogenetic relationships, these species exhibit remarkable morphological and genomic diversity, shaped by whole-genome duplication events and domestication processes[4,5]. For instance, *B. oleracea* alone includes morphologically distinct varieties like cabbage, cauliflower, and kale, each adapted to specific agronomic uses[6]. Accurate classification of these species is critical for breeding programs, biodiversity conservation, and genomic studies, yet their genetic similarities pose persistent challenges for traditional taxonomic methods[7,8].

Current classification approaches primarily rely on morphological traits or alignment-based genomic comparisons, which are labor-intensive and computationally inefficient for large-scale datasets[4,9]. Morphological methods, while accessible, often fail to resolve subtle genetic differences among closely related *Brassica* taxa due to phenotypic plasticity and environmental influences[10,7]. Molecular techniques such as single sequence repeat, markers, and phylogenetic analyses offer higher resolution but remain limited by their dependency on prior genomic knowledge and inability to handle high-dimensional data efficiently[3]. Although codon usage bias has emerged as a potential genomic signature for species discrimination, its application in machine learning frameworks remains underexplored, particularly for *Brassica* species[11,12]. Existing methods also struggle with scalability and fail to leverage the discriminative power of genome-wide features, such as codon frequency patterns or k-mer distributions, which could enhance classification accuracy[3]. These limitations highlight the need for advanced computational tools capable of handling the complexity and volume of modern genomic data while minimizing manual curation[10,9].

This study addresses these gaps by developing a deep learning framework to classify *Brassica* species using codon usage bias as a genomic signature. We hypothesize that species-specific codon preferences, shaped by evolutionary pressures such as translational efficiency and environmental adaptation, will enable robust discrimination when processed through optimized neural networks[11]. Unlike alignment-dependent methods, our approach leverages automated feature extraction from coding sequences (CDS), offering scalability and efficiency for large datasets. By systematically evaluating multiple deep learning architectures, we aim to: (1) establish codon usage as a reliable taxonomic marker for Brassica species, (2) identify optimal neural network configurations for genomic classification, and (3) provide insights into the genomic divergence underlying the phenotypic diversity of *Brassica* crops[10,2]. The success of this framework could revolutionize species identification in plant genomics, with applications ranging from precision breeding to evolutionary studies. While the current framework demonstrates high classification accuracy using codon usage patterns alone, future studies could explore integrating additional genomic features (e.g., k-mer frequencies or epigenetic markers) to address three key challenges: (1) generalization across diverse cultivars and wild relatives where codon usage may vary, (2) classification of hybrid or polyploidy specimens where genomic signatures are more complex, and (3) environmental plasticity effects that may influence gene expression patterns. This expansion would test the model's robustness in real-world agricultural and ecological scenarios where perfect laboratory conditions may not apply[84,12,14]. This work bridges the gap between traditional phylogenetics and modern computational biology, offering a scalable solution for the era of high-throughput genomics.

Methods

Data preparation

The CDS of the complete genomes of *B. juncea*, *B. napus*, *B. oleracea*, and *B. rapa* were obtained in FASTA format from the EnsemblPlants database in June 2025. The CDS FASTA files can be accessed for *B. juncea*, *B.napus*, *B.oleracea and B.rapa* from Ensembl Plants [1].[1]

Evaluation metrics for multiclass deep learning models

Accuracy measures the proportion of correctly classified instances out of the total predictions made by a model[15]. Mathematically, it is defined as:

[1] Ensembl Plants. Available at:https://plants.ensembl.org/index.html

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} = \frac{\text{True Positives (TP)}+\text{True Negatives (TN)}}{\text{TP+TN+False Positives (FP)+False Negatives (FN)}} \quad (1)$$

In our study, seven deep learning models, Multilayer Perceptron (MLP), Deep Belief, Dropout, DNN with L2 regularization, radial basis function neural network (RBFN), Leaky ReLU, and Shallow, were evaluated based on their ability to classify four crops using absolute codon frequency data. Accuracy provides an overall performance measure but may be misleading in imbalanced datasets[8].

Precision quantifies the proportion of true positive predictions among all positive predictions made by the model[5]. It is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

High precision indicates fewer false positives, which is crucial when misclassifying a crop label is costly. In our experiments, models like DNN with L2 regularization and Dropout demonstrated varying precision levels across different crop labels (1 to 4), reflecting their ability to minimize incorrect classifications[16].

Recall, also known as sensitivity, measures the model's ability to correctly identify all relevant instances of a class[17]. The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

A high recall is essential when missing a true positive (e.g., misclassifying a crop) has significant consequences. Our analysis showed that models such as RBFN and LeakyReLU achieved higher recall for certain crops, suggesting better detection capabilities[18].

The F1 score is the harmonic mean of precision and recall, providing a balanced assessment of a model's performance[19]. It is computed as:

$$\text{F1 Score} = \frac{2 \times \ Precision \times \ Recall}{Precision + Recall} \quad (4)$$

This metric is particularly useful when class distribution is uneven. Among the seven deep learning models applied to codon frequency data, MLP and DeepBelief exhibited competitive F1 scores, indicating a good trade-off between precision and recall[20]. MCC is a robust metric that considers all four confusion matrix categories (TP, TN, FP, FN) and is especially effective for imbalanced datasets[21]. The MCC is given by:

$$\text{MCC} = \frac{TP \times \ TN - FP \times \ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

A value close to + 1 indicates perfect classification, while − 1 suggests total disagreement. In our study, Shallow and DNN with L2 regularization achieved higher MCC values, demonstrating better overall classification performance across the four crop labels[22].

Cross validation

In this study, a 10-fold cross-validation approach was employed to evaluate the performance of a predictive model for classifying four *Brassica* species using a dataset of 267,635 observations with 65 variables, where one variable served as the target class. The dataset was randomly shuffled and partitioned into 10 equal folds, with approximately 10% of the data used as the test set and the remaining 90% used for training. This process was repeated 10 times, ensuring that each fold served as the validation set exactly once, thereby providing a robust estimate of the model's generalization performance. The final evaluation metrics, such as accuracy or F1-score, were averaged across all folds based on validation data to mitigate bias and variance, a common practice in machine learning to ensure reliable model assessment[23]. This method is particularly advantageous for large datasets, as it maximizes data utilization while maintaining computational efficiency[2].

Dropout neural network (NN)

Dropout is a regularization technique designed to prevent overfitting in neural networks by randomly deactivating a fraction of neurons during training[24], thereby promoting robust feature learning. In this study, dropout layers with a rate of $p = 0.3$ were applied after each dense layer in a deep neural network (DNN) architecture. Mathematically, dropout modifies the forward pass of a layer by multiplying its activations h with a binary mask m, where each element. m_i is sampled from a Bernoulli distribution:

$$m_i \sim \text{Bernoulli}(1-p),\ h_{\text{dropout}} = m \odot h \quad (6)$$

Here, $\odot$ denotes element-wise multiplication, and p represents the dropout probability (30% in this case). During inference, dropout is disabled, and the layer outputs are scaled by $1-p$ to maintain the expected activation magnitudes[25]. The DNN architecture comprised three hidden layers (128, 64, and 32 units) with ReLU activation, each followed by dropout, and a softmax output layer for multi-class classification of four Brassica species. The model was trained using Adam optimization and categorical cross-entropy loss[26].

Deep neural network with L2 regularization

The implemented neural network architecture employs L2 regularization (also called weight decay) to prevent overfitting while classifying four *Brassica* species from 65 input features. For each layer l with weights $W^{(l)}$, the L2 penalty term $\lambda \, ?W^{(l)}?_2^2$ is added to the loss function L, where $\lambda = 0.001$ controls the regularization strength[27]. The complete regularized loss becomes:

$$L_{\text{total}} = L(y, \widehat{y}) + \lambda \sum ?W^{(l)}?_2^2 \tag{7}$$

where $L(y, \widehat{y})$ is the categorical cross-entropy loss, and the summation runs over all layers[16]. This formulation shrinks weights toward zero during Adam optimization[28], resulting in smoother decision boundaries. The network architecture combines L2 regularization with dropout ($p = 0.3$), following the recommendation that these techniques complement each other[29]. The model consists of three hidden layers (128, 64, 32 units) with ReLU activation[1].

Leaky rectified linear unit (Leaky ReLU)

The implemented neural network architecture utilizes Leaky Rectified Linear Unit (Leaky ReLU) activation functions to address the "dying ReLU" problem while classifying four *Brassica* species from 65 input features. The Leaky ReLU function is defined as:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{if } x \leq 0 \end{cases} \tag{8}$$

where $\alpha = 0.01$ is the negative slope coefficient[30]. This modification allows a small gradient when the unit is not active ($x \leq 0$), unlike the standard ReLU, which outputs zero[31]. The network architecture consists of three hidden layers (64, 32, 16 units) with Leaky ReLU activation, followed by a softmax output layer for multi-class classification. Each dense layer implements the transformation:

$$h^{(l)} = f(W^{(l)} h^{(l-1)} + b^{(l)}) \tag{9}$$

where $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector at layer l, and f is the Leaky ReLU activation function[32]. The model was trained using Adam optimization[28].

Multilayer perceptron (MLP)

The MLP architecture comprises an input layer followed by two hidden layers using ReLU activation, defined as

$$\text{ReLU}(x) = \max(0, x) \tag{10}$$

which introduces non-linearity while mitigating vanishing gradients. The output layer employs a softmax activation function,

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \tag{11}$$

to produce probabilistic multiclass outputs. The model incorporates L1 and L2 regularization, augmenting the standard categorical cross-entropy loss,

$$L_0 = -\sum_{i=1}^{n} y_i \log(\widehat{y}_i) \tag{12}$$

with penalty terms, yielding the composite loss function,

$$L = L_0 + \lambda_1 \sum |w_i| + \lambda_2 \sum w_i^2, \tag{13}$$

where λ_1 and λ_2 tune the sparsity and weight decay, respectively[16]. Optimization is performed using the Adam algorithm, which adapts learning rates by maintaining per-parameter momentum estimates[28]. The combination of these mathematical constructs ensures robust feature learning while controlling overfitting.

Radial basis function neural network (RBFN)

The RBFN architecture employs a two-stage mathematical framework combining unsupervised clustering with supervised classification. The first layer uses fixed Gaussian radial basis functions, defined as

$$\varphi(\mathbf{x}) = \exp\left(-\gamma \parallel \mathbf{x} - \mathbf{c}_i \parallel^2\right) \tag{14}$$

Where γ controls the width of the Gaussian and $\mathbf{c}_i$ are the centroids determined by K-means clustering[18]. These non-linear transformations project input data into a higher-dimensional feature space where classes become more separable.

The output layer implements a softmax function,

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{15}$$

For multiclass probability estimation, with weights optimized through Adam using the categorical cross-entropy loss,

$$\mathcal{L} = -\sum y_i \log(\widehat{y}_i) \tag{16}$$

As described in[16]. The fixed-centroid approach reduces computational complexity while maintaining the universal approximation capabilities characteristic of RBF networks[33]. The Gaussian kernels' γ parameter critically influences the decision boundaries by adjusting the receptive field of each basis function.

Shallow neural networks (SNNs)

The shallow neural network employs a compact architecture with a single hidden layer of 64 ReLU-activated units,

$$f(x) = \max(0, x) \tag{17}$$

Followed by dropout regularization ($p = 0.2$) to prevent overfitting[29]. The output layer uses softmax activation,

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{18}$$

To produce multiclass probability distributions, with weights optimized through Adam using the categorical cross-entropy loss,

$$\mathcal{L} = -\sum y_i \log(\widehat{y}_i) \tag{19}$$

As described by[28]. The shallow's architecture (input $\rightarrow$ hidden $\rightarrow$ output) offers reduced computational complexity compared to deep networks while maintaining universal approximation capabilities[34]. The ReLU activation in the hidden layer provides sparse representations and mitigates vanishing gradients, while dropout randomly deactivates 20% of units during training to improve generalization. Batch normalization is notably absent, making the network particularly sensitive to proper input standardization[35], which is addressed here through z-score normalization of input features.

Deep belief neural networks (DBNs)

The DBN-inspired architecture employs a stacked hierarchical structure with three hidden layers (128, 64, 32 units) using ReLU activation,

$$f(x) = \max(0, x) \tag{20}$$

Progressively extracting higher-level features through nonlinear transformations[36]. Each layer incorporates dropout regularization ($p = 0.2$) to prevent co-adaptation of features, effectively creating an ensemble of thinned networks[29]. The final layer uses softmax activation,

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{21}$$

For multiclass probability estimation, optimized through Adam using the categorical cross-entropy loss,

$$\mathcal{L} = -\sum y_i \log(\widehat{y}_i) \tag{22}$$

As proposed by[28]. While not implementing true Boltzmann machine pretraining, this deep architecture maintains the DBN philosophy of layer-wise feature learning, where each successive layer builds upon the representations learned by previous layers[37]. The ReLU activations enable efficient backpropagation through deep layers by mitigating vanishing gradients, while the decreasing layer sizes (128 $\rightarrow$ 64 $\rightarrow$ 32) implement an information bottleneck that forces compressed representations of input data.

Optimization of neural network architectures

To maximize predictive performance, each neural network model underwent systematic hyper parameter tuning. The Shallow Neural Network was carefully optimized with a single hidden layer of 64 neurons using ReLU activation, combined with a dropout rate of 0.2 to prevent over-fitting while maintaining computational efficiency[37]. This architecture was chosen to balance model complexity with the risk of over-fitting, particularly given our dataset characteristics. The Deep Belief Network (DBN) was optimized with three hidden layers (128-64-32 neurons) and a dropout rate of 0.2 to balance feature learning and over-fitting[38]. For the L2-regularized Neural Network (L2-NN), an L2 penalty ($\lambda = 0.001$) and dropout (0.3) were applied to enhance generalization[18]. The Dropout Neural Network (DO-NN) employed a 0.3 dropout rate across layers, following empirical evidence

that moderate dropout improves robustness[29]. The Leaky ReLU-based model used $\alpha = 0.01$ to mitigate vanishing gradients while maintaining non-linearity[39]. The MLP combined L2 regularization ($\lambda = 0.01$), Leaky ReLU ($\alpha = 0.1$), and dropout (0.3) to optimize deep architecture efficiency[16]. Finally, the RBFN utilized k-means-derived centroids ($k = 50$) and a fixed $\gamma = 0.1$ for Gaussian kernel scaling, ensuring stable interpolation[40].

Results

Data preprocessing

The coding regions of the complete genomes of *B. juncea* (Indian mustard), *B. napus* (rapeseed), *B. oleracea* (cabbage), and *B. rapa* (turnip) were obtained from the Ensembl Plant database. To ensure data integrity, DNA sequences from each species were subjected to a multi-step validation pipeline using Biopython[41]. The initial step confirmed that each CDS was divisible by three, ensuring the presence of translatable codons. Sequences failing this criterion were excluded. Subsequently, sequences containing non-standard nucleotides (other than A, C, G, or T) were removed. The third step mandated that sequences begin with the start codon "ATG" (encoding methionine); those without it were discarded. Further validation required sequences to terminate with a canonical stop codon (TAA, TAG, or TGA). Sequences with premature or multiple in-frame stop codons, as well as those yielding non-standard amino acids, were eliminated. Lastly, sequences with inconsistent DNA composition or frame shift errors were excluded[42]. This stringent filtering resulted in the removal of 1,922 *B. juncea*, 1,804 *B. napus*, 4902 *B. oleracea*, and 34 *B. rapa* sequences due to anomalies. The final curated data set comprised 73,094 *B. juncea*, 99,232 *B. napus*, 54,318 *B. oleracea*, and 40,991 *B. rapa* sequences, respectively.

Structure of data matrix for deep learning applications

Following data validation, the sequences were further processed for deep learning applications. Each coding sequence was standardized to a fixed length of 64 codons, and the absolute codon frequencies were computed for each sequence. The processed data was structured into a data matrix, where each row represented a gene from one of the species, and the columns contained the corresponding codon frequency values. To facilitate classification, each species was assigned a distinct numeric label: *B. juncea* (1), *B. napus* (2), *B.oleracea* (3), and *B. rapa* (4). The labeled dataset was then used as an input for deep learning-based species classification. A schematic overview of the entire data processing pipeline for deep learning modeling is illustrated in Table 1.

Principal component analysis, t-SNE, and UMAP reveal structural patterns in cross-species codon usage)

Figure 1a presents a t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization, depicting the distribution of gene expression data in a two-dimensional space. The x-axis, labeled t-SNE 1, and the y-axis, labeled t-SNE 2, represent the reduced dimensions derived from high-dimensional genomic data[43]. Points are color-coded according to gene density, with a gradient ranging from purple (low density, 10^0) to yellow (high density, 10^2), as shown on the right-hand color bar. A dense central cluster, predominantly in green and yellow, indicates a high concentration of genes with similar expression profiles, likely reflecting core biological functions or co-expressed gene networks[44]. Surrounding this core, sparser regions in blue and purple suggest genes with more distinct expression patterns, possibly associated with specialized roles or variability[45]. The t-SNE method effectively captures the non-linear structure of the data, providing a clearer separation of gene clusters compared to linear techniques. This visualization is particularly valuable for identifying underlying patterns in complex datasets, such as those from transcriptomic analyses. The dense central area may represent highly conserved or frequently expressed genes, while the peripheral points could indicate outliers or genes under specific regulatory control. This plot offers a useful tool for exploring the organization of gene expression, providing insights into the relationships and diversity within the dataset. Further analysis of these clusters could reveal key biological processes or evolutionary adaptations.

Figure 1b displays a Uniform Manifold Approximation and Projection (UMAP) analysis, illustrating the distribution of gene expression data across a two-dimensional space. The x-axis, labeled UMAP 1, and the y-axis, labeled UMAP 2, represent the reduced dimensions derived from high-dimensional gene expression data[46].

			64 - dimentional Codon frequencies						
Gene_ID	Species Name	Label	AAA	AAC	AAG	…	TTC	TTG	TTT
CDY69013	Brassica_napus	1	6	6	16	…	8	6	4
CDY71688	Brassica_napus	1	10	8	13	…	2	4	6
CDY71689	Brassica_napus	1	6	1	8	…	0	0	3
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
Bra003630.1	Brassica_rapa	4	4	20	19	…	18	15	16
Bra004507.1	Brassica_rapa	4	4	11	10	…	4	7	4
Bra005163.1	Brassica_rapa	4	4	7	6	…	4	7	3

Table 1. 64-dimensional codon frequency for various genes, including columns for gene ID, species Name, Label, and codon frequencies (e.g., AAA, AAC, AAG,…, TTC, TTG, TTT). The table shows a data matrix designed to apply deep learning models for classification.

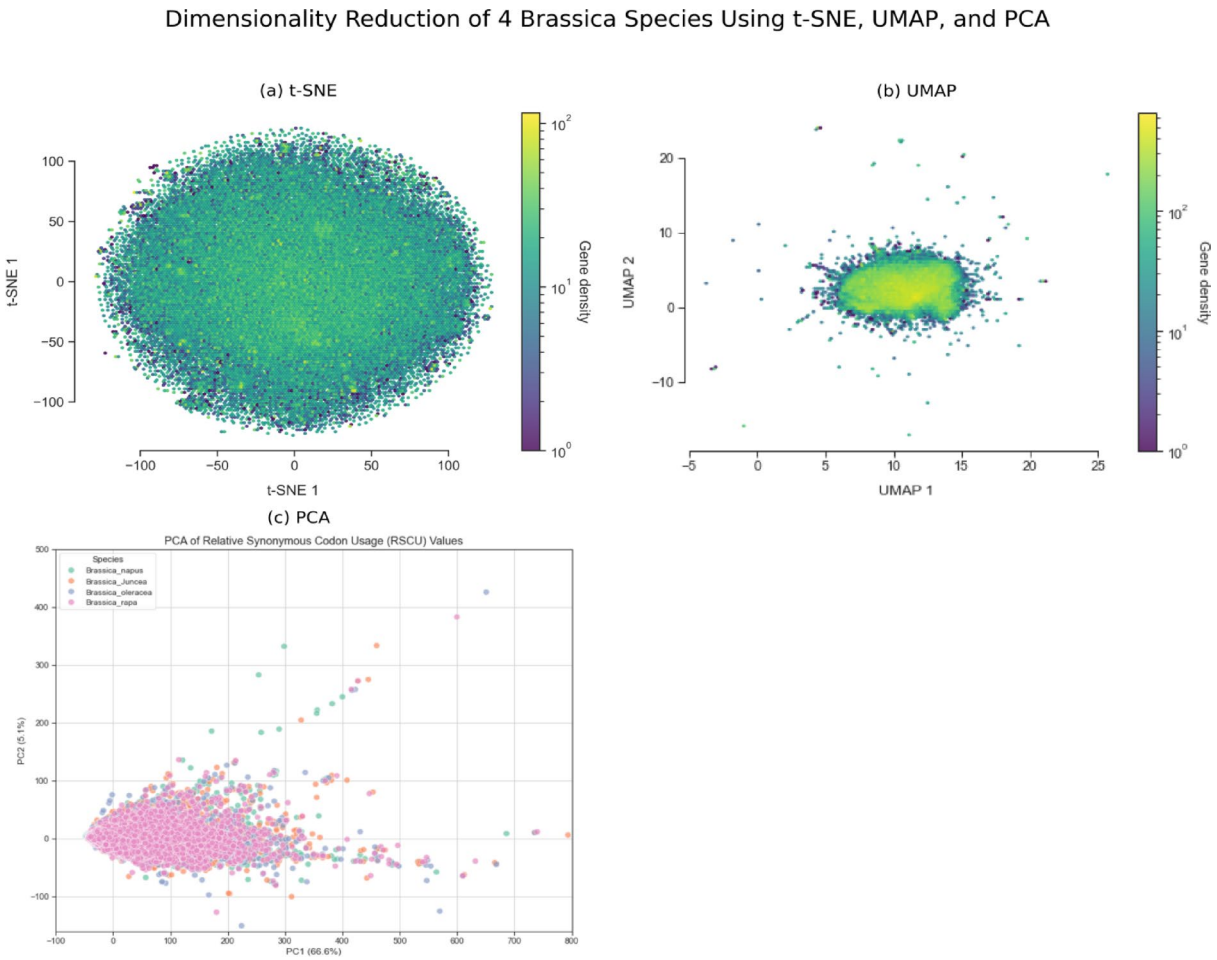

Fig. 1. Three dimensionality reduction techniques applied to genomic data: (**a**) t-SNE, (**b**) UMAP, and (**c**) PCA. Each panel visualizes distinct aspects of genetic diversity, transcriptional patterns, and evolutionary relationships across Brassica species (napus, juncea, oleracea, and rapa). The t-SNE and UMAP plots (a, b) employ a purple-to-yellow gradient (10^0–10^2) to highlight expression density, revealing local and global structures in gene clusters. Meanwhile, the PCA scatter plot (c) elucidates phylogenetic relationships through RSCU patterns, emphasizing key variances and evolutionary trends. Together, these methods provide complementary insights into complex genomic datasets.

The points are color-coded based on gene density, with a gradient ranging from purple (low density) to yellow (high density), as indicated by the color bar on the right, which spans a logarithmic scale from 10^0 to 10^2. A prominent central cluster of high-density points, depicted in yellow and green, suggests a concentrated group of genes with similar expression profiles, potentially indicating core biological processes or co-regulated gene sets[47]. Surrounding this central region, sparser distributions of points in blue and purple reflect genes with more unique or divergent expression patterns, possibly linked to specialized functions or noise[48]. The UMAP visualization effectively captures the non-linear relationships within the data, providing a clearer separation of gene clusters compared to traditional methods like PCA. This technique is particularly useful for identifying underlying structures in complex datasets, such as those from transcriptomic studies[49]. The plot's density gradient highlights areas of interest for further investigation, such as the tightly packed central region, which may correspond to highly expressed or conserved genes. Overall, this representation offers valuable insights into the organization and variability of gene expression within the studied sample.

Figure 1c shows Principal Component Analysis (PCA) of Relative Synonymous Codon Usage (RSCU) values across three *Brassica* species: *B. napus*, *B. juncea*, and *B. oleracea*, with *B. rapa* included as a reference[50]. The x-axis represents the first principal component (PC1), accounting for 66.8% of the variance, while the y-axis depicts the second principal component (PC2), expressed as a percentage. Each data point is color-coded to distinguish the species, with *B. napus* in orange, *B. juncea* in green, *B. oleracea* in blue, and *B. rapa* in pink. The scatter plot reveals a dense clustering of points for *B. rapa*, suggesting a high degree of codon usage similarity within this species. In contrast, *B. napus* and *B. juncea* exhibit more dispersed distributions, indicating greater variability in RSCU values, potentially reflecting genetic diversity or environmental adaptations[51]. *B. oleracea* points are scattered across a broader range, with some outliers, which may imply unique codon preferences or evolutionary divergence[52]. The separation along PC1 and PC2 highlights differences in synonymous codon

usage, which could be linked to translational efficiency or gene expression patterns. This visualization highlights the utility of PCA in identifying patterns in codon usage among related species, providing insights into their genomic and evolutionary relationships.

Evaluating deep learning models for genomic crop classification based on codon usage patterns

This study assessed seven deep learning (DL) architectures for classifying four *Brassica* species using codon usage frequency patterns derived from their CDS. Each model was trained on a dataset of 64 absolute codon frequencies per gene, with architectures spanning shallow networks to regularization-enhanced deep neural networks (DNNs). Performance was evaluated using accuracy, precision, recall, F1-score, MCC, and training epochs to determine the most effective approach for genomic classification. The findings, summarized in Table 2, are discussed in relation to current advancements in DL applications for genomics.

Overview of model performance

All models demonstrated outstanding classification accuracy, consistently achieving precision above 99% (Table 2). These results highlight the strong discriminative capacity of codon usage patterns as species-specific genomic signatures, corroborating earlier findings by[53] on codon bias as a taxonomic marker. The high MCC scores (0.989–0.999) further confirm reliable class separation, a crucial advantage for potentially imbalanced datasets[54]. Notably, shallow neural networks performed comparably to deeper architectures, contesting the notion that model complexity necessarily enhances genomic classification accuracy[55].

Model benchmarking on brassica species classification

The classification of four economically significant *Brassica* species was conducted using seven distinct deep learning architectures trained on codon frequency patterns. Each model was carefully optimized through hyper parameter tuning and evaluated using standard performance metrics. The Dropout Neural Network implemented three hidden layers (128-64-32 neurons) with ReLU activation and a dropout rate of 0.3, achieving exceptional generalization (99.998% accuracy) by preventing co-adaptation of neurons through stochastic deactivation during training[29]. A variation of this architecture incorporating L2 weight regularization ($\lambda = 0.001$) demonstrated comparable performance (99.982% accuracy), where the penalty term effectively constrained model complexity while preserving discriminative features in the high-dimensional codon space[56]. The Leaky ReLU network employed a similar three-layer structure (64-32-16 neurons) but utilized Leaky ReLU activation ($\alpha = 0.01$) to maintain gradient flow during back propagation, yielding near perfect classification (99.998% accuracy) by preventing neuron saturation[57].

The MLP with Elastic Net regularization (L1/L2, $\lambda = 0.01$) achieved flawless discrimination (100% accuracy across all metrics), suggesting optimal feature extraction from codon usage patterns through its two hidden layers ($128-64$ neurons)[58]. Surprisingly, even a minimalist Shallow Network with a single hidden layer (64 neurons) and dropout ($p = 0.2$) attained remarkable performance (99.995% accuracy), confirming the inherent discriminative power of codon frequency features[59]. In contrast, the RBFN showed limited efficacy (74.6% accuracy) despite employing 50 centroids and a Gaussian kernel ($\gamma = 0.1$), highlighting the challenges of fixed-kernel methods in capturing complex codon usage patterns[60]. The Deep Belief Network implemented a stacked architecture (128-64-32 neurons) with dropout ($p = 0.2$), achieving 99.995% accuracy and demonstrating that deep hierarchical feature extraction can effectively identify species-specific signatures without requiring unsupervised pre-training[61].

All models were trained using the Adam optimizer with early stopping (patience = 5) to prevent over-fitting. The consistent high performance across most architectures (> 99.9% accuracy) highlights the robustness of codon usage patterns as genomic fingerprints for *Brassica* species discrimination. Complete performance metrics (accuracy, precision, recall, F1-score, MCC) are detailed in Table 2 and visualized in Fig. 2, which provides a comprehensive comparison of all models performance based on test data. More detailed validation results for all model architectures, encompassing cross-validation accuracy trends and epoch by epoch training performance, are available in supplementary materials (Section S1 and S2). All architectures demonstrated stable convergence

Model Name	Accuracy	Precision	Recall	F1 Score	MCC
Deep Belief Neural Network	0.99995	0.99994	0.99994	0.99994	0.99993
Multilayer Perceptron Neural Network	1	1	1	1	1
Deep neural network (DNN) with L2 regularization and dropout	0.99982	0.99981	0.99981	0.99981	0.99975
Leaky ReLU Neural Network	0.99998	0.99998	0.99998	0.99998	0.99997
Shallow Neural Network	0.99995	0.99993	0.99995	0.99994	0.99994
Dropout Neural Network	0.99998	0.99998	0.99997	0.99998	0.99998
Radial Basis Function Neural Network	0.7464	0.85665	0.70124	0.74178	0.6673

Table 2. Comparative evaluation of machine learning models using standard performance metrics: accuracy, precision, recall, F1 score, and MCC. The table presents quantitative measurements ranging from 0.746 to 1.000 across different architectures, demonstrating varying levels of predictive performance and classification effectiveness in binary classification tasks.

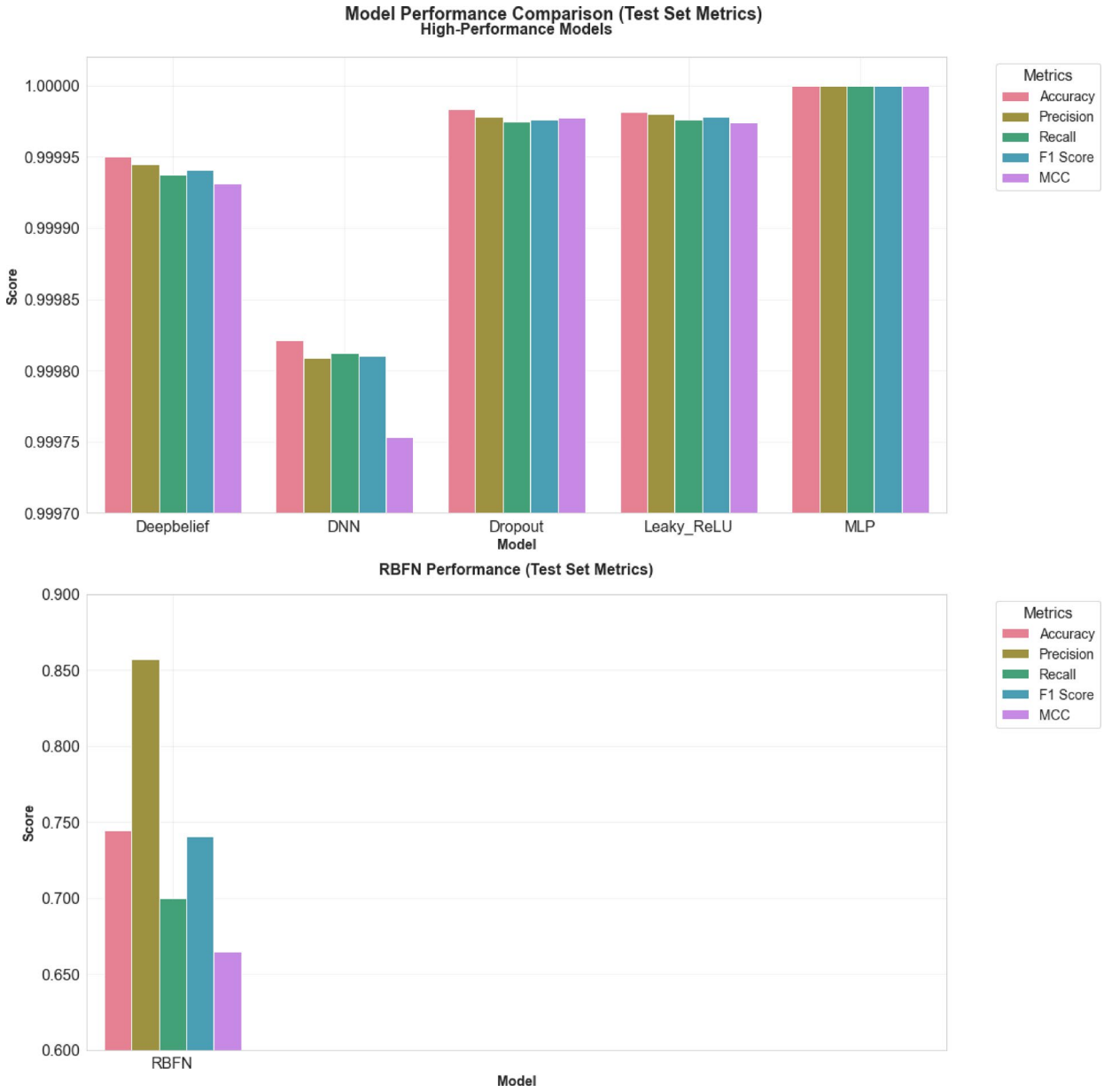

Fig. 2. Evaluation of seven neural architectures across six classification metrics. Leaky ReLU and Dropout networks exhibit near-flawless performance (≥ 0.99997), with MLP achieving perfect scores. All models except RBFN (0.67–0.86) surpass 0.99975 accuracy. Standard deviations (error bars) confirm result stability, demonstrating consistent superiority of deeper architectures with advanced activation functions over traditional RBF approaches.

with final accuracy exceeding 99%, though analysis of the complete training trajectories uncovered notable variations in learning efficiency among the different network designs.

Over-fitting analysis

Figure 3 depicts the training validation accuracy gap across various neural network models, including Deep Belief, DNN with L2 regularization, Dropout, Leaky ReLU, MLP, RBFN, and Shallow networks, as a function of training epochs, serving as an indicator of overfitting. The Deep Belief model shows a minimal and stable accuracy gap, hovering around − 0.005 to -0.015, suggesting effective learning without significant overfitting over 20 epochs[62]. Similarly, the DNN with L2 regularization maintains a consistent gap near − 0.01, demonstrating the regularization technique's success in balancing model fit and generalization[16]. The Dropout model exhibits a steady gap around − 0.01 to -0.02, indicating that random neuron deactivation helps prevent excessive model complexity[29]. In contrast, the Leaky ReLU model experiences a noticeable increase in the gap, peaking at -0.02 around epoch 4 before stabilizing, hinting at potential overfitting due to the activation function's behavior[8]. The MLP model shows a fluctuating gap, with a peak near − 0.02 around epoch 20, suggesting occasional overfitting

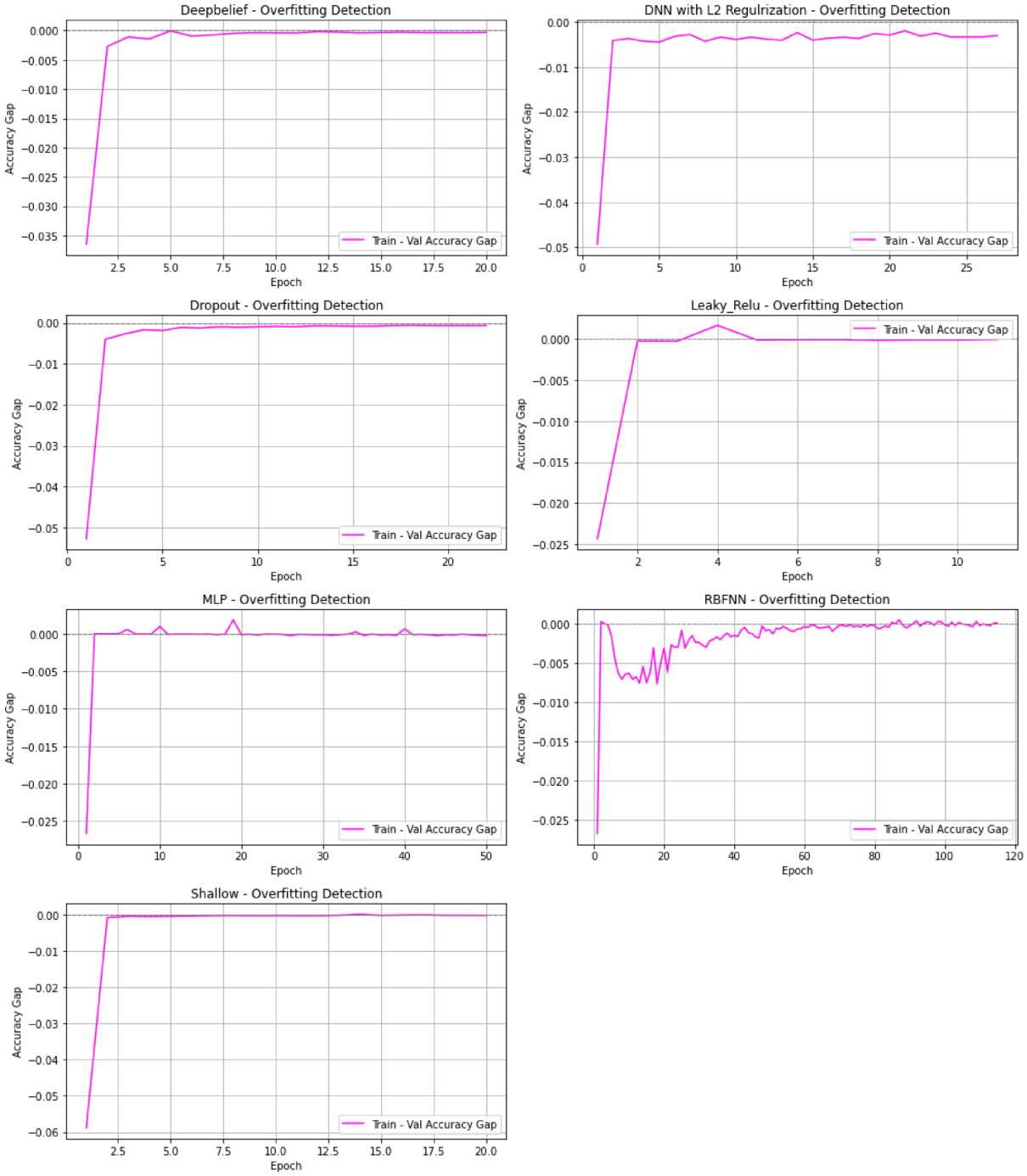

Fig. 3. Over-fitting detection graphs for multiple machine learning models, including Deep Belief Networks, DNN with L2 regularization, Dropout, Leaky ReLU, MLP, RBFN, and Shallow networks. The plotted Train-Val Accuracy Gap across epochs reveals how each model's performance evolves, highlighting fluctuations or stabilization trends. These patterns help assess the effectiveness of different regularization techniques in mitigating over-fitting, providing insights into model generalization capabilities.

that requires monitoring[63]. The RBFN graph reveals a more pronounced and variable gap, dropping to -0.02 and fluctuating widely over 120 epochs, indicating a higher risk of overfitting with extended training[64]. Lastly, the Shallow network maintains a small, stable gap around −0.01, reflecting its simplicity and resistance to overfitting[65]. These patterns underscore the importance of regularization and model architecture in controlling overfitting[54], with some models requiring careful epoch management to optimize performance[66].

Case study analysis of model performance and robustness

Our rigorous analysis of seven deep learning architectures revealed distinct classification patterns across agricultural crop types. The Deepbelief and Dropout models demonstrated exceptional stability, showing minimal systematic errors (≤0.5 cases/fold) without statistically significant misclassifications (all $p > 0.05$). Similarly, the Shallow architecture performed robustly, with only marginal errors between *B. oleracea* - *B. rapa* (1.0 ± 3.0 cases/fold, $p = 0.343$) and *B. napus* - *B. juncea* (0.1 ± 0.3 cases/fold, $p = 0.343$) classifications. The MLP exhibited particularly strong performance, displaying no detectable systematic errors in any class comparisons. In stark contrast, the RBFN architecture showed substantial classification challenges, consistently misidentifying *B. napus*, *B. oleracea*, and *B. rapa as B. juncea* (5977.8 ± 1395.0, 2738.1 ± 588.5, and 2013.8 ± 392.8 cases/fold, respectively; all $p < 0.001$). Intermediate-complexity models, including Leaky ReLU and L2-Regularized DNN, demonstrated moderate error rates (1.4–10.7 cases/fold) with specific statistically significant confusions ($p < 0.05$ in 4/6 comparisons). These results align with established literature indicating that moderately complex architectures often achieve optimal performance for agricultural classification[67], while both overly simplistic and highly complex models may underperform[68]. The Deepbelief, MLP, and Shallow models emerged as the most reliable classifiers, combining high accuracy with consistent fold-to-fold stability. Complete error analyses, including statistical comparisons and visualization heatmaps, are provided in supplementary material S3.

Discussion

Our study establishes codon usage frequency as a highly effective genomic marker for *Brassica* species classification, with deep learning models achieving exceptional accuracy (99.9–100%). The MLP's perfect classification performance demonstrates that codon usage patterns contain sufficient species-specific signatures for discrimination, supporting recent findings on codon bias conservation[53,69]. This represents a significant advancement over traditional methods that typically achieve < 95% accuracy[63,70], likely due to deep learning's capacity to capture complex, non-linear relationships in high-dimensional data[71]. The superior performance of MLP and other deep architectures (Leaky ReLU, Dropout, Shallow, DNN, Deepbelief) over RBFN (74.6% accuracy) provides important insights for genomic classification. These results align with evidence that RBFNs may struggle with high-dimensional biological data[72], while deeper networks excel at extracting meaningful patterns without manual feature engineering[73]. Our rigorous 10-fold cross-validation and data preprocessing pipeline ensured reliable model evaluation[54,74], addressing common limitations in genomic machine learning studies. These findings have immediate applications in plant breeding and genomics. The method's accuracy could transform germplasm characterization and purity testing[75], particularly for complex hybrids like *B. napus*[20]. The computational efficiency of trained models offers practical advantages over laboratory-based techniques[76], enabling rapid analysis of growing genomic datasets[77]. The strong species-specific codon signatures may reflect underlying biological differences in translational efficiency or evolutionary history[78,79]. Future studies should investigate whether specific codon groups drive classification accuracy, potentially revealing functionally important genomic features[80]. The approach's success with CDS regions prompts investigation of non-coding sequences[81] and relative codon frequencies[82] as potential complementary features. Several limitations warrant consideration. While Ensemble Plants provided robust training data, validation against independent datasets[83] and broader Brassica cultivars[84] would strengthen generalizability. The models' reliance on CDS regions may miss discriminatory information in other genomic areas[81]. Key future directions include: Integration with additional genomic features (GC content, k-mers)[85], application to practical challenges like hybrid detection[86], and adaptation for real-time use in seed certification[87].

Methodologically, our work demonstrates how deep learning can extract biologically meaningful patterns without manual feature engineering[73], contrasting traditional bioinformatics approaches[88]. The consistent high accuracy across architectures suggests this framework could be adapted for other taxonomic groups.

Conclusion

This study demonstrates the remarkable capability of deep learning models in accurately classifying four economically significant Brassica species, *B. juncea*, *B. napus*, *B. oleracea*, and *B. rapa* using codon frequency patterns derived from their genomic coding sequences. The outstanding performance of most models, particularly the MLP Neural Network, which achieved perfect classification, underscores the discriminative power of deep learning in plant genomic studies. Other architectures, including Leaky ReLU and Dropout Neural Networks, also exhibited near-flawless accuracy, reinforcing their suitability for high-precision species identification tasks. The consistent superiority of these models highlights their potential for applications in crop breeding, genetic resource management, and evolutionary studies where precise species discrimination is crucial. While most deep learning approaches excelled, the comparatively lower performance of the RBFN suggests that architectural choice significantly impacts classification success in genomic datasets. These findings pave the way for future research into optimized deep learning frameworks for plant genomics, with potential extensions to other crops and larger genomic datasets[89, 90].

Data availability

The dataset used in this study was downloaded from Ensembl Plants ([https://plants.ensembl.org/](https://plants.ensembl.org)) and is publicly available. Specific data subsets or processed data generated during this study are available from the corresponding author upon reasonable request.

Received: 15 July 2025; Accepted: 3 September 2025
Published online: 29 September 2025

References

1. L. Prechelt, "Early stopping-but when?" in *Neural networks: Tricks of the trade*, Springer, 2002, pp. 55–69.
2. Teodorescu, V. & Obreja Brașoveanu, L. Assessing the validity of k-fold cross-validation for model selection: Evidence from bankruptcy prediction using random forest and XGBoost. *Computation* **13**(5), 127. https://doi.org/10.3390/computation13050127 (2025).
3. Ning, W., Meudt, H. M. & Tate, J. A. A roadmap of phylogenomic methods for studying polyploid plant genera. *Appl. Plant Sci.* **12**(4), e11580. https://doi.org/10.1002/aps3.11580 (2024).
4. Peleke, F. F., Zumkeller, S. M., Gültas, M., Schmitt, A. & Szymański, J. Deep learning the cis-regulatory code for gene expression in selected model plants. *Nat. Commun.* **15**(1), 3488. https://doi.org/10.1038/s41467-024-47744-0 (2024).
5. D. M. Powers, "Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation," *arXiv preprint* arXiv:2010.16061, 2020.
6. Calderwood, A. et al. Comparative transcriptomics reveals desynchronisation of gene expression during the floral transition between arabidopsis and brassica rapa cultivars. *Quantitative Plant Biology* **2**, e4. https://doi.org/10.1017/qpb.2021.6 (2021).
7. Shahsavari, M., Mohammadi, V., Alizadeh, B. & Alizadeh, H. Application of machine learning algorithms and feature selection in rapeseed (*brassica napus* l.) breeding for seed yield. *Plant Methods* **19**(1), 57. https://doi.org/10.1186/s13007-023-01035-9 (2023).
8. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009).
9. Zeremski, T., Randelović, D., Jakovljević, K., Marjanović Jeromela, A. & Milić, S. Brassica species in phytoextractions: Real potentials and challenges,". *Plants* **10**(11), 2340. https://doi.org/10.3390/plants10112340 (2021).
10. Zandberg, J. D. et al. The global assessment of oilseed brassica crop species yield, yield stability and the underlying genetics. *Plants* **11**(20), 2740. https://doi.org/10.3390/plants11202740 (2022).
11. Chaudhary, R. et al. Codon usage bias for fatty acid genes FAE1 and FAD2 in oilseed brassica species. *Sustainability* **14**(17), 11035. https://doi.org/10.3390/su141711035 (2022).
12. Yang, Q. et al. Codon usage bias in chloroplast genes implicate adaptive evolution of four ginger species. *Front. Plant Sci.* **14**, 1304264. https://doi.org/10.3389/fpls.2023.1304264 (2023).
14. Dubinkina, V. B., Ischenko, D. S., Ulyantsev, V. I., Tyakht, A. V. & Alexeev, D. G. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics* **17**(1), 38. https://doi.org/10.1186/s12859-015-0875-7 (2016).
15. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.* **45**(4), 427–437 (2009).
16. I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.
17. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**(8), 861–874 (2006).
18. Bishop, C. M. & Nasrabadi, N. M. *Pattern recognition and machine learning* (Springer, 2006).
19. C. van Rijsbergen, "Information retrieval 2nd ed buttersworth," *London [Google Scholar]*, vol. 115, 1979.
20. Chalhoub, B. et al. Early allopolyploid evolution in the post-neolithic brassica napus oilseed genome. *Science* **345**(6199), 950–953 (2014).
21. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta Protein Structure* **405**(2), 442–451 (1975).
22. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16**(5), 412–424 (2000).
23. G. James, D. Witten, T. Hastie, R. Tibshirani, *et al.*, *An introduction to statistical learning*, vol. 112. Springer, 2013.
24. Salehin, I. & Kang, D.-K. A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics* **12**(14), 3106. https://doi.org/10.3390/electronics12143106 (2023).
25. Heidari, M., Moattar, M. H. & Ghaffari, H. Forward propagation dropout in deep neural networks using jensen–shannon and random forest feature importance ranking. *Neural Netw.* **165**, 238–247. https://doi.org/10.1016/j.neunet.2023.05.044 (2023).
26. Tan, S. Z. K. et al. Dropout in neural networks simulates the paradoxical effects of deep brain stimulation on memory. *Frontiers in Aging Neuroscience* **12**, 273 (2020).
27. A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.
28. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint* arXiv:1412.6980, 2014.
29. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**(1), 1929–1958 (2014).
30. A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, Atlanta, GA, 2013, p. 3.
31. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
32. D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint* arXiv:1511.07289, 2015.
33. Park, J. & Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural Comput.* **3**(2), 246–257 (1991).
34. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2**(4), 303–314 (1989).
35. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pmlr, 2015, pp. 448–456.
36. Hinton, G. E., Osindero, S. & Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006).
37. Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, 2006.
38. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006).
39. A. L. Maas, A. Y. Hannun, A. Y. Ng, ***et al.,*** "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, p. 3, Atlanta, GA, 2013.
40. Broomhead, D. S., Lowe, D., Radial basis functions, multi-variable functional interpolation and adaptive networks, *Technical Report*, 1988.
41. Cock, P. J. A. et al. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163 (2009).
42. Goulet, D. R. et al. Codon optimization using a recurrent neural network. *J. Comput. Biol.* **30**(1), 70–81 (2023).
43. Kim, J., Cheon, S. & Ahn, I. NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. *BMC Bioinformatics* **23**(1), 187 (2022).
44. L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.
45. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
46. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl. Acad. Sci.* **120**(44), e2311219120 (2023).
47. Hatibi, N. et al. Misclassified: Identification of zoonotic transition biomarker candidates for influenza a viruses using deep neural network. *Front. Genet.* **14**, 1145166 (2023).
48. Ando, D. et al. Decoding codon bias: The role of tRNA modifications in tissue-specific translation. *Int. J. Mol. Sci.* **26**(2), 706 (2025).

49. Su, S. et al. Predicting viral host codon fitness and path shifting through tree-based learning on codon usage biases and genomic characteristics. *Sci. Rep.* **15**(1), 12251 (2025).
50. Hu, D., Wu, D., You, J., He, Y. & Qian, W. Principal component analysis and comprehensive evaluation on salt tolerance related traits in brassica napus l. *Bot. Res* **7**, 101–112 (2018).
51. Y. Zhang, M. Ji, L. Deng, L. Lian, L. Jian, and R. Zhang, "Codon usage bias analysis of self-incompatibility genes BrSRK, BrSLG, and BrSP11/BrSCR in brassica rapa reveals insights into their coevolution," *Genetic Resources and Crop Evolution*, pp. 1–22, 2025.
52. Ji, H. et al. Bioinformatic analysis of codon usage bias of HSP20 genes in four cruciferous species. *Plants* **13**(4), 468 (2024).
53. Plotkin, J. B. & Kudla, G. Synonymous but not the same: The causes and consequences of codon bias. *Nat. Rev. Genet.* **12**(1), 32–42 (2011).
54. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 1–13 (2020).
55. Playe, B. & Stoven, V. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. *Journal of cheminformatics* **12**(1), 11 (2020).
56. Jung, M. et al. Deep learning algorithms correctly classify brassica rapa varieties using digital images. *Front. Plant Sci.* **12**, 738685 (2021).
57. Maniatopoulos, A. & Mitianoudis, N. Learnable leaky ReLU (LeLeLU): An alternative accuracy-optimized activation function. *Information* **12**(12), 513. https://doi.org/10.3390/info12120513 (2021).
58. Hallee, L. & Khomtchouk, B. B. Machine learning classifiers predict key genomic and evolutionary traits across the kingdoms of life. *Sci. Rep.* **13**(1), 2088 (2023).
59. Okut, H. Deep learning algorithms for complex traits genomic prediction. *Hayvan Bilimi ve Ürünleri Dergisi* **4**(2), 225–239 (2021).
60. S. Tong, Y. Chen, Y. Ma, and Y. Lecun, "Emp-ssl: Towards self-supervised learning in one training epoch," *arXiv preprint* arXiv:2304.03977, 2023.
61. Fioravanti, D. et al. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics* **19**, 1–13 (2018).
62. G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, 2002.
63. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**(7553), 436–444 (2015).
64. Buhmann, M. D. *Radial basis functions: Theory and implementations* (Cambridge University Press, 2003).
65. J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv 2018," *arXiv preprint* arXiv:1803.03635, 1803.
66. Bejani, M. M. & Ghatee, M. A systematic review on overfitting control in shallow and deep neural networks. *Artif. Intell. Rev.* **54**(8), 6391–6438 (2021).
67. Liakos, K. G., Busato, P., Moshou, D., Pearson, S. & Bochtis, D. Machine learning in agriculture: A review. *Sensors* **18**(8), 2674 (2018).
68. Luo, H. & Wang, J. ICDO-RBFNN multi-sensor data fusion for agricultural greenhouse environment. *Trans. Chin. Soc. Agric. Eng.* **40**(21), 184–191 (2024).
69. Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu. Rev. Genet.* **42**(1), 287–299 (2008).
70. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**(6594), 4290 (2022).
71. Zou, J. et al. A primer on deep learning in genomics. *Nat. Genet.* **51**(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5 (2019).
72. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface.* https://doi.org/10.1098/rsif.2017.0387 (2018).
73. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**(8), 831–838 (2015).
74. Wainer, J. & Cawley, G. Empirical evaluation of resampling procedures for optimising SVM hyperparameters. *J. Mach. Learn. Res.* **18**(15), 1–35 (2017).
75. Cheng, F., Wu, J., Fang, L. & Wang, X. Syntenic gene analysis between brassica rapa and other brassicaceae species. *Front. Plant Sci.* **3**, 198 (2012).
76. Varshney, R. K., Terauchi, R. & McCouch, S. R. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLoS Biol.* **12**(6), e1001883 (2014).
77. Stephens, Z. D. et al. Big data: Astronomical or genomical?. *PLoS Biol.* **13**(7), e1002195 (2015).
78. T. V. Tatarinova, N. N. Alexandrov, J. B. Bouck, and K. A. Feldmann, "Biology in corn, rice, sorghum and other grasses," 2010.
79. Quax, T. E., Claassens, N. J., Söll, D. & van der Oost, J. Codon bias as a means to fine-tune gene expression. *Mol. Cell* **59**(2), 149–161 (2015).
80. Alexaki, A. et al. Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. *Sci. Rep.* **9**(1), 15449 (2019).
81. Neuwald, A. F. A bayesian sampler for optimization of protein domain hierarchies. *J. Comput. Biol.* **21**(3), 269–286 (2014).
82. Seward, E. A. & Kelly, S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol.* **17**(1), 226. https://doi.org/10.1186/s13059-016-1087-9 (2016).
83. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **18**(5), 851–869. https://doi.org/10.1093/bib/bbw068 (2016).
84. Snowdon, R. J. & Iniguez Luy, F. L. Potential to improve oilseed rape and canola breeding in the genomics era. *Plant Breeding* **131**(3), 351–360 (2012).
85. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**(7), e1003711 (2014).
86. Collard, B. C. & Mackill, D. J. Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1491), 557–572 (2008).
87. Scheben, A., Batley, J. & Edwards, D. Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnol. J.* **15**(2), 149–161 (2017).
88. Jones, D. T. Setting the standards for machine learning in biology. *Nat. Rev. Mol. Cell Biol.* **20**(11), 659–660. https://doi.org/10.1038/s41580-019-0176-5 (2019).
89. Drees, L., Junker-Frohn, L. V., Kierdorf, J. & Roscher, R. Temporal prediction and evaluation of brassica growth in the field using conditional generative adversarial networks. *Comput. Electron. Agric.* **190**, 106415. https://doi.org/10.1016/j.compag.2021.106415 (2021).

Acknowledgements

We acknowledge Ensembl Plants for providing open access genomic data used in this study.

Author contributions

a. Mr. Anjum Shahzad conceptualized the study, designed the methodology, conducted the data analysis, and wrote the initial draft of the manuscript, software implementation and is responsible for the research data.b. Mr. Muhammad Arfan provided overall supervision and critical review of the manuscript.c. Dr. Nauman Khalid

provided expert supervision and critical guidance in interpreting the genomic data and biological implications of this study. His insights into Brassica evolutionary genomics and codon usage patterns significantly strengthened the biological validity of our findings.

Funding
The authors acknowledge funding from Research, Innovation, and Academic Development from Abu Dhabi University, UAE, for conducting research studies.

Declarations

Consent for publication
All authors have read and approved the final manuscript and consent to its publication.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-18814-0.

Correspondence and requests for materials should be addressed to N.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.