

# Comparative Evaluation of Hybrid CNN-LSTM Models for Salinity-Tolerance Gene Prioritization in Halophytes using Transcriptomic Data

Nikitha.V, Sree kala.T

*Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India, nikithavinothkumar@gmail.com*

*Department of Advanced Computing and Analytics, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India, sreekala.scs@vistas.ac.in*

## Abstract

Soil salinity is a major environmental challenge that significantly reduces agricultural productivity and threatens global food security. Although halophytes possess natural mechanisms to thrive under high-salinity conditions, identification of key salinity-tolerance genes remains a complex task due to the high dimensionality of transcriptomic data and the limited application of advanced deep learning approaches for gene prioritization. This study proposes a hybrid CNN–LSTM-based framework for the identification and prioritization of salinity-tolerance genes in halophytes using publicly available transcriptomic datasets. Gene expression data are collected from open-source repositories and subjected to pre-processing and differential expression analysis to identify significantly upregulated genes under salt stress. Functional enrichment analysis is performed to investigate the biological roles of candidate genes. Subsequently, trained and untrained hybrid CNN–LSTM models are comparatively evaluated to prioritize the most influential salinity-responsive genes. The proposed framework provides a robust and data-driven approach for salinity-tolerance gene discovery, contributing to computational plant stress biology and supporting future crop improvement.

**Keywords.** Halophytes, salinity-tolerance gene prioritization, transcriptomic analysis, hybrid CNN–LSTM, deep learning, differential gene expression.

## 1. INTRODUCTION

Soil salinity is a major environmental factor limiting agricultural productivity worldwide. Increasing salinization caused by climate change, seawater intrusion, and improper irrigation practices adversely affects plant growth, crop yield, and food security. Therefore, understanding the genetic mechanisms underlying salinity tolerance has become an important area of research in modern agriculture. Plants that are naturally able to tolerate high salinity levels are known as halophytes. Their unique features provide important insights into the molecular processes underlying salt stress tolerance. Recent advancements in transcriptomic technologies and publicly available gene expression repositories have enabled large-scale analysis of salinity-responsive genes in halophytic species [1,2]. However, identifying the most influential salinity-tolerance genes from high-dimensional transcriptomic datasets remains a challenging task. Traditional computational methods often fail to capture complex gene expression patterns, creating a need for more robust analytical approaches. Deep learning techniques, particularly hybrid CNN–LSTM models, offer significant potential for extracting meaningful biological features and improving gene prioritization accuracy [3].

The proposed framework provides the following advantages:

- Identification of significantly upregulated salinity-responsive genes in halophytes;
- Integration of transcriptomic analysis with hybrid CNN–LSTM deep learning models;
- Comparative evaluation of trained and untrained models for gene prioritization;
- Functional characterization of candidate genes through biological process analysis;
- Selection of top-ranked genes associated with salinity tolerance;
- Establishment of a data-driven framework for salinity-tolerance gene discovery.

### 1.1. Research Contributions

A transcriptomics-based framework is developed to identify salinity-responsive genes from plant expression datasets. Differential expression analysis is employed to prioritize candidate genes associated with salt stress, ANN, LSTM, and CNN–LSTM models are comparatively evaluated for gene prioritization, Biologically relevant salinity-tolerance genes are identified to support future crop improvement studies.

## 2. TRANSCRIPTOMIC DATA PROCESSING AND DEEP LEARNING FRAMEWORK

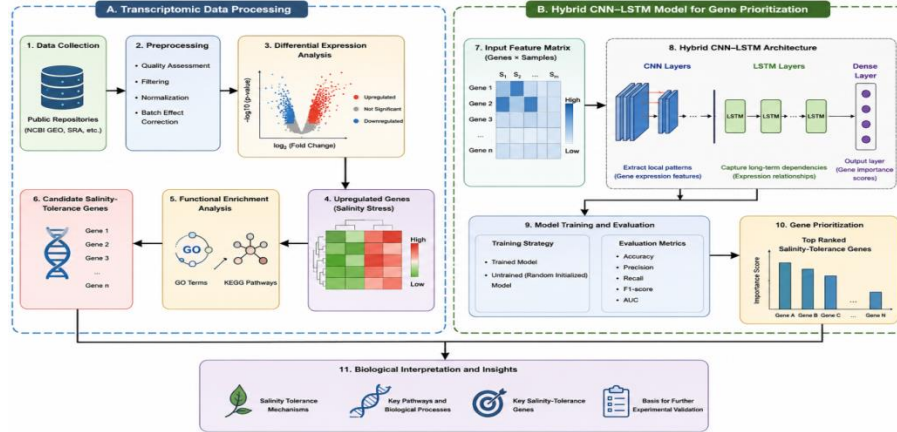


Figure 2.1. Proposed Hybrid CNN–LSTM-Based Framework for Salinity-Tolerance Gene Prioritization

### 2.1. Transcriptomic Data Collection and Gene Expression Analysis

Publicly available transcriptomic datasets of halophytic plant species were collected from gene expression repositories such as the NCBI Gene Expression Omnibus (GEO). The collected datasets contain gene expression profiles under normal and saline stress conditions. Data pre-processing, normalization, and quality assessment were performed to remove noise and ensure data consistency. Differential gene expression analysis was subsequently conducted for identifying significantly upregulated genes related to salinity tolerance [1,4].

#### 2.1.1. Salinity-Responsive Gene Expression

The fold change in gene expression was used to quantify differential expression between salt-stressed and control samples and can be represented as:

$$F_c = \frac{E_{\text{[stress]}}}{E_{\text{[control]}}} \quad 2.1$$

where  $FC$  denotes fold change,  $E_{\text{stress}}$  represents gene expression under salt stress conditions, and  $E_{\text{control}}$  represents gene expression under normal conditions. The identified upregulated genes were further subjected to biological process and functional enrichment analyses to determine their involvement in salinity adaptation mechanisms [5].

### 2.2. Hybrid CNN–LSTM Model for Gene Prioritization

Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) architecture was used to learn complex transcriptomic patterns and prioritize salinity-tolerance genes. The CNN component extracts local expression features, while the LSTM component captures long-range dependencies among gene expression patterns. The model was trained using transcriptomic feature matrices derived from the selected halophytic datasets.

#### 2.2.1. Model Performance Evaluation

The classification accuracy of proposed framework was evaluated using:

$$\text{ACCURACY} = \frac{TP+TN}{TP+TN+FP+FN} \quad 2.2$$

where  $TP$  denotes true positives,  $TN$  denotes true negatives,  $FP$  denotes false positives, and  $FN$  denotes false negatives.

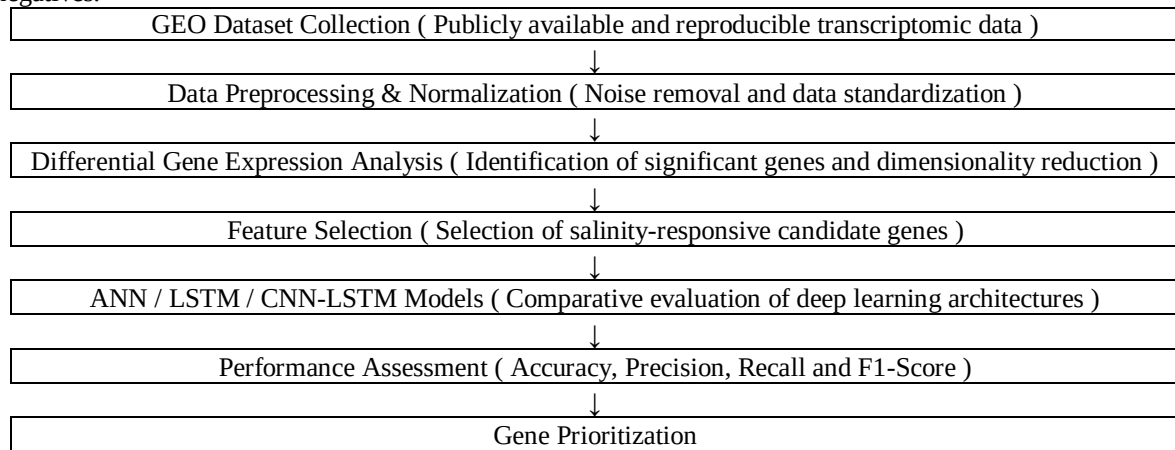


Figure 3.1. Pipeline from GEO data processing to deep learning-based gene prioritization.

Comparative analysis between trained and untrained CNN-LSTM models was performed to evaluate gene prioritization effectiveness. Based on the obtained performance metrics and biological relevance, the top salinity-responsive genes were identified for further investigation [6,7]. Workflow for Salinity-Tolerance Gene Prioritization Shown in [Figure 3.1].

### 3. RESULTS AND DISCUSSION

#### 3.1. Differential Gene Expression Analysis

It was performed on the GSE16765 transcriptomic dataset comprising twelve *Arabidopsis thaliana* samples subjected to control and 100 mM NaCl salt-stress conditions. Following preprocessing and normalization, statistical analysis has been carried out for identifying genes exhibiting significant expression differences between the two experimental groups. A total of 22,810 transcript probes were analyzed. Using a significance threshold of  $p < 0.05$  and an absolute log fold change ( $|\log FC|$ )  $> 1$ , four genes have been identified as significantly differentially expressed under saline conditions.

Table 3.1. Top Prioritized Salinity-Responsive Genes

| Probe ID  | LogFC | P-value  |
|-----------|-------|----------|
| 258498_at | 1.284 | 0.000617 |
| 262128_at | 1.223 | 0.006655 |
| 266901_at | 1.019 | 0.005559 |
| 254396_at | 1.016 | 0.00255  |

These genes demonstrated substantial transcriptional variation between control and salt-treated samples, suggesting their potential involvement in salinity-response mechanisms. These genes were selected for downstream deep learning analysis and salinity-tolerance gene prioritization [Table 3.1]

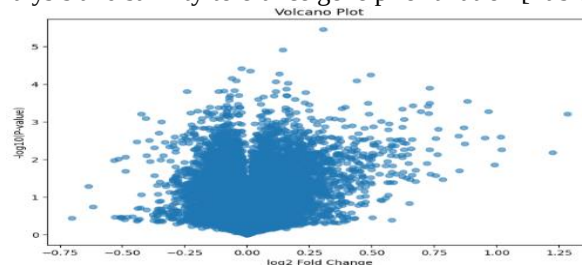


Figure 3.2. Volcano Plot of Differentially Expressed Genes under Salt Stress

Volcano plot illustrating the distribution of transcript probes according to fold change and statistical significance. Genes positioned in the upper-right region represent highly significant upregulated transcripts associated with salinity stress response. The volcano plot demonstrates that only a limited subset of genes exhibited statistically significant expression changes under salt treatment. This observation suggests that salinity tolerance may be regulated by a relatively small group of highly responsive genes rather than widespread transcriptomic alterations [Figure 3.2].

3.2. Expression Pattern Analysis of Significant Genes

To further investigate the expression behavior of the identified genes, a heatmap was generated using the expression values across all experimental samples.

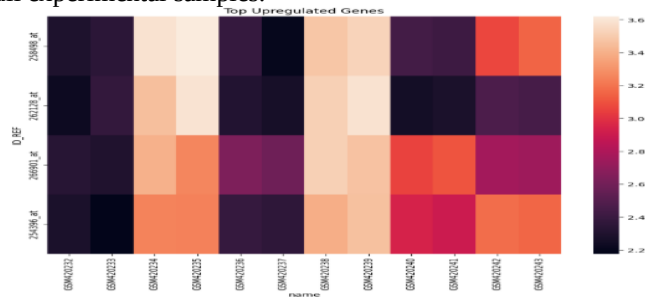


Figure 3.3. Heatmap of Top Differentially Expressed Salinity-Responsive Genes

The heatmap Figure 3.3, revealed distinct expression patterns between control and salt-treated samples. The four significant genes showed increased expression under salinity stress, suggesting their involvement in stress adaptation. The observed clustering pattern further provides the reliability of the differential expression analysis and biological relevance of the identified candidate genes. These findings suggest that the identified genes may serve as candidate biomarkers for salinity tolerance and warrant further biological investigation through pathway enrichment and functional validation studies.

3.3. Comparative Evaluation of Deep Learning Models

To evaluate the predictive capability of the identified salinity-responsive genes, three deep learning architectures were investigated, including Long Short-Term Memory (LSTM), Artificial Neural Network (ANN), Hybrid CNN-LSTM models. The selected significant genes were used as input features for model training and classification of control and salt-treated samples. Model performance has been assessed utilising classification accuracy.

Table 3.2. Comparative Performance of Deep Learning Models with Pre-existing Experiment

| Study                       | Dataset    | Methodology             | Outcome               | Accuracy (%) |
|-----------------------------|------------|-------------------------|-----------------------|--------------|
| Aziz et al. (Existing work) | Halophytes | Transcriptomic Analysis | Salt adaptation genes | —            |
| Proposed Work               | GSE16765   | ANN                     | Gene prioritization   | 100          |
| Proposed Work               | GSE16765   | LSTM                    | Gene prioritization   | 66.67        |
| Proposed Work               | GSE16765   | CNN-LSTM                | Gene prioritization   | 33.33        |

The comparative analysis showed clear differences among the evaluated models. The ANN achieved the highest accuracy of 100%, while LSTM and CNN-LSTM recorded 66.67% and 33.33%, respectively. The superior performance of ANN is likely due to the small feature space and limited dataset size, where simpler architectures perform better with fewer parameters and lower data requirements. In contrast, LSTM and CNN-LSTM are more effective for large-scale datasets with complex sequential patterns. These results suggest that increasing model complexity does not always improve performance in small transcriptomic datasets, and that selecting biologically relevant features can be more important than using deeper architectures. The results are summarized in the Transcriptomic and Deep Learning-Based Gene Prioritization Comparison [Table 3.2].

3.4. Biological and Computational Implications

The integration of differential expression analysis and deep learning enabled the identification of four salinity-responsive genes exhibiting significant expression changes under salt stress conditions. These genes effectively distinguished control and salt-treated samples and may serve as potential biomarkers for salinity tolerance.

Table 3.3. Putative Biological Significance of Prioritized Salinity-Responsive Genes

| Probe ID  | Expression Pattern | Putative Biological Significance                                       | Associated Pathway/Process               |
|-----------|--------------------|------------------------------------------------------------------------|------------------------------------------|
| 258498_at | Upregulated        | Potential involvement in osmotic stress adaptation and ion homeostasis | Salt stress response, osmotic regulation |
| 262128_at | Upregulated        | Associated with transcriptional regulation under abiotic stress        | Stress-responsive signaling pathways     |

|           |             |                                                                          |                                                    |
|-----------|-------------|--------------------------------------------------------------------------|----------------------------------------------------|
| 266901_at | Upregulated | May contribute to cellular protection against salt-induced damage        | Oxidative stress defense and redox regulation      |
| 254396_at | Upregulated | Potential role in maintaining cellular stability under saline conditions | Abiotic stress adaptation and metabolic regulation |

As summarized in Table 3.3, all four prioritized genes exhibited consistent upregulation under salinity stress, indicating their potential involvement in key adaptive responses. These genes are functionally associated with osmotic regulation, ion homeostasis, stress-responsive signaling, oxidative stress defense, and metabolic stability, which are critical mechanisms underlying plant salinity tolerance. From a biological perspective, the results suggest that these genes may play regulatory and protective roles in maintaining cellular function under high-salinity conditions. Computationally, the findings demonstrate that integrating transcriptomic feature selection with deep learning-based prioritization can effectively identify biologically relevant candidate genes associated with stress adaptation. Although the ANN model achieved the highest classification performance, the relatively small dataset highlights the need for further validation using larger and independent transcriptomic datasets. Overall, this study addresses a significant gap in salinity-tolerance gene discovery and provides a strong foundation for future multi-omics integration and functional validation studies.

#### 4. REFERENCES

- [1] Aziz, M. A., & Masmoudi, K. "Insights Into the Transcriptomics of Crop Wild Relatives to Unravel the Salinity Stress Adaptive Mechanisms", *Int. J. Mol. Sci.*, vol. 24, no. 12, 2023.
- [2] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. "A Survey of Best Practices for RNA-Seq Data Analysis", *Genome Biology*, vol. 17, no. 13, 2016.
- [3] Eraslan, G., Avsec, Ž., Gagneur, J. *et al.* "Deep learning: new computational modelling techniques for genomics", *Nat Rev Genet*, **20**, 389–403, 2019.
- [4] Frontiers Plant Science, "Comparative Transcriptome Analysis Reveals the Adaptive Mechanisms of Halophyte *Suaeda dendroides* Encountering High Saline Environment," *Front. Plant Sci.*, vol. 15, 2024.
- [5] John C, Sahoo J, Sajan IK, Madhavan M, Mathew OK., "CNN-BLSTM based deep learning framework for eukaryotic kinome classification: An explainability based approach", *Comput Biol Chem.* 2024.
- [6] Min, S., Lee, B., & Yoon, S. "Deep Learning in Bioinformatics", *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.
- [7] S. Toscano, D. Romano, & A. Ferrante, "Molecular Responses of Vegetable, Ornamental Crops, and Model Plants to Salinity Stress," *Int. J. Mol. Sci.*, vol. 24, no. 4, p. 3190, 2023.

#### Biographies

**Nikitha V** received the Master's degree in Bioinformatics and Data Science and is currently pursuing a Ph.D. in Computer Science at the Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. Her research interests include bioinformatics, deep learning, machine learning, transcriptomic data analysis, computational biology, and artificial intelligence applications in agriculture. She has authored and presented research papers in national and international conferences and is actively involved in interdisciplinary research focusing on genomic data analytics and sustainable agricultural technologies.

**Sree Kala T** is a Professor in the Department of Advanced Computing and Analytics, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. She has published numerous research articles in reputed journals and conferences and has extensive experience in guiding research scholars in advanced computing, data science, and interdisciplinary research domains.