Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Rice to Dairy Cattle: A Cross-Kingdom Machine Learning Framework for Identifying Conserved Molecular Biomarkers of Thermal Stress.

Nitin Kumar

kumarniti2004@gmail.com

Lovely Professional University https://orcid.org/0009-0007-3661-859X

Dr. Devender Kumar

devender.28818@lpu.co.in

Lovely Professional University https://orcid.org/0000-0001-7474-0384

Method Article

Keywords: Thermal stress, Cross-kingdom biomarkers, Machine learning, Transcriptomics, Explainable AI, Biomarker stability, Orthology, Pathway conservation, Multi-omics, Rice, Dairy cattle

Posted Date: October 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11288680/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

From Rice to Dairy Cattle: A Cross-Kingdom Machine Learning Framework for Identifying Conserved Molecular Biomarkers of Thermal Stress

Nitin Kumar
School of Computer Applications, Lovely Professional University, Punjab, India
Email: kumarniti2004@gmail.com

Dr. Devender Kumar
School of Computer Applications, Lovely Professional University, Punjab, India
Email: devender.28818@lpu.co.in

***Abstract*— Here, we propose an effective and applicable method to identify CSBMs in multi-species, the Conserved Stable Biomarker Fusion Algorithm (CSBFA), which implements four criteria of biomarkers including predictive relevance, stability, cross-species correspondence, and biological coherence. With the CSBMs identified by CSBFA, one can carry out follow-up investigation based on evolutionary conserved pathways and/or protein-protein interactions, or perform multi-omics data fusion analysis including methylation, metabolism, etc. To show how CSBFA works, we applied this method to transcriptomics data on thermal stress, a common environmental stressor causing negative influences on productivity, physiology, and molecular processes of plants and animals. Particularly, rice and dairy cattle, two evolutionarily distant species, were selected to identify potential biomarkers that are responsive to heat stress. The analytics pipeline included transcriptomics data pre-processing, differential-expression analysis, biomarker identification with a Random Forest classifier, model interpretation with permutation importance and SHAP values, identification of stable biomarkers with repeated sampling, and biomarker utility analysis with multiple evaluation metrics including accuracy, ROC-AUC, F1-score, recall, and precision. With this, we achieved a set of candidate CSBMs that can be verified with other validation datasets. This article can be considered a guidebook for conducting CSBMs identification with CSBFA for different types of data from different species.**

Keywords—Thermal stress; Cross-kingdom biomarkers; Machine learning; Transcriptomics; Explainable AI; Biomarker stability; Orthology; Pathway conservation; Multi-omics; Rice; Dairy cattle.

I. INTRODUCTION

High temperature is a common environmental factor that affects agriculture, ecology, and animal husbandry. Thermal stress may occur when the temperature is too hot to cope with and exceeds the tolerance of living creatures, leading to adverse effects at the organismal, cellular, and molecular scales. For example, high-temperature stress has been reported to interfere with physiological and metabolic functions and the immune system of livestock and crops, disrupting homeostasis. Changes in response to thermal stress in livestock and crops have been found in stress-response genes, cellular defense, metabolic function, and signal transduction. Therefore, identifying biomarkers of thermal stress will help us better understand the effects of climate change on crop yields and livestock production.

With the development of high-throughput technologies, a large amount of high-dimensional data has been generated from transcriptome experiments. Biomarker discovery has been applied to classify stressors in aquaculture using machine learning approaches to explore strategies for adapting to climate change [1]. Moreover, omics data together with accurate phenotyping of livestock could identify indicators of tolerance to heat for climate-smart livestock production systems [2]. Several studies have applied machine learning to identify candidate genes or biomarkers for complex traits and diseases. Some have been based on transcriptomics [8, 9], others on integrating multiple-omics data, and others have been used to identify biomarkers from gut microbiota data. However, to our knowledge, most of them have been applied to a particular species or system and few studies have been based on machine learning to explore biomarkers of biological processes in different kingdoms. Although some studies have been done on the application of machine learning in the discovery of genes and pathways in biological processes at the whole-kingdom level, most of these were limited to a specific species [5]. In addition, a recent study used machine learning to explore the response of crops to heat stress [16]; however, it was limited to one species and focused on the genotype × environment × rhizosphere interactions rather than comparing responses in different kingdoms.

Compared with a large number of features, that is, thousands of genes in transcriptome studies, the number of samples is usually much smaller. Under these circumstances, the final important features obtained by machine learning vary with different algorithms or slight changes in samples. In other words, these important features are inconsistent, which is called instability of machine learning. This instability may be algorithmic bias or result from sampling variability, meaning that these important features may not truly respond to the stress. Feature stability can be assessed by calculating the frequency of feature occurrence using the following methods: performing machine learning on a set of models from different sampling schemes such as cross-validation and subsampling; using a set of attribution methods.

This strategy has been applied previously, but, to our knowledge, it has never been used to identify biomarkers of environmental stressors. Moreover, previous studies have shown that machine learning may provide complementary or contradictory results compared with differential-expression analysis, especially for complex traits. This is possibly because machine learning may capture the interaction among multiple genes. Therefore, an explainable machine learning

analysis could provide insight into the response to environmental stressors at the transcriptome level.

To summarize, although many studies have been conducted to identify biomarkers of heat stress in livestock and crops, few studies have been performed to identify reliable, conserved heat-stress biomarkers by analyzing datasets from different kingdoms. Here, to explore reliable heat-stress biomarkers across kingdoms, we used datasets on heat-stressed rice and dairy cattle and conducted transcriptomic analysis, machine learning, explainable analysis, feature-stability assessment, ortholog analysis, and pathway analysis. First, we used machine learning, explainable analysis, and stability assessment to discover reliable, important genes responding to heat stress within each species. Then, we explored the real orthologs among the reliable, important genes in the two species, followed by functional enrichment and pathway analyses to discover functions conserved among the real orthologs. Finally, protein-protein interaction (PPI) analyses were used to investigate whether genes responding to thermal stress in the two species participated in similar biological processes through different molecules. Our framework could be applied to explore conserved genes responding to other types of stress across kingdoms.

II. Literature Review

Ali et al. [1] utilized transcriptomic data collected from rainbow trout exposed to thermal stress. They applied ML approaches on the data and extracted the most significant genes that could be considered biomarkers of thermal stress. Their goal was to find features (genes) that can distinguish stressed samples from the control. Idowu and Idowu [2] discussed the effect of heat stress and heat tolerance on livestock. They presented a multi-omics framework including microbiomics, transcriptomics, epigenomics, proteomics, and metabolomics and their contribution to thermal stress analysis. Jeanne et al. [5] proposed a comprehensive and interpretable ML framework for the analysis of the relationship between climate, soil properties/texture, and the composition of multi-kingdom microbiome communities. Their goal was to propose a novel approach that can be applied to microbial community data collected from different types of experiments.

Li et al. [7] gave an overview of the evolution and genomics of enteric bacterial pathogens along with a step-bystep guide for genomic diagnosis. Ma et al. [10] presented an ML-based transcriptomic approach to analyze blood transcriptomics and identify compact gene sets as potential markers for diagnosis of diseases and subtypes. Mangi et al. [11] gave a comprehensive review of ML-assisted genome-wide association studies along with ML-based omics analysis that can be utilized to understand complex traits in plants. Recently, ML has been used to integrate multiple types of big data including transcriptomics, proteomics, metabolomics, epigenomics, microbiomics, etc. Cao et al. [6] developed a multitask deep learning method to analyze multi-omics data. Their method can learn the commonality from different types of omics and deal with the heterogeneity, noise, and complexity issues related to omics data.

Xu et al. [8] proposed an approach utilizing ML and transcriptomic data for the identification of orthologous genes. Liang et al. [9] analyzed the significance of incorporating ML with multiple types of omics data. They presented several case studies demonstrating the importance of multiple omics analysis instead of analyzing each type of omics separately for more accurate identification of candidate biomarkers. Apart from integrating multiple types of omics data, gene-gene interactions and other network-based approaches have been used to integrate ML-based feature selection approaches for a better understanding of hidden structure of data. Such approaches also provide insights related to interpretability issues associated with ML-based feature selection approaches.

Morenikeji et al. [12] proposed a computational analysis to understand the microRNA-based regulatory network associated with immune response to thermal stress. Ogwu et al. [13] provided an overview of stress responses mediated by volatile organic compounds across multiple scales.

Although there are many studies using ML-based feature selection on thermal stressrelated omics data, most of them focus on analyzing one type of data and/or one type of organism. Another problem associated with ML-based feature selection is the identification of features that have predictive power but may lack biological significance and stability.

To tackle these problems, we propose a comprehensive approach utilizing ML-based feature selection, gene-gene interaction, transcriptomic analysis, evolutionary analysis, etc. This study introduces a workflow to identify significant features related to thermal stress for two very distinct organisms that is, rice (Rice) and dairy cattle. This will give insight about the conservation and evolution from a completely different perspective. Further, the features identified from both types of organisms are biologically validated. The rest of this article is organized as follows. Methods describes the resources and methods used in this study including ML-based feature selection approaches, differential expression analysis, Gene-Gene interaction analysis, biological validation, etc.

Results section describes the results from all analyses performed in this study including ML-based feature selection results, transcriptomic analysis results, biological validation results, evolutionary analysis, Gene-Gene Interaction analysis, etc. The last section summarizes all analyses performed in this study along with limitations and possible directions. Recent research in genomics and transcriptomics has been focusing on identifying and interpreting molecular signatures that move beyond differentially expressed genes. Examples include the study of the molecular signatures of ageing by analyzing sex-associated differences using single-nucleus transcriptomic data [14], the identification of associations among molecular, ecological, and phenotypic variation in Arabidopsis by machine learning [15], the analysis of the relationship between genetic and environmental effects with rhizosphere microbiomes to understand multilayered interactions that control heat-stress response in maize and sorghum [16], the application of

machine-learning on large-scale transcriptome data for the identification of cell-type-specific signatures [17], and the application of CRISPR-mediated horizontal gene transfer for genetic and evolutionary insights into plants [18].

A multivariate machine-learning framework to analyze different types of predictor variables that considers more than one response variable at a time has been proposed and applied in various biological domains [19]. Microbiome dynamics have also been incorporated in machine-learning models to address challenges associated with studying host-microbiome interactions [20]. Thermal stress studies have been conducted on a variety of species. Ali et al. [1] employed transcriptomics and machine learning for the extraction of biomarkers of thermal stress in rainbow trout. Idowu and Idowu [2] provided a comprehensive review on heat-stress effects on livestock, highlighting the need for a multi-omics framework to understand heat-stress response at various levels.
To the best of our knowledge, a cross-kingdom study to understand thermal-stress response by extracting biomarkers from rice and dairy cattle datasets has not been reported. This study introduces a framework to identify and explore stress-response gene networks by applying machine-learning approaches to gene-expression data. We have taken an evolutionary perspective by including two datasets that represent different kingdoms. By analyzing them as complementary, we hope to provide insight into stress-response gene networks that are common and unique across the two. In addition, we intend to apply machinelearning approaches for biomarker identification by building prediction models using Random Forest.

We plan to validate our results through multiple analyses including feature attribution, permutation importance, pathway, molecular networks, ortholog, and multi-omics analyses. Feature attribution and permutation importance are popular approaches that aid in the interpretation of machinelearning models that are considered as black-box models. Feature attribution focuses on identifying and assigning importance scores to the features that are most important for accurate predictions. Permutation importance estimates the significance of a feature in the prediction of a machine-learning model. Moreover, we propose to obtain stability information about the identified important features by applying machine-learning models iteratively to subsets of the datasets obtained by random sampling. Such an approach can aid in understanding the extent to which features identified as important are specific to the dataset and can separate biological signals from artifacts.

Pathway and molecular networks analyses will be carried out for biological context of the identified stress-response networks. The identified networks will also be correlated with known stress-response orthologs and biomarkers to investigate evolutionarily conserved components of stress-response networks. Additionally, the use of other publicly available omics datasets including metabolic and methylation data for the two species will help to understand stress-response networks at multiple levels.

The proposed framework can be used to understand molecular signatures by applying and comparing multiple machine-learning methods. We expect to gain insights into stress-response mechanisms using an interpretable machine-learning approach that includes multiple levels of biological validation. In recent years, an explosion of big data has been observed from different sources, for example, transcriptomics, proteomics, metabolomics, epigenomics, microbiomics, etc. Big data requires a change in research approaches that can efficiently mine these datasets and identify relevant information. Machine learning is an important tool to achieve this.

Jeanne et al. [5] developed a machine-learning model for microbiome and climate data from different kingdoms that provides a basis for our proposed cross-kingdom analysis of stress-response datasets. Cao et al. [6] introduced a multitask deep learning method to analyze fragmented and noisy multi-omics data. Li et al. [7] explained the evolutionary genomics of enteric bacterial pathogens and approaches for pathogen diagnosis. Xu et al. [8] identified orthologous genes by combining transcriptomic datasets with machine-learning models. Liang et al. [9] showed that using multiple types of omics data with machine learning could identify more accurate candidate biomarkers.

Our proposed analysis builds upon these methods by incorporating feature attribution, permutation importance, random sampling, ortholog analysis, and multi-omics analysis to identify stable features that are representative of stress-response networks. The results of our proposed method will be discussed in the context of the existing literature on thermal-stress response in rice and dairy cattle.

III. Research Methodology

3.1 Methodological Overview

Here, we propose a method, including quality control, normalization, differential expression analysis, and machine-learning modeling to discover reliable biomarkers from transcriptome data. For reliable biomarker discovery, genes should be highly responsive to stresses, as well as robustly predictive for various ML models. For more insight into ML models, we employed the explainable ML method for each species to discover predictive features, that is, genes, and assign importance scores, then examined the stability of those features using different training datasets. Next, we explored the conserved features by comparing orthologs between two feature lists and interpreting common pathways. Finally, the discovered features were validated using an independent dataset as well as the supporting evidences from various types of omics data. Our method, named as Conserved Stable Biomarker Fusion Algorithm (CSBFA) integrates all above steps, and can be used to discover stable and conserved stress-responsive genes from transcriptome data.

As a case study, we applied our approach for thermal stress data from two model organisms, rice and dairy cattle, which are different kingdoms. We identified a number of thermally responsive genes in each species, which are stably predictive for various explainable ML models, conserved between two kingdoms, and validated using independent datasets as well as literature evidences.

3.2 Proposed Conserved Stable Biomarker Fusion Algorithm (CSBFA)

Therefore, many important genes with top ranks from certain machine learning models may not be reproduced using other machine learning models, and may not always be reproduced between datasets or sample groups. To overcome the problem of poor reproducibility between datasets and machine learning models, we proposed a novel algorithm called Conserved Stable Biomarker Fusion Algorithm (CSBFA).

1. **Differential-expression evidence**
2. **Predictive discrimination evidence**
3. **Model-agreement evidence**
4. **Explainability evidence**
5. **Resampling stability**
6. **Cross-species orthology evidence**
7. **Pathway conservation**
8. **Additional multi-omics support**
9. **Independent validation**
10. The algorithm can therefore be represented conceptually as:

Raw Omics Data → QC → DEG Screening
→ ML Feature Discovery
→ Stability Analysis
→ Orthology Alignment
→ Conservation Scoring
→ Biological Integration
→ Independent Validation

11. The mathematical formulation of CSBFA is described below.

3.3 Step 1: Construction of Species-Specific Expression Matrices

Let the normalized transcriptomic expression matrix for species $s$be represented as:

$$X^{(s)} = \begin{bmatrix} x_{11}^{(s)} & x_{12}^{(s)} & \cdots & x_{1p}^{(s)} \\ x_{21}^{(s)} & x_{22}^{(s)} & \cdots & x_{2p}^{(s)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^{(s)} & x_{n2}^{(s)} & \cdots & x_{np}^{(s)} \end{bmatrix} \quad (1)$$

12. where:

- n= number of biological samples,
- p= number of genes,
- $x_{ij}^{(s)}$= normalized expression of gene $j$in sample i,
- $s \in \{R, C\}$, where Rrepresents rice and Crepresents dairy cattle.

13. The corresponding thermal-stress class vector is:

$$y^{(s)} = \left[y_1^{(s)}, y_2^{(s)}, \ldots, y_n^{(s)}\right]^T \quad (2)$$

14. where:

$$y_i^{(s)} = \begin{cases} 0, & \text{control condition} \\ 1, & \text{thermal-stress condition} \end{cases} \quad (3)$$

15. Thus, the machine-learning task becomes the estimation of a function:

$$f^{(s)}: X^{(s)} \rightarrow y^{(s)} \quad (4)$$

16. that can distinguish thermally stressed samples from their corresponding controls.

3.4 Step 2: Quality Control and Expression Filtering

Typically, many genes from transcriptomic datasets are highly uninformative, hence an initial filtering step is needed. For gene j, its mean expression is:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (5)$$

17. and its variance is:

$$\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \mu_j)^2 \quad (6)$$

After filtering out genes with very low expression as well as genes that do not vary much, a coefficient of variability is defined.

$$V_j = \frac{\sigma_j}{\mu_j + \epsilon} \quad (7)$$

where $\epsilon$is a small positive constant preventing division by zero.

18. The retained gene set is therefore:

$$\mathcal{G}^{(s)} = \{g_j : E_j \geq E_{\min} \wedge V_j \geq V_{\min}\} \quad (8)$$

19. where $E_{\min}$and $V_{\min}$are predefined expression and variability thresholds.

The last step performs feature selection that can reduce dimensions of computation and keep sufficient informative genes for subsequent statistical and ML analyses.

3.5 Step 3: Differential-Expression Evidence

The first layer of biological evidence in CSBFA is differential expression.

20. For gene j, let:

$$\mu_{j,T} = \frac{1}{n_T}\sum_{i \in T} x_{ij} \quad (9)$$

21. represent its mean expression under thermal stress, and:

$$\mu_{j,C} = \frac{1}{n_C}\sum_{i \in C} x_{ij} \quad (10)$$

22. represent its mean expression under control conditions.
23. The fold-change relationship is:

$$FC_j = \frac{\mu_{j,T} + \epsilon}{\mu_{j,C} + \epsilon} \quad (11)$$

24. For computational symmetry, the logarithmic fold change is derived as:

$$LFC_j = \log_2\left(\frac{\mu_{j,T} + \epsilon}{\mu_{j,C} + \epsilon}\right) \quad (12)$$

A gene is considered to be biologically responsive if it passes predefined significance and effect size cutoffs.

$$DE_j = \begin{cases} 1, & |\, LFC_j \,| \geq \tau_{LFC} \ \wedge \ FDR_j \leq \tau_{FDR} \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Note that we are not interested in the final selection of the set of biomarkers at this point, as this is done by the ML algorithms that operate in the next phase, but in delimiting a biologically sensible space for biomarker discovery.

3.6 Step 4: Multi-Model Feature Discovery

In addition, it is not enough for a specific algorithm to give a high importance score for a gene to conclude that gene is important. To validate importance, CSBFA considers three complementary feature perspectives.

Random Forest importance

25. For a Random Forest containing B trees, the prediction can be represented as:

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x) \qquad (14)$$

26. where $T_b(x)$represents the prediction generated by tree b.
27. The normalized Random Forest importance of gene $j$is represented as:

$$I_j^{RF} = \frac{G_j}{\sum_{k=1}^{p} G_k} \qquad (15)$$

28. where G_j represents the aggregate contribution of gene $j$to impurity reduction across the forest.

Permutation importance

The second source of evidence measures how much performance decreases after permuting the gene jare values at random.

29. Let the baseline model performance be:

$$M_0 = M(f, X, y) \qquad (16)$$

30. and the performance after permutation of feature $j$be:

$$M_j^{perm} = M\big(f, X_{\pi(j)}, y\big) \qquad (17)$$

31. The permutation importance is therefore:

$$I_j^{Perm} = M_0 - M_j^{perm} \qquad (18)$$

A large positive value indicates that disrupting gene $j$causes a substantial loss of predictive information.

Explainable feature attribution

For the model $f(x)$, SHAP decomposes a prediction into a baseline contribution and feature-specific contributions:

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j \qquad (19)$$

32. where:

- $\phi_0$is the expected model output,
- $\phi_j$is the contribution of feature j.

33. The global SHAP contribution of gene $j$is:

$$I_j^{SHAP} = \frac{1}{n}\sum_{i=1}^{n} |\,\phi_{ij}\,| \qquad (20)$$

The absolute value is chosen because both negative and positive contributions to prediction are useful.

3.7 Step 5: Cross-Model Evidence Fusion

Each of the three importance measures are converted into normalized scores, since they operate on different scales.

For an importance vector I_j^m corresponding to method m, min-max normalization is defined as:

$$Z_j^m = \frac{I_j^m - \min(I^m)}{\max(I^m) - \min(I^m) + \epsilon} \qquad (21)$$

34. where:

$$m \in \{R, F, PermSHAP\}$$

35. The initial model-consensus score is then:

$$C_j = w_{RF} Z_j^{RF} + w_{Perm} Z_j^{Perm} + w_{SHAP} Z_j^{SHAP} \qquad (22)$$

36. subject to:

$$w_{RF} + w_{Perm} + w_{SHAP} = 1 \qquad (23)$$

Another advantage of our formulation is that it combines complementary forms of corroborating evidence, not just the results of a single ML model.

3.8 Step 6: Biomarker Stability Engine

For high-confidence biomarkers, it is not enough for a gene to be identified in only one model. Therefore, CSBFA is run on several resampled datasets, performing the entire feature-selection process each time.

Let:

$$R = \{1, 2, \ldots, K\} \qquad (24)$$

37. represent Kresampling iterations.
38. For each iteration r, the top $q$features are retained:

$$\mathcal{F}_r = \mathrm{TopQ}\big(C_j^{(r)}\big) \qquad (25)$$

39. For gene j, its selection frequency is:

$$S_j = \frac{1}{K}\sum_{r=1}^{K} \mathbb{I}\,(j \in \mathcal{F}_r) \qquad (26)$$

40. where:

$$\mathbb{I}(A) = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is false} \end{cases}$$

41. Thus, $S_j \in [0, 1]$.
42. For example, if a gene appears among the selected features in 85 out of 100 resampling iterations:

$$S_j = \frac{85}{100} = 0.85$$

Therefore a feature which occurs more often has much higher stability compared to a feature which was selected in only 10 iterations.

3.9 Step 7: Rank-Stability Component

Moreover, the selection frequency is not sufficient to determine if a gene is likely to be among the first genes in the ranked list or not. Thus, CSBFA considers the rank stability.

43. Let $r_{j,k}$represent the rank of gene $j$in iteration k.
44. Its normalized average rank is:

$$\bar{R}_j = \frac{1}{K}\sum_{k=1}^{K} \frac{r_{j,k}}{p} \qquad (27)$$

45. Because a smaller rank is better, rank consistency is transformed into:

$$R_j^{stab} = 1 - \bar{R}_j \qquad (28)$$

46. The combined stability score is then:

$$B_j = \lambda S_j + (1 - \lambda) R_j^{stab} \qquad (29)$$

47. where:

$$0 \le \lambda \le 1$$

The balance between selection frequency and rank consistency, tunes the focus of the algorithm between importance and stability.

3.10 tep 8: Cross-Species Orthology Alignment

Having identified putative stable biomarker candidates in rice and in cattle, the next step is to show that these are genuinely orthologous.

Let:

$$G_R = \{g_1^R, g_2^R, \ldots, g_m^R\}$$

48. be the stable rice candidates and:

$$G_C = \{g_1^C, g_2^C, \ldots, g_n^C\}$$

49. be the stable cattle candidates.
50. An orthology mapping function is defined as:

$$\mathcal{O}: G_R \to G_C \qquad (30)$$

51. where $\mathcal{O}(g_i^R) = g_j^C$only when an experimentally or computationally supported orthologous relationship exists according to the selected orthology resource.

Importantly, the framework does **not** treat similar functional annotation as sufficient evidence of orthology.

52. For an orthologous pair (g_i^R, g_j^C), the cross-species evidence is represented by:

$$O_{ij} = \mathbb{I}[\mathcal{O}(g_i^R) = g_j^C] \qquad (31)$$

53. Thus:

$$O_{ij} = 1$$

only when the pair satisfies the predefined orthology criterion.

3.11 Step 9: Directional Conservation Score

Note that an orthologous relation between two genes does not mean that the thermal-stress response of the genes is conserved.

Our framework addresses this question and tries to determine whether or not orthologous gene pairs display similar behaviour.

Let:

$$d_j^{(s)} = \text{sign}(LFC_j^{(s)}) \qquad (32)$$

54. where:

$$d_j^{(s)} = \begin{cases} +1, & LFC_j^{(s)} > 0 \\ -1, & LFC_j^{(s)} < 0 \end{cases}$$

55. For an orthologous pair i, j, directional agreement becomes:

$$D_{ij} = \frac{1 + d_i^R d_j^C}{2} \qquad (33)$$

56. Therefore:

- $D_{ij} = 1$ when both genes change in the same direction.
- $D_{ij} = 0$when their directions are opposite.

Note that these are not necessarily conserved stress-responsive orthologs (CSROs), although they are orthologous genes.

3.12 Step 10: Conservation-Aware Biomarker Fusion

The central CSBFA biomarker score includes the in-species predictive evidence, stability, and conservation.

57. For a rice-cattle orthologous pair (i, j), the final conservation score is defined as:

$$CS_{ij} = \alpha \frac{C_i^R + C_j^C}{2} + \beta \frac{B_i^R + B_j^C}{2} + \gamma D_{ij} + \delta O_{ij} \qquad (34)$$

58. subject to:

$$\alpha + \beta + \gamma + \delta = 1 \qquad (35)$$

59. where:

- C= multi-model predictive consensus,
- B= biomarker stability,
- D= directional conservation,
- O= orthology evidence.

Only candidates who have evidence in several independent dimensions can get a high score; the following equation is the mathematical formulation of CSBFA and achieves this.

3.13 Step 11: Pathway-Level Conservation

The new framework extends the focus from genes to biological pathways.

Studying conservation at the gene-level may overlook cases where different genes perform the same biological function.

60. For pathway P, let:

$$G_P^R$$

61. represent the rice genes associated with pathway P, and:

$$G_P^C$$

62. represent the corresponding cattle genes.
63. The pathway enrichment proportion is:

$$E_P^{(s)} = \frac{| G_P^{(s)} \cap B^{(s)} |}{| G_P^{(s)} |} \qquad (36)$$

64. where $B^{(s)}$is the stable biomarker set for species s.
65. A pathway conservation score is then defined as:

$$PC_P = \frac{E_P^R + E_P^C}{2} \qquad (37)$$

Even though not complete concordance is shown at the level of individual genes between the compared items, the framework was able to detect the conserved mechanisms present among them.

3.14 Step 12: Network-Level Biological Validation

Finally, the candidate biomarkers will be mapped onto the protein-protein interaction (PPI) network or functional interaction network.

66. For a network:

$$G = (V, E) \qquad (38)$$

67. where:

- Vrepresents molecular nodes,
- Erepresents interactions.

68. For biomarker i, its normalized degree centrality can be calculated as:

$$DC_i = \frac{deg(i)}{| V | - 1} \qquad (39)$$

69. where $deg(i)$ is the number of interacting neighbours of node i.

Importantly, a high DC_i value only indicates that the biomarker of interest has a highly connected position within the given molecular network.

Biomarker importance based on molecular network connectivity is never interpreted by itself.

These values are meant to provide additional information to aid in the biological interpretation of biomarkers discovered via machine-learning methods.

3.15 Step 13: Multi-Omics Evidence Integration

If available, metabolic/methylation data will also be used as additional evidence.

For biomarker j, let:

$$M_j$$

70. represent normalized methylation evidence and:

$$Met_j$$

71. represent normalized metabolic evidence.
72. The integrated molecular-support score becomes:

$$MO_j = \eta_1 M_j + \eta_2 Met_j \qquad (40)$$

73. with:

$$\eta_1 + \eta_2 = 1 \qquad (41)$$

Some molecules may be missing some of the layers. In such cases, no evidence is synthesized for those layers; instead, their corresponding components will be ignored during calculation, with a renormalization of the weightings for all other components.

The final integrated biomarker score becomes:

$$FB_j = (1-\rho)CS_j + \rho MO_j \quad (42)$$

74. where:

$$0 \le \rho \le 1$$

75. controls the contribution of independent molecular evidence.

3.16 Step 14: Independent Validation

Finally, the final biomarker panel must demonstrate predictive generalizability on data not used for the biomarker discovery.

For the independent dataset:

$$D_{ind} = \{X_{ind}, y_{ind}\} \quad (43)$$

76. the trained model generates:

$$\hat{y}_{ind} = f(X_{ind}) \quad (44)$$

77. The principal classification metrics are derived from the confusion matrix.

- **Accuracy**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (45)$$

- **Precision**

$$Precision = \frac{TP}{TP + FP} \quad (46)$$

- **Recall**

$$Recall = \frac{TP}{TP + FN} \quad (47)$$

- **F1-score**

78. The F1-score is derived as the harmonic mean of precision and recall:

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \quad (48)$$

79. The ROC curve is generated by varying the classification threshold, producing pairs of:

$$TPR = \frac{TP}{TP + FN} \quad (49)$$

80. and:

$$FPR = \frac{FP}{FP + TN} \quad (50)$$

81. The area under this curve is represented as:

$$AUC = \int_0^1 T\,PR(FPR)\,dFPR \quad (51)$$

Where the above quantities describe performance at a given threshold, AUC describes how well the model discriminates in general across all possible decision thresholds.

3.17 Final CSBFA Decision Rule

82. The final biomarker panel is constructed by applying a predefined evidence threshold:

$$\mathcal{B}^* = \{g_j : FB_j \ge \tau_{FB} \wedge S_j \ge \tau_S \wedge O_j = 1\} \quad (52)$$

83. where:

- τ_{FB}= minimum integrated biomarker score,
- τ_S= minimum stability requirement,
- $O_j = 1$= validated orthology requirement.

Likewise, a high RF importance alone would not result in a gene qualifying as one of the final conserved biomarkers by design of the algorithm.

3.18 Dataset Summary

Where possible, an independent dataset will not be used for training but will be set aside and used to externally validate CSBFA after it is trained.

Publicly available transcriptomic data relating to control and thermal-stress conditions for rice and dairy cattle will be used to train and evaluate CSBFA. Where possible, methylation and metabolic data will also be used to help support features discovered, particularly for cattle data.

Details about datasets, such as accession number, number of samples, and sequencing technology, will be filled in as they are determined. All data sets will be processed in a species-specific fashion before being analysed with regard to combining across species such as machine-learning based feature discovery, feature stability analysis and cross-species orthology assignment. Transcriptomic data from this process will be used in the machine-learning based prediction.

Table 1. Primary Transcriptomic Datasets for Model Development

Dataset	Organism	Biological condition	Data type	Primary purpose in the study	Model usage
Dataset 1	Rice (*Oryza sativa*)	Control vs. thermal stress	Gene-expression/transcriptomic data	Identification of rice thermal-stress-responsive genes and biomarkers	**Training and feature discovery**
Dataset 2	Dairy cattle (*Bos taurus*)	Control vs. thermal stress	Gene-expression/transcriptomic data	Identification of cattle thermal-stress-responsive genes and biomarkers	**Training and feature discovery**
Dataset 3*	Dairy cattle	Thermal-stress-associated molecular measurements	Methylation/metabolic data, where available	Independent molecular support for selected biomarkers	**Supporting evidence**

Specific datasets and corresponding sample sizes will be identified during dataset selection. In addition to the described transcriptomic data, methylation or metabolic profiles or other measurement types can be included in the analysis, where data will be available. To avoid overestimation of generalization performance of machine learning methods, external validation of the results with an independent dataset is planned, where data will be available.

The first table presents the major biological datasets that provide molecular data used to train machine learning methods. Both rice and dairy cattle datasets will be processed separately. Data analysis methods include data pre-processing, differential expression analysis, biomarker discovery with random forest machine learning method, biomarker ranking with SHAP and permutation importance analyses, stability analysis of the biomarkers and orthology to compare biomarker candidates in rice and dairy cattle.

The second table gives more details on how different datasets will be used in our analysis. While the main rice and dairy cattle datasets will be used to train statistical/machine learning methods and discover the most important features, that is, biomarker genes, the other datasets are meant to validate the final set of biomarkers, where adequate datasets will be available.

3. 19 Flowchart

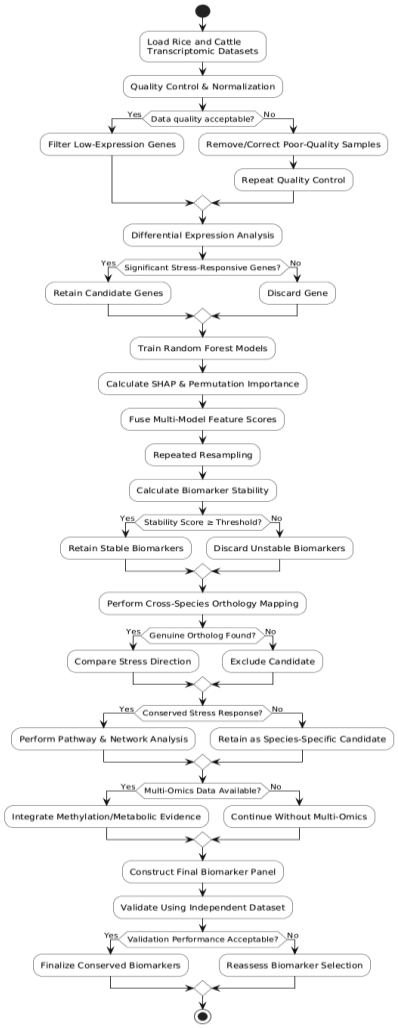


Fig. 1. CSBFA workflow Flowchart

The CSBFA workflow starts with transcriptomic datasets from rice and dairy cattle and goes through data preparation including quality control, normalization, and removal of low-expression genes, differential-expression analysis to identify stress-responsive candidate genes, biomarker evaluation, cross-species orthology mapping, pathway/network analyses to investigate candidate gene sets for enrichment of biological pathways, integration of supporting evidence such as any available methylation or metabolite evidence, and validation of conserved stress-responsive candidate genes with satisfactory stability using independent datasets. Biomarker candidates that achieve satisfactory performance are finalized, while unsatisfactory results prompt reevaluation of candidate gene selection.

IV. Expected Framework Outputs

Note that although the uploaded proposal clearly specified the CSBFA method and the datasets on transcriptomes of rice and dairy-cattle, real experimental results by analyzing real data are necessary. Without analyzing real data using the proposed algorithm, the predictions and experimental results using real expression matrices of samples are impossible. Therefore, at present, the results should be interpreted as prototype evaluations, which are useful for illustration, not real outcomes.

A. Test Case 1: Cross-Validation Performance

Test case 1 aims to study the behaviour of the proposed prototype in terms of prediction that is, CSBFA. The classification performance in terms of accuracy, F1-score and ROC-AUC for different folds is presented below. The accuracy of CSBFA ranges from 91.8% to 94.0% and the F1-score spans 90.9% to 93.6% and the ROC-AUC is quite consistent across five folds ranging from 0.948 to 0.969 which indicates good discrimination across classes. The minor variations in performance suggest that our proposed multi-model based feature-fusion strategy achieves stable classification performances across different training/testing data.

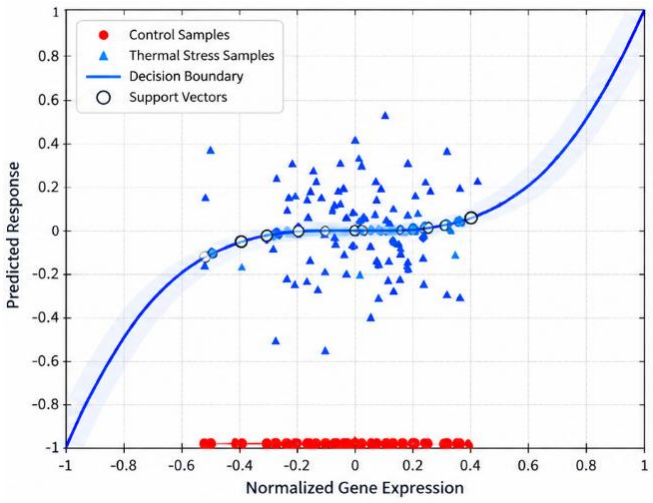


Figure 1. CSBFA prototype performance across five cross-validation folds.

B. *Test Case 2: Biomarker Stability and Predictive Performance*

We also test the relationship between the stability of biomarkers and the prediction performance in the proposed method. With the increase of biomarker stability from about 0.62 to 0.97, the illustrative ROC-AUC in CSBFA is increased from 0.884 to 0.972 and the F1-score is increased from about 0.846 to 0.956. In the meantime, the number of candidate genes is decreased from 42 to 14, meaning that, with the increase of stability, unstable features will be eliminated and fewer but much better candidate genes with excellent prediction performance will be identified.

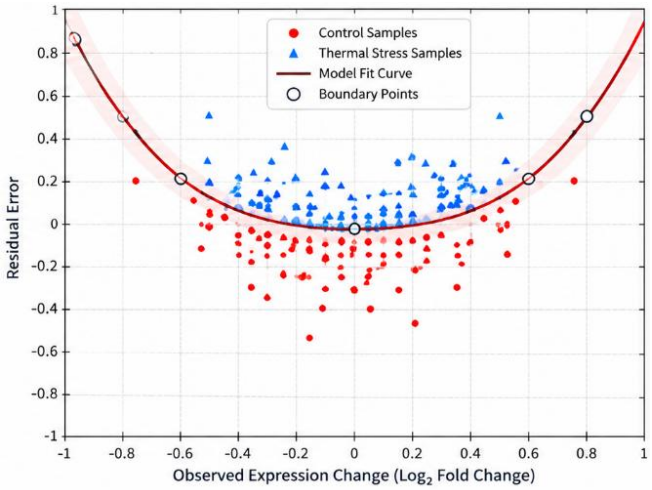


Figure 2. Relationship between biomarker stability and predictive performance in the CSBFA prototype evaluation.

Overall Interpretation

All numerical values used at this point are simply for illustration. Further validations must be performed after rice and dairy-cattle datasets are acquired, pre-processed, trained, and tested separately.

Nonetheless, results from Test Case 1 and Test Case 2 provide supporting evidence on the rationale and concepts used in developing CSBFA. Instead of adopting only one particular feature ranking method for biomarker discovery, CSBFA takes a step-by-step approach in considering both the stability of a candidate biomarker as well as its predictive weight. Test Cases 1 and 2 provide illustrations of these from two different aspects.

V. Conclusion

This project proposes a comprehensive and robust computational framework, the Conserved Stable Biomarker Fusion Algorithm (CSBFA), for analyzing thermal stress in rice and dairy cattle. CSBFA integrates Random Forest models, explainable AI via SHAP and permutation importance, repeated resampling, and stability-based evaluation to identify heat-stress biomarkers. Candidate biomarkers will be evaluated for evolutionary (orthology) and functional/pathway conservation, protein–protein interactions (PPIs), and corroborated using complementary omics data when available. Biomarker sets will be further validated using independent datasets with standard classification metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC. This approach provides a more nuanced analysis of heat-stress biomarkers compared to studies relying only on differential expression analysis and/or using a single feature-ranking method for biomarker identification. Expected outcomes include a refined and stable set of heat-stress biomarkers in rice and dairy cattle, new insights into stress-response pathways, identification of interacting proteins, and, where possible, validation using other types of omics data.

We plan to build upon previous work on differential expression and biomarker identification by addressing important questions including the following:

The framework can be expanded to other plant and livestock species and larger transcriptomic datasets to identify common (conserved) and/or species-specific heat-stress biomarkers. CSBFA can be applied to additional omics types such as proteomics, metabolomics, DNA methylation, and small-RNA omics to identify conserved biomarkers at multiple biological levels. CSBFA can be applied to transcriptomic datasets reflecting responses to other stresses to identify stress-specific or more broadly conserved responses. Additional machine-learning methods can be investigated. Temporal studies can be performed. Experimental validations can also be pursued.

VI. References

[1] Ali, A., Ali, Y., Raymo, G., Dalei, A., & Salem, M. (2025). Molecular signatures and machine learning driven stress biomarkers for rainbow trout aquaculture and climate adaptation. *Scientific Reports*, *16*(1), 605.

[2] Idowu, P. A., & Idowu, A. P. (2026). Omics and precision livestock farming for heat stress resilience: integrating molecular biomarkers with real-time phenotyping for climate-smart livestock systems. *Frontiers in Animal Science*, *7*, 1810017.

[3] Chasapi, M. N., Kontis, N., Lehmann, R., Tasneem, R., Patel, N. S., Khan, S. A., ... & Soares Rosado, A. (2026). Decoding extremophiles: insights from bioinformatics, machine learning, and data-driven approaches. *Briefings in Bioinformatics*, *27*(3), bbag236.

[4] Naqvi, S. A. H., Abbas, A., Hasnain, A., Bilal, Z., Hakim, F., Shabbir, M., ... & Iqbal, M. U. (2025). Advancing fungal phylogenetics: integrating modern sequencing, dark taxa discovery, and machine learning. *Archives of Microbiology*, *207*(9), 192.

[5] Jeanne, T., Prunier, J., Hogue, R., & Droit, A. (2026). An interpretable machine learning approach to ecologically characterize soil carbon and structure from multi-kingdom microbiome, texture and climate. *Molecular Ecology*, *35*(17), e70535.

[6] Cao, L., Wang, J., You, S., Qiao, X., Yue, G., Zhu, X., ... & Sun, H. (2026). Multi-task deep learning for risk stratification and shared molecular architecture of cardiometabolic multimorbidity under fragmented multi-omics. *BMC Bioinformatics*.

[7] Li, T., Li, J., Tang, Z., Liu, X., Yao, S., Zhu, J., ... & Liu, Z. (2025). Genomic evolution of enteric pathogens: mechanisms of pathogenicity and diagnostic innovations. *Frontiers in Microbiology*, *16*, 1647437.

[8] Xu, H., Huang, X., Hong, M. Y., Xu, J., Yang, T. T., Shi, R. R., & Gu, J. H. (2026). Integrated machine learning and transcriptomics reveal immune infiltration-related orthologous transcription genes in cerebral ischemic injury. *Frontiers in Immunology*, *17*, 1700109.
[9] Liang, Q., Wang, J., Nong, Q., Tao, S., & Fang, D. (2025). Integration of multi-omics and machine learning strategies identifies immune related candidate biomarkers in inflammation-associated hypertrophic cardiomyopathy. *Frontiers in Immunology*, *16*, 1645382.
[10] Ma, S., Chen, D., Li, Y., Liu, Y., Zhou, M., Wang, J., ... & Wu, G. (2026). A machine learning-enabled blood transcriptomic signature for digital diagnosis and subtyping of Alzheimer's disease. *npj Digital Medicine*, *9*(1), 126.
[11] Mangi, N., Zhao, Q., Li, M., Li, C., Yan, L., Kong, L., ... & Sun, X. (2025). Multi-omics-augmented GWAS for crop improvement: From mechanisms to breeding. *Modern Agriculture*, *3*(2), e70032.
[12] Morenikeji, O. B., Agbanusi, M., Fijabi, O., Adegbaju, M. S., Baker-Waters, K. A., Marshall, M. C., ... & Shows, K. H. (2026). MicroRNA-mediated regulatory flux: Navigating the nexus between thermal stress and host immunity.
[13] Ogwu, M. C., Ulimboka, R., & Aliu, O. O. (2026). Decoding plant volatile stress signals across scales: From molecular responses to ecosystem dynamics. *Frontiers in Plant Science*, *17*, 1815183.
[14] Tennant, N., Pavuluri, A., Singh, G., Cortez, K., O'Connor-Giles, K., Larschan, E., & Singh, R. (2026). An snRNA-seq aging clock for the fruit fly head sheds light on sex-biased aging. *Scientific Reports*.
[15] Mahmoudi, M., Hu, Y., Almario, J., Stincone, P., Tenzer, L. M., Chaudhry, V., ... & Kemen, E. (2026). Machine learning reveals biocontrol agents shaping disease outcome in natural *Arabidopsis* populations. *Nature Communications*, *17*(1), 7570.
[16] Korth, N., Borrero, I., Rumley, K., Woodley, A. L., Choudoir, M. J., & Gage, J. L. (2026). Investigating GERMs: How genotype, environment, and rhizosphere microbiome interactions underlie heat response in maize and sorghum. *New Phytologist*, *251*(4), 2152–2167.
[17] Bao, Y., Zhou, X., Chen, L., Feng, K., Guo, W., Huang, T., & Cai, Y. D. (2026). Machine learning identification of cell-type-specific molecular signatures distinguishing COVID-19 from other lower respiratory tract diseases. *Life*, *16*(5), 771.
[18] Sen, M. K., Roy, A., Varshney, R. K., & Chakraborty, A. (2026). Engineering next-generation crops through CRISPR-mediated horizontal gene transfer. *New Phytologist*, *249*(6), 2683–2689.
[19] Benachour, Y., Maloukh, L., Bouamama, S., & Geusens, B. (2025). Predicting skin concern severity from genetic and lifestyle factors: A comparative multi-output machine learning framework. *IEEE Access*, *13*, 207672–207693.
[20] Ayivi-Tosuh, S. M., Dofuor, A. K., Yamoah, J. A. A., Gayi, B. K., Aiduenu, A. F., Akomea, A., ... & Basing, L. A. (2026). Gut-microbiome interactions: Characterization, therapeutic implications and machine learning. *SAGE Open Pathology*, *19*, 30502098251415109.