UNIVERSITY OF LIFE SCIENCES ”KING MIHAI I” FROM TIMISOARA

”YOUNG PEOPLE AND MULTIDISCIPLINARY RESEARCH IN APPLIED LIFE SCIENCES”

05–06 November 2026



”Young people and multidisciplinary research in applied life sciences”

ANIMAL-LEVEL MACHINE LEARNING ASSESSMENT OF BOVINE RNA SEQUENCING QUALITY METRICS

Madalina MINCU-IORGA[1], Alexandru-Corneliu OLTEANU[2], Dinu GAVOJDIAN[1]*, Teddy LAZEBNIK[3]

[1] Cattle Production Systems and Welfare Laboratory, Research and Development Institute for Bovine, Balotesti, Romania
[2] Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Romania
[3] Computing Department, Jönköping University, Jönköping, Sweden
* Corresponding author: gavojdian_dinu@animalsci-tm.ro

Abstract

This study evaluated whether sample-level RNA-sequencing quality features could distinguish two seasonal study groups and established a reproducible framework for subsequent heat-stress research in dairy cattle. Fourteen whole-blood samples from seven Romanian Black and White Spotted cows (Holstein strain) underwent paired-end RNA sequencing, with one warm-season and one cold-season sample per animal. We harmonised GC content, Q30 and total alignment rate into a 14-by-3 feature matrix and compared six configurations of five machine-learning approaches using Leave-One-Cow-Out validation. CatBoost and a fully connected neural network (FCNN) achieved 85.7% accuracy, compared with 78.6% for Random Forest. These results support further investigation of seasonal group information in RNA-seq QC profiles and provide a reproducible foundation for modelling complete expression measurements.

Introduction

This study investigated whether RNA-sequencing quality metrics captured differences between two seasonal study groups in dairy cattle. We compared machine-learning classifiers using paired samples from seven cows and Leave-One-Cow-Out cross-validation to assess discrimination between BNR_A and BNR_S.

Materials and methods

The preceding workflow removed adapter-contaminated, N-rich and low-quality reads, then aligned retained reads to the bovine ARS-UCD1.2 reference using HISAT2. Gene quantification with featureCounts and differential-expression analysis with DESeq2 formed the wider bioinformatics analysis. For this study, we extracted predictors directly from the sample QC and mapping tables.

GC content	Percentage of guanine and cytosine bases. Describes sequence composition.
Q30	Percentage of bases at the Phred 30 threshold. Describes base-call quality.
Alignment rate	Percentage of analysed reads mapped to the reference. Describes mapping efficiency.

Animal-level validation and model specification

Seven Leave-One-Cow-Out folds trained on six cows (12 samples) and tested both samples of the remaining cow. Logistic regression, linear SVM and FCNN scaling used training-fold statistics. Random Forest and CatBoost used unscaled features. CatBoost used 200 depth-two trees. FCNN hidden layers contained 128, 64, 32 and 32 neurons, with ReLU activation and L-BFGS optimisation. All 128 within-cow label-swap assignments repeated the complete validation procedure.

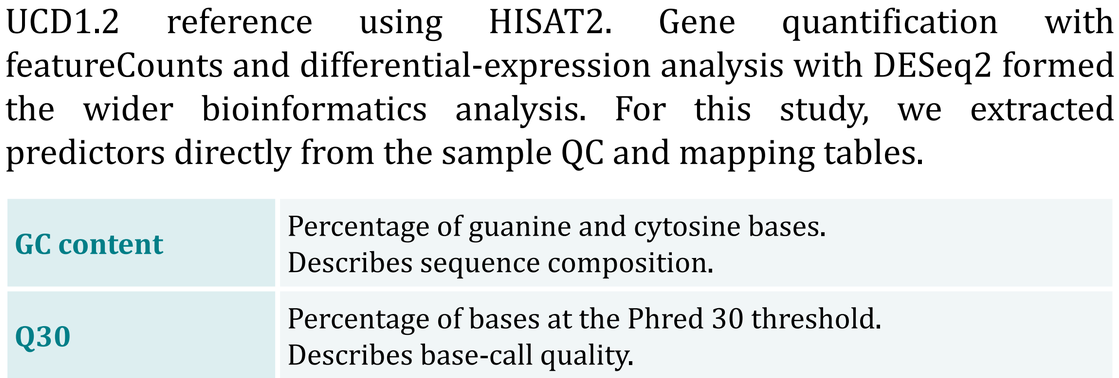


Figure 1. Input feature distributions by study group

Points show individual samples from the paired cohort. GC content had the largest mean absolute standardised SVM coefficient and mean Random Forest impurity importance. These describe contributions of QC variables to the fitted models.

Results and discussion

CatBoost and FCNN with α = 0.0001 each correctly classified 12/14 samples, one more than Random Forest. Their pooled ROC-AUC values were 0.959 and 0.939, respectively. Both retained 85.7% accuracy across six random seeds. The initial FCNN setting (α = 1) underfit, motivating a diagnostic follow-up with the library-default L2 strength and the same architecture.

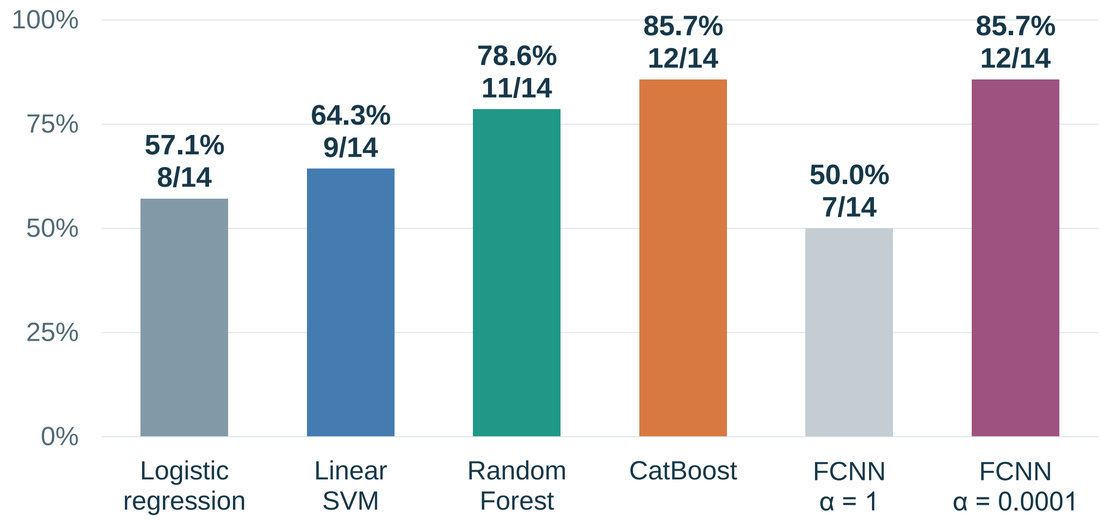


Figure 2. Classification accuracy across all configurations

Labels show the percentage and number of correctly classified samples. All predictions came from held-out cows. Both FCNN settings remain visible because the lower-L2 configuration followed the initial underfitting result.

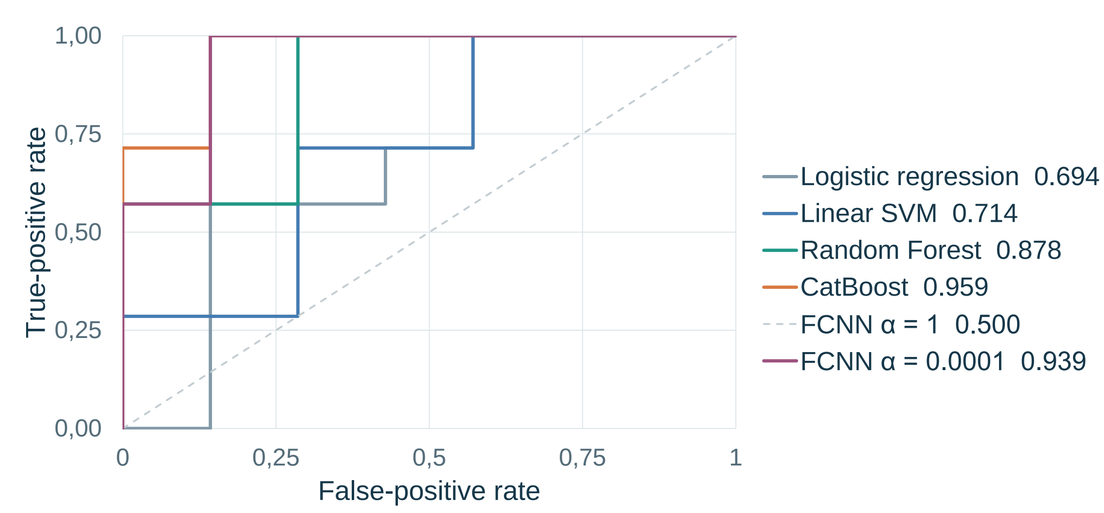


Figure 3. Discrimination from pooled out-of-fold scores

Legend values are pooled ROC-AUC. Curves combine scores from seven separately fitted folds and provide a descriptive comparison. The individual exact permutation p-values for CatBoost and the lower-L2 FCNN were both 0.03125.

Conclusions and next steps

The analysis established a reproducible QC benchmark with encouraging group discrimination. The maximum-across-six-configurations permutation test gave p = 0.0625, supporting an exploratory interpretation. QC metrics may reflect technical variation as well as RNA composition, so these results do not establish heat-stress biomarkers. Next steps are to quantify complete expression profiles from the existing FASTQs, evaluate gene-level predictors within animal-level validation and test the resulting signatures in independent animals.

Acknowledgement: This work was supported by COST Action CA22112, European Network on Livestock Phenomics (EU-LI-PHE), through a Virtual Mobility Grant (E-COST-GRANT-CA22112-aa042efe).