Supplementary Materials

0.1 Overview

This supplementary document provides additional performance results and analysis that complement the main manuscript. It includes extended classification results.

0.2 ROC Curve Analysis

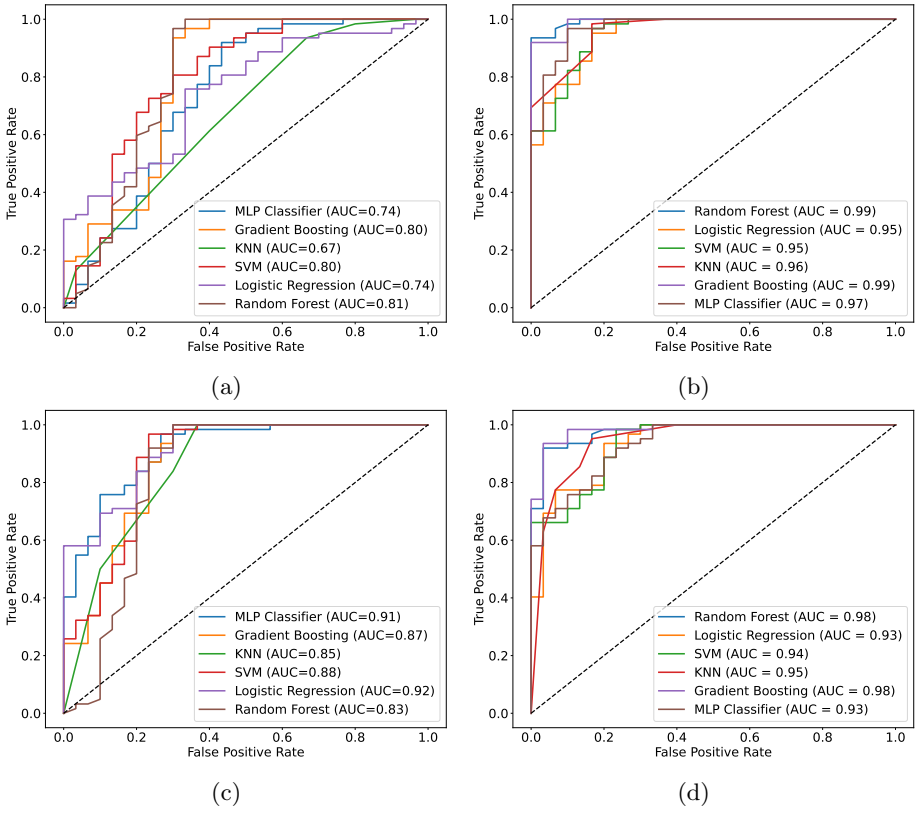

Figure S1: ROC curves comparing the classification performance of various models on poultry sausage datasets under different conditions. (a) Original poultry dataset with species variables. (b) Synthetic training with species variables. (c) Original dataset without species variables. (d) Synthetic training without species variables.

ROC Curve Analysis – Poultry Dataset

Figure S1 presents ROC curves comparing the classification performance of multiple machine learning models across different experimental configurations involving the poultry dataset. The AUC values are reported for each model to provide a quantitative measure of discriminative performance. In the scenario using the original dataset with species variables (Figure S1a), the Random Forest model achieved the highest AUC (0.81), followed by Gradient Boosting and SVM (both AUC = 0.80). In contrast, KNN showed the lowest AUC (0.67), suggesting limited predictive capability under this setting. When synthetic data were used for training (Figure S1b), all models exhibited significant performance improvements. Random Forest and Gradient Boosting reached AUC values of 0.99, while MLP and KNN achieved 0.97 and 0.96, respectively.

In the absence of species-level features, models trained and tested on the original data (Figure S1c) performed moderately well, with Logistic Regression (AUC = 0.92) and MLP (AUC = 0.91) showing the highest scores. However, in the synthetic-to-real transfer setting without species variables (Figure S1d), Random Forest and Gradient Boosting both reached AUC = 0.98, reinforcing the utility of synthetic data across feature-limited scenarios.

ROC Curve Analysis – Pork Dataset

Figure S2 shows the ROC curves for several machine learning models applied to the pork dataset under four different scenarios. In the original dataset with species variables (Figure S2a), Logistic Regression achieved the highest AUC (0.94), followed by MLP (0.90) and Random Forest (0.85). SVM and KNN performed less effectively, with AUCs of 0.71 and 0.75. In the synthetic-to-original transfer setting (Figure S2b), performance substantially improved, with Random Forest and Gradient Boosting reaching AUC = 0.98.

When excluding species-level features, models still performed well. Figure S2c shows KNN (AUC = 0.91), MLP, and Logistic Regression (both AUC = 0.90) as the top performers. With synthetic training (Figure S2d), AUC scores increased, and Random Forest achieved 0.97 while MLP reached 0.95. These findings highlight the robustness of synthetic data and its value in model generalization.

0.3 SHAP Analysis

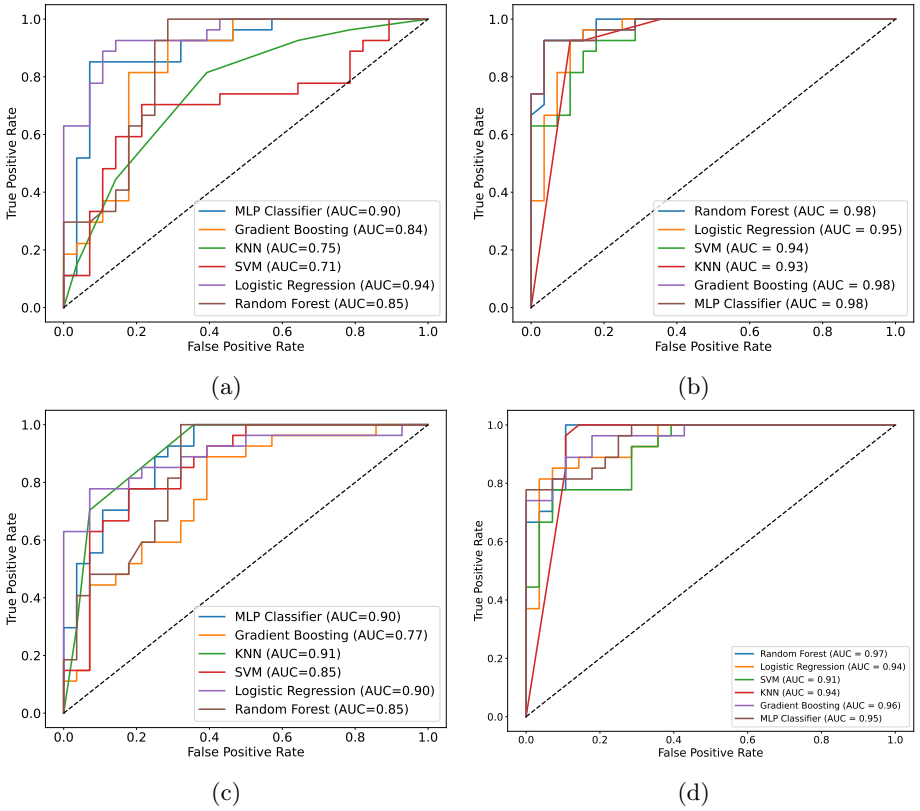

Figure S2: ROC curves comparing the classification performance of various models on pork sausage datasets under different conditions. (a) Original pork dataset with species variables. (b) Synthetic training with species variables. (c) Original dataset without species variables. (d) Synthetic training without species variables.

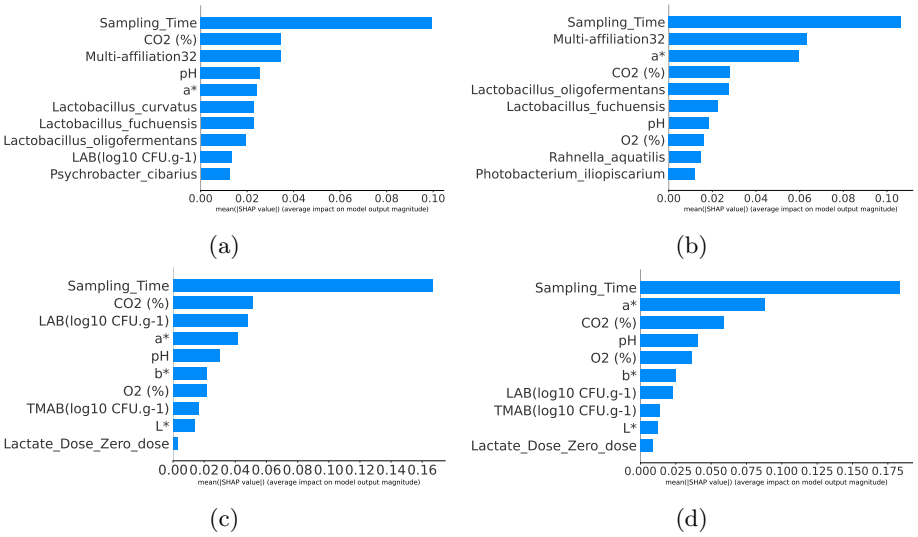

Figure S3: Bar plots showing the top 10 features contributing to model predictions in each poultry sausage dataset setting based on mean absolute SHAP values.

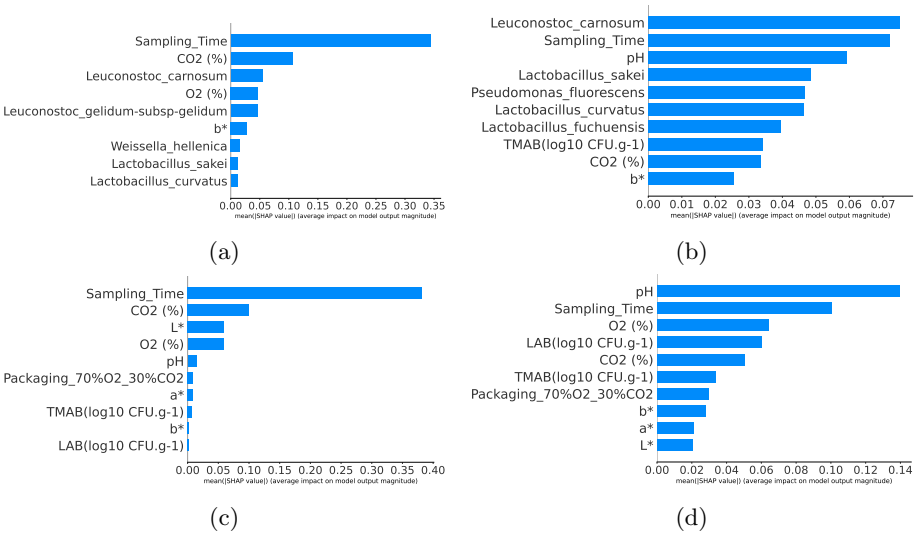

Figure S4: Bar plots showing the top 10 features contributing to the predictions of the highest-performing machine learning models in each pork sausage dataset setting, based on mean absolute SHAP values. The x-axis represents the average magnitude of SHAP values, indicating each feature's overall contribution to model predictions.