RESULTS

A] Logistic Regression Is As Effective As Complex Machine Learning Techniques In Predicting BPAs and Non- BPAs

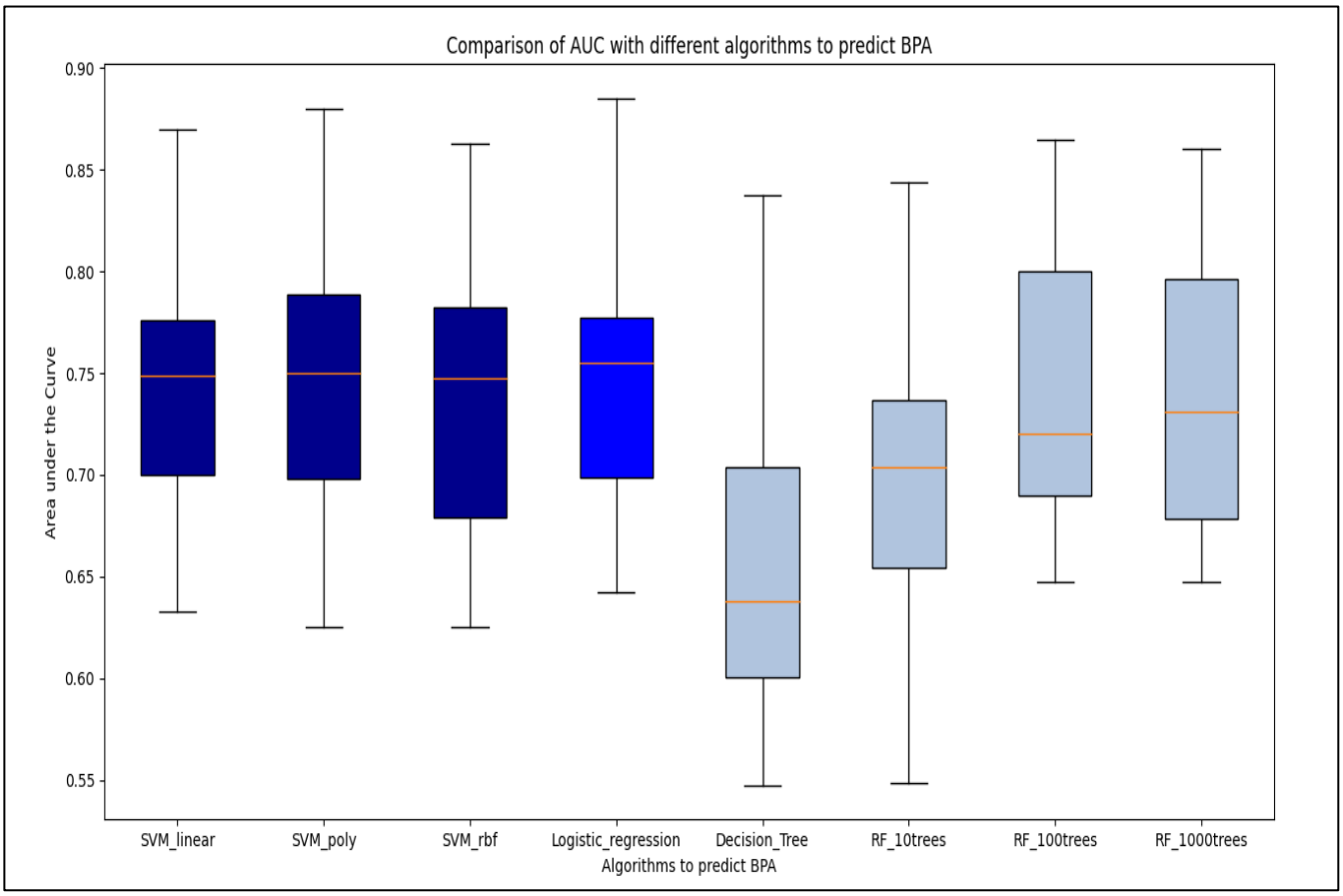

Figure 1: Comparison of area under the curve (AUCs) obtained in leave tenth out cross validation (LTOCV) of different machine learning algorithms predicting bacterial protective antigens (BPA) or non-BPAs from the curated dataset of BPAD200.

All classifiers were built using the previously documented top 10 features for BPA prediction in reverse vaccinology. Logistic regression (Mean ROC AUC: 0.75133) performed as well as complex ML techniques like SVM-RBF (Mean ROC AUC: 0.73966).

B] Pre-processing improves Classification in Reverse Vaccinology

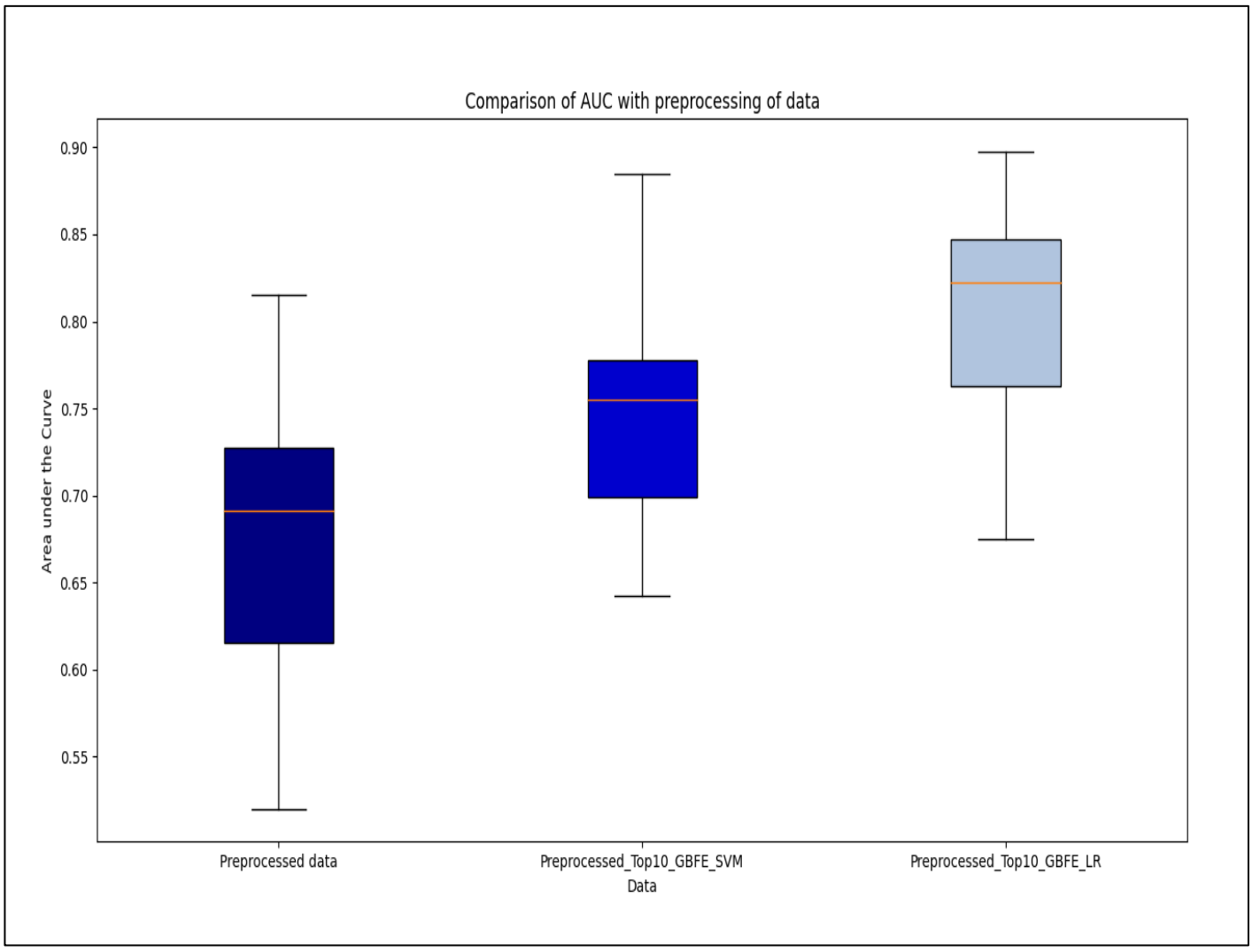

Figure 2: Comparison of how the pre-processing of the BPAD200 dataset impacts area under the curve (AUC) values obtained in reverse vaccinology (RV).

Boxplots of AUC values were obtained by leave tenth out cross validation (LTOCV) of logistic regression classifiers predicting bacterial protective antigens (BPA) or non-BPAs from BPAD200 dataset. From left to right; the raw dataset scaled between -1 and 1 for each feature, using ten features scaled between -1 and 1 selected using GBFE with a SVM-RBF, using 10 features scaled between -1 and 1 selected using greedy backward feature elimination (GBFE) with a logistic regression classifier.

C] Greedy Backward Feature Elimination and Permutation Analysis

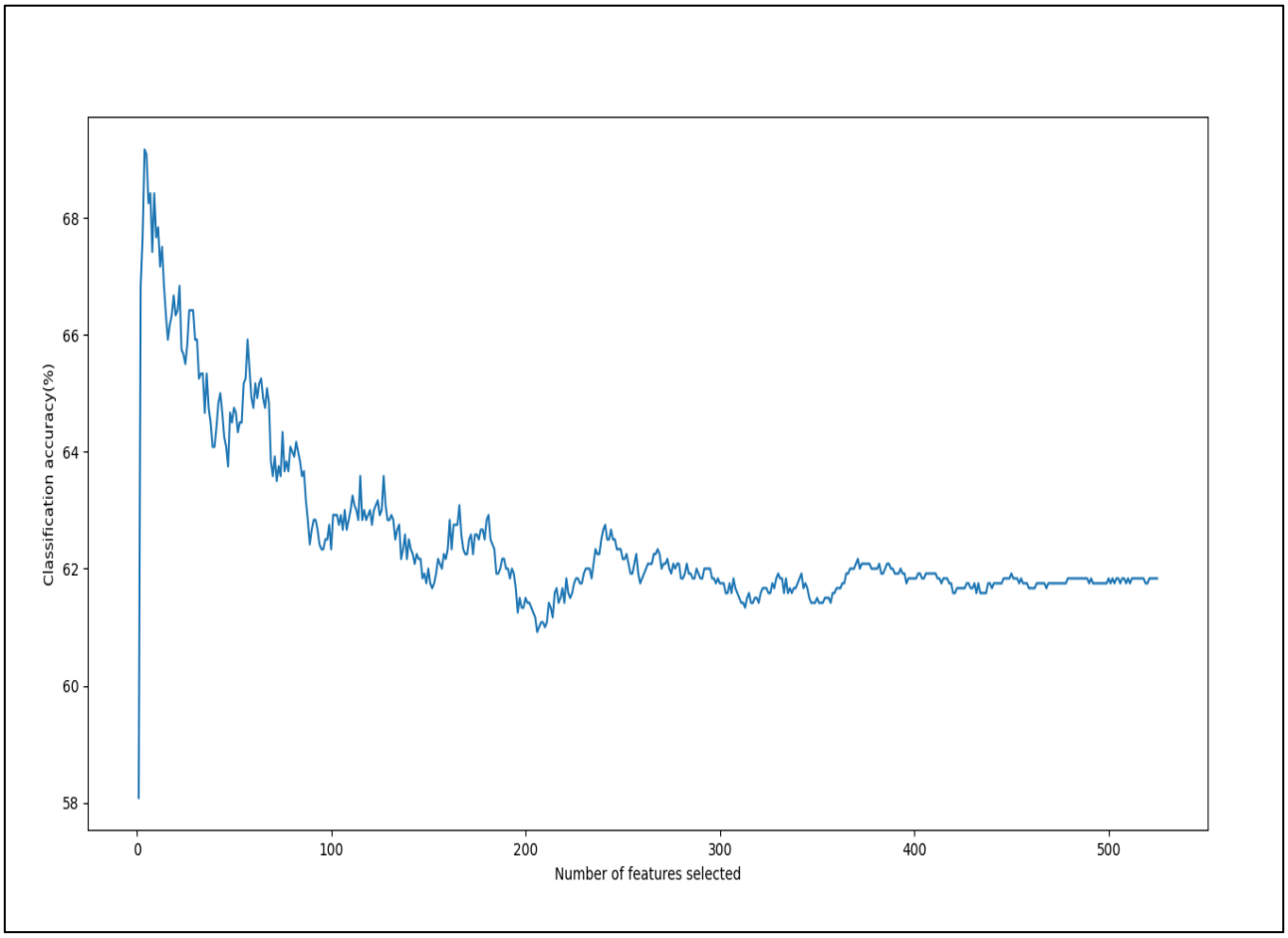

Figure 3: Accuracies obtained using a leave tenth out cross validation to predict bacterial protective antigens (BPAs) or non-BPAs from the dataset BPAD200.

Results from Greedy Backward Feature Elimination shows that increase in features from BPAD200 dataset results in reduction in accuracies when classifying BPAs from non- BPAs

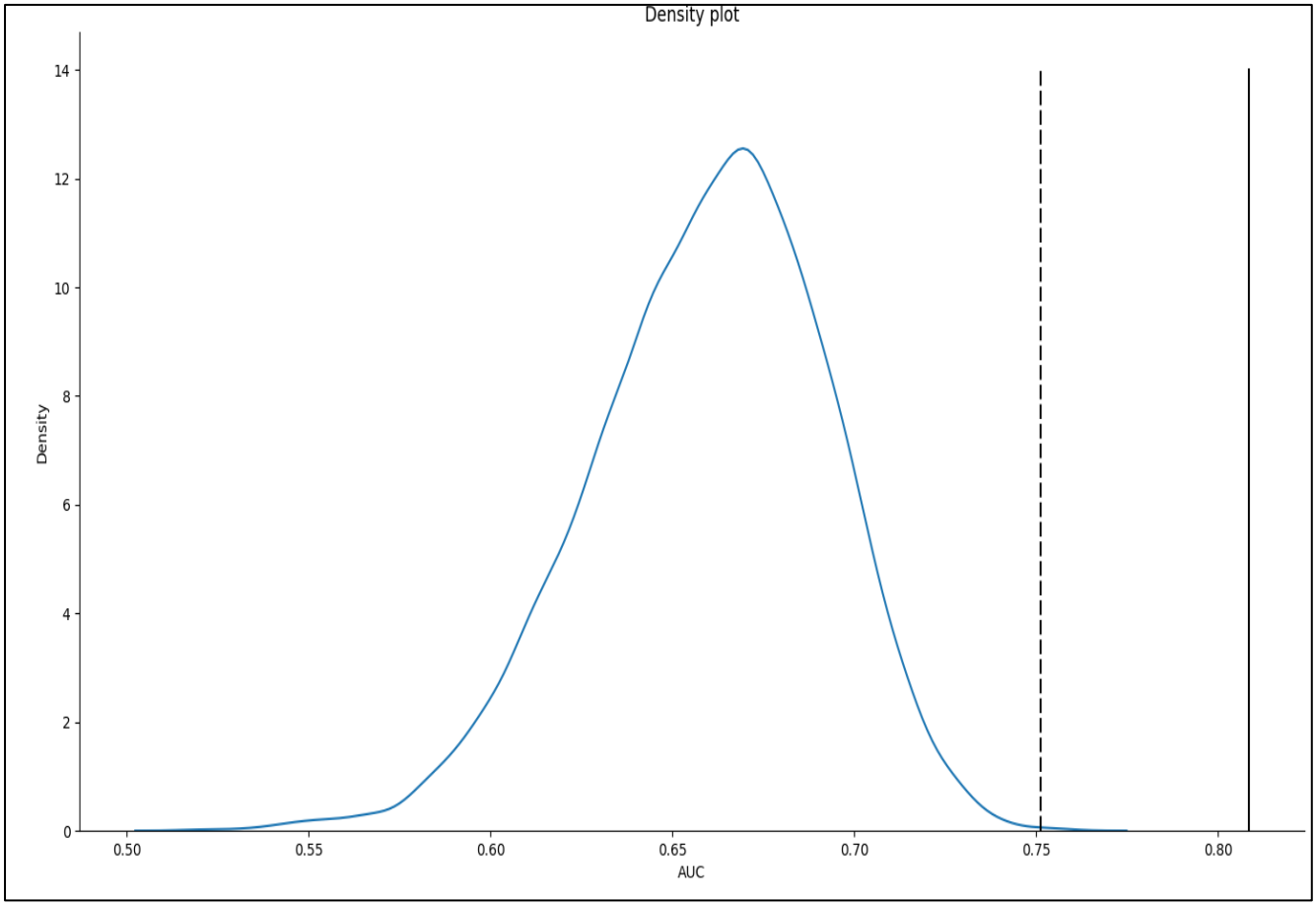

Figure 4: Area under the curve (AUC) obtained when classifying BPAs from non-BPAs using a logistic regression classifier comprised of ten features from the dataset BPAD200

10 features were selected randomly from BPAD200 dataset and logistic regression classifier was applied with Leave Tenth Out Cross Validation. This procedure was repeated 10000 times to get a frequency distribution.

The dotted vertical line represents the AUC obtained through LTOCV when using the top ten features obtained by GBFE and SVM-RBF technique while the solid vertical line is the AUC obtained from LTOCV using ten features derived from GBFE and logistic regression.

D] Leave One Bacteria Out Validation

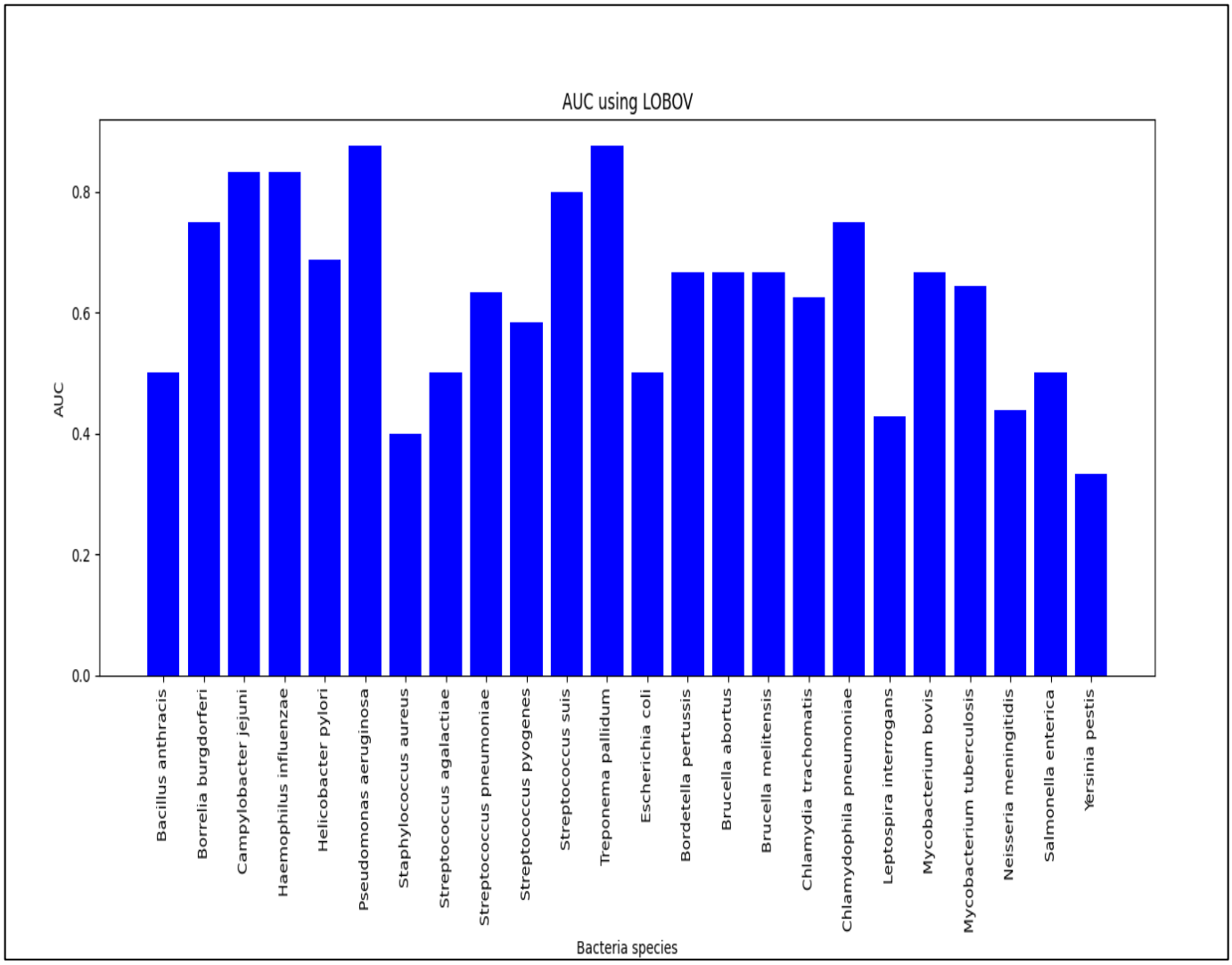

Figure 5: Area under the curve when predicting bacterial protective antigens (BPAs) and non-BPAs for each bacterial species that was left out of the training data and used as a test set, leave one bacteria out validation (LOBOV).