Machine Learning Enabled Early Diagnosis of HLB in Citrus

Dr. Yu Wang, Associate Professor

E-mail: yu.wang@ufl.edu

University of Florida, Institute of Food & Agricultural Sciences,
Citrus Research & Education Center

September. 25th, 2025

UF|IFAS
UNIVERSITY of FLORIDA

Early Detection for Citrus HuangLongBing(HLB)

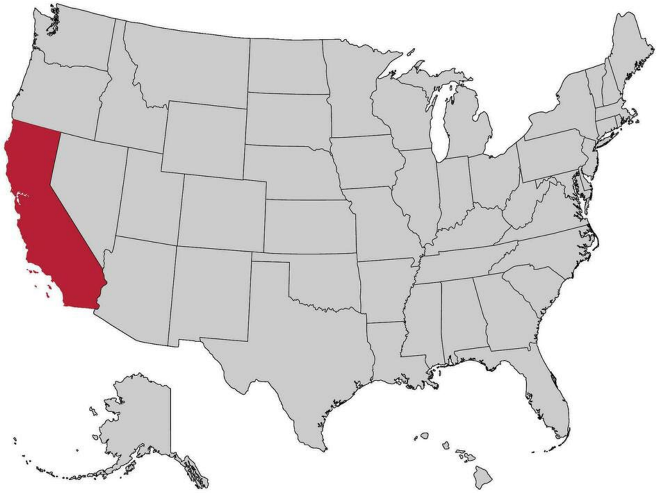
Destructive disease of citrus

- Bacterium: *Candidatus* Liberibacter asiaticus (CLas)
- Psyllid vector*: Diaphorina citri (D. citri)*

- Citrus production decreased from 250 million boxes in 2005 to 15 million boxes in 2024.
- qPCR for bacterial detection at 6 months

Early Detection for Citrus HuangLongBing(HLB)

CALIFONIA

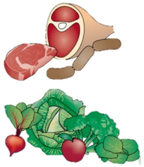


- Destroy infected trees
- Establish quarantines
- Prevent further spread

FLORIDA

- Evaluate novel treatments
- Identify tolerant/resistant cultivars
- Understand tree response

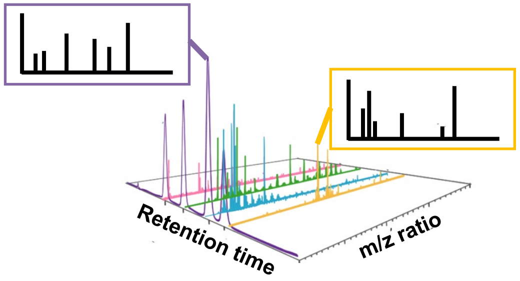
Nontargeted metabolomics workflow

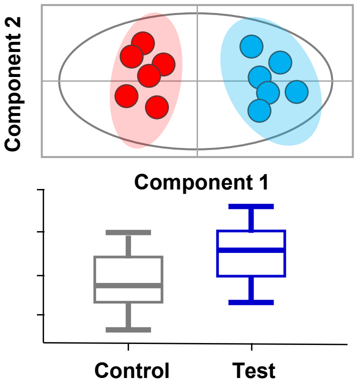


Sample preparation

Instrumental analysis

metabolites ~ 1,000
(tentatively identified)



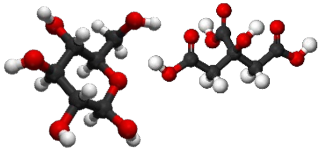


Raw data processing

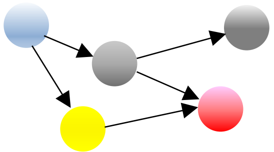
Metabolite identification

Database search

METLIN | mzCloud | NIST



Primary information



Hypothesis
(tentative confirmation)

Targeted metabolomics

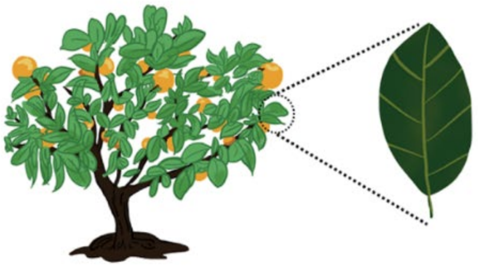
Plant materials

‘Midsweet’ sweet orange trees planted in greenhouse

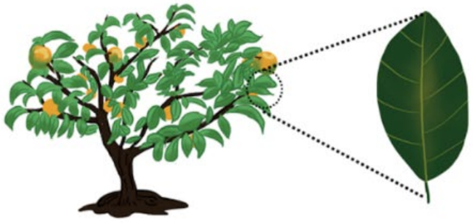
Budwoods source:

1) ‘Midsweet’ budwood was grafted onto US-802 rootstock & grown for one year
2) Seedlings were inoculated with scions from completely pathogen-free & seriously *Ca*Las-infected sour oranges;
3) After Passing qPCR test, budwood was cut for grafting

Sampling: 7 weeks post-exposure, 10 to 14 leaves were randomly collected from each of 12 individual healthy & infected trees.



healthy tree



HLB-affected tree

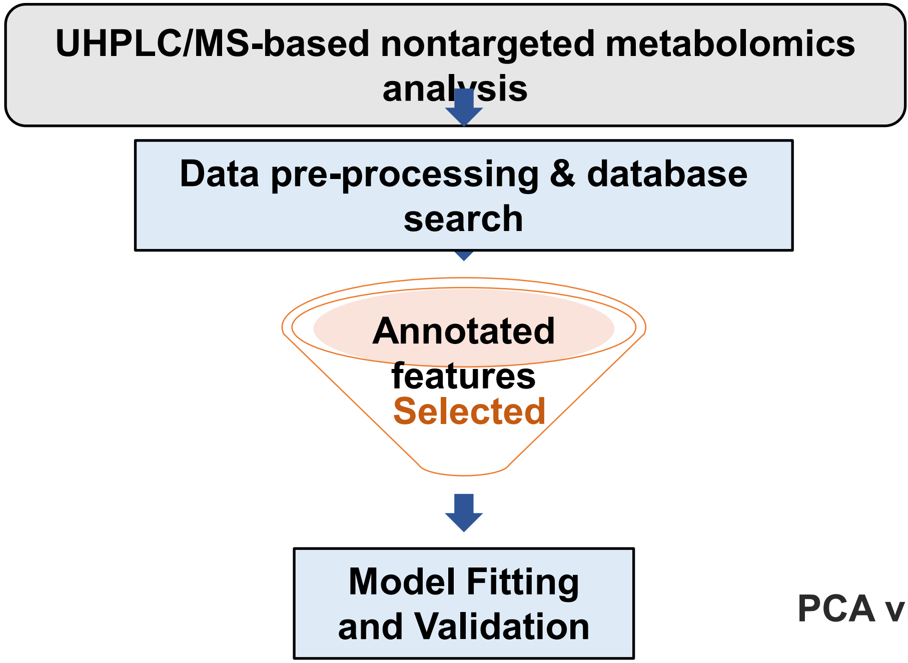
DATA CHALLENGE

Acquisition

No HLB-free trees in the grove in Florida anymore

Greenhouse: 24 Samples (12 HLB-affected, 12 healthy)

7 Weeks



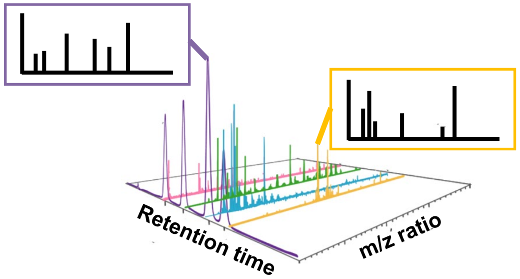
UHPLC/MS-based nontargeted metabolomics analysis

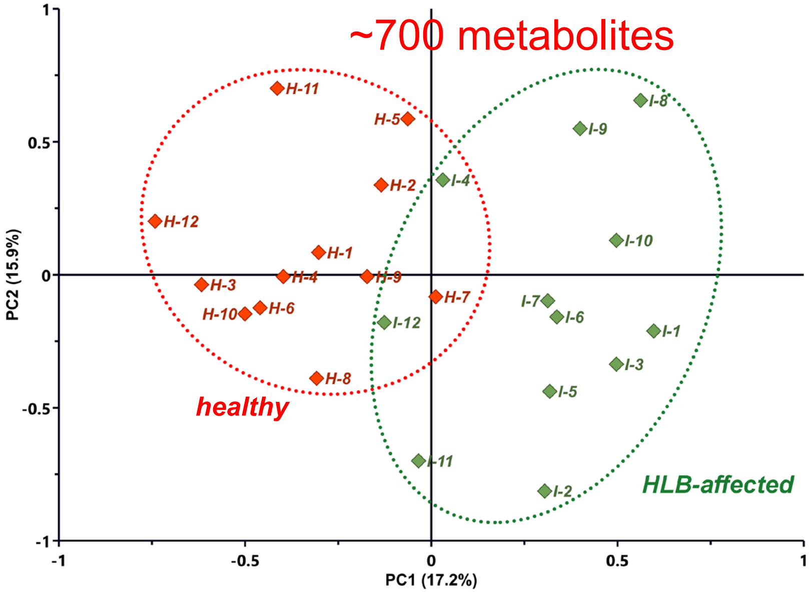
Data pre-processing & database search

Annotated features Selected

Model Fitting and Validation

Analysis



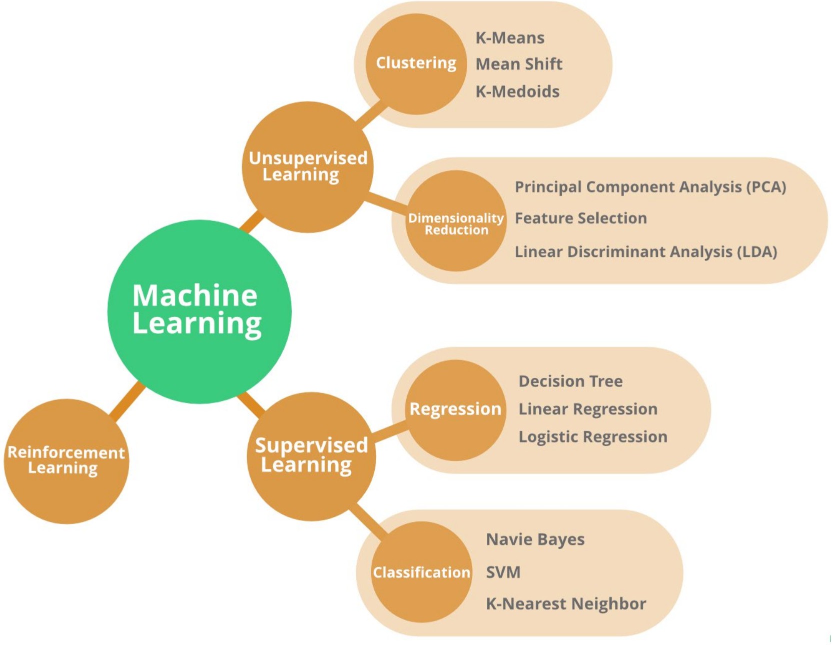


~700 metabolites

PCA visualizes differences between citrus trees of healthy group & HLB-affected group

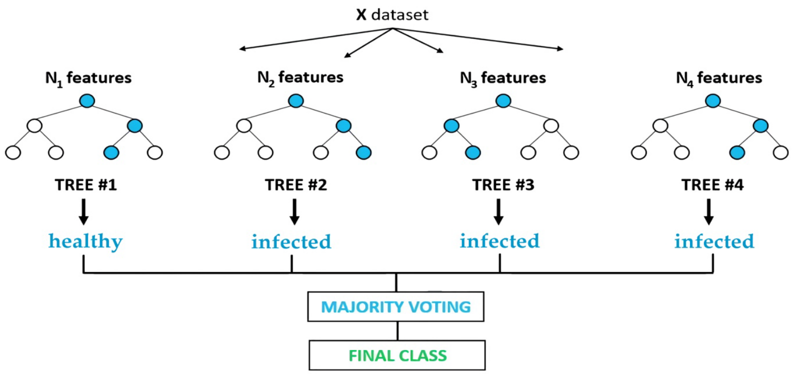
Machine learning (ML)

- **Manage ultra-high-dimensional data** – filter noise, select key features.
- **Detect complex, non-linear patterns** between biomarkers and disease.
- **Boost prediction accuracy** for early, reliable HLB detection.

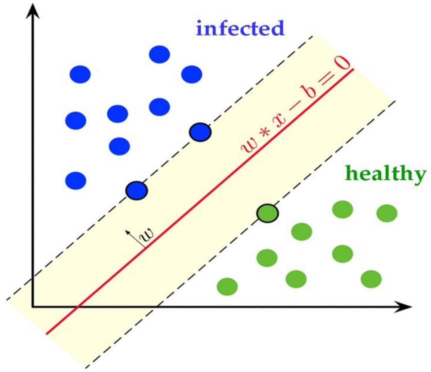


Model Selection

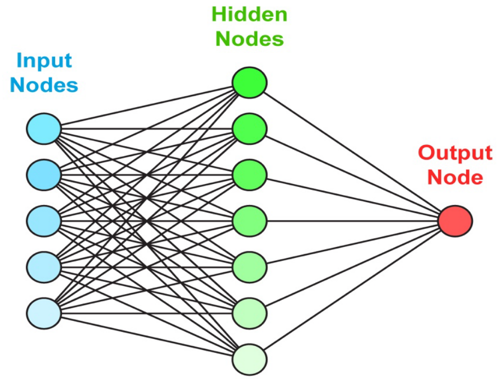
Select appropriate ML algorithms



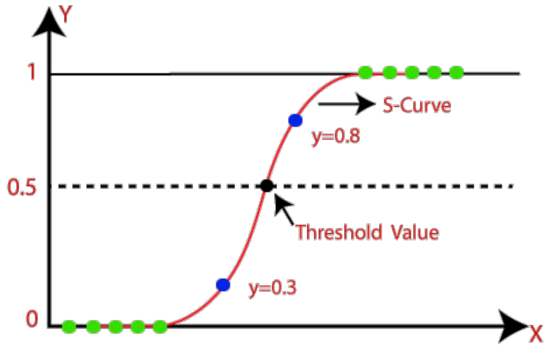
Random Forest



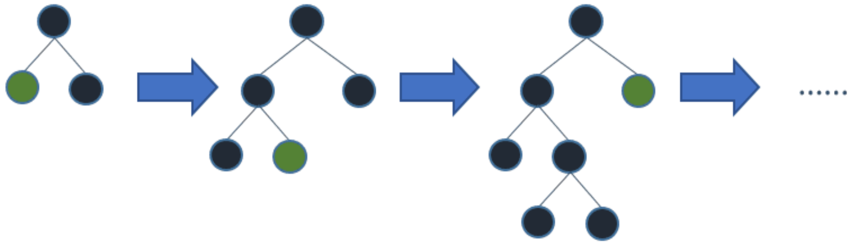
Support Vector Machine



Multi-Layer Perceptron

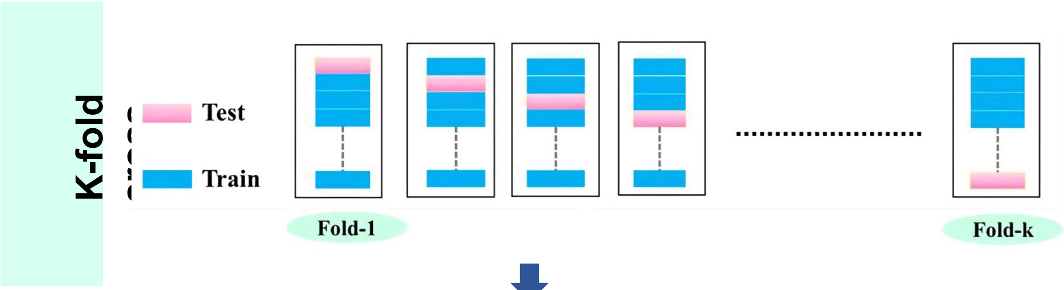


Logistic Regression (L1/L2)



Gradient-Boosted Decision Tree

Modeling & Performance Evaluation



Best threshold for ML modeling

Comparison of different state-of-the-art ML classifiers

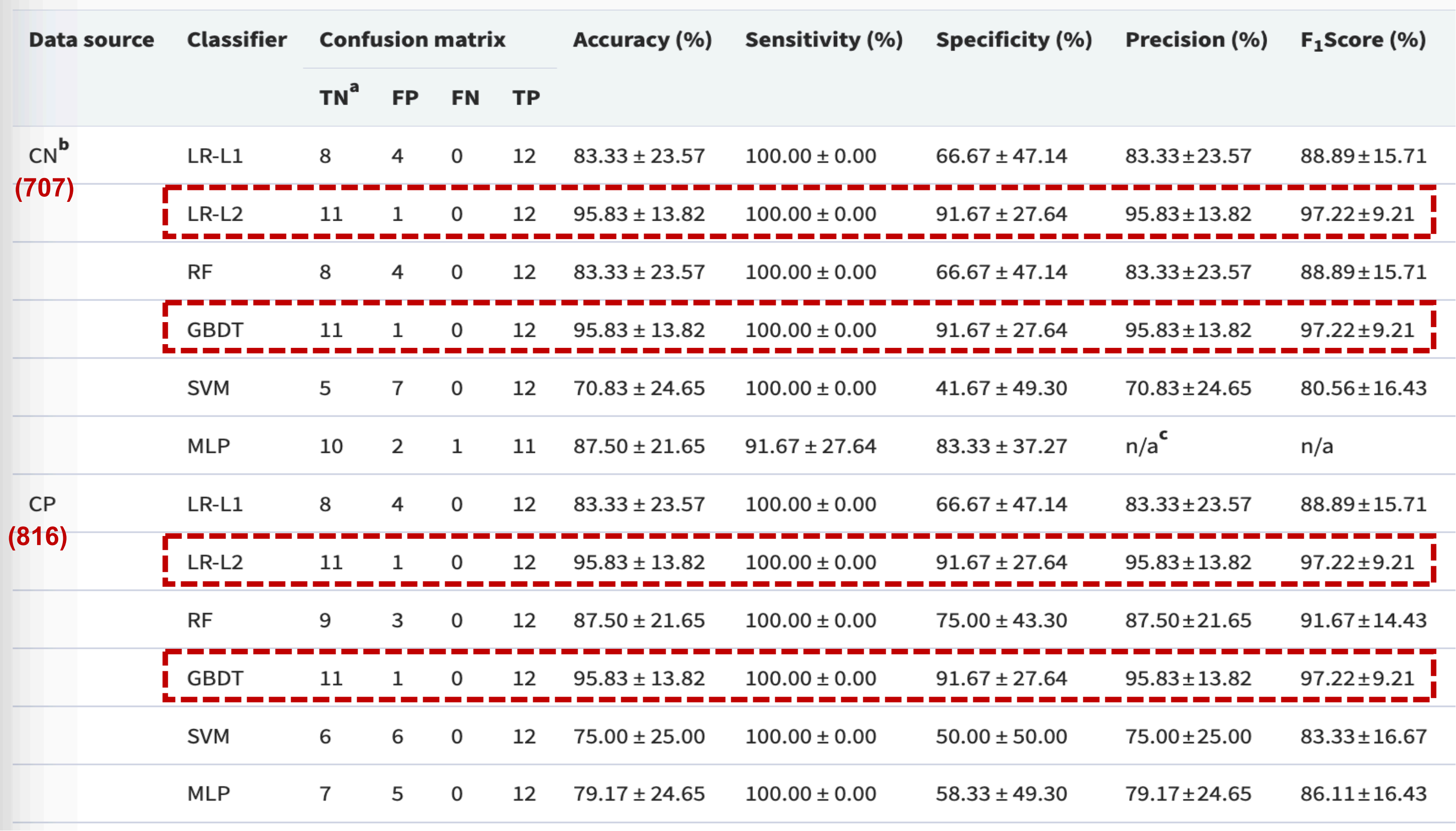
Accuracy | Sensitivity | Specificity | Precision | F_1 Score | MCC | AUC

Optimal classifier(s)

Top-ranked metabolite features (biomarkers)

Modeling & Performance evaluation

Mean performance metrics of six ML classifiers based on 1925 metabolite features in four datasets



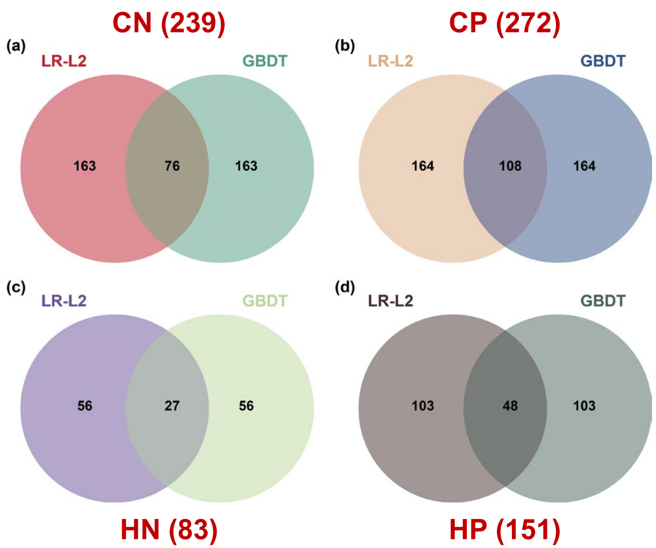
Data source	Classifier	Confusion matrix				Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F_1Score (%)
		TN[a]	FP	FN	TP					
CN[b] (707)	LR-L1	8	4	0	12	83.33 ± 23.57	100.00 ± 0.00	66.67 ± 47.14	83.33±23.57	88.89±15.71
	LR-L2	11	1	0	12	95.83 ± 13.82	100.00 ± 0.00	91.67 ± 27.64	95.83±13.82	97.22±9.21
	RF	8	4	0	12	83.33 ± 23.57	100.00 ± 0.00	66.67 ± 47.14	83.33±23.57	88.89±15.71
	GBDT	11	1	0	12	95.83 ± 13.82	100.00 ± 0.00	91.67 ± 27.64	95.83±13.82	97.22±9.21
	SVM	5	7	0	12	70.83 ± 24.65	100.00 ± 0.00	41.67 ± 49.30	70.83±24.65	80.56±16.43
	MLP	10	2	1	11	87.50 ± 21.65	91.67 ± 27.64	83.33 ± 37.27	n/a[c]	n/a
CP (816)	LR-L1	8	4	0	12	83.33 ± 23.57	100.00 ± 0.00	66.67 ± 47.14	83.33±23.57	88.89±15.71
	LR-L2	11	1	0	12	95.83 ± 13.82	100.00 ± 0.00	91.67 ± 27.64	95.83±13.82	97.22±9.21
	RF	9	3	0	12	87.50 ± 21.65	100.00 ± 0.00	75.00 ± 43.30	87.50±21.65	91.67±14.43
	GBDT	11	1	0	12	95.83 ± 13.82	100.00 ± 0.00	91.67 ± 27.64	95.83±13.82	97.22±9.21
	SVM	6	6	0	12	75.00 ± 25.00	100.00 ± 0.00	50.00 ± 50.00	75.00±25.00	83.33±16.67
	MLP	7	5	0	12	79.17 ± 24.65	100.00 ± 0.00	58.33 ± 49.30	79.17±24.65	86.11±16.43

HN (249) …

HP (453) …

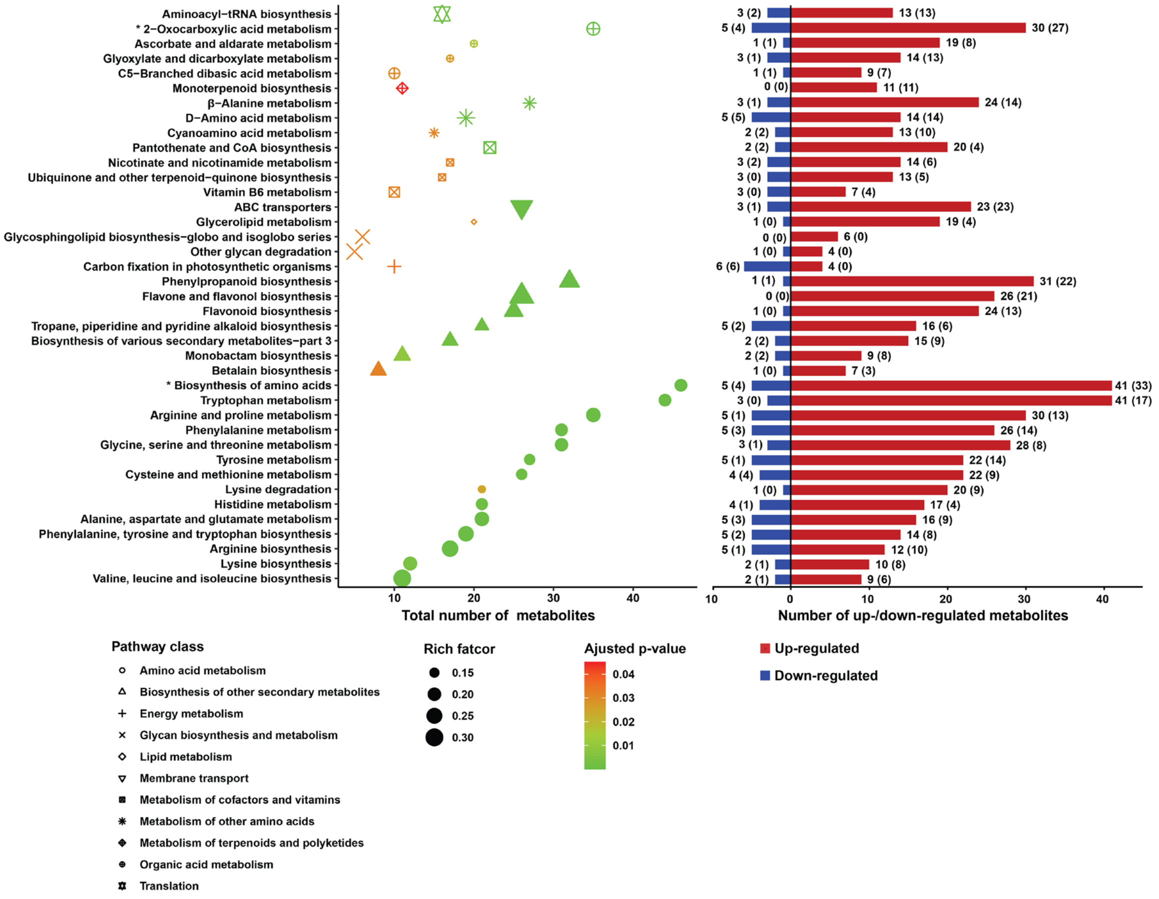
Feature Selection and Validation: Top-ranked metabolic biomarkers

Feature Selection by ML Models



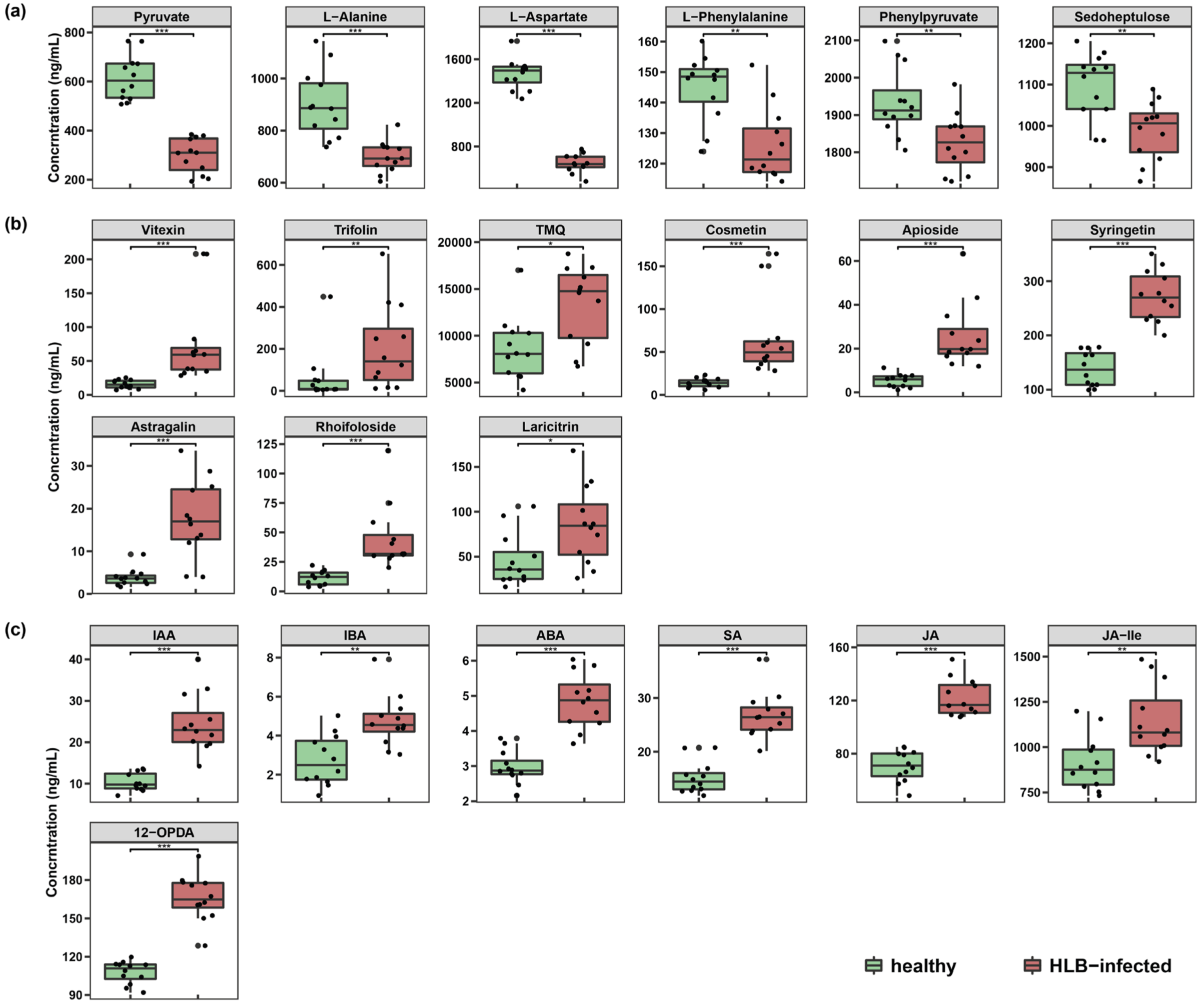
Counts of metabolic biomarkers resulted from two optimal ML classifiers

Structure Confirmation
Reduce Features by another 50%



Discovered 14 significant metabolic pathways related to HLB & number of up-/down-regulated metabolites

Model validation: Top-ranked metabolic biomarkers



Carbon fixation in photosynthetic organisms

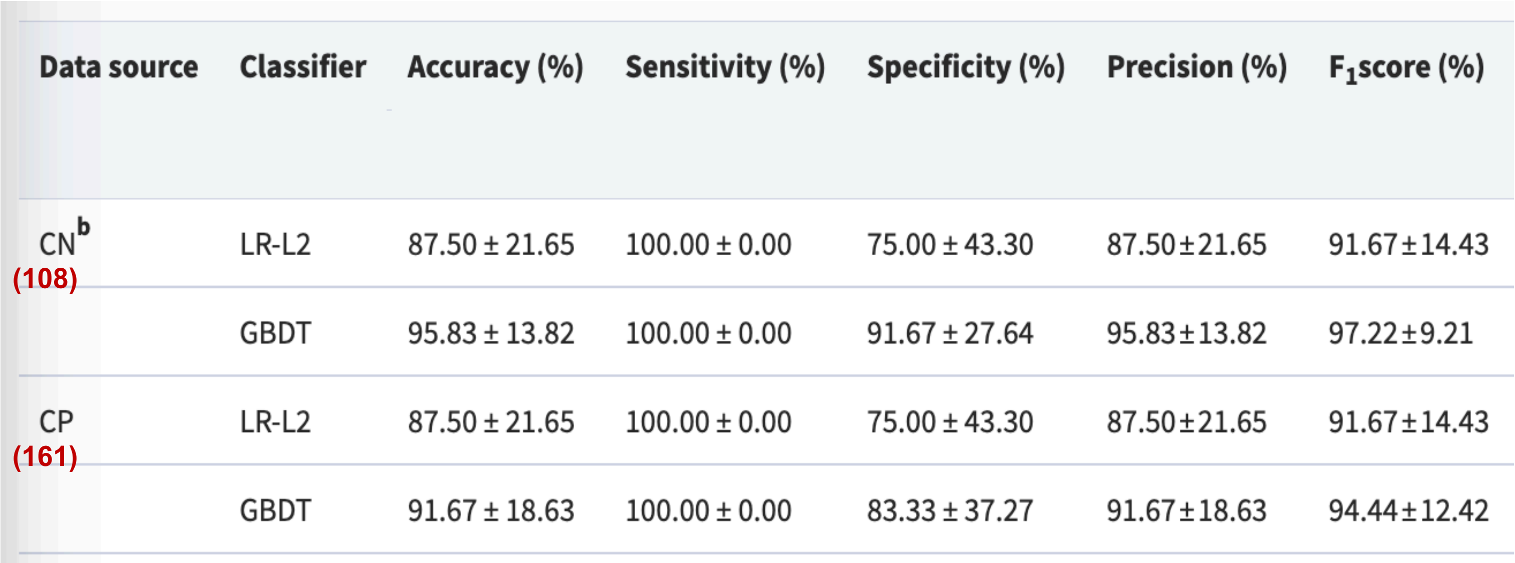
Flavone & flavonol biosynthesis

Plant hormone signal transduction

Content variation of some representative metabolites in three significant pathways

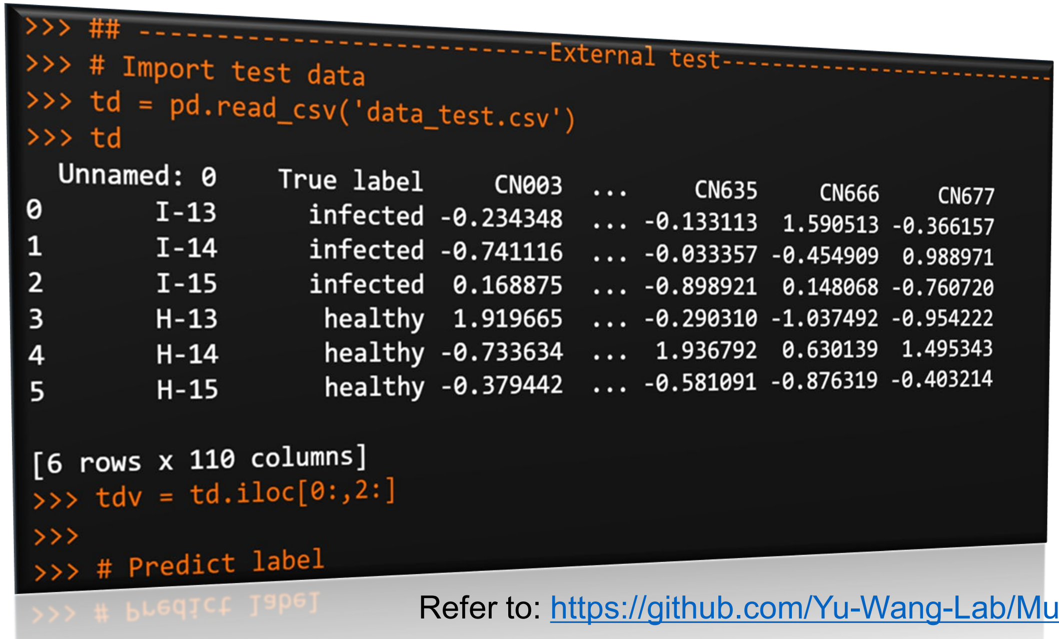
Final predictive model

Mean performance metrics of two optimal ML models based on identified metabolic biomarkers



Data source	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F_1score (%)
CN[b] (108)	LR-L2	87.50 ± 21.65	100.00 ± 0.00	75.00 ± 43.30	87.50±21.65	91.67±14.43
	GBDT	95.83 ± 13.82	100.00 ± 0.00	91.67 ± 27.64	95.83±13.82	97.22±9.21
CP (161)	LR-L2	87.50 ± 21.65	100.00 ± 0.00	75.00 ± 43.30	87.50±21.65	91.67±14.43
	GBDT	91.67 ± 18.63	100.00 ± 0.00	83.33 ± 37.27	91.67±18.63	94.44±12.42

External validation



```
>>> ## ------------------------------External test-----------------------------
>>> # Import test data
>>> td = pd.read_csv('data_test.csv')
>>> td
  Unnamed: 0  True label     CN003  ...     CN635     CN666     CN677
0       I-13    infected -0.234348  ... -0.133113  1.590513 -0.366157
1       I-14    infected -0.741116  ... -0.033357 -0.454909  0.988971
2       I-15    infected  0.168875  ... -0.898921  0.148068 -0.760720
3       H-13     healthy  1.919665  ... -0.290310 -1.037492 -0.954222
4       H-14     healthy -0.733634  ...  1.936792  0.630139  1.495343
5       H-15     healthy -0.379442  ... -0.581091 -0.876319 -0.403214

[6 rows x 110 columns]
>>> tdv = td.iloc[0:,2:]
>>>
>>> # Predict label
```

Refer to: https://github.com/Yu-Wang-Lab/Multiple_ML_modeling_for_HLB_prediction

Result of external validation for two optimal ML methods

Sample	True label	LR-L2		GBDT	
		Predicted label	HLB prob. (%)	Predicted label	HLB prob. (%)
I-13	infected	infected	96.47%	infected	99.88%
I-14	infected	infected	98.48%	infected	99.88%
I-15	infected	infected	99.75%	infected	99.88%
H-13	healthy	healthy	0.85%	healthy	0.12%
H-14	healthy	healthy	1.17%	healthy	1.98%
H-15	healthy	healthy	0.81%	healthy	1.12%

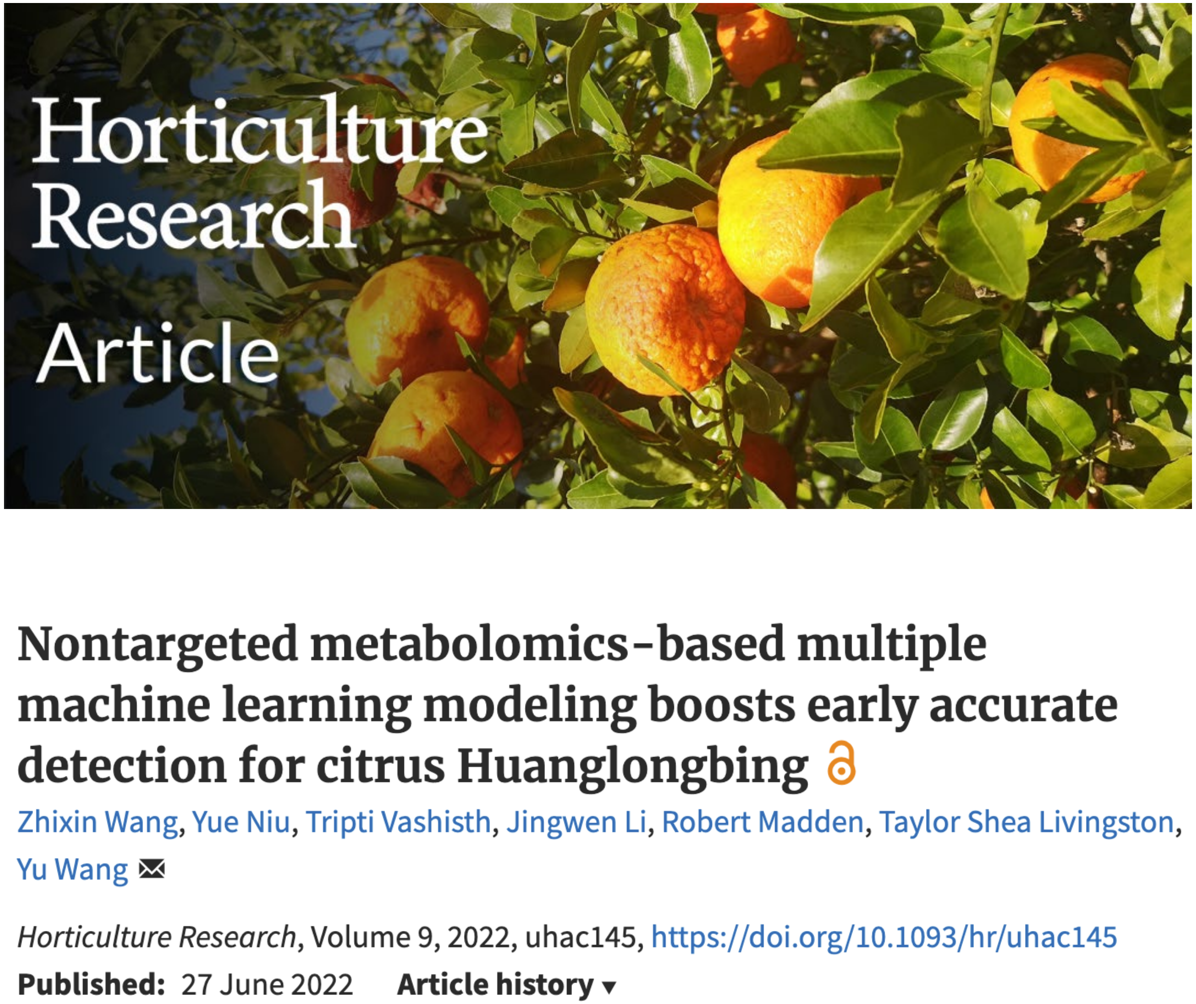
Take-Home Message

- A **new approach** combining UHPLC/MS-based metabolomics with machine learning for the **early detection of HLB** was developed.

- **Six ML models** were tested; **Logistic Regression (L2)** and **Gradient-Boosted Decision Trees** achieved the **best performance (95.8% accuracy)**.

- This strategy **overcomes limitations** of conventional methods (low sensitivity) and avoids issues with image-based ML.

- **Key biomarkers were confirmed** through pathway and differential analysis, showing strong consistency with previous studies.



Nontargeted metabolomics-based multiple machine learning modeling boosts early accurate detection for citrus Huanglongbing

Zhixin Wang, Yue Niu, Tripti Vashisth, Jingwen Li, Robert Madden, Taylor Shea Livingston, Yu Wang ✉

Horticulture Research, Volume 9, 2022, uhac145, https://doi.org/10.1093/hr/uhac145

Published: 27 June 2022 **Article history** ▾

ACKNOWLEDGEMENT

UF | IFAS
UNIVERSITY *of* FLORIDA

CRDF Citrus Research and Development Foundation, Inc.

USDA United States Department of Agriculture

New Varieties Development
NVDMC
& Management Corporation