

A Machine Learning-Based Multi-Objective Framework for Optimizing Fruit-Peel Waste-Derived Bio-Pesticide Formulations

Mithra R¹, Harini L²

M.Sc. Computer Science with Data Analytics, Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India¹

M.Sc. Computer Science with Data Analytics, Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India²

Abstract: Fruit-processing waste can be a cheap source of biologically active plant compounds, but turning it into a workable bio-pesticide depends on choosing the right composition and operating conditions. This study presents a prediction-and-optimization workflow for 200 fruit-peel formulations, in which regression models estimate Mortality % and Repellency %. Five models were compared: Random Forest, Gradient Boosting, XGBoost, Support Vector Regression (SVR), and Extra Trees. In the saved five-model evaluation, which used the available five-predictor CSV, XGBoost gave the best held-out test results for Mortality % ($R^2 = 0.854$, MAE = 3.049, RMSE = 3.621), and Gradient Boosting did best for Repellency % ($R^2 = 0.773$, MAE = 3.631, RMSE = 4.795). These figures still need to be reconciled with the seven-predictor feature set described in the original manuscript before final submission. The earlier Pareto analysis kept 19 non-dominated formulations and picked FPF_177 as a computational screening candidate, though biological confirmation is still needed.

Keywords: fruit-peel waste; bio-pesticide; machine learning; XGBoost; Support Vector Regression; Extra Trees; Pareto optimization.

I. INTRODUCTION

Crop losses from insect pests have kept farmers dependent on synthetic pesticides for a long time. Heavy, broad-spectrum use of these chemicals, though, brings environmental and ecological problems, which has pushed interest toward alternative control methods [1]. Plant-based pesticides are one such option, since botanical compounds can kill, deter, or repel pests. Even so, their development faces hurdles such as variable biological activity, inconsistent formulations, production cost, and regulatory requirements [2].

Meanwhile, fruit markets and processing units generate large volumes of peel waste. Citrus and other fruit peels contain essential oils, terpenoids, flavonoids, and phenolics, which makes them appealing for value-added pest management [3]–[6]. Recent reviews also describe improved extraction methods for recovering bioactive compounds from fruit and vegetable peels [7]. Experimental work has shown that essential oils from citrus peel waste have insecticidal activity [3], and a recent review examined the pesticidal potential of citrus fruits and their processing residues [3].

Formulation variables interact in nonlinear ways, which makes this problem a good fit for machine-learning models. Ensemble decision-tree methods work well with mixed tabular predictors because they pick up interactions without needing a biological equation specified beforehand [8]–[10]. XGBoost builds on boosted trees by adding regularization and an efficient implementation [11]. Recent reviews also point to the growing role of machine learning in precision crop protection and pest-management decision support [12]. Still, an accurate predictor does not by itself settle the practical trade-off between efficacy and resource use.

To address this, the study pairs regression-based prediction with Pareto screening. It aims to compare five regression models on the two efficacy responses, to retain formulations that offer good predicted efficacy at lower concentration and shorter exposure, and to identify which formulation variables drive the predictions most.

II. RELATED WORK

Research on botanical pest control shows that plant-derived products can act as insecticides, deterrents, or repellents, but success depends on practical formulation factors as well as biological activity [2], [13], [14]. For that reason, fruit-derived residues are better assessed through systematic candidate screening than through isolated trial-and-error experiments.

Three established ensemble strategies offer complementary ways to approach this prediction problem. Random Forest averages many randomized trees to generalize better [8], while gradient boosting builds trees one after another, each correcting the errors of the last [9]. XGBoost is a regularized, efficient version of boosted-tree learning [11]. These models suit formulation records that mix categorical choices with numerical operating conditions [10].

When several desirable outcomes compete, a single ranking score can hide important trade-offs. Pareto analysis avoids this by keeping a candidate whenever no alternative beats it on every objective at once. This non-dominated idea is central to multi-objective search methods [15]. Here it lets efficacy and resource use be weighed together, without imposing an arbitrary weighting scheme at the initial screening stage.

III. METHODOLOGY

A. Dataset and Preprocessing

The computational study used 200 formulation observations. Seven variables (Fruit peel, Extract type, Pest, Concentration, Exposure time, Formulation_Type, and Peel_Count) served as predictors, and Mortality % and Repellency % were treated as continuous responses. A data check found no missing records and no duplicate rows. Categorical fields were one-hot encoded with the first level dropped and unseen levels handled safely, and numerical fields were standardized. The records were split 80:20 into 160 training and 40 test observations, with random_state = 42. Table I summarizes the principal numerical variables.

TABLE I DESCRIPTIVE STATISTICS FOR THE PRINCIPAL NUMERICAL VARIABLES (N = 200)

Variable	Role	Min	Max	Mean	SD
Concentration (%)	Predictor	1.00	10.00	5.20	3.27
Exposure time (h)	Predictor	6.0	48.0	19.80	15.46
Mortality (%)	Target	28.50	79.80	59.01	9.60
Repellency (%)	Target	32.50	84.90	61.59	10.47

B. Predictive Modeling

Each response was modeled separately with Random Forest, Gradient Boosting, XGBoost, Support Vector Regression (SVR), and Extra Trees Regressor. The original manuscript reports five-fold GridSearchCV for the original three-model experiment. The SVR and Extra Trees results shown here come from the saved five-model evaluation, which used an 80:20 train-test split, and it has not been verified that their settings were tuned identically. To avoid data leakage, encoding and scaling should be done inside the training pipelines. The final comparison uses MAE, RMSE, and R² on the held-out test set (Table II), and Fig. 1 compares the XGBoost predictions with the actual values.

TABLE II HELD-OUT TEST PERFORMANCE OF FIVE REGRESSION MODELS

Model	Target	MAE	RMSE	R ²
Random Forest	Mortality	3.646	4.726	0.752
Gradient Boosting	Mortality	3.202	3.998	0.823
XGBoost	Mortality	3.049	3.621	0.854
SVR	Mortality	3.502	4.189	0.805
Extra Trees	Mortality	4.525	5.858	0.619
Random Forest	Repellency	4.052	5.425	0.709
Gradient Boosting	Repellency	3.631	4.795	0.773
XGBoost	Repellency	3.726	4.845	0.768
SVR	Repellency	4.624	5.637	0.686
Extra Trees	Repellency	4.479	5.845	0.662

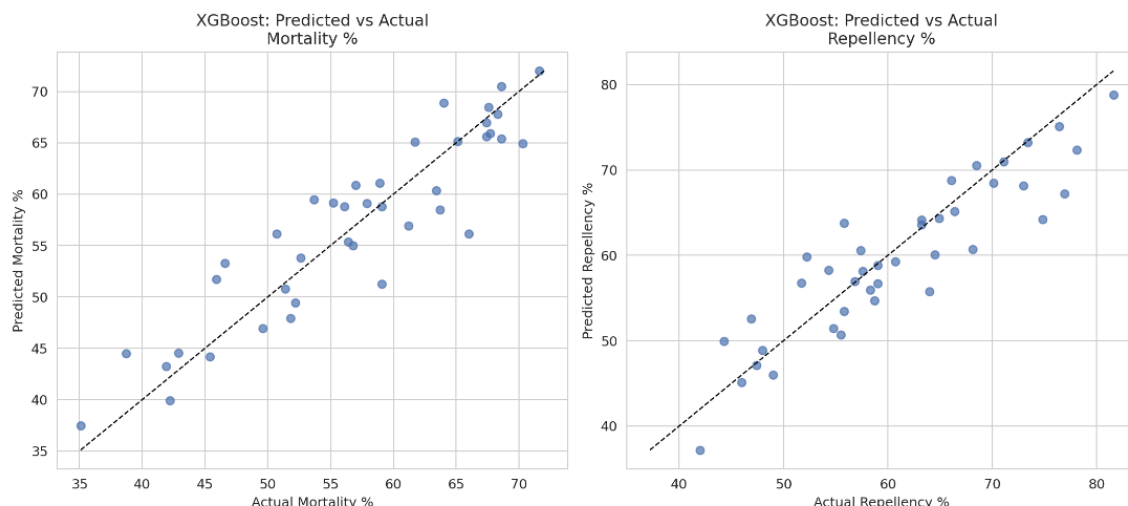


Fig. 1. XGBoost predicted vs. actual Mortality % (left) and Repellency % (right).

C. Multi-Objective Pareto Optimization

After model selection, XGBoost predictions were generated for all 200 formulation records. Predicted mortality and repellency were treated as objectives to maximize, while concentration and exposure duration were to be minimized. So that a standard minimization-based Pareto calculation could be used, the two efficacy predictions were sign-reversed and the resource variables were left as they were. Pairwise dominance testing then gave 19 non-dominated candidates, or 9.5% of the available formulations.

D. Feature-Importance Analysis

The selected mortality model was also checked for feature importance. Contributions from one-hot encoded levels were combined back under their original parent variables. The resulting percentages (Table III) show each predictor's impurity-based contribution to the model, so they reflect predictive importance and not evidence that a variable directly causes the biological response.

TABLE III RELATIVE FEATURE IMPORTANCE FOR MORTALITY % PREDICTION

Feature	Relative importance (%)
Fruit-Peel Combination	55.67
Pest	13.45
Formulation_Type	9.01
Concentration	8.90
Extract type	8.74
Exposure time	3.07
Peel_Count	1.15

IV. RESULTS AND DISCUSSION

A. Exploratory Patterns

Exploratory analysis showed that the two efficacy endpoints rose and fell together, with a Pearson correlation of $r = 0.8184$ between mortality and repellency. The composite Effectiveness Index also varied widely across formulation groups: Orange + Lemon had the highest mean (72.14) and Watermelon the lowest (47.07). Combined extraction averaged 63.98 against 54.26 for Aqueous extraction, and Mixed formulations averaged 63.98 against 57.42 for Single-extract formulations. Among pests, Aphid had the highest mean effectiveness (67.05) and Spodoptera litura the lowest (54.13). Together, these patterns suggest that formulation composition and pest identity are useful predictive variables.

B. Model Performance

In the saved five-model test-set comparison, XGBoost scored best for Mortality % ($R^2 = 0.854$, MAE = 3.049, RMSE = 3.621). Gradient Boosting scored best for Repellency % on the held-out set ($R^2 = 0.773$, MAE = 3.631, RMSE = 4.795),

with XGBoost close behind. These rankings apply only to the evaluated split and model settings, and they should not be read as a guarantee of general performance.

C. Pareto Optimization and Recommended Candidate

Nineteen formulations survived Pareto filtering. Some reached relatively strong predicted efficacy but needed higher resource settings, while others gave up some efficacy in exchange for lower concentration or shorter exposure. Orange + Lemon showed up repeatedly among the stronger candidates, which fits its high descriptive effectiveness. To arrive at a single recommendation, the non-dominated records were normalized and ranked by Euclidean distance from an ideal point (maximum efficacy, minimum concentration, minimum exposure), with equal weight on all four criteria. The selected formulation is listed in Table IV.

TABLE IV RECOMMENDED PARETO-OPTIMAL FORMULATION BASED ON EQUAL-WEIGHTED DISTANCE TO THE IDEAL POINT

Attribute	Recommended value
Formulation ID	FPF_177
Fruit peel	Orange + Lemon
Extract type	Combined Extract
Formulation type	Mixed
Target pest	Aphid
Concentration	5.0%
Exposure time	6 h
Predicted Mortality	72.83%
Predicted Repellency	76.29%

FPF_177 stands out as a screening candidate because it pairs the shortest tested exposure time (6 h) with a concentration of 5.0%, and its predicted responses sit above the dataset averages of 59.01% mortality and 61.59% repellency. These predictions are not evidence of field performance or safety. The formulation is better seen as a priority for controlled validation covering efficacy, stability, phytotoxicity, effects on non-target organisms, and ease of application.

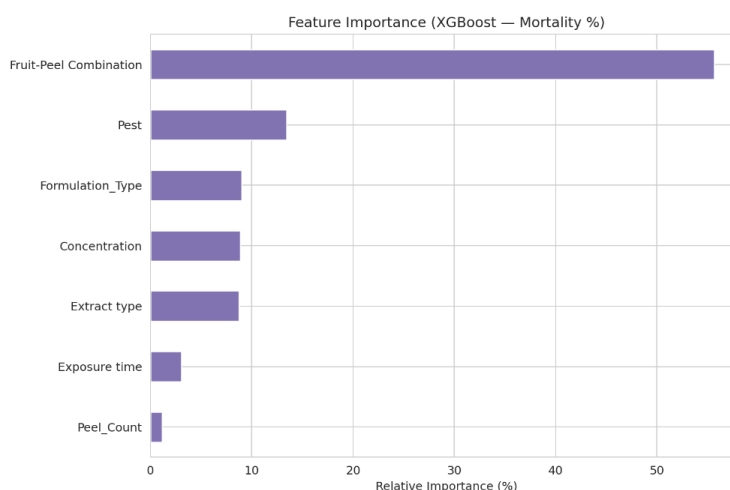


Fig. 2. Feature-importance ranking for the best Mortality % model.

The importance profile in Fig. 2 puts fruit-peel composition far ahead of the other predictors, with Fruit-Peel Combination at 55.67%. Next came Pest (13.45%), Formulation_Type (9.01%), Concentration (8.90%), Extract type (8.74%), Exposure time (3.07%), and Peel_Count (1.15%). Within this dataset, then, changing the botanical combination seems to affect the model's predictions more than changing exposure time alone. These percentages, however, describe how the model behaves and cannot establish biological causation.

V. CONCLUSION

This workflow connects machine-learning prediction with multi-objective candidate selection for developing fruit-peel bio-pesticides. In the expanded five-model test comparison, XGBoost came out ahead for mortality and Gradient Boosting for repellency under the settings evaluated. The earlier Pareto screening retained 19 non-dominated records and chose FPF_177 as a computational screening candidate. The framework is best seen as an experimental prioritization tool, and biological validation is required before any efficacy claims are made.

REFERENCES

- [1] D. Pimentel, "Amounts of Pesticides Reaching Target Pests: Environmental Impacts and Ethics", *J. Agric. Environ. Ethics*, vol. 8, pp. 17–29, 1995.
- [2] M. B. Isman, "Botanical Insecticides, Deterrents, and Repellents in Modern Agriculture and an Increasingly Regulated World", *Annu. Rev. Entomol.*, vol. 51, pp. 45–66, 2006.
- [3] S. Rafiq, R. Kaul, S. A. Sofi, N. Bashir, F. Nazir, and G. A. Nayik, "Citrus Peel as a Source of Functional Ingredient: A Review", *J. Saudi Soc. Agric. Sci.*, vol. 17, no. 4, pp. 351–358, 2018.
- [4] N. A. Sagar, S. Pareek, S. Sharma, E. M. Yahia, and M. G. Lobo, "Fruit and Vegetable Waste: Bioactive Compounds, Their Extraction, and Possible Utilization", *Compr. Rev. Food Sci. Food Saf.*, vol. 17, no. 3, pp. 512–531, 2018.
- [5] J. Banerjee, R. Singh, R. Vijayaraghavan, D. MacFarlane, A. F. Patti, and A. Arora, "Bioactives from Fruit Processing Wastes: Green Approaches to Valuable Chemicals", *Food Chem.*, vol. 225, pp. 10–22, 2017.
- [6] Y. K. Leong and J. S. Chang, "Valorization of Fruit Wastes for Circular Bioeconomy: Current Advances, Challenges, and Opportunities", *Bioresour. Technol.*, vol. 359, Art. no. 127459, 2022.
- [7] E. J. Rifna, N. N. Misra, and M. Dwivedi, "Recent Advances in Extraction Technologies for Recovery of Bioactive Compounds Derived from Fruit and Vegetable Waste Peels: A Review", *Crit. Rev. Food Sci. Nutr.*, vol. 63, no. 6, pp. 719–752, 2023.
- [8] L. Breiman, "Random Forests", *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [9] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [10] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review", *Sensors*, vol. 18, no. 8, Art. no. 2674, 2018.
- [11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
- [12] G. A. Mesías-Ruiz, M. Pérez-Ortiz, J. Dorado, A. I. de Castro, and J. M. Peña, "Boosting Precision Crop Protection Towards Agriculture 5.0 via Machine Learning and Emerging Technologies: A Contextual Review", *Front. Plant Sci.*, vol. 14, Art. no. 1143326, 2023.
- [13] E. Assadpour et al., "Application of Essential Oils as Natural Biopesticides; Recent Advances", *Crit. Rev. Food Sci. Nutr.*, vol. 64, no. 19, pp. 6477–6497, 2024.
- [14] I. Gupta et al., "Plant Essential Oils as Biopesticides: Applications, Mechanisms, Innovations, and Constraints", *Plants*, vol. 12, no. 16, Art. no. 2916, 2023.
- [15] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II", *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.