

Predicting Vitamin C Content in Vegetables Using Moisture and Ash Profiles: A Machine Learning Approach for Rapid Nutritional Screening

Md. Abdul Hamid Khan

24mschem114@student.cuet.ac.bd

Chittagong University of Engineering and Technology

Md. Mujahidul Islam

Chittagong University of Engineering and Technology

Rubaya Hasna

Islamic University Bangladesh

Shohanur Rahman

Islamic University Bangladesh

Research Article

Keywords: Vitamin C, vegetables, moisture, ash, machine learning, Bayesian optimization, HistGradientBoosting, SHAP, physicochemical screening, Bangladesh

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10809804/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting Vitamin C Content in Vegetables Using Moisture and Ash Profiles: A Machine Learning Approach for Rapid Nutritional Screening.

Md. Abdul Hamid Khan^{a,b*}, Md. Mujahidul Islam^a, Rubaya Hasna^b, Shohanur Rahman^b

^a *Department of Chemistry, Chittagong University of Engineering and Technology,
Chattogram-4349, Bangladesh*

^b *Department of Applied Chemistry and Chemical Engineering, Islamic University
Bangladesh, Kushtia-7003, Bangladesh.*

Corresponding Author: 24mschem114@student.cuet.ac.bd

Abstract

Direct iodometric titration is the reference method for measuring Vitamin C, but it becomes cumbersome for large-scale sample testing. This study explores a practical modeling approach to estimate Vitamin C concentration (mg/100 g) within a specific vegetable domain using routine, non-destructive physicochemical measurements specifically moisture and ash content without needing continuous titrations after model development. Data was collected from 350 samples representing the top seven vegetables available in the Philippines and Bangladesh (50 observations per group). The dataset was split into a 280-row training partition and a 70-row held-out test set. Tree-based regressors were optimized via Bayesian optimization and evaluated using five-fold stratified cross-validation. Based on a cross-validation-first ranking rule, HistGradientBoosting (HGB) achieved the top performance with a mean validation R^2 of 0.9205, a test R^2 of 0.9524, and test RMSE and MAE values of 6.00 and 3.49 mg/100 g, respectively. While Random Forest achieved the highest single test R^2 (0.9612), HGB was selected for its superior stability across repeated validations. SHAP analysis revealed that ash content had the highest global predictive contribution (72.67%), followed by moisture content (18.91%), ash standard deviation (7.53%), and moisture standard deviation (0.89%). Feature ablation confirmed this hierarchy: R^2 dropped to 0.3276 using moisture alone but reached 0.9348 when combining ash and moisture. The final framework successfully retained high-Vitamin-C predictions for broccoli and green bell pepper. This explainable screening tool accurately predicts nutritional values from physicochemical properties within the studied domain, pending external validation on unseen vegetable types.

Keywords: Vitamin C; vegetables; moisture; ash; machine learning; Bayesian optimization; HistGradientBoosting; SHAP; physicochemical screening; Bangladesh.

1. Introduction

Vegetables are often described in food-composition tables using a single value for moisture, minerals, vitamins, or other nutrients, although the composition of fresh plant tissue is rarely constant. Differences in species, cultivar, maturity, growing environment, soil conditions, post-harvest handling, storage temperature, exposure to oxygen and light, and the time between harvest and analysis can all alter the chemical profile of an edible vegetable [1]. This variability is separately relevant for water soluble and chemically sensitive nutrients such as Vitamin C. In Bangladesh, fresh vegetables obtained from local markets are consumed routinely and represent an important component of everyday diets. Knowing the Vitamin C concentration of these foods is therefore useful from both nutritional and food-quality perspectives. Conventional laboratory analysis can provide reliable sample-specific information, but analysing every nutrient separately becomes increasingly demanding when large numbers of samples have to be screened. This creates a practical analytical question that goes beyond ordinary composition reporting: can measurements that are already obtained during routine physicochemical analysis be used to estimate Vitamin C without requiring a separate Vitamin C determination for every subsequent sample? Vitamin C, or L-ascorbic acid, was selected as the nutritional response in the present study because it combines clear nutritional relevance with substantial variation in fresh plant foods [2]. Ascorbic acid participates in collagen synthesis, contributes to antioxidant defence, supports normal immune function and wound healing, and improves the absorption of non-haem iron. At the same time, it is one of the more labile micronutrients in vegetables, and its measured concentration can vary with biological maturity, processing, storage, oxygen exposure, light and temperature [3]. This variability was clearly present in the laboratory observations generated in the current work. Among the seven vegetables examined, laboratory mean Vitamin C concentration ranged from only 10.35 ± 0.25 mg/100 g in green banana to 80.26 ± 1.30 mg/100 g in broccoli, while green bell pepper also contained a comparatively high concentration of 78.54 ± 1.25 mg/100 g, Bitter melon contained 52.65 ± 1.10 mg/100 g, turnip 42.28 ± 0.95 mg/100 g, tomato 16.83 ± 0.45 mg/100 g, and ash gourd 14.80 ± 0.40 mg/100 g. At the individual-observation level, the complete 350-observation laboratory dataset covered an even broader Vitamin C range of approximately 9.82–85.33 mg/100 g. The target variable therefore contained sufficient biological and

analytical variation to pose a meaningful regression problem rather than a narrow prediction task in which most samples had essentially the same concentration.

The physicochemical measurements used as predictors were intentionally kept simple. Moisture content The fraction of fresh tissue that is water is described as the water fraction, while the inorganic residue is described as ashy fraction. the carbonaceous residue that is left after burning all of the organic material. In the present laboratory dataset, overall moisture averaged approximately 91.21%, although the seven vegetables showed distinct profiles. Green banana was clearly separated from the other groups, with values centred near 78%, while several vegetables had mean moisture contents above 93%. Ash occurred on a very different numerical scale. Across all 350 observations, mean ash content was approximately 0.694%, with individual values ranging from about 0.299% to 1.307%. Broccoli had the highest laboratory mean ash content at approximately 1.12%, whereas tomato was considerably lower at approximately 0.35%. These measurements do not represent Vitamin C directly. Moisture is not an ascorbic-acid measurement, and total ash does not quantify Vitamin C or any individual mineral. The scientific premise of the present work is therefore not that moisture or ash chemically determines ascorbic acid concentration. Instead, we asked whether the combined physicochemical pattern represented by water content, inorganic residue and their replicate variability contains sufficient information for a nonlinear model to distinguish samples with different Vitamin C concentrations [4]. This distinction becomes particularly important when the numerical relationships among the variables are examined. A simple linear association was present, but no individual physicochemical feature showed a sufficiently strong relationship with Vitamin C to provide an obvious single-variable calibration equation. In the 350 laboratory observations, Pearson correlation between ash content and Vitamin C was $r = 0.44$, while ash SD and Vitamin C showed $r = 0.40$. Moisture had a weaker positive relationship with the target ($r = 0.30$), and moisture SD was almost linearly unrelated to Vitamin C ($r = -0.07$). Moisture and ash themselves were only weakly negatively correlated ($r = -0.14$), whereas ash and ash SD were more strongly associated ($r = 0.75$). These values are informative because they show both signal and limitation. Ash clearly contains more direct predictive information than moisture when considered individually, yet an r value of 0.44 is far from a deterministic relationship [5]. The observations also form vegetable-specific clusters rather than following a single common straight line. A conventional linear model would therefore be required to impose one average slope across a data space in which different combinations of moisture and ash may correspond to markedly different Vitamin C concentrations. Tree-based machine-

learning models offer a different possibility: they can partition this space repeatedly, identify local thresholds and interactions, and learn nonlinear combinations of measurements without assuming beforehand that Vitamin C must increase or decrease at a constant rate with either moisture or ash [6] .

The use of predictive modelling for food-quality assessment has expanded rapidly for a similar reason. Recent studies have shown that Vitamin C and other nutritional attributes can be estimated from measurements that are easier or faster to collect than the reference chemical determination itself. Zushi et al. used machine-learning and regression approaches to predict ascorbic acid and antioxidant characteristics in strawberry using environmental, growth and fruit-quality information. Huang et al. combined visible and near-infrared spectroscopy with a one-dimensional convolutional neural network to estimate Vitamin C in citrus, while He et al. demonstrated that hyperspectral information and variable-selection methods could predict Vitamin C in sweet-potato roots [7] .Related work has also increasingly incorporated explainable artificial intelligence so that a high prediction score is not treated as sufficient evidence on its own; SHAP-based analyses, for example, have been used to identify the measurements that actually drive nutritional predictions. These studies illustrate an important direction in contemporary food analysis: reference chemistry remains essential for calibration and validation, but once a robust relationship has been established, information obtained from another measurement domain may be used to estimate the property of interest. The present study follows this general analytical principle but tests a substantially different and intentionally simpler input space. Rather than relying on hyperspectral cameras, visible–near-infrared instruments, dozens of environmental variables or a large panel of biochemical measurements, the prediction problem was restricted to moisture content, ash content, moisture SD and ash SD, all obtained from the routine laboratory physicochemical workflow [8] . This restricted predictor design is central to the practical motivation of the work. A highly accurate model based on many expensive measurements may be scientifically interesting but offer little reduction in analytical effort. In contrast, moisture and ash are already among the most familiar measurements in routine food-composition laboratories. In the present experiment, moisture was determined by oven drying and ash by muffle-furnace incineration, while Vitamin C was measured independently using iodometric titration. Importantly, the Vitamin C titration remained the reference analytical method throughout model development. Iodine volume, iodine concentration, titration endpoint information and calculated Vitamin C values were never supplied to the regression models as predictors [9] .The algorithms were allowed to see

only the four physicochemical descriptors. In this way, the machine-learning problem was deliberately separated from the reference calculation. The models were not being taught how to reproduce an iodine equation; they were being asked whether the patterns present in moisture and ash measurements could recover variation in a separately measured nutrient. The inclusion of moisture SD and ash SD also requires a careful interpretation [10]. These two variables are not additional chemical constituents of the vegetable. They summarize variability among laboratory replicate measurements. Their potential value may arise from differences in sample heterogeneity, repeatability or vegetable-specific analytical behaviour. In the full dataset, moisture SD averaged approximately 0.4304, whereas ash SD averaged 0.0329. This is because of their very different relationships with one another. The addition of Vitamin C was not warranted, however, and it was decided to test them separately. This was addressed By progressive feature ablations. This model was first trained with only moisture, and was later augmented with moisture plus First the ash, then moisture SD and lastly ash SD. The resulting performance differences provided a direct measure of how much predictive information entered at each stage. The laboratory dataset was also structured to allow all algorithms to be compared under the same conditions. A total of 350 laboratory observations were available, with 50 observations for each of the seven vegetable groups. The dataset was partitioned using a balanced 80:20 train–test design. From every vegetable, 40 observations were assigned to model development and 10 observations were retained for testing. Consequently, the training partition contained 280 observations and the untouched test partition contained 70 observations. 5 fold stratified cross-validation (CV) was performed only within the 280-observation training set. Each validation fold therefore contained 56 observations, including eight observations from each of the seven vegetables, while the corresponding fold-training portion retained 32 observations per vegetable. This balanced structure was chosen so that differences in algorithm performance would not simply reflect an uneven representation of one vegetable group. The held-out test set was not used during Bayesian hyperparameter optimization [11]. Rather than selecting a single algorithm in advance, ten tree-based regression approaches were evaluated: HistGradientBoosting, Random Forest, Gradient Boosting, Extra Trees, Decision Tree, Bagging, AdaBoost, XGBoost, LightGBM and CatBoost [12]. These algorithms differ substantially in the way they generate and combine decision trees, control variance, regularize model complexity and sequentially correct prediction errors. Their hyperparameters were optimized through Bayesian search using cross-validation performance from the training data [13]. This approach was preferred to an exhaustive grid because tree-based models may contain interacting parameters such as learning rate, tree depth, number of

estimators, leaf size, subsampling and regularization [14]. More importantly, final model selection was not based simply on whichever algorithm happened to produce the largest R^2 on the 70 test observations. Models were ranked first according to mean cross-validation validation R^2 , followed by validation stability and the training–validation gap, with held-out errors used as supporting evidence. This distinction proved important. Random Forest achieved the highest single test R^2 of 0.9612, but HistGradientBoosting produced the highest mean CV validation R^2 of 0.9205 and consequently ranked first under the predefined CV-first selection strategy. The selected HistGradientBoosting model still retained strong independent test performance, with $R^2 = 0.9524$, RMSE = 6.0028 mg/100 g and MAE = 3.4879 mg/100 g. The progression from moisture-only prediction to the complete model provided perhaps the clearest evidence that the predictive relationship was not simply a consequence of using many variables [15]. With moisture alone, the HistGradientBoosting model produced only $R^2 = 0.3276$ on the test data and an RMSE of approximately 22.57 mg/100 g. This result is consistent with the relatively weak moisture–Vitamin C correlation and shows that high water content by itself is not sufficient to distinguish high- and low-Vitamin-C samples. Now present analysis concerns prediction, not only biochemical causation. Using moisture + ash, test R^2 increased to 0.9348, while RMSE fell to approximately 7.03 mg/100 g. Adding moisture SD resulted in a smaller improvement to $R^2 = 0.9386$, and the final four-variable model reached $R^2 = 0.9524$ with RMSE 6.0028 mg/100 g. Thus, the largest improvement occurred not when the model was made increasingly complicated, but when ash was combined with moisture. The SD descriptors then refined an already strong physicochemical relationship. The explainability analysis supported this interpretation from a second, independent direction. For the selected HistGradientBoosting model, SHAP analysis attributed approximately 72.67% of the global absolute contribution to ash content, 18.91% to moisture content, 7.53% to ash SD and only 0.89% to moisture SD. The dominance of ash in SHAP analysis is consistent with the substantial increase in test R^2 observed when ash was introduced in the ablation experiment. The agreement between these analyses is more informative than either result alone. SHAP indicates how the fitted nonlinear model used each variable across individual predictions, whereas the ablation experiment shows what happened to predictive performance when those variables were withheld. Together, they indicate that ash was the principal source of predictive information in this dataset, moisture supplied an important secondary contribution, and the variability descriptors provided comparatively modest refinement. This should not be interpreted as evidence that increasing ash causes an increase in Vitamin C [16]. Ash is a bulk inorganic measurement and may act as

a marker of broader compositional differences among the vegetable tissues. Ash did not affect the situation.

The current analysis does not address the biochemical cause, but the prediction. Thus the practical interpretation of the model can be simply written. During model development, The values are used as a reference for the measurement of Vitamin C, which must be done experimentally. the algorithm learns. After the model is trained and well validated, a sample of the population should be able to be read. After the model is trained and properly validated, however, a sample of the population should be able to be read. The values of moisture and ash of the studied vegetable domain can represent its belonging to this domain. The values of moisture and ash of the studied vegetable domain can be a representation of its belonging to the studied domain. The characteristics were passed to the fitted nonlinear function to get an estimated Vitamin C concentration. This relation can be conceptually written as for the full model. $[\widehat{C}]_{\text{VitC}} = f(M, A, SD_M, SD_A)$, where $(\widehat{C})_{\text{VitC}}$ is the predicted The concentrations of Vitamin C, (M) is the moisture content, (A) is the ash content and (SD_M) and (SD_A) are the standard deviations of these content. each represent the variability in the corresponding replicates. This is NOT a typical chemical equation and should NOT be read as being one. data. It is the empirical function that is learned from the data, but it is nonlinear. laboratory observations. A held-out broccoli sample containing 91.02% moisture and 1.170% nitrogen had a nitrogen index of 1.07. The nitrogen index of a held-out sample of broccoli with 91.02% moisture and 1.170% N was 1.07. ash, moisture SD = 0.342 and ash SD = 0.055 was predicted by the optimized HistGradientBoosting The experimentally obtained reference value is 81.03mg/100g which is a value lower than the model value. value of 77.94 mg/100 g, giving an absolute error of approximately 3.09 mg/100 g or 3.97%. At the At the same time, there were some observations that were less accurately predicted, such as the one of ash-gourd, which had a value of 50. Meanwhile, in the same time there were fewer observations that were less accurately predicted such as the ash-gourd which has a value of 50. substantially larger error. These local breaks are significant due to the fact that they do not allow for a good overall R^2 . Against this background, the novelty of the present study does not rest simply on comparing ten machine-learning algorithms or reporting a high prediction coefficient. Its main contribution is the development and interrogation of an experimental physicochemical-to-nutritional prediction framework in which all 350 observations originated from laboratory analysis, Vitamin C was independently determined as the reference response, and only moisture- and ash-related descriptors were supplied to the models. Conventional

compositional studies generally stop after reporting the measured nutrient concentrations [17]. The present work extends this process by asking whether routine physicochemical measurements contain reusable predictive information, by optimizing several nonlinear regression approaches under a common validation design, and by using SHAP and progressive ablation to establish where that information originates. The model is therefore proposed as a calibrated screening tool within the investigated vegetable domain rather than as a chemical substitute for the iodometric method [18].

The study consequently addressed six connected objectives. First, the moisture, ash and experimentally determined Vitamin C profiles of seven selected vegetables from Bangladesh were characterized from the complete laboratory dataset. Second, linear and multivariable relationships among moisture content, ash content, moisture SD, ash SD and Vitamin C were examined. Third, ten tree-based regression algorithms were optimized by Bayesian search and compared under the same balanced train–test and five-fold cross-validation design. Fourth, the final model was selected according to repeated validation performance and stability rather than a single favourable test statistic. The fact that these forecasts were not "uniformly precise" is not evidence of that. Importance of features, SHAP The model was analysed and progressively ablated to identify measurements that led to the model. predictive performance [19]. Finally, the fitted model was tested as a useful framework for translating routine moisture and ash measurements in the lab into estimates for individual samples. The vitamin C content of the seven vegetable groups studied. By linking direct compositional analysis with the interpretable predictive modelling, the study will help to connect two types of modelling. nutrient measurements and data-supported screening, limitations of the underlying analysis. explicit.

2. Materials and methods

2.1. Sampling, sample preparation and laboratory procedures

Seven fresh vegetables, green banana, bitter gourd, green bell pepper, broccoli, turnip, ash gourd and Standardized processing was carried out on tomato—collected from the local markets in Bangladesh. laboratory procedure. After visible dirt and other surface impurities were removed, the samples were thoroughly washed with distilled water [20]. The edible portions were then separated, cut into smaller pieces, and homogenized to obtain a more uniform sample before subsampling. This common preparation procedure was followed for all seven vegetables so that the physicochemical measurements and the reference Vitamin C

concentrations used later for machine-learning model development were generated under the same laboratory workflow. Moisture and ash measurements provided the physicochemical descriptors, whereas Vitamin C was determined independently by iodometric titration and used as the experimental reference response [21]. The main stages of sample preparation, the prepared vegetable extracts, and the visible colour change associated with the endpoint of the iodometric Vitamin C assay are documented in **Fig. 1**, providing a direct link between the physical laboratory procedures and the numerical data subsequently used in the modelling analysis.

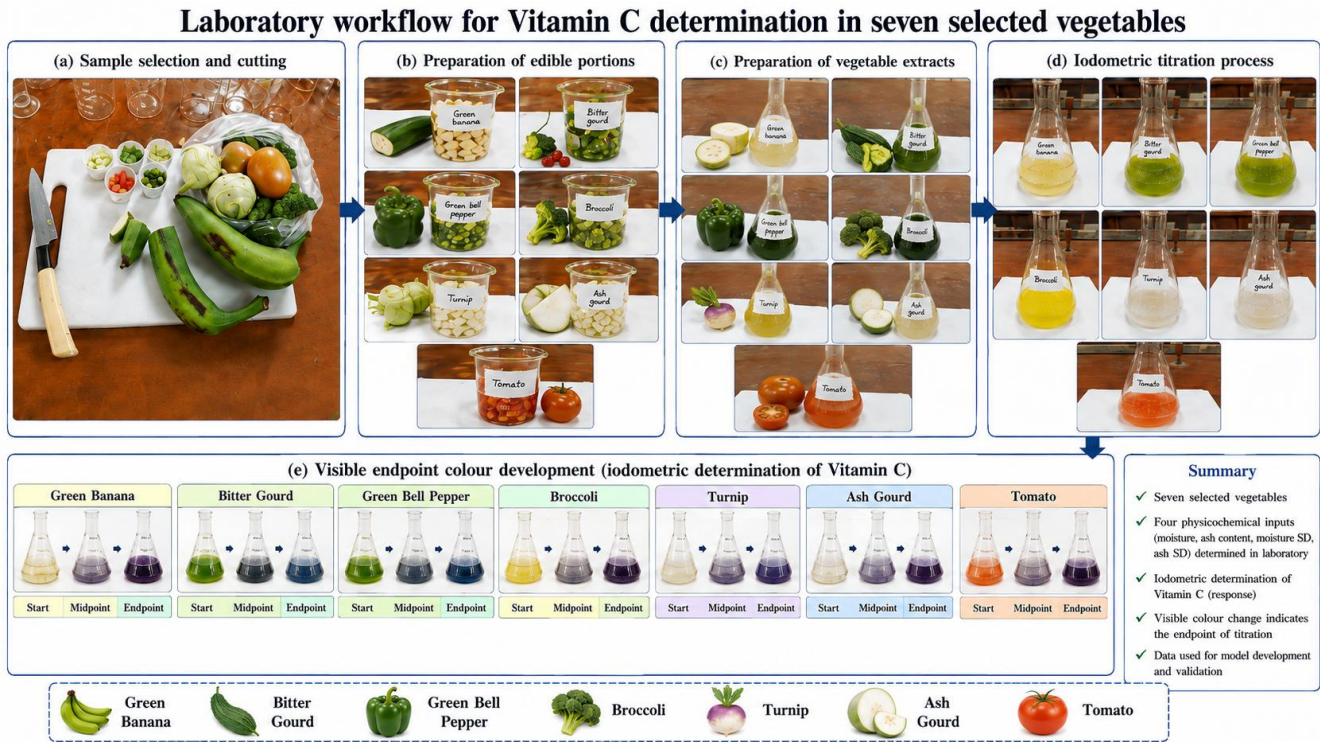


Fig. 1 Laboratory workflow used for the seven selected vegetables, showing the progression from sample selection and cutting to preparation of edible portions, production of vegetable extracts, and the visible colour development used to identify the endpoint during iodometric determination of Vitamin C.

The laboratory images in Fig. 1a provide the experimental starting point for the predictive analysis. The vegetables were not treated simply as numerical entries in a dataset; each observation originated from a sequence of physical preparation and laboratory measurement. The edible parts of the plants were prepared after selection and cutting. Homogenised prior to analytical measurements. Moisture and ash were determined as routine Vitamin C was determined independently by iodometric titration, while the others were determined by physicochemical procedures and were retained. As the reference value in model development.

Therefore, the images of the titration are useful for more than illustrating the procedure. They demonstrate the experimental step that the reference concentration for Vitamin C was taken. This was obtained as part of the data, along with the change from pre endpoint to persistent endpoint colour used in the iodometric assay. This separation of the reference measurement and the predictor measurements is key to this distinction. Research design of the study. During model development, the experimentally determined Vitamin C concentration provided the target against which prediction errors were calculated, but no information generated from the Vitamin C titration was allowed to enter the predictor matrix. Iodine volume, titrant concentration, endpoint information, and calculated Vitamin C concentration were therefore excluded from the model inputs. Instead, each sample was represented only by moisture content, ash content, moisture SD, and ash SD. This separation was intended to test whether information already contained in the routine physicochemical measurements was sufficient for a trained nonlinear model to recover variation in Vitamin C, rather than allowing the algorithm to reproduce information directly derived from the reference assay. The logic follows a broader development in food-quality prediction, where reference laboratory measurements are increasingly paired with another set of measurable characteristics to build calibrated prediction models [22]. For example, Vitamin C in citrus has been estimated using visible and near-infrared spectral information combined with regression and deep-learning models, while hyperspectral measurements have been used to predict Vitamin C in sweet-potato roots. Machine-learning models have also been developed to estimate ascorbic acid in strawberries from environmental, growth, and fruit-quality variables [23]. The present study adopts the same general calibration principle but deliberately uses a much simpler predictor space. Rather than requiring spectral instrumentation or a large collection of agronomic variables, the prediction framework is based on moisture- and ash-related measurements generated during conventional food analysis.

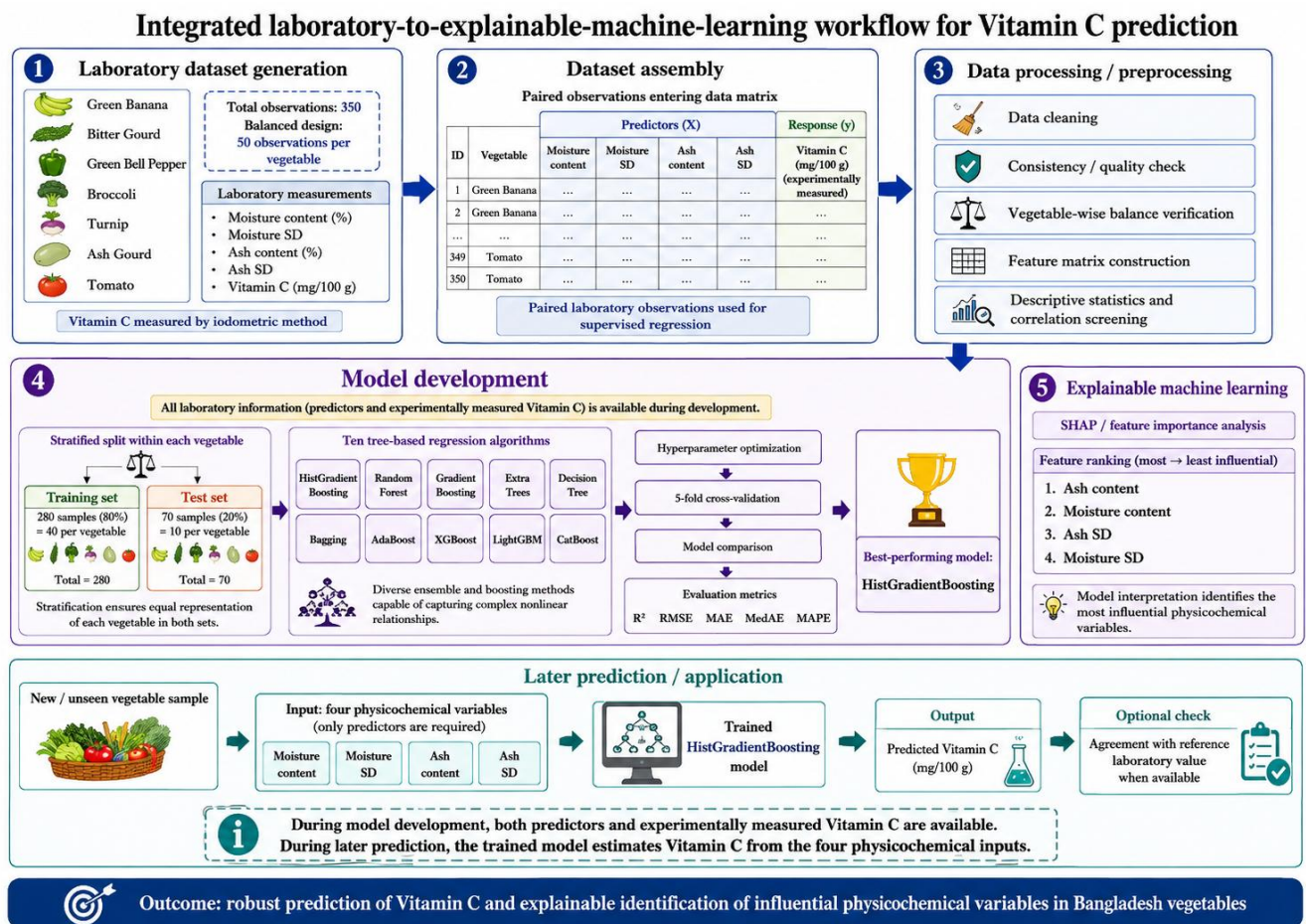


Fig. 2 Integrated laboratory-to-explainable-machine-learning workflow used for development and application of the Vitamin C prediction model.

Fig. 2 brings these experimental and computational stages together and, importantly, separates model development from later prediction. During model development, all relevant laboratory information is available. Moisture and ash characteristics form the predictor side of the dataset, while iodometrically determined Vitamin C provides the experimentally measured response. These paired observations are used to train the regression algorithms, optimize their hyperparameters, perform cross-validation, compare predictive performance, and select the final model [24]. In the present dataset, this process was based on four physicochemical inputs,

$$X = [M, A, SD M, SD A] \quad (1)$$

Where

M = Moisture content (%)

A = Ash content (%)

SD M = Standard deviation of moisture

SD A = Standard deviation of ash

where (M) denotes moisture content, (A) denotes ash content, and (SD M) and (SD A) represent the variability associated with their replicate measurements. The corresponding Vitamin C concentration, (C_{VitC}), was kept separate as the experimentally determined response. Once model fitting and validation have been completed, the pathway shown in **Fig. 2** changes in an important way. A new sample drawn from the same investigated vegetable domain does not need to provide an already measured Vitamin C concentration to the prediction model. Instead, its moisture- and ash-related characteristics are supplied to the previously trained function,

$$\hat{C}_{\text{VitC}} = f_{\text{ML}}(M, A, \text{SD } M, \text{SD } A) \quad (2)$$

Where:

$\hat{C}_{\text{VitC}}$ = concentration of Vitamin C predicted (mg/100 g) The value of f_{ML} is the trained nonlinear machine-learning prediction function. The model estimates the concentration of Vitamin C ($\hat{C}_{\text{VitC}}$). In other words, The iodometric determination is critical to defining the reference data set, to training the models, and The accuracy of the checking step is not an input used by the trained model, but the result of the Vitamin C titration step is an input used by the trained model is subsequently used for prediction. This is the actual transition shown in **Fig. 2** which is referred to as a practical transition. During the experimental stage the relationship is set and during the later screening stage that learned is applied. This separation is especially important if there is a relationship between the physicochemical measurements and a new sample. Important: Do not see the proposed approach as an alternative to the chemistry of. Vitamin C determination. The trained model does not fit ascorbic acid stoichiometrically from It does not mean ash is an equivalent or a chemical property of Vitamin C, but rather it learns an Experimental relationship between observations, experimentally matched pairs. Similar calibration-based reasoning lies behind recent studies of food quality using spectral or other indirectly measured data. The characteristics are used for the estimation of an attribute determined by chemical or physical methods [25] . In In the present case the question is formulated in a more limited manner: does a compact set of routine moisture- and Vitamin grams can be estimated by the measurements of ash-related. In the seven groups of vegetables studied, C was the highest in a vegetable. The workflow as presented in **Fig. 2**, thus, is a more. More than a series of laboratory and computational procedures. It captures the main idea in the practice, routine: reference chemistry is used to check and confirm the predictive relationship. That relationship is then applied via physicochemical measurements once the model has been trained. This separation also shields the analysis from leakage of the targets and increases the ease of the prediction. to interpret. [26] If moisture and ash and the variation

in moisture and ash can be considered acceptable accuracy, then the remaining data can be retained. The estimated Vitamin C value can be attributed to information learned from the supplied information to the model. The aforementioned physicochemical descriptors and not the titration result itself. The preparation step was standardized, since the quality of the resulting modelling analysis ultimately depends on the preparation step. The majority of the laboratory descriptors. Each vegetable was initially used as a food-analysis sample instead of as a food sample. computer-science record. Washing with distilled water decreased removable surface contamination, Cutting was used to make small chunks and homogenisation blended out any inter sample variation. before subsampling. To ensure consistency, the same seven vegetable identities were maintained throughout the workflow, with the laboratory labels and modelling groups and later the vegetable-wise interpretation all referring to the same identities. biological categories. The photographs in **Fig. 1** are thus not aesthetic in nature, but rather they are records of Physically moving from market vegetables to prepared food to extracts for use in. The Vitamin C results are compared with the reference Vitamin C assay. The analytical procedure has been designed to ensure that measurements are performed The samples of vegetables used as the description were not the same as the samples used to define the prediction target. Moisture and ash were obtained independently through gravimetric procedures, whereas Vitamin C was determined by iodometric titration and retained only as the experimental reference against which the models were trained and evaluated. [27] This distinction was important because the purpose of the modelling stage was to determine whether routine physicochemical measurements could carry enough information to estimate Vitamin C, not to reconstruct the titration calculation itself. Accordingly, no information originating from the Vitamin C assay—including iodine volume, titrant concentration, endpoint information, or any intermediate or final quantity derived from the titration—was supplied to the machine-learning algorithms. Each of the 350 laboratory observations was represented only by four predictor variables: moisture content, ash content, moisture SD, and ash SD, while the corresponding experimentally determined Vitamin C concentration was retained separately as the single continuous response. In matrix form, the modelling dataset therefore consisted of a 350×4 physicochemical predictor matrix paired with a 350-value Vitamin C response vector. Vegetable identity was also retained in the dataset, but only to preserve the balanced experimental structure of 50 observations for each of the seven vegetables during train–test partitioning and cross-validation; it was not used as a predictive feature. Keeping these elements separate ensured that the resulting model performance reflected information learned

from the moisture- and ash-related characteristics themselves rather than direct or indirect reuse of the reference Vitamin C measurement [28] .

2.2. Moisture determination

Moisture content was determined gravimetrically by oven drying, a conventional approach in food and agricultural analysis in which the change in sample mass before and after controlled drying is used to estimate the moisture fraction [29] .Oven-based methods commonly use temperatures around 105 °C and continue drying for a defined period or until a stable mass is obtained. For each determination in the present study, a known mass of homogenized fresh vegetable sample was transferred to a pre-weighed container and placed in an oven maintained at 105 °C [30] .Drying was continued until a constant mass was reached rather than terminating the procedure after an arbitrary single weighing. This step was important because residual moisture remaining in an incompletely dried sample would increase the final recorded mass and consequently lead to an underestimation of moisture content. After drying, the difference between the original fresh mass and the final constant dry mass was expressed relative to the original sample mass. The moisture content on a fresh-weight basis was calculated as:

$$Moisture(\%) = \left[\frac{W_{fresh} - W_{dry}}{W_{fresh}} \right] \times 100 \quad (3)$$

where W_{fresh} is the mass of the homogenized sample before oven drying and W_{dry} is the mass of the same subsample after drying to constant mass. Both measurements therefore refer to the same analytical portion. The number is the percentage of mass that is lost during the drying process. Conditions as compared to the initial fresh mass. This is a loss-on-drying method that is commonly used for moisture. This determination, however, requires controlled sample preparation for its analytical interpretation, The factors that affect the drying time, temperature, and weighing conditions. The same method and calculation was used To consistently replicate the moisture measurements to the next set of replicate measurements in order to model moisture values going into the next set of measurements Data from both of the datasets were collected on the same analytical basis. The measured moisture values were Significant differences between the vegetables. Experimental summary: green banana The moisture content of foods is contained lowest in ash gourd ($95.39 \pm$) and highest in carrot ($78.03 \pm$). 0.33%. This was extended to 75.31% to 98.74% by the complete set of 350 laboratory observations. 98.73%, with an overall mean moisture content of 91.2076% and an SD of 5.8221%. Most vegetable green banana was the only crop in the high-moisture zone above about 90%, while groups were mostly in the medium moisture zone. Clearly separated at the bottom, This distribution is more

applicable to the later prediction analysis. There was a broad physicochemical axis across the entire dataset provided by moisture, but the substantial variation in moisture was mainly responsible for the divergent tendencies in the data. This overlap of several of the high-moisture vegetables also indicated that moisture alone was not likely to be the cause of food getting contaminated. be able to tell the difference between samples that have very different Vitamin C concentrations. That expectation was The moisture-only HistGradientBoosting model was then used to validate the modelling results, as shown below. Predicted by produced the held out R^2 of 0.3276 and RMSE of 22.57 mg/100g, which suggests that water content by predicted produced by water content by model. It only caught a small proportion of the variability in Vitamin C. Information from the replicate measurements was not discarded after the mean moisture content had been calculated. Instead, moisture SD was retained as a separate analytical descriptor so that its possible contribution to prediction could be evaluated explicitly. Across the 350 observations, moisture SD ranged from 0.2000 to 0.8640, with a mean of 0.4304. This variable was not interpreted as an additional nutrient or physicochemical constituent; rather, it represented the observed variability associated with replicate moisture determinations. Retaining it allowed the modelling analysis to test whether variation among replicate measurements contained information beyond the moisture percentage itself. Importantly, this contribution was not assumed in advance. Progressive feature ablation was later used to determine whether adding moisture SD improved prediction after moisture and ash had already been included [31] .The resulting comparison showed only a modest improvement—from a test R^2 of 0.9348 for moisture plus ash to 0.9386 after moisture SD was added—supporting its interpretation as a secondary analytical descriptor rather than a dominant predictor of Vitamin C.

2.3. Ash determination

Ash content was determined gravimetrically by dry incineration [32] .A known mass of homogenized vegetable sample was transferred to a pre-weighed crucible and heated in a muffle furnace at 600 °C for 6 h. During this treatment, the combustible organic fraction of the sample was removed, leaving the non-combustible inorganic residue. After incineration, the crucible was allowed to cool and the remaining residue was weighed. The ash content was then determined as a percentage of the sample mass using **Equation (4)**.

Equation (4). Calculation of ash content

$$Ash(\%) = \left(\frac{W_{ash}}{W_{sample}} \right) \times 100 \quad (4)$$

456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489

where W_{ash} is the mass after combustion and W_{sample} is the mass of the sample. The of the sample used in the determination. The result of this calculation was considered as total Overall inorganic residue of the vegetable material—ash. Did not read as the the concentration of any particular element (no separate quantification of individual minerals) the present experiment. In the beginning, determining ash is a very different process than determining moisture. though both procedures are based on changes in sample mass. Moisture analysis measures the mass lost In controlled drying this material will be kept while ash analysis will be the material remaining after the high temperature drying. combustion. These two measurements represent different aspects of the vegetable in reality. With matrix moisture being water, matrix ash being the combined inorganic residue. That which is left over after the organic fraction is removed. This separation is especially pertinent to the Both the variables are not expected to directly influence the chemical and hence the present modeling framework. Rather, moisture and ash will give complementary physicochemical information as measures of Vitamin C. about the sample. The ash content from the seven samples was very different when measured in the laboratory. vegetables. In the Experimental summary the minimum mean Ash content was obtained in tomato, which is $0.35 \pm 0.02\%$, whereas broccoli showed the highest value at $1.12 \pm 0.05\%$. Across all 350 laboratory The ash content that was observed ranged from 0.299% to 1.307% with a mean of 0.6942% and SD of of 0.2389%. This is a fairly small moisture range, compared to moisture, content, the ash percentages were allocated systematically among the vegetable groups instead of being randomly scattered. This is shown better in the group-level pattern. Across the 50 laboratory In this study, the mean ash content of the tomato was about 0.351%, 0.469% for green bell pepper, 0.641% for ash gourd, 0.671% for green banana, 0.782% for turnip, 0.824% for bitter gourd, and 1.121% for broccoli. Therefore, only a few tenths of a difference in the reading of only a few ounces. A few of the vegetables had different physicochemical profiles that could be separated by a few percentage points. From a conventional food composition analysis point of view, these differences might appear small because ash is measured on a percentage scale. However, from a predictive point of view, Their value is less dependent on their absolute size and more dependent on their consistency in their occurrence. In combination with the other measured characteristics. Therefore, the data for ash content were not transformed into because the ash content was a continuous predictor. broad categories. Tree-based regression algorithms can make use of relatively small numerical differences; when such differences are repeated thresholds or when such differences interact with other variables in the predictor space [33] .Therefore, ash was

analysed along with moisture in the present analysis. content, moisture SD, and ash SD. The intent was not to imply that with greater ash content the chemical The Vitamin C content is higher in this. Rather the analysis focused on if ash is a part of a Physicochemical pattern that can be used to differentiate samples from different experimentally measured samples in a broader group. Vitamin C concentrations. This separation was important when the linear relationships between the variables were examined. Ash content had the highest positive correlation in Pearson correlation with Vitamin C among the The four physicochemical predictors provided $r = 0.44$ as compared to $r = 0.30$ for moisture content. An This association is of a magnitude that suggests that there were some useful pieces of information in the ash, but not enough. A simple explanation of the concentration of Vitamin C was attempted using only one variable (strong), which was not successful. The relationship was Therefore, further investigated with a combination of ash and the other physicochemical descriptors. The predictive value of ash was much better in the progressive feature-ablation analysis. When HistGradientBoosting was trained using moisture content alone, the held-out model achieved only $R^2 = 0.3276$, with an RMSE of approximately 22.57 mg/100 g. After ash content was added to the predictor set, test performance increased markedly to $R^2 = 0.9348$, while RMSE decreased to approximately 7.03 mg/100 g. This substantial improvement indicates that ash provided information that was not already captured by moisture alone. The magnitude of this change is especially informative because only one additional physicochemical variable was introduced.

A similar pattern emerged from the explainability analysis. For the final HistGradientBoosting model, SHAP analysis assigned approximately 72.67% of the global absolute contribution to ash content, compared with 18.91% for moisture content, 7.53% for ash SD, and 0.89% for moisture SD. [34] The agreement between the feature-ablation experiment and the SHAP analysis provides two complementary views of the same result. Ablation shows how much predictive performance changes when a variable is introduced, whereas SHAP describes how strongly the fitted nonlinear model uses that variable when producing individual predictions [35]. Both analyses therefore identified ash as the dominant predictive descriptor within the present laboratory dataset.

This result should nevertheless be interpreted as a predictive association rather than evidence of a direct biochemical mechanism [36]. Total ash is a bulk measurement of inorganic residue and does not reveal which specific minerals are present or in what concentrations [37]. The present data therefore cannot establish that a particular mineral influences Vitamin C concentration, nor can they show that experimentally increasing ash would cause Vitamin C to

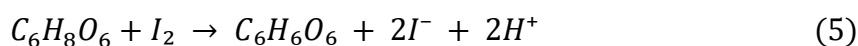
increase. A more cautious interpretation is that ash content captures part of the broader compositional variation among the vegetable samples, and that this variation also helps the model distinguish different Vitamin C profiles. Replicate variability associated with the ash measurement was retained separately as ash SD. Across the complete 350-observation dataset, ash SD ranged from 0.0120 to 0.0820, with an overall mean of 0.0329. As with moisture SD, this variable was not interpreted as a separate nutrient or chemical constituent. Instead, it summarized the observed variation among replicate ash determinations and was included so that the modelling analysis could determine whether analytical variability contained useful information beyond the mean ash percentage itself.

The later modelling results showed that ash SD contributed additional information, but its role was considerably smaller than that of mean ash content. Its global SHAP contribution was approximately 7.53%, compared with 72.67% for ash content. This distinction is useful because it separates the dominant compositional signal represented by the mean ash measurement from the more limited refinement associated with replicate variability. Rather than assuming that every additional predictor would improve prediction, the contribution of ash SD was therefore evaluated empirically through feature ablation and explainable machine-learning analysis [38] [39]. Taken together, the ash determination served two related purposes in the present study. First, it gave a traditional Laboratory determination of the amount of all inorganic matter in each vegetable sample. Second, it produced an idea that. reproducible physicochemical descriptor which can be tested as an independent predictor of The vitamin C content was determined experimentally. Its organized distribution in the seven vegetable groups, its moderate linear association with Vitamin C, the drastic improvement that follows its addition to the That all the SHAP contributions were strong and that ash carried a significant amount of moisture is apparent in the moisture-only model. Information that can be predicted in the area being studied. This is a predictive role, and it does not affect the chemical. It shows that a simple food-composition parameter can be used to demonstrate the meaning of the measurement; it does not demonstrate this. When taken with other data, contribute to a nonlinear data-driven estimate of Vitamin C Physicochemical properties that were measured experimentally.

2.4. The vitamin C content was determined using iodometric titration method.

Vitamin C was measured by iodometric titration and chosen as the response measured in experiments to develop and evaluate the machine-learning models later. The prepared

vegetable extract was placed in a titration flask, starch indicator was added to it and it was titrated with a standard iodine solution. Titration was continued until a persistent blue-black colour appeared, which was accepted to be the end-point of the titration. The volume of iodine used to reach this endpoint was measured and the iodine consumed used to calculate the amount of ascorbic acid in the sample analysed. The concentration of vitamin C was expressed in mg/100g of fresh vegetable sample, so that all the 7 vegetable groups are on the same ground of comparison. Iodometric method is based on a known redox reaction in which ascorbic acid is the reducing agent. The ascorbic acid is oxidized and the molecular iodine is reduced during the reaction. The reaction proceeds on a 1:1 molar basis, meaning that one mole of iodine corresponds stoichiometrically to one mole of ascorbic acid under the conditions of the assay [40]. The chemical reaction used as the basis of the determination is shown in **Equation (5)**. **Equation (5)**. In this reaction, ascorbic acid is used in exactly the same amounts as iodine is needed.



In this reaction ascorbic acid and iodine react in a stoichiometric ratio. The end point is an important component of the analytical process that is visible to the user. Until ascorbic acid is depleted. In the extract, the reaction of iodine with the sample is a redox reaction. When there is some ascorbic acid remaining, When reacted, there is also a slight surplus of iodine in solution which reacts with starch and forms the. characteristic persistent blue-black colour. [41] The colour transition therefore provides the practical endpoint used to determine how much iodine was consumed by the sample. In the present study, this visual change was documented in **Fig. 1a** as part of the experimental workflow.

The amount of iodine consumed during titration was calculated from the concentration of the standardized iodine solution and the volume required to reach the endpoint. This calculation is shown in **Equation (6)**.

Equation (6). Calculation of the amount of iodine consumed during titration

$$n(I_2) = C(I_2) \times V(I_2) \quad (6)$$

where $n(I_2)$ represents the total number of moles of iodine consumed, $C(I_2)$ is the total molar concentration of the iodine solution, and $V(I_2)$ is the corresponding titrant volume. Because the stoichiometric relationship between iodine and ascorbic acid is 1:1, the amount of iodine calculated from **Equation (6)** provides the basis for determining the amount of ascorbic acid represented by the titrated aliquot [42].

The calculated amount of ascorbic acid was then related back to the fresh vegetable sample represented by the analysed aliquot. This final conversion allowed Vitamin C concentrations

from samples of different vegetables to be reported on the same fresh-weight basis. The concentration was represented as mg of Vitamin C per 100 g of fresh sample according to **Equation (7)**.

Equation (7). Calculation of Vitamin C concentration on a fresh-weight basis

$$\text{Vitamin C } \left(\frac{\text{mg}}{100 \text{ g}} \right) = \frac{[m (\text{ascorbic acid in aliquot})]}{[m (\text{sample represented by aliquot})]} \times 100 \quad (7)$$

The iodometric procedure therefore provided a chemically determined Vitamin C value for every laboratory observation. These measurements were important because they established the reference target against which the machine-learning models could be trained and judged. The purpose of the modelling stage was not to reproduce the titration calculation itself, but to determine whether independently measured physicochemical characteristics could recover the variation observed in these experimentally determined Vitamin C values. This distinction continued to be observed across dataset construction. Although the information of endpoint, intermediate data etc did not affect the results, the iodine concentration, volume of iodine and data before the end point. The calculated quantities in the titration process were essential to get reference Vitamin C. All these variables were excluded from the machine-learning predictor matrix. The Only moisture content, ash content, moisture SD, and ash SD were given to the algorithms. The final Vitamin C The concentration measured by the iodometric method was stored independently as a continuous response variable. Thus predictor measurements and the determination of the target remained analytically independent. Maintaining this separation was particularly important for the interpretation of model performance. If iodine volume or another quantity directly calculated from the Vitamin C analysis If information regarding the attendance at this event had been among the input to the model, it could have used this information. Embedded in the reference calculation. Under those circumstances, high predictive accuracy would give There was little evidence that moisture and ash related measurements were informative. By excluding The model was to learn from the variables in the predictor set that were all obtained from titration. Unique physical and chemical properties of the samples [43] .

The Vitamin C measurements in experiments 2 were conducted over a broad range of concentrations with the 7 concentrations. vegetables. The lowest mean concentration was found in green banana (10.35 ± 0.25 mg/100 g) followed by by ash gourd at 14.80 ± 0.40 mg/100 g and tomato at 16.83 ± 0.45 mg/100 g. Avoid using turnip or bitter gourd. occupied the intermediate range, with mean concentrations of 42.28 ± 0.95 mg/100 g and 52.65 ± 1.10 mg/100 g, respectively. Significant higher levels were detected for green bell pepper at 78.54

± 1.25 mg/100 g and broccoli at 80.26 ± 1.30 mg/100 g. The lowest and highest value of the difference are The highest experimental value was thus near 70 mg/100 g, which was a good target value for the gradient. for regression modelling. The entire sample of 350 laboratory observations revealed a wider range of The range was from 9.82 to 85.33 mg/100 g in the samples. At the 50-observation group level, mean Vitamin C concentrations were approximately 10.37 mg/100 g for green banana, 14.81 mg/100 g for ash gourd, 16.87 mg/100 g for tomato, 42.37 mg/100 g for turnip, 52.74 mg/100 g for bitter gourd, 78.55 mg/100 g for green bell pepper, and 80.28 mg/100 g for broccoli. This broad and structured distribution was useful for model development because the algorithms were required to distinguish samples across a wide nutritional range rather than predict values concentrated within a narrow interval. The reference-assay strategy used here follows the general logic adopted in predictive food-quality research. A chemically or instrumentally measured property is first established as the reference response, while a separate set of measurements is evaluated for its ability to predict that response. The reference analysis remains essential during calibration and validation, but it is kept independent of the predictor measurements [44] .In the present work, iodometric titration played this reference role, whereas moisture and ash characteristics provided the independent information from which the models attempted to estimate Vitamin C. The practical significance of this distinction becomes clearer when the modelling stage is considered. **Equations (5)-(7)** describe how Vitamin C was determined experimentally from the iodometric assay. They were used to establish the laboratory reference values and were not recalculated by the machine-learning models during prediction. Once the experimental dataset had been established, the modelling analysis asked a different question: whether the variation in these chemically determined Vitamin C values could be learned from moisture content, ash content, moisture SD, and ash SD alone. The later prediction results indicated that meaningful information was indeed present in these independent physicochemical descriptors. Moisture alone provided limited predictive power, with a held-out R^2 of 0.3276 and an RMSE of approximately 22.57 mg/100 g. When ash was added, test performance improved markedly to $R^2 = 0.9348$, with RMSE decreasing to approximately 7.03 mg/100 g. Using all four physicochemical descriptors, the selected HistGradientBoosting model reached a test R^2 of 0.9524, an RMSE of 6.0028 mg/100 g, and an MAE of 3.4879 mg/100 g. Importantly, these results were obtained without including iodine consumption or any other variable originating from the Vitamin C titration among the predictors. The iodometric assay therefore had a clearly defined role in the study. It provided the chemically anchored reference Vitamin C concentrations required to train, validate, and

evaluate the prediction models. The subsequent machine-learning analysis did not replace this reference method or alter the underlying chemistry. Instead, it investigated whether routine moisture- and ash-related measurements contained sufficient independent information to estimate the experimentally established Vitamin C response within the seven vegetable groups studied.

2.5. Experimental reference values

Before presenting the machine-learning results, the experimentally measured composition of the seven vegetables was summarized to establish the laboratory reference profile of the dataset. **Table 1** presents moisture, ash, and Vitamin C as mean $\pm$ SD from the replicate determinations. Reporting these measurements first provides a direct view of the compositional range represented in the experiment and establishes the scale against which the later model predictions can be interpreted. Similar descriptive presentation of proximate composition and ascorbic acid has been widely used in vegetable-composition studies, particularly when the objective is to characterize differences among edible plant materials before further statistical or predictive analysis [45]. No inferential comparison, such as analysis of variance, was performed on these laboratory summary values. The differences presented in **Table 1** are therefore interpreted descriptively. Terms such as higher, lower, or intermediate are used to describe the observed measurements, whereas claims of statistical significance are deliberately avoided.

Table 1. Experimental moisture, ash, and Vitamin C values for the seven selected vegetables (mean $\pm$ SD).

Vegetable	Moisture (%)	Ash (%)	Vitamin C (mg/100 g)
Green Banana	78.03 $\pm$ 0.52	0.67 $\pm$ 0.03	10.35 $\pm$ 0.25
Bitter Gourd	93.08 $\pm$ 0.41	0.82 $\pm$ 0.04	52.65 $\pm$ 1.10
Green Bell Pepper	94.50 $\pm$ 0.36	0.47 $\pm$ 0.02	78.54 $\pm$ 1.25
Broccoli	89.74 $\pm$ 0.48	1.12 $\pm$ 0.05	80.26 $\pm$ 1.30
Turnip	94.48 $\pm$ 0.39	0.78 $\pm$ 0.03	42.28 $\pm$ 0.95
Ash Gourd	95.39 $\pm$ 0.33	0.64 $\pm$ 0.02	14.80 $\pm$ 0.40
Tomato	93.12 $\pm$ 0.44	0.35 $\pm$ 0.02	16.83 $\pm$ 0.45

Keeping **Table 1** separate from the later machine-learning summaries was intentional. The values in this table originate directly from the laboratory measurements and describe the experimental reference profile, whereas the subsequent modelling tables report sample-level prediction and model-performance information. Maintaining this separation allows ease of

tracking the food-composition analysis transition to the predictive phase of study. Prior to fitting any model, a broad Vitamin C gradient was already noted. The mean Vitamin C content of broccoli (80.26 ± 1.30 mg/100 g) was the highest, followed by green bell pepper (78.54 ± 1.25 mg/100 g). The concentration of bitter melon was comparatively high at 52.65 ± 1.10 mg/100 g, and turnip was in the middle at 42.28 ± 0.95 mg/100 g. Tomato, ash gourd and green banana had significantly lower mean concentrations with green banana having the lowest mean concentration (10.35 ± 0.25 mg/100 g). There was indeed a significant difference between the upper and lower ends of this range: the mean Vitamin C concentration, found in broccoli, was about 7.8 times greater than the mean concentration found in green banana. The range of responses was wide and supported the experimental evaluation of the ability of the physicochemical measurements to reproduce the observed variation by regression modelling. The laboratory profile also serves as an example of why moisture content was not sufficient to describe the variation of Vitamin C. The mean moisture contents of green bell pepper and turnip were found to be 94.50% and 94.48%, respectively, while the Vitamin C concentration was found to be different, 78.54 and 42.28 mg/100 g respectively. It was still more clearly seen in ash gourd. It had the highest mean moisture content of all the seven studied vegetables (95.39%) but the lowest concentration of Vitamin C (14.80 mg/100 g). Broccoli on the other hand, had at 89.74% moisture content and the highest concentration of Vitamin C, at 80.26 mg/100 g. These comparisons suggest that samples with moisture content and position in the water spectrum can easily be very different in their Vitamin C profile. The relationship between the measured physicochemical properties and Vitamin C was therefore unlikely to be represented adequately by a simple one-variable calibration. This observation provided the practical rationale for evaluating moisture together with ash and the replicate-variability descriptors within multivariable nonlinear regression models. Importantly, this interpretation is based on the observed composition pattern rather than on an assumed causal relationship between any individual physicochemical measurement and Vitamin C.

2.6. Modelling dataset and data integrity

The machine-learning analysis was based on the complete 350-observation modelling dataset, with an equal contribution of 50 observations from each of the seven vegetable groups. The recorded fields comprised Sample ID, Vegetable Code, Moisture Percent, Moisture SD, Ash Percent, Ash SD, and Vitamin C mg/100 g. These variables were not all treated in the same way during modelling. Sample ID was retained solely for traceability, allowing individual observations to be linked back to the underlying dataset, while Vegetable Code was used only

to preserve the balanced group structure during data partitioning and cross-validation. In both cases none of the data was used in the regression algorithms. This was done by purposefully including only four experimentally-determined physicochemical descriptors: moisture content, ash content, moisture SD and ash SD in the predictive space. Experimentally determined Vitamin C concentration was retained separately as the single continuous response. This separation was central to the design of the study. The algorithms were required to learn from the measured physicochemical characteristics of the samples rather than from vegetable identity or from any quantity originating from the Vitamin C titration. Excluding such variables also reduced the risk that apparently strong model performance would arise from information already closely tied to the target rather than from the intended predictor set. More broadly, machine-learning studies have shown that leakage between predictor construction, training, and evaluation can substantially inflate apparent predictive performance [46].

Before model fitting, the dataset was examined as an analytical object rather than passed directly to the algorithms without inspection. All 350 observations were complete for the recorded predictor and response fields, and no exact duplicate predictor–target rows were identified after Sample ID was excluded. The equal group size was particularly useful for partitioning because every vegetable could contribute the same number of observations to the development and held-out subsets. Consequently, the later 80:20 partition could be implemented as 40 development observations and 10 test observations for each vegetable, rather than allowing larger groups to dominate the evaluation. The complete observation-level dataset was retained in Supplementary **Tables S1A–S1G**, providing a transparent record of the values supplied to the modelling workflow. This is particularly important in a relatively compact experimental dataset because summary statistics alone cannot reveal the exact input–output combinations encountered by the algorithms. Retaining the full modelling table therefore supports reproducibility and allows the reported model behaviour to be interpreted against the underlying laboratory observations. An important boundary was also placed on the interpretation of the resulting performance estimates. All seven vegetable groups occur in both the development and test partitions. The held-out observations are therefore new samples within vegetable groups already represented during model development. The resulting test scores should consequently be interpreted as performance within the compositional space represented by these seven vegetables, rather than as evidence of successful prediction for an entirely new vegetable species [46]. This distinction between in-distribution and out-of-distribution evaluation is increasingly emphasized in machine-learning research because

models that perform well when training and test data share the same underlying groups may perform differently when confronted with genuinely unseen groups or domains. The overall distribution of the five modelling variables is summarized in **Table 2**. The descriptive statistics show that the predictors occupied substantially different numerical ranges, while Vitamin C provided a broad response distribution for regression analysis. **Table 2.** Overall descriptive statistics of the 350-row modelling dataset.

Variable	n	Mean	SD	Min	Q1	Median	Q3	Max	Skewness	Kurtosis
Moisture content (%)	350	91.2076	5.8221	75.3100	90.4600	93.0650	94.8225	98.7300	-1.4798	1.1457
Moisture SD	350	0.4304	0.1182	0.2000	0.3462	0.4135	0.5015	0.8640	0.7078	0.4586
Ash content (%)	350	0.6942	0.2389	0.2990	0.4840	0.6950	0.8215	1.3070	0.3985	-0.3749
Ash SD	350	0.0329	0.0150	0.0120	0.0210	0.0290	0.0410	0.0820	1.1089	1.0435
Vitamin C (mg/100 g)	350	42.2831	27.5590	9.8200	15.1500	42.1750	76.7825	85.3300	0.2420	-1.5264

The descriptive profile in **Table 2** confirms that the modelling problem combined variables with very different numerical characteristics. Water quality was focused at the upper end of its range. However the scale had a median of 93.0650% while ash content had a median of 0.6950%. The corresponding replicate-variability descriptors acted upon, on even smaller scales, with median values of 0.4135 for moisture SD and 0.0290 for ash SD. In comparison, Vitamin C ranged from 9.82 to 85.33. mg/100 g and had an overall mean of 42.2831 ± 27.5590 mg/100 g. The large response SD shows the: seemed to give good separation between the vegetable groups and gave the regression algorithms a useful Explain how concentration gradient is related to learning. The distributions of vegetables are based on the above and are included for further information that is not apparent from the pooled statistics. alone. **Table 3** indicates that there were 50 observations per group, but the locations and ranges of The moisture, ash and Vitamin C content were significantly different between vegetables. **Table 3.** Vegetable-wise distributional summary of the modelling observations (n=50 per vegetable).

Vegetable	Moisture, mean $\pm$ SD	Moisture range	Ash, mean $\pm$ SD	Ash range	Vitamin C, mean $\pm$ SD	Vitamin C range
Green Banana	78.20 ± 1.66	75.31–81.22	0.671 ± 0.054	0.576–0.779	10.37 ± 0.32	9.82–10.99
Bitter Gourd	93.02 ± 1.55	90.36–97.19	0.824 ± 0.063	0.700–0.949	52.74 ± 1.62	49.76–55.95
Green Bell Pepper	94.50 ± 1.86	91.59–98.49	0.469 ± 0.032	0.407–0.554	78.55 ± 2.08	74.26–83.37
Broccoli	89.81 ± 1.63	86.82–93.21	1.121 ± 0.083	0.957–1.307	80.28 ± 2.34	75.54–85.33
Turnip	94.49 ± 1.70	90.93–98.43	0.782 ± 0.051	0.677–0.906	42.37 ± 1.13	40.13–44.99

Ash Gourd	95.43 ± 1.76	91.91–98.73	0.641 ± 0.043	0.555–0.739	14.81 ± 0.50	13.95–15.82
Tomato	92.99 ± 1.66	90.35–97.14	0.351 ± 0.030	0.299–0.409	16.87 ± 0.53	15.97–17.82

773

774 For the following modelling strategy, there are a few features about **Table 3** that are relevant.

775 Green banana is a species with a clearly lower moisture region than the rest of the vegetables,

776 broccoli is distinguished by Relatively high ash levels. Some groups, however, overlap strongly

777 in one physicochemical. Although Vitamin C values are variable, they stay fairly distant from

778 each other, such as green bell pepper and turnip. Both of them have nearly the same mean

779 moisture percentage (94.50% and 94.49%) but almost the same mean Vitamin C (4.21% and

780 4.20%). The concentrations of these are 78.55 and 42.37 mg/ 100 g. story of ash gourd is that

781 it is also grown in high humidity. The Vitamin C content in region is 14.81 mg/100 g, which is

782 much lower compared to the other areas. These overlapping and Contrasting patterns

783 strengthen the argument for considering the prediction of Vitamin C as a multivariable problem

784 Instead, to assume that there is not one single measurement that will give an adequate

785 calibration signal. The ash profiles create another barrier to separation. The ash mean was the

786 highest for broccoli (1.121%), The levels of Vitamin C were higher in the sample than that of

787 other samples, while the mean of ash content was the lowest in tomato, 0.351%, It is

788 accompanied with a relatively low Vitamin C content. The pattern was not monotonic across

789 all vegetables, however. Green bell pepper had relatively low ash at 0.469% despite its high

790 Vitamin C concentration, and bitter gourd had higher ash than green bell pepper while

791 containing considerably less Vitamin C. The dataset therefore contains structured but non-

792 simple relationships, providing a reasonable basis for testing nonlinear tree-based regression

793 rather than imposing a single linear correspondence between individual physicochemical

794 measurements and Vitamin C [47] .

795 The difference in predictor scales was retained rather than removed simply for numerical

796 uniformity. This choice was compatible with the tree-based modelling framework used in the

797 study, in which predictions are constructed through feature-specific split thresholds rather than

798 through distances measured across features. Random-forest and histogram-gradient-boosting

799 implementations, for example, construct regression trees by searching feature-specific splits,

800 allowing predictors expressed on different measurement scales to be used directly. Retaining

801 the original units also made the later interpretation more intuitive: moisture remained expressed

802 as percentage moisture, ash as percentage residue, and the two SD descriptors in their recorded

803 laboratory scales. This was particularly useful for the subsequent explainability and feature-

ablation analyses. Rather than interpreting transformed or standardized quantities, the contribution of each variable could be related directly to the laboratory measurement from which it originated. The modelling dataset therefore preserved a close connection between the experimental workflow and the machine-learning workflow: the inputs remained physically interpretable measurements, Vitamin C remained an independently established reference response, and vegetable identity was retained only for balanced evaluation rather than used as a shortcut for prediction.

2.7. Balanced training-test split and five-fold cross-validation

A balanced 80:20 development–test split was used to preserve equal representation of all seven vegetable groups throughout model evaluation. The partition was generated using random state 42, resulting in 40 observations from each vegetable being assigned to the development set and the remaining 10 observations being reserved for final testing. This produced 280 development observations and 70 held-out test observations in total. Using a fixed random state ensured that the same partition could be reproduced in subsequent runs of the analysis. The held-out test set was kept completely separate from hyperparameter optimization. Model tuning was performed only within the 280-observation development set, while the 70 test observations were retained for the final assessment after model selection. This separation was important because repeatedly consulting the test set during hyperparameter selection can allow information from the final evaluation data to influence model development and produce overly optimistic performance estimates. Standard machine-learning evaluation practice therefore recommends retaining an independent test partition while using cross-validation within the development data for model selection.

Within the 280-observation development set, five-fold cross-validation was performed while preserving the vegetable-group distribution in every fold [48]. Because each vegetable contributed exactly 40 development observations, each validation fold contained eight observations from each of the seven vegetable groups. Consequently, every validation fold consisted of 56 observations, while the corresponding fold-training subset contained the remaining 224 observations, or 32 observations from each vegetable. Stratified cross-validation is designed to preserve the relative representation of predefined groups across folds, which was particularly useful here because the dataset had an exactly balanced seven-group structure.

Table 4. Balanced data allocation used for model development and testing.

Vegetable	Total	Training	Testing	Validation observations per 5-fold CV fold
Green Banana	50	40	10	8
Bitter Gourd	50	40	10	8
Green Bell Pepper	50	40	10	8
Broccoli	50	40	10	8
Turnip	50	40	10	8
Ash Gourd	50	40	10	8
Tomato	50	40	10	8
Total	350	280	70	56

836

837 The allocation shown in **Table 4** ensured that no vegetable group was disproportionately
838 represented during either model development or final testing. Because every group contributed
839 the same number of observations, differences in model performance were less likely to arise
840 simply from unequal group sizes. More importantly, the same balanced structure was retained
841 during cross-validation, so each model was repeatedly trained and validated using data drawn
842 from all seven vegetables rather than being evaluated on folds dominated by only a subset of
843 the groups. The five-fold procedure also made more efficient use of the available development
844 data than a single fixed validation subset. In each iteration, four folds were used for model ML
845 fitting and the remaining fold was used for validation, after which the validation role rotated
846 until every development observation had contributed once to validation. Cross-validation is
847 commonly used in this way during model selection because it provides performance estimates
848 across several resampled partitions while leaving the independent test set untouched for the
849 final evaluation.

850 The balanced structure also made the fold composition particularly transparent. Starting from
851 40 development observations per vegetable, division into five folds placed eight observations
852 from each vegetable in the validation subset and 32 in the corresponding fold-training subset.
853 Each cross-validation iteration therefore trained the candidate model on 224 observations and
854 evaluated it on 56 observations. Hyperparameter combinations were compared using
855 performance obtained across these five validation folds rather than on the final test set. This
856 design was intended to separate two different stages of model assessment. Cross-validation
857 within the development set was used to compare candidate configurations and assess how
858 consistently they performed across different internal partitions. The independent 70-
859 observation test set was then used only after model selection to provide the final held-out
860 performance estimate. Keeping these stages separate reduces the possibility that the final test
861 score merely reflects repeated adaptation of the model to the same evaluation data.

At the same time, the interpretation of this validation strategy has an important boundary. Stratification preserved vegetable representation rather than holding out entire vegetable groups. All seven vegetables were therefore present during model development as well as final testing. The resulting cross-validation and test scores describe predictive performance for new observations drawn from the seven vegetable groups already represented in the dataset. They do not establish how accurately the model would perform for a completely unseen vegetable species. Group-based validation procedures, such as leave-one-group-out designs, are specifically intended for that stronger form of evaluation because they exclude an entire predefined group from model fitting. For the purpose of the present study, however, the balanced 80:20 split combined with five-fold group-preserving cross-validation provided a controlled framework for comparing the ten regression algorithms within the defined seven-vegetable domain. Every model was evaluated under the same development–test allocation, every cross-validation fold retained equal representation of the vegetable groups, and the final test observations remained outside the hyperparameter-search process until model selection was complete. [49] .

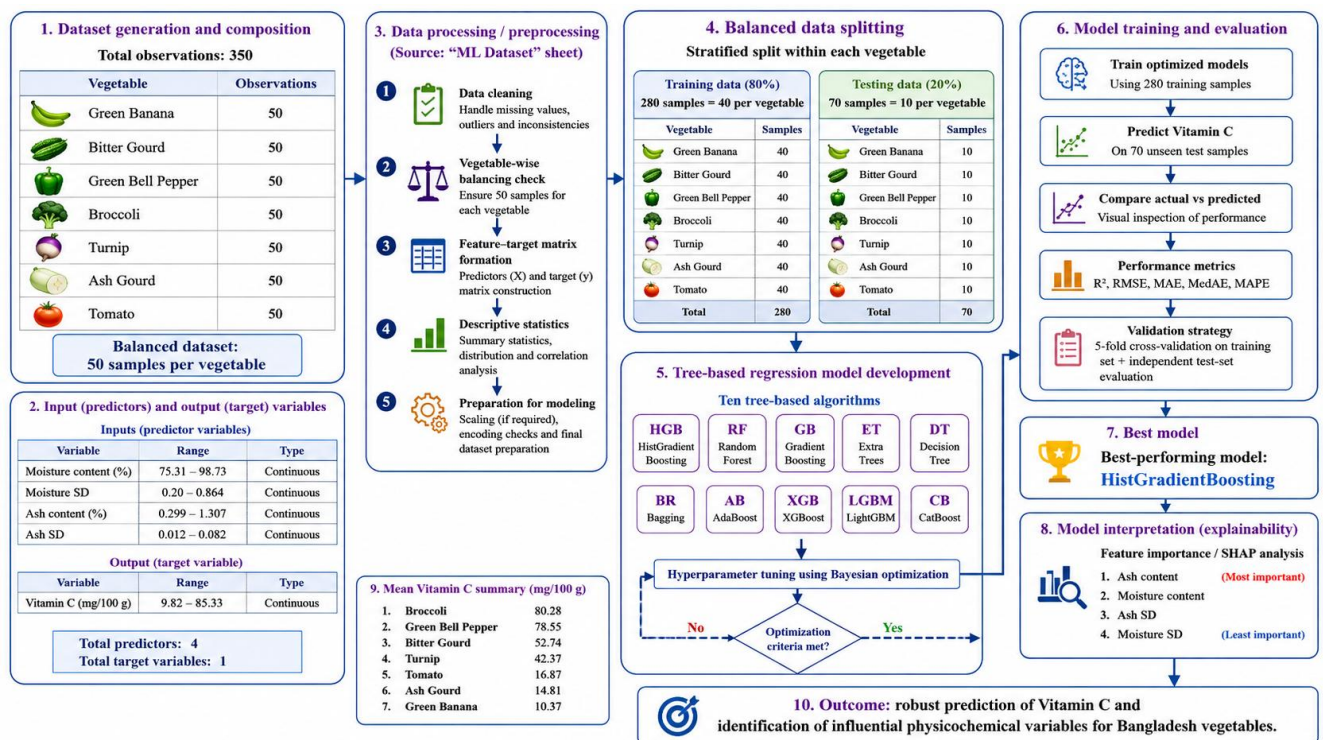


Fig. 3a. Balanced train–test allocation used for all seven vegetable groups. Five-fold cross-validation was performed exclusively within the 280-observation training partition, while the 70-observation test set remained outside model tuning and selection.

The 80:20 allocation was carried out separately within each vegetable group rather than by randomly dividing the 350 observations as a single pooled dataset. This ensured that every vegetable contributed exactly 40 observations to model development and 10 observations to the final test set. The resulting partition therefore contained 280 development observations and 70 held-out observations, while maintaining the original balance of 50 observations per vegetable. This group-preserving strategy was used to prevent the evaluation set from becoming disproportionately influenced by any one vegetable simply because of random sampling variation. The internal cross-validation stage was applied only after the 70 test observations had been set aside. Within the 280-observation development partition, five-fold cross-validation was stratified according to vegetable identity so that each fold retained the same seven-group composition. Because every vegetable contributed 40 development observations, each validation fold contained eight observations from each vegetable: eight green banana, eight bitter melon, eight green bell pepper, eight broccoli, eight turnip, eight ash gourd, and eight tomato observations. Each validation fold therefore contained 56 observations in total. The complementary fold-training subset consisted of 224 observations, corresponding to 32 observations from each of the seven vegetable groups. This balanced construction meant that every candidate model was trained and validated under the same group representation throughout the cross-validation process. Stratified fold construction is specifically designed to preserve the representation of predefined groups across partitions. The role of the cross-validation stage was different from that of the final test set. The five internal folds were used repeatedly during model development to evaluate candidate configurations and support hyperparameter optimization. In contrast, the 70-observation test set was not used to guide hyperparameter selection or to determine the ordering of candidate models. It remained untouched until the hyperparameter search and CV-based model comparison had been completed.

Keeping these stages separate was important for obtaining a more defensible estimate of predictive performance. When the same evaluation data are repeatedly consulted while selecting models or hyperparameters, the model-development process can gradually adapt to that particular dataset, even when the observations are not explicitly used for fitting. The resulting performance estimate may then become overly favourable. This concern has been demonstrated in methodological studies of cross-validation-based model selection, where reusing validation information for both optimization and performance estimation can introduce selection bias [50]. The evaluation workflow was therefore organized as a sequence rather than

as repeated inspection of the final test results. During each cross-validation iteration, candidate models were fitted on 224 observations and evaluated on the corresponding 56-observation validation fold. Validation performance across the five folds was then used to compare model configurations and assess their consistency across different internal partitions. Only after this process had been completed was the selected model evaluated against the independent 70-observation test set. This structure also allowed the reported performance measures to be interpreted at three distinct levels. Fold-training performance described how closely a model fitted the observations available during a particular cross-validation iteration. Fold-validation performance indicated how well the same model transferred to observations withheld within the development partition. The final held-out test performance provided a separate assessment using observations that had played no role in hyperparameter search or CV-based model ordering. Maintaining these distinctions made it possible to examine not only predictive accuracy but also the gap between training and validation behaviour, which was subsequently considered when evaluating model stability and selecting the final regression model. The balanced allocation shown in **Fig. 3a** was particularly useful for this dataset because Vitamin C concentrations differed markedly among the seven vegetable groups. A purely pooled random split could, by chance, have produced unequal representation of individual vegetables in the development or test partitions, making overall performance partly dependent on the composition of a particular random split. By fixing the allocation at 40 development and 10 test observations per vegetable and maintaining the same group balance within five-fold cross-validation, each algorithm was evaluated under a common and transparent sampling structure.

At the same time, this design defines the scope of the reported validation evidence. Because all seven vegetables occurred in both the development and test sets, the held-out assessment represents prediction for new observations within vegetable groups already encountered during model development. It does not represent a leave-one-vegetable-out experiment in which an entirely unseen vegetable species is presented to the model. The reported test performance should therefore be interpreted as within-domain prediction for the seven-vegetable experimental space rather than as evidence of universal generalization to vegetables absent from the training data.

2.8. Machine-learning models

Ten tree-based regression methods were compared under the same data partition: HistGradientBoosting, Random Forest, Gradient Boosting, Extra Trees, Decision Tree, Bagging, AdaBoost, XGBoost, LightGBM, and CatBoost. The group includes a single-tree

baseline, bagged and randomized ensembles, conventional gradient boosting, histogram-based boosting, and modern boosting libraries. Using this range made it possible to compare models with different bias-variance behaviour and different approaches to regularization rather than assuming that one family would be best in advance [51] .

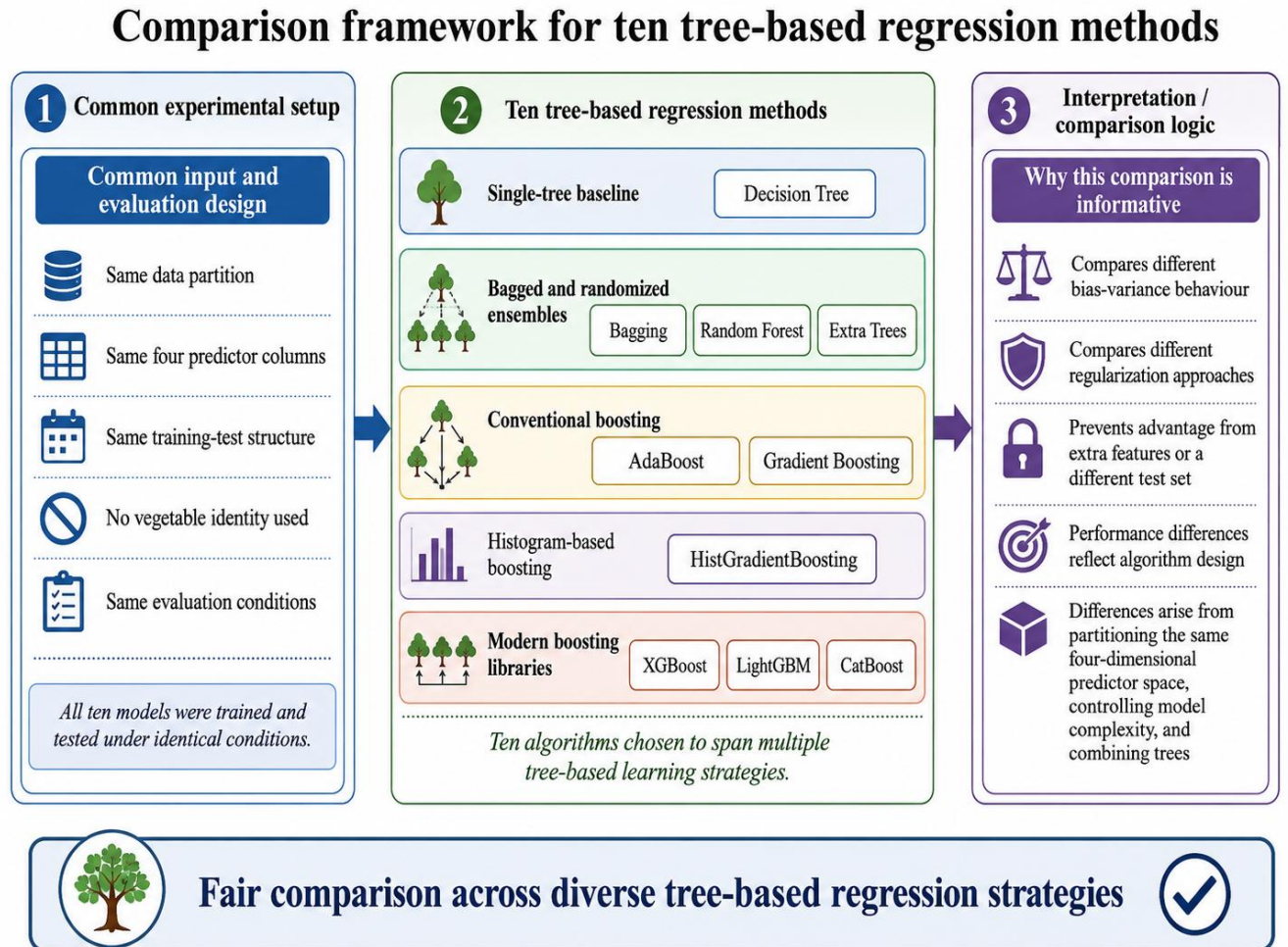


Fig. 3b. Comparison framework used for the ten tree-based regression methods. The figure summarizes the shared experimental setup, the five groups of tree-based learning strategies considered in this study, and the logic used to interpret performance differences under identical input and evaluation conditions.

To make the comparison among the ten regression algorithms transparent, a separate framework figure was included as **Fig. 3b**. This figure was used to summarize the logic of the model comparison before presenting optimization results and performance tables. Its main purpose was to show that all candidate models were evaluated under the same experimental conditions and that the comparison was therefore methodologically fair. In other words, the figure helps the reader understand that any difference in predictive performance did not arise

from different predictor sets, unequal train–test allocation, or privileged access to vegetable identity, but from the learning strategy of the algorithm itself.

The left panel of **Fig. 3b** highlights the common experimental setup used throughout the comparative analysis. All ten algorithms were trained on the same balanced data partition, received the same four predictor variables, followed the same training–test structure, and were evaluated under the same cross-validation and test conditions. Vegetable identity was intentionally excluded from the predictor matrix, and no variable derived from the Vitamin C titration was supplied to the regressors. This part of the figure therefore reinforces a central design principle of the study: the models were required to learn Vitamin C variation from physicochemical information alone [52] .

The central panel of **Fig. 3b** groups the ten algorithms according to their broader tree-based learning strategy. Decision Tree was treated as the single-tree baseline. Bagging, Random Forest, and Extra Trees represented bagged or randomized ensemble learning. AdaBoost and Gradient Boosting represented conventional boosting approaches based on sequential improvement. HistGradientBoosting was shown separately because it uses a histogram-based boosting strategy. XGBoost, LightGBM, and CatBoost were grouped as modern boosting libraries that extend gradient-boosted decision-tree learning through different optimization and regularization schemes. Presenting the models in this structured way makes the comparison easier to interpret than simply listing ten algorithm names in sequence [51] .

The right panel explains why this comparison is informative. Since all models were trained and tested under identical conditions, the reported differences can be interpreted as differences in algorithm design rather than artefacts of the dataset split. More specifically, the figure communicates that the comparison captures differences in bias–variance behaviour, regularization strategy, tree construction, and ensemble logic. A single decision tree partitions the predictor space through one recursive structure, whereas bagging-based models reduce instability by averaging multiple trees. Randomized ensembles introduce additional variability in split generation, while boosting-based models progressively refine prediction errors through sequential tree addition. Modern boosting libraries further adapt this general idea using different computational and regularization designs. The figure therefore serves as a visual bridge between the conceptual categories of tree-based regression and the empirical comparison presented later in the Results section. In the manuscript, **Fig. 3b** should be referred to when first explaining why ten tree-based methods were selected instead of relying on only one or two popular models. It can also be cited again in the Results discussion when explaining

that performance differences reflect how each algorithm handled the same four-dimensional predictor space. In this way, the figure functions not merely as an illustration, but as a compact summary of the comparative design underlying the entire modelling analysis.

2.9. Bayesian hyperparameter optimization

Bayesian optimization was always used to tune the model-specific hyperparameters of the ten tree-based regressors. The search was restricted to the 280-observation development partition established previously, while the 70 held-out test observations remained outside the optimization process. This separation was intentional: hyperparameters were selected from repeated evidence generated within the development data, rather than from the final test performance. Maintaining an independent test set is particularly important during model selection because repeated optimization against the same evaluation data can introduce selection bias and produce an overly favourable estimate of generalization performance [53].

Unlike an exhaustive grid in which a predetermined set of parameter combinations is evaluated without regard to the outcomes of earlier trials, Bayesian optimization uses information accumulated from previously evaluated configurations to guide the search toward parameter regions that appear more promising. This adaptive strategy is well established for machine-learning hyperparameter optimization, particularly when evaluating each candidate requires repeated model fitting and validation. Snoek et al. demonstrated the practical use of Bayesian optimization for automatically tuning machine-learning algorithms, while later work has continued to develop the approach for computationally expensive hyperparameter-search problems [54].

In the present study, the optimization criterion was the mean cross-validated mean squared error obtained from the five-fold procedure applied within the development partition. Hyperparameter selection can therefore be expressed by **Equation (8)**.

$$\theta^* = \arg \min_{\theta} CV - MSE(\theta) \quad (8)$$

Here, θ represents a candidate set of model-specific hyperparameters, and $CV - MSE(\theta)$ represents its mean validation MSE across the cross-validation folds. The optimization procedure sought the parameter configuration associated with the lowest cross-validated MSE. Importantly, test-set error was not part of this objective. The final test partition therefore remained independent of the numerical criterion used to select the hyperparameters. Bayesian optimization was particularly appropriate for the present comparison because tree-based algorithms contain interacting hyperparameters rather than a single isolated tuning control. For example, increasing tree depth can increase model flexibility but may also increase variance,

while reducing the learning rate in boosting methods often changes how many boosting iterations are needed to construct an adequate ensemble. Leaf-size restrictions, sampling parameters, and regularization controls can further alter the effective complexity of the fitted model. Hyperparameter optimization therefore involved searching combinations of settings rather than interpreting each parameter independently. Bayesian optimization is designed for this type of black-box tuning problem, where the relationship between a parameter configuration and validation loss is not available in closed form and each evaluation requires an actual model fit.

The retained optimization records showed that the search effort differed somewhat among algorithms. Candidate-level histories were available for nine of the ten models, with between 35 and 45 candidate configurations retained depending on the algorithm. The final Decision Tree package preserved the selected hyperparameter configuration but did not retain the numbered candidate history or optimization time. These metadata were therefore reported as not retained rather than estimated retrospectively. This choice preserves the distinction between information present in the archived modelling output and information that could only be reconstructed by rerunning or inferring the original search.

A further precaution was taken when interpreting the stored candidate histories. The exported candidate index was treated as an identifier for an evaluated configuration rather than as definitive evidence of wall-clock search order. The packaged tables do not establish that the candidate numbering preserves the precise chronological sequence in which Bayesian optimization evaluated the configurations. Consequently, the candidate-history plots are interpreted as records of the evaluated search landscape and selected candidate, not as exact temporal convergence trajectories.

The optimization and cross-validation results retained for each model are summarized in **Table 5**.

Table 5. Bayesian optimization and cross-validation summary for the ten regression models.

Model	Candidates	Opt. time (min)	CV val R ²	CV R ² SD	CV gap	CV val MSE	Test R ²	Test RMSE
HistGradientBoosting	45.00	2.60	0.92	0.05	0.05	60.31	0.95	6.00
Random Forest	40.00	5.16	0.91	0.05	0.08	68.25	0.96	5.42
Gradient Boosting	45.00	4.12	0.91	0.04	0.09	65.48	0.96	5.80
Extra Trees	40.00	3.63	0.91	0.04	0.07	70.22	0.95	5.91
Decision Tree	n/r	n/r	0.89	0.06	0.08	n/r	0.93	7.08
Bagging	40.00	3.77	0.89	0.04	0.07	81.20	0.94	6.68
AdaBoost	45.00	6.20	0.92	0.08	0.07	61.21	0.95	6.22
XGBoost	45.00	2.34	0.91	0.03	0.09	70.71	0.96	5.63
LightGBM	40.00	2.53	0.90	0.04	0.07	76.36	0.91	8.13
CatBoost	40.00	5.11	0.91	0.04	0.09	68.10	0.96	5.74

n/r = not retained in the final packaged Decision Tree optimization table. Missing metadata were not reconstructed or inferred.

Table 5 brings together two different types of information. The candidate counts and recorded optimization times describe the computational search that was retained in the final model packages, whereas the cross-validation statistics describe how the resulting configurations behaved across the development folds. Test R^2 and RMSE are included to provide the eventual held-out performance of each optimized model, but these values were not used as the Bayesian optimization objective. This distinction is preserved as it is important to keep these explicit since model tuning and final performance estimation differ. However, Cawley and Talbot demonstrated that optimizing a model-selection criterion can be subject to overfitting, further underscoring the need to distinguish evidence for model selection from actual performance evaluation. The candidate-level search behaviour that was retained, is displayed in **Fig. 4**. The figure does not require the plot to indicate a specific converging path in time, but instead illustrates the ranges of candidate configurations that could be contained within each of the optimization histories packaged, and designates the configuration labeled as selected. This is the interpretation that is consistent with what we are seeing in the candidate index that was archived, and it does not try to give a temporal interpretation that's not reflected in the data that are stored.

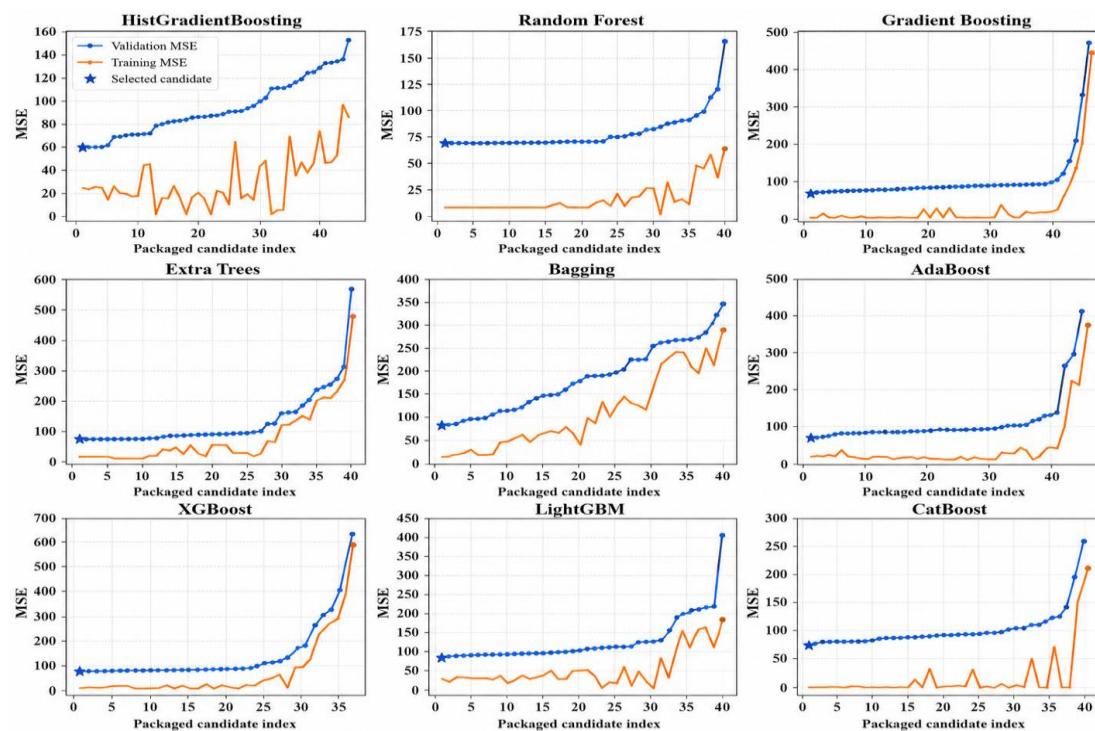


Fig. 4. Returned the nine final packages for the 9 models for which a package of Bayesian-search candidate histories were provided. MSE level of training & validation for candidates.

The X axis shows the exported candidate. The points with a star are candidate marked as index not a confirmed wall-clock sequence. Selected from the kept optimization history.

The computational effort associated with the searches is summarized separately in **Fig. 5**. Across the nine models for which these metadata were retained, the optimization histories contained between 35 and 45 candidate evaluations. XGBoost had the lowest recorded optimization time at 2.34 min, while AdaBoost had the highest at 6.20 min. HistGradientBoosting evaluated 45 retained candidate configurations in approximately 2.60 min. These values are useful for documenting the computational burden of the actual modelling workflow, but they should not be interpreted as a benchmark of the intrinsic speed of the algorithms.

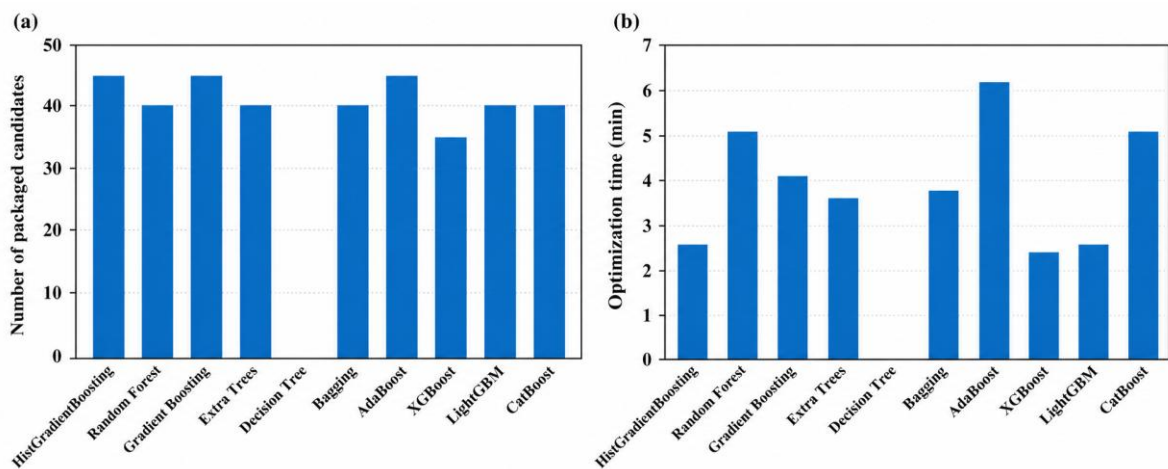


Fig. 5. Computational effort recorded for Bayesian hyperparameter optimization: (a) number of candidate evaluations retained in the final optimization packages and (b) recorded optimization time. Decision Tree is omitted from these comparisons because its final package did not retain candidate-count or optimization-time metadata.

Direct runtime comparisons require caution because the search spaces, numbers of evaluated candidates, hyperparameter definitions, software implementations, and computational operations differ across models. Bayesian optimization itself is often motivated by the cost of evaluating expensive hyperparameter configurations, and methodological work has emphasized that computational cost can vary considerably across configurations and algorithms. For this reason, the times in **Fig. 5** are reported as records of the present experiment rather than evidence that one modelling library is universally faster or more computationally efficient than another. The consistency of the procedure was more important than any actual duration. Each optimized model was selected based on repeated development partition evidence and the cross validated MSE (not the final test error) was the optimization target. The CV validation R^2 , the SD for the fold-to-fold CV, and the difference in absolute values between

CV training and validation now became available, not just as a measure of the predictive power, but also to measure stability and potential overfitting. This was relevant because the overall best model that was held out was not necessarily the model that had the most favourable score for that single fold in the internal validation. The optimization step was thus set up as a link between the first ten-model comparison and the subsequent model-ranking process. Bayesian search was employed to determine if there was a set of proper model specific hyperparameters in the training set, five-fold cross validation gave additional evidence from the training set as to the resulting set of model specific hyperparameters, and the test set was not used during the training phase. Missing optimization metadata were left explicitly unresolved rather than reconstructed, and the retained candidate indices were interpreted only to the extent supported by the archived files. Together, these choices allowed the optimization results to be reported as reproducible computational evidence without overstating what the stored search histories could demonstrate.

2.10. Selected hyperparameters

The Bayesian search did not converge on a common model structure across the ten algorithms. Instead, each method required a different combination of complexity, regularization, sampling, and ensemble-size controls to obtain its strongest cross-validated performance. This was expected because the candidate algorithms represent different tree-construction strategies, even though they were trained on exactly the same four physicochemical predictors. Random forests control prediction through the number and structure of randomized trees, whereas gradient-boosting methods additionally depend on the interaction between learning rate, number of boosting stages, tree depth, subsampling, and regularization.

HistGradientBoosting, which subsequently ranked first under the CV-first model-selection rule, was fitted with a learning rate of 0.01 and 568 boosting iterations. Its final configuration also used a maximum tree depth of 7, a maximum of 35 leaf nodes, a minimum of 26 observations per leaf, 177 histogram bins, and an L2 regularization value of 0.0783649896. These settings describe a model in which relatively small updates were accumulated over many boosting iterations while tree growth and terminal-node size were constrained. In histogram-based gradient boosting, parameters such as learning rate, number of iterations, leaf number, minimum leaf size, binning, and L2 regularization directly control how the ensemble is constructed and how complex individual fitted structures are allowed to become [55]. Random Forest reached its optimized configuration using 551 trees, a maximum tree depth of 12, and a

max_features value of approximately 0.8843. This is not necessarily a sign of more complex models than boosting models; the ensemble size is relatively large due to the fact that the models are many, and are actually many random trees that are averaged together. Breiman's formulation not only involves the number of trees and the degree of feature randomization but also the diversity and correlation between the trees and is part of a larger ensemble design, not just a specific tuning choice.

Gradient Boosting arrived at a noticeably different solution. The optimized model used 750 estimators with a learning rate of approximately 0.03011, maximum depth of 6, maximum leaf nodes of 15, minimum samples split of 16, and a subsample fraction of approximately 0.7989. The combination of a relatively small learning rate with a comparatively large number of boosting stages is consistent with the well-known trade-off between these two parameters in gradient boosting: reducing the contribution of each individual tree generally changes the number of sequential stages required to build the final additive model. Subsampling introduces another bias–variance control by fitting individual boosting stages on only a fraction of the available observations [56]. Taken together, these optimized configurations illustrate why the final analysis should not be described as a comparison of ten off-the-shelf or default algorithms. Although every model received the same moisture content, ash content, moisture SD, and ash SD inputs, the Bayesian search allowed each algorithm to adapt its own structural and regularization parameters to the development data. The resulting comparison therefore reflects optimized versions of the respective learning strategies rather than differences that might simply arise because one model happened to have more suitable default settings.

The complete selected hyperparameters for all ten regressors are reported in Supplementary Tables S2A–S2J. Keeping the full parameter sets in the Supplementary Information avoids interrupting the main text with lengthy model-specific lists while preserving enough detail for the fitted models to be reproduced and independently examined. Transparent reporting of datasets, model configurations, evaluation procedures, and computational choices is also emphasized in current recommendations for reproducible machine-learning-based science [57].

Retaining the settings for every candidate model was particularly important in the present study because model selection was not determined solely by the highest final test R^2 . HistGradientBoosting was selected under the predefined CV-first rule because it produced the strongest mean cross-validation R^2 together with a comparatively controlled train–validation gap. Random Forest, however, produced the highest single held-out test R^2 . Reporting only the

final HistGradientBoosting configuration would therefore conceal an important feature of the comparison: different algorithms appeared strongest under different evaluation criteria.

For this reason, the hyperparameter records in Supplementary **Tables S2A–S2J** should be viewed as part of the evidence supporting the model comparison rather than as secondary implementation details. They demonstrate that for each algorithm, it has been given the chance to run under the contextually optimised settings of that algorithm and that the following ranking can be interpreted in the light of those optimised settings. This is particularly true for flexible tree ensembles, where very subtle differences in learning rate, tree depth, constraints on the leaf size, feature sampling, or size of the ensemble can make a substantial difference in the fit vs. generalization trade off.

2.11. The assessment of the performances and the ranking of the final model.

Multiple complementary regression measures were used for the evaluation of the model performance, instead of a single goodness of fit measure. This was significant because two models may achieve the same R^2 , but have widely varying sizes, distribution or importance of prediction errors. Studies on the assessment of predictive-error have also stressed that no single measure of accuracy can tackle all aspects of model behavior and that that absolute, squared, and relative error measures ask different questions. In this study, the R^2 , MSE, RMSE, MAE, MedAE, and MAPE were therefore selected and analysed together and from various other complementary angles to determine the model performance. The coefficient of determination R^2 was used to quantify the proportion of the variation in the observed values of Vitamin C that could be reproduced by the predicted values, compared with a model that predicts the overall mean. The closer the values are to 1, the closer the agreement is with the observed response pattern; the closer the values are to 0, the greater a proportion of the variation is not accounted for by the fitted model [58]. R^2 was calculated using **Equation (9)**.

Equation (9). Coefficient of determination

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (9)$$

where y_i is the experimentally measured Vitamin C concentration for observation i , $\hat{y}_i$ is the corresponding model prediction, and $\bar{y}$ is the mean of the observed Vitamin C values.

Because R^2 alone does not describe the magnitude of prediction errors in the original response scale, error-based metrics were also calculated. Mean squared error assigns progressively greater weight to larger deviations because each residual is squared before averaging [58].

This makes MSE particularly sensitive to observations for which the prediction is far from the experimental reference value. It was calculated using **Equation (10)**.

Equation (10). Mean squared error

$$MSE = (1/n) \sum (y_i - \hat{y}_i)^2 \quad (10)$$

RMSE was obtained by taking the square root of MSE, as shown in **Equation (11)**. This retains the stronger penalty applied to large residuals while returning the error to the same unit as the response variable. In the present study, RMSE is therefore directly interpretable in mg/100 g of Vitamin C.

Equation (11). Root mean squared error

$$RMSE = \sqrt{[(1/n) \sum (y_i - \hat{y}_i)^2]} \quad (11)$$

MAE provided a complementary view of prediction error by averaging the absolute difference between experimental and predicted concentrations without squaring the residuals. As a result, individual large errors influence MAE less strongly than they influence MSE or RMSE. Because MAE also remains in mg/100 g, it provides a straightforward estimate of the average absolute distance between the model prediction and the laboratory reference value.

Equation (12). Mean absolute error

$$MAE = (1/n) \sum |y_i - \hat{y}_i| \quad (12)$$

Median absolute error was included to describe the typical prediction error in a way that is less influenced by a small number of unusually poor predictions. Rather than averaging all absolute residuals, MedAE identifies their median value. This distinction was useful in the present dataset because a model could perform closely for most samples while still producing a small number of comparatively large errors. In the present dataset this distinction proved to be very useful since a model may have a high correlation coefficient for the majority of the samples and a few relatively large errors.

Equation (13). Median absolute error

$$MedAE = median(|y_i - \hat{y}_i|) \quad (13)$$

To get a sense of how good or bad the model is, MAPE values were calculated, which each scales the absolute error against a different experimental measured Vitamin C value. When the same absolute error has a different magnitude at different points in the response range, percentage-based error measures can be useful, but need to be carefully interpreted because observations with lower reference values will be assigned more weight. Furthermore, Hyndman and Koehler have mentioned the general problems with percentage accuracy, especially close

to zero. In this particular dataset, the concentration of Vitamin C was not zero, so the division by zero did not occur, however, MAPE was not only used to determine the quality of the model; it was used as an additional parameter.

Equation (14). Mean absolute percentage error

$$MAPE = \left(\frac{100}{n}\right) \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (14)$$

This is particularly important for the Vitamin C range included in the study. The absolute error of ± 3 mg/100 g means much more for a response that shows a low mean at one end of the distribution, and a high mean at the other, than it does for a response that has a moderate-low mean at one end and a moderate-high mean at the other. For a sample containing approximately 80 mg/100 g Vitamin C, such an error represents less than 4% of the reference concentration, whereas the same 3 mg/100 g deviation for a sample containing approximately 10–15 mg/100 g represents a much larger proportional error. RMSE and MAE describe these cases in the original concentration unit, MedAE indicates the typical absolute error with reduced sensitivity to extreme residuals, and MAPE provides the additional relative perspective. Using the metrics together therefore gives a more informative description of model behaviour than any one value considered in isolation [59].

Final model selection was based on a predefined CV-first ranking strategy. The primary criterion was mean five-fold cross-validation validation R^2 , because this quantity summarized predictive performance across repeated internal validation partitions rather than on a single held-out sample. When models were close on mean validation R^2 , the fold-to-fold SD of validation R^2 was considered next, with lower variability indicating more consistent behaviour across the five validation folds. The absolute difference between mean training and validation performance was then considered as an additional indication of whether strong training performance was being maintained during validation.

This ordering was chosen to place greater weight on repeated validation evidence than on a single favourable test result. Methodological work on model selection has shown that the variability of cross-validation estimates matters because a noisy selection criterion can itself be overfitted during model comparison [60]. Cawley and Talbot therefore emphasize that both the expected performance and the variance of model-selection estimates deserve attention when comparing competing learning algorithms. This rationale is consistent with considering mean CV performance together with fold-to-fold variability and the training–validation gap in the present analysis.

The held-out test metrics remained important, but they served a different purpose. Test RMSE and MAE were used as supporting measures after CV-based comparison rather than as the primary inputs used to determine model rank. This distinction prevented the final model choice from being driven simply by whichever algorithm happened to obtain the most favourable score on one 70-observation test partition. The independent test set instead provided a final check of how the selected and competing models behaved on observations that had not participated in hyperparameter optimization. This selection proved to be important when the optimized models were compared. Random Forest had the highest R squared for a single test (0.96117) and lowest root mean square error (RMSE) for test (5.4236 mg/100 g). The test R^2 value for HistGradientBoosting was 0.95243 and RMSE was 6.0028 mg/100 g. Hence, Random Forest would have been preferred had they only used the last test R^2 to decide on the model. However, the repeated cross validation evidence provided a different ordering. HistGradientBoosting outperformed the other classifiers with the highest mean CV validation R^2 of 0.92050 for the same mean training–validation gap of 0.04755, whereas RF had an $R^2=0.90994$ and an absolute training–validation gap of 0.0768. HistGradientBoosting was then selected as the ultimate model for explaining, as specified by the CV-first rule. This choice was not meant to imply that Random Forest did poorly, rather it was a way to demonstrate that the held-out results of that method were of outstanding quality. Instead, however, it was chosen because it had the most convincing repeated validation evidence, rather than because of the best performance on a single test set. The last evaluation scheme accordingly separated the tasks of model fitting, model selecting, and final performance assessment. The main evidence for model ordering was provided by cross-validation, while the fold to fold variability and the training – validation gap were used to provide additional evidence of stability; the untouched test partition finally gave the model an independent check of performance. Such a distinction is relevant since performance evaluation and model selection are distinct but related tasks and if the same finite-sample evidence is used for both, then one can reach an overly positive conclusion about the selected model.

2.12. Introduce interpretability features of the model and SHAP. Present model interpretability features and SHAP.

The features were interpreted after comparing models, allowing the explanation step to be separated from the model selection step. A first-level analysis of the feature importance patterns was conducted over the ten fitted regressors in order to check if the same physicochemical features came up as the most important features repeatedly. The selected

model (HistGradientBoosting) was then detailedly interpreted by means of SHAP. This method is part of the general framework for explainable machine learning that Lundberg and Lee proposed, where each prediction is broken down into a baseline prediction and contributions for each individual feature[61] .

The relation between model output and expected output and the effect of each predictor for an individual observation was presented by **Equation (15)**.

Equation (15). SHAP decomposition for a single prediction of a model.

$$f(x) = E[f(x)] + \sum_j \varphi_j \quad (15)$$

In this case, the model baseline is denoted by $E[f(x)]$ and the contribution of predictor j to predicting a particular observation by the predictor is denoted by the function $\sum_i \varphi_j$. SHAP therefore provides an observation-level explanation of how individual features move a prediction away from the model baseline. For tree-based models, these local explanations can also be aggregated across observations to reveal broader patterns in how the fitted model uses its predictors.

Global importance was summarized from the mean absolute SHAP contribution of each feature according to **Equation (16)**.

Equation (16). Global SHAP feature importance

$$Global\ importance_j = (1/n) \sum_i |\varphi_{ij}| \quad (16)$$

Using absolute SHAP values makes it possible to summarize the overall magnitude of a feature's contribution without allowing positive and negative effects to cancel one another. This type of interpretation has increasingly been used in food-quality prediction studies. For example, SHAP has been applied to identify informative spectral regions for Vitamin C prediction in sweet potato and, more recently, to interpret boosting models developed for Vitamin C and other quality attributes in fruit juices. In the present study, however, SHAP values were interpreted only as evidence of how the fitted model used moisture- and ash-related information; they were not treated as evidence of a causal relationship with Vitamin C [62]. SHAP interpretation was complemented by progressive feature ablation. The two analyses were used together because they address slightly different questions. SHAP indicates how strongly the fitted model relies on each available predictor, whereas ablation tests whether prediction actually improves when additional measurements are introduced and the model is refitted. This distinction is useful because feature-importance values are model dependent and should not automatically be interpreted as proof that a variable provides independent predictive

information. Recent methodological work has likewise cautioned against relying on a single feature-importance measure without considering interactions, dependencies, and the predictive context in which the importance was calculated [63].

The ablation sequence was deliberately kept close to the laboratory workflow. In the first model, moisture was the only factor used. Ash was then included as second routine compositional measurement and moisture SD, and finally ash SD, to get the full four-predictor model. The R^2 and RMSE values for this reason were therefore interpreted as incremental improvements in the predictive performance of the model rather than four separate model fits. This stepwise sequence was also useful to assess the usefulness of the replicate-variability descriptors. Moisture SD and ash SD were not considered to be other chemical components of the vegetables, but just a representation of the differences between replicate laboratory determinations. They were thus justified in being included only if they brought in predictive information in addition to the mean moisture and ash. The interpretation strategy is a synthesis of three related perspectives of the same modelling problem. Cross-model feature importance was used to discover patterns, SHAP offered explanations both locally and globally, and progressive ablation was conducted to determine if the patterns provided measurable advantage for the model when it was retrained on increasingly informative laboratory measurements. This combination enabled the analysis to go beyond the ranking of features and also consider whether the variables found to be influential also enhance the practical prediction of experimentally measured Vitamin C.

2.13. A prediction formulation using the physicochemical characteristics was developed for Vitamin C.

The prediction problem was formalized separately from the chemical reference assay. Let M denote moisture content (%), A ash content (%), SD_M the moisture replicate standard deviation, and SD_A the ash replicate standard deviation. For a sample within the studied domain, the fitted machine-learning model estimates Vitamin C as a nonlinear function of these four quantities:

$$\hat{C}_{VitC} = f_{ML}(M, A, SD_M, SD_A) \quad (17)$$

where $\hat{C}_{VitC}$ is the predicted Vitamin C concentration (mg/100 g). Equation (15) is the central deployment equation of the study: the prediction uses moisture- and ash-related measurements only, whereas the iodometric result is required only as the reference target during model development and validation.

For HistGradientBoosting, the nonlinear prediction can be written in additive boosting form as:

$$\hat{CVitC}(x) = F_0 + \eta \sum_{t=1}^T h_t(x) \quad (18)$$

where F_0 is the initial model prediction, η is the learning rate, T is the number of boosting iterations, $h_t(x)$ is the contribution of the t -th fitted tree, and $x = [M, A, SD_M, SD_A]$. For the selected HistGradientBoosting model, $\eta = 0.01$ and $T = 568$. Unlike a conventional linear calibration equation, Eq. (16) represents a learned set of nonlinear splits and interactions and therefore cannot be reduced to one fixed coefficient for moisture and one fixed coefficient for ash [64].

A reduced screening model can also be written when replicate SDs are unavailable:

$$\hat{CVitC}_{reduced} = f_{ML}(M, A) \quad (19)$$

The reduced form was not treated as the final model, but its performance was evaluated in the ablation analysis to determine how much predictive information was already present in moisture and ash alone. Sample-level prediction error was calculated as:

$$Absolute\ error = |C_{ref} - \hat{CVitC}| \quad (20)$$

$$Percentage\ error\ (\%) = \left(\frac{|C_{ref} - \hat{CVitC}|}{C_{ref}} \right) \times 100 \quad (21)$$

where C_{ref} is the reference Vitamin C concentration obtained from the experimental dataset. These equations clearly separate the tasks of two stages: the chemical titration determines C_{ref} for model development and for checking, while the trained model calculates $\hat{CVitC}$ from the physicochemical inputs in the screening time. The deployment equations explicitly separate chemistry and prediction. The full model has $x = [M, A, SD_M, SD_A]$ with M and A being measured percentages and SD_M and SD_A being a summary of replicate variability. It is not possible to give one universal coefficient for ash and moisture in the fitted HistGradientBoosting function. Rather 568 boosting iterations make local changes around the model baseline, together with a learning price of 0.01. Additive model is therefore a mathematical representation that is additive at the ensemble level, but nonlinear with respect to the original four predictors. A simplified version with two arguments, $f(M, A)$ was kept as it has direct application. For a laboratory where reliable single moisture and ash measurements have been made, but where replicate SDs are not calculated, the reduced model provides an empirical, test $R^2 = 0.9348$ estimate. This improves to 0.9524 with the full 4 input formulation. This increase of R^2 of about 0.018 is real; and it is not a great deal compared with the increase of R^2 when ash is added to moisture, which is about 0.607. The comparison allows the paper to

distinguish the core predictive information from the incremental refinement provided by replicate variability.

3. Results and discussion

3.1. Physicochemical and Vitamin C profiles across the seven vegetables

The seven vegetables showed clearly differentiated physicochemical and Vitamin C profiles, although the degree of separation varied considerably among the measured variables. Green banana was most readily distinguished by its lower moisture content, whereas broccoli occupied the upper end of the ash distribution. Vitamin C showed the strongest group-wise differentiation, with broccoli and green bell pepper concentrated at the higher end of the response range and green banana, ash gourd, and tomato forming distinctly lower-value groups. This type of structured variation is important in predictive food-quality studies because a model is rarely learning from a completely homogeneous sample population; rather, it must recover the target from combinations of measurements that may overlap in one dimension while separating in another. Similar multivariate behaviour has been reported in food-quality prediction studies in which Vitamin C or other internal quality traits could not be adequately represented by a single measured feature [65].

The mean $\pm$ SD profiles of the five modelling variables provide an initial view of this structure. As shown in **Fig. 6**, the relative ordering of the vegetable groups changed depending on whether moisture, ash, or Vitamin C was considered. This lack of a common ordering across all three measurements is particularly relevant to the modelling problem because it suggests that no individual physicochemical descriptor can be expected to reproduce the full nutritional pattern on its own.

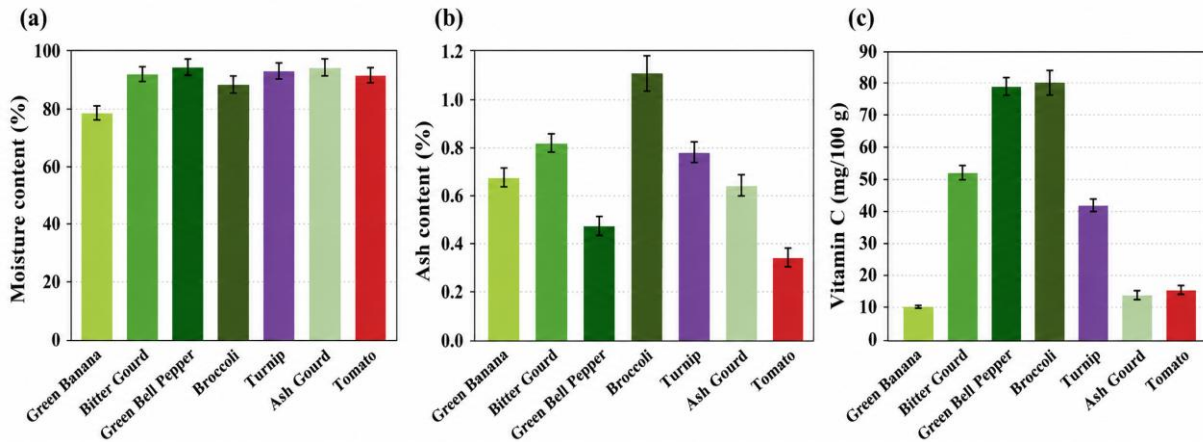


Fig. 6. Vegetable-wise mean $\pm$ SD profiles for (a) moisture content, (b) ash content, and (c) Vitamin C concentration in the 350-row modelling dataset.

The boxplots in **Figs. 7–9** provide a more detailed view of both within-group variability and between-group overlap. Moisture showed substantial overlap among several vegetables, particularly within the high-moisture region, indicating that water content alone offered limited discrimination among samples with very different Vitamin C concentrations. Ash provided clearer separation for some groups, although its relationship with Vitamin C was also not one-to-one. In contrast, the Vitamin C distributions formed much more distinct concentration bands. This combination of overlapping predictor distributions and more strongly separated target values provided the main statistical motivation for using multivariable nonlinear regression rather than a simple one-variable calibration [66]. Predictive studies of Vitamin C in citrus and sweet potato have likewise shown that useful nutrient prediction often depends on extracting information from combinations of indirect measurements rather than expecting a single measured feature to track Vitamin C consistently across all samples.

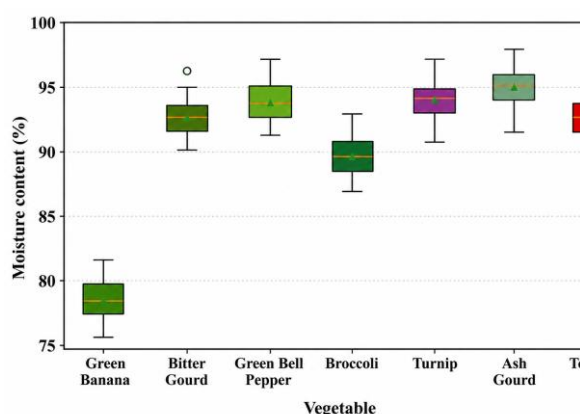


Fig. 7. Distribution of moisture content across the seven vegetable groups.

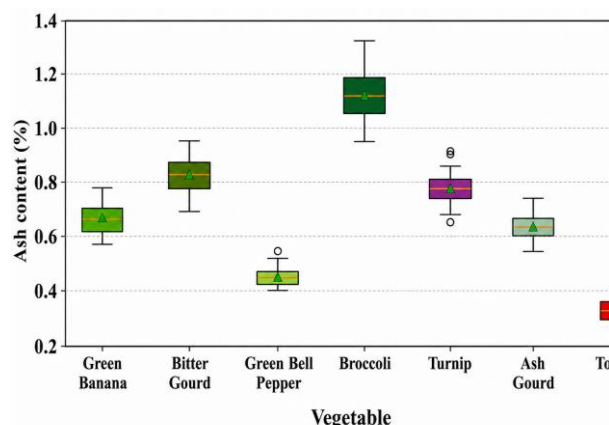


Fig. 8. Distribution of ash content across the seven vegetable groups.

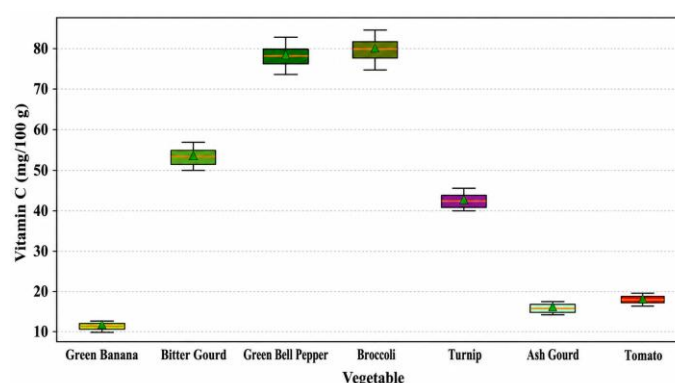


Fig. 9. Distribution of Vitamin C concentration across the seven vegetable groups.

The pooled distributions of the continuous variables are shown in **Fig. 10**. Viewing the overall distributions together with the vegetable-specific profiles is useful because pooled statistics can

conceal the structured group patterns present in the dataset. Moisture, ash, and their replicate SDs occupied very different numerical ranges, while Vitamin C extended across a broad response interval. The machine-learning models were therefore required to learn from a predictor space containing both scale differences and group-specific combinations rather than from one uniformly distributed population.

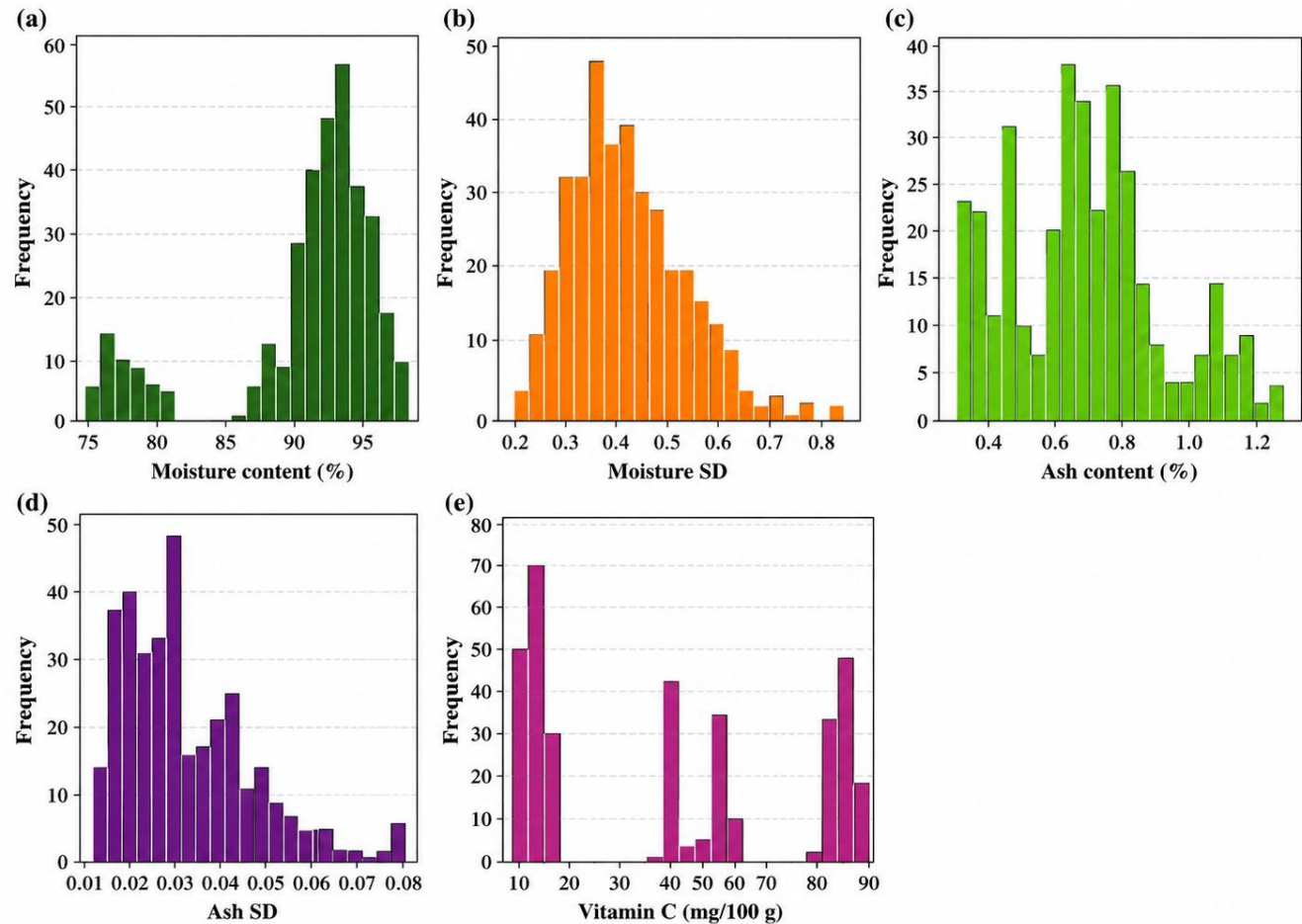


Fig. 10. Overall distributions of the five continuous variables used in the analysis: (a) moisture content, (b) moisture SD, (c) ash content, (d) ash SD, and (e) Vitamin C concentration.

Patterns seen in these numbers are reinforced with the numerical group summaries. The moisture content of the green banana was the lowest with the mean value being $78.20 \pm 1.66\%$, while that of ash gourd was the maximum with the mean value being $95.43 \pm 1.76\%$. Broccoli was different, it had a relatively low moisture content ($89.81 \pm 1.63\%$), and the highest ash concentration of all the 7 vegetables ($1.121 \pm 0.083\%$). Vitamin C produced the widest nutritional contrast, ranging from 10.37 ± 0.32 mg/100 g in green banana to 80.28 ± 2.34 mg/100 g in broccoli. The within-group ranges reveal why the relationship could not be reduced to a simple threshold in either moisture or ash. Broccoli Vitamin C values ranged from 75.54 to 85.33 mg/100 g, while green bell pepper ranged from 74.26 to 83.37 mg/100 g, resulting in

considerable overlap in the target variable. Their ash profiles, however, were markedly different, with group means of approximately 1.121% and 0.469%, respectively. The reverse situation was evident for bitter melon and turnip: their mean ash contents were relatively close at 0.824% and 0.782%, yet their mean Vitamin C concentrations were clearly separated at 52.74 and 42.37 mg/100 g. Similar indirect and multivariate relationships have been reported in food-quality modelling, where combinations of predictors can be informative even when individual variables do not follow the target monotonically [67].

The moisture distributions provide an even clearer example. Several vegetables clustered between approximately 91% and 97% moisture, yet their Vitamin C concentrations occupied very different ranges. Ash melon, green bell pepper and turnip were found to have high moisture content while mean Vitamin C content of all the above was found to be high. They had concentration values of about 14.81, 78.55, and 42.37 mg/100 g, respectively. High moisture. Hence, the observed data sample was not consistently related to either high or low Vitamin C. Taken together, **Figs. 6-10** demonstrate that the prediction problem was set in an overlapping manner. A much more tightly coupled Vitamin C response with physicochemical measurements. It was not just a case of the regression models identifying one high- or low-value threshold. Rather, the algorithms were required to incorporate the correlated patterns of moisture, ash and simulated variability. To separate samples with individual measurements that might seem close together, but have different Vitamin C concentrations differed substantially. This is the main one used for the subsequent analysis, of the reasons why the study was undertaken. Instead of using one single physicochemical measurement as a direct surrogate for the nonlinear multivariable model, the two approaches were compared. Instead, the two approaches were compared, treating any one physicochemical measurement as a direct surrogate for the nonlinear multivariable model. An alternative to Vitamin C.

3.2. Moisture-ash-Vitamin C relationships

Moisture and ash only had a weak correlation for all 350 observations. The slight negative trend suggests that the two measures reflect different aspects of the vegetable matrix and not interchangeable information. Loss of moisture is mass lost during drying while ash is mass lost during combustion. Loss of moisture is mass lost in drying, but ash is mass lost in combustion. represents the residue that remains after combustion which is inorganic. The vegetable-specific clusters in **Fig. 11** further show that the pooled relationship is not described well by a single straight line, which supports the later use of nonlinear tree-based models capable of learning local splits and interactions rather than imposing one global linear rule. Gradient-boosted trees

are specifically designed to build such nonlinear additive relationships from successive tree partitions [68] .

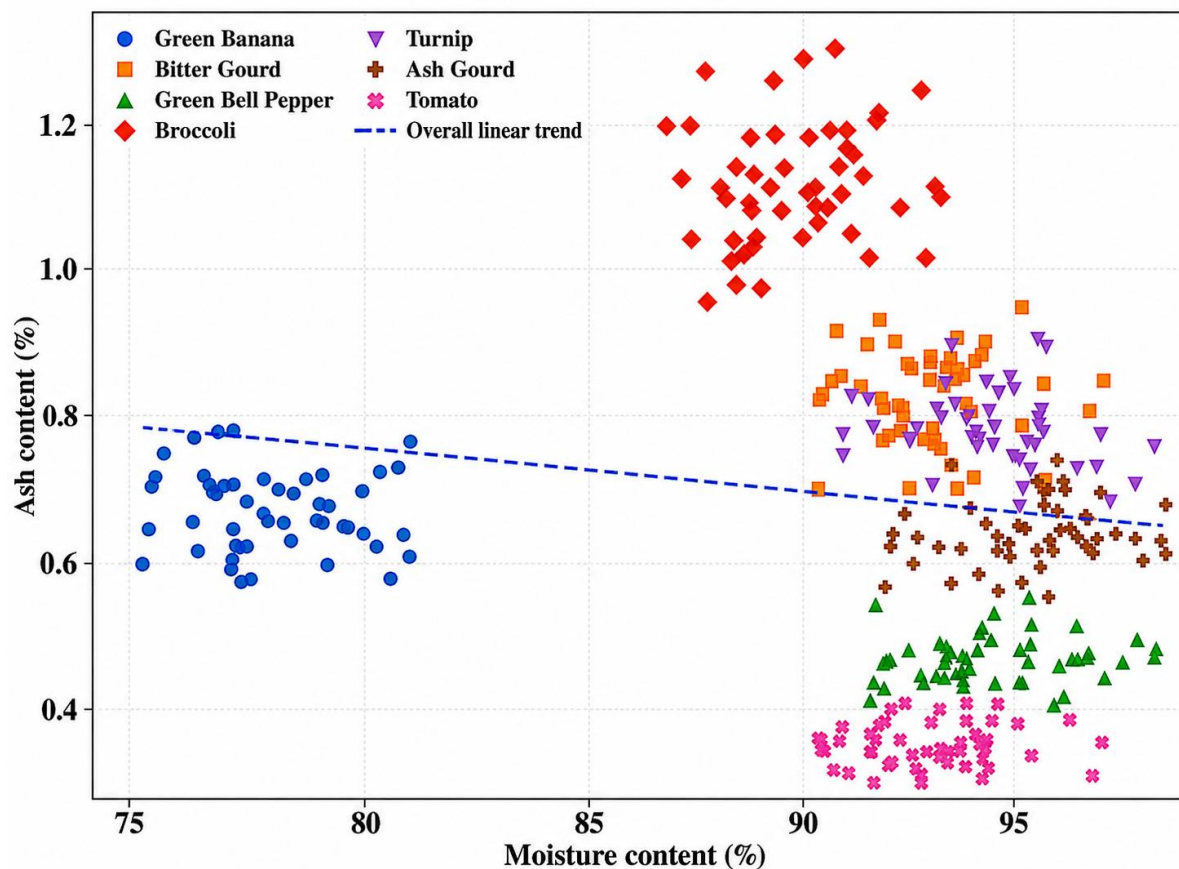


Fig. 11. Relationship between moisture and ash content across the seven vegetable groups; the dashed line represents the overall linear trend.

The two physicochemical measurements were also associated differently with Vitamin C. Moisture showed only a modest positive pooled tendency, whereas ash displayed a clearer positive relationship with the target. Even so, neither scatter pattern was sufficiently tight to support a one-variable calibration of Vitamin C. Instead, the useful predictive information appears to arise from the combined structure of moisture, ash, and their variability descriptors. A similar principle has been observed in recent food-quality prediction studies, where Vitamin C estimation improved when multiple informative measurements were considered jointly rather than interpreted as isolated one-to-one relationships [69] .

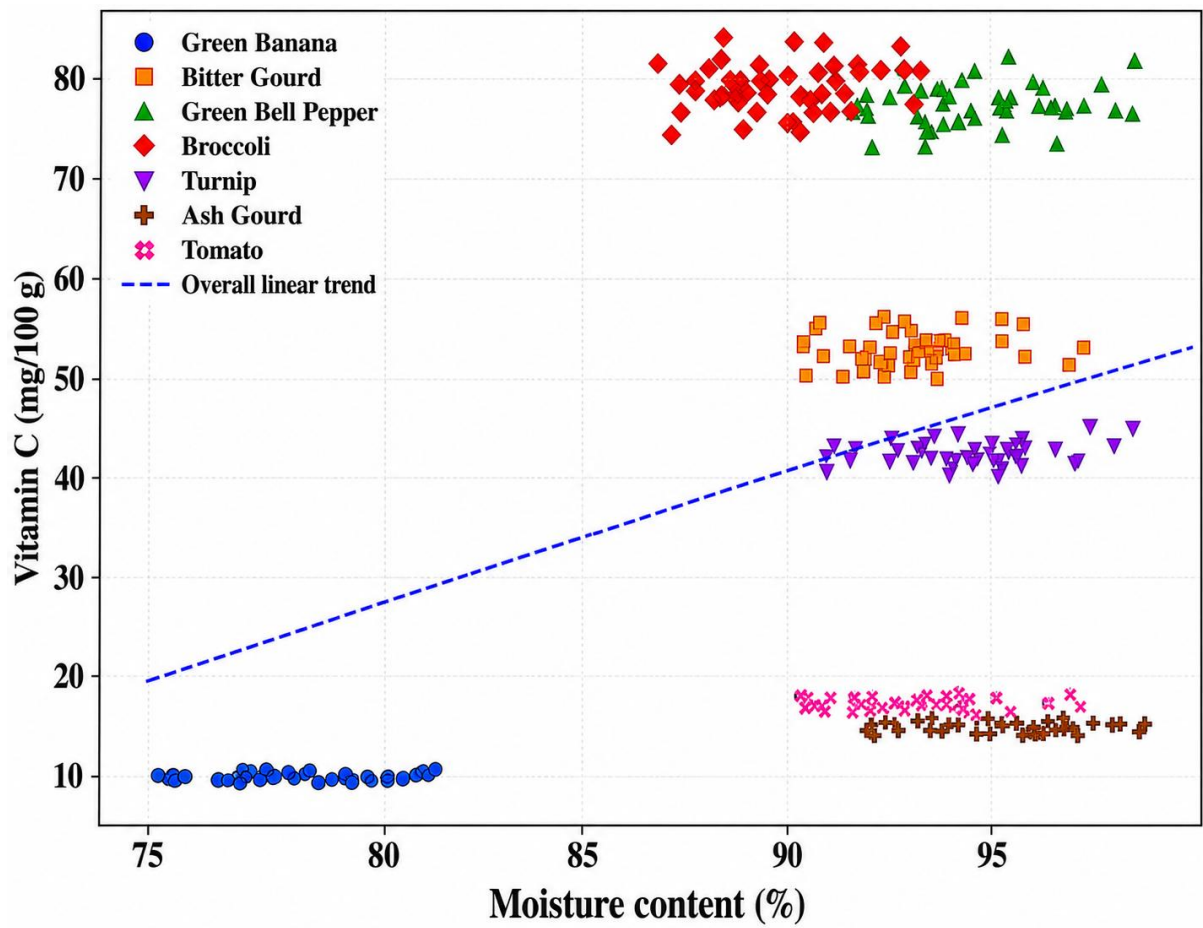


Fig. 12. Relationship between moisture content and Vitamin C concentration across the seven vegetable groups.

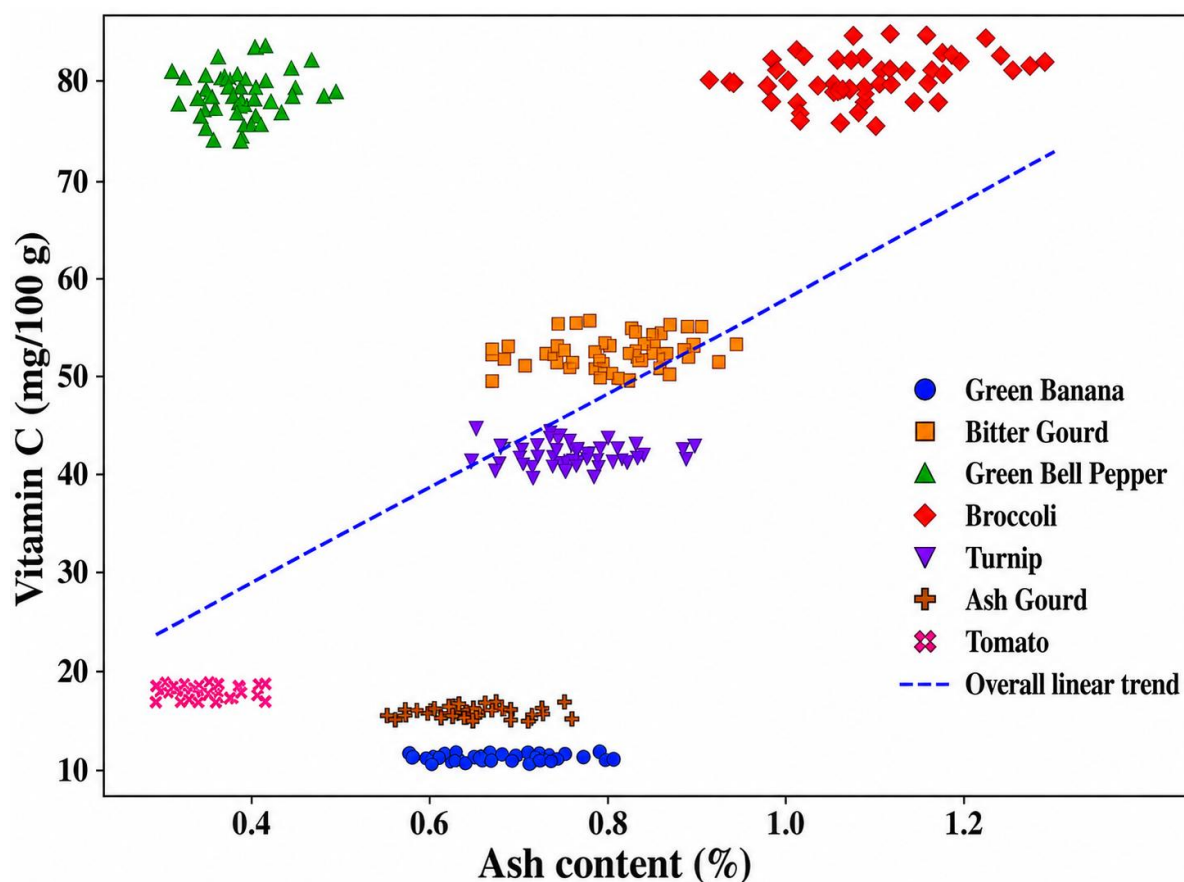


Fig. 13. Relationship between ash content and Vitamin C concentration across the seven vegetable groups.

The Pearson correlation matrix in **Fig. 14** provides a numerical summary of these pairwise relationships. Ash content showed the strongest positive linear association with Vitamin C at $r = 0.44$, followed by ash SD at $r = 0.40$ and moisture content at $r = 0.30$. Moisture SD was almost unrelated to Vitamin C on a linear basis, with $r = -0.07$. Among the predictors themselves, moisture and ash were weakly negatively correlated at $r = -0.14$, while ash and ash SD showed a much stronger positive relationship of $r = 0.75$, indicating that these two descriptors contain partly overlapping information.

Pearson Correlation Matrix of Physicochemical Variables and Vitamin C

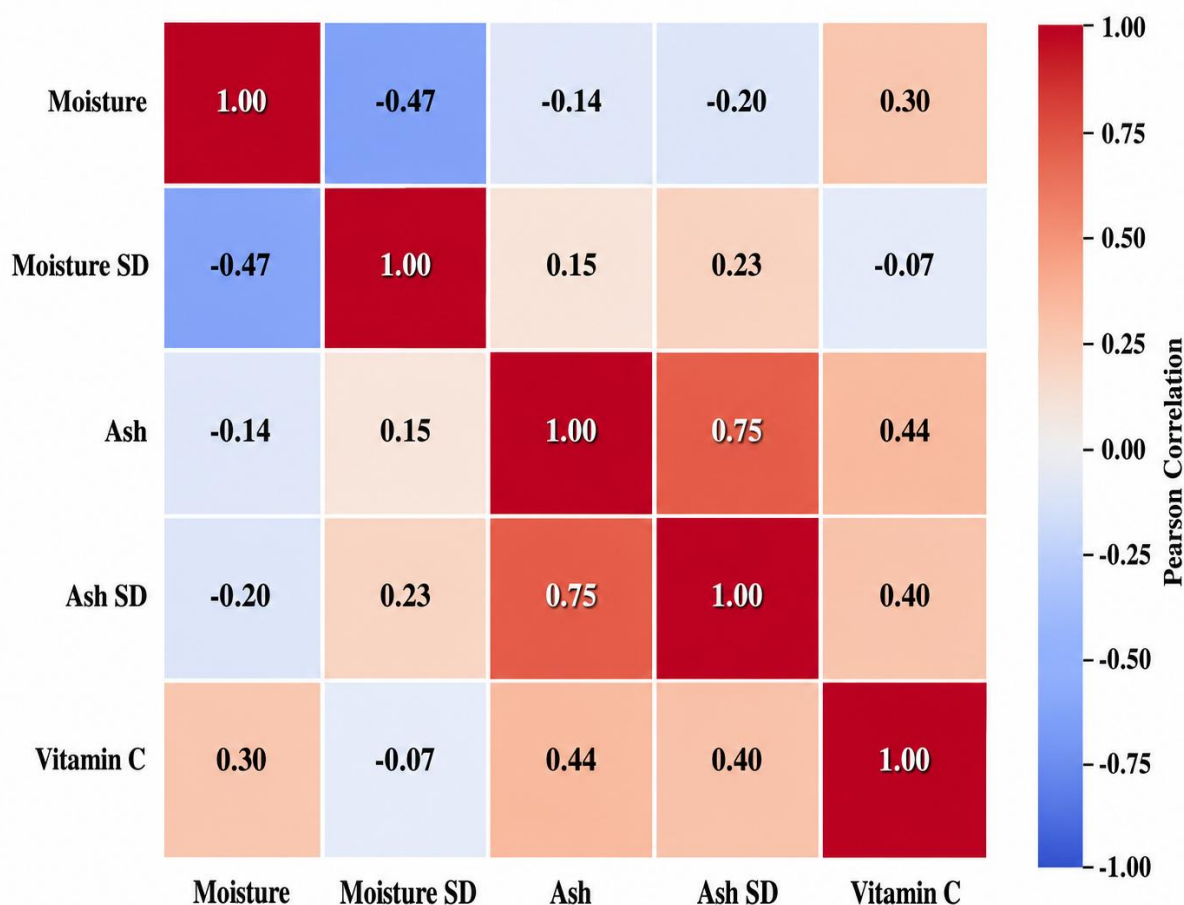


Fig. 14. Pearson correlation matrix for moisture content, moisture SD, ash content, ash SD and Vitamin C concentration.

The other predictor correlations were also fairly weak: moisture and moisture SD $r = -0.47$; moisture and ash SD $r = -0.20$; moisture SD and ash $r = 0.15$; moisture SD and ash SD $r = 0.23$. Combined, these coefficients indicate that none of these predictors is very strong in itself. This is in agreement with the scatter plots; moisture or ash samples may have very different Vitamin C ranges, depending upon what they have in common. Having no predictor with a target correlation greater than 0.50 doesn't contradict that several nonlinear predictors have test R^2 values higher than 0.95. Tree ensembles can include multiple predictors, segment the feature space into multiple regions and model interaction effects that can't be seen from any one pairwise Pearson correlation. This difference explains why tree-based methods can be successful even if individual linear correlation are found to be only fair [70]. The group structure that can be seen in the figures. While it is important to be cautious when interpreting **Figs. 11–13**, it is equally important to be cautious when interpreting **Figs. 11–13** Broccoli is in the high Vitamin C range and green bell pepper is in the high range, but the distribution of ash

is vastly different. It is therefore only in part a biochemical relationship—that is, the presence of pooled ash and Vitamin C is not merely a consequence of the relative position of the seven vegetable groups in the combined data set. Later, the SHAP analysis can be used to know how the fitted HistGradientBoosting model utilizes these patterns but it does not explain biological causality. [71] .

Overall, **Figs. 11–14** demonstrate that there are structured combinations of features that drive the problem, rather than a single dominant linear relationship. Each of these contribute separately to the structure of the sample, and the interaction of these three pattern components is nonlinear, leading to the high predictive power of the overall pattern later on.

3.3. Ten optimized tree-based models make comparative predictions

All 10 optimized regressors were able to explain a significant amount of the variance in Vitamin C, but the errors of the regressions were not exactly the same. This difference can be seen in the plots of actual versus predicted values in Figs. 15a and 15b. Most of the observations were within the range of the 1:1 reference line for the stronger ensemble models, with generally good agreement between the experimentally measured and the predicted concentrations throughout most of the Vitamin C range. Larger departures were however observed for some samples particularly in the low and intermediate concentration areas. This is an observation level deviation that warrants attention because it is possible to have a high overall R^2 and relatively large errors for individual observations. In food quality prediction studies, it is therefore common to use the plots of the predicted versus the reference values together with other measures of error, such as numerical measures, to compare regression models. [84] .

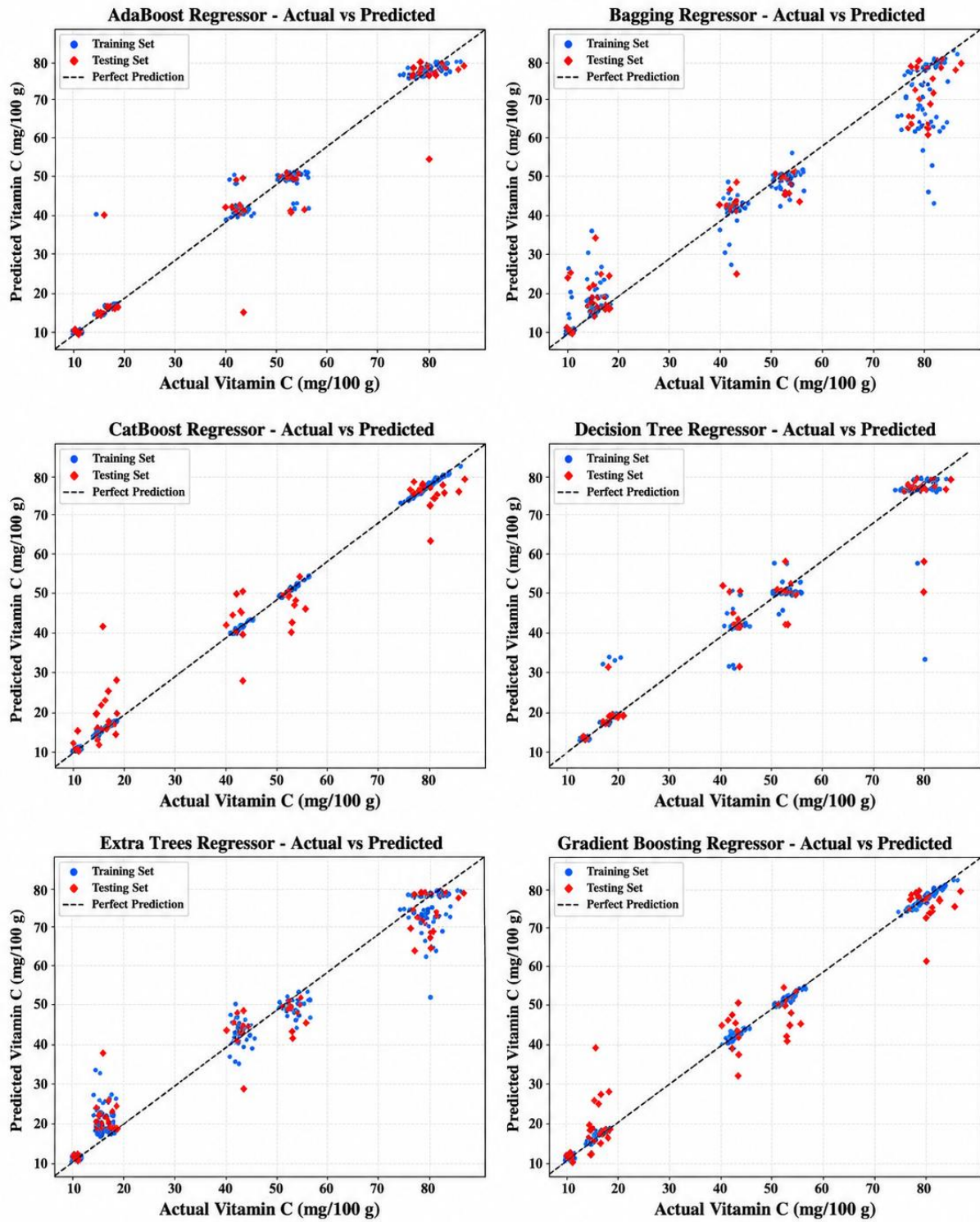


Fig. 15a. Actual versus predicted Vitamin C concentrations for the first six tree-based regression models in the ten-model comparison. Blue circles represent training observations and red diamonds represent held-out test observations.

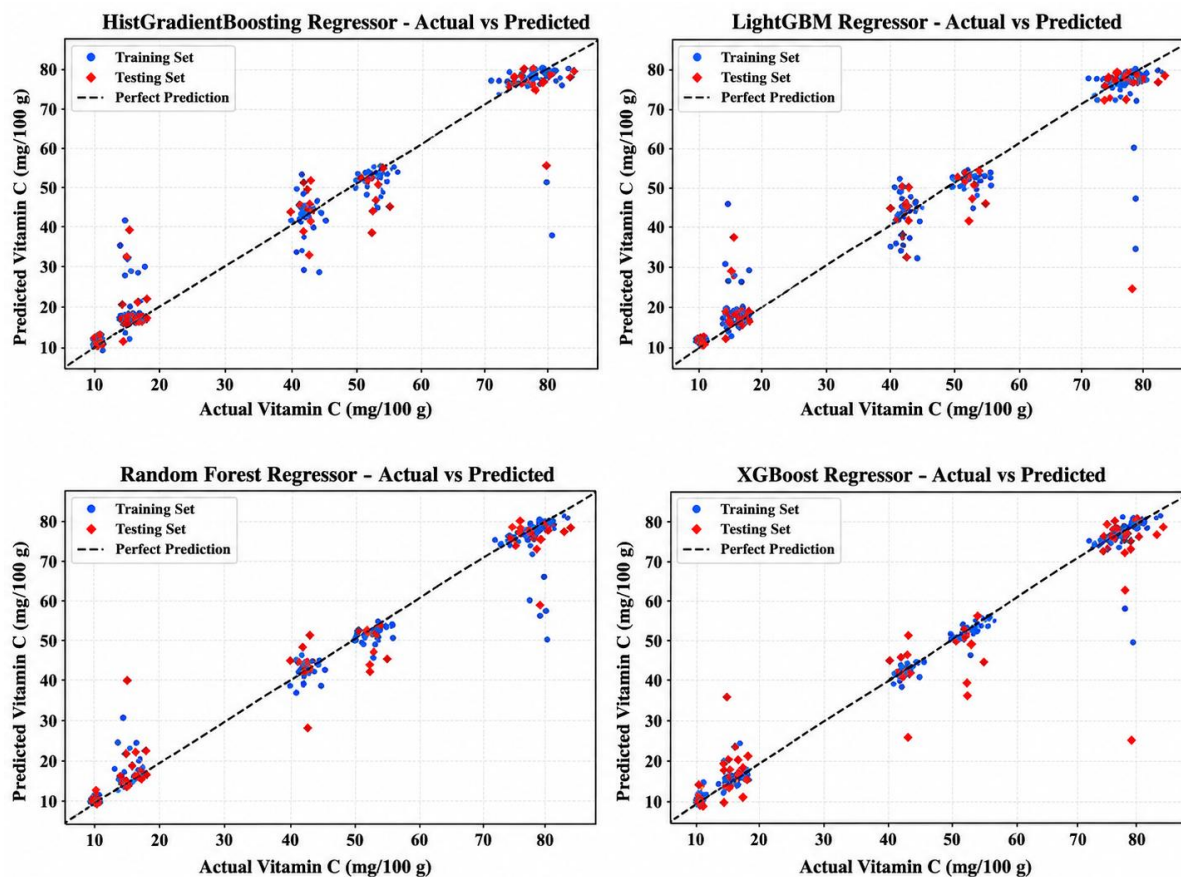


Fig. 15b. Actual versus predicted Vitamin C concentrations for the remaining four tree-based regression models in the ten-model comparison.

The full training and testing results are summarized in **Table 6**. Considering R^2 together with RMSE, MAE, MedAE, and MAPE provides a more complete view of model behaviour than ranking the algorithms by one statistic alone. This is particularly relevant here because several models produced very similar test R^2 values while differing more clearly in their absolute, median, or percentage errors. The use of multiple regression metrics is also common in predictive food and agricultural studies because squared-error and absolute-error measures emphasize different aspects of prediction performance.

Table 6. Comparative training and testing performance of the ten tree-based regression models.

Model	Dataset	R^2	MedAE	MAE	MSE	RMSE	MAPE (%)
HistGradientBoosting	Training	0.9696	1.0715	2.2654	23.0365	4.7996	8.0607
HistGradientBoosting	Testing	0.9524	1.9970	3.4879	36.0339	6.0028	11.9839
AdaBoost	Training	0.9896	0.4725	1.2968	7.8958	2.8099	3.5533
AdaBoost	Testing	0.9489	0.7900	2.8311	38.7174	6.2223	7.7941
Gradient Boosting	Training	0.9996	0.3758	0.4331	0.3111	0.5577	1.5799
Gradient Boosting	Testing	0.9556	2.8838	3.9283	33.6694	5.8025	14.0312
Random Forest	Training	0.9874	0.6309	1.4381	9.5118	3.0841	4.6466
Random Forest	Testing	0.9612	1.5767	3.1358	29.4153	5.4236	10.3718
CatBoost	Training	1.0000	0.0817	0.1113	0.0226	0.1505	0.4261
CatBoost	Testing	0.9566	2.2131	3.6385	32.9124	5.7369	13.1656

Extra Trees	Training	0.9734	2.0205	2.9482	20.1707	4.4912	10.4297
Extra Trees	Testing	0.9539	3.5443	4.3222	34.9371	5.9108	15.1658
XGBoost	Training	0.9919	0.4204	0.9101	6.1292	2.4757	2.9012
XGBoost	Testing	0.9581	2.0436	3.7641	31.7452	5.6343	13.6463
LightGBM	Training	0.9716	1.5720	2.6902	21.4743	4.6340	10.1462
LightGBM	Testing	0.9127	1.8842	3.9590	66.1383	8.1325	12.8971
Bagging	Training	0.9501	1.1365	3.3244	37.8006	6.1482	9.7064
Bagging	Testing	0.9411	3.0672	4.6008	44.5982	6.6782	16.8149
Decision Tree	Training	0.9693	0.7450	2.2206	23.2428	4.8211	6.7849
Decision Tree	Testing	0.9338	1.3586	3.5988	50.1862	7.0842	11.4265

1561 The held-out results show that no single algorithm dominated every metric. Random Forest
1562 achieved the highest test R^2 at 0.9612 and the lowest RMSE at 5.4236 mg/100 g. XGBoost
1563 followed closely with $R^2 = 0.9581$ and RMSE = 5.6343 mg/100 g, while CatBoost and Gradient
1564 Boosting produced test R^2 values of 0.9566 and 0.9556, respectively. Extra Trees and
1565 HistGradientBoosting also remained close, with test R^2 values of 0.9539 and 0.9524. AdaBoost,
1566 Bagging, Decision Tree, and LightGBM produced somewhat lower held-out values of 0.9489,
1567 0.9411, 0.9338, and 0.9127, respectively.

1568 The error metrics add an important layer to this ranking. HistGradientBoosting produced a test
1569 MAE of 3.4879 mg/100 g and a MedAE of 1.9970 mg/100 g, whereas AdaBoost achieved
1570 lower values for both measures, with MAE = 2.8311 mg/100 g and MedAE = 0.7900 mg/100
1571 g, despite having a lower overall test R^2 . Random Forest combined its high R^2 with an MAE of
1572 3.1358 mg/100 g and a MedAE of 1.5767 mg/100 g. At the other end of the comparison,
1573 Bagging produced the largest MAE at 4.6008 mg/100 g, while LightGBM showed the largest
1574 RMSE at 8.1325 mg/100 g. These differences illustrate why model quality cannot be reduced
1575 to a single number; RMSE places greater emphasis on larger errors, whereas MAE and MedAE
1576 describe different aspects of the typical prediction deviation [74].

1577 Training behaviour varied more strongly than test performance. CatBoost fitted the training
1578 data almost exactly, with $R^2 \approx 1.0000$ and RMSE = 0.1505 mg/100 g, while Gradient Boosting
1579 also produced a very tight training fit with $R^2 = 0.9996$. These results demonstrate high fitting
1580 capacity, but exceptionally small training error should not by itself be interpreted as superior
1581 predictive performance. The relevant question is whether that fit persists when the model is
1582 evaluated on data not used during fitting. HistGradientBoosting showed a less extreme training
1583 fit, with $R^2 = 0.9696$ and RMSE = 4.7996 mg/100 g, while still retaining strong held-out and
1584 cross-validation performance. Model-selection research has emphasized precisely this
1585 distinction between fitting performance and evidence of generalization from resampled or
1586 independent data [75].

1587 The multi-metric comparison is visualized further in **Figs. 16a and 16b**. The radar plots were
1588 used to display the relative training and testing behaviour of the algorithms across several

performance criteria simultaneously. Their purpose is not to establish a new ranking, but to make differences in the overall error profile easier to see when models have similar R^2 values [76].

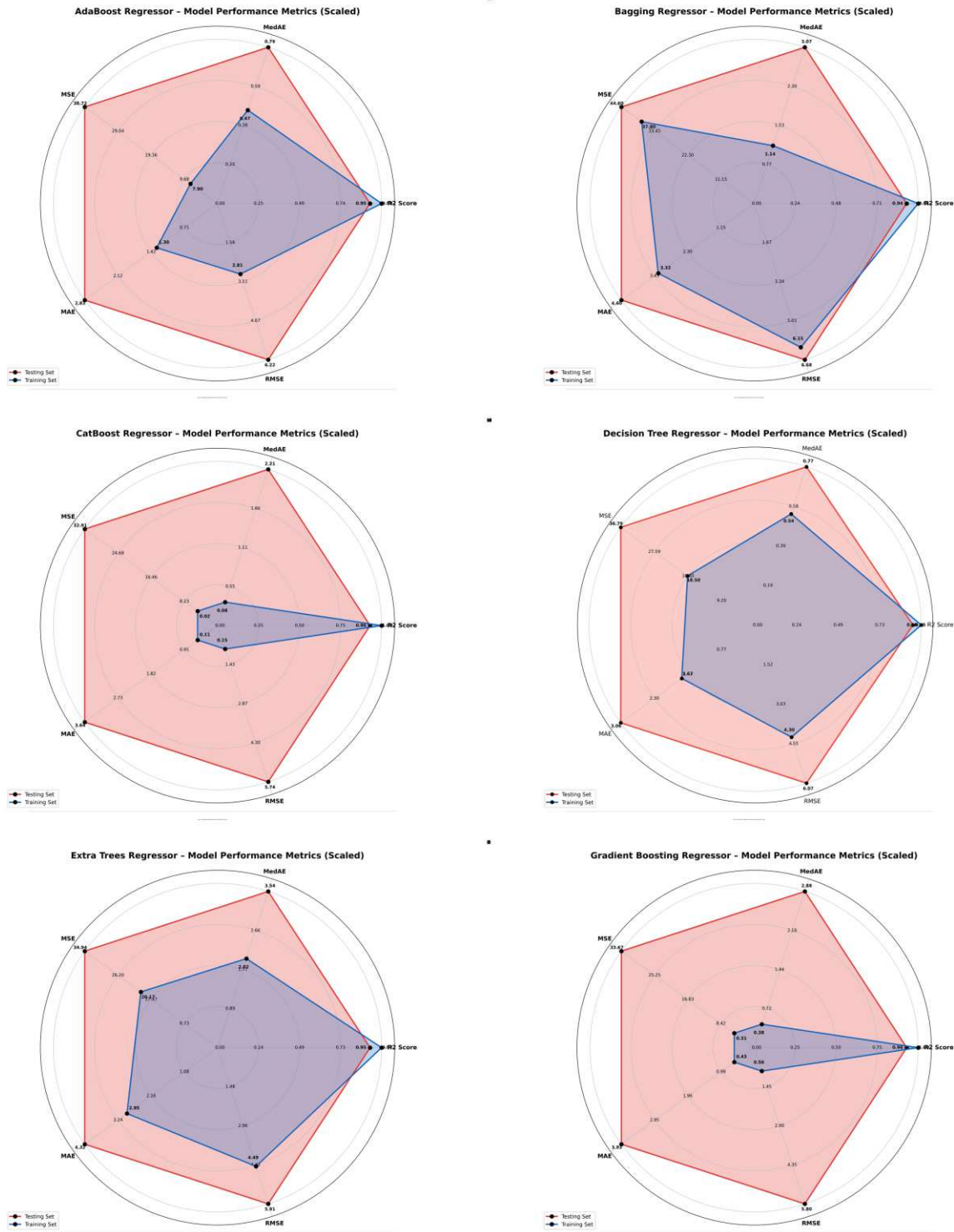


Fig. 16a. Radar comparison of scaled training and testing performance metrics for the first six regression models.

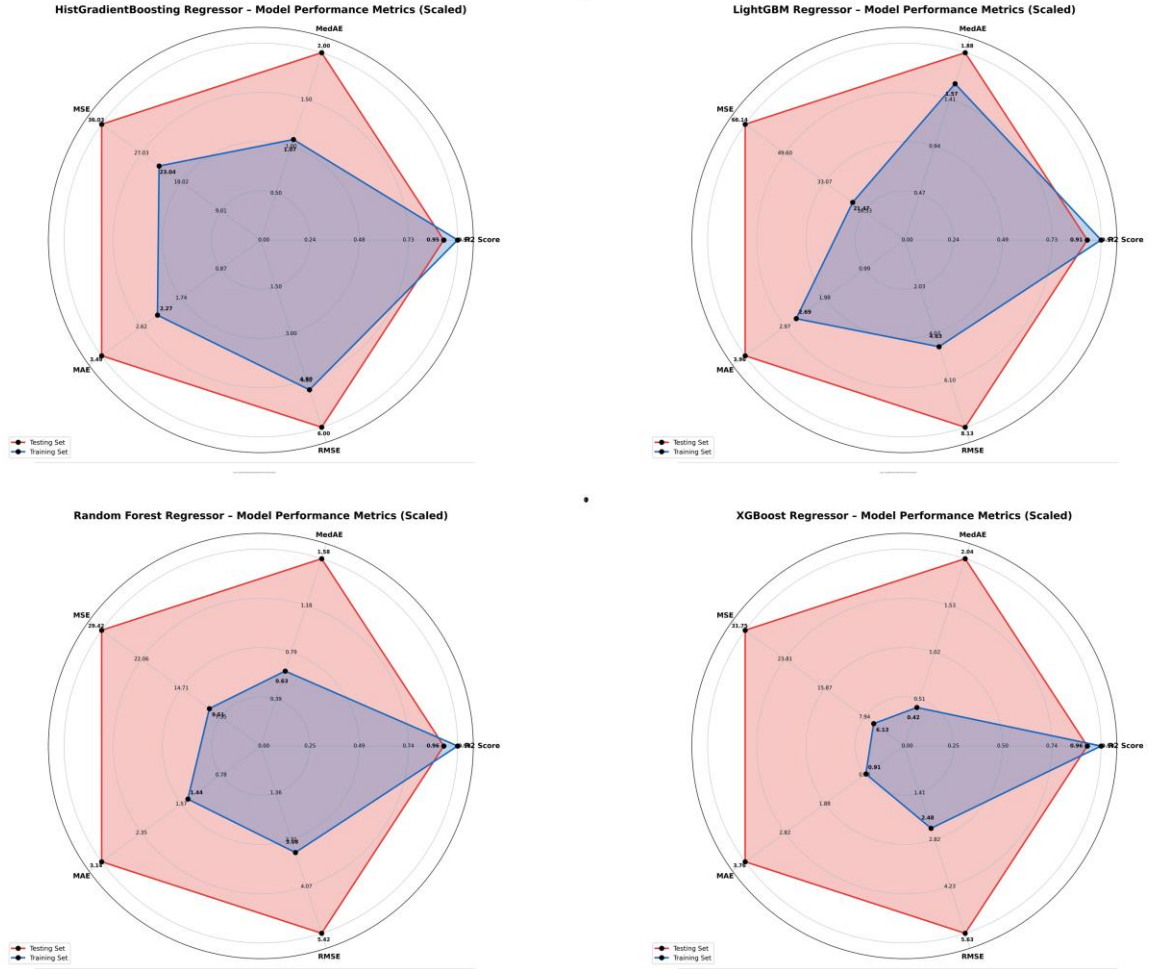


Fig. 16b. Radar comparison of scaled training and testing performance metrics for the remaining four regression models.

The distinction between repeated cross-validation and the final held-out result becomes clearer in **Fig. 17**. Random Forest produced the strongest single test score, but HistGradientBoosting achieved the highest mean cross-validation validation R^2 at 0.9205. Random Forest, by comparison, had a mean CV validation $R^2=0.9094$. This difference determined the final model choice because the ranking procedure had been specified in advance to prioritize repeated CV evidence rather than the most favourable result on one particular test partition. Model-selection literature has shown that performance estimates used for choosing among algorithms can themselves be affected by finite-sample variance, making repeated validation evidence particularly relevant when several models perform closely [77] .

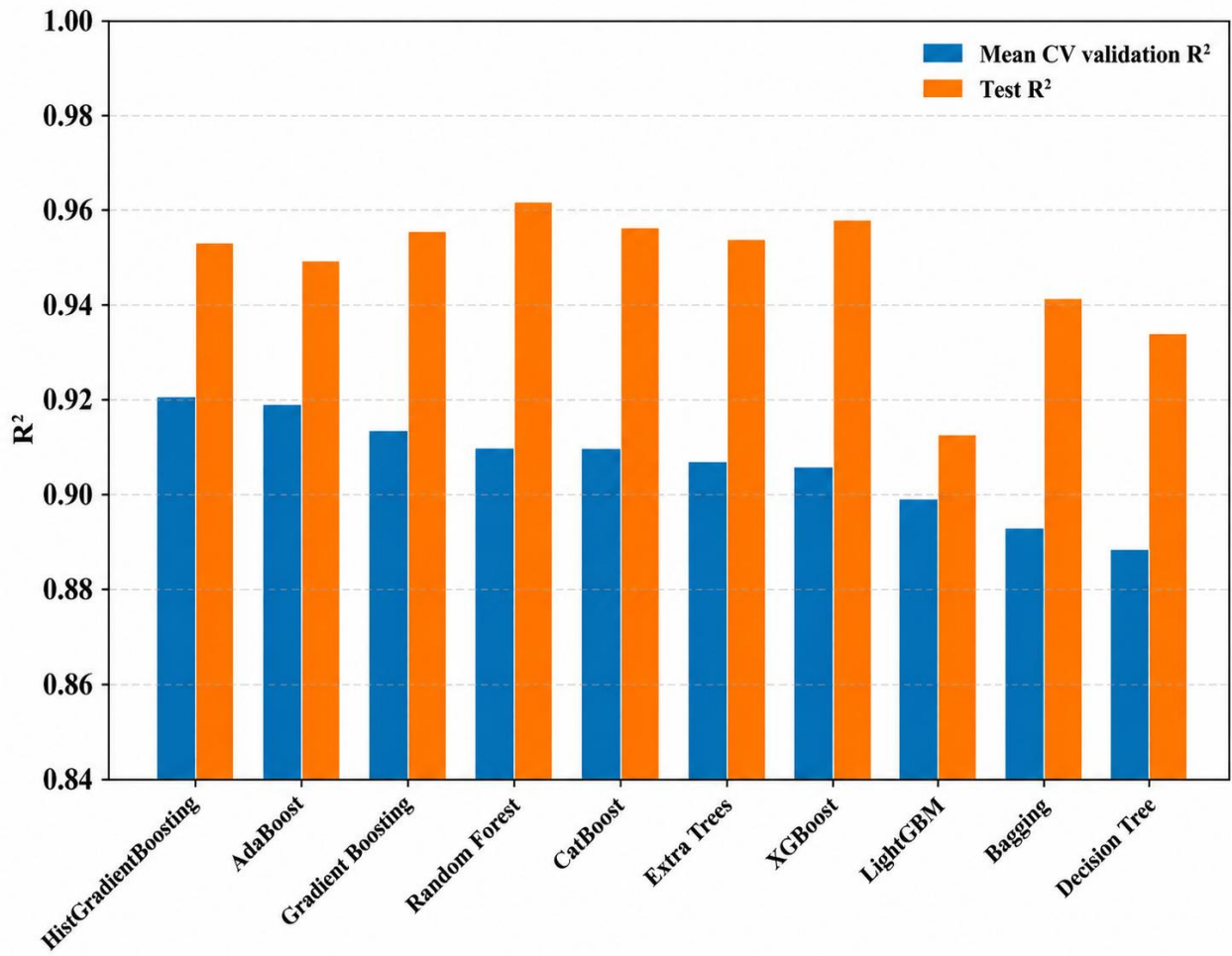


Fig. 17. Mean cross-validation validation R^2 compared with held-out test R^2 for all ten regression models.

HistGradientBoosting was therefore retained as the primary explanatory model despite not producing the numerically highest test R^2 . Its mean CV validation R^2 was 0.92050, and the final held-out evaluation remained strong at $R^2 = 0.95243$, RMSE = 6.0028 mg/100 g, and MAE = 3.4879 mg/100 g. Random Forest achieved a higher test R^2 of 0.96117 and lower RMSE of 5.4236 mg/100 g, but ranked lower under the predefined CV-first criterion. The result should therefore be read as a difference in selection philosophy rather than evidence that Random Forest was an inferior predictor [78]. The test-error profiles are summarized visually in **Fig. 18**, where RMSE, MAE, MedAE, and MAPE are displayed together. This view is useful because models that appear nearly interchangeable by R^2 can behave differently when the magnitude or proportional size of prediction errors is considered. Similar multi-metric reporting is used in contemporary food and agricultural prediction studies to avoid drawing conclusions from a single regression statistic [79].

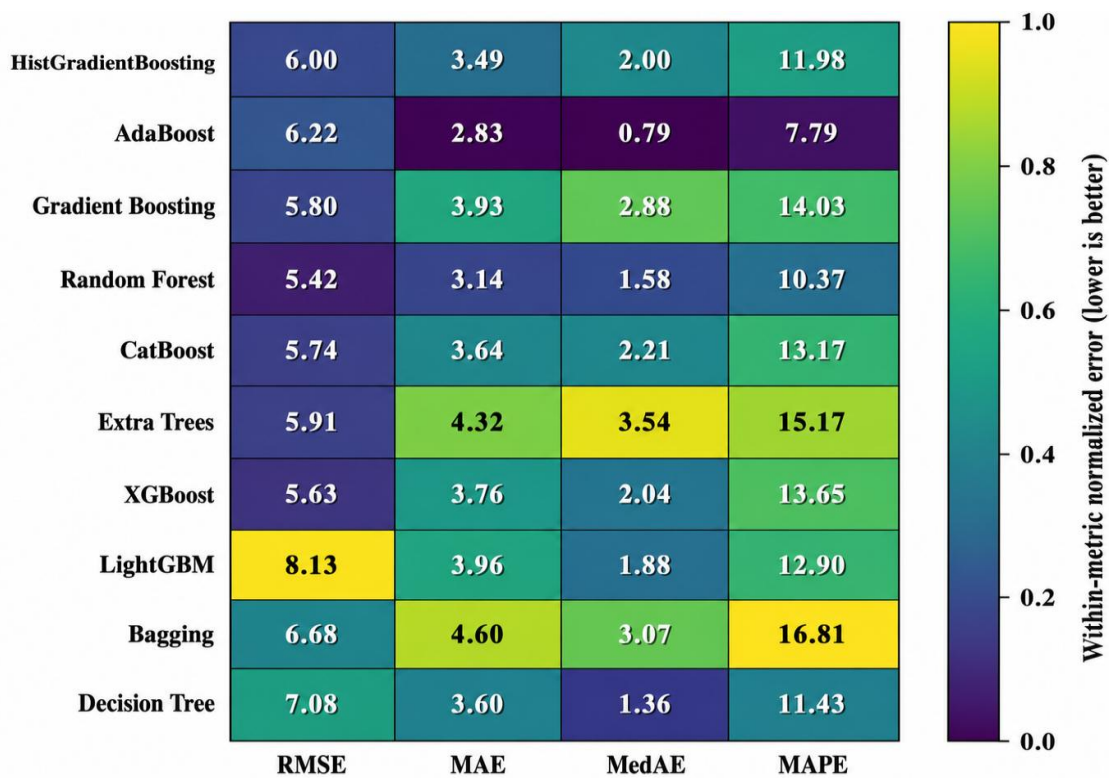


Fig. 18. Test-error heatmap showing RMSE, MAE, MedAE, and MAPE for each model. Cell colours are normalized within each metric, while the printed values represent the original error measurements.

Overall, most of the optimized ensembles performed strongly within the studied dataset, but their relative ordering depended on whether emphasis was placed on a single held-out score, absolute prediction errors, proportional errors, or repeated cross-validation. HistGradientBoosting provided the strongest result under the prespecified CV-first framework, while Random Forest provided the best single held-out R^2 and RMSE. Reporting both outcomes avoids presenting model selection as more clear-cut than the evidence supports [80].

Finally, the scope of these results remains defined by the sampling design. Every vegetable represented in the 70-observation test set was also represented during model development. The high test R^2 values therefore demonstrate strong prediction for held-out observations within the seven-vegetable domain studied here; they do not establish performance for a vegetable species that was completely absent from training. This distinction is important when interpreting the proposed model as a screening framework and supports the need for future external or leave-one-vegetable-out validation before broader species-level generalization is claimed [81].

3.4. Model ranking, stability, and the meaning of “best”

The final ranking was designed to reflect repeated validation performance rather than the outcome of a single favourable test split. Mean cross-validation R^2 was therefore used as the primary ranking criterion, while fold-to-fold variation and the absolute training–validation gap provided additional information on stability. The held-out test metrics were retained as supporting evidence. This distinction is important in comparative machine learning because model-selection criteria estimated from finite samples contain their own variability; selecting a model only because it performs best on one particular evaluation split can give an overly optimistic impression of its relative advantage. Cawley and Talbot similarly emphasize that the variance of a model-selection criterion can be as important as its average performance when competing algorithms are close.

Table 7. Ranking of final CV and held out supporting metrics.

Rank	Model	CV val R^2	CV R^2 SD	CV gap	Test R^2	RMSE	MAE	MedAE	MAPE (%)
1	HistGradientBoosting	0.9205	0.0463	0.0475	0.9524	6.0028	3.4879	1.9970	11.9839
2	AdaBoost	0.9193	0.0766	0.0666	0.9489	6.2223	2.8311	0.7900	7.7941
3	Gradient Boosting	0.9135	0.0413	0.0862	0.9556	5.8025	3.9283	2.8838	14.0312
4	Random Forest	0.9099	0.0470	0.0768	0.9612	5.4236	3.1358	1.5767	10.3718
5	CatBoost	0.9099	0.0406	0.0901	0.9566	5.7369	3.6385	2.2131	13.1656
6	Extra Trees	0.9074	0.0365	0.0698	0.9539	5.9108	4.3222	3.5443	15.1658
7	XGBoost	0.9065	0.0346	0.0857	0.9581	5.6343	3.7641	2.0436	13.6463
8	LightGBM	0.8992	0.0355	0.0683	0.9127	8.1325	3.9590	1.8842	12.8971
9	Bagging	0.8928	0.0433	0.0692	0.9411	6.6782	4.6008	3.0672	16.8149
10	Decision Tree	0.8887	0.0563	0.0824	0.9338	7.0842	3.5988	1.3586	11.4265

HistGradientBoosting was ranked first due to its high mean value of CV validation R^2 (0.9205), but a relatively small training–validation gap (0.0475) **Table 7**. AdaBoost was the second most successful with a mean CV R^2 of 0.9193, but a higher fold-to-fold variability of 0.0766. Random Forest, CatBoost, Extra Trees and XGBoost were the next most powerful models, with Gradient Boosting coming third. LightGBM, Bagging and Decision Tree took the remaining three spots. The ranking is thus not by the end of the day, but according to the prespecified validation strategy; it does not reflect the ordering of the algorithms by some other metric that was more favorable at the time the algorithms were run. This difference is particularly apparent when comparing to HistGradientBoosting and Random Forest. Random Forest provided the best single hold-out performance with test $R^2 = 0.9612$ and RMSE = 5.4236 (mg/100g), while HistGradientBoosting had test $R^2 = 0.9524$ and RMSE = 6.0028 (mg/100g). If the models were ranked according to test R^2 , then the Random Forest model would be selected. However, repeated cross validation results showed better results for Hgb, with an average R^2 validation of about 0.011 higher than Random Forest. The final result is thus based on the study's focus on multiple repetitions of performance rather than a small induced

advantage in one held-out partition. Consistent with methodological studies which have demonstrated that cross validation estimates are not immune from sampling variability and must be used with a critical eye on the stability in addition to the point estimate of the estimate. These other error metrics confirm the same thing: none of the models performed best for all of the criteria. Although AdaBoost was in second position, it obtained the smallest test MAE of 2.8311 mg/100 g and the smallest MedAE of 0.7900 mg/100 g. The last ranking one was Decision Tree, with a relatively low MedAE of 1.3586 mg/100 g for the CV-first procedure. The test R^2 for XGBoost and CatBoost was also higher than that of HistGradientBoosting, but it didn't outperform it when it comes to the key cross-validation parameter. The differences demonstrate that model performance varies depending on the aspect of prediction that is being emphasized: total variance in prediction, sensitivity to larger errors, mean absolute error, or stability of the development sets. [82] .

Hence, 4 cannot be said to be universally better than the other nine algorithms. A better reading of this is that it produced the most balanced result in the prespecified selection framework: it performed best on average over the cross-validated test sets, was well controlled over its training–test split, and yielded high test accuracy on the untouched test sets. The random forest model was also a strong competitor and outperformed other models numerically on the sole held-out R^2 and RMSE measures, and it performed best numerically on the absolute error measures among the models, whereas the AdaBoost model performed best on these measures. This difference reinforces, not undermines, the model comparison. It shows that the final model was not selected simply because one performance number happened to be the largest. Instead, the decision was based on a hierarchy of evidence defined before final interpretation, with repeated validation used for ranking and the independent test set used to confirm that the selected model retained strong predictive performance. Such separation between model selection and final performance assessment is important because optimizing and evaluating models on the same finite-sample evidence can introduce selection bias [83] .

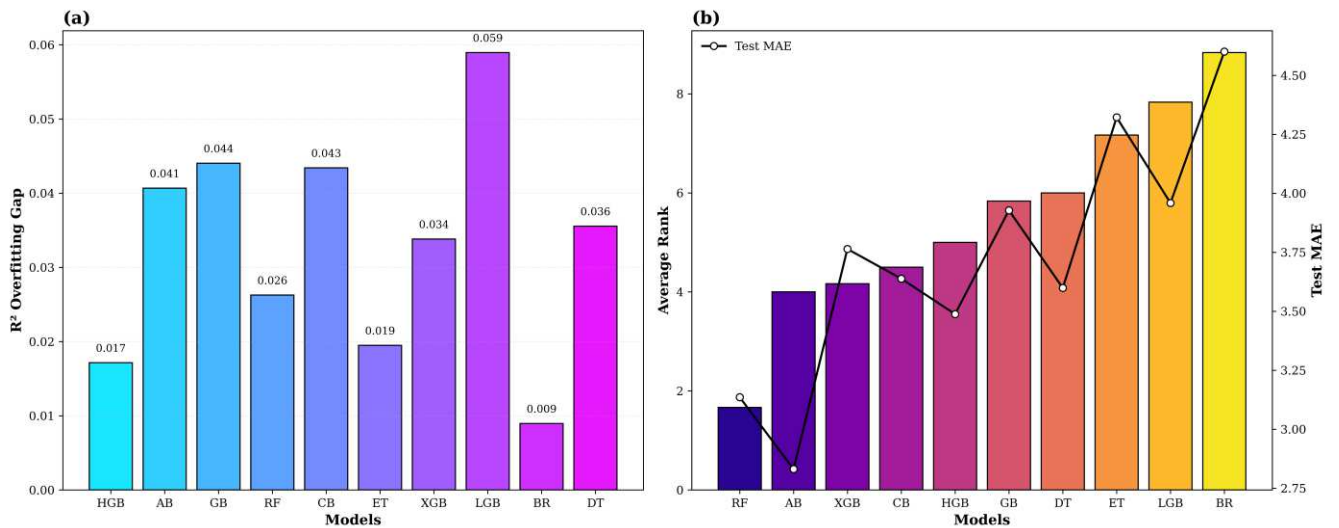


Fig. 19. Comparative model stability and ranking: train-test/CV performance gap and average-rank/test-MAE view of the ten regressors.

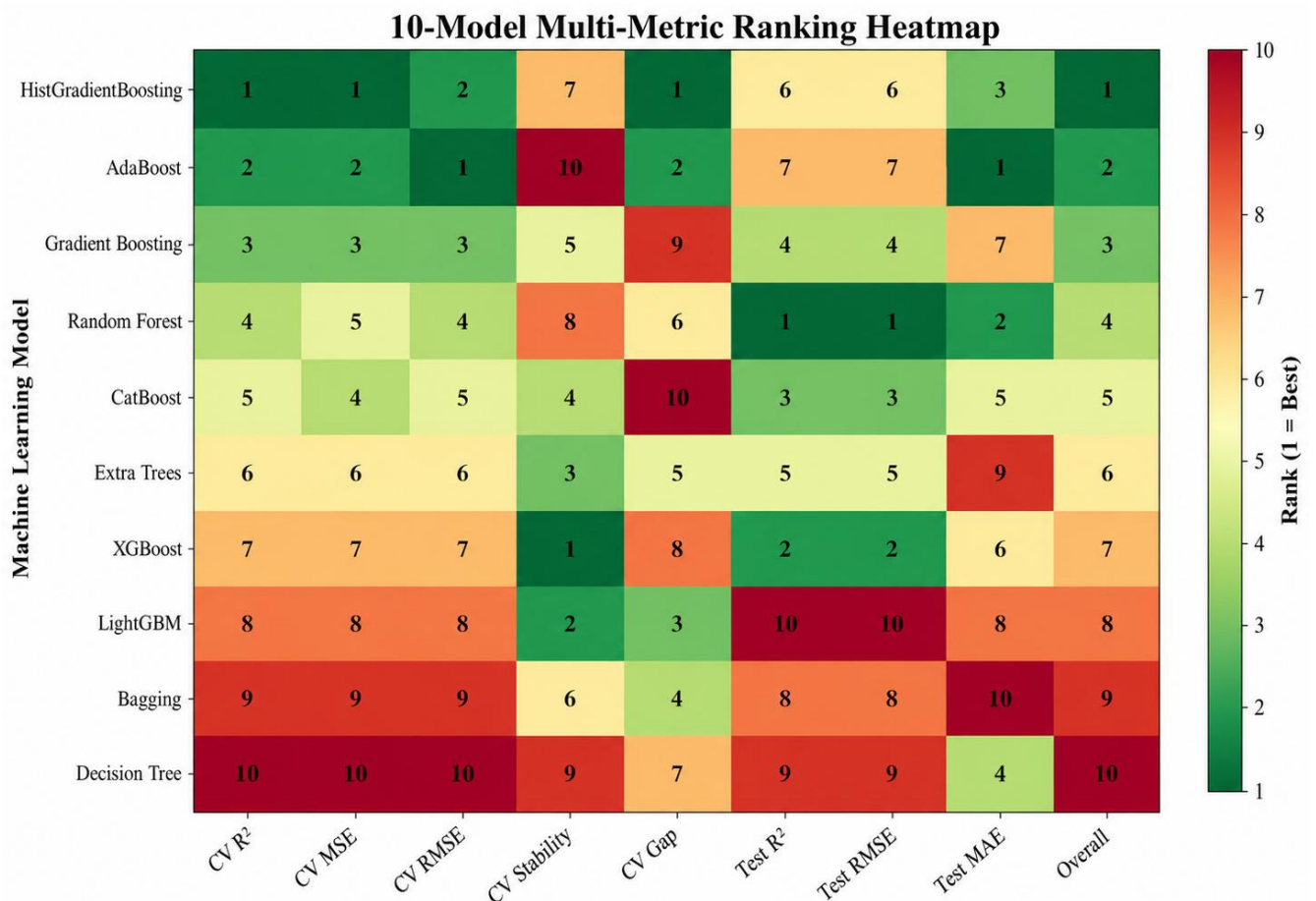


Fig. 20. Multi-metric ranking heatmap for the ten regression models.

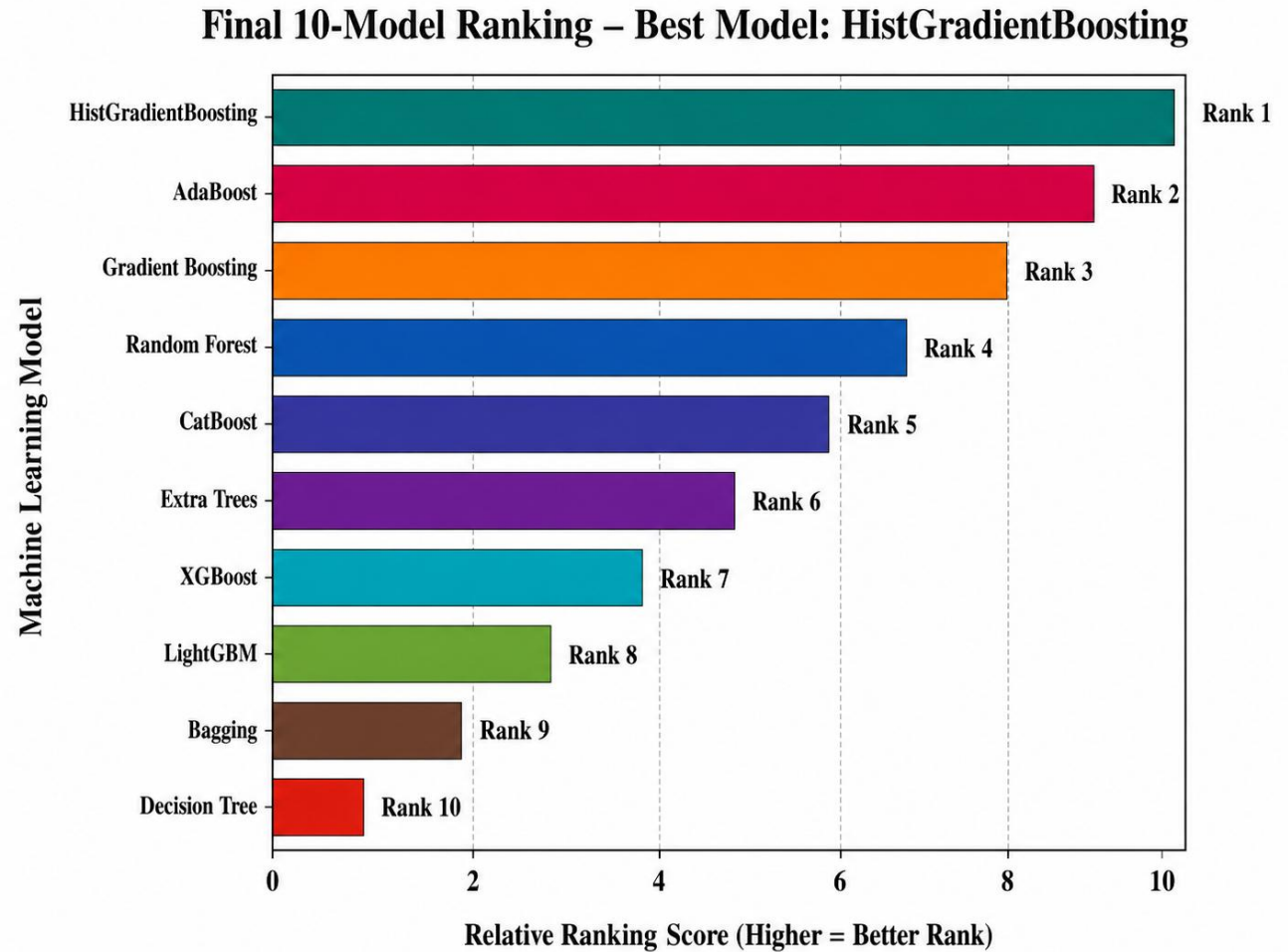


Fig. 21. Final model-ranking summary generated from the model-comparison workflow.

The cross-validation results provided a clearer basis for the final model ranking than the held-out test scores considered alone. HistGradientBoosting achieved the highest mean CV validation R^2 at 0.92050, followed closely by AdaBoost at 0.91926 and Gradient Boosting at 0.91351. Random Forest, although it produced the strongest final test R^2 , reached a lower mean CV validation R^2 of 0.90994 and therefore ranked fourth under the predefined CV-first rule. CatBoost followed almost identically at 0.90993, with Extra Trees at 0.90738, XGBoost at 0.90648, LightGBM at 0.89923, Bagging at 0.89275, and Decision Tree at 0.88874. The fold-to-fold variation added another dimension to this comparison. HistGradientBoosting showed a CV R^2 SD of 0.04632, whereas XGBoost had the smallest listed variability at approximately 0.0346. AdaBoost, despite its strong mean validation score, showed the largest fold-to-fold SD at 0.07656. The absolute training–validation gaps also varied among the models, ranging from

approximately 0.04755 for HistGradientBoosting to 0.0901 for CatBoost. These differences are relevant because cross-validation estimates themselves are subject to sampling variability; methodological studies have shown that model selection can become unstable when the selection criterion has high variance, even when its average performance appears favourable [84].

The resulting ranking should therefore not be interpreted as evidence that HistGradientBoosting was numerically superior under every performance measure. Rather, it emerged as the strongest overall choice under the prespecified hierarchy of evidence: mean CV validation R^2 was considered first, followed by fold-to-fold variability and the absolute training-validation gap. The held-out metrics were then used as supporting evidence rather than as the basis for retrospectively changing the ranking. This distinction is consistent with standard model-selection practice, where cross-validation is performed within the development data while a separate test set is retained for final evaluation. This consideration is especially important for the intended screening application. A future sample will arrive with moisture- and ash-related measurements, but its true Vitamin C concentration will not be known at the point of prediction. Model choice therefore needs to rely on evidence that is reasonably consistent across repeated development partitions rather than on whichever algorithm happens to perform best on one particular 70-observation hold-out set. Cross-validation cannot guarantee performance outside the population represented during development, but it provides a more defensible basis for selecting among closely competing models than repeatedly consulting the final test results [85].

Accordingly, the CV-first analysis in **Table 7** supports HistGradientBoosting as the final explanatory model while still recognizing the strong held-out performance of Random Forest and several other ensembles. The selected model should therefore be understood as the most balanced choice under the predefined validation framework, not as an algorithm that dominates every possible accuracy or stability criterion.

3.5. Feature importance across the ten algorithms

A broadly consistent feature pattern emerged across the ten optimized regressors. Ash content was repeatedly assigned the largest share of normalized importance, moisture generally ranked second, and the two replicate-SD descriptors contributed much less. The recurrence of this ordering across different tree-based learners is more informative than the result of any one model considered in isolation, because it suggests that the prominence of ash was not simply an artefact of a single fitted estimator. At the same time, feature-importance values are model

dependent, and methodological studies have shown that different importance definitions can behave differently and may be affected by the structure of the fitted model and the predictors themselves. Exact percentages should therefore be interpreted within each model rather than treated as directly interchangeable across algorithms [86].

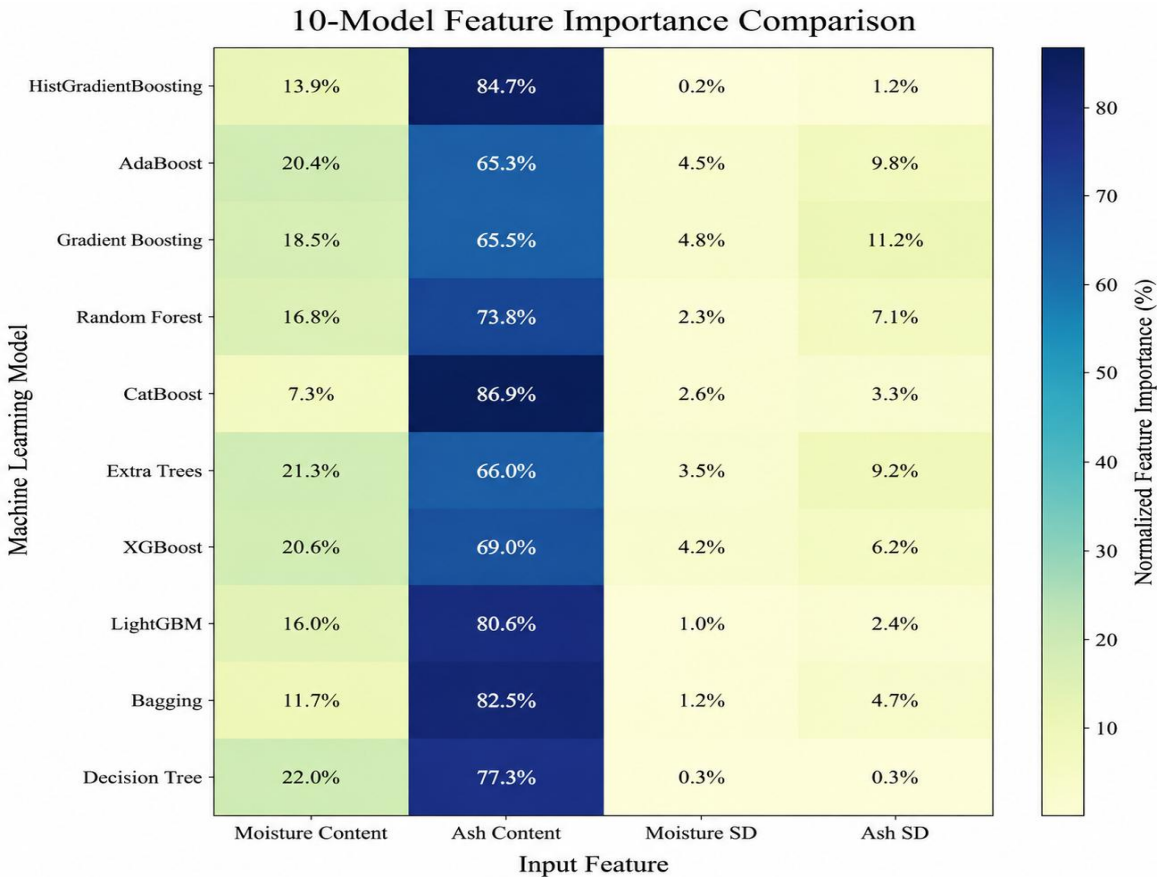


Fig. 22. Within-model normalized feature-importance heatmap across the ten machine-learning models.

The pattern in **Fig. 22** was strikingly consistent. Ash content accounted for approximately 84.7% of the normalized importance in HistGradientBoosting, 73.8% in Random Forest, 65.5% in Gradient Boosting, 66.0% in Extra Trees, 77.3% in Decision Tree, 82.5% in Bagging, 65.3% in AdaBoost, 69.0% in XGBoost, 80.6% in LightGBM, and 86.7% in CatBoost. Moisture content was usually the second-ranked predictor, ranging from approximately 7.3% in CatBoost to 22.0% in Decision Tree, while moisture SD and ash SD generally occupied smaller shares.

The individual importance plots were retained because the heatmap is intended to summarize a recurring pattern rather than imply that all ten algorithms measure importance on a common absolute scale. Tree-based importance can be constructed from different quantities, including

improvements in split criteria, changes in predictive performance after perturbation, or attribution-based explanations. These approaches answer related but not identical questions, and previous work has shown that their magnitudes and even rankings can differ under some data conditions.

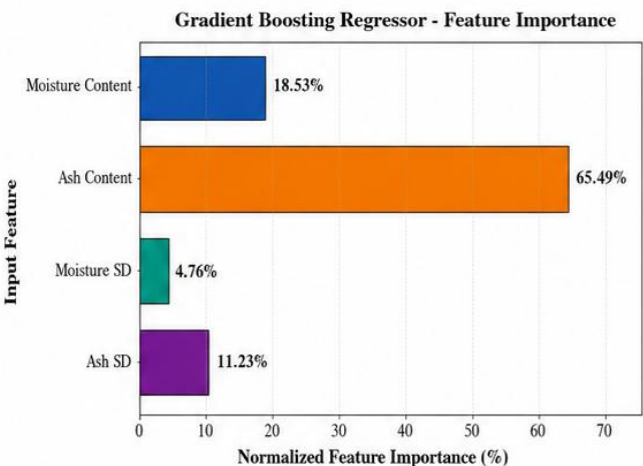
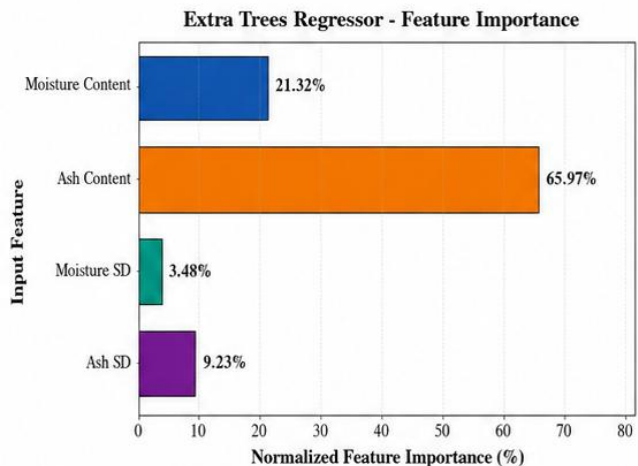
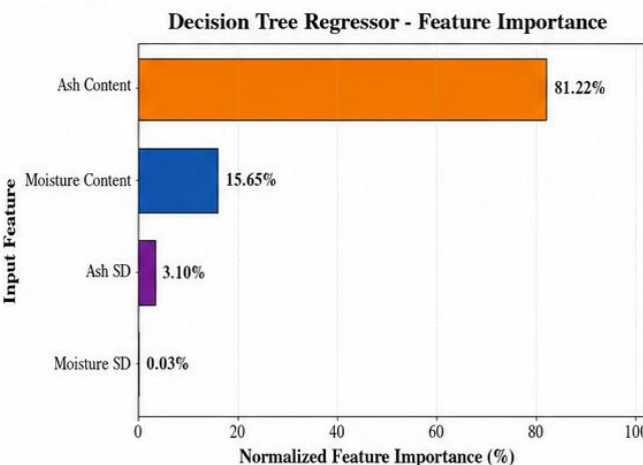
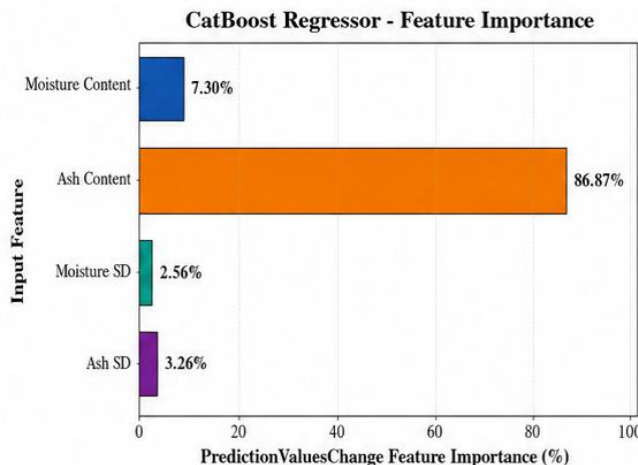
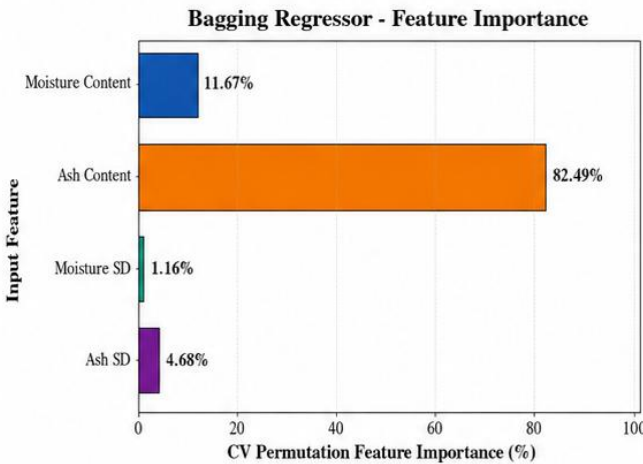
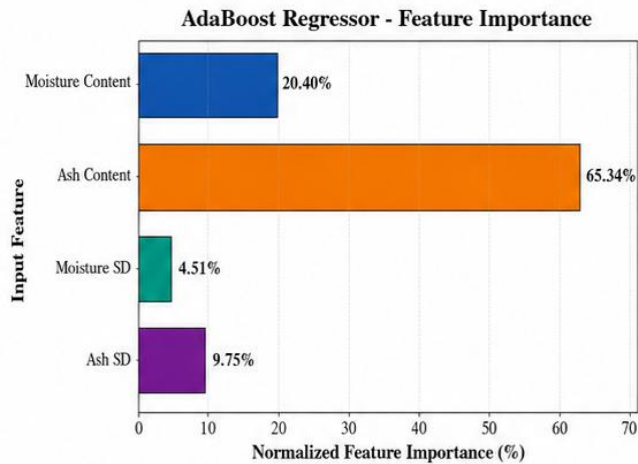


Fig. 23a. Individual normalized feature-importance profiles for the first six regression models.

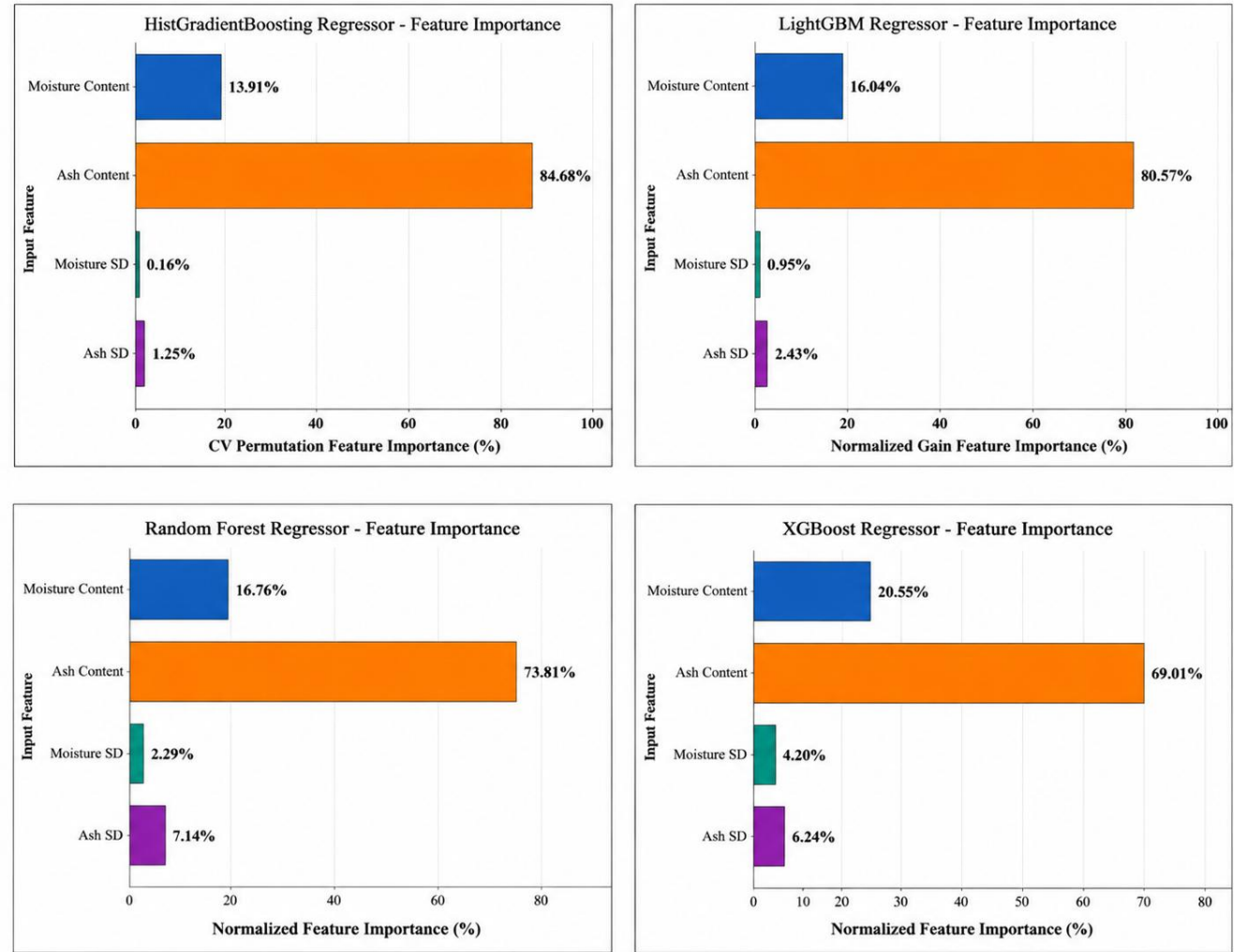


Fig. 23b. Individual normalized feature-importance profiles for the remaining four regression models.

For this reason, the main result in **Figs. 22–23** is the repeated ordering of the predictors rather than numerical equality of their importance percentages. The fact that ash remained the leading feature across all ten fitted models provides stronger evidence of a recurring predictive signal than would a single importance plot from the final model alone. Nevertheless, this should still be interpreted as a property of the fitted prediction models, not as evidence that ash has a direct causal or biochemical effect on Vitamin C. Moisture SD showed the opposite pattern. Its contribution remained small across most of the cross-model displays, and in the HistGradientBoosting normalized importance profile it accounted for only about 0.2%. The later SHAP analysis similarly assigned moisture SD less than 1% of the total mean absolute contribution. Agreement between these complementary analyses is useful because moisture SD

was included prospectively as a descriptor of replicate variability rather than on the assumption that every available laboratory variable would necessarily become an important predictor. The cross-model analysis therefore provided an initial robustness check for the interpretation developed later with SHAP. SHAP methods are designed to explain how fitted tree models distribute individual predictions among their input features and can be aggregated to reveal global model structure. In the present study, the cross-model importance plots first established the recurring ash-dominant pattern, while the subsequent HistGradientBoosting SHAP and ablation analyses were used to examine that pattern in greater detail and test whether the apparent importance translated into incremental predictive value [87] .

3.6. Explainable ML interpretation of HistGradientBoosting

SHAP provided the clearest view of how the selected HistGradientBoosting model distributed predictive information among the four physicochemical inputs. Unlike a simple correlation coefficient, which summarizes one pairwise linear relationship at a time, SHAP explains predictions from the fitted nonlinear model by assigning feature-specific contributions relative to a baseline prediction. This makes it particularly suitable for tree ensembles, where feature effects can vary across different regions of the predictor space and can depend on interactions with other variables [88] .

For the selected model, ash content was clearly the dominant contributor. Its mean absolute SHAP value was 18.0263, corresponding to 72.67% of the total normalized contribution. Moisture content ranked second with a mean absolute SHAP value of 4.6912 and an importance of 18.91%. Ash SD contributed 7.53%, while moisture SD accounted for only 0.89%. The resulting ranking is summarized in **Table 8**.

Table 8. Global SHAP feature importance for the best-ranked HistGradientBoosting model.

Rank	Feature	Mean SHAP	Importance (%)
1	Ash Content	18.0263	72.6707
2	Moisture Content	4.6912	18.9121
3	Ash SD	1.8679	7.5303
4	Moisture SD	0.2200	0.8869

This ordering was broadly consistent with the earlier cross-model importance analysis, where ash repeatedly emerged as the leading predictor, and it was also directionally compatible with the Pearson analysis, in which ash showed the strongest linear association with Vitamin C. The two analyses should nevertheless not be treated as equivalent. Pearson correlation describes the strength of a global linear association between two variables, whereas SHAP describes how an already fitted nonlinear prediction model uses each feature when generating its outputs.

TreeSHAP was developed specifically to retain local explanation while allowing those local contributions to be aggregated into a global description of tree-model behaviour.

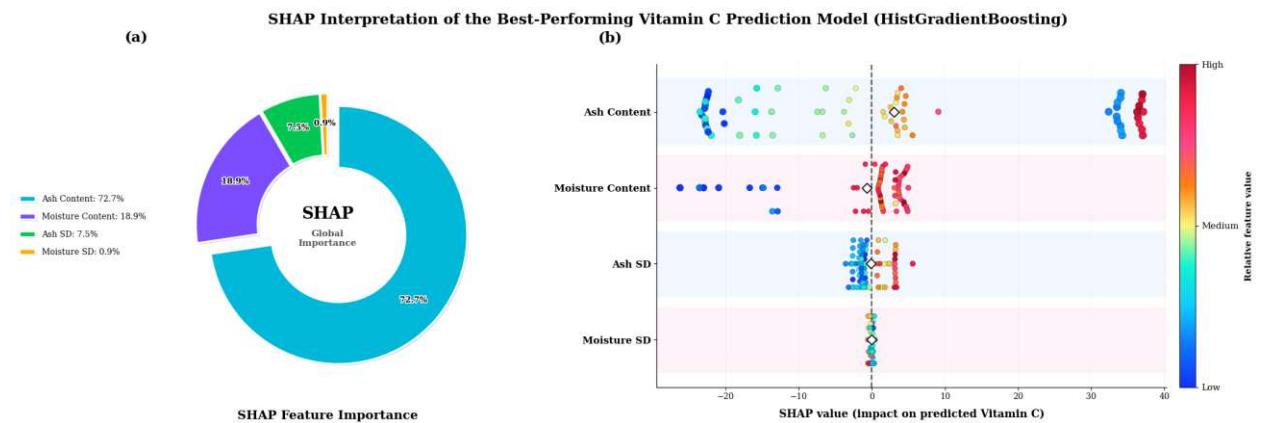


Fig. 24. SHAP interpretation of the selected HistGradientBoosting model, showing (a) global mean absolute SHAP contribution and (b) the beeswarm distribution of feature contributions across the evaluated observations.

The beeswarm component of **Fig. 24** adds information that the global percentages alone cannot provide. Mean absolute SHAP values describe the overall magnitude of a feature's contribution but remove its direction. In the beeswarm plot, individual SHAP values can be positive or negative depending on the sample's feature values and its position within the nonlinear structure learned by the model. Thus, ash does not contribute a fixed amount to every Vitamin C prediction. Its influence changes across observations according to the tree splits, interactions, and other predictor values encountered by the fitted ensemble. This local-to-global interpretation is one of the main advantages of SHAP for tree-based models [89].

The prominence of ash should also be interpreted carefully. A positive SHAP contribution indicates that, for a particular observation, the ash value helped move the model prediction above its baseline after accounting for the model's other inputs. It does not establish that experimentally increasing ash content would cause Vitamin C to increase. SHAP is fundamentally a model-attribution framework, and its feature contributions describe the behaviour of the fitted predictive function rather than a biochemical causal mechanism. This distinction is especially relevant here because ash and ash SD were themselves strongly correlated, indicating that part of their information is overlapping. The SHAP results also clarify the more limited role of the replicate-variability descriptors. Together, ash SD and moisture SD contributed approximately 8.4% of the total mean absolute SHAP importance, with most of that contribution coming from ash SD. Moisture SD remained negligible at less than 1%. This pattern is consistent with the broader feature-importance results and suggests

that the mean moisture and ash measurements carried most of the information used by the model, while replicate variability provided a smaller refinement [90] .

Finally, the SHAP interpretation agreed qualitatively with the progressive ablation analysis, although the two approaches answer different questions. SHAP describes how the final four-feature model distributes its predictions among the available inputs, whereas ablation tests what happens when the model is refitted after predictors are removed or added. Agreement between these complementary analyses therefore strengthens the interpretation that ash content carries the dominant predictive signal within the present seven-vegetable dataset, while moisture contributes a secondary signal and the SD descriptors provide comparatively modest additional information [91] .

3.7. Feature ablation: how much does ash actually add?

The ablation analysis provided the most direct test of how much predictive information was contributed by each stage of the physicochemical input set. Rather than interpreting feature importance only from a fitted model, HistGradientBoosting was refitted with progressively richer combinations of predictors. This type of reduced-versus-full feature comparison is widely used in predictive modelling to determine whether additional variables provide genuine incremental value or mainly add redundancy. Similar approaches have been used in food-quality studies to identify smaller predictor sets that retain most of the predictive performance of more complex models [92] .

Moisture content alone provided relatively weak prediction, with a test R^2 of 0.3276, RMSE of 22.5697 mg/100 g, and MAE of 17.0602 mg/100 g. The largest change occurred when ash content was added. Test R^2 increased sharply to 0.9348, while RMSE decreased to 7.0260 mg/100 g and MAE to 3.9764 mg/100 g. This step accounted for most of the improvement observed across the ablation sequence, indicating that ash supplied substantial information that was not available from moisture alone. Adding moisture SD produced a much smaller gain. Test R^2 increased from 0.9348 to 0.9386, while RMSE decreased from 7.0260 to 6.8182 mg/100 g. The final addition of ash SD produced the strongest overall configuration, with test R^2 reaching 0.9524, RMSE falling to 6.0028 mg/100 g, and MAE to 3.4879 mg/100 g. The full four-feature model was therefore retained as the preferred formulation, although the progression shows that most of the recoverable signal was already present once moisture and ash were considered together.

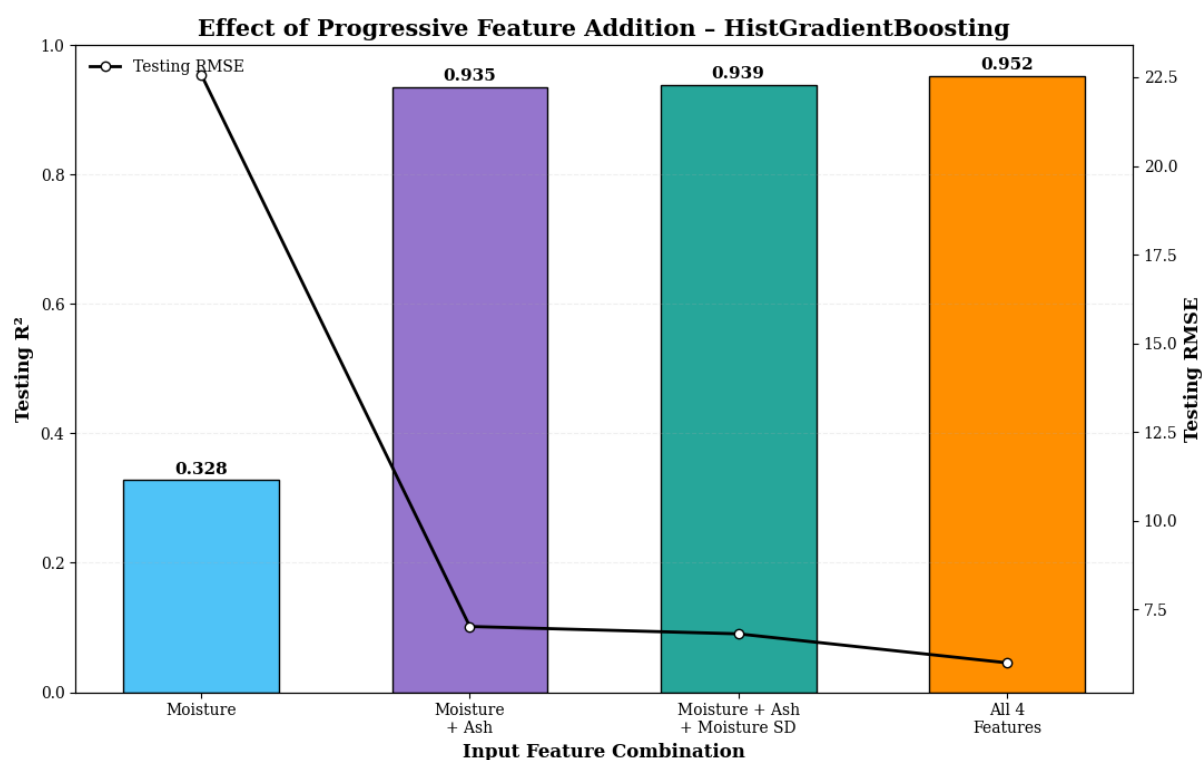
Table 9. Progressive feature-combination performance of HistGradientBoosting.

Feature set	Dataset	R ²	MAE	RMSE	MSE	MedAE	MAPE (%)
Moisture Content	Training	0.4400	15.6636	20.5923	424.0438	13.0609	58.5938
Moisture Content	Testing	0.3276	17.0602	22.5697	509.3919	17.7523	65.6662
Moisture + Ash	Training	0.9621	2.7088	5.3551	28.6769	1.2989	9.5398
Moisture + Ash	Testing	0.9348	3.9764	7.0260	49.3651	1.9541	14.7963
+ Moisture SD	Training	0.9641	2.7658	5.2134	27.1794	1.5459	9.8768
+ Moisture SD	Testing	0.9386	3.9293	6.8182	46.4883	1.9464	14.6995
All four features	Training	0.9696	2.2654	4.7996	23.0365	1.0715	8.0607
All four features	Testing	0.9524	3.4879	6.0028	36.0339	1.9970	11.9839

1877

1878 The same pattern is visible across the remaining error measures in **Table 9**. With moisture
1879 alone, test MedAE was 17.7523 mg/100 g and MAPE was 65.6662%. Once ash was included,
1880 these values fell to 1.9541 mg/100 g and 14.7963%, respectively. Moisture SD produced only
1881 a slight additional improvement, whereas the complete four-variable model reduced MAPE
1882 further to 11.9839%. The gain from the SD descriptors was therefore real but incremental
1883 compared with the much larger contribution associated with adding ash.

1884 The training results follow the same general progression. Training R² increased from 0.4400
1885 with moisture alone to 0.9621 after ash was introduced. Moisture SD raised this only slightly
1886 to 0.9641, and the final four-feature model reached 0.9696. The fact that the dominant
1887 improvement appears in both training and testing performance strengthens the interpretation
1888 that ash provides the major additional predictive signal, while the replicate-variability
1889 descriptors mainly refine an already strong moisture-plus-ash model [93] .



1890

1891 **Fig. 25.** Effect of progressive feature addition on HistGradientBoosting performance.

The pattern shown in **Fig. 25** is also consistent with recent food-quality modelling work in which feature-selection or ablation analyses were used to determine whether a reduced measurement set could preserve most of the predictive accuracy of the full model. For example, recent Vitamin C prediction studies have explicitly evaluated informative subsets rather than assuming that every available feature should automatically be retained. From a practical perspective, the ablation analysis suggests two possible levels of screening. The four-input formulation remains the preferred model when replicate measurements are available because it provides the strongest overall performance. However, the moisture-plus-ash model still achieved a test R^2 of 0.9348, indicating that a simpler two-input configuration retains much of the predictive information within the present dataset [94]. The increase from the two-feature model to the full model was approximately 0.018 R^2 units, whereas adding ash to moisture alone increased R^2 by approximately 0.607.

The reduced configuration should therefore be viewed as a lower-information screening option rather than an equivalent replacement for the full model. Its main value is practical: when replicate SDs are unavailable, moisture and ash alone may still provide a useful empirical estimate within the studied seven-vegetable domain. The full model remains preferable where replicate measurements can be obtained, while the ablation results make clear that the central predictive structure comes primarily from the combined moisture–ash profile rather than from the variability descriptors alone [95].

3.8. Practical calculation of Vitamin C from measured physicochemical inputs

The intended use of the fitted equation is illustrated with a held-out broccoli observation. The measured predictor values were moisture content = 91.02%, ash content = 1.170%, moisture SD = 0.342, and ash SD = 0.055. These four values form the input vector $x = [91.02, 1.170, 0.342, 0.055]$. The corresponding iodometric Vitamin C value was not supplied to the model; it was retained only as the reference value for checking the prediction.

$$F_0 = 42.299214, \quad \eta = 0.01, \quad M = 568$$

Therefore, the fitted HistGradientBoosting prediction for this input vector is:

$$\hat{C}_{VitC} = 42.299214 + 0.01 \sum_{m=1}^{568} h_m(91.02, 1.170, 0.342, 0.0550)$$

The cumulative learning-rate-adjusted contributions of the 568 boosting iterations can be grouped as:

$$14.968533 + 9.035311 + 8.875765 + 3.542410 + 1.726096 + 0.408999 \\ + 0.178070$$

Thus,

$$\hat{C}_{VitC} = 42.299214 + 14.968533 + 9.035311 + 8.875765 + 3.542410 \\ + 1.726096 + 0.408999 + 0.178070$$

$$\hat{C}_{VitC} = 81.0344 \text{ mg/100 g} \quad (25)$$

The corresponding reference value was $C_{ref} = 77.94 \text{ mg/100 g}$. The absolute error is therefore:

$$Absolute \text{ error} = |\hat{C}_{VitC} - C_{ref}|$$

$$= |81.0344 - 77.94|$$

$$= 3.0944 \text{ mg/100 g} \quad (26)$$

The percentage error is:

$$Percentage \text{ error} (\%) = \left(\frac{|\hat{C}_{VitC} - C_{ref}|}{C_{ref}} \right) \times 100$$

$$(3.0944 / 77.94) \times 100$$

$$\approx 3.97\% \quad (27)$$

The worked calculation shows exactly how the sample-level estimate should be interpreted. The ensemble begins from its fitted baseline, then the 568 boosting iterations provide small nonlinear corrections determined by the four physicochemical inputs. For this broccoli observation, the corrections raise the baseline from 42.299214 to 81.0344 mg/100 g. Comparison with the independent reference value of 77.94 mg/100 g gives an absolute difference of 3.0944 mg/100 g and a percentage error of approximately 3.97%. The calculation is intentionally shown in full because it makes clear that the final value is produced by the learned ensemble, not by a fabricated linear conversion between ash and Vitamin C.

To show how the same prediction equation behaves across all seven studied vegetable groups, **Eq. (15)** was applied to one held-out observation from each group. **Table 10A** condenses the four-element input vector used by the model and compares the resulting prediction with the corresponding reference Vitamin C target. The table is an illustrative calculation set; overall model accuracy continues to be based on the complete 70-observation held-out test set.

Table 10A. Equation-based Vitamin C predictions and agreement with reference values for one selected held-out observation from each vegetable.

Vegetable	Input vector $x = (M, A, SDM, SDA)$	Predicted Vitamin C	Reference Vitamin C	Absolute error	Error (%)
Green Banana	(76.47, 0.771, 0.491, 0.031)	12.3675	10.18	2.1875	21.49
Bitter Gourd	(92.54, 0.864, 0.454, 0.055)	54.4540	54.45	0.0040	0.01
Green Bell Pepper	(93.76, 0.473, 0.370, 0.012)	78.5299	80.13	1.6001	2.00
Broccoli	(91.02, 1.170, 0.342, 0.055)	81.0344	77.94	3.0944	3.97
Turnip	(93.34, 0.845, 0.398, 0.038)	50.7803	43.43	7.3503	16.92
Ash Gourd	(93.52, 0.733, 0.307, 0.031)	39.1998	15.71	23.4898	149.52
Tomato	(96.90, 0.309, 0.447, 0.018)	21.5121	17.77	3.7421	21.06

The calculation table shows both the strength and the limit of the screening equation. Bitter gourd, green bell pepper, and broccoli were especially close to their reference values, with sample-level percentage errors below 4%. Green banana, turnip, and tomato showed larger deviations, while the selected ash gourd observation was a clear failure case. This pattern is important: the equation-based model is capable of producing values close to the reference measurement for many observations, but closeness is not guaranteed for every sample.

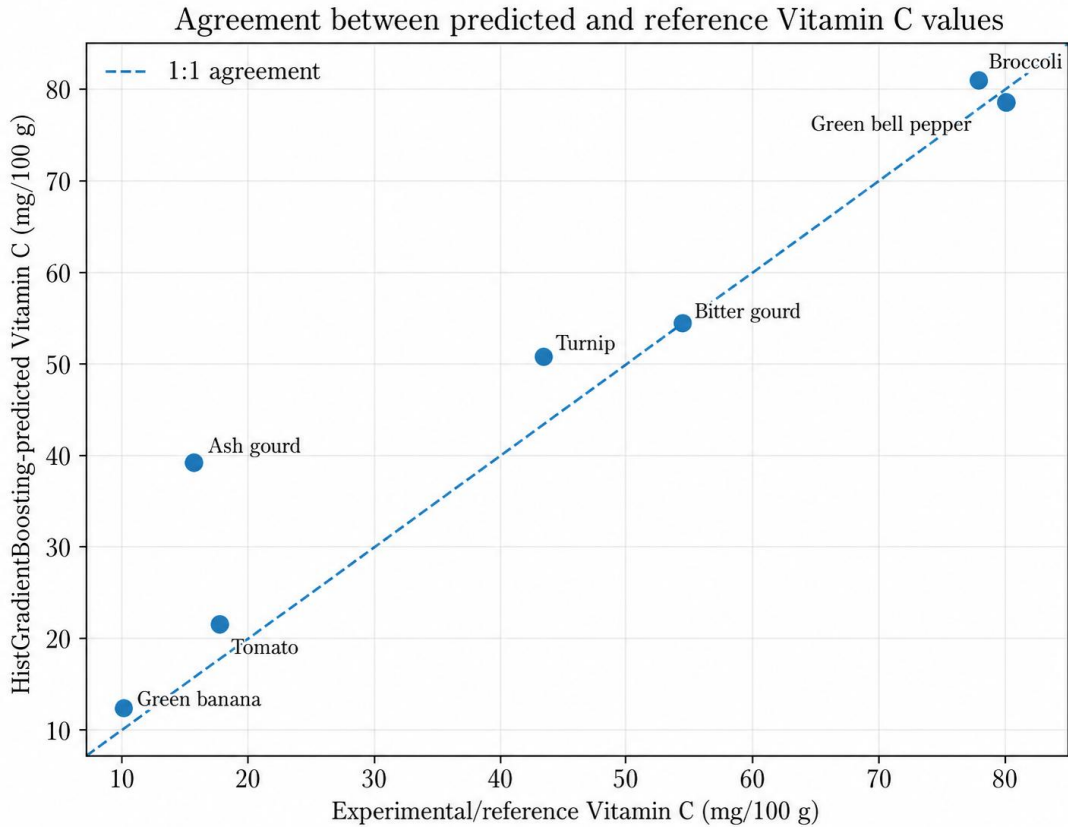


Fig. 26A. Agreement between predicted Vitamin C values based on HistGradientBoosting and true Vitamin C values for one sample (held-out sample) of each vegetable. The dashed 1:1 line represents perfect agreement – points closer to this line signify closer predictions at the sample. The seven points shown here are indicative, but should not be used to replace the full test-set metrics.

The total agreement for the held-out test set is more reliable because it represents overall agreement. Across all 70 test observations, HistGradientBoosting achieved $R^2 = 0.9524$, RMSE = 6.0028 mg/100 g, and MAE = 3.4879 mg/100 g [96]. Thus, the physicochemical-to-Vitamin C fit reproduced the greater part of the observed variation in the studied seven-vegetable domain and was generally close to the targets of the same physicochemical-to-Vitamin C; the ash gourd example illustrates the importance of direct Vitamin C analysis for verification and external validation purposes.

The calculation of the "ensemble" equation, used for the broccoli, gives a numerical example of the meaning of the ensemble equation. The fitted model begins from a baseline $F_0 = 42.299214$ mg/100 g. The selected input vector for the broccoli is [91.02, 1.170, 0.342, 0.055] and the following tree contributions are generated for this input vector, with learning rate = 0.01 and number of boosting iterations = 568. When the learning-rate-adjusted contributions are grouped across the 568 iterations, the cumulative additions are 14.968533, 9.035311, 8.875765, 3.542410, 1.726096, 0.408999, and 0.178070 mg/100 g. Adding these to the baseline gives 81.034398 mg/100 g, reported as 81.0344 mg/100 g.

The reference value for broccoli that was not measured was 77.94 mg/100 g. The absolute error is thus 3.0944 mg/100 g and the percentage error was approximately 3.97%. This is a useful worked example, as the calculation is sufficiently close to show that it is possible to apply the screening concept, but still has a visible non-zero error. The model is an estimation of a hidden titration formula, not a reproduction.

The 70 test metrics (observations) are more important than any worked case. The held out partition R^2 of the model was 0.9524 with RMSE = 6.0028 mg/100 g and MAE = 3.4879 mg/100 g across the whole held out partition. These aggregate values are the key evidence of performance in prediction; the broccoli calculation displays the passing of one sample through the fitted function.

3.9. A screening method for identifying plants with high vitamin C concentrations, as well as vegetable-specific prediction.

Although the scores were aggregated, the chosen model was also checked veggie-wise. There are some successful and unsuccessful examples included in Table 10 intentionally. The observations made on bitter gourd and green bell pepper were close and that of the selected ash gourd was not close. This contrast is significant in the context of a screening model because the high overall R^2 does not necessarily mean that the reliability is the same for each sample, or for each range of concentrations.

Table 10. Selected held-out Vitamin C values in actual/source versus HistGradientBoosting-predicted values are compared by types of vegetables.

Vegetable	Moisture	Ash	Moisture SD	Ash SD	Predicted	Actual	Abs. error	Error (%)
Green Banana	76.47	0.771	0.491	0.031	12.3675	10.18	2.1875	21.4879
Bitter Gourd	92.54	0.864	0.454	0.055	54.4540	54.45	0.0040	0.0074
Green Bell Pepper	93.76	0.473	0.370	0.012	78.5299	80.13	1.6001	1.9969
Broccoli	91.02	1.170	0.342	0.055	81.0344	77.94	3.0944	3.9702
Turnip	93.34	0.845	0.398	0.038	50.7803	43.43	7.3503	16.9246
Ash Gourd	93.52	0.733	0.307	0.031	39.1998	15.71	23.4898	149.5215
Tomato	96.90	0.309	0.447	0.018	21.5121	17.77	3.7421	21.0587

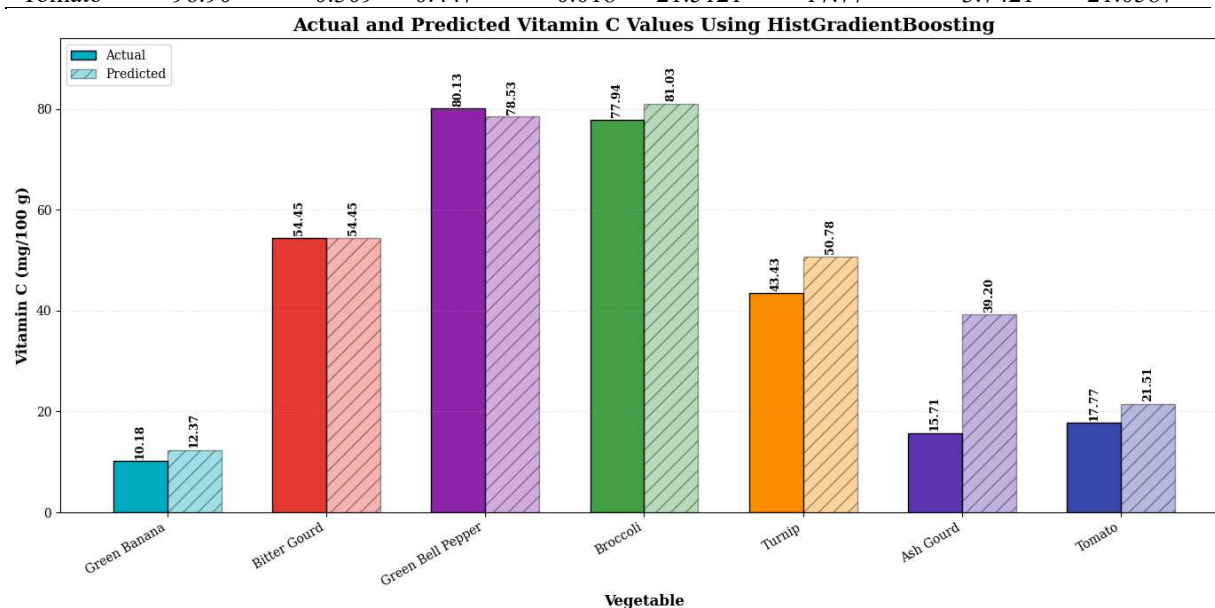


Fig. 26. The second figure shows the comparison of the actual/source and the predicted Vitamin C values of the selected vegetables using HistGradientBoosting.

The above example demonstrates that the use of the model as a substitute for the actual Vitamin C analysis would be a bad idea. The absolute error was of 23.49 mg/100 g and the percentage error was largely magnified due to the low reference value. The error region points to an area of the current predictor space in which there is an inadequate amount of moisture, ash and SDs.

2008 Additional descriptors such as maturity, cultivar, storage history, or other compositional
 2009 measurements may be needed to resolve such cases [97] .

2010 The model was also useful as a ranking tool. When the 70 held-out predictions were ordered
 2011 from highest to lowest, broccoli and green bell pepper dominated the upper part of the list. The
 2012 highest predicted value was 81.03 mg/100 g for broccoli Sample 154, and several other broccoli
 2013 observations followed closely. This agrees with the laboratory reference summary and suggests
 2014 that the fitted physicochemical model preserved the high-concentration end of the Vitamin C
 2015 distribution reasonably well [98] .

2016 **Table 11.** Fifteen highest HistGradientBoosting-predicted Vitamin C observations in the
 2017 held-out test set.

Vegetable	Sample ID	Actual Vitamin	Predicted Vitamin	Absolute Error
Broccoli	154	77.94	81.0344	3.0944
Broccoli	166	79.16	80.7753	1.6153
Broccoli	167	85.33	80.5307	4.7993
Broccoli	174	77.82	80.5307	2.7107
Broccoli	176	77.91	79.4794	1.5694
Broccoli	168	84.75	79.3433	5.4067
Broccoli	200	81.90	79.0778	2.8222
Broccoli	191	82.27	79.0450	3.2250
Broccoli	164	76.86	78.8394	1.9794
Green Bell Pepper	116	80.13	78.5299	1.6001
Green Bell Pepper	130	75.80	77.8524	2.0524
Green Bell Pepper	150	77.95	77.6173	0.3327
Green Bell Pepper	102	77.65	77.6173	0.0327
Green Bell Pepper	144	80.66	77.5507	3.1093
Green Bell Pepper	104	79.37	77.3554	2.0146

2018

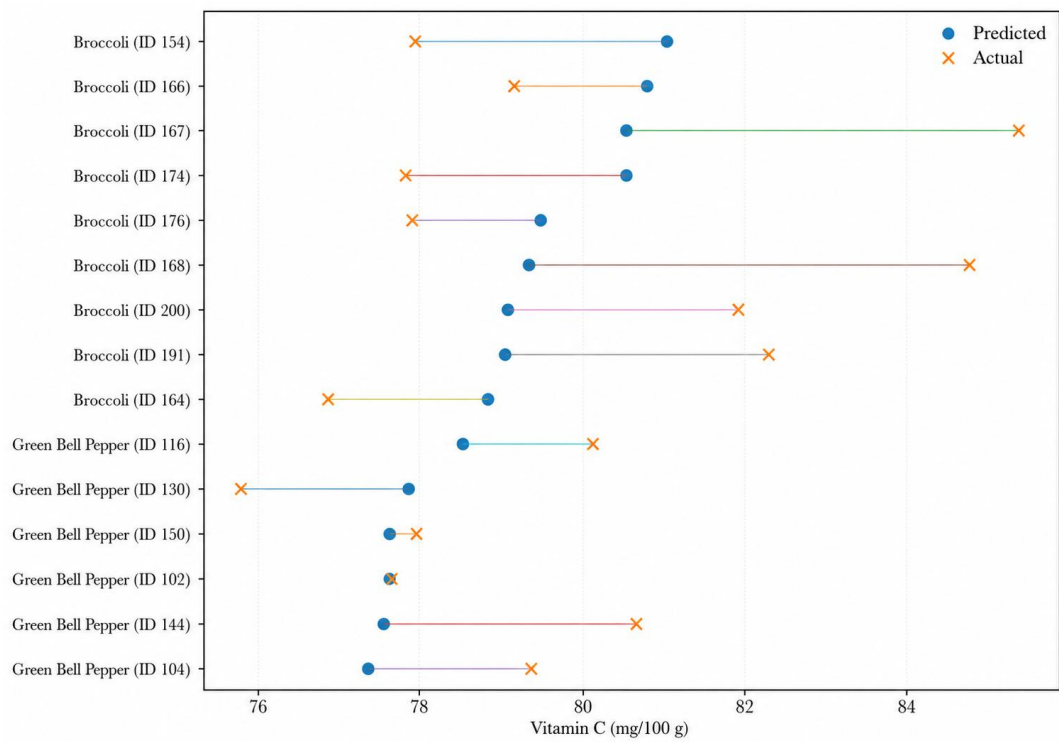


Fig. 27. Highest predicted Vitamin C observations in the held-out HistGradientBoosting test set, with actual values shown for comparison.

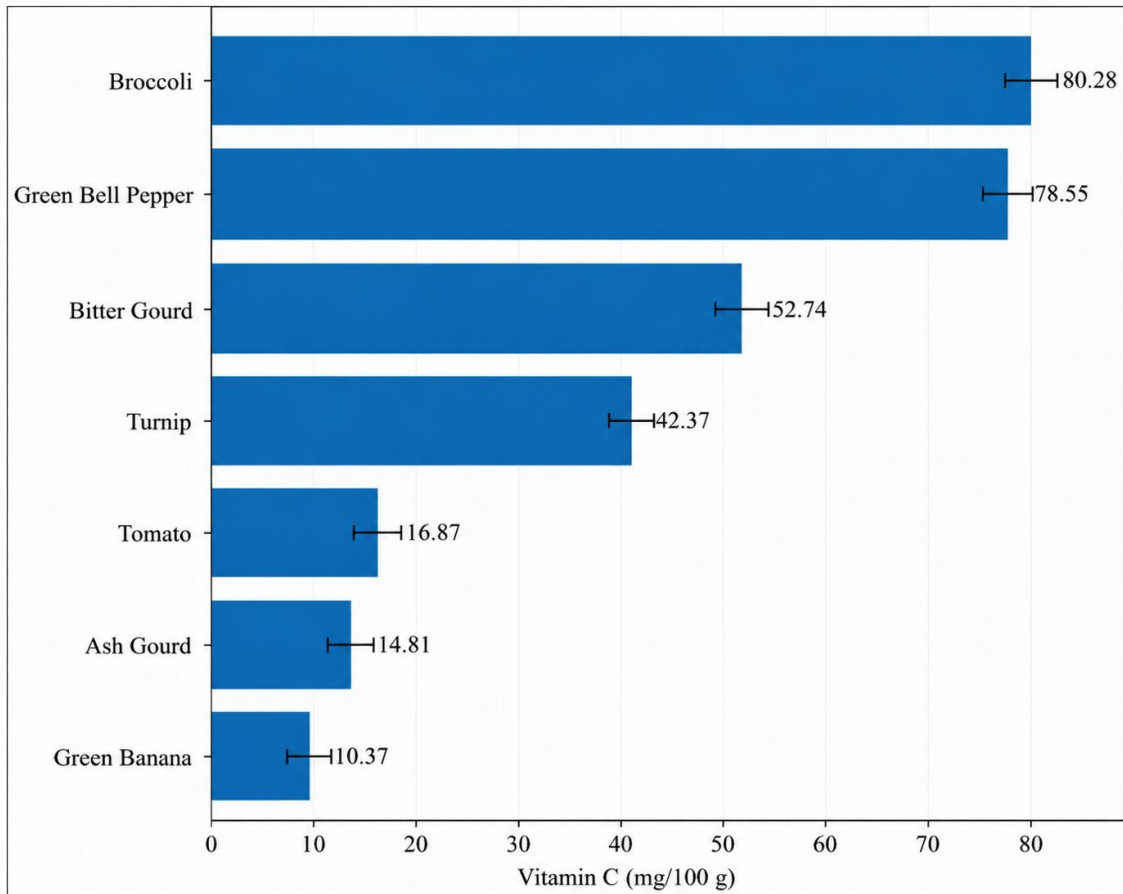


Fig. 28. Vegetable-level Vitamin C ranking based on the 50 modelling observations per vegetable (mean $\pm$ SD).

The seven selected examples in Table 10 demonstrate that sample-level accuracy is uneven. Bitter gourd was almost exact in the chosen row: 54.4540 mg/100 g predicted versus 54.45 mg/100 g reference, an absolute difference of only 0.0040 mg/100 g. Green bell pepper was also close at 78.5299 predicted versus 80.13 reference, and broccoli differed by 3.0944 mg/100 g. In contrast, the selected turnip row differed by 7.3503 mg/100 g, green banana by 2.1875, tomato by 3.7421, and ash gourd by 23.4898 mg/100 g. The ash gourd percentage error reached 149.52% because the reference concentration was low, 15.71 mg/100 g, while the model predicted 39.1998 mg/100 g. This example explains why MAPE must be interpreted cautiously in low-concentration observations. A moderate absolute error can become a very large percentage when the denominator is small. It also identifies a local part of the predictor space in which moisture, ash, and their SDs do not fully distinguish the sample's Vitamin C level. The ranking analysis was more stable at the high end. Among the fifteen highest predicted test observations, the first nine were broccoli samples, followed by green bell pepper observations. The highest prediction was 81.0344 mg/100 g for broccoli Sample 154. Other broccoli

predictions included 80.7753, 80.5307, 79.4794, 79.3433, 79.0778, 79.0450, and 78.8394 mg/100 g. Green bell pepper predictions in the displayed top group ranged from 77.3554 to 78.5299 mg/100 g. This ordering is consistent with the measured group means and supports the use of the model as a high-concentration screening aid within the represented vegetable domain [99] .

3.10. Nutritional interpretation and potential affordability extension

Within the seven vegetables studied, broccoli and green bell pepper were the clearest high-Vitamin-C groups, with bitter gourd occupying a lower but still substantial range. These rankings are specific to Vitamin C and should not be interpreted as an overall nutritional ranking, because fibre, carotenoids, phenolics, minerals, energy density, and other nutrients were outside the present model. Affordability remains a useful extension, particularly in a Bangladesh context, but it cannot be calculated from the present data because verified BDT/kg prices were not recorded. A study conducted in the future could use the same physicochemical-to-Vitamin C screening process coupled with current market prices and report the Vitamin C estimated per unit cost. This extension is not part of the current analysis which eliminates the need for assuming anything beyond what is already known about it. From a nutritional perspective, the group-level ranking is clear but should be kept narrow. Broccoli averaged 80.28 mg/100 g in the modelling table and green bell pepper 78.55 mg/100 g, followed by bitter gourd at 52.74 and turnip at 42.37. Tomato, ash gourd, and green banana were much lower at 16.87, 14.81, and 10.37 mg/100 g, respectively. These numbers describe Vitamin C only. They do not rank the vegetables for total nutritional quality, because fibre, carotenoids, phenolics, individual minerals, energy density, and other nutrients were not modelled.

The proposed affordability extension should therefore be quantitative rather than rhetorical. If verified market prices are later collected in BDT/kg for the same sampling dates and markets, the predicted or measured Vitamin C concentration could be converted into metrics such as mg Vitamin C per BDT or BDT required to obtain a fixed Vitamin C amount. No such price column exists in the present 350-row dataset, so the current manuscript does not calculate an affordability score. Keeping that boundary explicit prevents an appealing practical idea from being presented as a result that the data do not yet support.

3.11. What the model is learning and what it is not

There is a single predictive pattern coming from the various analyses. Ash is the primary model signal, moisture is a useful secondary signal, while the SD descriptors help to improve the

prediction. The following pattern should not be interpreted as a causal biochemical equation. The model has learned structures of associations between parameters in a fixed set of 7 vegetables and the main difference in tissue composition may be captured in the ash variable without being explicitly coded by any other biochemical variables. So the accuracy reported is answering a relatively practical question: within these seven groups of vegetables, how well can the moisture, ash and replicate variability recover the Vitamin C values originally determined by chemical analysis? It is not an accurate measurement for an eighth unseen species or for samples taken in other seasons, cultivars or markets. The model needs to be validated both with a leave-one-vegetable-out test and an independent experimental test set before it can be used as a transferable alternative to direct Vitamin C testing. Combined, the numerical data suggest a model that is robust in the space of the seven vegetables that were represented but not necessarily generalizable. The same groups are present in each of the sets (training vs. testing set), and a test $R^2=0.9524$ indicates an interpolation between the known vegetable profiles, rather than a transfer to an eighth unseen profile. The grouping of the data is clearly evident in the figures. The main ash contribution of 11-14 and ash gourd failure selected is in favour of this more conservative view. A leave-one-vegetable-out experiment would be more challenging, and an independent external data set from different markets, season or laboratories would be even better. The most helpful scientific inference is thus not that direct Vitamin C analysis is no longer necessary, but that direct analysis of a substance no longer requires a chemical reaction. The evidence serves to support a narrower claim: e.g., a small set of moisture- and ash-related descriptors can yield informative Vitamin C estimates for samples within the range of the samples studied for which there is evidence. The model can prioritize samples, define high prediction regions and minimize the need to consider each screening decision a new nutrient-specific assay. Reference titration is necessary to provide ground truth, to test for drift, to analyze failures and to tell if the model is applicable to data outside the distribution. [100].

In addition, the differences between the best and worst predictions in the local area further delineate the model's practical applicability. The difference between the bitter-gourd example and its reference was 0.0040 mg/100 g while in the case of the chosen ash-gourd example, it was 23.4898 mg/100 g. The broccoli worked case was 3.0944 mg/100 g different from the green-bell-pepper example, and the green-bell-pepper example was 1.6001 mg/100 g different from the broccoli worked case. These values explain why both the overall metrics and individual examples are reported in the manuscript. The $R^2=0.9524$ indicates the overall

structure recovery over 70 held out observations but it cannot predict the error for the next individual sample. The model is most useful for screening purposes in conjunction with periodic reference measurements, particularly for a new sample near the edges of moisture-ash combinations represented during training [101] .

4. Study limitations

The first one is the size of the domain. The fitted function is for a limited range of plant matrices and Vitamin C concentrations as only seven vegetables were represented. The more the number of species, cultivars, seasons and markets, the more difficult and challenging the physicochemical to nutritional relationship would be. Secondly, both the development and test partitions contained all the vegetable groups. This is valid for prediction within the range of domain covered by the studied species, but is not the same as testing a vegetable species that the model was in no way exposed to. However, external samples and leave one vegetable out evaluations are required prior to wide-spread implementation. Third, the set of predictors was purposely kept small. In addition to moisture and ash, Vitamin C is affected by maturity, cultivar, storage, oxygen, temperature, and various other biochemical properties, and measurements of these properties are routine. The SDs summarize the variability in replicates, but in addition to moisture and ash, measurements of the other properties have been made routinely: maturity, cultivar, storage, oxygen, temperature, and various other biochemical properties. The large error that is observed in one of the ash gourd sample indicates that the four inputs that are provided are not enough to solve every sample. Fourth, ash is a bulk residue measure and should not be considered as the concentration of a specific mineral. It must be noted that the predictive ability of ash is not enough to imply that it can be attributed to the action of any one mineral, namely calcium, potassium, magnesium, iron or any other mineral. Finally, direct Vitamin C analysis is still required with the present model. Training targets will need to be generated by titration or another validated reference method, and will need to be periodically confirmed by repeating the method, as well as evaluating independent samples. It is proposed to be used at screening, where trained model can provide an estimation of Vitamin C based on moisture- and ash-related measurements, without having to re-titrate for each estimation. Another extension is also affordability, since there wasn't any data available on market price verified. The sample-level modelling table in the computational workflow is essentially a reproduction of the 350 row data set. If the number of independent biological collections is claimed, this is consistent with the laboratory work carried out and the sampling design.

5. Conclusions

The 350-row dataset is reproduced as the sample-level modelling table used in the computational workflow. Any claim about the number of independent biological collections should remain consistent with the underlying laboratory records and sampling design.

The data for the routine analyses of vegetables was reformulated as a prediction problem, that is, can moisture, ash and the replicate variability of these vegetable analyses predict the Vitamin C content after training a model with chemically determined reference values? HistGradientBoosting slightly outperformed other ten Bayesian-optimized tree-based regressors in terms of CV first validation $R^2=0.9205$, test $R^2=0.9524$, and repeated validation balance (0.9354). The single test R^2 for Random Forest was slightly higher (0.9612) which further motivated against adopting a model based on a single hold-out score. The two SD descriptors made minor contributions, and the interpretation analyses were consistent: ash was the dominant descriptor (and adding ash to moisture gave larger increases in test R^2 from (0.3276 to 0.9348), and the two SD descriptors made smaller refinements, giving $R^2 = 0.9524$ in the four-input model. The practical contribution is a map of the physicochemical to Vitamin C in **Eqn. (15)-(20)**. For samples within the domain of the trained nonlinear model (the seven vegetables), the measured moisture, measured ash, moisture SD, and ash SD can be fed into the trained nonlinear model to get an estimate for Vitamin C without using the direct titration measurement as a prediction input. This is displayed by the broccoli example and the large ash gourd error displays its current border. The method should thus be presented as an "explainable screening" and not as a "validated universal replacement" to chemical analysis. Before routine deployment, a number of additional experiments are required to be done by independent groups and tests need to be conducted on vegetables that are not visible and other markets and seasons.

CRediT Authorship Contribution Statement

CRediT Authorship Contribution Statement

Md. Abdul Hamid Khan: Conceptualization, Data Curation, Formal analysis, Methodology, Field investigation, Visualization, Data Analysis, Software, Writing - original draft, Writing - Review and Editing; **Md. Mujahidul Islam:** Conceptualization, Data Curation, Formal analysis, Methodology, Field investigation, Visualization, Data Analysis, Writing - Review

and Editing; **Rubaya Hasna:** Conceptualization, Data Curation, Formal analysis, Methodology, Field investigation, Visualization, Data Analysis, Writing - Review and Editing; **Shohanur Rahman:** Conceptualization, Data Curation, Formal analysis, Methodology, Field investigation, Visualization, Data Analysis, Writing - Review and Editing

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Funding

We have not received any funding from anywhere to support the research.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors acknowledge that some parts of this manuscript were prepared using the generative artificial intelligence ChatGPT (OpenAI) and Grammarly tools. The tasks that were supported by these tools included grammar refinement, stylistic editing, and content organization. The authors formulated all scientific concepts, data interpretations and conclusions and were responsible to the integrity and accuracy of the work. The originality, integrity and intellectual contribution of the manuscript were not affected by the use of AI.

6. References:

- [1] N. S. Gruda, G. Samuolienė, J. Dong, and X. Li, “Environmental conditions and nutritional quality of vegetables in protected cultivation,” *Compr Rev Food Sci Food Saf*, vol. 24, no. 2, p. e70139, Mar. 2025, doi: 10.1111/1541-4337.70139.

- [2] K. Verma, P. K. Nema, P. C. Badgujar, K. Venkatesh, A. Rajendran, and G. L. Loushigam, "Vitamin C in fruits and vegetables: stability, processing effects, analytical detection, nutrient interactions, bioavailability, and health implications," *Food Chemistry*, vol. 505, p. 148058, Mar. 2026, doi: 10.1016/j.foodchem.2026.148058.
- [3] C. S. Johnston, "Chapter 9 - Vitamin C," in *Present Knowledge in Nutrition (Eleventh Edition)*, B. P. Marriott, D. F. Birt, V. A. Stallings, and A. A. Yates, Eds., Academic Press, 2020, pp. 155–169. doi: 10.1016/B978-0-323-66162-1.00009-3.
- [4] R. Bhandari, H. Achhami, B. B. Pachhai, S. Chaudhary, S. P. Pant, and L. R. Bhatt, "Physicochemical, proximate composition, and antioxidant activity of fresh and fermented juices of selected wild edible fruits from the Gulmi District, Western Nepal," *LWT*, vol. 231, p. 118307, Sep. 2025, doi: 10.1016/j.lwt.2025.118307.
- [5] M. Abdellatief, M. E. Abd-Elmaboud, M. Mortagi, and A. M. Saqr, "A convolutional neural network-based deep learning approach for predicting surface chloride concentration of concrete in marine tidal zones," *Sci Rep*, vol. 15, p. 27611, Jul. 2025, doi: 10.1038/s41598-025-12035-1.
- [6] "(PDF) Tree-Based Methods: A Tool for Modeling Nonlinear Complex Relationships and Generating New Insights from Data." Accessed: Aug. 16, 2026. [Online]. Available: https://www.researchgate.net/publication/362096004_Tree-Based_Methods_A_Tool_for_Modeling_Nonlinear_Complex_Relationships_and_Generating_New_Insights_from_Data
- [7] K. Zushi *et al.*, "Machine learning and multiple linear regression models can predict ascorbic acid and polyphenol contents, and antioxidant activity in strawberries," *Journal of the Science of Food and Agriculture*, vol. 105, pp. 1159–1169, Sep. 2024, doi: 10.1002/jsfa.13906.
- [8] "Computational Chemistry as Applied in Environmental Research: Opportunities and Challenges | ACS ES&T Engineering | ACS Publications." Accessed: Aug. 16, 2026. [Online]. Available: <https://pubs.acs.org/aeecco/article-abstract/4/1/66/352041/Computational-Chemistry-as-Applied-in?redirectedFrom=fulltext>
- [9] M. Ojo, *MODIFICATION OF IODINE TITRATION METHOD FOR THE DETERMINATION OF VITAMIN C*. 2006.

- 2226 [10] “Practical feature filter strategy to machine learning for small datasets in chemistry |
2227 Scientific Reports.” Accessed: Aug. 17, 2026. [Online]. Available:
2228 <https://www.nature.com/articles/s41598-024-71342-1>
- 2229 [11] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, “Data leakage
2230 inflates prediction performance in connectome-based machine learning models,” *Nat*
2231 *Commun*, vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.
- 2232 [12] “1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking,”
2233 scikit-learn. Accessed: Aug. 17, 2026. [Online]. Available: [https://scikit-](https://scikit-learn/stable/modules/ensemble.html)
2234 [learn/stable/modules/ensemble.html](https://scikit-learn/stable/modules/ensemble.html)
- 2235 [13] B. Bischl *et al.*, “Hyperparameter optimization: Foundations, algorithms, best
2236 practices, and open challenges,” *WIREs Data Mining and Knowledge Discovery*, vol.
2237 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.
- 2238 [14] “HistGradientBoostingRegressor,” scikit-learn. Accessed: Aug. 17, 2026. [Online].
2239 Available: [https://scikit-](https://scikit-learn/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html)
2240 [learn/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html](https://scikit-learn/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html)
- 2241 [15] T. Parr, J. Hamrick, and J. D. Wilson, “Nonparametric feature impact and importance,”
2242 *Information Sciences*, vol. 653, p. 119563, Jan. 2024, doi: 10.1016/j.ins.2023.119563.
- 2243 [16] G. Shmueli, “To Explain or to Predict?,” *Statistical Science*, vol. 25, no. 3, pp. 289–
2244 310, Aug. 2010, doi: 10.1214/10-STS330.
- 2245 [17] “USDA FoodData Central.” Accessed: Aug. 17, 2026. [Online]. Available:
2246 https://fdc.nal.usda.gov/?utm_source=chatgpt.com
- 2247 [18] G. Shmueli, “To Explain or to Predict?,” *Statistical Science*, vol. 25, no. 3, pp. 289–
2248 310, Aug. 2010, doi: 10.1214/10-STS330.
- 2249 [19] Y. Hu, R. Sandt, and R. Spatschek, “Practical feature filter strategy to machine
2250 learning for small datasets in chemistry,” *Sci Rep*, vol. 14, no. 1, p. 20449, Sep. 2024,
2251 doi: 10.1038/s41598-024-71342-1.
- 2252 [20] K. J. Mila *et al.*, “Nutritional composition, bioactive compounds and antioxidant
2253 potentiality of some indigenous vegetables consumed in Bangladesh,” *Sci Rep*, vol. 14,
2254 p. 27699, Nov. 2024, doi: 10.1038/s41598-024-78625-7.
- 2255 [21] K. J. Mila *et al.*, “Nutritional composition, bioactive compounds and antioxidant
2256 potentiality of some indigenous vegetables consumed in Bangladesh,” *Sci Rep*, vol. 14,
2257 p. 27699, Nov. 2024, doi: 10.1038/s41598-024-78625-7.

- 2258 [22] V. N. Okafor, D. O. Omokpariola, and D. E. Enenche, “Determination of vitamin c in
2259 raw fruit and vegetable homogenates: dietary exposure and health effects of excess
2260 intake in adults and children”, doi: 10.32394/rpzh.2024.0294.
- 2261 [23] K. Zushi *et al.*, “Machine learning and multiple linear regression models can predict
2262 ascorbic acid and polyphenol contents, and antioxidant activity in strawberries,”
2263 *Journal of the Science of Food and Agriculture*, vol. 105, no. 2, pp. 1159–1169, 2025,
2264 doi: 10.1002/jsfa.13906.
- 2265 [24] B. Bischl *et al.*, “Hyperparameter optimization: Foundations, algorithms, best
2266 practices, and open challenges,” *WIREs Data Mining and Knowledge Discovery*, vol.
2267 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.
- 2268 [25] H.-J. He *et al.*, “Improved prediction of vitamin C and reducing sugar content in
2269 sweetpotatoes using hyperspectral imaging and LARS-enhanced LASSO variable
2270 selection,” *Journal of Food Composition and Analysis*, vol. 132, p. 106350, Aug. 2024,
2271 doi: 10.1016/j.jfca.2024.106350.
- 2272 [26] “Data leakage inflates prediction performance in connectome-based machine learning
2273 models | Nature Communications.” Accessed: Aug. 17, 2026. [Online]. Available:
2274 https://www.nature.com/articles/s41467-024-46150-w?utm_source=chatgpt.com
- 2275 [27] K. J. Mila *et al.*, “Nutritional composition, bioactive compounds and antioxidant
2276 potentiality of some indigenous vegetables consumed in Bangladesh,” *Sci Rep*, vol. 14,
2277 p. 27699, Nov. 2024, doi: 10.1038/s41598-024-78625-7.
- 2278 [28] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, “Data leakage
2279 inflates prediction performance in connectome-based machine learning models,” *Nat*
2280 *Commun*, vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.
- 2281 [29] I. H. Mondal, L. Rangan, and R. V. S. Uppaluri, “Effect of oven and intermittent
2282 airflow assisted tray drying methods on nutritional parameters of few leafy and non-
2283 leafy vegetables of North-East India,” *Heliyon*, vol. 5, no. 11, p. e02934, Nov. 2019,
2284 doi: 10.1016/j.heliyon.2019.e02934.
- 2285 [30] “Application of Near-Infrared Spectroscopy in Moisture Detection of Carrot Slices
2286 During Freeze-Drying.” Accessed: Aug. 17, 2026. [Online]. Available:
2287 https://www.mdpi.com/2304-8158/15/7/1256?utm_source=chatgpt.com
- 2288 [31] Y. Hu, R. Sandt, and R. Spatschek, “Practical feature filter strategy to machine
2289 learning for small datasets in chemistry,” *Sci Rep*, vol. 14, no. 1, p. 20449, Sep. 2024,
2290 doi: 10.1038/s41598-024-71342-1.

- 2291 [32] I. H. Mondal, L. Rangan, and R. V. S. Uppaluri, "Effect of oven and intermittent
2292 airflow assisted tray drying methods on nutritional parameters of few leafy and non-
2293 leafy vegetables of North-East India," *Heliyon*, vol. 5, no. 11, p. e02934, Nov. 2019,
2294 doi: 10.1016/j.heliyon.2019.e02934.
- 2295 [33] "From local explanations to global understanding with explainable AI for trees | Nature
2296 Machine Intelligence." Accessed: Aug. 17, 2026. [Online]. Available:
2297 <https://www.nature.com/articles/s42256-019-0138-9>
- 2298 [34] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate Shapley
2299 value feature attributions," *Nat Mach Intell*, vol. 5, no. 6, pp. 590–601, Jun. 2023, doi:
2300 10.1038/s42256-023-00657-x.
- 2301 [35] S. M. Lundberg *et al.*, "From local explanations to global understanding with
2302 explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi:
2303 10.1038/s42256-019-0138-9.
- 2304 [36] G. Shmueli, "To Explain or to Predict?," *Statistical Science*, vol. 25, no. 3, pp. 289–
2305 310, Aug. 2010, doi: 10.1214/10-STS330.
- 2306 [37] "Official Methods of Analysis 2023," AOAC INTERNATIONAL. Accessed: Aug. 19,
2307 2026. [Online]. Available: [https://www.aoac.org/resources/official-methods-of-](https://www.aoac.org/resources/official-methods-of-analysis-2019/)
2308 [analysis-2019/](https://www.aoac.org/resources/official-methods-of-analysis-2019/)
- 2309 [38] Y. Hu, R. Sandt, and R. Spatschek, "Practical feature filter strategy to machine
2310 learning for small datasets in chemistry," *Sci Rep*, vol. 14, no. 1, p. 20449, Sep. 2024,
2311 doi: 10.1038/s41598-024-71342-1.
- 2312 [39] S. M. Lundberg *et al.*, "From local explanations to global understanding with
2313 explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi:
2314 10.1038/s42256-019-0138-9.
- 2315 [40] "Titrations involving iodine are widely used in analytical chemistry." Accessed: Aug.
2316 19, 2026. [Online]. Available:
2317 https://staff.buffalostate.edu/nazareay/che112/iodine.htm?utm_source=chatgpt.com
- 2318 [41] "Your I3." Accessed: Aug. 19, 2026. [Online]. Available:
2319 [https://www.vanderbilt.edu/AnS/Chemistry/courses/chem104/experiment6/back2.htm?](https://www.vanderbilt.edu/AnS/Chemistry/courses/chem104/experiment6/back2.htm?utm_source=chatgpt.com)
2320 [utm_source=chatgpt.com](https://www.vanderbilt.edu/AnS/Chemistry/courses/chem104/experiment6/back2.htm?utm_source=chatgpt.com)
- 2321 [42] "Titrations involving iodine are widely used in analytical chemistry." Accessed: Aug.
2322 19, 2026. [Online]. Available:
2323 https://staff.buffalostate.edu/nazareay/che112/iodine.htm?utm_source=chatgpt.com

- 2324 [43] T. Mucci, “What is Data Leakage in Machine Learning? | IBM.” Accessed: Aug. 19,
2325 2026. [Online]. Available: [https://www.ibm.com/think/topics/data-leakage-machine-](https://www.ibm.com/think/topics/data-leakage-machine-learning)
2326 learning
- 2327 [44] “What is Data Leakage in Machine Learning? | IBM.” Accessed: Aug. 19, 2026.
2328 [Online]. Available: [https://www.ibm.com/think/topics/data-leakage-machine-](https://www.ibm.com/think/topics/data-leakage-machine-learning?utm_source=chatgpt.com)
2329 learning?utm_source=chatgpt.com
- 2330 [45] “INFOODS: Tables and databases.” Accessed: Aug. 19, 2026. [Online]. Available:
2331 [https://www.fao.org/infoods/infoods/tables-and-](https://www.fao.org/infoods/infoods/tables-and-databases/en/?utm_source=chatgpt.com)
2332 databases/en/?utm_source=chatgpt.com
- 2333 [46] “12. Common pitfalls and recommended practices — scikit-learn 1.9.0
2334 documentation.” Accessed: Aug. 19, 2026. [Online]. Available: [https://scikit-](https://scikit-learn.org/stable/common_pitfalls.html?utm_source=chatgpt.com)
2335 learn.org/stable/common_pitfalls.html?utm_source=chatgpt.com
- 2336 [47] “HistGradientBoostingRegressor — scikit-learn 1.9.0 documentation.” Accessed: Aug.
2337 19, 2026. [Online]. Available: [https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html?utm_source=chatgpt.com)
2338 learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.h
2339 tml?utm_source=chatgpt.com
- 2340 [48] “StratifiedKFold,” scikit-learn. Accessed: Aug. 19, 2026. [Online]. Available:
2341 [https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html)
2342 learn/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
- 2343 [49] “GroupKFold — scikit-learn 1.9.0 documentation.” Accessed: Aug. 19, 2026.
2344 [Online]. Available: [https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html?utm_source=chatgpt.com)
2345 learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html?utm_so
2346 urce=chatgpt.com
- 2347 [50] “GroupKFold,” scikit-learn. Accessed: Aug. 19, 2026. [Online]. Available:
2348 [https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html)
2349 learn/stable/modules/generated/sklearn.model_selection.GroupKFold.html
- 2350 [51] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001,
2351 doi: 10.1023/A:1010933404324.
- 2352 [52] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-
2353 learning-based science,” *Patterns (N Y)*, vol. 4, no. 9, p. 100804, Aug. 2023, doi:
2354 10.1016/j.patter.2023.100804.
- 2355 [53] “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance
2356 Evaluation.” Accessed: Aug. 19, 2026. [Online]. Available:
2357 https://www.jmlr.org/papers/v11/cawley10a.html?utm_source=chatgpt.com

- [54] J. Snoek, H. Larochelle, and R. Adams, “Practical Bayesian Optimization of Machine Learning Algorithms,” *Advances in Neural Information Processing Systems*, vol. 25, 2012, Accessed: Aug. 19, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html?utm_source=chatgpt.com
- [55] “HistGradientBoostingRegressor,” scikit-learn. Accessed: Aug. 19, 2026. [Online]. Available: <https://scikit-learn/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html>
- [56] “Human Verification.” Accessed: Aug. 19, 2026. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2001AnSta..2903451F?utm_source=chatgpt.com
- [57] “Moving towards reproducible machine learning,” *Nat Comput Sci*, vol. 1, no. 10, pp. 629–630, Oct. 2021, doi: 10.1038/s43588-021-00152-6.
- [58] “3.4. Metrics and scoring: quantifying the quality of predictions,” scikit-learn. Accessed: Aug. 19, 2026. [Online]. Available: https://scikit-learn/stable/modules/model_evaluation.html
- [59] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, Oct. 2006, doi: 10.1016/j.ijforecast.2006.03.001.
- [60] G. C. Cawley and N. L. C. Talbot, “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation,” *Journal of Machine Learning Research*, vol. 11, no. 70, pp. 2079–2107, 2010.
- [61] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017, Accessed: Aug. 19, 2026. [Online]. Available: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html?utm_source=chatgpt.com
- [62] H.-J. He *et al.*, “Improved prediction of vitamin C and reducing sugar content in sweetpotatoes using hyperspectral imaging and LARS-enhanced LASSO variable selection,” *Journal of Food Composition and Analysis*, vol. 132, p. 106350, Aug. 2024, doi: 10.1016/j.jfca.2024.106350.
- [63] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, “Problems with Shapley-value-based explanations as feature importance measures,” in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Nov. 2020, pp. 5491–

2391 5500. Accessed: Aug. 19, 2026. [Online]. Available:
2392 <https://proceedings.mlr.press/v119/kumar20e.html>

2393 [64] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *The*
2394 *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi:
2395 10.1214/aos/1013203451.

2396 [65] C. Liu, S. X. Yang, and L. Deng, “Determination of internal qualities of Newhall navel
2397 oranges based on NIR spectroscopy using machine learning,” *Journal of Food*
2398 *Engineering*, vol. 161, pp. 16–23, Sep. 2015, doi: 10.1016/j.jfoodeng.2015.03.022.

2399 [66] H.-J. He *et al.*, “Improved prediction of vitamin C and reducing sugar content in
2400 sweetpotatoes using hyperspectral imaging and LARS-enhanced LASSO variable
2401 selection,” *Journal of Food Composition and Analysis*, vol. 132, p. 106350, Aug. 2024,
2402 doi: 10.1016/j.jfca.2024.106350.

2403 [67] H.-J. He *et al.*, “Improved prediction of vitamin C and reducing sugar content in
2404 sweetpotatoes using hyperspectral imaging and LARS-enhanced LASSO variable
2405 selection,” *Journal of Food Composition and Analysis*, vol. 132, p. 106350, Aug. 2024,
2406 doi: 10.1016/j.jfca.2024.106350.

2407 [68] “Bootstrap Confidence Regions Computed from Autoregressions of Arbitrary Order |
2408 Journal of the Royal Statistical Society Series B: Statistical Methodology | Oxford
2409 Academic.” Accessed: Aug. 19, 2026. [Online]. Available:
2410 [https://academic.oup.com/jrsssb/article-](https://academic.oup.com/jrsssb/article-abstract/62/2/461/7083242?redirectedFrom=fulltext)
2411 [abstract/62/2/461/7083242?redirectedFrom=fulltext](https://academic.oup.com/jrsssb/article-abstract/62/2/461/7083242?redirectedFrom=fulltext)

2412 [69] A. Sánchez, D. Vélez, and V. Devesa, “Processes influencing the toxicity of
2413 microplastics ingested through the diet,” *Food Chemistry*, vol. 456, p. 139947, Oct.
2414 2024, doi: 10.1016/j.foodchem.2024.139947.

2415 [70] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001,
2416 doi: 10.1023/A:1010933404324.

2417 [71] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model
2418 Predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017,
2419 Accessed: Aug. 19, 2026. [Online]. Available:
2420 [https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767](https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html?utm_source=chatgpt.com)
2421 [-Abstract.html?utm_source=chatgpt.com](https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html?utm_source=chatgpt.com)

2422 [72] R. T. Forestano, K. T. Matchev, K. Matcheva, and E. B. Unlu, “Supervised Machine
2423 Learning Methods with Uncertainty Quantification for Exoplanet Atmospheric

2424 Retrievals from Transmission Spectroscopy,” Aug. 07, 2025, *arXiv*: arXiv:2508.04982.
 2425 doi: 10.48550/arXiv.2508.04982.

2426 [73] P.-P. D. Breuck, M. L. Evans, and G.-M. Rignanese, “Robust model benchmarking and
 2427 bias-imbalance in data-driven materials science: a case study on MODNet,” *J. Phys.:
 2428 Condens. Matter*, vol. 33, no. 40, p. 404002, Oct. 2021, doi: 10.1088/1361-
 2429 648X/ac1280.

2430 [74] A. L. Suárez-Cetrulo, A. Kumar, and L. Miralles-Pechuán, “Modelling the COVID-19
 2431 virus evolution with Incremental Machine Learning,” Apr. 21, 2021, *arXiv*:
 2432 arXiv:2104.09325. doi: 10.48550/arXiv.2104.09325.

2433 [75] J. Camacho, “A Set of Rules for Model Validation,” *Journal of Chemometrics*, vol. 40,
 2434 no. 2, p. e70110, Feb. 2026, doi: 10.1002/cem.70110.

2435 [76] “Spider Chart Explained: Examples, Uses & How to Create.” Accessed: Aug. 19,
 2436 2026. [Online]. Available: <https://algoscale.com/blog/what-is-spider-chart/>

2437 [77] L. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model
 2438 selection: a primer with examples from ecology,” Mar. 09, 2022, *arXiv*:
 2439 arXiv:2203.04552. doi: 10.48550/arXiv.2203.04552.

2440 [78] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model
 2441 selection: A review with examples from ecology,” *Ecological Monographs*, vol. 93,
 2442 no. 1, p. e1557, 2023, doi: 10.1002/ecm.1557.

2443 [79] K. Halder *et al.*, “High-resolution maize yield mapping across Africa using earth
 2444 observation and machine learning, deep learning, and foundation model,” *Science of
 2445 Remote Sensing*, vol. 13, p. 100344, Jun. 2026, doi: 10.1016/j.srs.2025.100344.

2446 [80] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine
 2447 Learning,” Nov. 11, 2020, *arXiv*: arXiv:1811.12808. doi: 10.48550/arXiv.1811.12808.

2448 [81] A. G. Khoei, Y. Yu, and R. Feldt, “Domain Generalization through Meta-Learning: A
 2449 Survey,” Aug. 22, 2024, *arXiv*: arXiv:2404.02785. doi: 10.48550/arXiv.2404.02785.

2450 [82] T. Pan, V. Z. Yu, V. Devanarayan, and L. Tian, “A Honest Cross-Validation Estimator
 2451 for Prediction Performance,” Oct. 09, 2025, *arXiv*: arXiv:2510.07649. doi:
 2452 10.48550/arXiv.2510.07649.

2453 [83] “(PDF) On Over-fitting in Model Selection and Subsequent Selection Bias in
 2454 Performance Evaluation.” Accessed: Aug. 20, 2026. [Online]. Available:
 2455 [https://www.researchgate.net/publication/220320908_On_Over-
 2456 fitting_in_Model_Selection_and_Subsequent_Selection_Bias_in_Performance_Evalua
 2457 tion](https://www.researchgate.net/publication/220320908_On_Over-fitting_in_Model_Selection_and_Subsequent_Selection_Bias_in_Performance_Evaluation)

2458 [84] C. Eve, G. Varoquaux, and T. Moreau, “Crossing the Validation Crisis: Cross-
2459 Validation Reduces Benchmarking Variance Surprisingly Well,” Jun. 10, 2026, *arXiv*:
2460 arXiv:2606.12552. doi: 10.48550/arXiv.2606.12552.

2461 [85] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model
2462 selection: A review with examples from ecology,” *Ecological Monographs*, vol. 93,
2463 no. 1, p. e1557, 2023, doi: 10.1002/ecm.1557.

2464 [86] A. Suter, “Importance measures derived from random forests: characterisation and
2465 extension,” Jun. 21, 2021, *arXiv*: arXiv:2106.09473. doi: 10.48550/arXiv.2106.09473.

2466 [87] “Be careful when interpreting predictive models in search of causal insights — SHAP
2467 latest documentation.” Accessed: Aug. 20, 2026. [Online]. Available:
2468 [https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20](https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html)
2469 [when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20](https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html)
2470 [insights.html](https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html)

2471 [88] S. M. Lundberg *et al.*, “Explainable AI for Trees: From Local Explanations to Global
2472 Understanding,” May 11, 2019, *arXiv*: arXiv:1905.04610. doi:
2473 10.48550/arXiv.1905.04610.

2474 [89] M. Brenndorfer, “SHAP (SHapley Additive exPlanations): Model Interpretability.”
2475 Accessed: Aug. 20, 2026. [Online]. Available:
2476 [https://mbrenndorfer.com/writing/shap-shapley-additive-explanations-complete-](https://mbrenndorfer.com/writing/shap-shapley-additive-explanations-complete-guide-model-interpretability-feature-attribution)
2477 [guide-model-interpretability-feature-attribution](https://mbrenndorfer.com/writing/shap-shapley-additive-explanations-complete-guide-model-interpretability-feature-attribution)

2478 [90] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent Individualized Feature
2479 Attribution for Tree Ensembles,” Mar. 07, 2019, *arXiv*: arXiv:1802.03888. doi:
2480 10.48550/arXiv.1802.03888.

2481 [91] Y. Zhao, S. Chen, and N. Su, “When Does Multi-Agent Collaboration Help? An
2482 Entropy Perspective,” Jun. 04, 2026, *arXiv*: arXiv:2602.04234. doi:
2483 10.48550/arXiv.2602.04234.

2484 [92] D. Mowery, C. Bryan, and M. Conway, “Feature Studies to Inform the Classification
2485 of Depressive Symptoms from Twitter Data for Population Health,” Jan. 28, 2017,
2486 *arXiv*: arXiv:1701.08229. doi: 10.48550/arXiv.1701.08229.

2487 [93] S. Kircali Ata *et al.*, “Predicting the Textural Properties of Plant-Based Meat Analogs
2488 with Machine Learning,” *Foods*, vol. 12, no. 2, p. 344, Jan. 2023, doi:
2489 10.3390/foods12020344.

- [94] D. Kirk, C. Catal, and B. Tekinerdogan, "Predicting Plasma Vitamin C Using Machine Learning," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2042924, Dec. 2022, doi: 10.1080/08839514.2022.2042924.
- [95] S. Kircali Ata *et al.*, "Predicting the Textural Properties of Plant-Based Meat Analogs with Machine Learning," *Foods*, vol. 12, no. 2, p. 344, Jan. 2023, doi: 10.3390/foods12020344.
- [96] R. T. Forestano, K. T. Matchev, K. Matcheva, and E. B. Unlu, "Supervised Machine Learning Methods with Uncertainty Quantification for Exoplanet Atmospheric Retrievals from Transmission Spectroscopy," Aug. 07, 2025, *arXiv*: arXiv:2508.04982. doi: 10.48550/arXiv.2508.04982.
- [97] "MAPE (Mean Absolute Percentage Error): complete guide in January 2026," Openlayer. Accessed: Aug. 20, 2026. [Online]. Available: <https://www.openlayer.com/blog/mape-mean-absolute-percentage-error>
- [98] D. Boldini, L. Friedrich, D. Kuhn, and S. A. Sieber, "Machine Learning Assisted Hit Prioritization for High Throughput Screening in Drug Discovery," *ACS Cent Sci*, vol. 10, no. 4, pp. 823–832, Mar. 2024, doi: 10.1021/acscentsci.3c01517.
- [99] "Absolute Error, Percentage Error, And Mean Absolute Percentage Error (mape)," FasterCapital. Accessed: Aug. 20, 2026. [Online]. Available: [https://fastercapital.com/topics/absolute-error,-percentage-error,-and-mean-absolute-percentage-error-\(mape\).html/1](https://fastercapital.com/topics/absolute-error,-percentage-error,-and-mean-absolute-percentage-error-(mape).html/1)
- [100] B. Shon *et al.*, "Clinical Laboratory Parameter–Driven Machine Learning for Participant Selection in Bioequivalence Studies Among Patients With Gastric Cancer: Framework Development and Validation Study," *JMIR AI*, vol. 4, p. e64845, May 2025, doi: 10.2196/64845.
- [101] "Clinical Laboratory Parameter–Driven Machine Learning for Participant Selection in Bioequivalence Studies Among Patients With Gastric Cancer: Framework Development and Validation Study - PMC." Accessed: Aug. 20, 2026. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12223687/>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.docx](#)