Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Multimodal Deep Learning for Predicting Apple (Malus domestica Borkh.) Graft Performance in Ethiopian Highland Nurseries: Cross-Method Validation of Empirical Agronomic Relationships

Simachew Melaku Ayalew

smmb2009sm@gmail.com

Bahir Dar University https://orcid.org/0000-0002-8952-1966

Research Article

Keywords: Apple, graft performance, Ethiopian highlands, chilling requirement, deep learning, ordinal classification, SHAP, explainable artificial intelligence

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11648409/v1

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

Additional Declarations: The authors declare no competing interests.

Explainable Multimodal Deep Learning for Predicting Apple (Malus domestica Borkh.) Graft Performance in Ethiopian Highland Nurseries: Cross-Method Validation of Empirical Agronomic Relationships

[1]Anteneh Birhan Yenew, [1]Simachew Melaku Ayalew, [1]Abrham Tegegne Bogale, [1]Agerie Belete Aderaw, [2]Yibeltal Tilahun Ayele,
[1]Department of Information Technology, College of Engineering and Technology, Injibara University, Ethiopia
[2]Department of Horticulture, College of Agriculture, Injibara University, Ethiopia
anteneh.birhan@inu.edu.et, simachew.melaku@inu.edu.et , abrham.tegegne@inu.edu.et , agerie.belete@inu.edu.et, yibeltal290@gmail.com
"Corresponding Author= Simachew.melaku@inu.edu.et"

Abstract

Background: *Apple production is growing in the highlands of Ethiopia. People who grow apples do not have good tools to help them pick the right apple variety or to predict how well a graft will work. This study aimed to create and test a deep-learning system that can explain its predictions about apple graft performance and to check if the explanations match rules that farmers use.*

Objective: *The goal of this study was to build a deep-learning framework that can predict the performance level of apple grafts and to see whether the explanations given by the model agree with simulated farming rules.*

Methods: *A simulator that uses the climate and soil conditions of Injibara District, Ethiopia (altitude 2,500–2,700 m and 400–800 chill hours) was made. The simulator produced 1,500 apple trees from five apple cultivars. The data set contained tables of numbers a 52-week series of plant growth and a spot for a picture. Each tree was put into one of five performance levels. After checking for and fixing any data-leak problems a neural network called GraftNet was tested with five random starts. Its results were compared with those from machine-learning methods. Two ways of looking at the model's explanations—SHAP and grouped permutation importance—were used.*

Results: *GraftNet gave a weighted kappa of 0.954 ± 0.003 an accuracy of 0.799 ± 0.011 and a balanced accuracy of 0.775 ± 0.018. However GraftNet did not do better than a forest or a logistic regression. Adding the 52-week growth series did not help the classification. A simpler method, ridge regression predicted how fast the trees would grow accurately (RMSE 1.7 cm versus 6.1 cm). The important factors were the chilling requirement and the chill deficit. Variables that were not part of the simulator's rule set were almost unimportant. GraftNet worked well on the simulated data but it did not give an edge over traditional machine-learning models. The work shows that explainable machine learning can help explore apple graft performance. Still more tests are needed with nursery and field data, from many apple varieties, rootstocks and different farming areas.*

Keywords: Apple; graft performance; Ethiopian highlands; chilling requirement; deep learning; ordinal classification; SHAP; explainable artificial intelligence.

1. Introduction

Apple (Malus domestica Borkh.) is a fruit crop that's very important for the economy. Apple has a lot of potential for adding variety to fruit farming. Apple can help make life better in the Ethiopian highlands. The successful growing of apple in tropical highland areas depends on things. Apple varieties must work well with winter cold. The choice of rootstock for apple is also very important. Soil conditions for apple also matter a deal. Earlier research in Ethiopia has shown that apple varieties differ in their timing of growth. Apple timing can vary with flowering. Apple fruit yield can also differ across highland areas. These differences in apple show that it is important to choose the apple variety, for each local environment (Melke and Fetene, 20s14; Melke, 2019)**.** These considerations are especially important for Injibara District, in Northwest Ethiopia because the high altitude environment gives a chance to look at how temperate fruit production might grow. I see that the fit of each cultivar and rootstock combination must be checked locally of just using broad agro-ecological recommendations. In Injibara District the high altitude. Temperate fruit production require evaluation of cultivars and rootstock combinations instead of just using broad agro-ecological recommendations.

Winter chilling is one of the important environmental factors that control when apple trees come out of dormancy start to break buds, flower and develop fruit. Different apple varieties have chilling needs. When apple trees do not get winter chilling it can lead to late and uneven bud break longer flowering periods, weak fruit set and lower overall productivity (González Noguer *et al*., 2023). The Cold Hours, Utah and Dynamic models are often used to calculate how much cold weather accumulates during the winter even though their results can be different because of how the temperature changes, the type of plant and the way cold's measured (Richardson et al., 1974; Fishman et al., 1987; González Noguer et al., 2023). In addition to the climate being right choosing the rootstock affects how the upper part of the tree grows how it takes in nutrients, how it handles water and how it produces flowers and fruit. Research done in Ethiopia has shown that apple rootstocks such, as MM106, MM111 and M9 react differently to soil that is too acidic.. Managing the pH of the soil by adding lime can affect how the tree grows and how much fruit it produces (Kebede *et al*., 2016). Therefore it is important to consider the requirements of the apple varieties the characteristics of the rootstocks and the soil management conditions when looking at how well apple trees will grow through grafting in new areas like Injibara. Apple varieties need amounts of cold to grow properly. Rootstocks affect how the tree grows and how it handles the soil. Soil management includes things, like keeping the soil healthy and well-drained. All these things matter when trying to grow apples in places where apple growing's new. Injibara is one of those places where people are starting to grow apples. So looking at all these factors helps to know if the grafting will work well.

Apple production in Ethiopian highland environments is becoming more popular. However apple production still has limited seasonal data to predict how different cultivars and rootstocks will perform. Current suggestions for apple production usually rely on cultivar screening data and only a few site-specific observations. In addition the agronomic guide that is used in this apple production study states that its cultivar performance rankings are provisional and need verification. Another problem for apple production is the lack of a labelled nursery-level dataset that has systematically measured graft performance results. Because of these gaps in apple production knowledge there is a need for reproducible computational methods that can gather current apple production information explore relationships between important

apple production variables and propose hypotheses, for future field and nursery testing of rootstock performance.

Machine learning (ML) offers chances to study connections among environmental, physiological and management variables. Machine learning (ML) also helps build models that support decisions in agriculture. Multimodal learning is a type of Machine learning (ML) that can merge kinds of information. This includes agronomic variables, phenological time-series observations and images all, within one unified predictive framework (Baltrušaitis *et al*., 2019)**.** Gated multimodal architectures let you learn how much each input modality matters when you mix them together. Gated multimodal architectures help decide which input modalities are more important, during information fusion (Arevalo *et al*., 2017). The success of these methods relies on having input data. Both, in amount and quality. It's important to test them against machine learning models to see how they really do. In fact past tests have shown that tree-based ensemble methods often match or even beat learning models when working with medium-sized tabular data (Grinsztajn *et al*., 2022; Shwartz-Ziv and Armon, 2022).

Beyond accuracy it is very important that the model can be understood when machine-learning applications are used to help make decisions in farming. Nursery managers, horticultural experts and extension professionals need to know which specific cultivar, climate, rootstock and soil factors affect the predictions made by the model. Explainable artificial intelligence (XAI) methods, SHapley Additive exPlanations (SHAP) show the contributions of each feature. These contributions help explain predictions and the overall behavior of the model (Lundberg and Lee, 2017). Permutation-based feature importance gives another way to understand which features matter most. It looks at how the model works when the values of a feature are changed.. Both methods need to be thought about carefully. This is especially true when the features are linked or have information (Hooker *et al*., 2021)**.** Model explanations should not be automatically taken as proof of relationships or, as confirmation of agronomic mechanisms. Just because a model points to a factor doesn't mean that factor is actually causing the outcome. We need to be careful not to assume causality just because a model highlights a connection. The same goes for mechanisms—just because a model suggests a process doesn't mean it reflects whats really happening in the field. These explanations are helpful. They must be tested and validated through real-world experiments and careful analysis (Rudin, 2019). Evaluating explanation methods, against datasets that have explicitly known data-generating mechanisms gives a chance to see if explanation methods can find the relationships that are built into the underlying data.

To tackle these challenges we tested a repeatable rule-based simulation and machine-learning system that predicts apple graft performance in the agro-ecological setting of Injibara District, Northwest Ethiopia. Because there was no labelled nursery-level data set we made a synthetic data set of 1,500 virtual grafted trees. We used information on how different apple cultivars perform and agricultural rules taken from an existing guide. The simulated data set included agricultural variables 52-week phenological time-series observations and a spot for image input. We did a data leakage audit to find and fix any transfer of information between the time-series variables and the performance labels before evaluating the model. GraftNet, a gated three-branch network, with an ordinal prediction head was then compared with standard machine-learning and statistical models. We also used explainability methods to look at whether the

important predictors matched what we know from the simulator. So this study gives a controlled assessment, not a field validation, of apple graft performance.

The specific contributions of this study are as follows:

1. **Development of a simulation pipeline:** A rule-based framework was developed to transform information from a Word-format agronomic guide into a synthetic dataset of 1,500 virtual apple grafted trees incorporating structured agronomic variables and phenological time-series observations together with a data leakage audit and correction procedure.

2. **Development of GraftNet:** A gated three-branch neural network with a prediction head was developed to investigate the integration of heterogeneous agricultural information for predicting five-tier apple graft performance.

3. **Model evaluation:** GraftNet was benchmarked against conventional machine-learning and statistical approaches using consistent data partitions and evaluation procedures for ordinal classification and growth-kinetics prediction.

4. **Evaluation of model explainability:** SHAP and grouped permutation importance were applied to examine predictors, assess agreement, between explanation methods and evaluate whether model explanations recovered the known relationships encoded in the simulated dataset.

The study was especially relevant to the emerging potential for apple production in Injibara District and other Ethiopian highland environments. However its findings should be understood within the limits of the simulated data and the assumptions made in the guide used. The study does not confirm graft compatibility, cultivar adaptation or real field productivity under Injibara conditions. Instead it offers a computational framework for organizing agronomic knowledge assessing predictive modeling methods and highlighting key areas for future nursery and field research. Independent validation using grafting results actual phonological observations and image data, across multiple cultivars, rootstocks, growing seasons and agro-ecological zones is necessary to determine its practical value.

2. Background and related work

2.1. Apple chilling requirements and graft performance

Apple is a temperate fruit tree whose yearly growth cycle is guided by the mix of winter cold and later warmth. In winter low temperatures help Apple to break out of dormancy. Then when spring heat builds up enough Apple will start buds, flower and grow further. Apple varieties differ a lot in how chill and heat they need. Choosing Apple varieties that match the climate is key, for a successful orchard (González Noguer *et al*., 2023). The Chilling Hours model calculates how much chilling builds up by looking at temperatures that're between 0 and 7.2 degrees Celsius. It focuses on this range because that’s when the chilling effect is most effective. Temperatures, outside this range don’t count toward the total. This method is simple but widely used in agriculture to predict when trees or plants will break dormancy. It helps farmers plan for bloom times and manage crops better (Weinberger, 1950), whereas the Utah model assigns

different chilling efficiencies to temperature ranges and incorporates the negative effects of relatively warm temperatures (Richardson *et al*., 1974). The Dynamic model uses a two-stage process to estimate accumulation, in chill portions and the Dynamic model carefully estimates chilling accumulation to determine chill portions (Fishman *et al*., 1987). These models are used widely. However cultivar-specific differences and climatic variations can affect the accuracy of these models. This effect is especially strong, in regions that have mild winters (Luedeling, 2012; González Noguer *et al*., 2023). We think published chilling requirements are best used as clues, to whether a cultivar will thrive, not as hard rules that fit every situation. Chilling requirements should guide, not dictate.

In highland areas how well apple cultivars grow depends on the mix of altitude, temperature, chilling hours and how farmers manage their orchards. These factors work together to shape the success of apple trees in these regions. (Melke and Fetene (2014) emphasized that cultivar selection, in highlands should consider chilling requirements. Melke and Fetene (2014) also stated that inadequate chilling can lead to bud break, irregular flowering and irregular fruit development. Similarly Ethiopian studies have shown that the performance of cultivars and their adaptability can vary in production environments. This confirms that location-specific evaluation of crops is essential (Melke, 2019). Rootstock selection is another consideration. Rootstocks influence scion vigor. Rootstocks influence how scions acquire nutrients. Rootstocks influence water relations of scions. Rootstocks influence development of scions. Rootstock choice matters a lot In Ethiopia, Jobir et al. (2016) found that the growth and fruiting of Anna and Jonagored apple varieties changed depending on which rootstock they were grafted onto. The rootstocks used were MM111, MM106 and M9. These results show how important it is to look at the features of the rootstock and how the soil is managed when checking how well apple trees grow in highland areas. The study highlights the role of rootstock characteristics and soil conditions when testing apple grafts in highland farming.

2.2. Multimodal learning and information fusion

Multimodal machine learning is about mixing data from different sources so that a single model can learn from all of them. The data can be numbers that are neatly organized, sequences of numbers that change over time or pictures. Each type of data gives its clues, about what will happen and when these modalities are used together Multimodal machine learning can make better predictions and decisions (Baltrušaitis *et al*., 2019). Structured variables can show conditions and management conditions and they are useful for understanding plant development. Phenological time series can show changes in plant development over time. They help us see how plant development moves through seasons. Images can show information about plant growth, graft development and physiological responses and they give us a clear picture of how plants look and react. When we integrate variables phenological time series and images we may improve predictive performance. Each modality has complementary signals and combining the modalities gives us better predictive performance because the signals, from each modality add up.

Gated multimodal architectures offer a way to combine information from sources. They learn how each type of input, like text, images or sound adds to an understanding. The model decides which parts of the input are important and how they should be used together. This helps in building an more accurate representation. (Arevalo *et al*. (2017) introduced gated multimodal units that use gates to regulate the contribution of individual modalities during neural network learning. Multimodal units give a way to combine different kinds of agricultural data.. The success of multimodal fusion depends on how good, complete and useful each input modality is. Adding modalities does not always make the model work better especially when the data set is small, noisy or has redundant information. So we need to compare multimodal and unimodal setups to see if combining information really gives a predictive benefit.

2.3. Machine learning for structured agricultural data

Machine-learning approaches are being used more and more for classification, regression, forecasting and decision support in research. Conventional algorithms such as regression, random forests, gradient boosting and multilayer perceptron give different ways to find links between predictor variables and target outcomes. Random forests can especially find links and interactions among predictor variables by using ensemble learning. Deep learning methods give shapes for modelling complex patterns but often need larger data sets and a lot of work, on model tuning.

The performance of machine learning methods versus deep learning methods depends on how big the data is, what kind of predictors are used, what the model looks like and how it is trained. Different data sizes can make one approach better, than the other. The features that the predictors have also matter a lot. The way the model is built and how it is improved during training are important too. All these factors together decide which method works better. Grinsztajn *et al*. (2022), Used a benchmark of 45 datasets. Showed that tree-based models often keep strong predictive performance on medium-sized tabular datasets and tree-based models remain reliable, on such data. Similarly, Shwartz-Ziv and Armon (2022) showed that tree-based ensemble methods can beat many deep learning approaches when the data is structured. These results remind us that we must test deep learning models against methods before we think that more complicated architecture means better prediction. In this paper we compare GraftNet with regression, random forest, gradient boosting, multilayer perceptron and regression-based approaches to see how well GraftNet performs compared to these established techniques.

2.4. Explainable artificial intelligence and validation of model explanations

Explainable artificial intelligence tries to make it easier to understand how machine-learning models make predictions. SHapley Additive exPlanations, also known as SHAP was created by Lundberg and Lee (2017). SHAP offers a way to figure out how each input feature affects the results a model gives. SHAP values can help look at each prediction and also see which features are most important overall. Permutation importance is another way to check how important a feature is, by seeing what happens to the models performance when the values of that feature are mixed up randomly. Using explanation methods together can help check if different models agree on which features matter the most.

Feature-importance methods have limitations and feature-importance methods are especially limited when predictor variables are strongly correlated. Hooker et al. (2021) showed that permutation-based methods that are not restricted can give importance scores when the relationships, between features are broken during the permutation process. Also when there are predictors that are related to each other it becomes hard to see the effect of each individual variable. So when looking at model explanations it is important to think about how the data's structured and how the features are connected. Also feature importance shows how variables are connected to predictions and how the model works. It does not always mean that it has found a cause-and-effect relationship. Rudin (2019) said that it is very important to tell the difference between explanations of models and approaches that are naturally understandable. In decision-making applications explanations of models must be clear and reliable. The same is true, for interpretable approaches because people need to trust and see how decisions are made.

I think one way to look into the reliability of explanation methods is to test those explanation methods on datasets that have clearly defined ways of making data. In those settings the factors that affect the result are known by design letting researchers look at whether explanation methods can find the expected connections. Consistency between explanation methods can give extra proof of interpretability though it does not on its own prove agricultural or biological truth. This way is especially useful when real data, with labels is not available and when the goal is to check method features before doing real-world testing.

In this study SHAP and grouped permutation importance were used to check how consistent the important predictors were in the simulated apple graft performance dataset. The explanations were compared to the rules that were used to create the performance labels. Special focus was placed on requirements, chill deficit, traits related to the apple cultivar, soil pH, lime application and rootstock effects. This method gives a controlled way to assess how interpretable the model is. It also helps to spot issues that come from correlated predictors and differences, in how models behave. However the explanations reflect how well the model follows the simulation rules. They should not be seen as proof of the agronomic processes that affect apple graft performance.

3. Materials and methods

3.1. Knowledge base

Knowledge base explained apple growing conditions in Injibara District, Awi Nationality Administrative Zone, Amhara Region, Ethiopia. The collected agro ecological profile showed an elevation of 2,500–2,700 m a temperature of 10–18 °C, 400–800 chill hours yearly rainfall of 1,400–2,200 mm and soils that are moderately acidic Nitisols and Alisols (pH 5.0–6.0). The production guide listed five apple varieties. Ranked them based on expected apple performance, in this area (Table 1).

The script pulled cultivar information from the guide. The script used the guide as the source, for chilling requirements, performance tiers and disease-susceptibility flags.

Table 1. Cultivar characteristics extracted from the apple production guide

Cultivars	Chilling requirement (h)	Performance tier (ordinal)	Susceptibility flags
Anna	200–300	Optimal (4)	Fire blight; powdery mildew
Bonded Red	300–500	Moderate (3)	Sap-sucking vectors
Crispin	500–600	Suppressed (2)	Blister spot; apple scab
Jonagold	600–700	Severe decrease (1)	Fire blight; storage disorders
Granny Smith	700–800 (open-ended “700+” capped at 800)	Lowest index (0)	Powdery mildew; apple scab

3.2. Multimodal Dataset and Experimental Design

A dataset contains 1,500 trees. There are 300 trees for each cultivar. The random seed used is 2025. The variables include rootstock type. Rootstock type has MM106 with 0.60, M9 with 0.25 and seedling with 0.15. The variables also include cultivar chilling requirement. The site chill ranges from 400 to 800 hours. The soil pH ranges from 5.0 to 6.0. The lime amount ranges from 2 to 3 kilograms. The compost amount ranges from 10 to 15 kilograms. The NPS amount ranges from 200 to 300 grams. The urea amount ranges from 150 to 250 grams. The elevation ranges from 2,500 to 2,700 meters. The rainfall ranges from 1,400 to 2,200 millimeters. The graft-union height ranges from 10 to 15 centimeters. Chill deficit is calculated as site chill, minus cultivar chilling requirement.

The ranking for how well each tree performed was created using a method based on the type of tree and a measure of how stress it experienced. The score had penalties added for problems. There was a penalty of 1.0 if the tree did not get cold weather 0.5 if the soil was too acidic 0.3 if there was not enough lime in the soil and 0.4 if the roots were not healthy. Some random numbers were added to the scores to make them more realistic. Then the scores were rounded and limited to a range of 0 to 4. The final data had 566 trees in the performance group 211 in the next 218 in the third 289 in the fourth and 216 in the highest group. The average performance, for each type of tree matched what the guide had predicted: Anna trees had an average of 3.70 Bonded Red trees had 2.67 Crispin trees had 1.31 Jonagold trees had 0.26. Granny Smith trees had 0.00.

Tree growth followed a curve over fifty-two weeks using an asymptote (K) a growth rate (r) and a midpoint (t0). The K parameter included a label effect, management effects, rootstock adjustments, soil-pH effects, chill-deficit effects and Gaussian noise (SD = 22 cm). In the corrected dataset the growth rate (r) and the midpoint (t0) depended on nitrogen supply and chill deficit of the performance tier. Eight weekly channels were added: plant height, girth, graft-union diameter, node count, chlorophyll content (SPAD) temperature, cumulative chill units and callus score. The image modality contained placeholder tensors, specifically random noise for fifty trees and zeros, for the remaining trees and therefore supplied no meaningful image information.

The dataset was split into three parts using an approach: 80 percent for training 10 percent for validation and 10 percent for testing. That means 1,200 trees were used for training 150 for validation and 150, for testing. The split was done in a way that kept the distribution of cultivars balanced across all subsets. No cultivar was left out. Separated during the splitting process. This helps ensure that the model sees all types of trees in each phase.

3.3. Leakage audit and development history

The development process involved two corrections. At first K was completely determined by the performance tier, which made the model favor extreme classes. To fix this the K-generation rule was updated, which reduced the correlation between labels and K to 0.431. Later a leakage audit revealed connections, between weekly SPAD values, growth parameters and performance tiers. In the dataset SPAD was partly generated from the tier and both r and t0 also relied on the tier. Because of this a random forest model using SPAD summaries achieved an accuracy of 1.000 showing clear target leakage.

These relationships were. The dataset was regenerated as described in Section 3.2. In the corrected dataset the SPAD- classification reached a balanced accuracy of 0.27. Models that used summaries of all channels achieved about 0.31 to 0.35 compared to the chance level of 0.20. All results mentioned in Section 4 were based on the corrected dataset. The earlier results were found to belong to the simulator version.

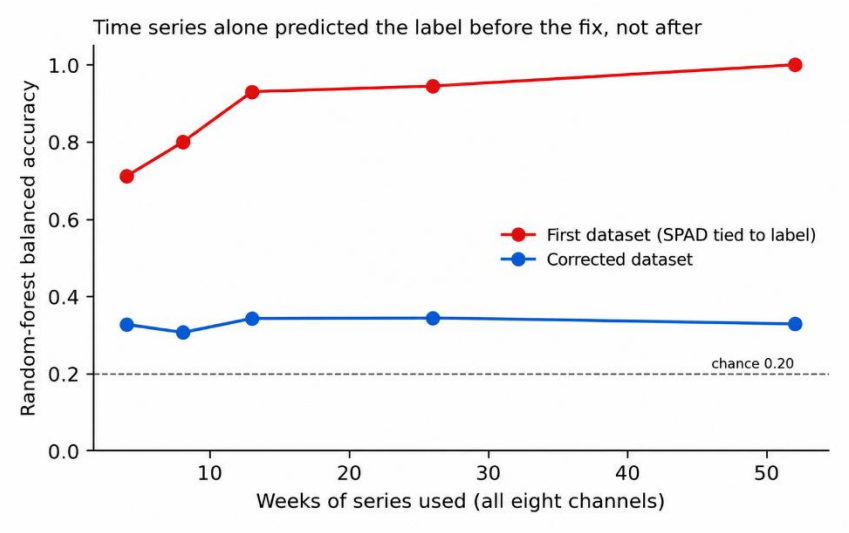


Figure 1. Leakage audit comparing random-forest classification based on weekly-series summaries before and after dataset correction. The corrected data produced substantially lower balanced accuracy, approaching the chance level of 0.20.

3.4. GraftNet architecture

GraftNet had three encoders, for different types of input data: tabular, temporal and image. The tabular part took in 25 features. Passed them through a two-layer multilayer perceptron. Each layer had 128 hidden units. There was a linear skip connection to help with information flow. The temporal branch used two one-convolutional layers with a kernel size of 5 and 64 channels each. After that it applied average pooling to reduce the data. The image branch used three layers to process the image inputs. Each of these three encoders produced a 64- feature representation.

A gated fusion layer calculated modality weights using half of a softmax of learned weights plus a half uniform prior guaranteeing that each modality receives least one sixth of the weight. The fusion layer output was projected to ninety six dimensions. Then sent to two output heads: an ordinal classification head with four cumulative binary classifiers and a regression head, for predicting K and r. The fusion layer has one hundred five thousand five hundred thirteen trainable parameters.

3.5. Training and calibration

GraftNet was trained with the AdamW optimizer. The base learning rate was set to 0.001. The weight decay was set to 0.0001. The batch size was set to 64. Learning-rate multipliers of 5 and 0.5 were applied to the tabular branch and the image branch, respectively. Training ran for a maximum of 100 epochs. Early stopping was used after 15 epochs with no improvement. Class-weighted loss was used. Cultivar-feature dropout was set to 0.10. Branch dropout was set to 0.15. Regression-loss weight was set to 0.5. All training

was performed on a CPU. The regression targets, K and r were standardised using the training-set means and standard deviations. Model checkpoints were chosen using a combined validation score.

The combined validation score was calculated as QWK. 0.25 Times the sum of R^2 for K and R^2 for r. This selection criterion was chosen because using QWK alone produced insufficient training for the regression head. Five random seeds, numbered 0 to 4 were used. For each seed separate stratified splits of 80 % training 10 % validation and 10 % test were created. The best epochs for the five seeds were 87, 6 84 41 and 45. Probability calibration was performed with temperature scaling, on each validation set. Expected calibration error, ECE was computed using 10 bins.

3.6. Baseline models

Classification performance was compared against regression, random forest (300 trees with balanced class weights) histogram gradient boosting, MLP (128 and 64 hidden units) and majority-class predictor. Each classifier used either tabular features tab) or tabular features combined with summary statistics of all eight weekly channels (tab+series). The summary statistics included the value, mean, maximum and standard deviation. Regression baselines, for K and r included a predictor, ridge regression, random forest and MLP. All baseline models were evaluated using the five stratified data splits as GraftNet. Series- baseline models were not rerun after dataset correction.

3.7. Explainability and cross-method validation

SHAP values were calculated using a kernel estimator for the seed-0 GraftNet model. The output being explained was the expected performance tier, which was determined as the probability-weighted mean of five classes. The analysis relied on 25 tabular features. The temporal input was set to the training median. The image input was set to zero. The method used 50 background trees, 100 test trees, 300 coalition samples and three repeated runs. Spearman correlation across the repeated runs was used to assess explanation stability.

Grouped permutation importance was computed as the drop in accuracy, after shuffling each feature group. Seven feature groups were evaluated using both forest and MLP models across five data splits. Each group was shuffled 10 times. Weekly time-series features were excluded from this analysis.

The explanation methods were tested against the simulator's known label-generation rules. Positive controls included cultivar tier, requirement, soil pH, lime and rootstock. Negative controls included elevation, rainfall, graft-union height, NPS and urea. Higher importance assigned to controls compared to negative controls was taken as evidence that the explanation methods aligned with the simulator's label-generation structure.

3.8. Evaluation metrics

We measured classification performance using accuracy, accuracy, macro-F1, quadratic weighted kappa (QWK) and ordinal mean absolute error (MAE). We measured regression performance using root square error (RMSE) and coefficient of determination (R^2) for K and r. We measured calibration performance using expected calibration error (ECE). We reported results as or, minus standard deviation across five

seeds or paired splits. For the GraftNet model we also calculated 95 percentile bootstrap confidence intervals using 5 000 resamples of the 150 test trees.

4. Results

4.1. Classification performance

Across five seeds GraftNet achieved an accuracy of 0.799 ± 0.011. GraftNet also reached an accuracy of 0.775 ± 0.018, a macro-F1 of 0.765 ± 0.014, a QWK of 0.954 ± 0.003 and an ordinal MAE of 0.204 ± 0.011. 99.7% Of predictions fell within one performance tier of the actual class. Most classification errors occurred between neighbouring tiers. Random forest, regression, gradient boosting and MLP models achieved comparable or slightly better results using tabular features alone (Table 2 and Figure 2). Random forest recorded the accuracy of 0.807 ± 0.021 and the highest QWK of 0.956 ± 0.006. Adding time-series summaries did not consistently improve classification performance. The small differences between GraftNet and tabular baselines did not demonstrate an advantage for the deep-learning model. For the GraftNet model 120 of 150 test trees were correctly classified. The bootstrap 95% confidence intervals were 0.733–0.860 for accuracy 0.685–0.836 for accuracy and 0.929–0.968, for QWK. Class-specific performance and the confusion matrix are presented in Table 3. Figure 3.

Table 2. Classification performance on the corrected dataset (mean ± SD across five seeds)

Model	Features	Accuracy	Balanced accuracy	QWK
GraftNet (calibrated)	tab+series+image	0.799 ± 0.011	0.775 ± 0.018	0.954 ± 0.003
Random forest	Tab	0.807 ± 0.021	0.779 ± 0.024	0.956 ± 0.006
Random forest	tab+series	0.804 ± 0.024	0.775 ± 0.027	0.955 ± 0.005
Logistic regression	Tab	0.800 ± 0.014	0.776 ± 0.019	0.954 ± 0.005
Logistic regression	tab+series	0.781 ± 0.020	0.758 ± 0.027	0.942 ± 0.013
HistGB	Tab	0.791 ± 0.029	0.740 ± 0.033	0.950 ± 0.009
HistGB	tab+series	0.803 ± 0.038	0.760 ± 0.046	0.954 ± 0.011
MLP	Tab	0.791 ± 0.012	0.746 ± 0.020	0.952 ± 0.002
MLP	tab+series	0.781 ± 0.030	0.739 ± 0.037	0.944 ± 0.018
Majority class	Tab	0.379 ± 0.003	0.200 ± 0.000	0.000 ± 0.000

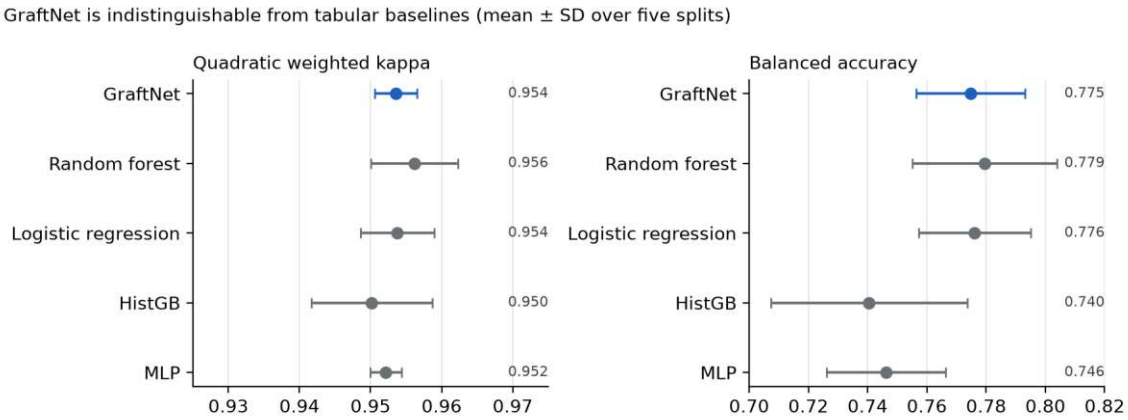


Figure 2. QWK and balanced accuracy of GraftNet and baseline classifiers across five splits

Table 3. Seed-0 GraftNet classification performance and confusion matrix (test n = 150)

Tier	Precision	Recall	F1-Score	Support	Predicted 0	1	2	3	4
0 (Lowest)	0.93	0.93	0.93	57	53	3	1	0	0
1 (Severe decrease)	0.64	0.67	0.65	21	4	14	3	0	0
2 (Suppressed)	0.71	0.68	0.70	22	0	5	15	2	0
3 (Moderate)	0.78	0.72	0.75	29	0	0	2	21	6
4 Optimal	0.74	0.81	0.77	21	0	0	0	4	17

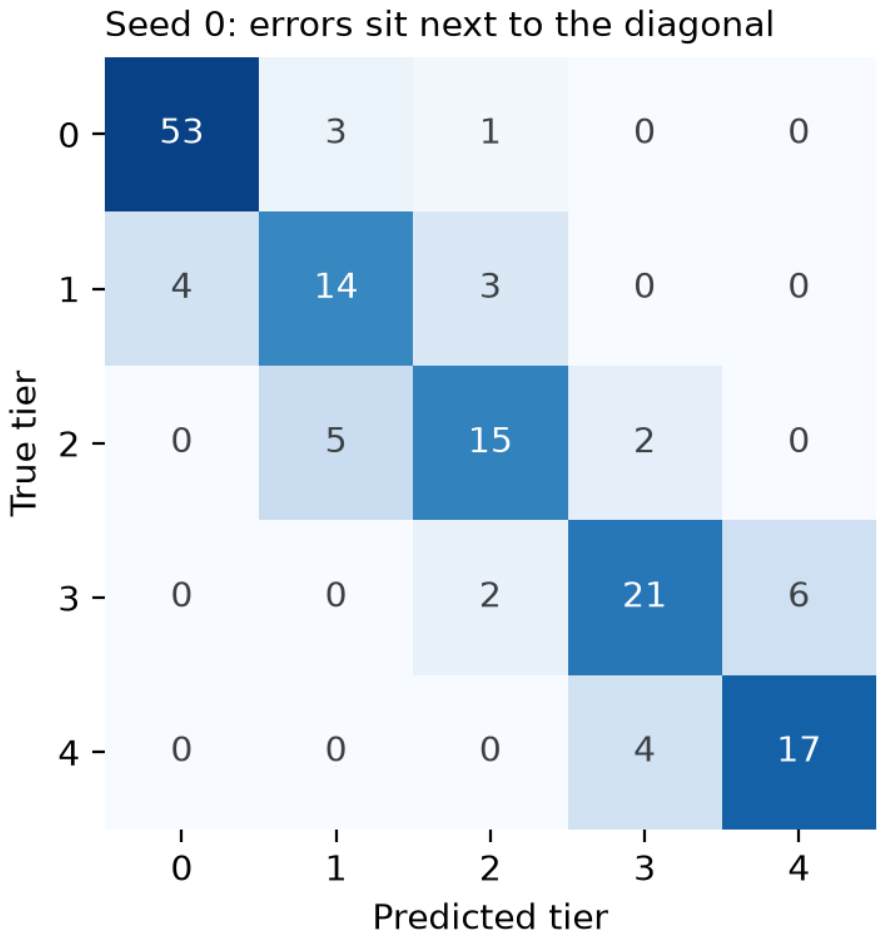


Figure 3. Figure 3. Confusion matrix of the calibrated seed-0 GraftNet model

4.2. Growth-kinetics regression

I tested GraftNet across five seeds. Found that GraftNet had an RMSE of 6.13 ± 0.36 cm and an R^2 of 0.952 ± 0.007 for K. When I looked at r I saw that r had an RMSE of 0.0050 ± 0.0020 and an R^2 of 0.67 ± 0.28. R prediction changed a lot from seed to seed and seed 1 was very bad with an R^2 of 0.12. Ridge regression with full-series summaries performed better giving an RMSE of 1.66 ± 0.12 cm for K. An RMSE of 0.0024 ± 0.0002 for r (Figure 4). Random forest and MLP also beat GraftNet's regression heads. These findings show that simpler models are better, at predicting the growth parameters. Still the high regression scores mainly show that the models can recover the parameters that were used to create the growth curves not that they can predict growth outcomes that were measured separately.

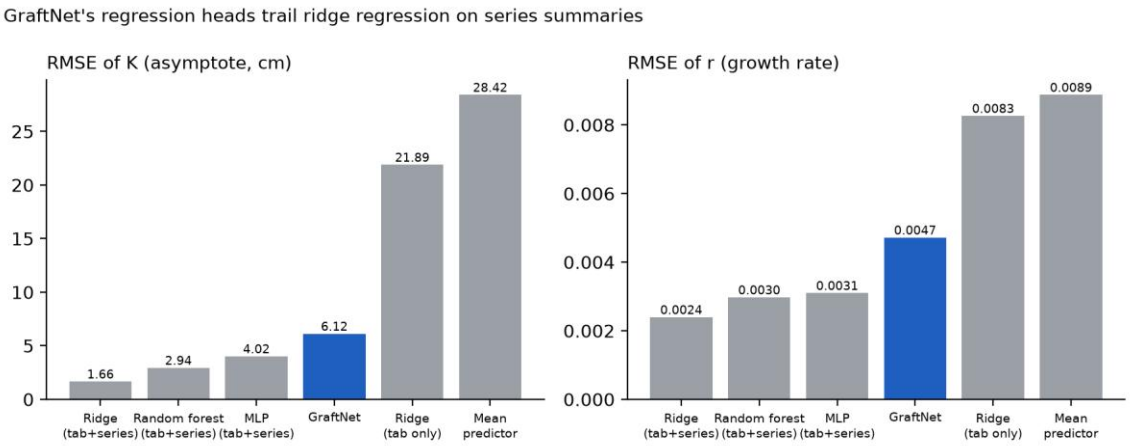


Figure 4. Test RMSE for growth asymptote (K) and growth rate (r) across GraftNet and regression baselines.

4.3. Calibration

Before temperature scaling GraftNet recorded an ECE of 0.108 ± 0.028 showing that the predictions were too confident. After temperature scaling the ECE dropped to 0.064 ± 0.012. The mean scaling temperature was 1.88 ± 0.33. However the calibration estimates were based on 150 test trees, per split so they require careful interpretation.

4.4. Modality contribution

The mean fusion weights were 0.465 ± 0.038 for tabular data 0.353 ± 0.026 for time-series data and 0.183 ± 0.019, for image data as shown in Figure 5. The image weight stayed near the minimum of 1/6 because the image inputs were just placeholders. Although the temporal branch had a high fusion weight the time-series features did not always improve classification performance. The contribution of time-series features seemed to matter more for growth-parameter prediction. No branch-ablation test confirmed this interpretation. Therefore the real performance of image-based and multimodal methods was not tested.

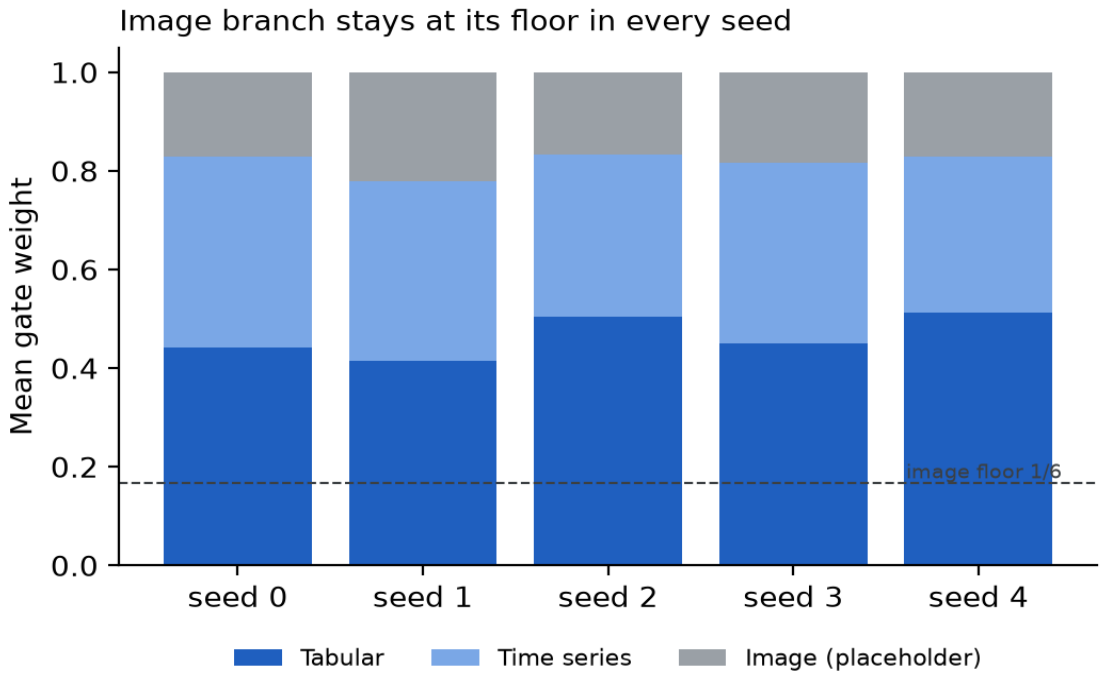


Figure 5. Mean fusion-gate weights across five random seeds

4.5 Explainability results

SHAP analysis: The SHAP analysis using seed-0 showed that the rankings stayed very stable across three repetitions with Spearman correlation values, between 0.96 and 0.98. Chilling requirement had the mean absolute SHAP value at 0.410 followed by chill deficit at 0.258. These were the two variables. Disease-susceptibility indicators came next in importance. Soil pH ranked sixth. Lime ranked eighth. Variables that were not part of the label-generation rule generally fell between 17th and 25th in importance. Their impact was much lower compared to the variables included in the rule. Rootstock effects were only partially captured: M9 ranked 12th seedling ranked 16th and MM106 ranked 24th. These results mostly matched the effects set by the simulator. However the SHAP analysis only looked at tabular features. The temporal inputs were fixed at their training median so the analysis did not explain the branch.

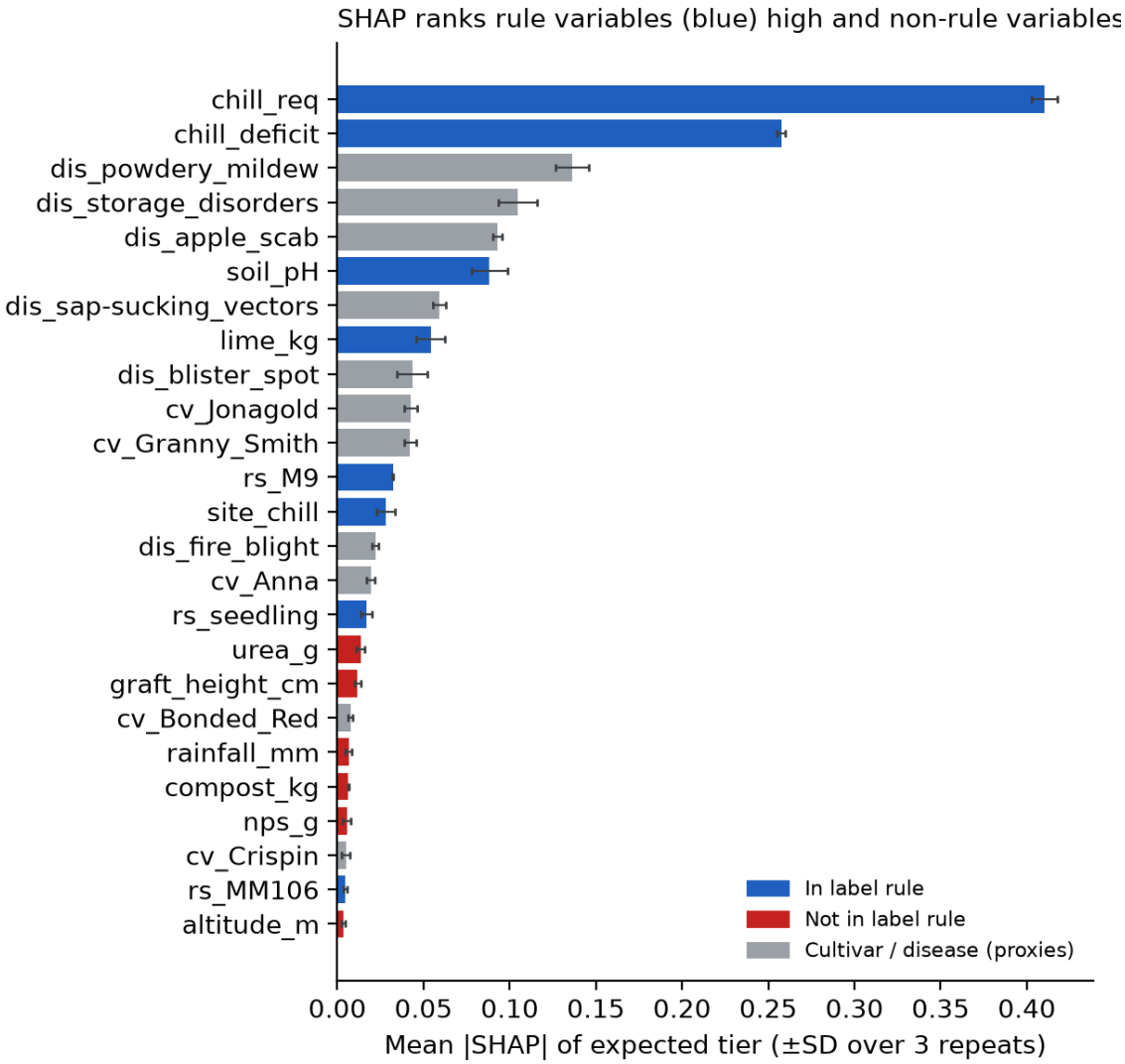


Figure 6. Mean absolute SHAP values for tabular features in the seed-0 GraftNet model

Permutation importance: Grouped permutation analysis found that chilling variables were the important group for random forest, with a value of 0.238 ± 0.046. The next important groups were disease flags, cultivar identity and soil and liming variables. For MLP disease flags had the importance, 0.197 ± 0.040 followed by cultivar identity at 0.174 ± 0.019. Chilling variables came in third for MLP. Site and fertilizer variables had importance values to zero for both random forest and MLP as shown in Table 4 and Figure 7.

Table 4. Grouped permutation importance (change in balanced accuracy, mean ± SD across five splits)

Feature group	In label rule?	Random forest	MLP
Chilling (requirement, site, deficit)	Yes (weight 1.0)	+0.238 ± 0.046	+0.118 ± 0.022
Disease susceptibility flags	No (cultivar proxy)	+0.062 ± 0.012	+0.197 ± 0.040
Cultivar identity	Yes (baseline tier)	+0.056 ± 0.011	+0.174 ± 0.019

Feature group	In label rule?	Random forest	MLP
Soil and liming (pH, lime, compost)	Yes (weights 0.5, 0.3)	+0.045 ± 0.023	+0.034 ± 0.025
Rootstock	Yes (weight 0.4)	+0.018 ± 0.012	+0.035 ± 0.010
Site (altitude, rainfall, graft height)	No	−0.011 ± 0.003	+0.002 ± 0.009
Fertilizer (NPS, urea)	No (series only)	+0.002 ± 0.009	0.000 ± 0.007

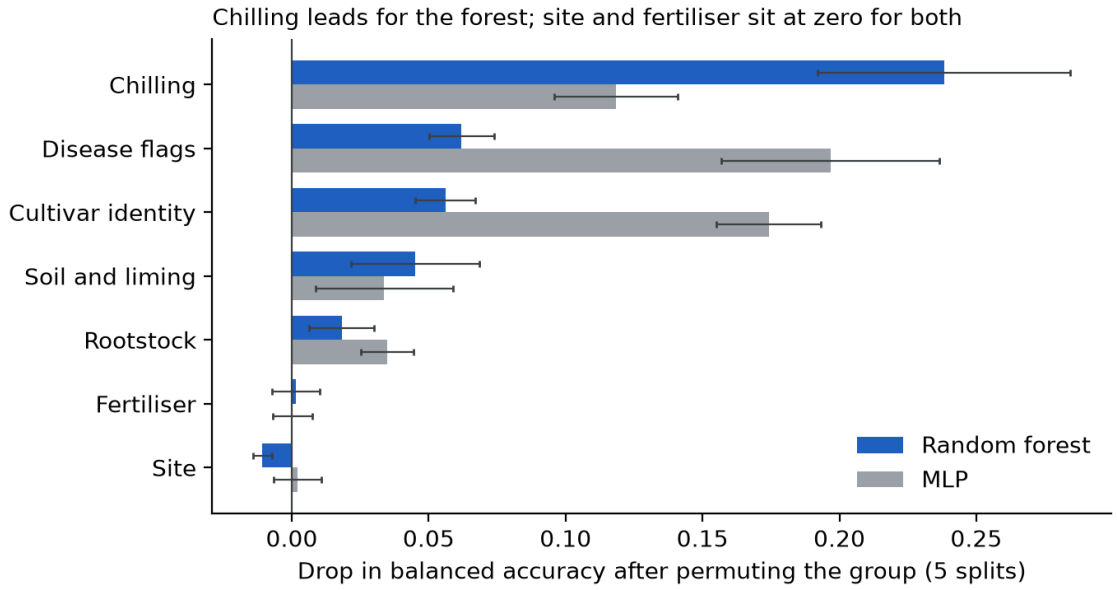


Figure 7. Grouped permutation importance for random forest and MLP models

4.6. Cross-method validation

Comparison of SHAP and permutation importance showed that both methods agreed broadly with the simulator's label-generation rules (Table 5). Requirement and chill deficit were consistently ranked high in importance. Soil and liming variables were given importance. Variables that were not part of the label rule generally had importance. However the MLP model ranked cultivar and disease-susceptibility indicators than chilling variables. This was different from what SHAP and random forest found. Rootstock effects were also not consistently captured, for MM106.

Overall the explanation methods did capture key relationships from the simulator.. The feature rankings differed across models. This difference came from dependencies, between cultivar, requirement and susceptibility indicators.

Table 5. Comparison of explanation methods with the simulator's known label-generation rules

Mechanism	Empirical Data	GraftNet SHAP	RF permutation	MLP permutation	Verdict
Chilling deficit and requirement	Strongest rule term	Rank 1–2	Rank 1	Rank 3	Recovered; order model-dependent
Cultivar and susceptibility flags	Sets baseline tier	Flags rank 3–14; one-hots rank 10–23	Flags rank 2, cultivar rank 3	Flags rank 1, cultivar rank 2	Recovered, but confounded (flags are cultivar proxies)
Soil pH and lime	Small penalties	Rank 6 and 8	Rank 4	Rank 5	Recovered
Rootstock	Small penalty	M9 rank 12, seedling 16, MM106 24	Rank 5	Rank 4	Weakly recovered
Altitude, rainfall, graft height	Absent from label rule	Rank 18, 20, 25	About zero	About zero	Agree
NPS, urea	Affect series, not label	Rank 22 and 17	About zero	About zero	Agree

5. Discussions

The main contribution of this study is methodological. It focuses on identifying and fixing data leakage. It also involves an evaluation of model explanations against known data-generating mechanisms. In the corrected simulated dataset SHAP analysis of the network and grouped permutation importance across two model families consistently highlighted chilling-related variables as the most influential predictors. Variables that were not part of the label-generating rules received importance. This agreement, between two explanation methods—SHAP and grouped permutation importance—provides strong evidence that the models captured the key predictive relationships that were built into the simulator. SHAP offers a way to explain how each feature contributes to a models prediction Permutation importance looks at how important a feature's by checking how much the models performance drops when the features values are shuffled.. These methods show how the model works not what causes what. Soil pH and lime application had effects in the simulation and they ranked in the middle with SHAP, which matches how much they influenced the results. Still SHAP didn't capture the effect of rootstock MM106, which had the effect, in the simulation data.

These findings show again how important it is to look at feature dependencies when we read machine-learning explanations. Susceptibility indicators seemed strong showing up as one of the top feature groups for the multilayer perceptron (MLP) and falling between third and fifth in SHAP rankings. Yet those susceptibility indicators were simply functions of cultivar identity, in the simulated data set so they did not reveal any independent causal effects. Chilling requirements were also created inside ranges so they too hinted at cultivar identity. These feature dependencies make it hard to assign importance to correlated

variables because the model may spread importance in different ways across features that share overlapping information (Strobl *et al*., 2008; Molnar *et al*., 2021) . I believe that the differences in feature-importance rankings between forest and MLP should be understood by looking at the underlying dependencies. The differences, in feature-importance rankings do not prove that there are different agronomic mechanisms.

Regarding performance GraftNet did not consistently outperform simpler machine-learning approaches. On the corrected dataset random forest and logistic regression achieved classification results that were comparable to or better than GraftNet when using tabular inputs. Including the 52-week phenological series did not lead to consistent improvements. This result aligns with benchmarking studies, which have shown that tree-based ensemble methods often outperform or remain competitive with deep neural networks, on medium-sized tabular datasets (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022).

GraftNet's regression heads did not perform well compared to ridge regression when using time-series summaries to predict the growth parameter K. The RMSE for GraftNet was 6.1 cm while ridge regression achieved a better RMSE of 1.7 cm. Also checkpoint selection in GraftNet showed instability depending on the seed used. Still GraftNet had some value from a standpoint. It was able to predict the ranking of growth with high accuracy. 99.7% Of predictions were within one performance tier of the actual reference labels. It also managed to predict both K and r in a single model framework, which is a useful capability. Overall the results suggest that random forest models combined with SHAP or permutation-based explanations are currently more practical for handling structured nursery-related data. However this conclusion should be taken with caution. The findings are based on a simulated dataset, not real-world nursery data. So they do not prove the model works in nursery settings. There is no validation across real-world conditions yet. Moving forward improvements, to the multimodal framework should focus on including grafting images and actual measured phenological data. After that independent validation should be carried out across cultivars, rootstocks, growing seasons and agroecological environments.

6.Limitations

Even though apple production in Injibara District is becoming more important this study still has some problems. All performance labels were made with rules that came from one production guide; because of that the fake results do not show if they would work in real nurseries or, on real fields. The explainability results only show that the simulation matches the simulator; they do not prove farming methods. Also the predicted performance levels show how different apple varieties rank. They do not measure real graft-union details like how many grafts survive, how callus forms or how many trees live after grafting. The test set was 150 trees each time the data were split differently for five seeds and there was no test that did not depend on a specific apple variety; all of this made it hard to know if the model would work elsewhere. Because apple variety, how susceptible it is and how much chill it needs are all linked it was hard to see which features were most important. So it is still important to test with nursery and field data from many apple varieties, rootstocks, seasons and different environments before we can use it in practice.

7. Conclusions

This study. Tested an explainable deep learning system to predict how well apple grafts perform in simulated conditions in Ethiopia's highlands especially in the Injibara District. A problem with data leakage was found in the original time series simulator. It was fixed. With the corrected data GraftNet had a QWK of 0.954 ± 0.003 across five seeds but did not do better than traditional models like random forest and

logistic regression. Ridge regression that used summaries from the time series predicted growth patterns accurately. The image part of the system was not tested because real images were not available. SHAP and permutation importance showed that chilling needs, chill deficit and specific types of apple trees were factors. However the small effect of MM106 was not clearly seen. In traditional machine learning models performed as well as the new system but with less complexity. More testing with data from nurseries and fields was needed to make sure the system can be used reliably in Injibara and similar areas, in Ethiopia.

Declartion

- **Consent for publication:** Not applicable.

- **Funding Information:** The authors state no funding involved.

- **Conflict of Interest:** The authors declare no conflicts of interest regarding the publication of this manuscript.
- **Data Availability Statement:** Data available on request from the authors

References

Arevalo, J., Solorio, T., Montes-y-Gómez, M. and González, F.A., 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), pp.423-443.

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Cao, W., Mirjalili, V. and Raschka, S., 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, pp.325-331.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), p.213.

Fishman, S., Erez, A. and Couvillon, G.A., 1987. The temperature dependence of dormancy breaking in plants: mathematical analysis of a two-step model involving a cooperative transition. *Journal of Theoretical Biology*, *124*(4), pp.473-483.

Grinsztajn, L., Oyallon, E. and Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data?. *Advances in neural information processing systems*, *35*, pp.507-520.

González Noguer, C., Delgado, A., Else, M. and Hadley, P., 2023. Apple (Malus× domestica Borkh.) dormancy–a review of regulatory mechanisms and agroclimatic requirements. *Frontiers in Horticulture*, *2*, p.1217689.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q., 2017, July. On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Hooker, G., Mentch, L. and Zhou, S., 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), p.82.

Kebede, J., Masresha, F. and Sheleme, B., 2015. Effect of rootstock type and low soil pH on growth and yield of'Anna'and'Jonagored'cultivars.

Luedeling, E., 2012. Climate change impacts on winter chill for temperate fruit and nut production: a review. *Scientia horticulturae*, *144*, pp.218-229.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Melke, A., 2019. Studies on the economic benefits and extent of drought adaptation of apple (Malus domestica Borkh.) *genotypes introduced to Ethiopia* [Doctoral dissertation, Addis Ababa University].

Melke, A. and Fetene, M., 2014. Apples (Malus domestica, Borkh.) phenology in Ethiopian highlands:

plant growth, blooming, fruit development and fruit quality perspectives.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M. and Bischl, B., 2020. Pitfalls to avoid when interpreting machine learning models.

Richardson, E.A., Seeley, S.D. and Walker, D.R., 1974. A model for estimating the completion of rest for 'redhaven'and 'elberta'peach trees1. *HortScience*, *9*(4), pp.331-332.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), pp.206-215.

Shwartz-Ziv, R. and Armon, A., 2022. Tabular data: Deep learning is not all you need. *Information fusion*, *81*, pp.84-90.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. and Zeileis, A., 2008. Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), p.307.

Weinberger, J.H., 1950. Chilling requirements of peach varieties.