



Daily strawberry yield forecasting using artificial intelligence in commercial horticulture

Salman Ahmed ^{*}, Nicholas HM Caldwell 

Digital Futures Institute, Research and Knowledge Exchange Directorate, University of Suffolk, Ipswich, England, UK

ARTICLE INFO

Keywords:

Strawberry yield forecasting
Machine learning
TabPFN
Tabular data modelling

ABSTRACT

Accurate forecasting of yield is critical for commercial strawberry production, where daily harvest estimates directly influence labour allocation, logistics planning, and market supply commitments. Nevertheless, farm-scale forecasting operates under significant constraints, including limited multi-year history, seasonal discontinuities, seasonal zero-yield boundaries, and heterogeneous operational data sources. While most existing forecasting studies are evaluated on extensive datasets, this work specifically investigates yield prediction performance under limited-data conditions. We employ a leakage-aware, biologically informed framework for daily strawberry yield forecasting using six consecutive production years (2020–2025) from a commercial UK farm. The modelling pipeline integrates rolling environmental summaries and multi-horizon yield lag features while preserving seasonal boundaries and enforcing strict temporal alignment. A comprehensive benchmarking study evaluates linear models, distance-based methods, tree ensembles, gradient boosting frameworks, neural architectures, and a tabular foundation model (TabPFN) under three validation approaches: random train–test splitting, Leave-One-Out Cross-Validation, and forward-year generalisation. The results show that the evaluation design materially affects the apparent performance. Under deployment-relevant forward-year validation (train 2020–2024, test 2025), TabPFN achieves MAE of 438.53 and coefficient of determination (R^2) of 0.696, outperforming classical baselines. Tree-based ensembles also exhibit strong generalisation, whereas neural sequence models trained from scratch underperform in the limited-data. The findings indicate that biologically motivated temporal features capture substantial predictive structure and that robust inductive bias is more beneficial than architectural complexity in small seasonal datasets.

1. Introduction

Accurate yield forecasting remains one of the most challenging problems in commercial horticulture. For high-value, labour-intensive crops such as strawberries, daily production estimates directly inform harvest scheduling, labour allocation, logistics planning, storage management, and contractual supply commitments. Even modest forecasting errors at daily resolution can result in labour inefficiencies and post-harvest waste. The economic motivation is substantial: the global fresh strawberry market has been estimated at approximately USD 20.7 billion in 2024, with projections exceeding USD 30 billion by the mid-2030s, reflecting sustained growth in year-round demand and supply-chain scale [1]. In the United Kingdom, official statistics report the value of the strawberry crop at £389 million in 2024, showing the commercial relevance of improving forecasting accuracy within domestic horticulture [2]. As digital record-keeping and environmental sensing become increasingly embedded in farming operations, there is growing interest in

deploying Machine Learning (ML) and Deep Learning (DL) methods to predict historical production into reliable yield forecasts. Similar studies across other agricultural crops have further demonstrated the value of combining meteorological information with statistical and machine-learning approaches for yield forecasting, including comparative evaluations of linear, penalised-regression, and neural-network models for rice and sugarcane production [3–5].

Strawberry production presents particular modelling difficulties. The yield is not a sequence of independent observations, but a biologically structured time series governed by plant physiology, phenology, environmental exposure, and management practice. A typical production cycle follows a recognisable lifecycle: output begins at zero prior to harvest onset, increases progressively as plants enter peak fruiting, and eventually declines to zero as the season concludes. This rise–peak–decline behaviour is shaped by cultivar characteristics, planting density, accumulated temperature and humidity exposure, and field-level operational decisions. Consequently, daily yield must be understood as the result

^{*} Corresponding author.

E-mail address: s.ahmed@uos.ac.uk (S. Ahmed).

<https://doi.org/10.1016/j.atech.2026.102583>

Received 25 March 2026; Received in revised form 9 September 2026; Accepted 17 September 2026

Available online 24 September 2026

2772-3755/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

of accumulated biological processes rather than as an isolated daily outcome. Without careful temporal alignment, environmental variables recorded on the day of harvest or even after harvest events can sometimes be used as predictors. Such errors may lead to information leakage and artificially inflate the reported model performance. Similarly, if zero-yield periods occurring before the start and after the end of harvest are removed or aggregated, this will eliminate biologically meaningful seasonal boundaries and encourage models to learn an incomplete production lifecycle. A realistic forecasting framework must therefore satisfy three conditions. First, it must preserve the natural lifecycle structure of the crop, including genuine zero-yield periods. Second, it must ensure strict temporal alignment between environmental features and yields based on historical records for prediction. Third, it must evaluate model performance under validation schemes that reflect operational deployment, particularly forecasting into unseen future seasons rather than interpolating within the same season.

Another practical challenge concerns data availability. In commercial horticulture, reliable daily yield records and associated environmental data are often available for only a few years. Models trained on such limited temporal diversity may overfit to season-specific patterns and fail to generalise. As additional seasons accumulate, interannual variability in weather, planting density, and management practices introduces richer learning signals but also greater complexity. This study addresses these issues through the development and evaluation of a biologically grounded, leakage-aware framework for daily strawberry yield forecasting using real field-level data spanning six consecutive production years (2020–2025). The modelling pipeline integrates daily yield observations with environmental measurements using lag-based yield features to capture short-, medium-, and long-term production memory. Rolling environmental indicators are used to represent accumulated exposure to temperature and humidity, reflecting physiologically relevant growth drivers. All genuine zero-yield periods are retained, preserving the natural onset and termination of harvest cycles rather than imposing some notional and potentially untrue view of continuous production. Fig. 1 shows strawberry plants observed during an on-site visit to the partner commercial Farms.

The novelty of this study lies in the combination of these elements within a single commercial forecasting framework. In contrast to previous strawberry yield studies that have largely focused on big horticulture datasets, this study investigates whether a tabular foundation model can provide useful inductive bias for a small, seasonal, operational horticultural dataset when compared with conventional machine-learning, boosting, and neural approaches.

The contributions of this work are fourfold. First, it establishes a lifecycle-aware data preparation strategy that maintains biological plausibility while avoiding information leakage. Second, it demonstrates the value of biologically motivated lag features in capturing production memory and improving predictive stability. Third, it provides a forward-year evaluation of multiple modelling paradigms on daily strawberry yield data, highlighting the practical strengths and limitations of each approach. Fourth, it presents evidence that multi-year temporal diversity substantially enhances model generalisation under realistic forecasting conditions. The remainder of this paper is organised as follows. Section 2 reviews the existing literature. Section 3 describes the dataset and preprocessing procedures. Section 4 details the modelling approaches and validation design. Section 5 presents empirical results, including forward-year performance analysis. Sections 6 and 7 discuss implications for agricultural forecasting and directions for future research.

2. Related work

Strawberry yield prediction has a longer history than many modern precision-horticulture applications because the crop's commercial value makes even small improvements in planning accuracy economically meaningful. Early empirical work focused on relating weather and phenology to yield at seasonal horizons. Døving and Måge proposed



Fig. 1. Strawberry plants during a farm visit to commercial farms.

yield prediction approaches for strawberry production in Norway, shows the long-standing use of climatic and seasonal signals for anticipating harvest volumes [6]. Pathak et al. developed a meteorology-based yield forecasting approach for strawberries, demonstrating that correlations between meteorological variables and yield can support lead-time prediction [7]. Maskey et al. extended this line of work by evaluating statistical baselines alongside machine learning models (including Random Forests and neural networks) for field-scale strawberry yield prediction using weather variables. They reported that ML methods can outperform principal component regression and emphasised that validation across fields and seasons remains essential [8]. Both works are directly relevant to farm deployment contexts, as they highlight two persistent realities: weather signals contain meaningful predictive information, yet performance estimates are highly sensitive to how temporality, seasonality, and generalisation are handled. Sim et al. used greenhouse environmental and plant growth data to predict strawberry yield, showing that regression models can make effective use of microclimate variables in soil-grown greenhouse systems [9]. In greenhouse settings, information about cultivar and growth stage often improves prediction performance. For example, Jo et al. showed that cultivar-specific growth models produce stronger agreement between predicted and observed yields than a single pooled model, supporting the view that cultivar identity and phe-

nology should be treated as meaningful factors rather than incidental covariates [10].

More recently, strawberry yield forecasting has increasingly incorporated imaging to capture crop states such as flower counts, fruit counts, and phenological stages that cannot be reliably inferred from weather data alone. Liu et al. developed a visual-specific strawberry yield forecasting system that predicts next-week yield by combining machine-vision-based fruit monitoring with environmental data, demonstrating that multi-modal inputs can improve short-horizon operational forecasting for indeterminate, repeatedly harvested crops [11]. Similarly, Yang et al. proposed a knowledge-guided computer-vision framework (KGCV-strawberry) that assimilates sequential observations to improve fruit growth and yield prediction in greenhouse settings, reporting improved prediction accuracy as observation sequences lengthen [12]. Yoon et al. [13] explored yield prediction using image-based receptacle detection, arguing that receptacle signals are directly related to the eventual formation of fruit and can support yield estimation when detection uncertainty.

IoT-based sensing has enabled higher frequency measurement of microclimate conditions and has encouraged the use of nonparametric and ML approaches that exploit short-term environmental fluctuations. Yamazaki et al. proposed a strawberry yield prediction scheme using data from general-purpose IoT nodes and demonstrated improved accuracy when incorporating differenced (change-based) features into non-parametric regression models, particularly generalised additive models [14]. However, even IoT-focused strawberry studies frequently report a practical limitation: sensor histories are short, and data completeness varies across seasons and farms. One response is the synthetic augmentation of sensor histories rather than yield values. Ayall et al. proposed a backcasting approach that generates synthetic IoT sensor data to extend limited on-farm histories and improve strawberry yield forecasting when real and generated signals are combined [15]. The work indicates that the farms often have incomplete temporal coverage, so forecasting pipelines must explicitly accommodate discontinuities, seasonal gaps, and sparse multi-year records rather than assuming long, stationary time series.

A core challenge in farm-scale strawberry forecasting is that the effective sample size is constrained by the number of harvest days and seasons, and the resulting time series contains genuine structural zeros at season boundaries. Under these conditions, high-capacity sequence models can be unstable, while conventional tabular models may overfit if evaluation procedures permit information leakage. Bergmeir et al. analysed when cross-validation is valid for autoregressive forecasting and clarified why naïve cross-validation choices can yield biased performance estimates [16]. Cerqueira et al. compared multiple strategies for estimating forecasting performance and showed that the most reliable approach can vary under non-stationarity and temporal dependence [17]. Roberts et al. argued for blocked or structured cross-validation when data are temporally or spatially correlated, highlighting that random splitting can substantially overstate generalisation performance [18]. These concerns are highly relevant to strawberry yield data, as strong seasonality and within-season autocorrelation make it easy to introduce information leakage through random train test splits or improperly aligned weather variables. Table 1 shows that, while strawberry yield forecasting has been explored across field, greenhouse, and IoT-enabled environments, our study explicitly targets multi-year farm datasets with strict forward-season evaluation.

3. Dataset

3.1. Project scope

This work was undertaken within the *Agri-KG* project, an industry-led research and development initiative supported through Innovate UK collaborative R&D funding. The project context is explicitly deployment-oriented: the objective is to develop a practical forecasting capability

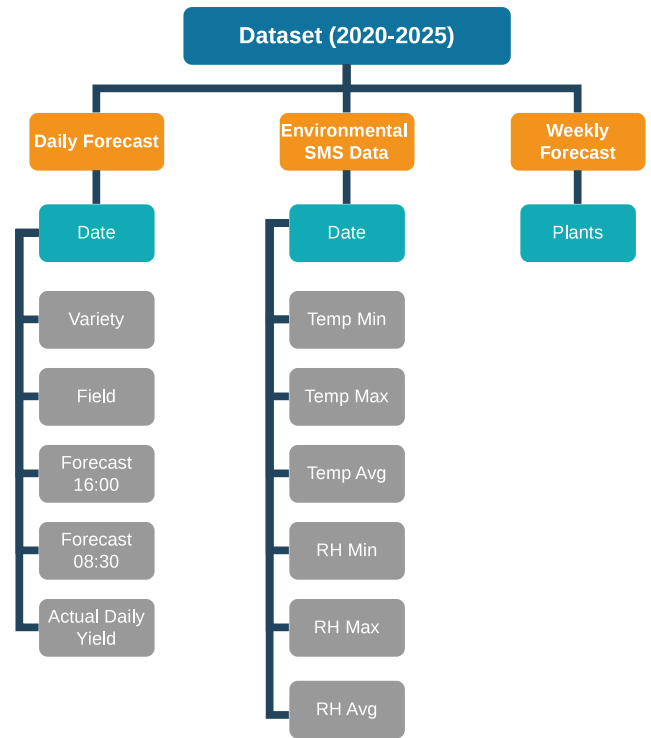


Fig. 2. Operational data sources used to construct the 2020–2025 strawberry forecasting dataset.

that can be integrated into farm decision-making workflows (e.g., harvest planning) and, in the longer term, into an AI-enabled agricultural knowledge infrastructure. In this setting, methodological choices are driven not only by prediction but also by operational constraints typical of commercial horticulture, including short historical records, discontinuous seasonal production, missing sensor observations, and evolving field management practices. The dataset used in this study originates from a partner farm, a commercial soft-fruit producer in the United Kingdom, and reflects real operational conditions rather than controlled trials. The data were not collected solely for research purposes; they also consist of routine production records and sensor exports used in day-to-day farm management. Consequently, the dataset exhibits practical challenges that determine the industrial viability of forecasting methods, including zero-yield periods at season boundaries and interannual variability driven by environmental data.

3.2. Data sources

The modelling dataset was constructed by integrating three primary data sources, shown in Fig. 2:

- (i) **Weekly forecast dataset** A weekly planning dataset provides field- and variety-level attributes in wide format. The columns include identifiers
 - **Field_Name:** Farm field/block identifier used in planning records.
 - **Variety:** Strawberry cultivar/variety grown in the field.
 - **Area_ha:** Cultivated area of the field in hectares.
 - **Plants:** Number of plants in the field (planting density proxy).
 - **Kg/plant:** Productivity metric (yield per plant; derived from totals).
 - **Kg/m2:** Productivity metric (yield per square metre; derived from totals/area).
 - **Kg/Ha:** Productivity metric (yield per hectare; derived from totals/area).

Table 1
Comparison of representative strawberry yield forecasting studies and evaluation protocols.

Study	Setting	Forecast Horizon	Input Signals	Dataset Duration	Validation Strategy	Reported Limitation
Døving & Måge (2001) [6]	Field	Seasonal	Climate variables		Holdout / empirical evaluation	Coarse aggregation; limited ML methods
Pathak et al. (2016) [7]	Field	Seasonal / Lead-time	Meteorological variables		Train–test split	Weather-only features; limited cross-season validation
Maskey et al. (2019) [8]	Field-scale	Seasonal	Weather indices + ML		Random and holdout splits	Limited temporal diversity; risk of optimistic splits
Sim et al. (2020) [9]	Greenhouse	Within-season	Environmental + growth data	Single experiment period	Within-dataset validation	Controlled setting; limited inter-annual variability
Jo et al. (2022) [10]	Tunnel greenhouse	Seasonal	Cultivar-specific growth variables	Short experimental window	Train–test split	Strong cultivar dependency; limited seasons
Liu et al. (2025) [11]	Greenhouse	Next-week	Vision + environmental data	Multiple waves within season	Standard holdout	Requires imaging hardware; short-horizon focus
Yang et al. (2024) [12]	Greenhouse	Short-term	Knowledge-guided vision + climate	Sequential observations	Train–test split	Primarily controlled environment; limited long-term generalisation
Yamazaki et al. (2025) [14]	IoT-enabled field	Short-term	IoT climate nodes	Limited sensor history	Holdout	Sparse historical depth; small dataset approach
Sugiyama et al. (2025) [19]	Mechanistic simulation	Seasonal	Physiological DM partition model	Experimental calibration data	Simulation validation	Parameter-sensitive; not data-driven ML
This Study	Field-scale (commercial farm)	Daily	Yield lags + rolling weather + field attributes	Six consecutive seasons (2020–2025)	Forward-year validation (2025 unseen) + leakage-aware design	Targets small multi-year approach with strict temporal alignment

- **Kg/field**: Total yield for the field (seasonal aggregate).
- **EstvsActual**: Operational comparison indicator between estimated and actual totals.
- **Totals_Check**: Consistency/check column used in planning spreadsheets.
- **25-Feb, 03-Mar, ..., 01-Dec**: Week-indexed seasonal columns recording weekly planned/estimated/recorded yield values (wide format).
- **%_picked**: Proportion of expected crop already harvested/picked (progress indicator).

Only **Plants** is merged into the modelling dataset as a causally meaningful agronomic predictor. Aggregate productivity indicators and season-total columns (e.g., **Kg/field**, **Kg/Ha**) are excluded from predictive features to avoid target leakage.

(ii) **Daily forecast dataset**. The daily forecasting table records production outcomes and operational forecasts at daily resolution with the following columns:

- **Product**: Crop category (used to filter strawberries where applicable).
- **Variety**: Strawberry cultivar/variety.
- **Field_No**: Numeric field/block identifier used in daily production logs.
- **Week_Start_Date**: Start date of the corresponding production week (weekly anchor).
- **Weekday**: Day-of-week label for operational reporting.
- **Date**: Calendar date of the daily record.
- **Forecast_1600**: Farm operational yield forecast recorded at 16:00 (day before harvest).
- **Forecast_0830**: Farm operational yield forecast recorded at 08:30 (same day of harvest).
- **Actual_Daily_Yield**: Actual harvested yield for that day (target variable, kg).

Actual_Daily_Yield is the supervised learning target. **Forecast_0830** and **Forecast_1600** are retained as operational baselines and may optionally be used as predictors only when the research question concerns forecast refinement.

(iii) **Environmental sensor measurements (SMS platform)**. Environmental measurements were obtained from on-farm sensors accessed via the SoilMoistureSense platform (<https://www.soilmoisturesense.com>) deployed at the partner farms. The on-farm sensor exports available for this research provide temperature and relative humidity summaries at daily resolution. Other agronomically relevant drivers, such as rainfall, precipitation and solar radiation/daylight exposure were not available in the operational dataset and therefore could not be included as predictors.

- **Serial No.:** Unique identifier of the IoT sensor/station.
- **Label:** Human-readable sensor/station label.
- **Local Date:** Local calendar date of the sensor record.
- **Local Time:** Local time of the sensor record.
- **Max Temp1:** Maximum temperature from sensor channel Temp1 over the reporting period.
- **Min Temp1:** Minimum temperature from sensor channel Temp1 over the reporting period.
- **Ave Temp1:** Average temperature from sensor channel Temp1 over the reporting period.
- **Max Temp2:** Maximum temperature from sensor channel Temp2 over the reporting period.
- **Min Temp2:** Minimum temperature from sensor channel Temp2 over the reporting period.
- **Ave Temp2:** Average temperature from sensor channel Temp2 over the reporting period.
- **Max RH:** Maximum relative humidity over the reporting period.
- **Min RH:** Minimum relative humidity over the reporting period.
- **Ave RH:** Average relative humidity over the reporting period.

Sensor measurements are aggregated/aligned to daily resolution and merged by date with the yield records. Environmental predictors are temporally aligned to prevent leakage (i.e., ensuring that only biologically plausible pre-harvest information is used for forecasting).

4. Methodology

The forecasting dataset was gathered from operational farm records covering six production years (2020–2025). Three data sources were integrated to form the modelling dataset: weekly forecasting spreadsheets, daily forecast records, and environmental sensor exports from the SoilMoistureSense platform. Each source differed in format and temporal resolution and therefore required restructuring prior to merging. The weekly data were provided as six annual spreadsheets in wide format. These files contained field identifiers, agronomic descriptors, and week-indexed seasonal projections. From these spreadsheets, only the plant count variable was extracted and merged into the daily dataset. Daily forecast data consisted of 189 individual files covering the full study period. These files contain realised daily yield together with operational forecasts recorded at 08:30 and 16:00. All files were consolidated into a unified long-format dataset. Date formats were standardised, and field and variety identifiers were harmonised to ensure consistency across years. Environmental measurements were recorded at sub-daily frequency and included temperature and relative humidity observations with timestamps. These data were aggregated to daily summaries, including maximum, minimum, and average values. Aggregation was performed strictly within calendar-day boundaries to ensure temporal consistency with daily production records. Strawberry production is seasonal and follows a defined biological cycle within each year. Yield increases after harvest onset, reaches a peak phase, and declines to zero at the end of the season. Zero-yield periods before and after harvest were retained, as they reflect real production boundaries. The dataset therefore, consists of repeated seasonal trajectories indexed by field, variety, and year rather than a single continuous time series. This structure informed the choice of a regression-based modelling approach with temporally informed features. Fig. 3 illustrates the general workflow of the methodology.

4.1. Data preprocessing

To ensure consistent representation of seasonal boundaries, a fixed number of seven zero-yield days were appended at the beginning and end of each harvest cycle where appropriate. These padded observations represent biologically plausible pre-harvest and post-harvest periods rather than artificial mid-season gaps. The dataset contained 2280 daily observations without boundary padding. After inclusion of padded zero-yield days, the total number of records increased to 3,540. This augmentation was applied solely to reinforce season onset and termination signals and to improve model stability in learning harvest boundaries. No artificial values were introduced within active harvest periods. Numeric fields were cleaned to remove non-numeric characters and standardise data types. Missing environmental values were imputed using median statistics to prevent unnecessary loss of daily observations while maintaining robustness against extreme values. Sensor readings were inspected for implausible outliers using distributional checks and removed where clearly inconsistent with expected operating ranges.

Plant count was not recorded in the daily forecasting files and was instead obtained from the farm's annual weekly planning spreadsheets, where plant numbers are maintained as part of routine commercial reporting. Plant count was treated as a static field-level attribute and mapped deterministically onto daily records via the common field identifier. In addition to pre- and post-harvest zero-yield boundaries, the dataset contains genuine in-season zero-yield observations reflecting operational resting periods and differing spring versus autumn cycle

Table 2

Annual composition of the final modelling dataset.

Year	Observations	Production locations	Location–variety combinations
2020	569	15	15
2021	608	14	18
2022	768	13	19
2023	556	11	15
2024	460	10	11
2025	579	11	12
Total	3,540	–	90

lengths across years. These zeros were retained and treated as valid production outcomes rather than missing values, that the modelling task reflects both the magnitude and timing structure of commercial harvest dynamics.

The annual composition of the final modelling dataset is summarised in Table 2, which reports the number of observations, production locations, and location–variety combinations for each year from 2020 to 2025.

4.2. Feature engineering

Throughout preprocessing, strict temporal and causal validity was maintained. Variables derived from realised seasonal outcomes, such as yield per hectare, percentage picked, and seasonal totals, were excluded from the feature set to prevent indirect encoding of the target variable. Environmental measurements were aligned strictly by calendar date, ensuring that no information from future timestamps entered the forecasting pipeline. Feature engineering was designed to capture biological production memory and environmental influence. Yield persistence was represented using rolling averages of prior yield over lag windows of 1, 3, 7, 14, and 21 days. These features reflect short-term harvest momentum and medium-term fruit development cycles [20]. Strawberry yield exhibits strong temporal dependence driven by biological growth cycles. To capture this behaviour, lagged rolling mean features are computed for yield at multiple temporal horizons within each field–variety combination. Short-term lags (e.g., 1 day) capture immediate momentum, weekly lags (7 days) represent short growth cycles, and longer lags (14 and 21 days) reflect early-season and cumulative growth trends. These features enable the model to learn whether production is increasing, stabilising, or declining, providing substantial predictive power in a biologically realistic manner. Plant response is governed not only by daily weather conditions but also by accumulated environmental exposure. Accordingly, rolling window features are derived to capture short- and medium-term environmental stress, including 7-day rolling temperature averages, 14-day rolling humidity conditions. These features model the cumulative effects of heat and moisture, improving the model's ability to anticipate yield spikes, plateaus, and drops. Temporal feature generation may be reproduced by following these steps.

1. Sort records by FIELD_KEY, Variety, and Date.
2. Shift realised yield by one observation before constructing historical-yield features.
3. Calculate 1-, 3-, 7-, 14-, and 21-observation moving-average yield features.
4. Calculate 7- and 14-observation rolling averages of average temperature and relative humidity.
5. Construct the predictor matrix(X) and Actual_Daily_Yield target(y).

4.3. Experimental setup

All experiments were implemented in Python using scikit-learn, XGBoost, LightGBM, CatBoost, PyTorch, and TabPFN. Where applicable, models were executed on GPU hardware to provide consistent computational comparison across frameworks. A unified preprocessing pipeline

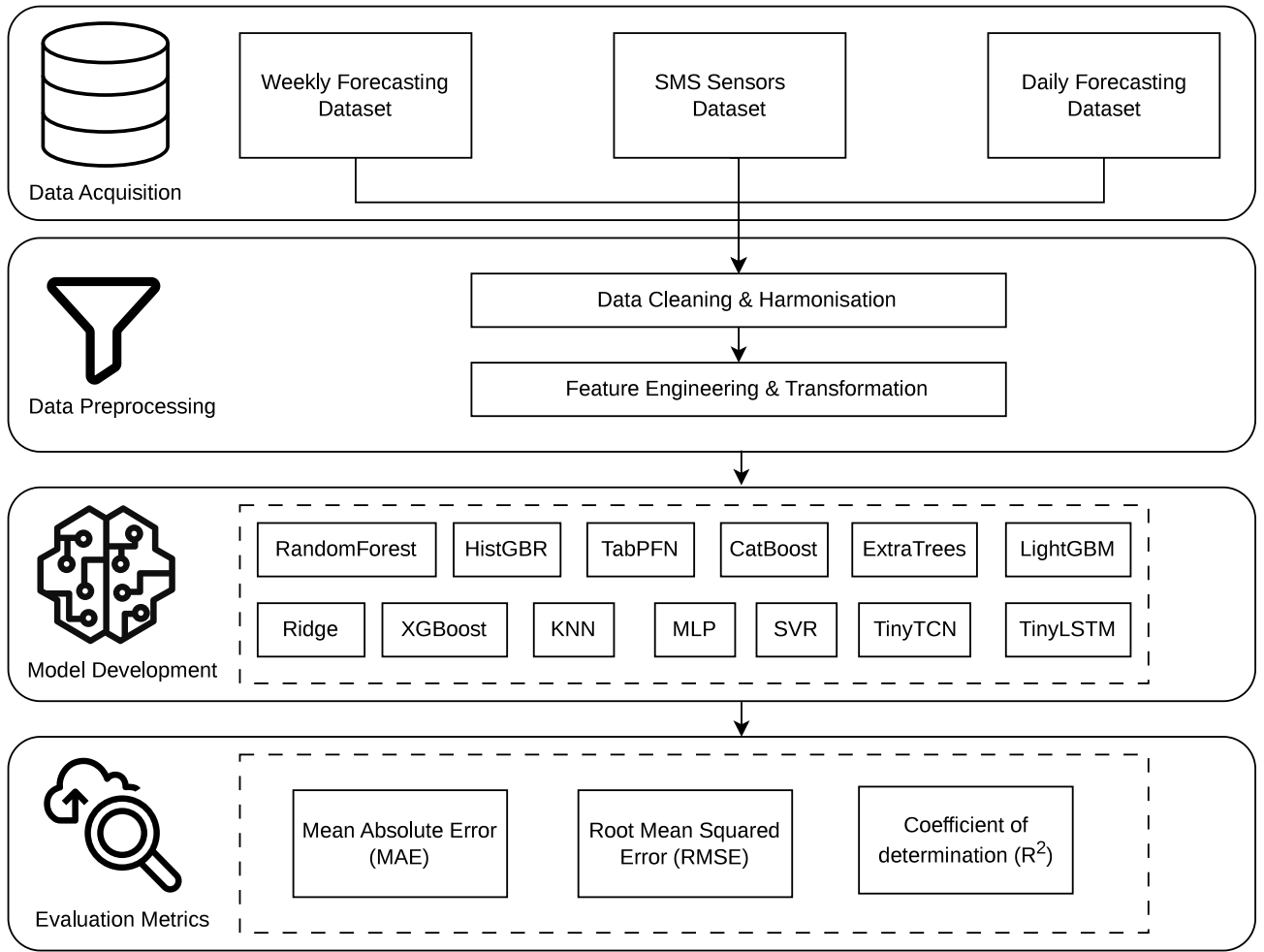


Fig. 3. Overall workflow for data integration, preprocessing, model development, and performance evaluation.

Table 3
Model configurations used in the comparative evaluation.

Model	Experimental configuration
Ridge	$\alpha = 1.0$; one-hot encoding and numerical standardisation.
KNN	$k = 8$ neighbours; one-hot encoding and numerical standardisation.
SVR-RBF	RBF kernel; $C = 10$; one-hot encoding and numerical standardisation.
Random Forest	300 estimators; <code>random_state=42</code> ;
Extra Trees	300 estimators; <code>random_state=42</code> ;
HistGradientBoosting	Maximum depth = 8; <code>random_state=42</code> ;
XGBoost	400 estimators; histogram tree method; CUDA execution; <code>random_state=42</code> ; pre-transformed feature matrix.
LightGBM	400 estimators; GPU execution; <code>random_state=42</code> ; pre-transformed feature matrix.
CatBoost	400 iterations; GPU execution; <code>random_state=42</code> ; pre-transformed feature matrix.
MLP	128–64–1 architecture; ReLU activation; Adam optimiser; learning rate 10^{-3} ; MSE loss; 150 epochs; CUDA execution.
TinyLSTM	Single LSTM layer with hidden size 32 followed by a linear output layer; Adam optimiser; learning rate 10^{-3} ; MSE loss; 150 epochs; CUDA execution.
TinyTCN	Two Conv1D layers with 32 and 16 channels, respectively; kernel size = 2; ReLU activation; global averaging; Adam optimiser; learning rate 10^{-3} ; MSE loss; 150 epochs; CUDA execution.
TabPFN	<code>TabPFNRegressor(device="cuda")</code> ; TabPFN v6.0.5; constructor parameters retained the defaults of version 6.0.5.

was applied across all models. Categorical variables were encoded using one-hot encoding with unknown categories ignored at inference time.

The categorical encoding and numerical standardisation transformations were fitted using the corresponding training partition within each evaluation approach and subsequently applied unchanged to the associated test data.

The exact configurations used in the experiments are summarised in Table 3. Variety was the categorical predictor and was transformed using `OneHotEncoder`; numerical predictors were standardised using

`StandardScaler`. NumPy and PyTorch random seeds were fixed at 42, with `random_state=42` specified where supported.

Dedicated Transformer-based and other high-capacity time-series architectures were not included because the available data comprise only six production seasons with relatively short and discontinuous field-variety trajectories. Although such architectures can be applied to agricultural forecasting, their substantially greater data and sequence requirements would increase the risk of overfitting and unstable estimation in the present limited-data setting [21]. Temporal dependence was

TabPFN - Tabular Prior-data Fitted Network

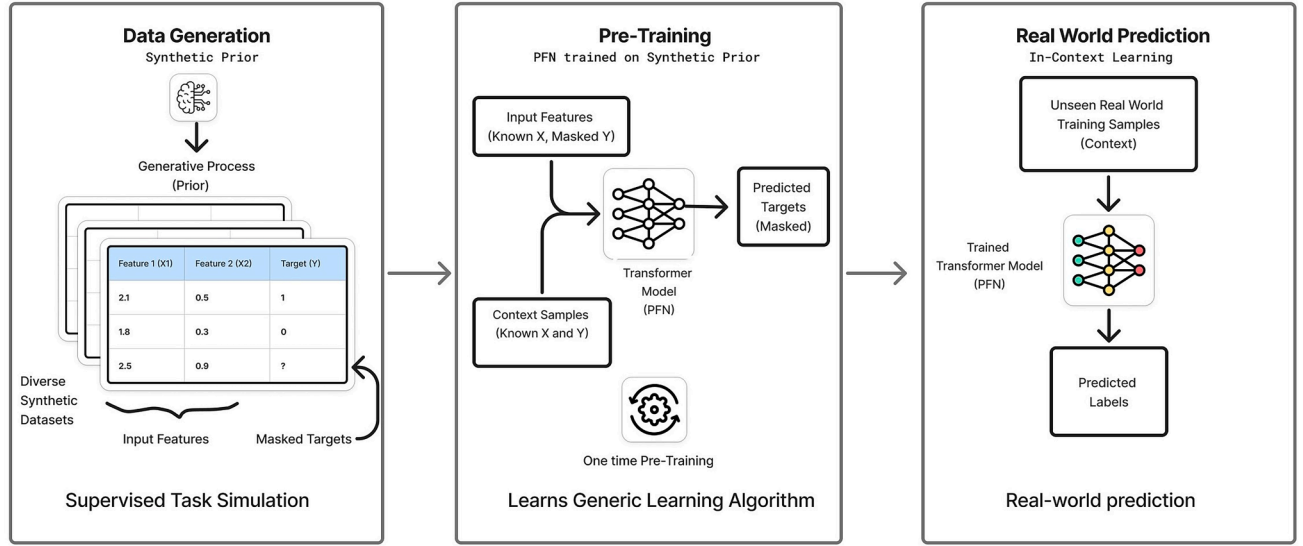


Fig. 4. Overview of the TabPFN framework for pretrained tabular prediction [22].

therefore represented through biologically motivated lag and rolling-window features and evaluated using compact neural baselines alongside conventional machine-learning and tabular foundation-model approaches.

Neural models were trained for 150 epochs using the Adam optimiser (learning rate 10^{-3}) and mean squared error loss. The Multi-layer Perceptron consisted of two hidden layers (128 and 64 units) with ReLU activation. The LSTM model used a single recurrent layer (hidden size = 32), and the Temporal Convolutional Network used two convolutional layers followed by global averaging. Neural computations were performed on GPU where available. For GPU gradient boosting models, histogram-based tree methods were used with 400 estimators and fixed random states. Classical tree ensembles used 300 estimators. Hyperparameters were selected using stable default configurations rather than aggressive tuning to evaluate robustness under limited data conditions. TabPFN was trained using the same preprocessing pipeline and executed on GPU (Fig. 4).

4.4. Feature importance evaluation

Feature importance was evaluated using a permutation-based approach. Permutation importance is a model-agnostic method that measures the contribution of each input variable by quantifying the degradation in predictive performance when the values of that feature are randomly permuted. The intuition behind this method is that if a feature contains important predictive information, shuffling its values breaks the relationship between that feature and the target variable, resulting in a measurable increase in prediction error. The analysis was conducted using the trained forecasting pipeline, which includes both preprocessing operations and the predictive model. Categorical variables were encoded using one-hot encoding, and numerical variables were standardised through z-score normalisation within the preprocessing stage. Permutation importance was then computed directly on the complete pipeline to ensure that the perturbation of input variables occurred before preprocessing, thereby preserving consistency with the original model input structure.

Feature importance was evaluated using the unseen test dataset corresponding to the 2025 production season, which served as the forward-year validation set in the primary experimental design. For each feature, the values in the test set were randomly permuted while all other

variables remained unchanged. The model was then used to generate predictions, and the resulting change in prediction error was recorded. This process was repeated multiple times ($n=10$) for each feature to reduce the influence of random fluctuations. MAE was used as the performance metric when evaluating permutation importance. The importance score for each feature was defined as the average increase in MAE observed after permuting that feature. Larger increases in MAE indicate that the model relies more strongly on the corresponding variable for accurate prediction. This procedure provides an interpretable ranking of predictors and enables examination of whether the forecasting model is driven by variables consistent with biological and environmental drivers of strawberry production.

4.5. Evaluation metrics

Model performance was assessed using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Let y_i denote the observed daily yield for observation i , $\hat{y}_i$ the corresponding predicted yield, $\bar{y}$ the mean observed yield in the evaluation set, and n the number of observations.

4.5.1. Mean absolute error (MAE)

MAE measures the mean absolute magnitude of the prediction errors:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (1)$$

MAE is expressed in the same units as the target variable, with lower values indicating better predictive performance.

4.5.2. Root mean squared error (RMSE)

RMSE is the square root of the mean squared prediction error:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (2)$$

RMSE is expressed in the same units as the target variable. Because prediction errors are squared before averaging, RMSE gives greater weight to larger errors than MAE and therefore provides a complementary measure of predictive performance.

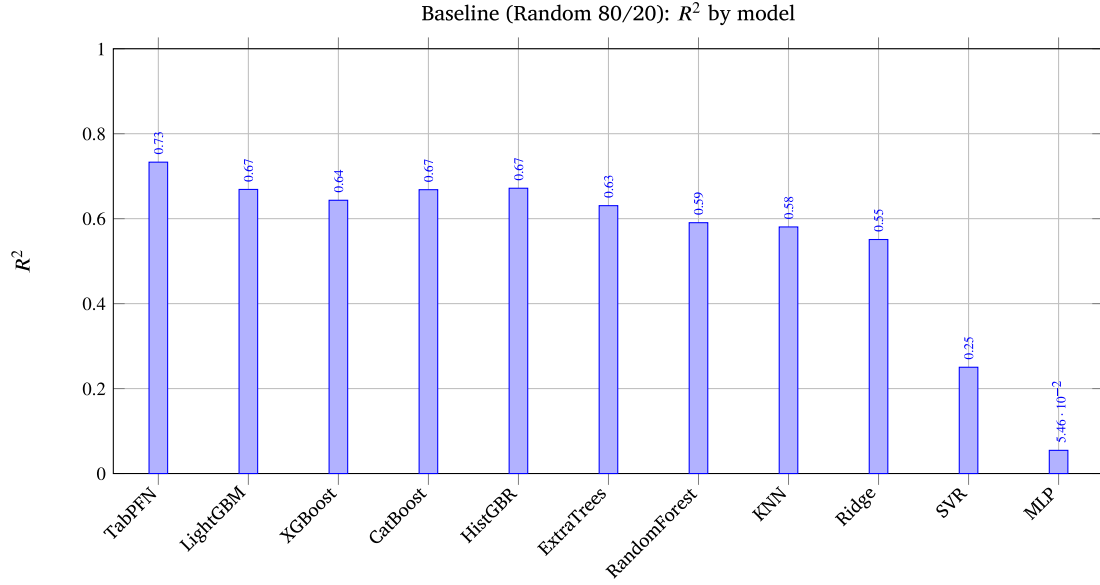


Fig. 5. Baseline R^2 comparison under random 80/20 splitting. TabPFN achieves the highest R^2 , followed by tree-based and boosting models, while neural baselines exhibit markedly lower explanatory power.

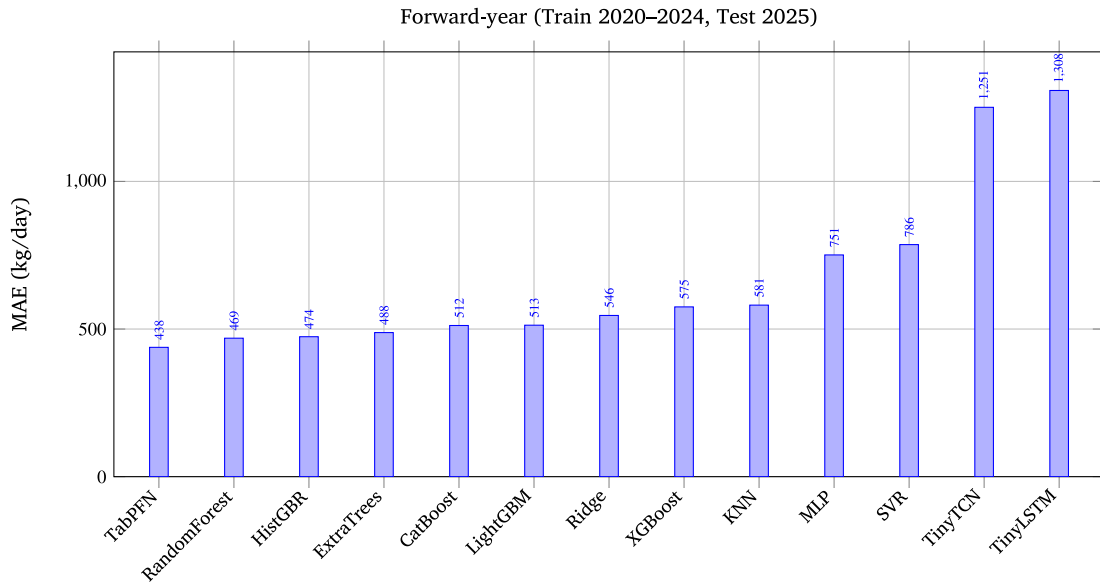


Fig. 6. Forward-year MAE comparison (Train 2020–2024, test 2025). TabPFN remains best overall; robust tree ensembles (Random Forest, HistGBR, Extra Trees) are competitive; neural sequence baselines perform poorly in this limited-data setting.

4.5.3. Coefficient of determination (R^2)

The coefficient of determination is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (3)$$

R^2 is dimensionless and measures predictive performance relative to a constant predictor equal to the mean observed yield in the evaluation set. A value of $R^2 = 1$ indicates perfect predictions, $R^2 = 0$ indicates the same sum-of-squares performance as the mean-yield reference, and $R^2 < 0$ indicates poorer performance than that reference.

These three metrics provide complementary information for operational forecasting. MAE represents the average absolute difference between the predicted and actual harvested yield and therefore gives a direct indication of the expected daily forecasting error in kilograms. RMSE gives greater weight to larger prediction errors and is partic-

ularly informative when substantial errors occur during high-yield or peak-harvest periods. R^2 provides a dimensionless measure of how well the predictions perform relative to using the mean observed yield as a reference.

4.6. Validation design

Three evaluation approaches were implemented.

4.6.1. 80/20 Random split

First, an 80% / 20% random train–test split was used as a baseline comparison. The dataset was randomly partitioned with shuffling enabled and fixed random state (42) to ensure reproducibility. This conventional split provides a reference benchmark for model behaviour under standard machine learning practice.

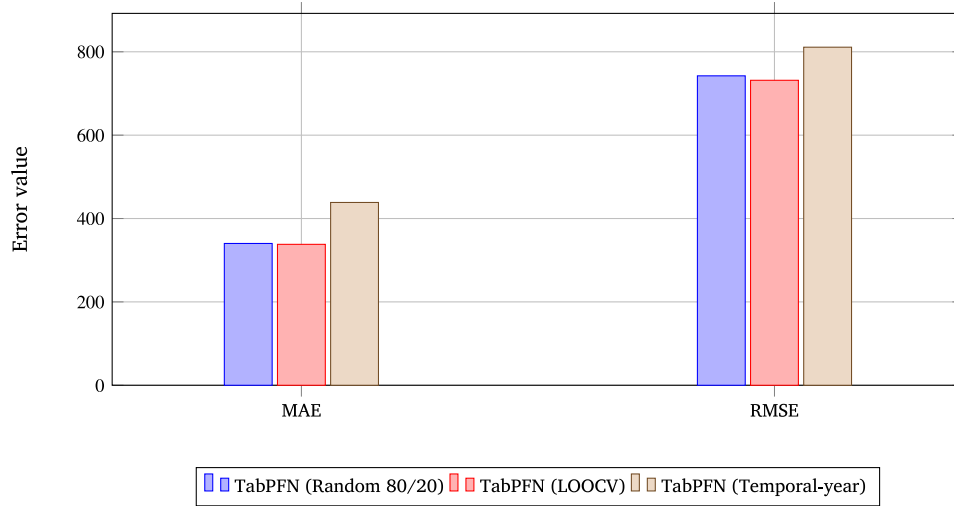


Fig. 7. TabPFN MAE and RMSE under increasing validation strictness.

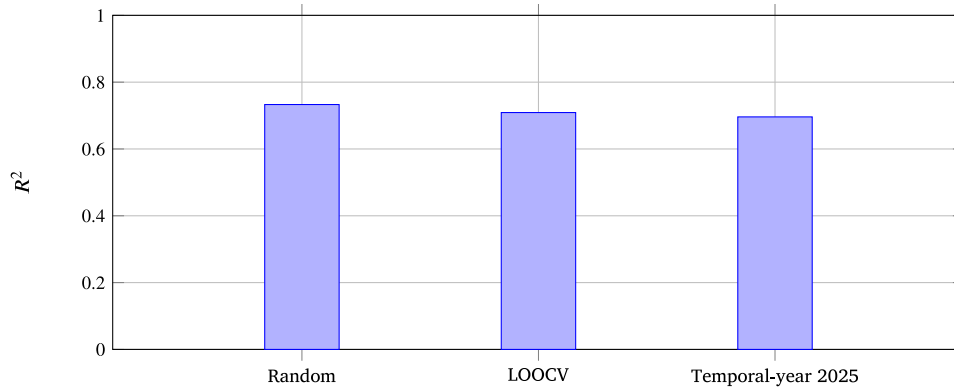


Fig. 8. TabPFN (R^2) under increasing validation strictness.

4.6.2. Leave-one-out cross-validation

To assess model stability under extreme data scarcity, Leave-One-Out Cross-Validation (LOOCV) was performed using the full modelling dataset. Under this approach, each observation is used once as a test sample while all remaining observations serve as the training set. This procedure results in N training-evaluation cycles for a dataset containing N observations, providing an exhaustive internal resampling estimate without reliance on a single arbitrary train-test split. LOOCV was implemented using scikit-learn's `LeaveOneOut` iterator. For each iteration, the preprocessing pipeline (one-hot encoding for categorical variables and standardisation for numerical variables) was fitted exclusively on the training subset to prevent leakage.

4.6.3. Forward-year validation (2020–2024 → 2025)

While LOOCV provides a comprehensive internal validation estimate, it does not replicate operational deployment conditions where models must generalise to future seasons. Therefore, a forward-year validation protocol was adopted as the primary evaluation strategy. Under this approach, all observations from production years 2020–2024 were used for training, and the entire 2025 season was reserved as an unseen test set. This approach prevents temporal leakage and ensures that the model is evaluated on a season characterised by potentially different weather patterns, management decisions, and production dynamics. The same preprocessing pipeline was fitted on the training years and applied unchanged to the 2025 test data. Models were trained once on the historical data and evaluated directly on the future season.

5. Results

This section reports forecasting performance under three validation approaches of increasing operational realism: (i) a conventional random 80/20 split (baseline), (ii) Leave-One-Out Cross-Validation (LOOCV) for exhaustive internal validation, and (iii) forward-year evaluation in which models are trained on 2020–2024 and tested on the unseen 2025 season.

5.1. Random 80/20 split

This is a standard check of predictive ability when train and test may share similar seasonal segments. It often appears strong because the model sees mixed conditions during training. Table 4 summarises the baseline results under a random 80/20 split. Under this conventional evaluation approach, the tabular foundation model (TabPFN) achieved the lowest error (MAE = 340.18, RMSE = 742.32, R^2 = 0.733). Among classical machine learning baselines, GPU-accelerated gradient boosting methods were consistently strong: LightGBM provided the best MAE among non-TabPFN models (MAE = 416.49), closely followed by XGBoost (MAE = 418.08) and Extra Trees (MAE = 420.40). Histogram Gradient Boosting achieved a similar MAE (MAE = 423.48) while being computationally efficient. The neural baselines were comparatively weak in this approach. The MLP exhibited substantially higher error (MAE = 738.21) and near-zero variance explanation (R^2 $\approx$ 0.055), while the sequence-style baselines (TinyTCN and TinyLSTM) produced negative R^2 , indicating performance worse than a constant mean pre-

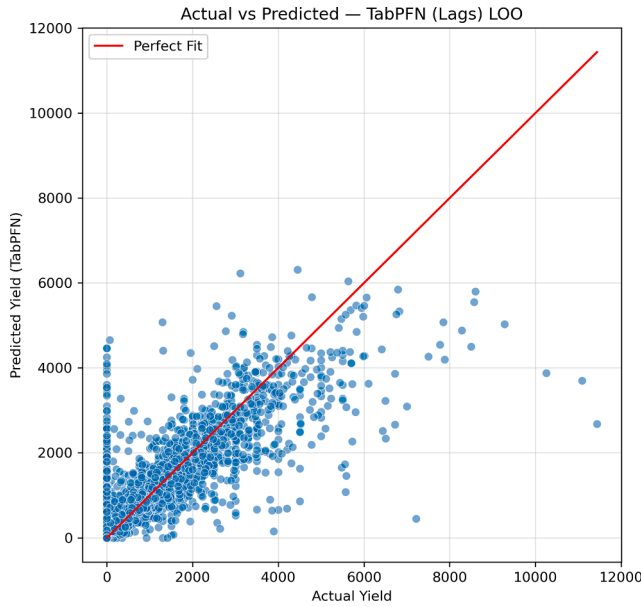


Fig. 9. Actual versus predicted daily yield for the TabPFN model under LOOCV evaluation. The red line indicates perfect prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4

Baseline benchmarking results under random 80/20 split. Lower MAE/RMSE and higher R^2 indicate better performance. Runtime is reported where available from the benchmarking script.

Model	MAE	RMSE	R^2	Runtime (s)
LightGBM_GPU	416.49	827.04	0.6688	1.13
XGBoost_GPU	418.08	858.31	0.6433	0.77
ExtraTrees	420.40	873.40	0.6306	3.82
CatBoost_GPU	420.59	827.77	0.6682	0.85
HistGBR	423.48	823.60	0.6716	0.27
RandomForest	442.46	919.68	0.5905	10.50
KNN	462.55	930.77	0.5805	0.02
Ridge	538.60	963.11	0.5509	0.02
SVR_RBF	622.62	1244.37	0.2502	0.53
PyTorch_MLP_GPU	738.21	1397.34	0.0546	0.24
TinyTCN_GPU	891.01	1661.45	-0.3366	0.97
TinyLSTM_GPU	912.22	1700.59	-0.4003	0.60
TabPFN	340.18	742.32	0.733	0.61

dictor. In practical terms, these outcomes are consistent with a limited-data (2020–2025) approach, where the inductive biases of tree-based learners and tabular foundation models are more consistent than those of DL models. Fig. 5 reports baseline R^2 values under random 80/20 splitting.

5.2. Zero-padding ablation and sensitivity analysis

Following the baseline comparison in Section 5.1, TabPFN was held constant to examine the effect of harvest-cycle boundary padding independently of model choice. To assess whether predictive performance was sensitive to the originally adopted seven-day zero-padding interval, four configurations were evaluated: no padding (0 days), 3-day padding, the original 7-day padding, and 14-day padding. The corresponding datasets contained 2,280, 2,820, 3,540, and 4800 records, respectively.

All other elements of the experimental pipeline, including feature construction, preprocessing, model configuration, randomisation controls, validation procedure, and performance metrics, were kept unchanged across the four configurations. Performance was evaluated using MAE, RMSE, and R^2 . Because increasing the padding interval

changes both the number of observations and the target distribution, MAE, RMSE, and R^2 were interpreted jointly rather than relying on any single metric in isolation.

Sensitivity of TabPFN performance to zero-padding length.

Padding	Records	Added records	MAE	RMSE	R^2
0 days	2280	0	610.90	1,089.22	0.486
3 days	2820	540	446.99	815.14	0.673
7 days	3540	1260	340.18	742.32	0.733
14 days	4800	2520	282.58	712.96	0.646

Performance improved substantially when padding was introduced. Compared with no padding, the 3-day and 7-day configurations progressively reduced MAE and RMSE and increased R^2 . The original 7-day configuration achieved the highest R^2 (0.733), together with substantially lower MAE (340.18) and RMSE (742.32) than the unpadded and 3-day configurations.

Extending the padding interval to 14 days further reduced MAE and RMSE to 282.58 and 712.96, respectively; however, R^2 decreased to 0.646. Therefore, increasing the amount of padding did not produce a consistent improvement across all evaluation metrics. Among the configurations evaluated, the original 7-day padding provided the highest explained variance while retaining substantial improvements in absolute prediction error.

5.3. Leave-one-out cross-validation

To validate the baseline Random 80/20 split, models were evaluated using leave-one-out cross-validation (LOOCV). LOOCV repeatedly trains on all days except one and tests on the held-out day. Table 5 reports LOOCV performance across all evaluated models. TabPFN achieved the lowest MAE (338.09) and highest R^2 (0.709) among all evaluated models, substantially outperforming classical machine learning baselines. The margin of improvement over the strongest gradient boosting model (LightGBM_GPU, MAE = 400.79) is considerable, indicating that the tabular foundation model exhibits superior inductive bias under repeated retraining in the limited-data approach. It is important to note that LOOCV did not yield a meaningful improvement in predictive performance for the TabPFN model. The dataset spans multiple production seasons, providing sufficient temporal variability for the model to learn stable and generalisable relationships. Hence the model becomes less sensitive to individual observations, and the fine-grained resampling provided by LOOCV offers limited additional benefit. Tree-based boosting methods (LightGBM, XGBoost, ExtraTrees, HistGBR, CatBoost) constitute a clear second tier in performance, demonstrating stable non-linear modelling capabilities. In contrast, neural architectures trained from scratch (MLP, TCN, LSTM) exhibit negative R^2 , indicating predictive performance below a naïve mean baseline. This behaviour reflects the sensitivity of high-capacity sequence models to limited seasonal depth and repeated retraining under small sample conditions.

5.4. Temporal hold-out: train 2020–2024, test 2025

This validation represents the most realistic production scenario: training on historical seasons and predicting an unseen future season. The experiment trained models on data from 2020–2024 and evaluated performance on the held-out 2025 season. Results are reported in Table 6. TabPFN achieved the strongest forward-year performance (MAE = 438.53, RMSE = 811.04, R^2 = 0.696). Among classical baselines, tree-based ensemble methods performed well: Random Forest achieved the lowest MAE among non-TabPFN models (MAE = 469.57), followed closely by Histogram Gradient Boosting (MAE = 474.54) and Extra Trees (MAE = 488.32). GPU-based gradient boosting methods (CatBoost and LightGBM) remained competitive but did not outperform the best-performing tree ensembles under this year-shift evaluation. Deep learn-

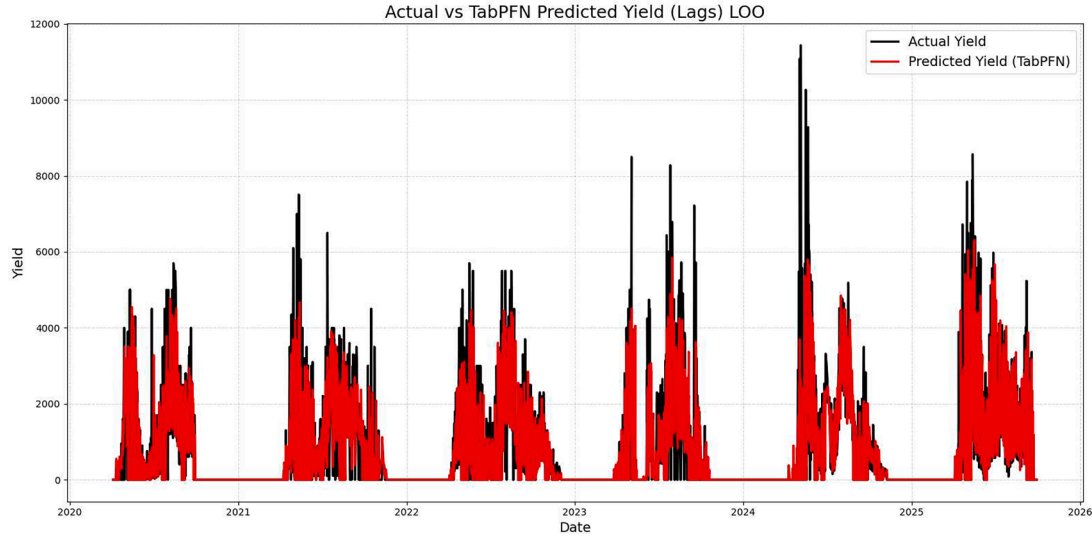


Fig. 10. Multi-year comparison of actual and predicted yield across all seasons (2020–2025).

Table 5

Leave-one-out cross-validation (LOOCV) performance across all evaluated models, ordered by increasing MAE (lower is better).

Model	MAE	RMSE	R^2	Runtime (s)
LightGBM_GPU	400.79	792.55	0.659	3722.58
XGBoost_GPU	403.62	822.31	0.633	1185.59
ExtraTrees	404.63	821.38	0.633	15829.11
HistGBR	415.11	820.10	0.634	943.02
CatBoost_GPU	415.20	803.54	0.649	2971.21
RandomForest	421.72	853.99	0.604	43924.65
KNN	439.93	867.58	0.591	44.26
Ridge	549.74	957.11	0.502	46.99
SVR_RBF	586.71	1132.13	0.303	2261.32
PyTorch_MLP_GPU	868.74	1602.78	−0.396	263.91
TinyTCN_GPU	873.43	1612.67	−0.413	296.06
TinyLSTM_GPU	873.65	1613.17	−0.414	465.22
TabPFN	338.09	731.65	0.709	65825.26

Table 6

Forward-year validation results (Train 2020–2024, test 2025). This is the primary deployment-relevant evaluation.

Model	MAE	RMSE	R^2	Runtime (s)
RandomForest	469.57	873.33	0.6475	9.99
HistGBR	474.54	850.38	0.6657	0.31
ExtraTrees	488.32	874.18	0.6468	3.65
CatBoost_GPU	512.58	891.05	0.6330	0.86
LightGBM_GPU	513.62	876.06	0.6452	0.91
Ridge	546.41	897.58	0.6276	0.01
XGBoost_GPU	575.97	1015.05	0.5238	0.89
KNN	581.70	986.09	0.5505	0.02
PyTorch_MLP_GPU	751.09	1301.83	0.2166	1.17
SVR_RBF	786.45	1397.28	0.0975	0.52
TinyTCN_GPU	1251.73	1901.31	−0.6709	0.43
TinyLSTM_GPU	1308.69	1967.39	−0.7891	0.67
TabPFN	438.53	811.04	0.696	0.60

ing baselines again showed limited generalisation. Although the multi-layer perceptron (MLP) achieved a positive R^2 (0.217), its absolute error was substantially higher than that of tree-based and foundation models. Sequence-based models (TinyLSTM and TinyTCN) produced negative R^2 values and large errors, reinforcing the conclusion that training temporal neural networks from scratch is ineffective in this setting with limited data and discontinuous seasonal structure. Fig. 6 shows Forward-year validation results.

5.5. Cross-approach comparison

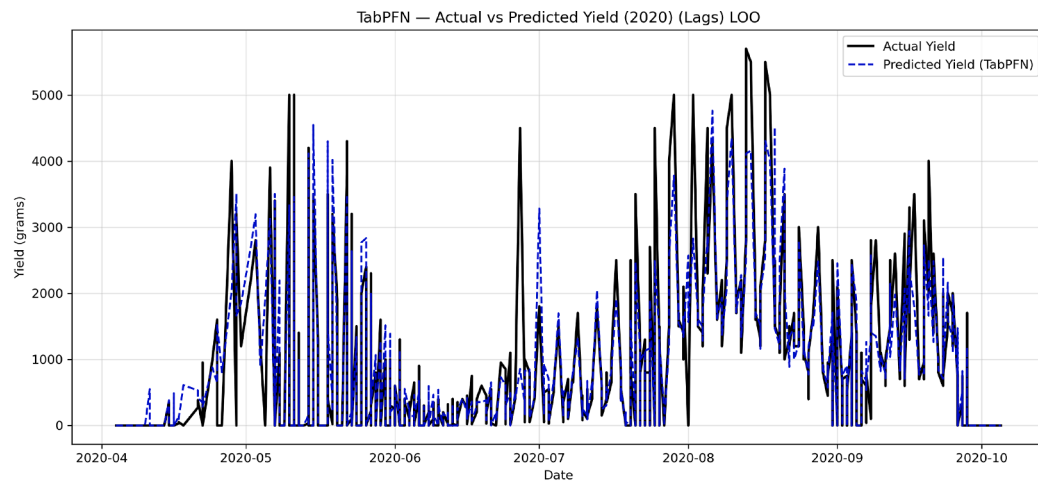
TabPFN shows similar performance under random-split and LOOCV (MAE $\approx$ 340), while forward-year evaluation increases MAE to 438.53, reflecting the challenge of generalising to an unseen production year. In this approach, robust tree ensembles outperform GPU-based boosting methods, whereas deep learning baselines remain weak due to limited data and discontinuous seasonal structure. Figs. 7 and 8 further show that TabPFN's errors increase and R^2 decreases under forward-year validation, indicating sensitivity to interannual temporal shift.

The relationship between actual and predicted yield is illustrated in Fig. 9. The model demonstrates a clear positive correlation, indicating that it successfully captures the underlying production dynamics. Predictions are relatively accurate within low to moderate yield ranges, where most observations are concentrated.

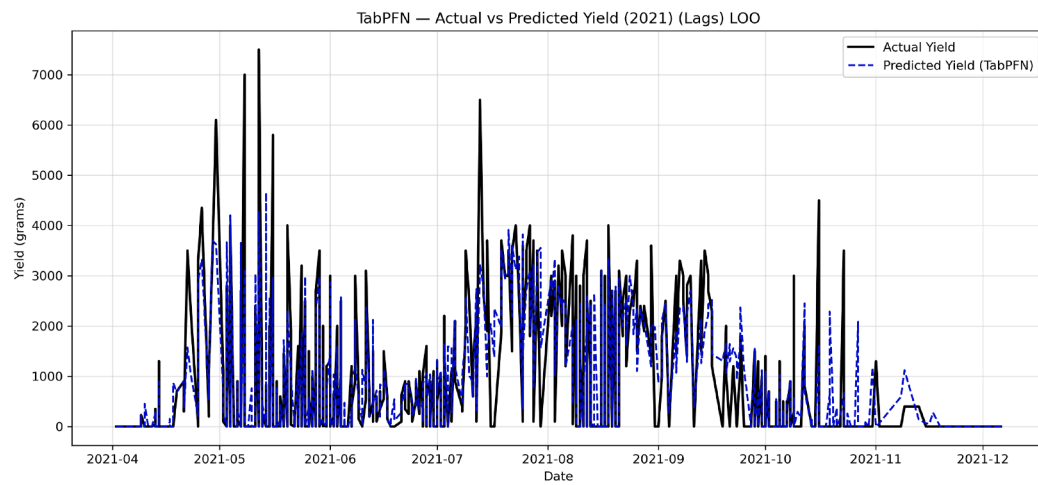
Despite this, the model exhibits a tendency to underestimate high-yield events, as evidenced by the increasing deviation from the perfect fit line at higher yield values. This behaviour suggests that peak production periods are more difficult to capture, likely due to limited environmental variables and the absence of key drivers such as solar radiation and accumulated heat units. Likewise, the temporal prediction performance of the TabPFN model is illustrated in Figs. 10 and 11. Across all seasons, the model successfully captures the general structure of strawberry production cycles, including the onset of harvesting, mid-season peaks, and late-season decline. The alignment between predicted and actual yield indicates that the model effectively learns underlying temporal dynamics. Although, deviations are observed during high-yield periods, where the model tends to underestimate peak production. This behaviour is consistent across multiple seasons and suggests that extreme yield events are more difficult to model, potentially due to the absence of key environmental drivers such as solar radiation and accumulated heat. Despite this limitation, the model demonstrates robust performance in tracking overall seasonal trends and variability.

5.6. Feature importance results

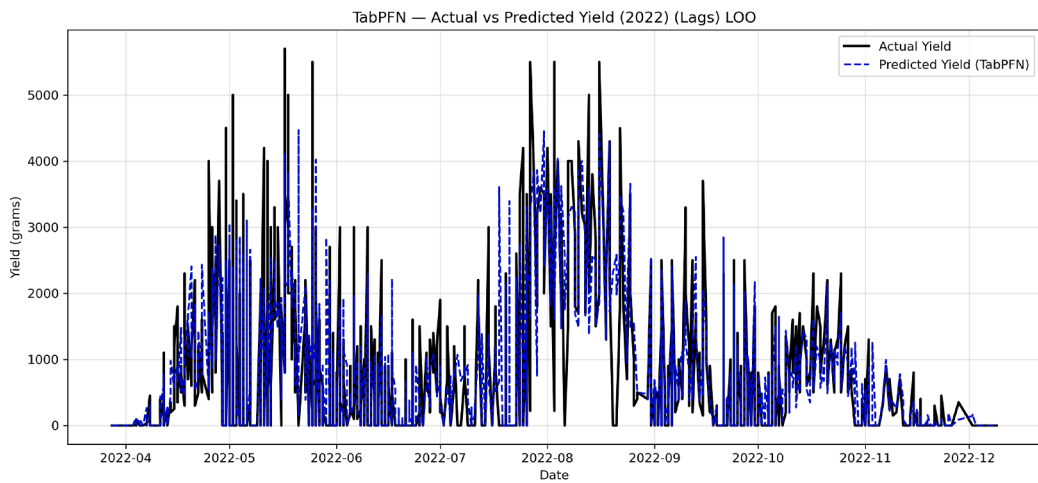
Fig. 12 presents the resulting feature importance ranking based on the increase in MAE observed when each variable was randomly permuted. Features producing larger increases in MAE are interpreted as having greater influence on the predictive performance of the model. The results show that short-horizon yield memory features dominate the importance ranking. In particular, the most recent lag variables (Yield_lag1, Yield_lag3, Yield_lag7, Yield_lag14, and



(a) 2020

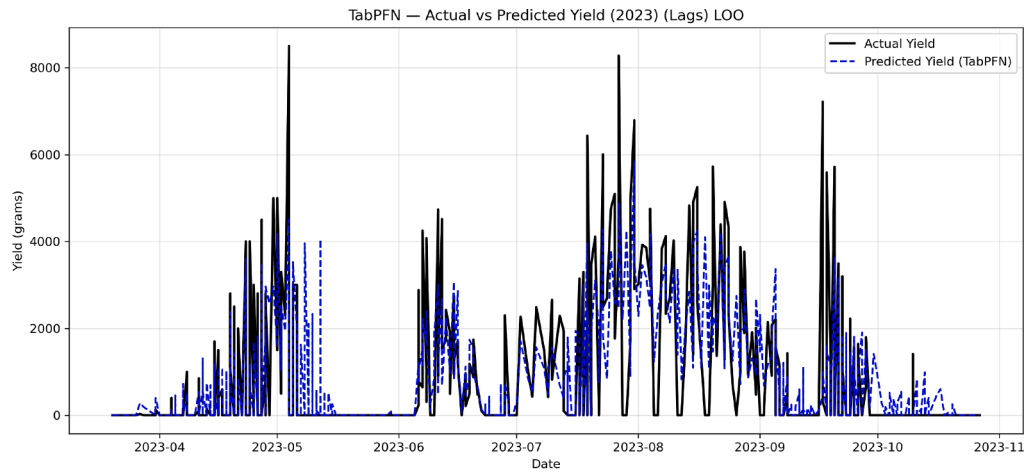


(b) 2021

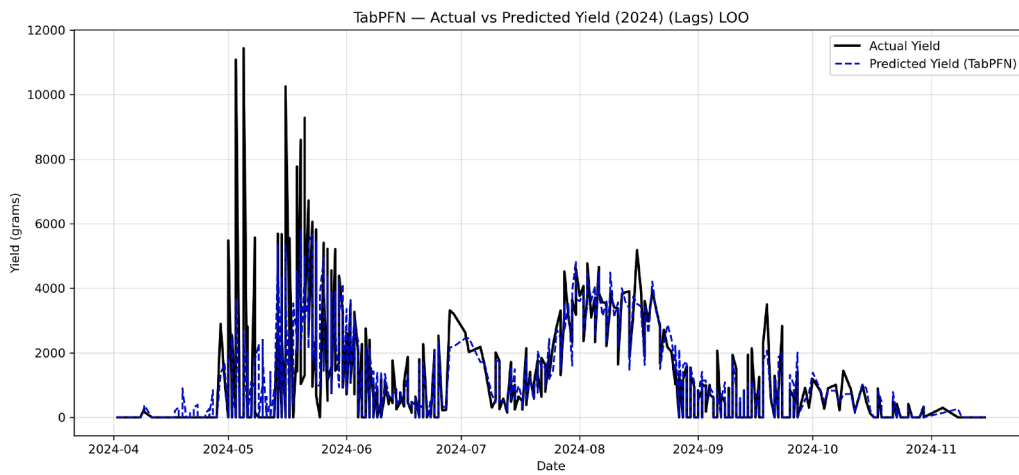


(c) 2022

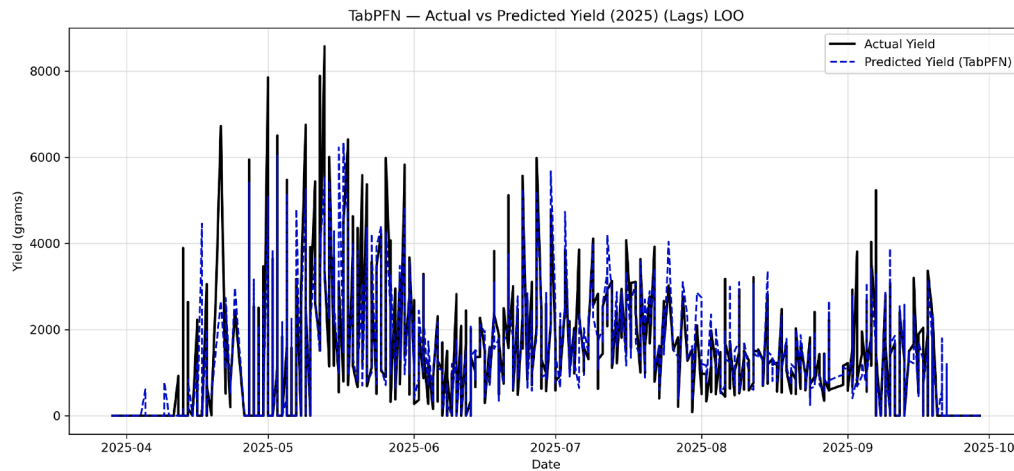
Fig. 11. Year-wise comparison of actual and predicted yield using TabPFN under LOOCV evaluation.



(d) 2023



(e) 2024



(f) 2025

Fig. 11. (Continued).

`Yield_lag21`) contribute strongly to the model predictions. This indicates that daily strawberry production exhibits substantial temporal persistence, where harvest volumes on a given day are strongly influenced by production levels observed in preceding days. Such behaviour reflects the biological nature of strawberry harvesting, where fruit maturation

and harvesting typically occur across consecutive days rather than in isolated events. Environmental variables also play a measurable role in the forecasting process. Temperature and relative humidity variables appear among the important predictors, particularly through interaction terms such as `Yield1_Temp` and `Yield1_RH`. These features combine re-

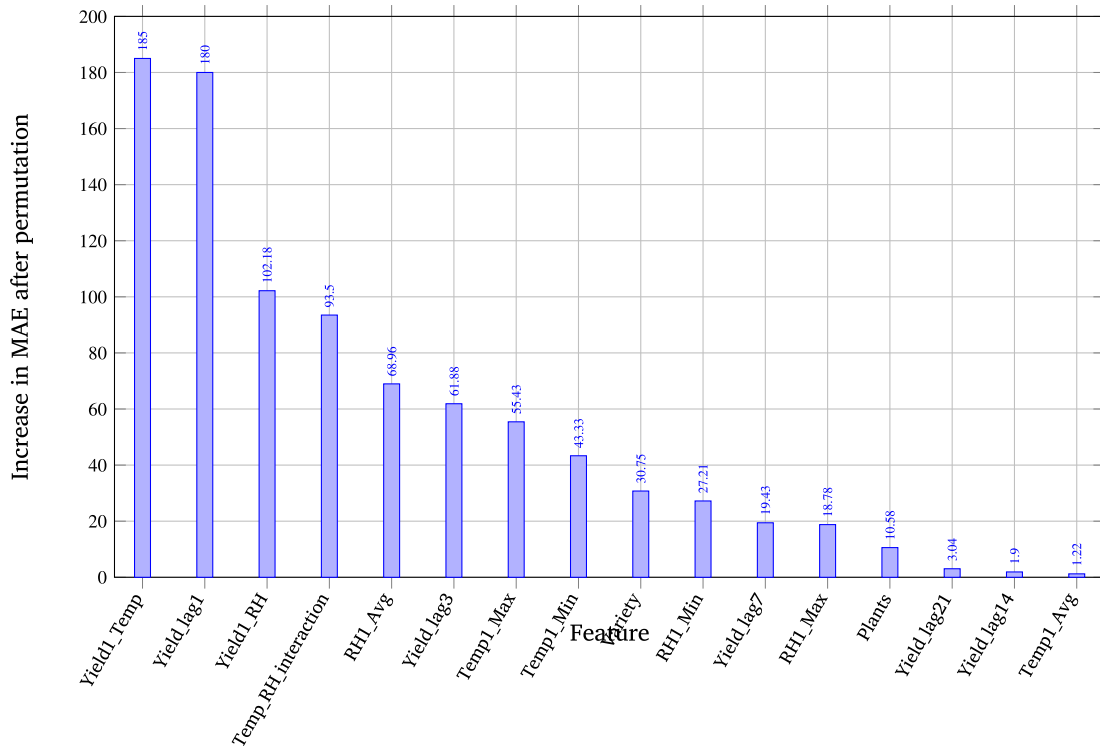


Fig. 12. Permutation-based feature importance for the TabPFN model evaluated on the 2025 test season. The bars represent the increase in mean absolute error when each feature is randomly permuted. Higher values indicate greater contribution to model performance.

cent production levels with current environmental conditions, allowing the model to capture how weather conditions influence fruit development and harvesting dynamics. This suggests that environmental conditions modulate production intensity rather than acting as independent drivers of yield.

6. Discussion

This study demonstrates that credible daily strawberry yield forecasting is achievable under realistic commercial constraints, but that apparent performance is highly sensitive to evaluation design. Random train-test splits produce optimistic error estimates, while forward-year validation exposes the true difficulty of generalising across seasons. The observed increase in error under year-ahead testing is therefore not a weakness, but an indication that the evaluation protocol captures genuine interannual variation in weather, phenology, and management. As a result, forward-year evaluation provides the most reliable estimate of deployment performance in seasonal farm settings. In contrast, tree ensembles offer robust non-linear modelling with implicit regularisation, while TabPFN benefits from a pretrained tabular prior that performed effectively on the small commercial dataset examined here. Model behaviour under limited seasonal data highlights the importance of inductive bias. Tree-based ensembles and the tabular foundation model TabPFN consistently outperformed neural sequence models trained from scratch. This outcome is expected in small, discontinuous time series where strong autocorrelation, structural zeros, and short histories limit the ability of high-capacity temporal models to learn stable representations. In contrast, tree ensembles offer robust non-linear modelling with implicit regularisation, while TabPFN benefits from pretrained tabular structure that transfers effectively to small agricultural datasets.

While weekly plant count provides useful context as a static capacity indicator, it is not a direct proxy for fruit availability on a given day. Forecasting peak yield is likely to benefit from phenology-aware signals that track the developing fruit load, such as flower or truss counts. Prior

work supports the mechanistic relevance of flowering as a yield precursor because flowers directly precede fruit formation, and recent operational systems integrate phenological or visual crop-status signals for improved yield prediction. In practical terms, augmenting the current pipeline with routinely collected flowering/truss observations (manual or computer vision) is a plausible route to improving peak capture without compromising temporal validity.

The dominance of short-horizon yield-history variables, particularly `Yield1_lag1` and `Yield1_lag3`, reflects the continuity of strawberry production across neighbouring harvest days, where fruit development and ripening occur within overlapping production cohorts. The smaller contribution of the longer 7-, 14-, and 21-day yield-history features suggests that broader crop-cycle position remains informative, although recent production state is considerably more influential for daily forecasting. The strong importance of `Yield1_Temp` further indicates that environmental conditions interact with the current production state of the crop. Temperature is an established regulator of strawberry flowering, plant development, and post-bloom fruit growth, with responses also interacting with photoperiod [23,24]. Relative humidity forms part of the crop microclimate together with temperature, soil moisture, and radiation, and these conditions have been shown to influence strawberry fruiting and phenological development [25]. Their interaction with recent yield may therefore help the model identify whether an established production trajectory is likely to persist or change.

Importantly, these features respect temporal causality and avoid leakage, allowing models to generalise meaningfully to unseen seasons. Explaining a substantial fraction of yield variance under forward-year validation indicates that daily strawberry production contains learnable structure even under sparse data conditions.

A recurring error mode is under-prediction around peak harvest days. One plausible explanation is that the available feature set captures recent yield dynamics and accumulated temperature conditions, but does not fully represent several environmental and management factors that influence peak production. Solar radiation affects the energy available for crop growth, day length provides an important seasonal

and photoperiod signal, irrigation inputs influence crop water availability, and growing degree days (GDD), a cumulative measure of thermal time typically calculated from daily temperature relative to a crop-specific base temperature, provide an indicator of accumulated crop development. These factors may be particularly important during rapid increases in production, when similar recent yield, temperature, and humidity conditions can still be associated with different peak-yield potential. Their absence may therefore contribute to the observed underestimation at high production levels, although their individual effects cannot be quantified from the present dataset.

Several limitations should be noted. Although the dataset spans six consecutive production seasons and includes multiple strawberry cultivars grown in both polytunnel and glasshouse environments, all observations originate from a single commercial farm. The present study therefore demonstrates temporal generalisation to an unseen season within the same commercial setting, rather than external generalisation across independent farms or geographical regions. The proposed forecasting pipeline, including causal feature construction, leakage-aware preprocessing, conventional machine-learning models, TabPFN, and forward-year validation, can in principle be applied to other production environments; however, the relationships learned by the fitted models should not be assumed to transfer unchanged. Differences in regional climate and season timing, cultivar-specific phenology and productivity, and production-system characteristics may alter the relationship between historical yield, environmental conditions, and subsequent production. For example, closed hydroponic systems may be influenced more strongly by controlled irrigation, fertigation, and greenhouse climate management than traditional tunnel production. Application of the pipeline to such settings would therefore require independent external validation and, where necessary, local recalibration or retraining using data from the target production environment.

The forward-year evaluation implies exposure to inter-annual weather variability and operational shift, which can change the relationship between temperature/humidity and yield. Because only a subset of environmental drivers was available, the models necessarily learn conditional relationships under partial observability. This does not invalidate the results; rather, it clarifies the operational approach being targeted: forecasting with the sensors farms typically deploy. Expanding sensing to include radiation and precipitation proxies is expected to reduce unexplained variance and improve responsiveness to rapid environmental changes.

Despite these limitations, the permutation importance analysis indicates that the forecasting model relies primarily on variables that are consistent with known biological and environmental processes governing strawberry yield formation. This provides evidence that the model captures meaningful production dynamics rather than relying on spurious statistical relationships within the dataset.

7. Conclusion and future work

This study examined daily strawberry yield forecasting under realistic commercial farm conditions characterised by limited historical depth, discontinuous seasonal production, and heterogeneous operational data. A leakage-aware forecasting framework was developed that combines biologically informed yield lags, rolling environmental summaries, and strict temporal validation. Model performance was evaluated at daily resolution under progressively more rigorous protocols, culminating in forward-year generalisation from 2020–2024 to an unseen 2025 production season. The results show that meaningful and stable predictive performance is achievable despite these constraints. Under forward-year validation, which best reflects operational deployment, TabPFN achieved the strongest performance ($\text{MAE} = 438.53$, $R^2 = 0.696$), outperforming classical machine-learning baselines. Tree-based ensemble models also generalised reliably, while neural sequence architectures trained from scratch performed poorly. These findings indicate that, in small-to-medium farm-scale datasets with seasonal discontinu-

ities, strong inductive bias and regularisation are more important than model depth or temporal complexity. A key methodological contribution of this work is the explicit demonstration that evaluation design substantially affects apparent forecasting performance. Random splitting was shown to overestimate generalisation in seasonal agricultural data, whereas strict forward-year validation provides a more conservative and deployment-relevant assessment. Retaining genuine zero-yield periods and enforcing causal feature alignment further strengthened the credibility of the evaluation. Together, these results highlight the importance of leakage-aware modelling practices in agricultural forecasting. From an operational perspective, the findings suggest that daily yield forecasting can be integrated into commercial decision-support systems using routinely collected farm data. Although forecasting errors remain non-trivial, the achieved predictive performance indicates that daily yield dynamics contain substantial predictable structure. The robustness of tree ensembles and tabular foundation models under interannual shift supports their suitability for real-world deployment where data are limited and evolving.

Several directions for future work emerge naturally. Multi-farm evaluation is required to assess external validity across different geographical regions, cultivars, and production systems, including polytunnel, glasshouse, and closed hydroponic production. Future studies should compare direct application of the forecasting pipeline with local recalibration or retraining using data from the target farm. Additional agronomic and environmental variables should be investigated in future work, particularly solar radiation, day length, irrigation inputs, and growing degree days. Solar radiation could be obtained from on-farm or meteorological measurements, while day length could be approximated from date and geographical location or suitable meteorological records. Irrigation timing and volume could be incorporated from farm-management records, and GDD could be derived from historical temperature measurements. The value of these additional variables should then be evaluated under forward-year validation, with particular attention to forecasting performance during peak production periods.

CRedit authorship contribution statement

Salman Ahmed: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Nicholas HM Caldwell:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Human animal statement

This study did not involve human participants or animals.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Salman Ahmed and Nicholas Caldwell reports financial support was provided by University of Suffolk. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by Innovate UK, supported under Grant Nos. 10122783 (*Agri-F2P: Feasibility Project for AI-Driven Forms Management in Agriculture*) and 10129638 (*Agri-KG: Transforming Agriculture with an AI-Powered Knowledge Graph*). We would also like to thank our industry partner, Diometer, particularly Kevin Gooding, Dave Chen, and Alan Jones, for their invaluable collaboration, and our farm partner, Boxford Farms Ltd, for providing domain expertise and access to operational datasets.

Data availability

The data that has been used is confidential.

References

- [1] Expert Market Research, Fresh strawberry market report and forecast 2025–2034, 2025. Global fresh strawberry market value reported for 2024 with forecast to 2034, <https://www.marketresearch.com/Expert-Market-Research-v4220/Fresh-Strawberry-Forecast-41222944/>.
- [2] Department for Environment, Food & Rural Affairs (DEFRA), Horticulture statistics 2024, 2025. Published 26 June 2025. Includes UK strawberry crop value for 2024, <https://www.gov.uk/government/statistics/latest-horticulture-statistics/horticulture-statistics-2024>.
- [3] P. Setiya, A. Satpathi, A. Nain, Predicting rice yield based on weather variables using multiple linear, neural networks, and penalized regression models, *Theor. Appl. Climatol.* 154 (2023) 1–11. <https://doi.org/10.1007/s00704-023-04563-5>
- [4] A. Satpathi, N. Chand, P. Setiya, R. Ranjan, A.S. Nain, D.K. Vishwakarma, K. Saleem, A.J. Obaidullah, K.K. Yadav, O. Kisi, Evaluating statistical and machine learning techniques for sugarcane yield forecasting in the tarai region of North India, *Comput. Electron. Agric.* 229 (2025) 109667. <https://doi.org/10.1016/j.compag.2024.109667>
- [5] P. Setiya, A.S. Nain, A. Satpathi, Comparative analysis of SMLR, ANN, elastic net and LASSO based models for rice crop yield prediction in Uttarakhand 2024, *MAUSAM* 75 (1) (2024) 191–196. <https://doi.org/10.54302/mausam.v75i1.3576>
- [6] A. Døving, F. Måge, Prediction of strawberry fruit yield, *Acta Agric. Scand. B - Soil Plant Sci.* 51 (1) (2001) 35–42. <https://doi.org/10.1080/090647101317187870>
- [7] T.B. Pathak, S.K. Dara, A. Biscaro, Evaluating correlations and development of meteorology based yield forecasting model for strawberry, *Adv. Meteorol.* 2016 (2016) 9525204. <https://doi.org/10.1155/2016/9525204>
- [8] M.L. Maskey, T.B. Pathak, S.K. Dara, Weather based strawberry yield forecasts at field scale using statistical and machine learning models, *Atmosphere* 10 (7) (2019) 378. <https://doi.org/10.3390/atmos10070378>
- [9] H.S. Sim, et al., Prediction of strawberry growth and fruit yield based on environmental and growth data in a greenhouse for soil cultivation with applied autonomous facilities, *Hortic. Sci. Technol.* (2020). Publisher PDF, <https://www.hst-j.org/articles/pdf/Qbad/kshs-2020-038-06-8.pdf>.
- [10] J.S. Jo, et al., Prediction of strawberry fruit yield based on cultivar-specific growth models in the tunnel-type greenhouse, *Hortic. Environ. Biotechnol.* (2022). <https://doi.org/10.1007/s13580-021-00416-0>
- [11] S. Liu, Y. Ampatzidis, C. Zhou, W.S. Lee, AI-driven time series analysis for predicting strawberry weekly yields integrating fruit monitoring and weather data for optimized harvest planning, *Comput. Electron. Agric.* 233 (2025) 110212. <https://doi.org/10.1016/j.compag.2025.110212>
- [12] Q. Yang, L. Liu, J. Zhou, M. Rogers, Z. Jin, Predicting the growth trajectory and yield of greenhouse strawberries based on knowledge-guided computer vision, *Comput. Electron. Agric.* 220 (2024) 108911. <https://doi.org/10.1016/j.compag.2024.108911>
- [13] S. Yoon, J.S. Jo, S.B. Kim, H.S. Sim, S.K. Kim, D.S. Kim, Prediction of strawberry yield based on receptacle detection and Bayesian inference, *Heliyon* 9 (3) (2023) e14546. <https://doi.org/10.1016/j.heliyon.2023.e14546>
- [14] S. Yamazaki, M. Aono, Y. Kiriwa, A strawberry yield prediction scheme suitable for non-parametric regression using general-purpose IoT nodes, *Discov. Internet Things* 5 (2025) 135. <https://doi.org/10.1007/s43926-025-00219-0>
- [15] A. Ayall, et al., Enhancing strawberry yield forecasting with backcasted IoT sensor data and machine learning, 2025. arXiv preprint, <https://arxiv.org/html/2504.18451v1>.
- [16] C. Bergmeir, R.J. Hyndman, B. Koo, A note on the validity of cross-validation for evaluating autoregressive time series prediction, *Comput. Stat. Data Anal.* 120 (2018) 70–83. <https://doi.org/10.1016/j.csda.2017.11.003>
- [17] V. Cerqueira, L. Torgo, I. Mozetič, Evaluating time series forecasting models: an empirical study on performance estimation methods, *Mach. Learn.* (2020). <https://doi.org/10.1007/s10994-020-05910-7>
- [18] D.R. Roberts, V. Bahn, S. Ciuti, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, *Ecography* 40 (8) (2017) 913–929. <https://doi.org/10.1111/ecog.02881>
- [19] T. Sugiyama, et al., Simulation of strawberry yield using dry matter distribution based on the potential growth of the sink-source organs, *Front. Plant Sci.* (2025). <https://doi.org/10.3389/fpls.2025.1544735>
- [20] M. Beddows, G. Leontidis, A multi-farm global-to-local expert-informed machine learning system for strawberry yield forecasting, *Agriculture* 14 (6) (2024) 883. <https://doi.org/10.3390/agriculture14060883>
- [21] R. Shao, X. Bi, Transformers meet small datasets, *IEEE Access* 10 (2022) 118454–118464. <https://doi.org/10.1109/ACCESS.2022.3221138>
- [22] N. Hollmann, S. Müller, K. Eggensperger, F. Hutter, TabPFN: a transformer that solves small tabular classification problems in a second, 2023. [arXiv:2207.01848](https://arxiv.org/abs/2207.01848) [cs.LG]
- [23] S. Serçe, J.F. Hancock, The temperature and photoperiod regulation of flowering and running in the strawberries, *Fragaria chiloensis*, *F. virginiana*, and *F. x ananassa*, *Sci. Hortic.* 103 (2) (2005) 167–177. <https://doi.org/10.1016/j.scienta.2004.04.017>
- [24] S.Y. Wang, M.J. Camp, Temperatures after bloom affect plant growth and fruit quality of strawberry, *Sci. Hortic.* 85 (3) (2000) 183–199. [https://doi.org/10.1016/S0304-4238\(99\)00143-0](https://doi.org/10.1016/S0304-4238(99)00143-0)
- [25] M.A. Lee, A. Monteiro, A. Barclay, J. Marcar, M. Miteva-Neagu, J. Parker, A framework for predicting soft-fruit yields and phenology using embedded, networked microsensors, coupled weather models and machine-learning techniques, *Comput. Electron. Agric.* 168 (2020). <https://doi.org/10.1016/J.COMPAG.2019.105103>