Machine learning in tea industry: data-driven approaches for quality and sustainability

Fuquan Gao[#], Shuyan Wang[#] and Xiaomin Yu[*] [iD]

Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Haixia Institute of Science and Technology and College of Horticulture, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[#] Authors contributed equally: Fuquan Gao, Shuyan Wang
[*] Corresponding author, E-mail: xmyu0616@fafu.edu.cn

Abstract

Tea holds significant cultural, economic, and nutritional value globally, yet the industry faces persistent challenges including quality inconsistency, climate change impacts, labor shortages, and inefficient conventional production methods. Machine learning (ML) has emerged as a powerful solution to address these challenges through data-driven decision-making, automation, and resource optimization across the entire production chain. This review systematically examines ML applications throughout tea industry, from cultivation and harvesting to processing and quality assessment. We analyze the development of comprehensive ML pipelines encompassing data acquisition, feature engineering, model development, optimization, evaluation, and deployment. Key technological advancements include automated monitoring systems for enhanced productivity, non-destructive spectroscopic and imaging techniques for improved quality assessment, and precision resource management for sustainable production. Despite its transformative potential, widespread ML adoption faces implementation barriers, including limited data availability, scalability issues, and integration with established practices. To overcome these barriers, we highlight strategic approaches such as advanced preprocessing techniques and domain-specific feature engineering to mitigate data limitations, resource-efficient ML architectures tailored for constrained environments, and user-centered interfaces that effectively bridge computational insights with traditional expertise. By synthesizing theoretical frameworks with practical implementation strategies, this review provides researchers, industry stakeholders, and practitioners with essential knowledge to advance sustainable and efficient tea industry through targeted ML integration.

Introduction

Tea, one of the world's most consumed non-alcoholic beverages, holds significant cultural and economic importance globally. Originating in ancient China, tea cultivation has evolved into a major international industry spanning over 60 countries and regions, with annual production reaching 6.7 million metric tons as of 2022 (FAOSTAT, 2024). Beyond its economic significance, tea is renowned as a rich source of bioactive compounds, including polyphenols, theanine, terpenes, and caffeine, which contribute to its numerous health benefits[1,2]. Recognizing its profound cultural heritage and nutritional value, the United Nations designated May 21st as International Tea Day in 2019.

Despite its prominence, the tea industry faces critical challenges that threaten its sustainability. Quality variability, which is driven by complex interactions between agronomic practices, varietal diversity, harvesting methods, and post-harvest processing, erodes consumer trust and market stability[3–5]. These issues are compounded by climate change, as tea plants exhibit acute sensitivity to temperature fluctuations, droughts, and erratic rainfalls, leading to unpredictable yields and persistent quality decline[6–8]. Concurrently, labor shortages and rising operational costs increasingly strain traditional labor-intensive cultivation practices[9]. Conventional production methods, marked by outdated infrastructure, excessive agrochemical application, and energy-intensive processing, further undermine economic and environmental resilience[10].

Machine learning (ML), a transformative subfield of artificial intelligence (AI), has emerged as a promising solution to these challenges. By analyzing complex, multidimensional datasets and identifying latent patterns, ML enables automation, predictive decision-making, and precision resource management in agriculture[11]. Prior studies have demonstrated ML's efficacy in critical agricultural applications such as pest and disease detection[12,13], yield forecasting[14], resource optimization[15], and environmental monitoring[16], showcasing its potential to modernize farming while promoting sustainability.

Recent reviews have highlighted AI's expanding role in advancing precision agriculture, quality control, and IoT-driven automation in tea industry[17–19]. Computer vision (CV) and spectroscopy integrated with ML have automated tea bud detection and non-destructive quality assessment, while blockchain technologies and IoT platforms have enhanced traceability and real-time monitoring. However, practical guidance on implementing ML in real-world tea industry remains limited.

This review provides a focused examination of ML applications across the tea industry value chain. By synthesizing theoretical frameworks with practical methodologies–from data acquisition and feature engineering to model optimization and deployment–this review equips researchers, industry stakeholders, and practitioners with the essential knowledge to enhance efficiency, sustainability, and quality throughout tea industry systems. Through addressing key implementation challenges including scalability, integration with traditional practices, and resource constraints, this review aims to catalyze the transition from experimental ML frameworks to practical, adaptable agricultural solutions with measurable impacts on productivity and sustainability.

Overview of ML fundamentals and common algorithms in tea research

ML in tea research centers on three core methodologies: traditional machine learning (TML), deep learning (DL), and ensemble learning (EL), each addressing distinct challenges across the tea industry chain (Fig. 1).

TML

TML models relationships between input data and target variables through statistical learning, prioritizing interpretability and efficiency with structured datasets[20]. Its modular workflow, involving preprocessing, feature extraction, and classification, encompasses supervised, unsupervised, semi-supervised, and reinforcement learning approaches[21–23].

Supervised learning algorithms dominate tea research applications. Support Vector Machines (SVM) identify optimal hyperplanes to maximize margins between classes in high-dimensional space, making them effective for tea classification[24–27], disease detection[28,29], quality assessment[30], adulteration identification[31], and plantation mapping[32]. Random Forests (RF) combine multiple decision trees (DTs) through bootstrap aggregation, mitigating overfitting while providing feature importance rankings. This approach has been applied to predict tea yields[33], monitor fermentation stages[34], classify tea origins[35] and evaluate hazard risks[36]. Linear Discriminant Analysis (LDA) projects features onto axes that maximize inter-class variance, simplifying multidimensional chemical profiles[37–39]. K-Nearest Neighbors (KNN) classifies samples based on majority votes from their k nearest neighbors, using distance metrics to grade tea quality with small datasets, while Naïve Bayes (NB) assumes feature independence to compute posterior probabilities, enabling rapid tea classification[40–43].

Unsupervised learning extracts patterns from unlabeled data. Principal Component Analysis (PCA) reduces dimensionality by projecting features onto orthogonal axes of maximum variance, resolving multicollinearity in tea spectral and chemical datasets[44,45]. K-Means Clustering partitions data into k groups by minimizing intra-cluster variance, with applications in automated tea leaf harvesting systems[46]. Hierarchical clustering builds nested clusters by iteratively merging or splitting groups, useful for identifying relationships among tea varieties based on chemical profiles[47].
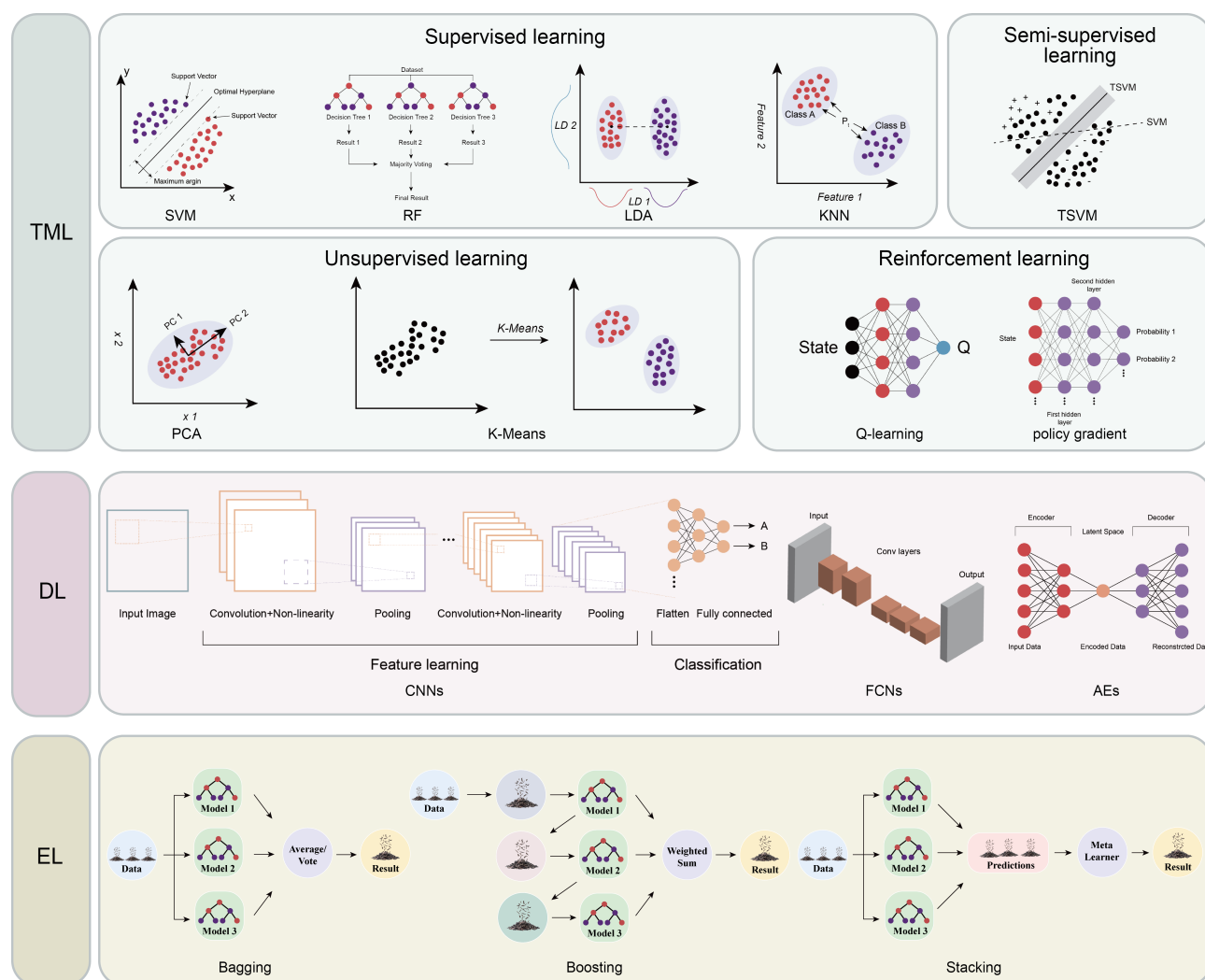
Fig. 1 Overview of machine learning (ML) methodologies. Visualization of ML approaches organized into three principal categories: traditional machine learning (TML), deep learning (DL), and ensemble learning (EL). TML encompasses supervised learning (exemplified by SVM, RF, LDA, and KNN), unsupervised learning (featuring PCA and K-Means Clustering), semi-supervised learning (represented by TSVM), and reinforcement learning (depicting Q-learning and policy gradient). DL demonstrates specialized neural network architectures, including CNNs, FCNs, and AEs. EL portrays integration strategies including bagging, boosting, and stacking.

Semi-supervised methods bridge labeled and unlabeled data to address scenarios where annotation is costly or requires expertise[48]. Operating on cluster and manifold assumptions, these methods improve generalization while minimizing labelling efforts. Self-training iteratively labels high-confidence predictions to expand the training set while co-training leverages multiple views of the data to enhance model robustness. Graph-based approaches propagate labels through similarity networks, particularly effective for spectral data classification[49]. Transductive Support Vector Machines (TSVM) extend traditional SVMs by optimizing decision boundaries to maximize margins in low-density regions of both labeled and unlabeled data, maintaining accuracy even when labeled samples are limited[50]. Semi-supervised learning has been combined with image processing to identify tender tea leaves under varying conditions[51], though broader adoption in tea-related tasks remains underexplored.

Reinforcement learning (RL) focuses on training agents to learn optimal decision-making policies by interacting with environments, aiming to maximize cumulative rewards over time[23]. Unlike supervised learning, RL operates on a trial-and-error framework, where agents iteratively refine strategies based on feedback from states and reward signals, which is crucial for applications requiring foresight, such as robotics, resource allocation, and game AI[52]. Recent methodological advancements, including Q-learning for value-based optimization, Monte Carlo techniques for episodic sampling, and policy gradient methods for direct policy updates, have strengthened RL's ability to handle stochastic and partially observable agricultural environments. Deep reinforcement learning (DRL) architectures, which integrate DL networks with RL decision-making (e.g., frameworks like DeepSeek), further enhance scalability and generalization for complex tasks[53]. In tea-harvesting robotics, a modified pointer network trained via the REINFORCE policy gradient (augmented by a critic network baseline) has demonstrated efficiency in solving picking sequence planning[54].

DL

DL methods process unstructured data like images and sensor streams through layered neural networks, combining feature design and modeling into unified systems that reduce manual preprocessing[55]. These systems learn patterns directly from raw inputs, addressing complex agricultural challenges such as identifying disease markers on tea leaves or distinguishing cultivars through spectral variations.

Convolutional Neural Networks (CNNs), foundational for image-based tasks in tea research, specialize in spatial feature extraction through learnable filters that capture hierarchical patterns in images. These filters detect low-level features in initial layers and progressively build more complex representations in deeper layers. Several landmark CNN architectures have been adapted for tea applications: LeNet established the convolutional-pooling pattern for tea leaf disease detection[56]; ResNet introduced residual connections to mitigate vanishing gradients in deeper networks, enabling superior feature extraction for disease classification[57]; YOLO's unified detection framework improved real-time efficiency for tea bud[58,59] and disease detection[60,61]; and R-CNN enhanced precision in tea pest localization and disease spot identification through region proposal networks[62].

Fully Convolutional Networks (FCNs) extend CNNs for pixel-level segmentation by replacing fully connected layers with convolutional layers, allowing end-to-end training and dense predictions on arbitrary-sized inputs[63]. This approach has facilitated precise determination of tea plucking points under variable field conditions[64] and supported the development of automated guidance systems for tea plantations[65].

Recurrent Neural Networks (RNNs) process sequential data through feedback connections that maintain internal memory, making them ideal for time-series analysis in agricultural applications[66]. Unlike feedforward networks, RNNs share parameters across time steps while modeling temporal dependencies. In tea research, RNNs have been integrated with CNNs for accurate tea plantation mapping from multi-temporal imagery[67]. Long Short-Term Memory (LSTM) networks, an advanced RNN variant designed to capture long-term dependencies, have been applied to tea disease detection[68], growth monitoring[69,70], and pest control[71].

Autoencoders (AEs) compress input data into compact latent representations through encoder-decoder structures, proving valuable for dimensionality reduction, anomaly detection, and noise filtering. Variational autoencoders (VAEs), an extension of traditional AEs, introduce probabilistic encoding by learning distributions of latent variables, yielding flexible and generalizable representations of complex data patterns. These architectures have demonstrated effectiveness in recognizing tea diseases[72,73] and differentiating tea clones[74], even under noisy or distorted conditions.

Artificial Neural Networks (ANNs), such as Multi-Layer Perceptrons (MLP), are supervised models inspired by biological neural networks, comprising input, hidden, and output layers[11]. Their functionality is determined by activation functions, network structure, and learning algorithms. ANNs have been widely applied to structured data tasks, including authenticating tea origins[5], predicting antioxidant activity from biochemical profiles[75], optimizing tea drying conditions[76], and classifying tea types and grades[77,78].

EL models

EL enhances predictive accuracy and robustness by combining multiple algorithms, drawing on their strengths while reducing individual weaknesses. Common EL techniques include bagging (e.g., RF), boosting (e.g., AdaBoost, XGBoost, and LightGBM), and stacking. Bagging reduces variance by aggregating predictions from bootstrapped models, boosting iteratively refines weak learners by focusing on misclassified samples, and stacking integrates heterogeneous models through meta-learners to optimize final prediction[79]. These approaches are particularly effective for both classification and regression tasks, especially when base models exhibit high variance or bias.

In tea research, EL has delivered superior performance across cultivation and production tasks. RF, Bagging, and AdaBoost have outperformed conventional models in estimating daily evapotranspiration rates, a critical metric for irrigation planning in tea plantations[80]. For complex nonlinear problems like carbon flux prediction, hybrid approaches such as the nonlinear ensemble-generalized regression neural network (NLE-GRNN) have significantly improved the prediction accuracy for the net ecosystem exchange[81]. Gradient Boosting (GB) has been effective for the geographic origin authentication of Wuyi rock tea, achieving high accuracy with reduced data complexity by iteratively optimizing feature interactions[5].

Applications of ML in the tea industry

ML integrates seamlessly across the entire tea industry chain, spanning cultivation, processing, and quality control (Fig. 2). In cultivation, ML-powered systems analyze imaging and sensor data to automate disease detection, pest identification, yield forecasting, tea harvest, and resource allocation. For processing and classification, ML streamlines tea categorization by type, variety, and origin, ensuring authenticity and standardization. In quality evaluation, ML models replace subjective assessments with data-driven analyses of sensory traits and chemical profiles, establishing objective metrics for quality assurance.
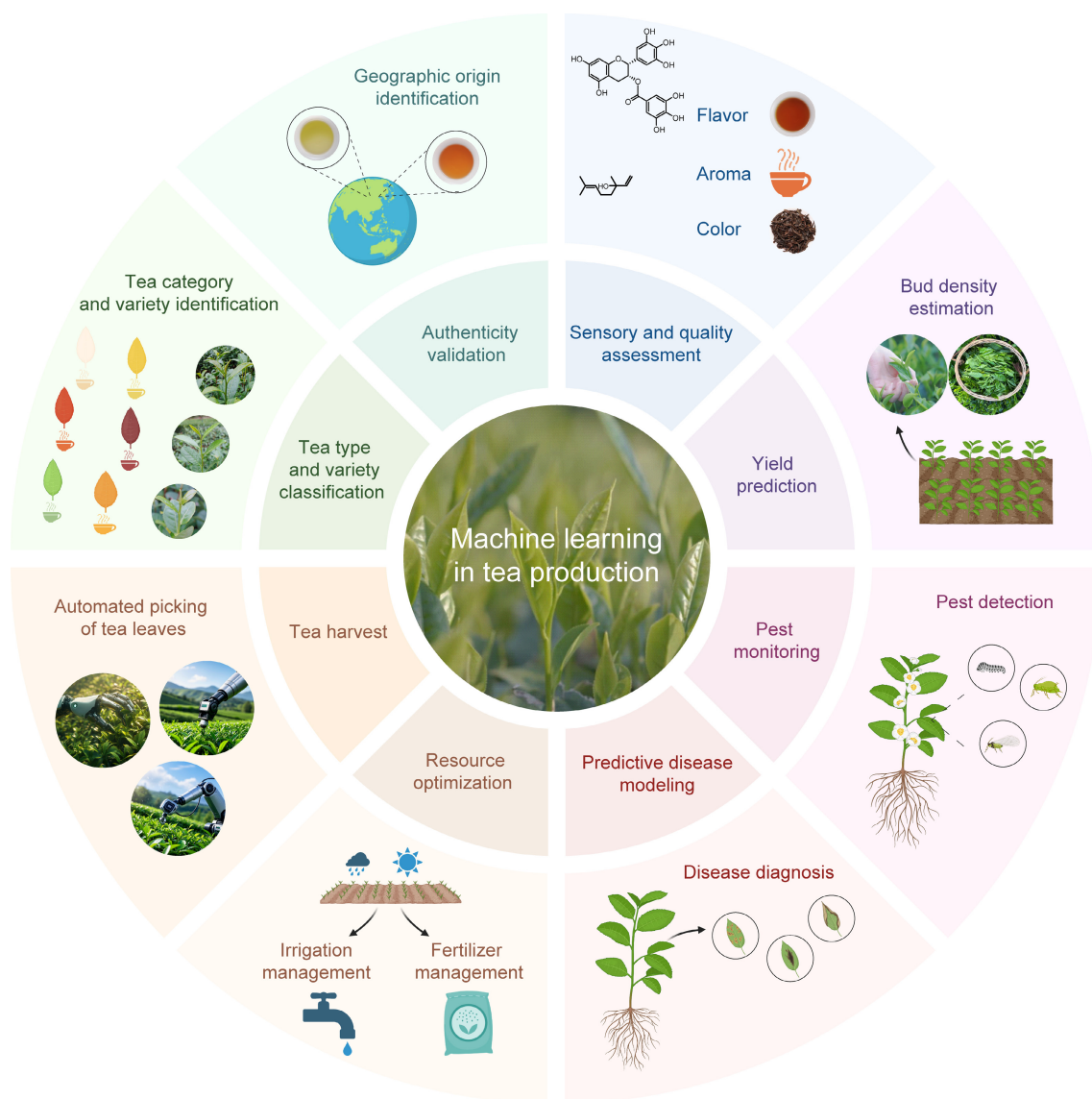
Fig. 2 ML applications in tea industry systems. This schematic depicts key applications of MLs throughout the tea value chain, including geographic origin authentication, tea variety classification, sensory quality assessment, yield prediction, resource optimization, pest monitoring, disease diagnosis, and automated harvesting.

Tea quality and sensory evaluation

The integration of ML with non-invasive spectroscopic and imaging technologies has significantly improved tea quality assessment, enabling rapid, objective, and holistic evaluations that enhance precision in multi-sensory profiling.

Traditional methods for tea color assessment, which rely on human judgment, have been revolutionized by ML-driven approaches. For instance, CV combined with DT AdaBoost algorithms has achieved 98.50% prediction accuracy in color grading of needle-shaped green tea[82]. Similarly, MLP models applied to colorimetric data have improved the reliability of instrumental color assessments for Sichuan Dark Tea[83].

In flavor evaluation, advancements in electronic sensing technologies paired with ML have yielded remarkable results. A multi-task back-propagation neural network (MBPNN) integrated with electronic nose data has attained 99.83% accuracy for organic green tea grading and price prediction, surpassing conventional RF and SVM models[84]. Similarly, ant colony optimization (ACO)-SVM hybrid methods using electronic tongue and near-infrared spectroscopy (NIRS) data have achieved 93.56% accuracy in black tea grading[30].

Research combining electronic nose, tongue, and eye technologies with chemometrics has reported 100% accuracy in tea grade classification and superior prediction of chemical components ($R^2 > 0.97$), with RF models on fused signals outperforming single-sensor approaches[85]. More recently, a multi-modal fusion of CV and NIRS with a temporal convolutional network (TCN) has elevated black tea quality assessment to 98.2% accuracy by synthesizing appearance, chemical, and spectral data[86], underscoring the power of data fusion in quality control.

Hyperspectral imaging (HSI) has become essential for tea quality analysis, offering unique insights through simultaneous spectral and spatial feature extraction when combined with ML models. For example, HSI with spectral-texture fusion and LSSVM models has achieved 99.57% accuracy in Dianhong black tea evaluation[87]. The integration of HSI with olfactory visualization systems and SVM has improved green tea classification accuracy to 92%, outperforming single-sensor methods[88]. In a cross-category application, the feature fusion of electronic nose and HS data optimized via XGBoost has demonstrated robust performance ($R^2 = 0.900$, RMSE = 1.895) for polyphenol content estimation across black, green, and yellow

teas[89]. Collectively, these technological advances establish objective, reproducible analysis systems that complement human expertise while overcoming the limitations of traditional evaluation methods.

Tea harvest automation

The integration of ML into tea harvesting has shifted traditional non-selective mechanical methods toward precision-based selective systems, driven by advancements in intelligent target identification and automation. ML methodologies now underpin three core components of the harvesting pipeline: bud detection, picking point determination, and integrated system optimization, which enhance efficiency and accuracy.

Accurate bud identification forms the foundation of selective harvesting. Recent innovations include an optimized YOLO v3 algorithm achieving over 90% detection accuracy across four distinct morphological categories of tea targets[90]. For real-time classification, a DenseNet-20 based CNN model has attained 96.9% accuracy using augmented real-time datasets (262 images), outperforming SVM (73.1%)[91]. In resource-constrained environments, a K-Means Clustering combining with image morphology processing has attained over 80% recognition accuracy while minimizing computational costs[92].

Precise localization of picking points is critical for minimizing mechanical damage. One approach has integrated YOLO-v3 detection with Fast-SCNN semantic segmentation, skeleton extraction, and minimum bounding rectangle algorithms, achieving 83% accuracy in picking point localization[93]. More advanced approaches have adopted a cascaded Faster R-CNN and FCN framework, improving accuracy to 84.9% through hierarchical stem segmentation[64]. Mask R-CNN has subsequently streamlined this process while delivering comparable robustness in complex environments[94].

Modern systems unify detection, localization, and motion planning into cohesive workflows. A compressed YOLO-v3 network fused with point cloud processing and spatial path planning has demonstrated success rates of 85.16% for bud detection, 78.90% for picking point localization, and 80.23% for motion planning[95]. These integrated systems represent cutting edge ML applications in tea harvesting, achieving precision and efficiency unattainable with conventional methods.

Yield prediction, resource optimization, and growth monitoring

ML has become integral to tackling issues in tea cultivation, particularly in yield forecasting, resource management, and plant health tracking. In yield prediction, the XGBoost regressor has delivered higher accuracy with fewer inputs (MAE: 0.0093 t/ha, RMSE: 0.120 t/ha) compared to traditional crop simulation models[96]. In small-scale hillside tea farms, YOLOv5 paired with on-site sampling has analyzed bud density with 29.56% relative error, supporting early management decisions[97]. A spatiotemporal hybrid model (DRS-RF) integrating dragonfly optimization (DR) with support vector regression has reduced prediction errors to 11%[33]. Remote sensing integration has further advanced scalability. Sentinel-2 satellite imagery fused with UAV-derived RGB data has been used to develop season-specific ML models for field-scale yield estimation, achieving reasonable annual accuracy (R^2 = 0.47, RMSE = 168.7 kg/acre)[98].

For resource optimization, ML has improved irrigation efficiency. An SVM model enhanced with the Bald Eagle Search algorithm has predicted soil moisture levels with high precision (R^2 = 0.9435)[99]. RF-based methods have mapped land suitability to identify optimal cultivation zones[100]. NIRS-HSI combined with optimization algorithms have allowed pixel-level moisture monitoring, enabling targeted irrigation strategies[101].

In growth monitoring, multi-sensor UAV platforms combining LiDAR, multispectral, and RGB data with ML algorithms have demonstrated strong predictive capabilities for key phenotypic parameters. RF models have estimated leaf chlorophyll (R^2 = 0.87) and nitrogen contents (R^2 = 0.65), whereas SVM has performed well in assessing leaf area index (R^2 = 0.90) and plant height (R^2 = 0.82)[102]. The Random Forest with Parameter Optimization (RFPO) algorithm, which merges UAV-derived vegetation indices with meteorological data, has mapped spatial variations in leaf dry matter and nitrogen levels (R^2 = 0.80 and 0.79, respectively), aiding precision nutrient management[103].

Disease and pest management

ML has contributed to more effective disease and pest control in tea cultivation by improving early detection accuracy and guiding targeted treatment strategies. For disease management, CNN-based systems have classified seven common tea diseases with over 90% accuracy[104]. Pooling-based CNNs combined with weighted RF models have raised detection rates to 92.47%[105]. Advanced segmentation techniques like SLIC_SVM have improved the extraction of tea leaf disease saliency maps from complex backgrounds, achieving 95% accuracy in evaluations of precision, recall, and F1-score[106]. Researchers have also expanded focus to severity quantification: VGG16 networks using Retinex enhancement and Faster R-CNN have improved detection precision by 6% and severity grading accuracy by 9% over traditional methods[107]. For occluded leaves in field conditions, ellipse restoration paired with SVM segmentation has reached a mean F1-score of 84%, outperforming standard CNNs[108]. Architectural comparisons have shown DenseNet169 surpassing ResNet50 (95.2%), and VGG16 (90.6%) with 99% accuracy, faster convergence, and lower loss values[109].

In pest management, CNN-based models have identified 14 pest types with 97.75% accuracy, exceeding TML approaches[110]. Faster R-CNN has detected lesions caused by three diseases and four pests at 89.4% accuracy, providing real-time diagnostic support[111]. Feature selection combined with Incremental Backpropagation Learning Network (IBPLN) has achieved perfect classification accuracy (100%) for pest identification[112]. A modified Yolov7-tiny model, enhanced with deformable convolution and dynamic attention mechanisms, has optimized pest detection efficiency (93.23% accuracy) while reducing computational time and costs[113].

Classification of tea categories and cultivars

ML has strengthened the classification of tea types and cultivars to ensure product consistency and aid breeding efforts. In tea category classification, fluorescence hyperspectral technology combined with MSC-RF-RFE-SVM models has identified oolong tea varieties with 98.73% accuracy[41]. For cross-category differentiation, NIRS-HSI paired with redundant wavelet transforms and a lightweight CNN (L-CNN) have distinguished black, green, and yellow teas at 98.73% accuracy[114]. A hybrid approach merging spectral and texture features with Lib-SVM has classified five Chinese tea categories at 98.39% accuracy[115].

For cultivar identification, hyperspectral reflectance data from tea plant canopies analyzed with SVM, LDA, and ANNs have differentiated six out of nine tea varieties with 75%−80% accuracy, even under natural field conditions[116]. CNN-based methods like DenseNet201 have achieved 96.38% accuracy in cultivar recognition, rising to 97.81% with optimized training and data augmentation[117].

Tea origin authentication and fraud detection

ML integrated with spectroscopy and metabolomics has improved tea traceability and fraud detection. In origin authentication,

gas chromatography–time-of-flight mass spectrometry (GC-TOFMS) analysis of 333 Wuyi rock tea samples has detected geographic origins with over 90% accuracy using an MLP model[5]. Nuclear magnetic resonance (NMR) spectroscopy paired with RF and SVM has identified black tea origins at 92.7% and 91.8% accuracy, respectively[118], while NIRS-HSI with PCA-SVM has classified green tea origins at 97.5% accuracy[119].

For adulteration detection, fluorescence hyperspectral imaging combined with an SVM model has identified Tieguanyin tea adulterated with lower-grade Benshan tea, achieving 100% accuracy for pure samples and 94.27% accuracy in mixed batches using SNV-RF-SVM[120]. Spectrophotometry coupled with neural network modeling has detected the banned additive carmine in black tea with 100% accuracy, closely matching HPLC results ($R^2 > 0.97$)[121]. A smartphone-based micro-NIRS device with SVM has classified adulterated green tea at 97.47% accuracy, while an SVR model has quantified contaminants like sugar and glutinous rice flour[122].

Building and applying ML pipelines in the tea industry

The development of ML pipelines offers a structured approach to transforming raw data into actionable insights within the tea industry, addressing critical tasks such as yield prediction, quality assessment, and pest detection. This section provides comprehensive guidelines for developing effective ML pipelines, using tea leaf disease detection as an illustrative case study while emphasizing adaptable principles applicable to diverse tea-related applications.

Guidelines for developing ML pipelines

Automatic Machine Learning (AutoML), pioneered by organizations such as Google and prominent academic institutions, simplifies the adoption of ML by automating essential tasks including data preprocessing, feature selection, and hyperparameter optimization[123,124]. This automation benefits practitioners across experience levels by enhancing accessibility and operational efficiency. Complementary technologies such as MobileNetV2[125] and multi-scale feature extraction methodologies[126] further streamline computational requirements while preserving model accuracy.

A well-structured ML pipeline typically encompasses four fundamental stages: data preparation and feature engineering, model training and optimization, model evaluation, and deployment. Figure 3 illustrates these sequential stages from initial data collection through to operational deployment, while Table 1 summarizes the key components, their specific relevance to tea applications, and commonly employed tools and techniques to guide implementation.

### Data preparation and feature engineering

Robust data preparation constitutes the foundation for developing reliable ML models. Platforms such as H2O-AutoML automate critical preprocessing procedures, including data normalization, scaling, and outlier management, ensuring consistency across heterogeneous datasets. In tea-related applications, diverse data sources such as leaf imagery, soil quality parameters, and meteorological variables can be integrated to create comprehensive datasets for modeling complex relationships.

To enhance generalization capabilities and reduce overfitting, data augmentation strategies such as brightness modulation, geometric transformations, and controlled noise introduction can be implemented. These techniques effectively simulate the variations frequently encountered in field conditions. For instance, in tea leaf disease detection systems, such augmentation ensures consistent model performance across varying illumination conditions and orientation differences.

Feature engineering remains equally critical for model performance and interpretability. AutoML platforms frequently automate feature extraction using methods such as feature importance scoring and PCA. For tea quality assessment, characteristics including leaf coloration, texture patterns, and biochemical compositions are commonly emphasized. Advanced feature selection methods like mutual information analysis and recursive feature elimination effectively reduce data redundancy while preserving the most informative attributes[127,128].

Model training and optimization

The model training phase involves the selection and refinement of algorithms to achieve optimal performance while mitigating overfitting risks. AutoML platforms streamline this process through automated model selection and hyperparameter tuning, minimizing both manual intervention and error potential. In tea-related applications, training objectives may include detecting subtle quality variations in premium tea leaves, differentiating rare cultivars, or
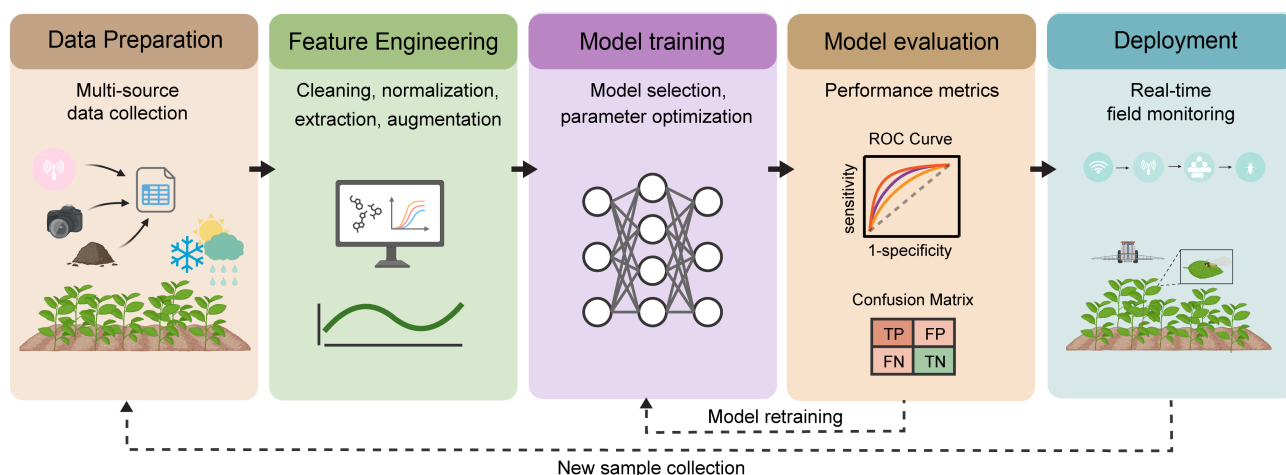
Fig. 3 Structured workflow of ML pipeline implementation for tea-specific applications. Systematic representation of the end-to-end ML pipeline encompassing sequential stages: multimodal data acquisition and preprocessing, domain-specific feature engineering, model architecture selection and hyperparameter optimization, comprehensive performance evaluation, and deployment with edge computing integration. The pipeline incorporates feedback mechanisms for continuous model refinement based on operational data and evolving environmental conditions.

Table 1. Key components of an ML pipeline for tea-related tasks.

Stage	Description	Relevance to tea tasks	Example tools/techniques
Data preparation and feature engineering	Data integration, preprocessing (e.g., scaling, normalization), augmentation, feature extraction, and selection	Integrates diverse tea-specific data: leaf images, NIR/hyperspectral scans, soil parameters, and weather metrics, and electronic tongue/nose readings. Augmentation simulates field conditions to improve model robustness.	H2O-AutoML, PCA, mutual information, recursive feature elimination, image rotation/brightness adjustment
Model training and optimization	Model selection, hyperparameter optimization, iterative refinement	Optimizes models for tea-specific tasks: disease detection, quality grading, yield prediction, and fermentation monitoring. Balances accuracy with computational efficiency for field deployment.	Bayesian optimization, evolutionary algorithms, MobileNetV2, CNNs for disease detection, RNNs for fermentation monitoring
Model evaluation	Performance metrics (accuracy, precision, recall, F1-score, R^2 and RMSE), cross-validation, error analysis	Validates model reliability across different tea varieties, regions, and seasonal conditions. Identifies specific error patterns in tea-related predictions and classifications.	K-fold cross-validation, confusion matrix analysis, performance comparison tools, learning curve visualization
Deployment	Architecture optimization, model compression, edge computing implementation, IoT integration	Enables field-level applications: real-time disease detection, harvest optimization, and processing monitoring through lightweight models on portable devices and sensor networks.	NAS, AutoML for Model Compression (AMC), MobileNetV2, edge-AI frameworks, IoT sensor integration for tea estate monitoring

achieving accelerated inference speeds for real-time monitoring systems.

To improve model generalization and reduce overfitting, practitioners should implement strategies such as regularization techniques (e.g., dropout and L2 regularization), early stopping protocols, and utilization of sufficiently diverse training datasets[129,130]. Incorporating variations across multiple tea cultivars, developmental stages, and environmental conditions can strengthen model robustness. Additionally, advanced hyperparameter optimization approaches, including Bayesian optimization and evolutionary algorithms, can further refine model performance. Bayesian optimization predicts optimal hyperparameters using probabilistic models, making it particularly suitable for complex, high-dimensional parameter spaces[131]. Evolutionary algorithms iteratively refine configurations by simulating natural selection processes, balancing exploration of diverse model structures with exploitation of promising solutions, thereby improving both performance metrics and generalization capabilities[132].

Model evaluation

Comprehensive model evaluation is essential for assessing reliability and generalization potential. AutoML platforms typically implement *k*-fold cross-validation, a methodical technique that systematically partitions the data into multiple subsets to evaluate model performance across diverse conditions. For classification tasks such as disease identification or quality categorization, key performance metrics including accuracy, precision, recall, and F1-score provide a quantitative assessment of model efficacy. Visualization tools like confusion matrices offer insights into misclassification patterns, such as differentiating healthy leaves from early-stage disease manifestations. For regression-based applications prevalent in tea research, such as yield prediction, biochemical content estimation, and continuous quality assessment, traditional evaluation metrics like coefficient of determination (R^2) and Root Mean Square Error (RMSE) are particularly valuable. R^2 quantifies the proportion of variance in the dependent variable that is predictable from independent variables, providing insights into model fit quality, while RMSE measures the average magnitude of prediction errors in the original variable's units, offering practical interpretability for agricultural stakeholders. There metrics are especially relevant when predicting continuous variables such as tea yield quantities, antioxidant levels, or precise flavor compound concentrations. Furthermore, many AutoML frameworks support comparative performance analysis across multiple models using these diverse evaluation criteria, allowing practitioners to select optimal solutions based on application-specific requirements.

Deployment-ready solutions

Effective deployment ensures that trained models operate efficiently in real-world environments. AutoML tools facilitate this transition by automating architecture selection, model compression, and hardware optimization. While certain platforms support advanced functionalities such as Neural Architecture Search (NAS) for complex optimization scenarios, most emphasize balancing model accuracy with resource efficiency to address practical constraints[133]. These platforms enable the deployment of high-performing models for field applications such as monitoring tea plantations using resource-constrained edge devices. Techniques including reinforcement learning and evolutionary algorithms embedded within AutoML frameworks assist in discovering efficient architectures while reducing computational demands. Automated pruning methodologies, such as AutoML for Model Compression (AMC), further optimize performance by reducing model complexity without significant accuracy degradation[134].

Case study: ML pipelines for tea plant disease detection

To demonstrate the practical implementation of an ML pipeline, we present a case study on tea plant disease detection[135]. A diverse image dataset of tea leaves was collected under varying environmental conditions using mobile imaging devices. Preprocessing procedures included standardized image resizing, normalization, and data augmentation (e.g., brightness adjustment and rotation) to enhance model robustness.

The dataset was split into training and validation subsets in a 7:3 ratio to ensure a balanced evaluation. Feature engineering techniques focused on extracting multi-scale texture patterns and morphological characteristics. The Coordinate Attention mechanism was implemented to emphasize disease-specific regions, while Multi-branch Parallel Convolution modules expanded the receptive field of the network, improving generalization across varying disease severities. The MobileNetV2 architecture, refined through NAS and optimized using AMC, achieved a classification accuracy of 96.12% across five common tea leaf diseases. This approach surpassed previous methods by balancing computational efficiency with model accuracy. Further improvements could include expanding the dataset to cover a broader spectrum of leaf conditions, incorporating supplementary data sources such as soil health parameters, and implementing periodic model retraining to adapt to evolving agricultural environments. Additionally, integrating explainable AI techniques would improve model interpretability for agricultural practitioners, potentially accelerating adoption in commercial tea industry settings.

Key challenges and solutions in implementing ML in the tea industry

The integration of ML in the tea industry presents significant innovation opportunities despite challenges posed by seasonal variability, quality differentiation, and geographic diversity. Addressing these challenges is crucial for maximizing ML potential in cultivation, processing, and quality assessment.

Data quality, availability, and resource constraints

ML applications in tea industry require high-quality datasets to model critical processes such as leaf growth, disease progression, fermentation stages, and quality assessment. However, fragmented cultivation practices often yield inconsistent data. Smallholder farms frequently lack infrastructure to collect comprehensive data on critical variables such as soil health, pest outbreaks, or yield trends, undermining model accuracy.

Several approaches can address these constraints. Advanced preprocessing techniques can manage missing or inconsistent data by transforming raw inputs into more reliable features. Domain-specific knowledge helps bridge gaps and generates novel features that accurately reflect seasonal variations and regional differences in tea industry systems. Additionally, cost-effective tools like smartphone imaging systems and portable NIR spectrometers offer practical approaches for localized data acquisition. For instance, handheld NIR devices have been deployed to correlate spectral patterns with flavor compounds in tea, directly linking field-based data collection to quantifiable quality outcomes[136,137].

In addition to enhancing data collection methods, lightweight ML models optimize performance in resource-constrained environments. I-MobileNetV2 designed for tea disease identification is a prime example. By incorporating Coordinate Attention mechanisms and Multi-branch Parallel Convolution algorithms, this model has achieved a 40% reduction in computational complexity while maintaining 96.12% accuracy[135]. Similarly, lightweight CNNs with knowledge distillation have shown promise for quality control during black tea fermentation[138]. These optimized models provide an effective balance between computational efficiency and predictive performance, making them well-suited for environments with limited computational resources.

Cloud platforms and open-source tools enhance data processing capabilities. A Platform-as-a-Service (PaaS) cloud system utilizing CNNs has enabled real-time tea disease prediction, allowing farmers to upload plant images for rapid analysis[139]. This capacity for real-time analysis helps address challenges posed by fluctuating weather conditions and emergent pest outbreaks, allowing for timely decision-making and more effective disease management protocols.

Moreover, strategic partnerships with technology providers can improve access to specialized tools and expertise. These collaborative arrangements enable the selection and implementation of data collection and processing systems specifically tailored to the tea sector, ensuring that smallholder operations can adopt appropriate technologies despite resource limitations. Such partnerships also provide ongoing technical support in data analysis, model deployment, and troubleshooting, thus overcoming technological barriers commonly faced by smallholder producers.

Finally, pre-trained models offer practical solutions for smallholder agricultural operations by allowing fine-tuning with smaller, region-specific datasets. These models, initially trained on large, diverse datasets, such as CNNs for disease detection or sensor data analysis frameworks, require substantially less data to adapt to local conditions. This approach reduces the need for extensive data collection, accelerates ML implementation, and eases the burden on smallholder producers.

Security and privacy concerns

Digitization of plantation data underpins ML applications but introduces risks to sensitive information like proprietary processing techniques and blending formulas. As such, robust security measures are essential to protect data integrity and maintain privacy.

To address these risks, privacy-preserving technologies including data encryption protocols and secure access controls should be implemented[140,141]. Multi-factor authentication and role-based access systems can restrict data access to authorized personnel. Furthermore, industry-specific data-sharing protocols, such as secure cloud-based platforms, can enable collaborative research without compromising proprietary information.

Human-machine collaboration and transparency

The tea industry's historical reliance on traditional knowledge and artisanal expertise can create resistance to ML adoption. Agricultural producers may distrust ML-generated predictions that contradict their experiential knowledge, while professional tasters and quality graders might perceive ML-based quality assessments as lacking the nuanced sensory discrimination essential for tea evaluation.

To foster effective human-machine collaboration, ML systems should be designed to augment rather than replace human expertise[142]. User-centric interfaces that seamlessly integrate ML-derived insights into existing operational workflows can bridge the gap between traditional practices and technological innovation. Customized training programs are equally crucial for building trust, demonstrating how ML complements rather than supplants traditional methodologies. For example, ML algorithms can analyze historical yield data in conjunction with farmers' observational records to deliver more accurate and context-aware recommendations.

Transparency in ML model development and operation is equally important for improving stakeholder acceptance. Explainable AI techniques that clearly articulate how models arrive at specific decisions can boost trust in automated systems[143]. For instance, a crop disease prediction model could provide clear visualizations of affected plant regions and explicit explanations of the factors driving its diagnosis. Clear communication of both the capabilities and limitations of ML systems in accessible language will build confidence and encourage widespread adoption across the industry.

Academic and industrial integration

The effective implementation of ML technologies hinges on bridging the persistent gap between academic research initiatives and industrial practice. Limited collaboration between academic institutions and industry stakeholders frequently results in technological solutions that inadequately address the practical challenges faced by tea producers. Furthermore, the absence of comprehensive interdisciplinary training programs has created a significant shortage of professionals possessing expertise in both ML methodologies and tea industry.

Strengthening collaborative partnerships between universities, research institutions, and commercial tea producers is therefore essential. Integrated research initiatives should specifically focus on addressing industry-relevant challenges, such as developing ML models to assess the impact of climate change on yields or predicting optimal blending ratios for consistent flavor. Additionally, interdisciplinary educational curricula that integrate ML with agricultural science principles can better prepare future professionals to tackle these challenges. For example, specialized training programs combining ML techniques with sensory evaluation methods commonly used in tea quality assessment could facilitate the development of tools that align more precisely with industry requirements.

Conclusions and future perspectives

The tea industry is undergoing a profound transformation driven by increasingly complex demands for efficiency, quality consistency, and sustainability across global production systems. ML has emerged as a pivotal tool for addressing multifaceted challenges within the sector, from optimizing cultivation and processing workflows to facilitating non-destructive quality assessment. By harnessing comprehensive data-driven insights, ML models have demonstrably contributed to smarter cultivation practices, improved quality control, and refined sensory profiling of diverse tea products. Despite these advances, substantial implementation barriers persist, including prohibitive initial investment costs, infrastructure limitations particularly in developing regions, and a critical shortage of professionals with interdisciplinary expertise. Overcoming these systematic obstacles will necessitate sustained collaborative initiatives among academic researchers, industry stakeholders, and technology developers to create contextually appropriate solutions.

Looking forward, several promising research trajectories hold significant potential for further advancing the integration of ML technologies throughout the tea value chain (Fig. 4):

(1) Advanced robotic plucking systems with sophisticated ML integration. Future research initiatives should prioritize the development of advanced ML algorithms for autonomous harvesting systems capable of highly selective plucking operations. These next-generation systems must effectively account for complex variables including leaf maturity indices, environmental heterogeneity, and variable canopy architectures while simultaneously optimizing precision in leaf selection and minimizing mechanical damage to delicate tea shoots. Particular emphasis should be placed on developing adaptive learning capabilities that can accommodate diverse cultivars and production contexts.

(2) Real-time data processing through edge computing and ML integration. The strategic convergence of edge computing technologies with lightweight ML models presents opportunities for improving decision-making processes at the plantation level. Research priorities should focus on optimizing resource-efficient ML architectures that can effectively process multimodal sensor data on-site, enabling real-time interventions across critical domains including precision irrigation management, early-stage pest and disease detection, and dynamic resource allocation based on spatiotemporal needs.

(3) Climate-adaptive ML models for resource efficiency and environmental sustainability. Developing sophisticated ML models capable of integrating diverse data streams, including high-resolution climate data, multispectral satellite imagery, and field-level measurements can support advanced predictive capabilities for weather pattern forecasting, resource optimization, and comprehensive soil
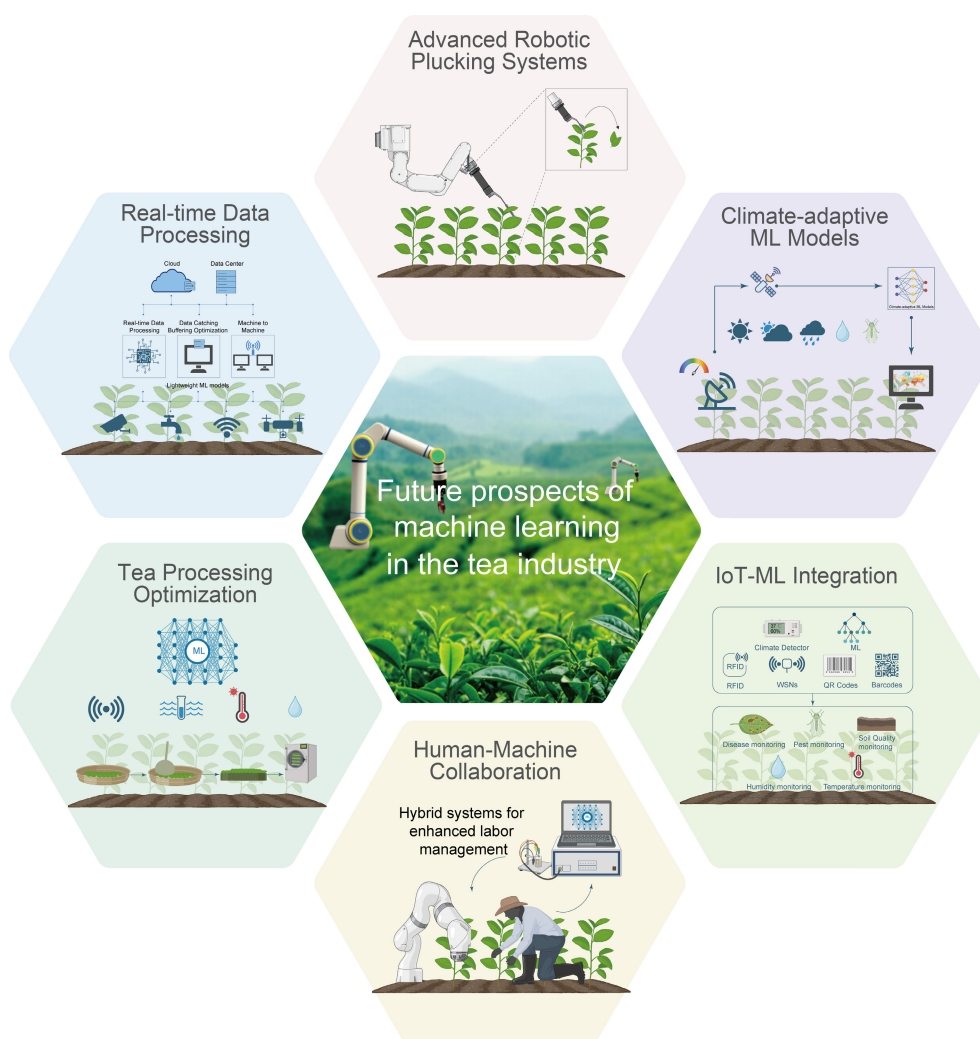
Fig. 4 Future prospects of ML applications in tea industry. Key future ML applications in tea industry include robotic plucking, real-time data processing, climate-adaptive models, processing optimization, IoT integration, and human-machine collaboration.

health management. These integrated models could substantially contribute to climate adaptation strategies, such as precision water management, targeted pest control, and optimized soil nutrient management while reducing environmental impacts through minimized carbon emissions and water consumption.

(4) Optimization of tea processing through ML-driven monitoring systems. The application of ML methodologies in tea processing operations offers considerable potential for enhancing product consistency and quality attributes. Research endeavors should concentrate on developing real-time monitoring systems that use multiparametric sensors and advanced ML algorithms to assess critical processing parameters such as leaf moisture dynamics, oxidation kinetics, fermentation progression, and precise drying conditions. Such systems would ensure consistent product quality across production batches while potentially identifying novel processing optimizations not apparent through traditional methods.

(5) IoT-ML integration for holistic estate management. With the accelerating adoption of IoT technologies on tea estates, establishing standardized protocols for integrating diverse IoT devices with specialized ML models becomes increasingly essential. Such integrated systems can enable continuous monitoring of environmental conditions, soil health, and operational factors, thereby optimizing agricultural practices and promoting evidence-based sustainable estate management decisions aligned with certification requirements.

(6) Human-machine collaboration frameworks for labor optimization. As persistent labor shortages continue to impact the tea industry, particularly in traditional production regions, ML can assist in managing human resources more effectively. Research should explore hybrid systems that synergistically combine ML capabilities with human expertise, offering real-time guidance, predictive analytical insights, and task optimization recommendations, particularly for manual tasks such as selective harvesting and critical processing stages requiring sensory evaluation.

In conclusion, ML technologies present substantial opportunities for enhancing productivity, product quality, and sustainability within the tea industry. Targeted interdisciplinary research focusing on autonomous harvesting systems, real-time analytical capabilities, climate adaptation strategies, processing optimization methodologies, comprehensive IoT integration frameworks, and human-machine collaborative systems is essential for unlocking the full transformative potential of ML applications in this sector. Sustained collaborative partnerships among academia, industry, and technology developers will remain critical to ensuring successful implementation, continuous improvement, and widespread adoption of these innovations across diverse production settings.

Author contributions

The authors confirm contributions to the paper as follows: conceptualization and writing: Gao F, Wang S, Yu X; figure and table modification: Gao F, Wang S, Yu X; review and editing: Gao F, Wang S, Yu X. All authors reviewed the results and approved the final version of the manuscript.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Acknowledgments

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

References

1. Yu X, Xiao J, Chen S, Yu Y, Ma J, et al. 2020. Metabolite signatures of diverse *Camellia sinensis* tea populations. *Nature Communications* 11:5586

2. Yang G, Meng Q, Shi J, Zhou M, Zhu Y, et al. 2023. Special tea products featuring functional components: health benefits and processing strategies. *Comprehensive Reviews in Food Science and Food Safety* 22:1686−721

3. Zhao F, Chen M, Jin S, Wang S, Yue W, et al. 2022. Macro-composition quantification combined with metabolomics analysis uncovered key dynamic chemical changes of aging white tea. *Food Chemistry* 366:130593

4. Gao T, Shao S, Hou B, Hong Y, Ren W, et al. 2023. Characteristic volatile components and transcriptional regulation of seven major tea cultivars (*Camellia sinensis*) in China. *Beverage Plant Research* 3:17

5. Peng Y, Zheng C, Guo S, Gao F, Wang X, et al. 2023. Metabolomics integrated with machine learning to discriminate the geographic origin of Rougui Wuyi rock tea. *NPJ Science of Food* 7:7

6. Wang L, Yao L, Hao X, Li N, Wang Y, et al. 2019. Transcriptional and physiological analyses reveal the association of ROS metabolism with cold tolerance in tea plant. *Environmental and Experimental Botany* 160:45−58

7. Shen J, Wang Y, Chen C, Ding Z, Hu J, et al. 2015. Metabolite profiling of tea (*Camellia sinensis* L.) leaves in winter. *Scientia Horticulturae* 192:1−9

8. Chen S, Shen J, Fan K, Qian W, Gu H, et al. 2022. Hyperspectral machine-learning model for screening tea germplasm resources with drought tolerance. *Frontiers in Plant Science* 13:1048442

9. Liu J, Zhang C, Hu R, Zhu X, Cai J. 2019. Aging of agricultural labor force and technical efficiency in tea production: evidence from Meitan County, China. *Sustainability* 11:6246

10. Xie S, Feng H, Yang F, Zhao Z, Hu X, et al. 2019. Does dual reduction in chemical fertilizer and pesticides improve nutrient loss and tea yield and quality? A pilot study in a green tea garden in Shaoxing, Zhejiang Province, China. *Environmental Science and Pollution Research* 26:2464−76

11. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. 2018. Machine learning in agriculture: a review. *Sensors* 18:2674

12. Kasinathan T, Singaraju D, Uyyala SR. 2021. Insect classification and detection in field crops using modern machine learning techniques. *Information Processing in Agriculture* 8:446−57

13. Sujatha R, Chatterjee JM, Jhanjhi N, Brohi SN. 2021. Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocessors and Microsystems* 80:103615

14. van Klompenburg T, Kassahun A, Catal C. 2020. Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture* 177:105709

15. Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, et al. 2021. Machine learning in agriculture: a comprehensive updated review. *Sensors* 21:3758

16. Hamrani A, Akbarzadeh A, Madramootoo CA. 2020. Machine learning for predicting greenhouse gas emissions from agricultural soils. *Science of The Total Environment* 741:140338

17. Wang H, Gu J, Wang M. 2023. A review on the application of computer vision and machine learning in the tea industry. *Frontiers in Sustainable Food Systems* 7:1172543

18. Xu Q, Zhou Y, Wu L. 2024. Advancing tea detection with artificial intelligence: strategies, progress, and future prospects. *Trends in Food Science & Technology* 153:104731

19. Wei Y, Wen Y, Huang X, Ma P, Wang L, et al. 2024. The dawn of intelligent technologies in tea industry. *Trends in Food Science & Technology* 144:104337

20. Wang P, Fan E, Wang P. 2021. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters* 141:61−67

21. Wang J, Ma Y, Zhang L, Gao RX, Wu D. 2018. Deep learning for smart manufacturing: methods and applications. *Journal of Manufacturing Systems* 48:144−56

22. Khan S, Sajjad M, Hussain T, Ullah A, Imran AS. 2020. A review on traditional machine learning and deep learning models for WBCs classification in blood smear images. *IEEE Access* 9:10657−73

23. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. 2020. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology* 9:14

24. Wu D, Yang H, Chen X, He Y, Li X. 2008. Application of image texture for the sorting of tea categories using multi-spectral imaging technique and support vector machine. *Journal of Food Engineering* 88:474−83

25. Wang S, Yang X, Zhang Y, Phillips P, Yang J, et al. 2015. Identification of green, oolong and black teas in China via wavelet packet entropy and fuzzy support vector machine. *Entropy* 17:6663−82

26. Liu L, Hu P, Yang F, Song M. 2020. Application of terahertz time-domain spectroscopy combined with support vector machine to determine tea and pesticide samples. *Materials Express* 10:1646−53

27. Ahmad H, Sun J, Nirere A, Shaheen N, Zhou X, et al. 2021. Classification of tea varieties based on fluorescence hyperspectral image technology and ABC-SVM algorithm. *Journal of Food Processing and Preservation* 45:e15241

28. Hossain S, Mou RM, Hasan MM, Chakraborty S, Razzak MA. 2018. Recognition and detection of tea leaf's diseases using support vector machine. *Proc. of 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, 2018*. pp. 150−54. US: IEEE. doi: 10.1109/CSPA.2018.8368703

29. Prabu S, Bapu BRT, Sridhar S, Nagaraju V. 2022. Tea plant leaf disease identification using hybrid filter and support vector machine classifier technique. In *Recent Advances in Internet of Things and Machine Learning*, eds Balas VE, Solanki VK, Kumar R. Cham: Springer. Volume 215. pp. 117−28 doi: 10.1007/978-3-030-90119-6_10

30. Ren G, Zhang X, Wu R, Yin L, Hu W, et al. 2023. Rapid characterization of black tea taste quality using miniature NIR spectroscopy and electronic tongue sensors. *Biosensors* 13:92

31. Amsaraj R, Mutturi S. 2023. Rapid detection of sunset yellow adulteration in tea powder with variable selection coupled to machine learning tools using spectral data. *Journal of Food Science and Technology* 60:1530−40

32. Liang L, Wang J, Deng F, Kong D. 2023. Mapping Pu'er tea plantations from GF-1 images using Object-Oriented Image Analysis (OOIA) and Support Vector Machine (SVM). *PLoS One* 18:e0263969

33. Jui SJJ, Masrur Ahmed AA, Bose A, Raj N, Sharma E, et al. 2022. Spatiotemporal hybrid random forest model for tea yield prediction using satellite-derived variables. *Remote Sensing* 14:805

34. Dao DH, Tang NC, Pham BT. Monitoring and evaluating the fermentation level of black tea using the random forest model. In *Advances in Engineering Research and Application*, eds Nguyen DC, Vu NP, Long BT, Puta H, Sattler KU. Cham: Springer. pp. 739−53 doi: 10.1007/978-3-030-92574-1_76

35. Deng X, Liu Z, Zhan Y, Ni K, Zhang Y, et al. 2020. Predictive geographical authentication of green tea with protected designation of origin using a random forest model. *Food Control* 107:106807

36. Han Y, He Y, Liang Z, Shi G, Zhu X, et al. 2023. Risk assessment and application of tea frost hazard in Hangzhou City based on the random forest algorithm. *Agriculture* 13:327

37. Diniz PHGD, Pistonesi MF, Alvarez MB, Band BSF, de Araújo MCU. 2015. Simplified tea classification based on a reduced chemical composition profile *via* successive projections algorithm linear discriminant analysis (SPA-LDA). *Journal of Food Composition and Analysis* 39:103−10

38. Mohammadi N, Esteki M, Simal-Gandara J. 2024. Machine learning for authentication of black tea from narrow-geographic origins: combination of PCA and PLS with LDA and SVM classifiers. *LWT* 203:116401

39. Lin J, Zhang P, Pan Z, Xu H, Luo Y, et al. 2013. Discrimination of oolong tea (*Camellia sinensis*) varieties based on feature extraction and selection from aromatic profiles analysed by HS-SPME/GC-MS. *Food Chemistry* 141:259−65

40. Gan N, Sun M, Lu C, Li M, Wang Y, et al. 2022. High-speed identification system for fresh tea leaves based on phenotypic characteristics utilizing an improved genetic algorithm. *Journal of the Science of Food and Agriculture* 102:6858−67

41. Hu Y, Xu L, Huang P, Luo X, Wang P, et al. 2021. Reliable identification of oolong tea species: nondestructive testing classification based on fluorescence hyperspectral technology and machine learning. *Agriculture* 11:1106

42. Wu X, Yang J, Wang S. 2018. Tea category identification based on optimal wavelet entropy and weighted k-Nearest Neighbors algorithm. *Multimedia Tools and Applications* 77:3745−59

43. Wijaya DR, Handayani R, Fahrudin T, Kusuma GP, Afianti F. 2024. Electronic nose and optimized machine learning algorithms for noninfused aroma-based quality identification of gambung green tea. *IEEE Sensors Journal* 24:1880−93

44. Xu M, Wang J, Zhu L. 2021. Tea quality evaluation by applying E-nose combined with chemometrics methods. *Journal of Food Science and Technology* 58:1549−61

45. Hu Y, Huang P, Wang Y, Sun J, Wu Y, et al. 2023. Determination of Tibetan tea quality by hyperspectral imaging technology and multivariate analysis. *Journal of Food Composition and Analysis* 117:105136

46. Shao P, Wu M, Wang X, Zhou J, Liu S. 2018. Research on the tea bud recognition based on improved k-means algorithm. *MATEC Web of Conferences* 232:03050

47. Shevchuk A, Jayasinghe L, Kuhnert N. 2018. Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of LC-MS data. *Food Research International* 109:387−402

48. Zhu X, Goldberg AB. 2009. Overview of semi-supervised learning. In *Introduction to Semi-Supervised Learning*. Cham: Springer. pp. 9−19 doi: 10.1007/978-3-031-01548-9_2

49. Weng Y, Zhang Y, Wang W, Dening T. 2024. Semi-supervised information fusion for medical image analysis: recent progress and future perspectives. *Information Fusion* 106:102263

50. Kondratovich E, Baskin II, Varnek A. 2013. Transductive support vector machines: promising approach to model small and unbalanced datasets. *Molecular Informatics* 32:261−66

51. Yang J, Chen Y. 2022. Tender leaf identification for early-spring green tea based on semi-supervised learning and image processing. *Agronomy* 12:1958

52. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518:529−33

53. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, et al. 2025. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv* 00:2501.12948

54. Lin G, Xiong J, Zhao R, Li X, Hu H, et al. 2023. Efficient detection and picking sequence planning of tea buds in a high-density canopy. *Computers and Electronics in Agriculture* 213:108213

55. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436−44

56. Gayathri S, Wise DCJW, Shamini PB, Muthukumaran N. 2020. Image analysis and detection of tea leaf disease using deep learning. *Proc of 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020*. US: IEEE. pp. 398−403. doi: 10.1109/ICESC48915.2020.9155850

57. Deepa C, Sanjay SM, Mosses V, Yugeshwaran M, Sam Jebaraj A. 2024. *Camellia sinensis* (tea) plant disease classification using RESNET. *Proc of 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024*. US: IEEE. pp. 1−4 doi: 10.1109/ICSTEM61137.2024.10560945

58. Bai B, Wang J, Li J, Yu L, Wen J, et al. 2024. T-YOLO: a lightweight and efficient detection model for nutrient buds in complex tea plantation environments. *Journal of the Science of Food and Agriculture* 104:5698−711

59. Shi M, Zheng D, Wu T, Zhang W, Fu R, et al. 2024. Small object detection algorithm incorporating swin transformer for tea buds. *PLoS One* 19:e0299902

60. Xue Z, Xu R, Bai D, Lin H. 2023. YOLO-tea: a tea disease detection model improved by YOLOv5. *Forests* 14:415

61. Ye R, Shao G, Yang Z, Sun Y, Gao Q, et al. 2024. Detection model of tea disease severity under low light intensity based on YOLOv8 and EnlightenGAN. *Plants* 13:1377

62. Li H, Shi H, Du A, Mao Y, Fan K, et al. 2022. Symptom recognition of disease and insect damage based on Mask R-CNN, wavelet transform, and F-RNet. *Frontiers in Plant Science* 13:922797

63. Masoud KM, Persello C, Tolpekin VA. 2020. Delineation of agricultural field boundaries from sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks. *Remote Sensing* 12:59

64. Chen YT, Chen SF. 2020. Localizing plucking points of tea leaves using deep convolutional neural networks. *Computers and Electronics in Agriculture* 171:105298

65. Lin YK, Chen SF, Kuo YF, Liu TL, Lee SY. 2021. Developing a guiding and growth status monitoring system for riding-type tea plucking machine using fully convolutional networks. *Computers and Electronics in Agriculture* 191:106540

66. Zhu N, Liu X, Liu Z, Hu K, Wang Y, et al. 2018. Deep learning for smart agriculture: concepts, tools, applications, and opportunities. *International Journal of Agricultural and Biological Engineering* 11:32−44

67. Yao Z, Zhu X, Zeng Y, Qiu X. 2023. Extracting tea plantations from multitemporal Sentinel-2 images based on deep learning networks. *Agriculture* 13:10

68. Mao Y, Li H, Xu Y, Wang S, Yin X, et al. 2024. Early detection of gray blight in tea leaves and rapid screening of resistance varieties by hyperspectral imaging technology. *Journal of the Science of Food and Agriculture* 104:9336−48

69. Li H, Mao Y, Wang Y, Fan K, Shi H, et al. 2022. Environmental simulation model for rapid prediction of tea seedling growth. *Agronomy* 12:3165

70. Huang Y, Jiang H, Wang W. 2022. Research on tea tree growth monitoring model using soil information. *Plants* 11:262

71. Chen X, Hassan MM, Yu J, Zhu A, Han Z, et al. 2024. Time series prediction of insect pests in tea gardens. *Journal of the Science of Food and Agriculture* 104:5614−24

72. Krishnan Jayapal S, Poruran S. 2023. Enhanced disease identification model for tea plant using deep learning. *Intelligent Automation & Soft Computing* 35:1261−75

73. Zhang J, Guo H, Guo J, Zhang J. 2023. An information entropy masked vision transformer (IEM-ViT) model for recognition of tea diseases. *Agronomy* 13:1156

74. Zilvan V, Ramdan A, Heryana A, Krisnandi D, Suryawati E, et al. 2022. Convolutional variational autoencoder-based feature learning for automatic tea clone recognition. *Journal of King Saud University - Computer and Information Sciences* 34:3332−42

75. Cimpoiu C, Cristea VM, Hosu A, Sandru M, Seserman L. 2011. Antioxidant activity prediction and classification of some teas using artificial neural networks. *Food Chemistry* 127:1323−28

76. Kalathingal MSH, Basak S, Mitra J. 2020. Artificial neural network modeling and genetic algorithm optimization of process parameters in fluidized bed drying of green tea leaves. *Journal of Food Process Engineering* 43:e13128

77. Chen Q, Zhao J, Vittayapadung S. 2008. Identification of the green tea grade level using electronic tongue and pattern recognition. *Food Research International* 41:500−04

78. Li X, He Y. 2008. Discriminating varieties of tea plant based on Vis/NIR spectral characteristics and using artificial neural networks. *Biosystems Engineering* 99:313−21

79. Mienye ID, Sun Y. 2022. A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access* 10:99129−49

80. Geng J, Li H, Luan W, Shi Y, Pang J, et al. 2023. Estimation of daily actual evapotranspiration of tea plantations using ensemble machine learning algorithms and six available scenarios of meteorological data. *Applied Sciences* 13:12961

81. Raza A, Hu Y, Lu Y. 2024. Improving carbon flux estimation in tea plantation ecosystems: a machine learning ensemble approach. *European Journal of Agronomy* 160:127297

82. Li J, Li Q, Luo W, Zeng L, Luo L. 2024. Rapid color quality evaluation of needle-shaped green tea using computer vision system and machine learning models. *Foods* 13:2516

83. Zou Y, Ma W, Tang Q, Xu W, Tan L, et al. 2020. A high-precision method evaluating color quality of Sichuan Dark Tea based on colorimeter combined with multi-layer perceptron. *Journal of Food Process Engineering* 43:e13444

84. Liu H, Yu D, Gu Y. 2019. Classification and evaluation of quality grades of organic green teas using an electronic nose based on machine learning algorithms. *IEEE Access* 7:172965−73

85. Xu M, Wang J, Zhu L. 2019. The qualitative and quantitative assessment of tea quality based on E-nose, E-tongue and E-eye combined with chemometrics. *Food Chemistry* 289:482−89

86. Liang J, Guo J, Xia H, Ma C, Qiao X. 2025. A black tea quality testing method for scale production using CV and NIRS with TCN for spectral feature extraction. *Food Chemistry* 464:141567

87. Ren G, Wang Y, Ning J, Zhang Z. 2021. Evaluation of Dianhong black tea quality using near-infrared hyperspectral imaging technology. *Journal of the Science of Food and Agriculture* 101:2135−42

88. Li L, Xie S, Ning J, Chen Q, Zhang Z. 2019. Evaluating green tea quality based on multisensor data fusion combining hyperspectral imaging and olfactory visualization systems. *Journal of the Science of Food and Agriculture* 99:1787−94

89. Yang B, Qi L, Wang M, Hussain S, Wang H, et al. 2020. Cross-category tea polyphenols evaluation model based on feature fusion of electronic nose and hyperspectral imagery. *Sensors* 20:50

90. Yang H, Chen L, Chen M, Ma Z, Deng F, et al. 2019. Tender tea shoots recognition and positioning for picking robot using improved YOLO-V3 model. *IEEE Access* 7:180998−1011

91. Jayanthy S, Sathyendraa VM, Sumedh KP, Suresh S. 2023. Tea leaf disease classification and tea bud identification. *Proc. 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022*. US: IEEE. pp. 1−5 doi: 10.1109/ICERECT56837.2022.10059683

92. Wang G, Wang Z, Zhao Y, Zhang Y. 2022. Tea bud recognition based on machine learning. *Proc. 2022 41st Chinese Control Conference (CCC), Hefei, China, 2022*. US: IEEE. pp. 6533−37 doi: 10.23919/CCC55666.2022.9902610

93. Chen C, Lu J, Zhou M, Yi J, Liao M, et al. 2022. A YOLOv3-based computer vision system for identification of tea buds and the picking point. *Computers and Electronics in Agriculture* 198:107116

94. Wang T, Zhang K, Zhang W, Wang R, Wan S, et al. 2023. Tea picking point detection and location based on Mask-RCNN. *Information Processing in Agriculture* 10:267−75

95. Hassoun A, Aït-Kaddour A, Abu-Mahfouz AM, Rathod NB, Bader F, et al. 2023. The fourth industrial revolution in the food industry—Part I: Industry 4.0 technologies. *Critical Reviews in Food Science and Nutrition* 63:6547−63

96. Batool D, Shahbaz M, Shahzad Asif H, Shaukat K, Alam TM, et al. 2022. A hybrid approach to tea crop yield prediction using simulation models and machine learning. *Plants* 11:1925

97. Xu H, Ma W, Tan Y, Liu X, Zheng Y, et al. 2022. Yield estimation method for tea based on YOLOv5 deep learning. *Journal of China Agricultural University* 27:213−20

98. Liu H, Liu Y, Xu W, Wu M, Wang L, et al. 2025. A seasonal fresh tea yield estimation method with machine learning algorithms at field scale integrating UAV RGB and Sentinel-2 imagery. *Plants* 14:373

99. Huang Y. 2023. Improved SVM-based soil-moisture-content prediction model for tea plantation. *Plants* 12:2309

100. Xing W, Zhou C, Li J, Wang H, He J, et al. 2022. Suitability evaluation of tea cultivation using machine learning technique at town and village scales. *Agronomy* 12:2010

101. Sun J, Zhou X, Hu Y, Wu X, Zhang X, et al. 2019. Visualizing distribution of moisture content in tea leaves using optimization algorithms and NIR hyperspectral imaging. *Computers and Electronics in Agriculture* 160:153−59

102. Li H, Wang Y, Fan K, Mao Y, Shen Y, et al. 2022. Evaluation of important phenotypic parameters of tea plantations using multi-source remote sensing data. *Frontiers in Plant Science* 13:898962

103. Jiang J, Ji H, Zhou G, Pan R, Zhao L, et al. 2025. Non-destructive monitoring of tea plant growth through UAV spectral imagery and meteorological data using machine learning and parameter optimization algorithms. *Computers and Electronics in Agriculture* 229:109795

104. Chen J, Liu Q, Gao L. 2019. Visual tea leaf disease recognition using a convolutional neural network model. *Symmetry* 11:343

105. Heng Q, Yu S, Zhang Y. 2024. A new AI-based approach for automatic identification of tea leaf disease using deep neural network based on hybrid pooling. *Heliyon* 10:e26465

106. Sun Y, Jiang Z, Zhang L, Dong W, Rao Y. 2019. SLIC_SVM based leaf diseases saliency map extraction of tea plant. *Computers and Electronics in Agriculture* 157:102−9

107. Hu G, Wang H, Zhang Y, Wan M. 2021. Detection and severity analysis of tea leaf blight based on deep learning. *Computers & Electrical Engineering* 90:107023

108. Hu G, Wei K, Zhang Y, Bao W, Liang D. 2021. Estimation of tea leaf blight severity in natural scene images. *Precision Agriculture* 22:1239−62

109. Deng X, Photong C. 2024. Evaluation of tea leaf disease identification based on convolutional neural networks VGG16, ResNet50, and DenseNet169 image recognitions. *2024 12th International Electrical Engineering Congress, Pattaya, Thailand, 2024*. US: IEEE. pp. 1−4 doi: 10.1109/iEECON60677.2024.10537865

110. Chen J, Liu Q, Gao L. 2021. Deep convolutional neural networks for tea tree pest recognition and diagnosis. *Symmetry* 13:2140

111. Lee SH, Lin SR, Chen SF. 2020. Identification of tea foliar diseases and pest damage under practical field conditions using a convolutional neural network. *Plant Pathology* 69:1731−39

112. Samanta RK, Ghosh I. 2012. Tea insect pests classification based on artificial neural networks. *International Journal of Computer Engineering Science* 2:1−13

113. Yang Z, Feng H, Ruan Y, Weng X. 2023. Tea tree pest detection algorithm based on improved Yolov7-tiny. *Agriculture* 13:1031

114. Cui Q, Yang B, Liu B, Li Y, Ning J. 2022. Tea category identification using wavelet signal reconstruction of hyperspectral imagery and machine learning. *Agriculture* 12:1085

115. Ning J, Sun J, Li S, Sheng M, Zhang Z. 2017. Classification of five Chinese tea categories with different fermentation degrees using visible and near-infrared hyperspectral imaging. *International Journal of Food Properties* 20:1515−22

116. Nidamanuri RR. 2020. Hyperspectral discrimination of tea plant varieties using machine learning, and spectral matching methods. *Remote Sensing Applications: Society and Environment* 19:100350

117. Zhang Z, Yang M, Pan Q, Jin X, Wang G, et al. 2025. Identification of tea plant cultivars based on canopy images using deep learning methods. *Scientia Horticulturae* 339:113908

118. Cui C, Xu Y, Jin G, Zong J, Peng C, et al. 2023. Machine learning applications for identify the geographical origin, variety and processing of black tea using ^{1}H NMR chemical fingerprinting. *Food Control* 148:109686

119. Liu Y, Huang J, Li M, Chen Y, Cui Q, et al. 2022. Rapid identification of the green tea geographical origin and processing month based on near-infrared hyperspectral imaging combined with chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 267:120537

120. Hu Y, Kang Z. 2022. The rapid non-destructive detection of adulteration and its degree of Tieguanyin by fluorescence hyperspectral technology. *Molecules* 27:1196

121. Wei L, Yang Y, Sun D. 2020. Rapid detection of carmine in black tea with spectrophotometry coupled predictive modelling. *Food Chemistry* 329:127177

122. Li L, Jin S, Wang Y, Liu Y, Shen S, et al. 2021. Potential of smartphone-coupled micro NIR spectroscopy for quality control of green tea. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 247:119096

123. Hutter F, Kotthoff L, Vanschoren J. 2019. *Automated machine learning: methods, systems, challenges*. Cham: Springer. doi: 10.1007/978-3-030-05318-5

124. Raschka S, Patterson J, Nolet C. 2020. Machine learning in Python: main developments and technology trends in data science, machine learning, and artificial intelligence. *Information* 11:193

125. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. 2018. MobileNetV2: inverted residuals and linear bottlenecks. *Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018*. pp. 4510−20 doi: 10.1109/CVPR.2018.00474

126. Chen H, He G, Peng X, Wang G, Yin R. 2024. A multi-scale feature fusion deep learning network for the extraction of cropland based on landsat data. *Remote Sensing* 16:4071

127. Cang S, Yu H. 2012. Mutual information based input feature selection for classification problems. *Decision Support Systems* 54:691−98

128. Jeon H, Oh S. 2020. Hybrid-recursive feature elimination for efficient feature selection. *Applied Sciences* 10:3211

129. Agliari E, Alemanno F, Aquaro M, Fachechi A. 2024. Regularization, early-stopping and dreaming: a hopfield-like setup to address generalization and overfitting. *Neural Networks* 177:106389

130. Ying X. 2019. An overview of overfitting and its solutions. *Journal of Physics: Conference Series* 1168:022022

131. Wu J, Chen XY, Zhang H, Xiong LD, Lei H, et al. 2019. Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology* 17:26−40

132. Tani L, Rand D, Veelken C, Kadastik M. 2021. Evolutionary algorithms for hyperparameter optimization in machine learning for application in high energy physics. *The European Physical Journal C* 81:170

133. Wang X, Zhu W. 2024. Advances in neural architecture search. *National Science Review* 11:nwae282

134. He Y, Lin J, Liu Z, Wang H, Li LJ, et al. AMC: AutoML for model compression and acceleration on mobile devices. In *Computer Vision – ECCV 2018*, eds Ferrari V, Hebert M, Sminchisescu C, Weiss Y. Cham: Springer. Vol. 11211. pp. 815−32. doi: 10.1007/978-3-030-01234-2_48

135. Li Z, Li Y, Yan C, Yan P, Li X, et al. 2024. Enhancing tea leaf disease identification with lightweight MobileNetV2. *Computers, Materials & Continua* 80:679−94

136. Wang J, Zareef M, He P, Sun H, Chen Q, et al. 2019. Evaluation of matcha tea quality index using portable NIR spectroscopy coupled with chemometric algorithms. *Journal of the Science of Food and Agriculture* 99:5019−27

137. Sun Y, Wang Y, Huang J, Ren G, Ning J, et al. 2020. Quality assessment of instant green tea using portable NIR spectrometer. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 240:118576

138. Ding Z, Yang C, Hu B, Guo M, Li J, et al. 2024. Lightweight CNN combined with knowledge distillation for the accurate determination of black tea fermentation degree. *Food Research International* 194:114929

139. Lanjewar MG, Panchbhai KG. 2023. Convolutional neural network based tea leaf disease prediction system on smart phone using paas cloud. *Neural Computing and Applications* 35:2755−71

140. Zhang G, Chen X, Feng B, Guo X, Hao X, et al. 2022. BCST-APTS: blockchain and CP-ABE empowered data supervision, sharing, and privacy protection scheme for secure and trusted agricultural product traceability system. *Security and Communication Networks* 2022:2958963

141. Liu Z, Guo J, Yang W, Fan J, Lam KY, et al. 2022. Privacy-preserving aggregation in federated learning: a survey. In *IEEE Transactions on Big Data*. US: IEEE. pp. 1−20 doi: 10.1109/TBDATA.2022.3190835

142. Deng C, Ji X, Rainey C, Zhang J, Lu W. 2020. Integrating machine learning with human knowledge. *iScience* 23:101656

143. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, et al. 2020. Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020*. New York, NY, USA: Association for Computing Machinery. pp. 648−57 doi: 10.1145/3351095.3375624