ENHANCING BRIX VALUE PREDICTION IN STRAWBERRIES USING MACHINE LEARNING: A FUSION OF PHYSIOCHEMICAL AND COLOR-BASED FEATURES FOR IMPROVED SWEETNESS ASSESSMENT

Ameetha Junaina T. K [1], R. Kumudham[1*], Ebenezer Abishek B.[2], and Mohamed Shakir[3]

[*1]The Department of Electronics and Communication Engineering, Vels Institute of Science, Technology, and Advanced Studies
Chennai, 6000 117, India

[2]The Department of Electronics and Communication Engineering
VelTech MultiTech Dr. Rangarajan and Dr. Sakunthala Engineering College
Chennai, 600062, Tamil Nadu, India

[3]White-Box Analytics
Sydney, NSW 2000, Australia

Corresponding author: kumudham.se@velsuniv.ac.in*

ABSTRACT

This study contributes to the ongoing wave of artificial intelligence integration by applying machine learning techniques to automate the assessment of strawberry quality. This research focuses on determining if the sweetness of strawberries can be predicted using a combination of physiochemical variables, their interaction parameters, and color-based features extracted from image data. This research used a 150-sample collection of strawberry images and physiochemical characteristics such as salinity, specific gravity, pH, and Brix. Normalized raw and derived feature variables and selected dataset transformations were done. We then split the dataset into mutual exclusivity training and test sets. Exponential Gaussian Process Regression (GPR) suited well due to low validation errors. This best model predicted Brix values for the remaining test samples. The Mean Absolute Percentage Error(MAPE) showed 98.783% forecast accuracy (Acc). We also examined the model's coefficient of determination (R^2) values, which were 0.78 and 0.9739 for training and testing, respectively. The Mean Square Error (MSE) and Mean Absolute Error (MAE) obtained after training were 0.32994 and 0.0453, and testing was 0.35286 and 0.0663. Using input characteristics with high Acc and low error rates, deep learning models like Recurrent Neural Network (RNN) and its derivatives were constructed. Using physiochemical and visual data, machine learning and deep learning models successfully predict strawberry sweetness. This prediction accuracy shows the complex link between internal components and Brix readings, enabling high-quality strawberry production.

Keywords: Brix Value Prediction; Feature Engineering; Gaussian Process Regression; Machine Learning; Physio-chemical Parameters; Predictive Model; Strawberry Image Analysis; Sweetness Assessment.

1.0 INTRODUCTION

The modern agricultural landscape is undergoing a transformation driven by the escalating demand for organic, fresh, and nutritious food. However, this paradigm shift is met with challenges in the traditional farming workforce due to urbanization, leading to a shortage of skilled laborers. To meet evolving consumer demands and ensure the production of high-quality fresh produce, the agricultural sector is compelled to embrace technological advancements and implement smart farming practices. Computer vision and machine learning methodologies are utilized in the automation procedures of agriculture, encompassing both pre- and post-harvesting stages. As an example, an overview on the utilization of image processing and machine learning techniques in automated grading systems for fresh produce is presented in [1]. The term ‘smart farming’, characterized by automation and the integration of artificial intelligence, has made remarkable progress across various agricultural domains, including harvesting, sowing, weeding, etc. Automated robots [2], [3], [4], guided by artificial intelligence, have become integral to contemporary agriculture, mitigating cultivation losses and providing valuable insights to farmers. Despite the evident advantages of automation in agriculture, certain nations, such as India, have been slow in adopting cutting-edge agricultural technologies due to factors like limited awareness, high costs, or a lack of understanding of the benefits they offer. Smart farming technologies hold the potential to address challenges like labor shortages and evolving consumer preferences, ultimately resulting in increased productivity, higher yields, reduced labor costs, and expedited delivery of fresh produce to consumers [5-14]. Fruit quality-based sorting and

grading is a pivotal aspect of agricultural processing, significantly influencing preservation, transportation, and product valuation [15-21]. In this study, we turn our attention to strawberries, celebrated for their sweet taste, vibrant red color, fragrant aroma, and juicy texture. These berries are a rich source of essential nutrients, including vitamins C and K, fiber, folic acid, manganese, and potassium. By automating the prediction of sweetness level in strawberries, our goal is to ensure the consistent availability of high-quality, highly nutritious strawberries in the market, thereby meeting consumer expectations and fostering satisfaction.

In the agricultural sector, predicting the sweetness level, or Brix values, which quantify the sugar content in strawberries and other fruits, holds paramount importance [22], [23], [24]. Growers and distributors depend on accurate Brix value predictions to make informed decisions regarding harvesting, sorting, and pricing, enabling them to assess strawberry quality and sweetness precisely. Recent advancements in computer vision and machine learning techniques offer new possibilities for forecasting Brix values by analyzing physiochemical properties, interaction parameters, and visual cues from strawberry images.

Despite progress in Brix value prediction, there exists a research gap in comprehensively integrating a wide range of features that collectively account for the impact of physiochemical properties, interaction parameters, and visual characteristics of strawberries. Previous studies have often focused on limited feature sets, neglecting potential interconnections and synergistic effects among various traits. Furthermore, visual characteristics, such as color, have been underexplored despite their significance in predicting Brix readings.

This research aims to bridge this gap by developing an innovative regression analysis leveraging machine learning techniques. Our analysis will incorporate physiochemical properties, their interaction parameters, and the visual features of strawberries to accurately forecast Brix values. By utilizing a comprehensive feature set that includes established physiochemical properties, novel interaction parameters capturing intricate relationships, and visual features extracted from high-resolution images, we seek to significantly enhance prediction Acc. Moreover, this approach will shed light on the factors influencing strawberry sweetness, potentially paving the way for optimizing strawberry-growing conditions.

The implications of this study extend to the food industry, particularly the export market for strawberry products such as juice. Growers and producers can harness machine learning techniques to predict strawberry sweetness, thereby improving the consistency and quality of their output, enhancing profitability, and ensuring long-term sustainability. In the following section, we review existing literature that utilizes machine learning techniques to predict Brix values for various fruits, highlighting the relevance and significance of our research.

The paper is organized into the following sections: Section 1 provides an overview and examines previous research in the field. Section 2 provides a detailed account of the materials and methods employed in the study. It encompasses a comprehensive description of the datasets utilized, including their acquisition methods, as well as an elucidation of the pre-processing procedures undertaken. Additionally, the mechanism for predicting Brix is also outlined in the study. Section 3 of the document encompasses the presentation of the results and subsequent discussion. In the fourth section, we present our concluding remarks on the findings.

2.0 RELATED WORKS

Numerous research endeavors have been dedicated to the prediction of sugar levels in various fruits, leveraging the power of machine learning techniques. In this section, we present two notable works from the literature that exemplify the integration of machine learning and sensory data to forecast Brix values in fruits.

In 2021, Gomes *et al*. [25] undertook a study focused on predicting sugar content in vintage port wine grapes using a combination of machine learning and hyperspectral imaging (HSI) techniques. Vintage port wine grape berries exhibit distinct enological features crucial for assessing their maturity, making them an ideal candidate for predictive analysis. The study introduced a framework for developing on-the-fly, non-invasive sensing technologies, offering a promising avenue for precision viticulture. The research employed four different machine learning techniques: ridge regression, partial least squares, neural networks, and convolutional neural networks (CNNs). These techniques were applied to predict sugar content in grapes, a key determinant of grape maturity. The study encompassed an evaluation of the models' generalization capabilities across various vintages and grape types as test samples, which were not included in the training dataset. The results revealed robust performance across all approaches, with prediction errors falling within acceptable margins. Particularly noteworthy was the superior performance of the 1D CNN model, demonstrating its ability to more accurately and efficiently estimate sugar content compared to the other three methods. Furthermore, the study underscored the impact of terroir and grape variety variations on predictive models, highlighting the need for comprehensive considerations in grape ripening stage prediction. This research showcased the potential of combining hyperspectral imaging technology

with machine learning approaches as a viable tool for non-destructive and non-invasive assessment of grape quality. The study's findings hinted at the possibility of accurately gauging the sugar levels in maturing wine grapes, thereby contributing to enhanced wine production and quality standards.

In 2018, Sangsongfa *et al*. [26] embarked on an innovative project aimed at predicting the sweetness of pineapples using deep learning algorithms and pineapple images. The authors sought to determine pineapple sweetness without the need for invasive procedures, relying solely on external images of the fruit. They considered visual cues such as shade variations, fruit size, skin surface characteristics, and fruit color luster. Interestingly, the study challenged the conventional belief that fruit hue directly correlates with sweetness, emphasizing the multifaceted nature of sweetness determination in pineapples. To achieve accurate predictions, the researchers integrated multiple variables and leveraged machine learning methodologies for training and validation. The results indicated that sweetness could be predicted with a satisfactory degree of accuracy, as evidenced by metrics such as Acc and Root Mean Squared Error (RMSE) obtained from experimental data. The study introduced an innovative approach for evaluating pineapple sweetness, utilizing smartphone photography on the Android platform. This approach offered simplicity and speed in assessing pineapple sweetness without damaging the fruit. The methodology demonstrated an Acc rate of 80.15%, a RMSE of 0.0156, and a reliability rate of 95%. However, it is worth noting that the study primarily focused on predominantly green pineapples, with only a minor percentage of orange ones. Predictions for fully orange pineapples yielded results with a significant error margin. Despite this limitation, the study provided valuable insights and laid the foundation for potential future innovations, including the development of predictive tools for pineapples affected by mealybug wilt and Android-based applications.

These two studies exemplify the capacity of machine learning and image-based techniques to revolutionize fruit quality assessment, offering non-invasive, efficient, and accurate methods for predicting Brix values and enhancing the overall quality of fresh produce.

3.0 MATERIALS AND METHODS

In the pursuit of enhancing Brix value prediction in strawberries, our methodology leverages machine learning regression models, a cornerstone in the field of supervised learning for continuous output prediction using input features. Constructing a machine learning-based regression model involves a structured sequence of steps [27]. Initially, the dataset is assembled, comprising both independent variables and the target variable to be predicted. Subsequently, this data undergoes a rigorous cleaning and pre-processing phase to address issues such as missing values and outliers while standardizing it to a format suitable for modeling. The pre-processed data is then divided into three subsets: training, validation, and testing. The training set is employed to train the model, while the validation set serves to evaluate its performance and fine-tune its parameters if necessary. Following the training phase, the model undergoes evaluation using the validation set, with adjustments made to its parameters to optimize performance. Finally, the trained model is evaluated with the designated testing set, and if it demonstrates satisfactory performance, it is deployed for predictions on novel data in real-world scenarios. Continuous monitoring and potential model refinements are essential for ensuring sustained effectiveness.

3.1 Dataset Used

Our study encompasses the acquisition of two distinct datasets: a strawberry image dataset and a dataset containing instrumental or physiochemical features obtained from strawberry juice extracts. This dataset was meticulously compiled by procuring strawberries exclusively from local fresh markets during their seasonal availability, resulting in a dataset comprising 150 samples of 'Mahabaleshwar Strawberry' collected between October 2020 and February 2021. Our analysis specifically targets image features related to the color and physiochemical properties of strawberries, aiming to explore their potential influence on the overall quality and sensory attributes of the fruit. The central focus of our investigation is to determine whether these properties hold the potential to impact the sweetness of strawberries, a parameter quantified using the Brix value. To assess the physiochemical properties, we consider salinity, specific gravity, and pH values as predictor features.

3.2 Acquisition of Image Dataset and Processing

For the acquisition of strawberry images, we established a controlled studio setup, employing an HD camera. The images used for our investigation were obtained through the use of a Logitech C920 HD Pro camera. The camera provides a maximum resolution of 1080p at a frame rate of 30 frames per second (fps), enabling Full HD resolution. The device is equipped with a sensor that has a resolution of 3 megapixels, enabling the capture of images with

high levels of detail. Furthermore, the camera has a diagonal field of view (dFoV) measuring 78 degrees, which allows for a broad and inclusive viewpoint when capturing images. The aforementioned parameters played a pivotal role in our research endeavors since they were important in acquiring visual data of exceptional quality. Fig 1 illustrates the home studio setup designed for capturing images of the strawberry samples, also featuring the refractometers used for collecting instrumental parameters [28].

Fig. 1: Image acquisition: home studio setup [28]

To enhance the dataset's quality, we executed several pre-processing steps, including segmentation and image augmentation. After meticulously capturing images of 150 strawberry samples, we applied a segmentation technique to extract and eliminate the background from these strawberry images, as demonstrated in Fig 2.

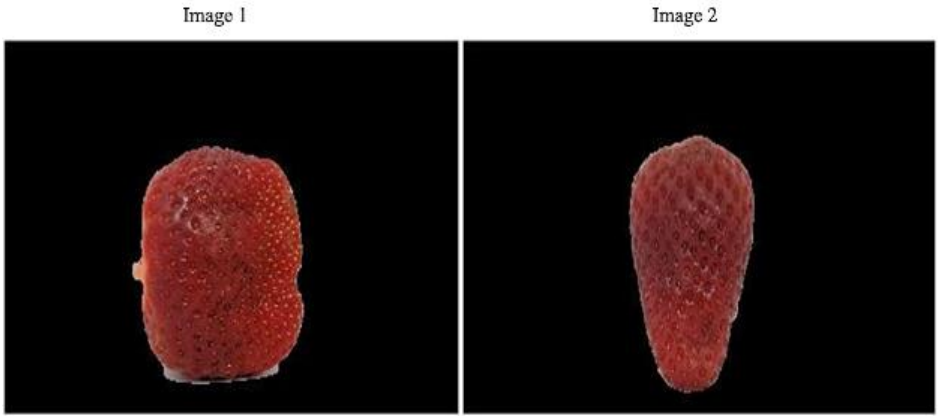

Fig. 2: Segmented and background removed strawberries

Subsequently, we employed an image augmentation technique to expand the dataset, resulting in a total of 300 images. Data augmentation plays a crucial role in mitigating the limitations associated with insufficient data availability and enhancing the model's ability to generalize to real-world variations [29]. In our research, we employed the imageDataAugmenter function in MATLAB to apply diverse transformations, such as random rotations, translations along the x and y axes, and other augmentation methods, generating augmented images that represent various perspectives and orientations of strawberries. This approach not only bolsters dataset diversity but also improves the model's robustness and precision in generating forecasts based on the expanded dataset.

The subsequent phase involved the creation of an image datastore, encompassing a cumulative quantity of 300 images, comprising 150 original images and 150 augmented ones. From this dataset, we extracted histogram-based color features (CF). These features provide valuable insights into an image's color

composition and intrinsic attributes, effectively representing the distribution of color values and offering quantitative assessments of attributes like color predominance, variety, disparity, equilibrium, and similarity or dissimilarity. Histogram-based color features are extensively utilized in diverse image analysis applications, encompassing object recognition, image retrieval, and color-based image classification [30]. We extracted a total of 256 color features from the dataset of 300 images, resulting in the creation of a 300x256 matrix representing image features. A glimpse of a few of these color features extracted from five image samples is presented in Table 1. These color features, when combined with the physiochemical features, form the basis for our Brix prediction analysis.

Table 1: Histogram-based color image features from strawberry samples

ID	CF 1	CF 2	CF3	CF 4	CF5	CF6	CF7
1	0.4451	0.03408	0.03635	0.04271	0.0395	0.0428	0.0458
2	0.6540	0.00095	0.00217	0.00195	0.0024	0.0027	0.0032
3	0.7285	0.59869	0.78272	0.97324	0.9884	0.9881	0.9884
4	0.7462	0.76494	0.83852	0.85162	0.6981	0.6895	0.6232
5	0.8640	0.33589	0.37982	0.41362	0.4006	0.3937	0.4067

3.3 Acquisition of Instrumental Dataset and Processing

In parallel with the image dataset, we conducted instrumental assessments of the strawberries to acquire physiochemical parameters critical for our analysis. These parameters include Brix (°Bx), salinity, pH values, and specific gravity. To determine sweetness, saltiness, acidity, and density levels, strawberries were sliced using a sharp knife, and handheld refractometers, namely the Brix Refractometer, Salinity Refractometer, and Specific Gravity Refractometer, were utilized for measurement.

The Calibration of the refractometers with distilled water ensured precise measurements. The data acquisition process also involved recording atmospheric temperature and pressure, contributing to comprehensive data collection.

The BRIX Measuring Refractometer [31] used in the experimental setup of our research is shown in Fig. 3.

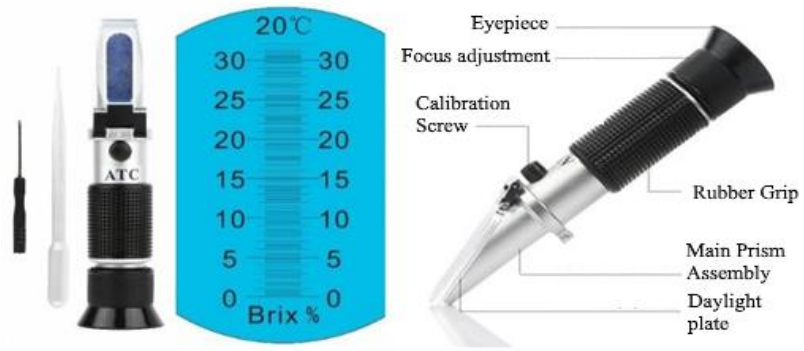

Fig. 3: BRIX measuring refractometer [31]

Additionally, we employed a pH measuring kit to assess pH levels, which is shown in Fig. 4.

Fig. 4: pH measuring device

The pH of a strawberry fruit can be influenced by the composition and concentration of organic acids present, including citric acid, malic acid, and tartaric acid. Typically, the pH levels of strawberry fruit fall within the range of 3.0 to 3.5, reflecting its acidic nature. Factors such as soil pH, water quality, and temperature during fruit development can affect strawberry fruit pH. Salinity levels in strawberries can be influenced by the concentration of salts in both the soil and irrigation water. Specific gravity is influenced by the fruit's density and composition, including water content, sugar content, and other solids. The specific gravity of strawberry fruit typically ranges from 0.94 to 1.03, with higher values indicating a higher concentration of solids. Factors affecting specific gravity include plant nutrition, water availability, and temperature during fruit development.

A meticulous data collection process was conducted on a sample size of 150 strawberries, resulting in a dataset with 150*4 instrumental readings (150 samples, 4 parameters each). These parameters were measured using the respective refractometers and a pH meter, and the recorded instrumental values were tabulated and saved. The instrumental readings obtained for a few strawberry samples are provided in Table 2. The acquired physiochemical parameters were increased in size to 300 by duplicating and then aligned in the same order as those of the samples to match the entire number of image samples, including the augmented ones.

Table 2: Instrumental readings of a few strawberry samples

Strawberry ID	BRIX	Salinity	Specific Gravity	pH
1	5	40	1.03	3
2	5	50	1.04	3
3	7.5	55	1.04	4
4	7	60	1.045	3
5	8.9	77	1.058	5
6	3.9	30	1.025	3
7	6.2	55	1.04	4
8	5	45	1.035	2
9	5.6	45	1.035	2
10	7	61	1.046	3
11	6	50	1.04	2
12	6	42	1.032	3

Overall, the physiochemical parameters of a strawberry fruit are influenced by an array of factors, encompassing genetics, environmental conditions, and post-harvest handling practices. Effective management of these factors is instrumental in optimizing strawberry fruit quality and consistency. The following section delves into a detailed exploration of the considered physiochemical parameters, including Brix, salinity, specific gravity, and pH, and their acquisition, tabulation, and analysis, particularly in relation to the target variable, Brix.

The Brix scale (°Bx) has long been employed in various industries, including wine, sugar, fruit, and honey, to quantify sucrose content, often referred to as sugar content. The Brix unit is used to measure the refractive index of solutions, with the refractive index of pure water conventionally designated as the baseline (assigned a value of "0"). This baseline provides a reference point for determining the optimal ripeness stage for fruit harvesting. Brix values are commonly measured using a Brix (°Bx) refractometer, and the relationship between the Brix scale and sucrose concentration is defined as 1°Bx = 1% Brix, making it a unit for quantifying the mass percentage of a liquid. Sweetness in fruit is assessed by measuring the percentage of sugar expressed in degrees Brix (°Brix) through soluble solids in the juice.

The study involved various physiochemical measurements, including Brix, salinity, specific gravity, and pH. Salinity was measured using a salinity refractometer with a scale ranging from 0-100 PSU (practical salinity unit), while specific gravity ranged from 1.000 to 1.070, representing the fruit's density compared to water. All instruments were initially calibrated using distilled water, with temperature and pressure also recorded.

The pH values of the strawberries were determined to assess their acidity. Fruits with pH values less than 7 were considered acidic, while those above 7 were alkaline. The acidity increases with decreasing pH values. This measurement was conducted by pressing firmly cut strawberry pieces against pH paper strips, and the color change of the paper strip indicated the pH value. This information was tabulated for each strawberry.

After gathering the instrumental dataset parameters from 150 strawberry samples, scatter plots were created to explore the relationships between the predictor variables (salinity, specific gravity, and pH) and the response variable (Brix). These scatter plots provided valuable insights into the associations between these parameters. While salinity and specific gravity exhibited linear relationships with Brix, pH levels showed distinct vertical lines at values of 2, 3, and 4, suggesting only a notable correlation between pH levels and the perceived sweetness of strawberries.

3.4 Proposed Method

The acquired dataset parameters from 150 strawberry samples were saved in tabular form and imported into Matlab for further processing. To prepare the data for machine learning, several preprocessing steps were applied. These included data normalization to ensure that all features were standardized to a common scale, preventing any single feature from having undue influence. Feature engineering techniques were employed to create interaction parameters from the physiochemical measurements [32]. These interactions captured nonlinear correlations and relationships among the variables, enhancing the model's predictive capabilities. To identify the most influential features for predicting Brix values, a correlation-based feature selection (CFS) method was utilized. This method computes the correlation between each feature and the Brix values. The top 25 features with the highest correlation scores were selected for further analysis. These features included a combination of physiochemical measurements, derived interaction parameters, and histogram based-color features.

The research process commenced with the implementation of various regression models for predicting Brix values in strawberries, aiming to identify the model that offers superior accuracy and reliability. Comparative analysis of these models unveiled the Exponential GPR model as the optimal choice, boasting a minimal RMSE of 0.57441 and a coefficient of determination (R-squared) value of 0.78 during training. This signifies the model's exceptional ability to capture and predict the sweetness of strawberries. Table 3 serves as a valuable reference for assessing the performance attributes of the different regression models, including training time, RMSE, MSE, R-squared, and MAE. The selection of the Exponential GPR model was based on its outstanding performance metrics, further highlighting its potential for practical applications.

Table 3. Comparison of various trained regression models' statistics using our dataset

Sl. No.	Regression Model	Training Time (sec)	RMSE	MSE	R-Squared	MAE
1	Linear Regression	23.483	0.93444	0.87317	0.42	0.62271
2	Linear SVM	24.5859	0.86687	0.75146	0.50	0.61913
3	Boosted Trees Ensemble	9.2808	0.73195	0.53575	0.64	0.56385

4	Fine Tree	15.6883	0.66667	0.44445	0.71	0.50208
5	Gaussian Process (Exponential)	11.559	0.57441	0.32994	0.78	0.35286

To gain deeper insights into the model's performance, we utilized visualizations such as response plots as illustrated in Fig. 5 and actual vs. predicted Brix plots as shown in Fig. 6. These visual aids provide an intuitive understanding of the model's behavior and how well it aligns with the ideal regression model. The Exponential GPR model consistently demonstrated its capability to predict Brix values closely in line with actual measurements.

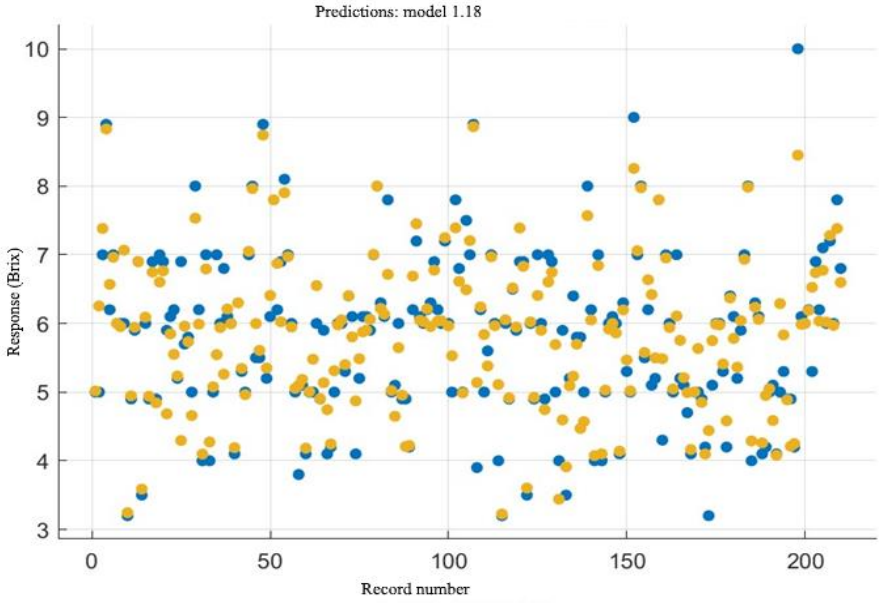

Fig.5: Response plot of training samples of the Exponential GPR model

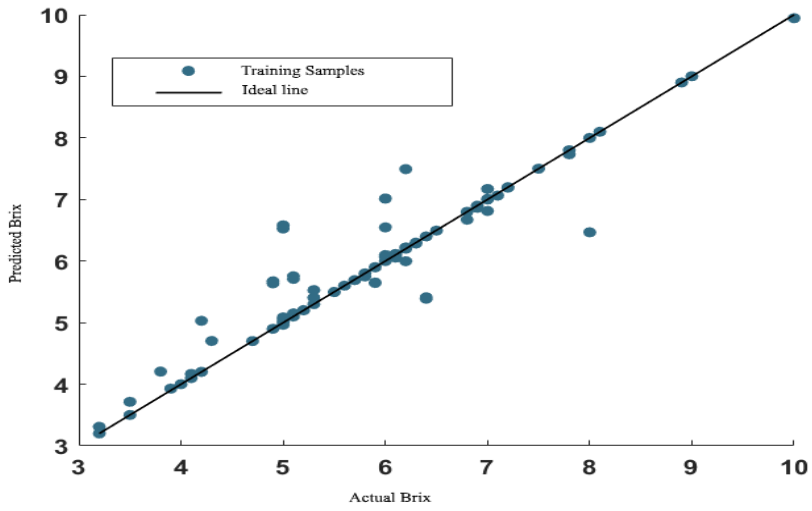

Fig.6: Actual vs predicted brix plot of training samples of the Exponential GPR model

Once the Exponential GPR model was identified as the best performer, it was exported for deployment in practical scenarios. This model, leveraging an Exponential kernel function and a constant basis function, embodies GPR, a non-parametric, Bayesian approach with a unique capability to evaluate prediction uncertainty and excel even with limited datasets.

The research then shifted its focus to the rigorous testing and validation of the Exponential GPR model. For this, the remaining mutually exclusive dataset, which was reserved for testing purposes, was utilized. Below is the code used for testing the model in Matlab.

```
YPred = trainedModel.predictFcn(Test_Data);
```

The prediction performance of the model was then evaluated, which is discussed below in detail.

4.0 RESULTS AND DISCUSSION

The remaining 30% of the dataset, designated as the 'test dataset,' served as the basis for evaluating the model's predictive prowess. The 'predict' function in Matlab facilitated the prediction of Brix values for the test dataset, and the results were meticulously recorded in a 'RESULTS' Table 4 as given below.

Table 4: RESULTS table: Actual vs. predicted brix values of a few samples

Strawberry ID	Actual BRIX	Predicted BRIX
'3.png'	7.50000000000000	7.49713350055414
'5.png'	8.90000000000000	8.89728907768484
'10.png'	7	7.17207262164017
'11.png'	6	5.99990997796704
'15.png'	7	6.99996529664081
'18.png'	6	6.00001601506904
'23.png'	6.90000000000000	6.91743320434474
'27.png'	4.90000000000000	4.90006676491666
'31.png'	4.90000000000000	4.89982462541431
'32.png'	7	6.99954288901365
'33.png'	6.90000000000000	6.82601379657835
'35.png'	4	4.44720437722777
'46.png'	5.70000000000000	5.69990758026706
'50.png'	8	7.99969762932555
'51.png	6.20000000000000	6.19974947527340

The assessment of the model's efficacy was carried out by computing various statistical parameters, including MAE, MSE, RMSE, and MAPE. These metrics served as quantitative indicators of the model's predictive accuracy and reliability. The calculated MAPE value of 1.2169 and the impressive Acc rate of 98.7831% underscored the model's remarkable ability to predict Brix values with precision. These findings confirm that the model's predictions closely align with the actual Brix values, instilling confidence in its applicability for assessing strawberry sweetness.

The research did not stop at model selection and validation. It delved into feature engineering, demonstrating how the incorporation of a modified dataset, derived from normalized values of the acquired dataset, significantly enhanced the model's predictive capabilities. This process added relevance and precision to the model's predictions, a testament to the importance of feature engineering in machine learning applications.

The analysis extended to data visualization, which played a pivotal role in understanding the model's behavior. Scatter plots (Fig. 7), line plots (Fig. 8), and histograms of residuals (Fig. 9) were employed to gain insights into patterns, trends, and the model's overall performance. These visualizations provided a tangible means of assessing model assumptions and behavior.

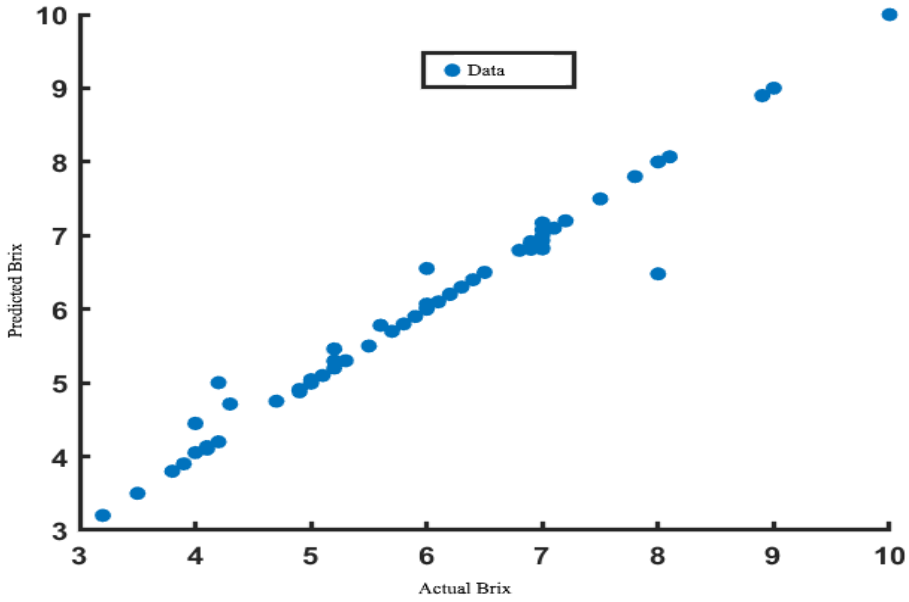

Fig.7: Scatter plot of actual vs predicted brix output

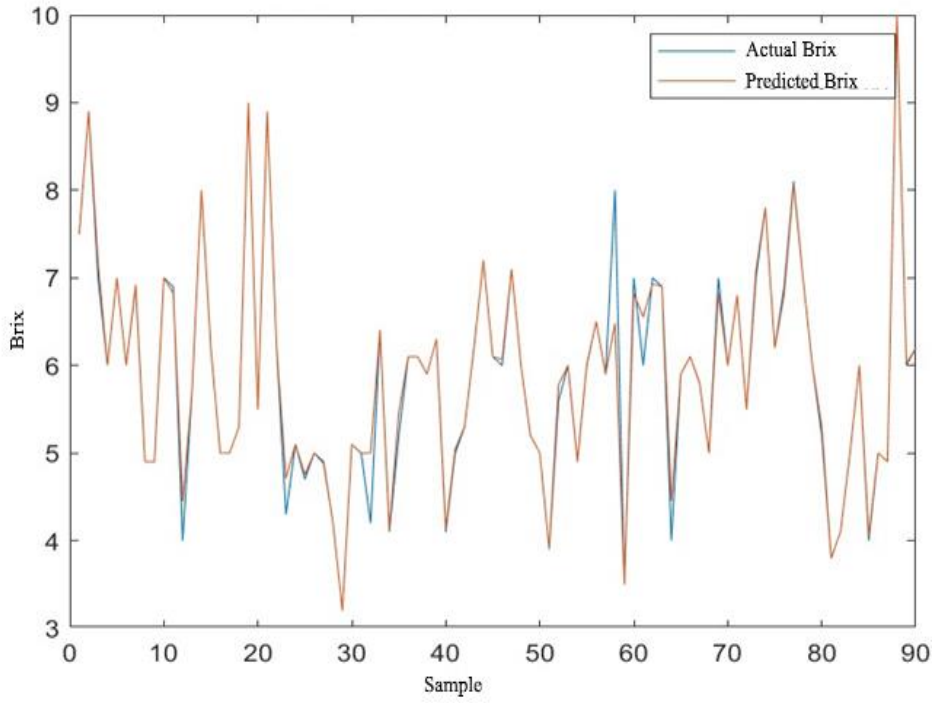

Fig.8: Line plot of actual vs predicted brix output

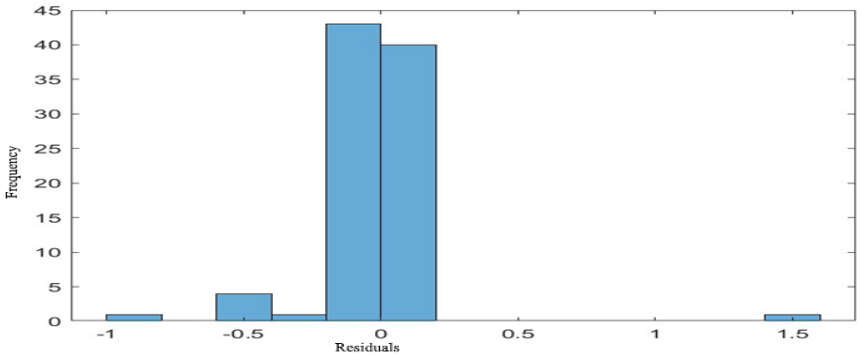

Fig.9: Histogram of residuals plot

Finally, the research culminated in a comparative analysis, pitting the proposed Brix prediction system against existing algorithms in the literature. Two key studies, one on grape sugar content prediction [25] and the other on strawberry quality prediction [33], served as benchmarks for comparison. The results showcased the superior performance of the proposed model, indicating higher R-squared values and a lower RMSE, further validating its effectiveness in assessing internal quality parameters.

The performance of the proposed method that predicts the Brix values of strawberries using the modified dataset generated from the normalized values of the acquired dataset can be compared with existing algorithms in the literature that predict the total soluble solution in strawberry juice. Brix levels and total soluble solids are internal quality metrics that provide vital insight into strawberries' sweetness and general quality. While Brix readings show the sugar concentration of fruit juices, total soluble solids include all dissolved solids, including sugars, organic acids, and other substances. Since these parameters share a fundamental connection in assessing the internal quality of strawberries, both algorithms can be compared for better performance.

As referred to earlier, the study mentioned in [25] aimed to develop and compare four distinct machine learning techniques, namely ridge regression, partial least squares, neural networks, and convolutional neural networks, for the purpose of predicting the sugar content in grapes. This parameter is widely recognized as a key indicator of grape maturity. This study primarily examined the generalization capacity of various vintages and varieties that were not included in the training model. The prediction errors obtained by each method fell within acceptable ranges, indicating that the overall performance achieved was highly satisfactory in terms of robustness. Moreover, the findings of this study demonstrate that the suggested one-dimensional convolutional neural network (1D CNN) structure can effectively be utilized for the estimation of sugar content in wine grape berries. The 1D CNN outperforms the alternative approaches of ridge regression, partial least squares, and neural networks in terms of performance.

In a work developed by Jayanta Kumar Basak *et al*. [33] in 2022, the determination of internal quality parameters like soluble solids (TSS in 0Brix) and pH in strawberry cultivation was performed. Their primary goal was to create a nondestructive method using a machine learning algorithm to predict TSS and pH in strawberries. A hundred samples from different ripening stages were taken randomly for dataset creation using biometrical features such as length, diameters, weight, TSS, and pH values. Using an image processing approach, an image of each strawberry fruit was collected for color feature extraction. Multiple linear regression (MLR) and support vector machine regression (SVM-R) models were developed using RGB, HSV, and HSL channels as input variables. Two statistical metrics, RMSE and R2, were used to assess the performance of both models. The study found that the SVM-R model using HSV color space performed marginally better than the MLR model for TSS and pH prediction. The HSV-based SVM-R model, which is considered the best model for their dataset, could explain a maximum of 84.1% and 79.2% of the variances in the measured and predicted TSS (0 Brix) data, respectively.

The present study compares the performance measures, specifically the coefficient of determination (R2p) and RMSE, attained by the most effective methods outlined in the references [25] and [33] for predicting the Brix values of Port Wine grapes and the Total Soluble Solid Content of Strawberries (internal quality parameter), respectively. These measures are then contrasted with the Brix value prediction algorithm proposed in this study, as depicted in Fig. 10. As it is known, R-squared (coefficient of determination) is a measure of a regression model's goodness of fit. A higher R-squared value indicates a better fit of the model to the data. Similarly, RMSE is a measure of the deviation of predicted values from the true values, with a lower RMSE indicating better performance. The below graph plot shows that our proposed regression technique outperformed other models in

terms of a higher R^2 value and a lower RMSE for internal quality prediction of strawberry extract during the prediction phase.

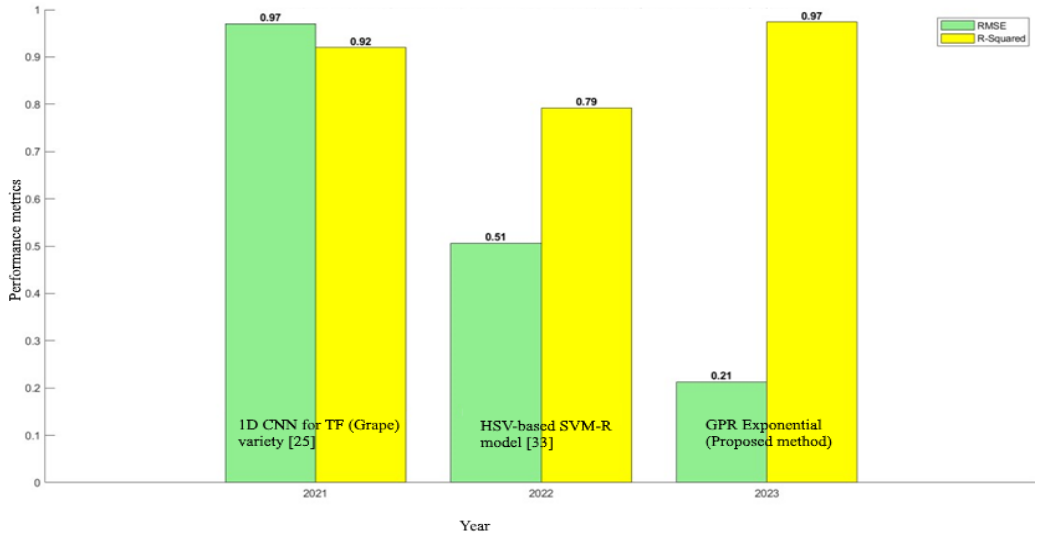

Fig.10: Comparison of performance of the proposed brix prediction system with other prediction systems in literature [25], and, [33]

This research successfully leveraged a fusion of physiochemical and color-based features, coupled with advanced machine learning techniques, to enhance the prediction of Brix values in strawberries. The Exponential GPR model emerged as the optimal choice, offering precision, reliability, and practical applicability.

Additionally, Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), Gated Recurrent Unit (GRU), Bidirectional Gated Recurrent Unit (BiGRU) etc. were implemented using these input features and the results obtained are displayed in Table 5 below [34]. The training and testing data were created by extracting features from the fusion dataset, which consisted of histogram-based color picture characteristics, physiochemical data, and interaction parameters.

Table 5: Performance metrics table of the test samples using variants of RNNs

MODEL	MSE	RMSE	MAE	R-SQUARED	MAPE	Acc
RNN	0.5543	0.744	0.6214	0.6953	0.1168	99.88%
LSTM	0.7284	0.853	0.6443	0.5997	0.1172	99.88%
BiLSTM	0.6995	0.836	0.6343	0.6155	0.1119	99.89%
GRU	0.7038	0.838	0.6026	0.6132	0.1017	99.90%
BiGRU	0.8762	0.936	0.7081	0.5184	0.1197	99.88%

RNNs are used to do regression tasks, particularly predicting the Brix values based on input features. A customized RNN model was developed for the purpose of regression analysis. The model's particular objective was to accurately forecast Brix values by using the collection of input characteristics. The process started with data preparation, whereby input features and associated target variables were retrieved from both the training and testing datasets. The RNN architecture was built by combining one input layer, two fully linked hidden layers, and a regression output layer. Significantly, the second concealed layer integrated recurrent connections. The training process was streamlined using the Adam optimization method. This involves providing parameters such as the

maximum number of epochs and the mini-batch size. After completing training, the model was deployed to make predictions of Brix values for the testing dataset. The predictions were then assessed to evaluate the model's performance and efficacy in capturing the underlying connections between the input characteristics and target variables whose performance metrics values are given in above table.

An LSTM model was used to forecast Brix levels using the input features. The LSTM architecture was designed with layers that are specially customized for processing sequence data. The input layer was initialized with the size of the input features, followed by an LSTM layer with a specific number of hidden units, which was set up to produce sequences as output. Afterwards, a fully connected layer was used to provide regression output, which is then followed by a regression layer to calculate the loss. The layer configuration is structured in a layer graph, which acts as the plan for the LSTM model. The training process utilizes the Adam optimization method, including specific parameters such as maximum epochs and mini-batch size that are specified in the training options. The model is trained using the given training data. After being trained, the LSTM model is used to forecast Brix values for the testing dataset. Subsequently, the predictions are subjected to further analysis and review to assess the model's ability in making accurate predictions as displayed in Table 5.

A BiLSTM model was used to forecast Brix values using the input features. The design has a sequential input layer, which is then followed by a BiLSTM layer with a predetermined number of hidden units. The output layers consist of a fully connected layer that produces regression output and a regression layer that calculates the loss. The training process utilizes Adam optimization with predefined parameters, while the progress is tracked using validation data. After being trained, the model uses the testing dataset to make predictions of Brix values. The performance of the model is then assessed using several metrics including MAPE, Acc, R-squared, MSE, RMSE, and MAE. These metrics provide valuable information about the model's Acc and its ability to accurately represent the variability in the Brix values. These performance measures are displayed in Table 5.

The implementation of the GRU model for forecasting Brix values included setting the input size according to the dimensions of the training data. Additionally, a sequence input layer was provided to handle the input characteristics. The GRU layer was configured with 100 hidden units to effectively capture intricate patterns within the data. A singular output class was constructed for the regression job. The training process used the Adam optimization method, using a maximum of 100 epochs and a mini-batch size of 32. To mitigate the issue of bursting gradients during training, a gradient threshold of 1 was implemented. Additionally, the initial learning rate was adjusted to 0.01. At each epoch, data shuffling was carried out to guarantee resilient training. Validation data was used to monitor the performance of the model, and the progress of the training was represented using plots. Following the completion of training, the model produced predictions on the test dataset. To evaluate the model's performance, many metrics were computed, including MAPE, Acc, R-squared, MSE, RMSE, and MAE as given in Table 5.

The Brix prediction model was constructed using the BiGRU architecture, which adhered to the following specifications: The input and output sequences were arranged, and the BiGRU architecture was established utilizing a sequence input layer, followed by forward and backward GRU layers with 100 hidden units each. A depth concatenation layer merges the outputs of the forward and backward GRU layers, which are then sent through a fully connected layer to provide regression output. The training parameters were configured using Adam optimization, with a maximum of 100 epochs, a mini-batch size of 32, and an initial learning rate of 0.01. The model underwent training, and then generated predictions on the test data. The model's performance was assessed by computing evaluation measures as displayed in Table 5.

Upon evaluating the results of the ML model and the DL models on the strawberry dataset, it is clear that both approaches demonstrated exceptional performance with high accuracy and low error rates. The GPR model has given strong performance, with an Acc of 98.783% after testing. These results indicate that the model has excellent predictive potential. Additionally, the DL models exhibited impressive accuracy, with the RNN obtaining a remarkable Acc rate of 99.88%. Similarly, RNN variants like LSTM, BiLSTM, GRU, and BiGRU displayed similarly high performance.

5.0 CONCLUSION AND FUTURE WORK

Our study underscores the significant advancements that automation and machine learning have brought to the agricultural sector. We have successfully developed a robust automation model that predicts Brix values in strawberries with an impressive testing Acc of 98.7% using a machine learning model and Acc above 99% using deep learning models with lower error rates. The GPR model gave a very satisfactory performance, achieving an

Acc of 98.783% after testing. These findings demonstrate a high level of predictive capability. In addition, the deep learning models gave exceptional accuracy, with the RNN achieving an outstanding Acc rate of 99.88%. Similarly, other variants such as LSTM, BiLSTM, GRU, and BiGRU have also shown notably superior performance in terms of accuracy and other performance metrics. Leveraging a unique dataset encompassing various predictors and advanced feature selection techniques, our model demonstrates exceptional precision in assessing the sweetness and maturity of strawberries, a vital factor for growers and stakeholders in the strawberry industry. This research not only contributes to the field of agricultural research and quality control but also provides valuable tools for enhancing product quality and minimizing wastage.

Looking ahead, there are exciting prospects for future work. Expanding the feature set to include environmental and weather data could offer a more comprehensive view of fruit quality. Developing dynamic models that adapt to changing conditions in real-time could revolutionize farming practices. Additionally, extending this methodology to other crops and commodities could have a broad-reaching impact on agriculture. As we continue to refine and expand upon these findings, we have the potential to further optimize farming processes, improve product quality, and contribute to a more sustainable and efficient agricultural sector.

ACKNOWLEDGEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Mrs. T K Ameetha Junaina *et al.*, "A Survey on Fresh Produce Grading Algorithms Using Machine Learning and Image Processing Techniques", in *IOP Conference Series Materials Science and Engineering, December 2020, 981, 042084.* https://doi.org/10.1088/1757-899X/981/4/042084.

[2] F. Qingchun, Z. Wengang, Q. Quan, J. Kai and G. Rui, "Study on strawberry robotic harvesting system.", *2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE),2012, Zhangjiajie, China, pp. 320-324.* https://doi.org/10.1109/CSAE.2012.6272606.

[3] Q. Feng, J. Chen, M. Zhang and X. Wang, "Design and Test of Harvesting Robot for Table-top Cultivated Strawberry", *WRC Symposium on Advanced Robotics and Automation (WRC SARA), 2019, Beijing, China,* pp. 80-85. https://doi.org/10.1109/WRC-SARA.2019.8931922.

[4] Tianhai Wang, Bin Chen, Zhenqian Zhang, Han Li, Man Zhang, "Applications of machine vision in agricultural robot navigation: A review", *Computers and Electronics in Agriculture,* Volume 198, 107085, ISSN 01681699, 2022, https://doi.org/10.1016/j.compag.2022.107085.

[5] Xu (Annie) Wang, Julie Tang, Mark Whitty, "Data-centric analysis of on-tree fruit detection: Experiments with deep learning", *Computers and Electronics in Agriculture,* Volume 194, 10674, 2022. https://doi.org/10.1016/j.compag.2022.106748.

[6] Fangfang Gao, Wentai Fang, Xiaoming Sun, Zhenchao Wu, Guanao Zhao, Guo Li, Rui Li, Longsheng Fu, Qin Zhang, "A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in the modern orchard", *Computers and Electronics in Agriculture,* Volume 197, 107000, 2022 https://doi.org/10.1016/j.compag.2022.107000.

[7] Anand Koirala, Kerry B. Walsh, Zhenglin Wang, Cheryl McCarthy, "Deep learning – Method overview and review of use for fruit detection and yield estimation", *Computers and Electronics in Agriculture,* Volume 162, July 2019, pp. 219-234, https://doi.org/10.1016/j.compag.2019.04.017.

[8] Tim Van De Looverbosch, Jiaqi He, Astrid Tempelaere, Klaas Kelchtermans, Pieter Verboven, Tinne Tuytelaars, Jan Sijbers, Bart Nicolai, "Inline nondestructive internal disorder detection in pear fruit using explainable deep anomaly detection on X-ray images", *Computers and Electronics in Agriculture,* Volume 197, 106962, June 2022, https://doi.org/10.1016/j.compag.2022.106962.

[9] Lucas Costa, Yiannis Ampatzidis, Charles Rohla, Niels Maness, Becky Cheary, Lu Zhang, "Measuring pecan nut growth utilizing machine vision and deep learning for the better understanding of the fruit growth

curve." ,*Computers and Electronics in Agriculture,* Volume 181, 105964, February 2021. https://doi.org/10.1016/j.compag.2020.105964.

[10] Shenglian Lu, Wenkang Chen, Xin Zhang, Manoj Karkee, "Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation." *Computers and Electronics in Agriculture,* Volume 193, 106696, February 2022, https://doi.org/10.1016/j.compag.2022.106696.

[11] Suharjito, Gregorius Natanael Elwirehardja, Jonathan Sebastian Prayoga, "Oil palm fresh fruit bunch ripeness classification on mobile devices using deep learning approaches", *Computers and Electronics in Agriculture,* Volume 188, 106359, September 2021. https://doi.org/10.1016/j.compag.2021.106359.

[12] W. C. Seng and S. H. Mirisaee, "A New Method for Fruits Recognition", in *International Conference on Electrical Engineering and Informatics, 5-7 August 2009, Selangor, Malaysia.*

[13] S. Naskar and T. Bhattacharya, "A Fruit Recognition Technique using Multiple Features and Artificial Neural Network", *International Journal of Computer Applications*, vol. 116, no. 20, pp. 23-28, April 2015.

[14] Nima Teimouri, Mahmoud Omid, Kaveh Mollazade, Ali Rajabipour, "A novel artificial neural networks assisted segmentation algorithm for discriminating almond nut and shell from background and shadow." *Computers and Electronics in Agriculture,* Volume 105, pp. 34-43. https://doi.org/10.1016/j.compag.2014.04.008, July 2014.

[15] Zareiforoush, H., Minaei, S., Alizadeh, M.R. *et al*., "Qualitative classification of milled rice grains using computer vision and metaheuristic techniques", *Journal of Food Science and Technology*, 53, 118–131, January 2016. https://doi.org/10.1007/s13197-015-1947-4.

[16] Shen, F., Zhang, B., Cao, C., & Jiang, X., "Online discrimination of storage shelf-life and prediction of postharvest quality for strawberry fruit by visible and near-infrared spectroscopy", *Journal of Food Process Engineering*, 41(7), e12866, September 2018, https://doi.org/10.1111/jfpe.12866.

[17] Phongsakhon Tongcham, Pichaya Supa, Peerapong Pornwongthong, Pitcha Prasitmeeboon,"Mushroom spawn quality classification with machine learning", *Computers and Electronics in Agriculture,* Volume 179, December 2020, 105865. https://doi.org/10.1016/j.compag.2020.105865.

[18] K Koyama, M Tanaka, Cho B-H, Y Yoshikawa, S Koseki, "Predicting sensory evaluation of spinach freshness using machine learning model and digital images.", PLoS ONE, 16(3): e0248769. March 19, 2021, https://doi.org/10.1371/journal.pone.0248769.

[19] D. S. Prabha and J. Satheesh Kumar, "Assessment of Banana Fruit Maturity by Image Processing Technique.", Journal of Food Science and Technology, vol. 52, no. 3, pp. 1316-1327, March 2015.

[20] A. AL- Marakeby, A. A. Aly and F. A Salem, "Fast Quality Inspection of Food Products using Computer Vision", International Journal of Advance Research in Computer and Communication Engineering, vol. 2, no. 11, pp. 4168-4171, November.

[21] R. R. Parmar, K. R. Jain, and C. K Modi, "Unified Approach in Food Quality Evaluation using Machine Vision", *Communications in Computer and Information Science (CCIS)*, book series, volume *192*, July.

[22] Ayesha Zeb, Waqar S. Qureshi, Abdul Ghafoor, Amanullah Malik, Muhammad Imran, Javaid Iqbal, Eisa Alanazi, "Is this melon sweet? A quantitative classification for near-infrared spectroscopy", *Infrared Physics & Technology,* Volume 114, 103645, May 2021. https://doi.org/10.1016/j.infrared.2021.103645.

[23] Wanhyun Cho, Myunghwan Na, Sangkoon Kim, and Wonbae Jeon, "Automatic prediction of Brix and acidity in stages of ripeness of strawberries using image processing techniques", in *34th International*

Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), JeJu, Korea (South), 2019, pp. 1-4, https://doi.org/10.1109/ITC-CSCC.20198793349.

[24] Ameetha Junaina T. K, R. Kumudham, Ebenezer Abishek. B, Mohamed Shakir, "Using Deep Learning-Based Features and Image Augmentation to Predict Brix Values of Strawberries for Quality Control", *International Journal of Engineering Trends and Technology*, vol. 71, no. 7, pp. 326-342. 2023. https://doi.org/10.14445/22315381/IJETT-V71I7P231.

[25] Gomes, Véronique, Marco S. Reis, Francisco Rovira-Más, Ana Mendes-Ferreira, and Pedro Melo-Pinto, "Prediction of Sugar Content in Port Wine Vintage Grapes Using Machine Learning and Hyperspectral Imaging", *Processes 9*, no. 7: 1241. https://doi.org/10.3390/pr9071241

[26] Sangsongfa, Adisak & Am-Dee, Nopadol & Meesad, Phayung,"Prediction of Pineapple Sweetness from Images Using Convolutional Neural Network", *EAI Endorsed Transactions on Context-aware Systems and Applications, 7, 165518, 2018.* https://doi.org/10.4108/eai.13-7-2018.165518.

[27] Sarker, I.H, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science*., 2, 160, March 2021. https://doi.org/10.1007/s42979-021-00592-x.

[28] Ameetha Junaina, T.K., Ebenezer Abishek, B., Kumudham, R., Mohammed S, *Lecture Notes in Electrical Engineering*, IN (eds) Futuristic Communication and Network Technologies. VICFCNT 2021, Springer, Singapore, June 2023, Chapter 13, vol 995, PP- 153–163, "Maturity Level Detection of Strawberries: A Deep Color Learning-Based Futuristic Approach.". https://doi.org/10.1007/978-981-199748-8_13

[29] Shorten, C., & Khoshgoftaar, T. M, "A survey on Image Data Augmentation for Deep Learning", *Journal of Big Data*, 6(1), 1-48, July 2019. https://doi.org/10.1186/s40537-019-0197-0.

[30] S. Sergyan, "Color histogram features based image classification in content-based image retrieval systems", *6th International Symposium on Applied Machine Intelligence and Informatics, Herlany, Slovakia*, 2008, pp. 221224. doi: 10.1109/SAMI.2008.4469170.

[31] Amazon Website, Brix Refractometer.https://www.amazon.in/Refractometer-Hydrometer-Measuring-Saccharimeter-calibration/dp/B0B6DRM1DH/ 2024.

[32] Sun, S. Wang, Z. Li, and X. Li, "Feature Engineering for Machine Learning: Principles and Techniques", *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 5, pp. 1162-1179, May 2016.

[33] Basak, J.K.; Madhavi, B.G.K.; Paudel, B.; Kim, N.E.; Kim, H.T, "Prediction of Total Soluble Solids and pH of Strawberry Fruits Using RGB, HSV and HSL Colour Spaces and Machine Learning Models", *Foods*, 2022, 11, 2086. https://doi.org/10.3390/foods11142086.

[34] RNNS, https://medium.com/@salmantahir717/recurrent-neural-networks-rnns-long-short-term-memory-lstm-and-gated-recurrent-unit-gru-a-ec8150a369ce, 2024.