2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE).
Mexico City, Mexico. November 10-12, 2021

Multimodal Deep Learning via Late Fusion for Non-destructive Papaya Fruit Maturity Classification

Cinmayii A. Garillos-Manliguez
Dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan 804
Dept. of Math, Physics, and Computer Science
University of the Philippines Mindanao
Davao City, Philippines 8000
Email: cgmanliguez@up.edu.ph

John Y. Chiang
Dept. of Computer Science and Engineering
National Sun Yat-sen University
Dept. of Healthcare Administration
and Medical Informatics
Kaohsiung Medical University
Kaohsiung, Taiwan 804
Email: chiang@mail.cse.nsysu.edu.tw

***Abstract*—Maturity of fruits significantly affects various areas of the agriculture industry such as the quality assurance of agricultural products, supply chain, and marketing. However, classifying papaya fruit maturity given six ripeness stages with precision remains a challenge since most changes happen inside the fruit rather than the external characteristics, which are quite similar between stages. Using internal properties in classification would require destructive and time-consuming laboratory tests. With the emergence of deep learning and imaging technologies, data with high dimensions, which correlates with internal and external characteristics of an object such as those produced by hyperspectral cameras, can be processed to perform a high-level intelligent classification task without impairing the fruit. In this paper, we present an AI-derived non-destructive approach that utilizes hyperspectral and visible-light images in estimating the papaya fruit maturity stage and implements multimodality via late fusion of imaging-specific networks. The proposed multimodal architecture is composed of imaging-specific deep convolutional neural networks as base learners and a meta-learner that executes late fusion of the dual unimodal networks. Multiclass logistic regression and averaging are explored as the meta-learners of the multimodal fused network that generates the final classifications. Experimental results of the proposed multimodal-late fused models are compared with the multimodal-feature concatenation approach for estimation of papaya fruit maturity, and our proposed model framework obtained an improved F1-score of up to 0.97. This indicates that multimodal-late fused architecture and multimodal imaging systems have great potential for agricultural and other industrial applications.**

I. Introduction

Fruit grading plays a significant role in post-harvest operations due to its effect on the differentiation of market types in the export industry. Maturity is among the grading criteria which evaluates the continuous ripening of the produce after the harvest. It affects valuable qualities such as the flavor, nutrient content, color, texture, and size. Other dependent variables include perishability and susceptibility to damages that increase as the fruit ripens and these can also contribute to 40% of food loss at post-harvest and processing levels [1]. Hence, harvesting at the optimal time of maturity is very critical especially in tropical fruits like papaya fruit.

In practice, an expert must inspect the papaya fruit maturity or grade and destructive methods, e.g. juice extraction in the laboratory, are conducted to quantify internal properties such as total soluble solids (TSS) content, starch, and acidity [2]. These traditional processes are invasive, time-consuming and prone to human and instrumental error. Recently, research studies implementing smart agriculture with non-invasive methods through computer vision (CV) techniques, artificial intelligence (AI), sensors and imaging technologies provided promising results while keeping the fruit in good shape, which help reduce post-harvest losses [3], [4]. Imaging technology with CV and AI is a very viable alternative to fruit grading and ripeness estimation approaches since it only uses images of the samples as input to an AI model which generates the prediction. Hyperspectral imaging, for instance, has become an emerging scientific instrument for non-destructive fruit and vegetable quality assessment in recent years due to the comprehensive spectral and spatial data that it can acquire and its high correlation with physical properties, internal characteristics, and nutrient contents e.g. firmness, texture, TSS, moisture content, acidity, ascorbic acid, etc. [3], [5]. However, this kind of data has high dimensions which makes it computationally challenging.

In deep learning, the modes of learning have advanced from unimodal learning to multimodal learning. Unimodal learning allows a deep learning model to train and perform a task using only a single type of dataset like texts, audio or voice, images, and videos alone. Multimodal learning, on the contrary, enables a model or a method to utilize a combination of input data types such as texts+audio, image+videos, texts+images, and others to accomplish a certain task. Recent studies in agriculture, for example, are now exploring multimodal deep learning or multimodality with data obtained from various imaging technologies such as near infrared (NIR)+red-green-blue (RGB) to detect fruits in the plantation [6] and hyperspectral (HS)+RGB to estimate fruit ripeness [7].

In this work, we present a high-performance multimodal deep learning through late fusion of imaging-specific deep CNNs for classification of papaya fruit maturity using two

978 − 1 − 6654 − 0029 − 9/21/$31.00 ©2021 IEEE

modalities: hyperspectral data cubes and visible-light (VL) images i.e. RGB. The main contributions of this study include the following: 1) a novel multimodal deep learning via late fusion architecture for non-destructive papaya fruit maturity estimation using HS and VL images, 2) three deep CNNs, namely, AlexNet, VGG16 and VGG19 are experimented as the imaging-specific base learners and two methods, namely, averaging and multiclass logistic regression as meta-learners of our multimodal late-fused networks, and 3) a comparison of performance of the proposed approach and some recent works on multimodality in agriculture and evaluation of results in terms of F1 score, top-2 error rate, and computational time.

The remaining parts of this paper is organized as follows: Section (2) presents the studies related to this paper; Section (3) provides a description of a multimodal deep learning architecture with late fusion implementation for papaya fruit maturity estimation using two modes: hyperspectral and visible-light images; experiment results and discussion of the comparative experiments are presented in Section (4); and finally, Section (5) summarizes the observations and points out the future works.

II. Related Works

Machine learning, a primary contributor in the advancement of AI, has progressed from statistical methods, backpropagation, random forest, and support vector machines to more sophisticated supervised, unsupervised, and even semi-supervised learning methods such as ensemble methods, long short-term memory recurrent neural networks, stacked autoencoders, deep belief network, and deep convolutional neural networks. Deep learning has paved its way to agricultural applications such as detection and classification of various fruits in the field using Faster R-CNN [6], estimation of fruit maturity or ripeness using state-of-the-art deep CNNs and transfer learning [7], [8], and detection of bruise or damages on fruits using ResNet/ResNeXt [10]. Although there is a need for intensive data gathering in deep learning, yet with limited dataset resources it is a viable and non-destructive alternative especially that agricultural products are time-critical since they spoil through time and results in wastage.

Furthermore, multimodal deep learning (MDL) is a relatively new approach in agriculture that is germinating and must be cultivated from a research perspective. MDL involves learning representations from at least two different kinds of data. For instance, a deep convolutional neural networks architecture is modified to utilize RGB and NIR images to detect and classify fruits in the field [6]. This system uses two mechanisms of utilizing the RGB and NIR images, which are early fusion and late fusion. The former alters the input layer of the network to provide 4 channels (red, green, blue, and NIR layers) instead of 3, while the latter combines the classification decisions from the output layers of the two networks in which one accepts RGB as input and the other is NIR. With k-fold cross validation, the late fusion with F1 score of 0.838 performed better than 0.799 of the early fusion counterpart. Another recent study used RGB and HS images with a total of 153 combined channels to estimate the papaya fruit maturity [7]. The multimodal input dataset consisting of morphological and spectral features were used to train deep learning models and achieved up to 0.90 F1-score, which is higher than the late fusion result of [6], and 1.45% top-2 error rate.

Imaging technologies like hyperspectral imaging (HSI) system and visible-light imaging have long been studied as non-destructive alternatives for applications in agriculture and food production [11][12]. In [11], studies on hyperspectral or multispectral imaging systems measure the following parameters for fruit grading: quality, defect segmentation, bruise detection, canker detection, light correction, fly infestation, and rottenness detection. On the other hand, visible-light e.g. RGB or traditional imaging measures parameters that include peel defect detection, segmentation, quality inspection, tuber detection, ripeness inspection, and maturity evaluation. With technological advances in both hardware and software, HSI systems with low-cost and fast-detecting properties are anticipated [5] and thereby eliminating the challenge on the cost of HSI systems. Conducting explorational research on the use of hyperspectral images for fruit maturity classification or fruit detection poses an interesting challenge to researchers and practitioners because this bridges two well-established research fields: remote sensing and computer vision.

III. Multimodal Deep Learning Architecture for Papaya Fruit Maturity Estimation

In this section, a multimodal deep learning via late fusion (Multi-DeepLLaF) framework for non-destructive classification of papaya fruit maturity will be discussed. Fig. 1 presents the structure of the proposed method implementing the late fusion approach, which consists of two major levels of learning: level 0 or base learners and level 1 or meta-learner. Two deep CNNs accept HS and VL images as input and establish robust classifiers for the system at the level 0, while the meta-learner in level 1 learns furthermore the patterns of the probabilities produced by each individual base classifier to generate the final classification result. As shown in Fig. 1, one deep CNN (deep learning model or 'DL Model') will be trained using the preprocessed RGB dataset, while the other DL Model will be executing a challenging hyperspectral image classification using the preprocessed HS data cubes of the papaya fruit samples.

A. Hyperspectral Image Classification

Hyperspectral image classification (HIC) can be seen in remote sensing where a single pixel of H bands in a hypercube dataset or an $m \times n \times h$ three-dimensional image data represents the spectral data of the objects in a square meter (m^2) area or $t\ m^2$ depending on the imaging system resolution. HIC in the past decade is gaining popularity as nondestructive techniques in food quality and safety inspection and most of all in agricultural produce especially on tasks involving fresh fruits, etc. In the latter, unlike in remote sensing application, the whole hypercube may only contain one object to be classified. This is the case in this study: each

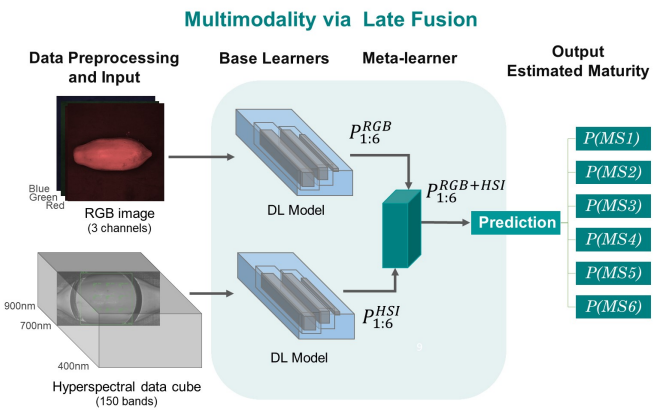


Fig. 1. Multimodal deep learning via late fusion framework for non-destructive papaya fruit maturity classification.

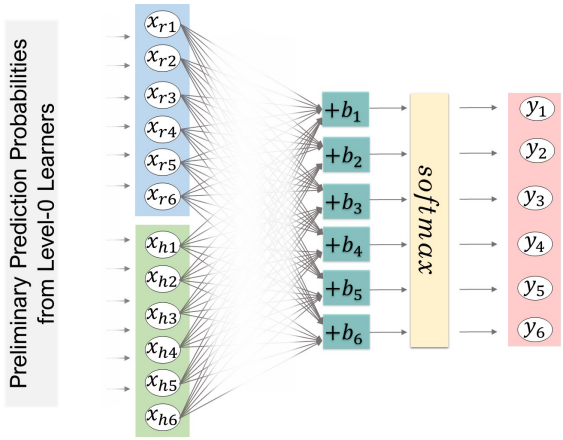


Fig. 2. Multinomial logistic regression structure as meta-learner in the multimodal ensemble deep learning architecture for six-staged fruit maturity classification.

HS data cube is categorized into one maturity stage, which is a very difficult task due to its high dimensionality and requires a large memory capacity. Nevertheless, deep learning e.g. deep CNN has gained reputation in classifying HS images in remote sensing benchmark datasets and most especially with VL images in various tasks and datasets. However, as with other applications, the limited training samples remains a challenge to the researchers and practitioners in remote sensing and agriculture. Hence, this study will contribute to this body of knowledge.

B. Multimodality

One main advantage of deep learning in multimodality is that a deep CNN automatically extracts domain-specific features from the input sample regardless of the preprocessing operations. This also implies that learning the multimodal relationships can happen from low- to high-levels of abstraction in the deep CNN hierarchy of representations to accomplish a designed task [14] [15].

Multimodal learning strategies can be employed in two major ways: (1) feature concatenation (FC) and (2) ensemble method (EM). FC carries out early fusion of input data from multiple modalities to achieve a single feature vector. EM, on the other hand, is a late fusion method that trains and learns the features of each individual modality first before integrating the results to yield the final class estimation [14]. FC is computationally efficient when the complexity of the early fusion setup has been managed meticulously. EM may consume almost twice the resources of FC depending on the constructed architecture, however, this can exploit the synergy of the modality-specific branches in the network that might exceed the former's performance.

In this study, we implemented deep CNNs with modality-specific functions followed by a classifier, multinomial logistic regression, that inherits the predictive powers of these lower-level classifiers in the learning hierarchy to obtain highly accurate classification of papaya fruit maturity. More details will be explained in the following subsection.

1) Multinomial Logistic Regression: Multinomial logistic regression (MLR) generalizes the logistic regression for multiclass problems as represented in the equation below, where $a_k = w_k^T\theta + b_k$ is the activation for k = 1,.., K classes, $w_k = [w_k1, w_k2, \ldots, w_{kM}]^T$ for M input variables, $\theta = [\theta_1, \theta_2, \ldots, \theta_M]^T, C_k$ represents the 1-of-K scheme for each class k. This will learn a set of K weight vectors $w_1, w_2, \ldots, w_K$ and biases $b_1, b_2, \ldots, b_K$.

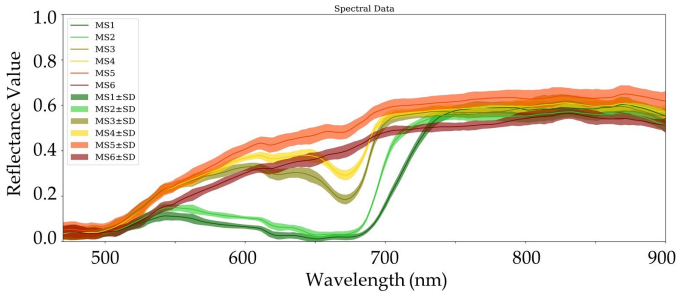
$$p(C_k|\theta) = y_k(\theta) = (exp(a_k))/\sum_j (exp(a_j)) \quad (1)$$

Fig. 2 shows the implementation structure of MLR as the level-1 classifier or meta-learner of Multi-DeepLLaF. The base learners calculated the probabilities i.e. $x_r = [x_{r1}, x_{r2}, \ldots, x_{r6}]$ from level-0 learner or deep CNN of VL images and another set of probabilities $x_h = [x_{h1}, x_{h2}, \ldots, x_{h6}]$ from level-0 learner of HS data cubes. Hence, in the implementation of this study, the probabilities x_r and x_h will be transmitted to the meta-learner as input for training of MLR and adjustment of weights w_k and biases b_k as shown in this equation: $a_k = w_k^T x_r + w_k^T x_h + b_k$.

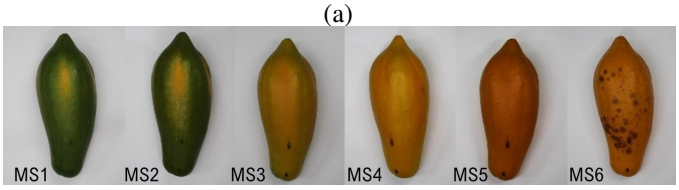
MLR uses Softmax function as an activation function, which transforms the data into positive values and then maps them within the range of 0 to 1. These probabilities are distributed to each output node and then, argmax is used for the final classification of the fruit maturity. To further improve them, the loss function must be minimized through stochastic gradient descent, which calculates the gradient for the randomly selected sample features.

C. Experimental Design

A novel non-destructive multimodal deep learning via late fusion architecture is developed in this study to estimate papaya fruit maturity. Papaya fruit (*Carica papaya L.*) samples, a total of 253, were obtained from Kaohsiung Market and were ripened in a controlled environment. At each ripeness level, the hyperspectral data cubes (HSd) and visible-light images (VLi), specifically RGB, were obtained using Imec SNAPSCAN



(a)



(b)

Fig. 3. Sample HSd and VLi from the six maturity stages of a sample papaya fruit: (a) mean and standard deviation of the hyperspectral data cubes and (b) large-scale RGB images.

visible/near-infrared (VNIR) hyperspectral imaging system and Canon EOS 100D camera, respectively. [7]. The six labels or maturity classes of papaya fruits are described as MS1 or green fruits with a trace of yellow, MS2 or more green than yellow, MS3 or mix of green and yellow, MS4 or more yellow than green, MS5 or fully ripe, and MS6 or overripe. Regarding the distribution of images for each maturity stage: MS1 has 576 data entries for each HSd and VSi data sets, MS2 has 666, MS3 has 792, MS4 has 720, MS5 has 909, and MS6 has 945. The dataset includes 4,608 VLi with 3 channels (red, green, blue) and 4,608 HSd with 150 channels in between 470 nm to 900 nm wavelength data range, and thus having a total of 9,216 data entries. The ground truth classification standards is based on [19], which also complies with the Philippine National Standard (PNS/BAFPS 33:2005). Both HSd and VLi have 32 × 32 pixel dimensions for each channel to reduce redundancy and computation cost. Figure 3a presents the mean and standard deviation of the reflectance values of a sample fruit for each channel of the 150-band HSd for each maturity stage, while Fig.3b displays the VLi of the same sample papaya fruit at six maturity stages.

The training samples set is composed of 2,741 randomly selected data entries, i.e. 60 percent of the total number of data entries from each set of HSd and VLi. The rest of the dataset is divided into 30% or 1,383 data entries for the validation set and 10% or 484 data entries for the test set. A laptop computer with Intel Core i5-9300H 2.40 GHz CPU (8 CPUs), NVIDIA GeForce GTX 1660Ti 6GB GPU, and 8 GB memory space running on Windows 10 Home 64-bit (10.0, Build 18362) system trained the proposed models.

D. Parameter Setting and Learners Experiments

Finding the best set of parameters or hyperparameter optimization is a way to improve model performance. These parameters should fit on the model and its dataset. The high dimensionality of the HSd requires careful selection of parameters. Batch size (B), learning rate (lr), and number of epochs (E) values were under experiment in this study. Best-first search approach was used to obtain the best batch size number from B = (32, 64, 96, 128, 160) for the first pass, and after obtaining the best B value in the first pass, the interval was reduced from 32 to 16 to the left and right of the value for the second pass. The epochs and the learning rate are then studied after selecting B, and the selected values are E = (100, 200, 300) and lr = (0.001, 0.0001, 0.00001), respectively. The summary of parameters and respective values are presented in Table I. As shown in the results of [7], using one of the values in the given sets produced promising results for some deep learning algorithms and thus, incremental values of the mentioned parameters were used in this experiment.

TABLE I
HYPERPARAMETERS FOR MULTI-DEEPLLAF PARAMETER SETTING EXPERIMENT.

Hyperparameters	Values
Batch size (B)	(32, 64, 96, 128, 160)
Number of Epochs (E)	(100, 200, 300)
Learning rate (lr)	(0.001, 0.0001, 0.00001)

Regarding the learners experiments, AlexNet, VGG16, and VGG19 were selected as base learners for comparison because of their promising performance [7], [20], while averaging and MLR were selected as meta-learners for comparison since the input values involved are only vectors of probabilities. To quantitatively assess the performance of the proposed models, F1-score, precision, recall, top-2 accuracy, and logistic regression score (for models with MLR) are used. For the computational time, the program records the training and the prediction time of the models.

IV. RESULTS AND DISCUSSION

This paper proposed a Multi-DeepLLaF approach to classify fruit maturity using two modalities: hyperspectral data cubes and visible light images. In this section, 1) we inspect the parameter setting experiment for the batch size, learning rate, and number of epochs, 2) we examine the training and prediction performance of the three deep CNNs i.e. AlexNet, VGG16 and VGG19 when integrated Multi-DeepLLaF architecture, and 3) we compare the Multi-DeepLLaF method with the feature concatenated – multimodal deep learning (FC-MDL) approach of [7] and evaluate performance in terms of F1 score, top-2 error rate, and computational time.

A. Inspection of Parameter Setting Results

The final list of values tested for the batch size value are enumerated in Table II. For each run, the training time (TT), prediction time (PT), average prediction time (Avg. PT), top-2 error rate, and F1-score were recorded for comparison. Based on the results shown, a batch size of 64 produced the highest macro-average F1-score of 0.96, and the lowest top-2 error rate of 0.005%. The high F1-score, low error rate, and relatively

TABLE II
CLASSIFICATION RESULTS OF MULTI-DEEPLLAF OBTAINED BY VARYING BATCH SIZE ON HSD-VLI DATASET.

Batch Size	TT (s)	PT (s)	Avg. PT (10^{-2} s)	Top-2 (10^{-2})	F1-score
32	885.08	**25.03**	**1.81**	1.74	0.91
64	793.26	27.03	1.95	**0.51**	**0.96**
96	1505.16	28.91	2.09	1.23	0.93
128	**736.12**	29.74	2.15	1.30	0.91
160	1802.25	30.01	2.17	2.17	0.86

TABLE III
CLASSIFICATION RESULTS OF MULTI-DEEPLLAF OBTAINED BY VARYING THE NUMBER OF EPOCHS ON HSD-VLI DATASET.

Epochs	TT (s)	PT (s)	Top-2 error (10^{-2})	F1-score
100	**793.26**	27.03	**0.506**	0.96
200	1148.74	**23.62**	0.578	**0.97**
300	1670.00	30.96	**0.506**	**0.97**

low computation time of B = 64 makes it the best value for this parameter of the model.

In the number of epoch experiment, the selection of the best parameter value considers the computing resources besides F1-score. Table III presents the results of this part of the experiment, and we can see how fast is the overall training time of Multi-DeepLLaF with epoch of 100 compared to the other variables. It also obtained the lowest error rate with minimal margin from the other values.

Table IV displays the experiment results of using variable learning rate on Multi-DeepLLaF. The macro-average F1-scores in this table is as high as the results obtained in Table III. However, with a learning rate value of 0.00001, the model obtained competitive performance with a top-2 error rate of 0.005%, overall prediction time of 27 seconds, and overall training time of 13 minutes and 13 seconds.

B. Performance of Multi-DeepLLaF for Non-destructive Papaya Fruit Maturity Classification

After evaluating the parameter experiments, the best parameters are then used to set the different Multi-DeepLLaF algorithms. Tables V and VI show the individual result of the base learners and the summary of the classification results of the five Multi-DeepLLaF algorithms, respectively. Four algorithms used averaging for level-1: multimodal-late fusion (ML) AlexNet, ML-VGG16, ML-VGG19, and ML-VGG16-AlexNet (MLVA), while the last ML-VGG16-AlexNet used multinomial logistic regression (MLVA-MLR). From table V to VI, the F1-scores improved after late fusion though mostly the VLi-specific learners produced better results that the HSd-specific learners. Due to the complexity of the HSd, the modified Keras deep CNNs were not able to obtain very high accuracy. Among the five algorithms, three exhibited the highest performance in terms of F1-score namely, ML-VGG16, ML-VGG19, and MLVA-MLR, which obtained 0.97 F1-score. In terms of the top-2 error rate or logistic regression (LR) score, ML-AlexNet got the lowest error rate of 0.% followed by ML-VGG16 with 0.007%, while the logistic regression (LR) score of MLVA-MLR is as high as 0.97 i.e. close to 1.0. Moreover, MLVA-MLR is also leading in terms of overall training time followed by ML-AlexNet, which are 705 seconds (11 minutes and 45 seconds), and 793 seconds, respectively. ML-VGG16 also remained competitive in terms of prediction time because in just about 22 seconds this Multi-DeepLLaF algorithm is capable of completing the task it is programmed to do. However, MLVA-MLR outperform the other algorithms in terms of overall training and prediction time i.e. it was done training in about 12 minutes and classifying in just about 10 seconds.

TABLE IV
CLASSIFICATION RESULTS OF MULTI-DEEPLLAF OBTAINED BY VARYING LEARNING RATE ON HSD-VLI DATASET.

Learning Rate	TT(s)	PT(s)	Top-2 (10^{-2})	F1-score
0.00001	648.83	26.07	5.278	0.77
0.0001	793.26	27.03	**0.506**	**0.96**
0.001	649.28	**21.75**	1.085	**0.96**

TABLE V
CLASSIFICATION RESULTS OF UNIMODAL BASE LEARNERS ON HSD AND VLI DATASETS.

HSd-specific learner	VLi-specific learner	TT(s)	PT(s)	Top-2	F1-score
AlexNet		672.53	13.82	0.1005	**0.66**
(AVG)	AlexNet	120.73	13.21	0.0065	0.95
VGG16		780.11	10.34	0.1157	0.60
(AVG)	VGG16	285.01	11.41	0.0086	**0.97**
VGG19		910.43	14.05	0.1395	0.55
(AVG)	VGG19	358.94	12.17	0.0108	0.96
VGG16		803.51	14.92	0.1106	0.64
(AVG)	AlexNet	**114.47**	14.21	0.0087	0.92
VGG16		**461.80**	**8.29**	0.1287	0.60
(MLR)	AlexNet	243.50	**1.95**	0.0116	0.96

TABLE VI
CLASSIFICATION RESULTS OF MULTIMODAL DEEP LEARNING VIA LATE FUSION APPROACHES ON HSD-VLI DATASET.

Base Learners (HSd-specific - VLi-specific)	Meta-Learner	TT(s)	PT(s)	Top-2 (10^{-2}) / LR Score	F1-score
AlexNet-AlexNet	AVG	793.26	27.03	**0.50**	0.96*
VGG16-VGG16	AVG	1065.12	21.75	0.72	**0.97***
VGG19-VGG19	AVG	1269.37	26.22	1.16	**0.97***
VGG16-AlexNet	AVG	917.98	29.13	1.08	0.92*
VGG16-AlexNet	MLR	**705.30**	**10.24**	0.97	**0.97***

* statistically significant at $p < 0.001$ using McNemar's Test

C. Performance Comparison of Multi-DeepLLaF and Other Related Works

For comparison with FC-MDL [7] for papaya fruit maturity estimation, MEVA-MLR will be used by virtue of majority-wins voting rule due to its excellent performance in terms of computation time, prediction time, top-2 error rate and

TABLE VII
PERFORMANCE COMPARISON OF MULTI-DEEPLLAF AND FC-MDL IN TERMS OF PRECISION, RECALL, F1-SCORE, AND ACCURACY.

Deep Learning Methods	Precision	Recall	Top-2	F1-score
MEVA-MLR^{z}	0.97	0.97	0.9711	0.97
MD-VGG16^{z} [7]	0.90	0.90	0.9855	0.90
5-model LF^{y} [23]	0.97	0.97	0.9731	0.97
5-model LF^{x} [23]	0.98	0.98	0.9877	0.98
Faster R-CNN EF^{w} [6]	-	-	-	0.80
Faster R-CNN LF^{w} [6]	-	-	-	0.84

Datasets: ^{z}Papaya Fruits, ^{y}Plant Seedlings, ^{x}CNU Weeds, ^{w}Sweet Pepper

macro-average F1-score. Table VII shows the algorithms and their results in terms of precision, recall, accuracy, and F1-score. From the table, MEVA-MLR, with F1-score of 0.97, of this study is superior to the early fusions of MD-VGG16 [7] and Faster R-CNN [6], with F1-scores of 0.90 and 0.80, respectively; and late fusion of Faster R-CNN [6] with 0.84 F1-score. As seen in the table, the F1-score of the proposed model improved the performance drastically compared to the other existing multimodal framework, i.e. MD-VGG16 [7], using the same dataset. Furthermore, in comparison with 5-model late fusion algorithms [23], the proposed MEVA-MLR model in this study achieved a superior performance at par with these algorithms considering that this study is only a 2-model approach.

V. CONCLUSIONS AND RECOMMENDATIONS

In fruit maturity classification, specifically for Papaya fruit, we have introduced five multimodal deep learning with late fusion algorithms that were implemented in this study. Results showed that the proposed algorithms are very promising for this kind of application. Although, there are still great challenges that the research community in hyperspectral classification has to address. The very high dimensionality of data produced by hyperspectral imaging limits the experiments that can be done e.g. lower hyperparameter values should be used for parameter setting experiments, reduced image sizes, etc. The cost of the imaging system is also a main problem especially when adopting the technology for its applications, however there are already low-cost hyperspectral imaging being introduced in the market today to help companies benefit from the technology while minimizing migration cost.

REFERENCES

[1] L. E. Mopera, "Food Loss in the Food Value Chain: The Philippine Agriculture Scenario," p. 9.

[2] F. F. Calegario, R. Puschmann, F. L. Finger, and A. F. S. Costa, "Relationship Between Peel Color and Fruit Quality of Papaya (Carica papaya L.) Harvested at Different Maturity Stages," 1997. /paper/Assessment-on-the-skin-color-changes-of-Carica-L./ea9a9885e88d8259df4e330149858f90acb7cc70 (accessed Aug. 22, 2020).

[3] B. Li, J. Lecourt, and G. Bishop, "Advances in Non-Destructive Early Assessment of Fruit Ripeness towards Defining Optimal Time of Harvest and Yield Prediction—A Review," Plants, vol. 7, no. 1, p. 3, Mar. 2018.

[4] D. Surya Prabha and J. Satheesh Kumar, "Assessment of banana fruit maturity by image processing technique," J. Food Sci. Technol., vol. 52, no. 3, pp. 1316–1327, Mar. 2015.

[5] Y.-Y. Pu, Y.-Z. Feng, and D.-W. Sun, "Recent Progress of Hyperspectral Imaging on Quality and Safety Inspection of Fruits and Vegetables: A Review," Compr. Rev. Food Sci. Food Saf., vol. 14, no. 2, pp. 176–188, 2015.

[6] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "DeepFruits: A Fruit Detection System Using Deep Neural Networks," Sensors, vol. 16, no. 8, Art. no. 8, Aug. 2016.

[7] C. A. Garillos-Manliguez and J. Y. Chiang, "Multimodal Deep Learning and Visible-Light and Hyperspectral Imaging for Fruit Maturity Estimation," Sensors, vol. 21, no. 4, Art. no. 4, Jan. 2021.

[8] S. K. Behera, A. K. Rath, and P. K. Sethy, "Maturity status classification of papaya fruits based on machine learning and transfer learning approach," Inf. Process. Agric., May 2020.

[9] N. Vélez Rivera et al., "Early detection of mechanical damage in mango using NIR hyperspectral images and machine learning," Biosyst. Eng., vol. 122, pp. 91–98, Jun. 2014.

[10] Z. Wang, M. Hu, and G. Zhai, "Application of Deep Learning Architectures for Accurate and Rapid Detection of Internal Mechanical Damage of Blueberry Using Hyperspectral Transmittance Data," Sensors, vol. 18, no. 4, Apr. 2018.

[11] A. Bhargava and A. Bansal, "Fruits and vegetables quality evaluation using computer vision: A review," J. King Saud Univ. - Comput. Inf. Sci., Jun. 2018.

[12] Y. Liu, H. Pu, and D.-W. Sun, "Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications," Trends Food Sci. Technol., vol. 69, pp. 25–35, Nov. 2017.

[13] F. Mattern and C. Floerkemeier, "From the Internet of Computers to the Internet of Things," in From Active Data Management to Event-Based Systems and More, vol. 6462, K. Sachs, I. Petrov, and P. Guerrero, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 242–259.

[14] V. Radu et al., "Multimodal Deep Learning for Activity and Context Recognition," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 1, no. 4, p. 157:1-157:27, Jan. 2018.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, Art. no. 7553, May 2015.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," p. 8, 2011.

[17] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," in Proceedings of the 2017 ACM International Symposium on Wearable Computers, New York, NY, USA, Sep. 2017, pp. 158–165.

[18] K. M. Ting and I. H. Witten, "Issues in Stacked Generalization," J. Artif. Intell. Res., vol. 10, pp. 271–289, May 1999.

[19] J. Barragán-Iglesias, L. L. Méndez-Lagunas, and J. Rodríguez-Ramírez, "Ripeness indexes and physicochemical changes of papaya (Carica papaya L. cv. Maradol) during ripening on-tree," Sci. Hortic., vol. 236, pp. 272–278, Jun. 2018.

[20] Z. Gao, Y. Shao, G. Xuan, Y. Wang, Y. Liu, and X. Han, "Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning," Artif. Intell. Agric., vol. 4, pp. 31–38, Jan. 2020.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," ArXiv14091556 Cs, Apr. 2015, Accessed: Oct. 02, 2020. [Online]. Available: http://arxiv.org/abs/1409.1556.

[23] V. Hoang Trong, Y. Gwang-hyun, D. Thanh Vu, and K. Jin-young, "Late fusion of multimodal deep neural networks for weeds classification," Comput. Electron. Agric., vol. 175, p. 105506, Aug. 2020.

[24] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," Neural Comput., vol. 10, no. 7, pp. 1895–1923, Oct. 1998.