agronomy

MDPI

Article

Utilisation of Deep Learning with Multimodal Data Fusion for Determination of Pineapple Quality Using Thermal Imaging

Maimunah Mohd Ali [1], Norhashila Hashim [1,2,*], Samsuzana Abd Aziz [1,2] and Ola Lasekan [3]

[1] Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia
[2] SMART Farming Technology Research Centre, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia
[3] Department of Food Technology, Faculty of Food Science and Technology, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia
* Correspondence: norhashila@upm.edu.my; Tel.: +60-3-97694336; Fax: +60-3-89466425

Abstract: Fruit quality is an important aspect in determining the consumer preference in the supply chain. Thermal imaging was used to determine different pineapple varieties according to the physicochemical changes of the fruit by means of the deep learning method. Deep learning has gained attention in fruit classification and recognition in unimodal processing. This paper proposes a multimodal data fusion framework for the determination of pineapple quality using deep learning methods based on the feature extraction acquired from thermal imaging. Feature extraction was selected from the thermal images that provided a correlation with the quality attributes of the fruit in developing the deep learning models. Three different types of deep learning architectures, including ResNet, VGG16, and InceptionV3, were built to develop the multimodal data fusion framework for the classification of pineapple varieties based on the concatenation of multiple features extracted by the robust networks. The multimodal data fusion coupled with powerful convolutional neural network architectures can remarkably distinguish different pineapple varieties. The proposed multimodal data fusion framework provides a reliable determination of fruit quality that can improve the recognition accuracy and the model performance up to 0.9687. The effectiveness of multimodal deep learning data fusion and thermal imaging has huge potential in monitoring the real-time determination of physicochemical changes of fruit.

Keywords: deep learning; thermal imaging; fruit quality; convolutional neural network; multimodal data fusion

Citation: Mohd Ali, M.; Hashim, N.; Abd Aziz, S.; Lasekan, O. Utilisation of Deep Learning with Multimodal Data Fusion for Determination of Pineapple Quality Using Thermal Imaging. *Agronomy* **2023**, *13*, 401. https://doi.org/10.3390/agronomy13020401

Academic Editors: Daniel García Fernández-Pacheco, José Miguel Molina Martínez and Dolores Parras-Burgos

Received: 26 December 2022
Revised: 20 January 2023
Accepted: 27 January 2023
Published: 30 January 2023

1. Introduction

Fruit is partly responsible for human health, as it provides an abundant source of vitamins, minerals, fibres, and nutrients. Pineapple (*Ananas comosus*) is one of the major tropical fruits worldwide and possesses high nutritional composition, pharmacological content, and excellent flavour, as well as growing commercial value [1]. The global pineapple production reached 28.18 million metric tons in 2019, where Costa Rica ranked first as the top pineapple producer worldwide, generating about 3.33 million metric tons [2]. The huge demand for pineapple fruit has further aided various research studies on the postharvest handling and grading/sorting processes, as well as fruit production and cultivation. Efficient postharvest handling should start with producing excellent quality fruits upon harvest, with appropriate measures along the commercial chain until consumption [3]. Typically, the fruit quality is monitored based on physiological growth, where the fruit continues to ripen even after harvesting. Due to the nature of conventional methods that are time-consuming and destructive, recent advances in non-destructive techniques are being developed as an alternative to solve this issue [4]. Recent advances acquired from imaging-based techniques employ multimodal data fusion based on deep learning, which

is beneficial in various food and agricultural applications, since abundant information is retrieved to obtain accurate results. The accurate, rapid, and non-destructive determination of physicochemical changes in fruit can provide a guideline to obtain a better quality of pineapples.

Imaging-based techniques have been developed as one of the non-destructive means to substitute the conventional methods that depend on pattern matching by comparing the input image with reference data to compute the correlation [5]. Thermal imaging has been applied in various food and agricultural applications, including fruit detection by means of differentiating the fruit based on the temperature or thermal properties [6]. Furthermore, thermal imaging can generate the simultaneous monitoring of the dynamic differentiation of temperature by providing a high correlation with the internal quality attributes of the fruit [7]. Döner et al. [8] investigated the effect of ohmic thawing on the temperature distribution of minced beef using thermal imaging. Badia-Melis et al. [9] explored the potential of thermal imaging in predicting the surface temperature of apples in two different packaging boxes with an error of 0.41 °C (plastic boxes) and 0.086 °C (cardboard boxes), respectively. Wang et al. [10] evaluated the potential of a thermal imaging system to investigate the temperature distribution of stem lettuce slices during microwave freeze drying and pulse-spouted microwave freeze drying. In addition, Xu et al. [11] evaluated the temperature distribution of rapeseed using thermal imaging during dielectric drying. Manickavasagan et al. [12] utilised the thermal imaging technique to classify eight different wheats that obtained classification accuracies of 95% using the quadratic discriminant method. Jiang et al. [13] reported the temperature distribution of banana chips using thermal imaging during microwave drying. The findings demonstrated that the temperature distribution of the samples was uniform during the sublimation drying phase. Ma et al. [14] investigated the reaction of plasma-activated water on the decay and quality of Chinese bayberries using the thermal imaging technique. Mohd Ali et al. [15] reported the application of thermal imaging in predicting pineapple quality, which achieved a coefficient of determination (R^2) of 0.94 using a partial least squares model. This study continued further using different machine learning classifiers in order to discriminate the pineapple varieties that obtained the highest classification accuracy of 100% using support vector machines [16].

Deep learning has been extensively applied to solve various classification tasks due to its ability to deal with large datasets in various food and agricultural sectors. A recent work performed by Guo et al. [17] utilised thermal imaging to detect bruises on strawberries using a convolutional neural network (CNN) with an accuracy of 98%. Xie et al. [18] reported an application of deep learning that utilised five different CNN models to identify defective carrots with the highest accuracy of 97%. Benmouna et al. [19] developed a CNN model to classify the ripening stages of apples with a high accuracy rate of 96%. Lin et al. [20] employed a deep convolutional neural network (DCNN) model for the discrimination of three different types of rice kernels. The findings signified that the DCNN model obtained the highest classification accuracy of 99%. Assadzadeh et al. [21] employed deep learning segmentation for predicting grain quality. The results obtained a great prediction performance, with a R^2 of 0.99 for sound grain. Apart from that, Zhou et al. [22] investigated the application of near-infrared spectroscopy coupled with deep learning for the classification of six different types of powdery food, which obtained the highest accuracy of 98%.

Multimodal data fusion is one of the emerging methods in deep learning, which utilises knowledge from one modality that supports and reinforces another [23]. The fundamental concept of multimodal data fusion is to extract features from different layers of combined CNN models to further enhance the capability of the classifier model. The multimodal deep learning method allows retrieving the information from one modality obtained by the data features with minimal human engineering [4]. The utilisation of multimodal data learns representations from different modalities because of the information embedded in the dataset in order to enhance the classification accuracy [23,24]. Although considerable

effort has been made to investigate the chemical and physical characteristics of pineapples, to date, no studies have been conducted to develop classification tasks by developing multimodal data fusion. Deep learning can be deployed in a realistic setting such that feature extraction can produce different classification rates and boost the model accuracy.

In this case, multimodal input data comprised of thermal images are fed into the network architecture to create data features for desired applications. In multimodal data fusion for pineapples, the quality changes of the fruit, along with feature extraction, should be considered when developing a deep learning network model. Utilising the sensitivity of the thermal imaging approach, the performance of multimodal deep learning could leverage the model performance in classifying pineapple varieties according to the physico-chemical changes of the fruit. Thermal imaging integrated with the powerful ability of deep learning offers a non-contact way of generating the temperature distribution of the fruit sample. As a branch of artificial intelligence, deep learning is capable of analysing a huge amount of data by providing more robust analyses with high performance in the thermal information. In this sense, the deep learning approach provides efficient and precise results compared to conventional and routine laboratory analyses. Hence, the aim of this study is to develop a multimodal data fusion framework based on a deep learning approach where multi-network architectures were combined with the feature extraction of pineapples using the thermal imaging technique.

2. Materials and Methods

2.1. Experimental Design

A total of 1080 pineapple fruits from MD2, Morris, and Josapine varieties were harvested from an orchard in Simpang Renggam, Johor, Malaysia. The fruit samples were stored at three different storage temperatures: 5, 10, and 25 °C and a relative humidity of 85–90%. A handheld thermal imaging device (FLIR E60, FLIR systems, King Hills, United Kingdom) was used to obtain the thermal images ranging from 0.7–1.4 μm with an infrared resolution of 320 × 240 pixels and temperature control of −20 °C to +650 °C. For each sample, three different sides of the fruit were acquired. In the case of sampling, the thermal images were obtained that were used for discrimination using deep learning methods, including ResNet, VGG16, InceptionV3, and multimodal data fusion, respectively.

2.2. Data Acquisition

For the thermal imaging acquisition, the setting of the camera device used was FLIR E60 with integrated 3.1 megapixels resolution, 320 × 240 pixels array size, and thermal sensitivity <0.05 °C. The image acquisition of thermal images was conducted in a laboratory room at an ambient temperature. The camera device was positioned perpendicularly at 40 cm above the surface of the fruit. The images were captured and connected to a computer for storing the thermal images. A schematic diagram of image acquisition using the thermal imaging system is illustrated in Figure 1. The image acquisition process was repeated for three replications, which obtained a total of 3240 thermal images. The images stored in the dataset were gathered in order to increase the input data to the neural network and reduce the model overfitting.

The image processing procedures were conducted to extract the information from the thermal image before input into the deep learning models. Firstly, the thermal image of pineapple was converted to a grayscale image. The image processing procedures comprised background removal, which was segmented from the thermal image using the thresholding technique to select the region of interest. The Otsu thresholding technique was used to obtain the threshold value to convert the grayscale image into a binary image. Feature extraction was selected for each region of interest image based on the pixel intensity and shape feature values of the thermal images. All feature values are described in pixel count. The image processing and analysis were performed using MATLAB software (Version R2020a, The MathWorks, Natick, MA, USA).

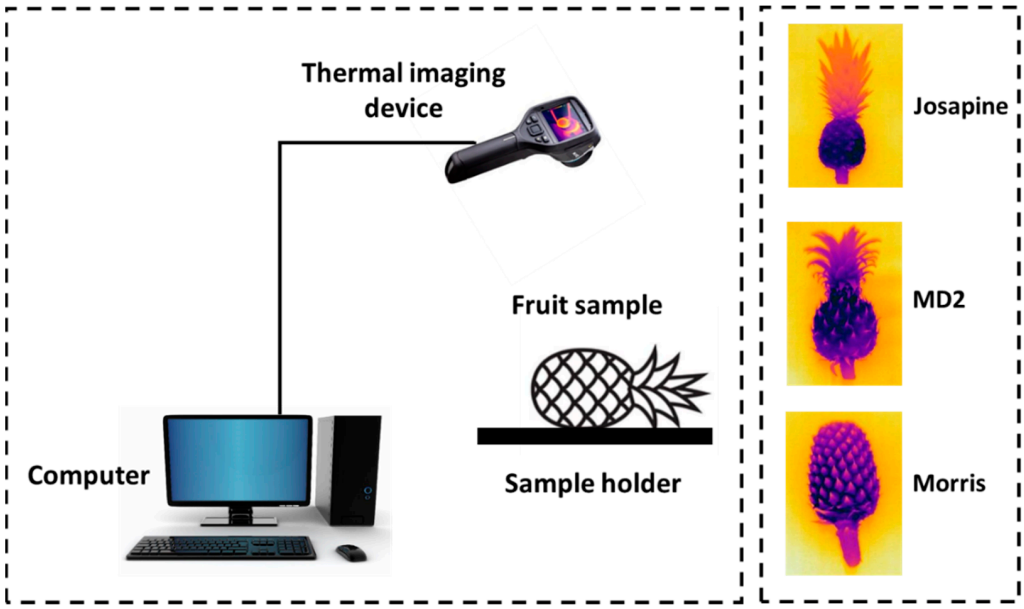

Figure 1. Schematic diagram of image acquisition using the thermal imaging system.

2.3. Fruit Quality Determination

After filtering through Whatman paper no. 1, the soluble solids content (SSC) value was determined from the pineapple juice using a digital refractometer (Pal-1, Atago Co., Tokyo, Japan). The SSC value was computed from the average of three readings and expressed as a percentage. Using a penetrometer (GY-1, G-tech Co., Ltd., Guangdong, China) with a 3.5 mm diameter plunger tip, the firmness of the pineapple flesh was assessed. Three different areas of the pineapples, including bottom, middle, and top, were evaluated for firmness. To get an average value from the samples, maximum force was inserted into the pineapple flesh. Oven drying at 105 °C was used to measure the moisture content until a consistent weight was achieved. In a metal dish, a pineapple cube of 3 cm^3 was dried in an air-drying oven. The percentage based on a wet basis was measured to compute the moisture content.

2.4. Multimodal Data Fusion

After image pre-processing and analysis, the thermal images were integrated for multimodal data fusion. The data fusion based on the selected image parameters was fed into the deep learning network architecture for the classification of the pineapple varieties. Deep learning network architecture signifies a significant role in establishing a multimodal data fusion framework. The advantage of multimodal data fusion is associated with its capability in representation learning, as it could transform raw multimodal input into higher input data. In this sense, the network architectures do not require complicated pre-processing procedures in order to extract the specific features from each model network. Multimodal data fusion is designed by concatenating the extracted features from three CNNs of ResNet, VGG16, and InceptionV3, which were connected to a convolutional layer for the classification task. These state-of-the-art CNN architectures used convolution and pooling operations in several layers with different widths, depths, and cardinality [23]. The convolutional layer assists the network to learn from the concatenated features extracted from those three CNN architectures [25]. In this way, the CNN models could learn the integrated features that are beneficial in the classification tasks. Additionally, it could also be implied that representation learning is significantly faster than a network trained by a single model [26]. Compared with a single model, the accuracy of multimodal data fusion is improved by enhancing the interpretability of the model recognition results. Figure 2 shows the multimodal data fusion of the deep learning framework for the determination of pineapple varieties.

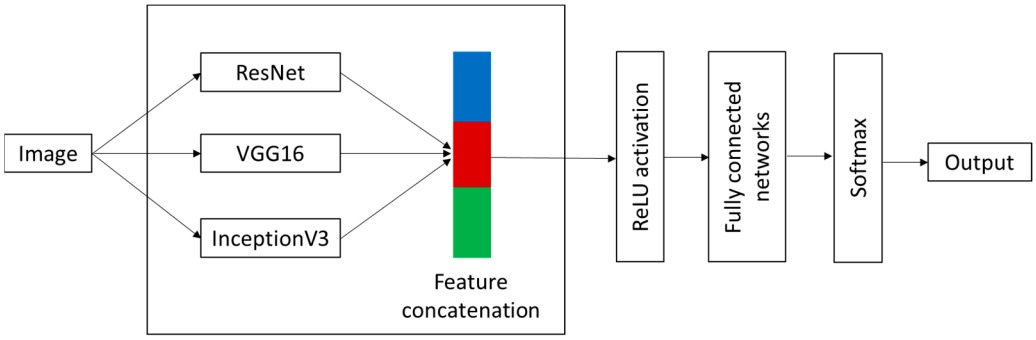

Figure 2. The multimodal data fusion of the deep learning framework.

The same dataset was applied to train the multimodal networks in which three outputs from these three architectures were used as the input data. The utilisation of multimodal data fusion could enhance the classification accuracy by using quality attributes information of the fruit to assist in the recognition task. The recognition of the fruit varieties was defined based on the combination of the fruit information using the multimodal data fusion framework of the deep learning method. The properties of the CNN network architectures and multimodal data fusion are presented in Table 1. Among the deep network architectures used, multimodal data fusion has the highest number of parameters, whereas ResNet has the lowest, with approximately 173.8 million and 11.5 million parameters, respectively. The deep learning network architecture was developed using the framework Keras Version 2.1.4 [27] with TensorFlow Version 1.4 [28].

Table 1. The properties of the CNN network architectures and multimodal data fusion.

Network Architecture	Number of Parameter (Million)	Depth
ResNet	11.5	18
VGG16	138.4	16
InceptionV3	23.9	48
Multimodal data fusion	173.8	82

The image datasets were collected for classification tasks to facilitate the determination of the training and testing processes. During the training and testing processes, the proposed network architectures could automatically extract the image features based on fruit labelling. The sample dataset was randomly divided into a training set and testing set, respectively. In order to assess the performance of the proposed deep learning network architecture for the detection of pineapple varieties, a total of 3240 images was split into 80%, 10%, and 10% for training, validation, and testing, respectively. A stratified sampling was used with an 80:10:10 split ratio, which attempted to keep the same percentages of classes in each split to obtain more representative results. It should be noted that the pineapple surface was fixed in the centre of the sample holder in such a way that the region of interest information was extracted based on the position of the fruit when the thermal image was acquired. The processor for the subsequent training and testing processing platform was built using an Intel® Core™ i5 vPro with 32GB DDR4 memory and a 2TB hard disk.

2.5. Data Analysis

The multimodal data fusion and three single CNN models were evaluated based on precision, recall, F1-score, and accuracy. Precision is defined as the ratio of correct prediction out of the total prediction scores [29]. Recall is denoted as the fraction of the correct prediction that the samples can be classified accurately out of all the actual samples per class [30]. Meanwhile, F1-score is described as the mean recall and precision evaluation that signifies the balance between the classifier from both metrics [31]. In the

fruit industry, it is worth mentioning that recall and precision should be greatly considered due to the quality and safety precautions in obtaining high-quality products. Accuracy is denoted as the fraction of correctly categorised samples among the overall amount of whole samples [32]. In this case, accuracy offers objective decisions for classification tasks with a stable number of testing datasets for each category. These performance metrics are generally used in deep learning applications to analyse the performance of the network architectures for both the training and testing datasets, as described in Equations (1)–(4):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} - \text{score} = 2\frac{\text{TP} \times \text{FP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

where TP (true positive) are the sample datasets categorised by the neural network and ground truth, FP (false positive) is not related to the ground truth in which the sample datasets are categorised by the neural network only, and FN (false negative) is used to count the negative number of sample datasets.

3. Results and Discussion

3.1. Measurement of Fruit Quality

The physicochemical properties measured in the training and testing datasets from three different pineapple varieties are presented in Table 2. The highest firmness values were obtained by Morris for both the training set (1.47 N) and the testing set (2.47 N), respectively. These findings were in agreement with Siow & Lee [33], with the firmness values of the pineapple fruits ranging from 0.20 to 3.50 N. A low firmness value in pineapple was observed in the testing set, which was obtained by Josapine at 0.63 N. For the training set, the highest SSC values were obtained by MD2 (13.50%), whereas the highest SSC value of 12.50% was achieved by Josapine for the testing set, respectively. The research work conducted by Padrón-Mederos et al. [34] revealed an increasing trend in SSC values of pineapples ranging from 11 to 15%. On the contrary, the highest moisture content values were obtained by Morris (91.75%) for the training set and MD2 (88.78%) for the testing set, respectively. The range values of the physicochemical properties of different pineapple varieties were sufficiently large in relation to the training and testing sets. The training sets obtained the highest range compared to the testing sets for MD2 and Morris, which could be utilised to enhance the performance and classification accuracy of the model.

Table 2. Physicochemical properties measured in the training and testing sets from three pineapple varieties.

Variety	Dataset	SSC (%)	Firmness (N)	Moisture Content (%)
Josapine	Training set	12.30 ± 1.16	1.01 ± 0.06	90.76 ± 0.68
	Testing set	12.50 ± 0.95	0.63 ± 0.52	87.27 ± 0.39
MD2	Training set	13.50 ± 1.48	1.43 ± 0.58	85.72 ± 0.84
	Testing set	12.10 ± 0.26	1.01 ± 0.05	88.78 ± 0.89
Morris	Training set	8.20 ± 0.54	1.47 ± 0.93	91.75 ± 0.01
	Testing set	9.70 ± 0.42	2.47 ± 0.15	70.72 ± 1.94

Results are presented as the mean ± standard deviation.

3.2. Evaluation of Trained Networks

The proposed multimodal data fusion was trained using a training dataset of 2592 thermal image samples. The hyperparameters used in the multimodal data fusion CNN model based on the optimal classification accuracy are shown in Table 3. The proposed multimodal data fusion model was validated after the tenth training iteration. The multimodal data fusion was trained by a stochastic gradient descent with momentum (SGDM) and cross-entropy loss function. In this network, the initial learning rate was 0.0001; a batch size of 25 with the maximum epoch was set to 100, which showed the effectiveness of the model during training. These values were selected based on the hardware limitations and several trials that reported better results using these parameter settings.

Table 3. The hyperparameters in the multimodal data fusion CNN model.

Hyperparameters	Values
Classes	3
Batch size	25
Learning rate	0.0001
Epochs	100
Loss function	Cross-entropy
Momentum	0.9
Weight decay	0.0005
Optimizer	Stochastic gradient descent with momentum

In order to train all the deep learning models by optimising the biases and weights, the SGDM algorithm was applied, with cross-entropy as the loss function. In this study, there was no overfitting, since the dropout layer was added, which randomly set the output features of a layer to zero. Furthermore, it was enough to train the multimodal data fusion for 100 epochs to avoid data overfitting. This was because the training accuracy did not converge and the training loss did not change significantly after 100 epochs. Specifically, the multimodal data fusion accuracy did not significantly influence the relationship between the image size and the model's end inference, which could influence the accuracy of the fruit classification. The uncertainty factor associated with the classification accuracy from the multimodal data fusion was determined based on the number of classified image datasets [35].

The comparison in accuracy and loss for the training and validation performance of deep learning models is demonstrated in Figure 3. The trend observed in the training and validation results is shown in the estimation of each deep learning model on the testing dataset. For the training dataset, the multimodal data fusion outperformed the other deep learning models with an accuracy of 94.76% and loss of 0.55, respectively. ResNet, InceptionV3, and VGG16 also achieved promising results, with an accuracy and loss of 87.23–92.45% and 0.61–0.80, respectively. Likewise, a similar trend was also demonstrated in the validation dataset. The multimodal data fusion obtained the highest accuracy of 92.84%, followed by VGG16 (90.58%), InceptionV3 (92.84%), and ResNet (85.06%). In this case, the accuracy obtained from the multimodal data fusion according to feature concatenation distinctly denoted the highest performance compared to the single deep learning models. The multimodal data fusion was more stable compared to the single CNN models. During the learning process, the irrelevant portion of a task was prone to noise, which can enhance the generalisation ability of the models. In this case, multimodal data fusion can improve the model performance to a certain degree that has proven to be superior to the single CNN models. By optimising the shared portion of an epoch twice during fine-tuning, the training loss of the models declined rapidly and became more stable. By comparing different deep learning models with the same hyperparameters, it could determine the superiority of the models under specific configurations. To better understand how the deep learning models are generalised for other issues and domains may be helpful in understanding features training or fine-tuning.

Figure 3. Comparison in accuracy and loss for the training and validation performance of deep learning models.

In terms of model training, the proposed multimodal data fusion has the longest computation time due to the complexity of the model network. Despite using feature datasets based on a single CNN architecture, multimodal data fusion utilised a new recognition model that does not require the execution of finding features from the image datasets. As a result, the performance of multimodal data fusion improved the classification accuracy of the model, since it offers high accuracy rates for the fruit classification task. Additionally, the results demonstrated a strong performance of multimodal data fusion compared to each unimodal architecture designed in the fine-tuning phase. It is vital to choose the features taken from the multimodal data fusion despite having a small number of labelled datasets available during the training process [36].

The comparison of the deep learning models in terms of the consumed time for training is displayed in Figure 4. The multimodal data fusion obtained the longest consumed time (78.3 min) for the training, considering the fact that the model comprised feature concatenation of three different CNN architectures. InceptionV3 took the shortest training time with 33.7 min, followed by VGG16 (40.9 min) and ResNet (61.4 min), respectively. Multimodal data fusion has a huge number of parameters, which causes a long training time, slow convergence speed, and vast storage capacity in practical applications. The weights of the convolution layers are pre-trained by multimodal data fusion in order to avoid a long training time. In this case, dropout is applied to reduce the training time and control overfitting due to a large number of parameters. It should be emphasised that the

model training time of the deep learning models relied on hardware resources that could be shortened by using an advanced graphics processing unit (GPU).

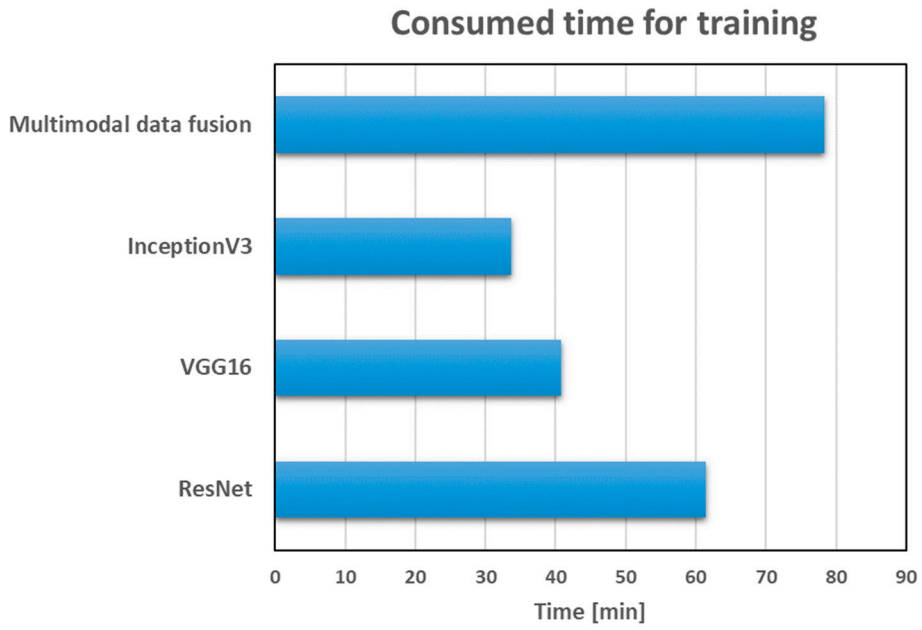

Figure 4. The consumed time for training.

3.3. Modal Comparison of Different Deep Learning Models

Different deep learning network architectures may vary in terms of feature recognition. By fusing these multimodal networks, a high classification accuracy could be achieved compared to a single network. The multimodal data fusion, along with the three CNN architectures, including ResNet, VGG16, and InceptionV3, was developed for fruit quality detection as the basis for the deep learning approach. The methods for the network architectures are evaluated under the same parameter settings and configurations. Based on the findings, it is noted that the proposed multimodal data fusion method has a high accuracy compared to other single models. Figure 5 shows the confusion matrices associated with the correct classification rate (green) and misclassification rate (pink) by different deep learning models. It was observed that the classification results were consistent for MD2 with 100% accuracy for all deep learning models. For multimodal data fusion, it can be seen that MD2 and Josapine were correctly distinguished with a classification rate of 100%, respectively. InceptionV3 and VGG16 also performed well with 99% and 98% accuracies for correctly classified Josapine. On the other hand, ResNet obtained the lowest accuracy compared to the other deep learning models, with 92% accuracy for correctly classified Morris.

The performance comparison of the multimodal data fusion and single CNN architectures in terms of precision, recall, F1-score, and accuracy is illustrated in Table 4. The performance metrics have been evaluated for the estimated labels and ground truth labels used in the pineapple datasets. In terms of precision, the model performance in ascending order was multimodal data fusion (0.9495), VGG16 (0.91110), InceptionV3 (0.9049), and ResNet (0.8932). For recall, multimodal data fusion obtained the highest value of 0.9580, followed by InceptionV3 (0.8963), ResNet (0.8812), and VGG16 (0.8555), respectively. Likewise, the other performance metrics such as F1-score and accuracy showed promising results, with the highest performance values obtained by multimodal data fusion for the classification of pineapple varieties. From the results, multimodal data fusion outperformed the other single CNN architectures in relation to the performance metrics, which were based on the concatenation of the features. Sa et al. [37] proposed a multimodal deep CNN using near-infrared and RGB images for the identification of sweet peppers, which provided a F1-score of 0.838 by considering both the precision and recall performances.

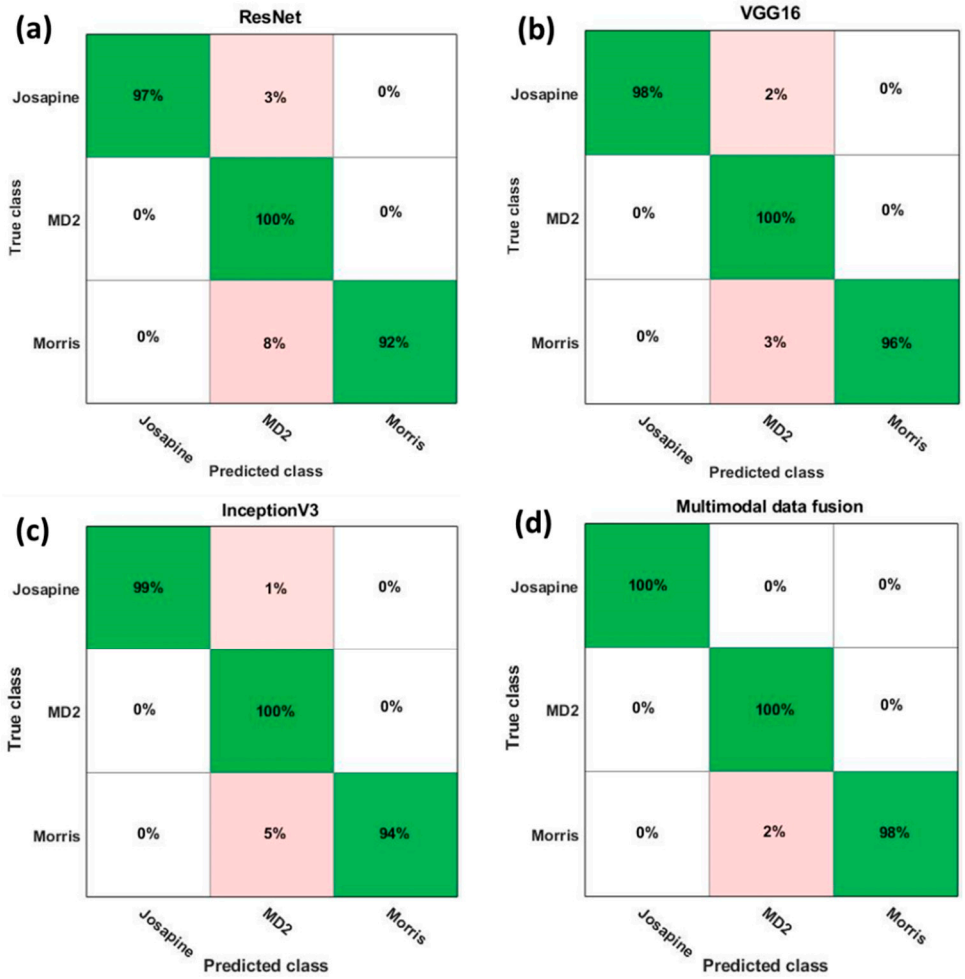

Figure 5. The confusion matrix of deep learning models for (**a**) ResNet, (**b**) VGG16, (**c**) InceptionV3, and (**d**) multimodal data fusion.

Table 4. Performance metrics comparison of the deep learning models.

Deep Learning Models	Precision	Recall	F1-Score	Accuracy
ResNet	0.8932	0.8812	0.9205	0.8385
VGG16	0.9110	0.8555	0.9299	0.8999
InceptionV3	0.9049	0.8963	0.9258	0.9256
Multimodal data fusion	0.9495	0.9580	0.9473	0.9687

Generally, the conventional method for feature extraction is obtained based on the low-level and high-level features. For this reason, the relevant information related to the pineapple samples was correlated depending on the discriminative features of the fruit in terms of shape and pixel value features. The deep learning-based approach was performed to identify the significant feature extractions with minimal data pre-processing steps. The efficacy of the deep learning approach was suitable for related classification tasks in order to extract the optimal features which described the low- and high-level information [25]. Apart from that, the feature concatenation of multimodal data fusion achieved a better performance compared to the single CNN deep learning models, which could improve the interpretability of the required tasks. Through this multimodal data fusion, the performance accuracy of fruit detection was significantly improved by employing three different network models for monitoring the quality changes of pineapple fruit.

3.4. Performance of Multimodal Data Fusion

Among all the model networks, the multimodal data fusion outperformed the single CNN architectures (ResNet, VGG16, and InceptionV3). Transfer learning was not performed due to the multimodal nature of the input image dataset, which was different from the available pre-trained models. Apart from that, overfitting might occur because of the depth of the CNN models that contributed to the strongly adjusted training variables based on the image sizes [4]. These results are in agreement with the studies conducted by Ganesh et al. [38], who developed multimodal segmentation based on the mask region-based CNN method using orange images for fruit detection in the orchard. Zhang et al. [39] proposed a multimodal fusion using a neural network combined with the weight information of fruits in terms of the shape, colour, and texture. Gené-Mola et al. [6] developed a fruit detection system with multimodal images of apples using the radiometric ability of Kinect v2 based on the depth, colour, and range-corrected infrared intensity dataset.

Over the past years, the thermal imaging technique has emerged as a new modality for the quality and safety inspection of various food and agricultural products [40]. The changes in the morphological properties can be determined with thermal images that deliver high sensitivity and high correlation with the internal attributes of the fruit. The thermal images were correlated based on the morphological changes of the fruit during storage when building the multimodal data fusion framework. The proposed multimodal data fusion can be employed for real-time data collection and image analysis from different data representations. Generally, different fields vary in terms of ranges and distributions due to the diverse domains obtained from multiple types of representations and descriptions of multimodal data [41]. In a real-world application, multimodal data fusion has different structures either in the form of unstructured or high-dimensional data. The multi-layer and deep feature representations of multimodal data fusion are explored to fuse the network mechanism with the qualitative information from different storage treatments in relation to the quality attributes of pineapples. The deeper layer could learn high-level features based on the previous layers by generating them from the feature image datasets of each group [42]. For this reason, the last fully connected layers of the trained network achieved from the feature visualisation signify a better representation of the different mechanisms associated with the pineapple varieties.

The performance of the multimodal data fusion was better than the single CNN architectures on all evaluation metrics, which was also improved via unpaired training datasets, as well as multiple modalities. The effectiveness of multimodal data fusion could further learn the feature representation through the conversion between modalities during the training process in order to enhance the encoding paths [23]. Driven by the feasibility of non-destructive techniques and the adaptability of the model network, the fusing of multimodal data features enhanced the fruit classification tasks [24]. Ahn et al. [5] designed multimodal architectures from the hyperspectral images to select feature representation according to the deep neural networks in monitoring the food nutrient estimation. It is worth highlighting that the deep learning architectures are useful in multimodal data recognition tasks. In another study, Garillos-Manliguez & Chiang [4] suggested multimodal classification using a deep CNN based on the feature concatenation of hyperspectral and RGB images of papayas to determine the fruit maturity. Weng et al. [43] explored the potential of multi-feature fusion in terms of texture, morphology, and spectroscopic features from hyperspectral images in order to classify rice varieties.

Having the ability of feature representation learning, no pre-processing is required for extracting the domain-specific features from the input datasets. Due to the high level of learning representation in multimodal data fusion, the corresponding raw datasets could automatically be trained at both low and high levels of abstraction [44]. Multimodal data fusion helps in improving the classification performance of deep learning models that are well suited for diverse highly engineered detection purposes [45]. As the multimodal data fusion integrates information from multiple sources for classification tasks, it is advantageous in terms of functional continuity and robustness of the model. Through

data integration, multimodal fusion combines low-level features from each modality via correlations in order to synchronise among numerous input datasets [46].

4. Conclusions

The deep learning method is capable of reducing the complexity of the feature extraction and enhancing the networks through the representation learning models despite having the high dimensionality of multimodal input datasets. Through feature concatenation of the selected image parameters, the image datasets were fused to create multimodal data fusion during data processing. The selected deep learning model architectures used the multimodal input data and demonstrated promising results for the determination of pineapple quality. Among all the deep learning architectures, multimodal data fusion achieved the highest classification accuracy of 0.9687, followed by InceptionV3, VGG16, and ResNet, which exhibited comparable outcomes. Multimodal data fusion also took the longest training time of up to 78.3 min compared to the single CNN models. Taking into consideration the benefits of multimodal data fusion coupled with powerful deep learning models, the proposed approach delivers a huge potential for the identification of physicochemical changes of fruits in various storage treatments. This study further enhances the foundation of the deep learning method, especially for the data collection of different fruit varieties at various growth stages.

Author Contributions: Data curation, M.M.A.; investigation, M.M.A.; software, M.M.A.; writing—original draft preparation, M.M.A.; visualisation, M.M.A.; supervision, N.H., S.A.A. and O.L.; validation, N.H.; writing—reviewing and editing, N.H.; formal analysis, S.A.A.; and conceptualisation, O.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Putra Grant, GP-IPB (Project code: GP-IPB/2020/9687800).

Data Availability Statement: Not applicable.

Acknowledgments: The authors are thankful for the support and facilities provided by the Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Steingass, C.B.; Grauwet, T.; Carle, R. Influence of Harvest Maturity and Fruit Logistics on Pineapple (*Ananas Comosus* [L.] Merr.) Volatiles Assessed by Headspace Solid Phase Microextraction and Gas Chromatography-Mass Spectrometry (HS-SPME-GC/MS). *Food Chem.* **2014**, *150*, 382–391. [CrossRef] [PubMed]
2. Statista Leading Countries in Pineapple Production Worldwide in 2019. Available online: https://www.statista.com/statistics/298517/global-pineapple-production-by-leading-countries/ (accessed on 3 May 2021).
3. Joy, P.P.; Rajuva, T.A.R. *Harvesting and Postharvest Handling of Pineapple*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
4. Garillos-Manliguez, C.A.; Chiang, J.Y. Multimodal Deep Learning and Visible-Light and Hyperspectral Imaging for Fruit Maturity Estimation. *Sensors* **2021**, *21*, 1288. [CrossRef] [PubMed]
5. Ahn, D.H.; Choi, J.Y.; Kim, H.C.; Cho, J.S.; Moon, K.D.; Park, T. Estimating the Composition of Food Nutrients from Hyperspectral Signals Based on Deep Neural Networks. *Sensors* **2019**, *19*, 1560. [CrossRef] [PubMed]
6. Gené-Mola, J.; Vilaplana, V.; Rosell-Polo, J.R.; Morros, J.R.; Ruiz-Hidalgo, J.; Gregorio, E. Multi-Modal Deep Learning for Fuji Apple Detection Using RGB-D Cameras and Their Radiometric Capabilities. *Comput. Electron. Agric.* **2019**, *162*, 689–698. [CrossRef]
7. Du, Z.; Zeng, X.; Li, X.; Ding, X.; Cao, J.; Jiang, W. Recent Advances in Imaging Techniques for Bruise Detection in Fruits and Vegetables. *Trends Food Sci. Technol.* **2020**, *99*, 133–141. [CrossRef]
8. Döner, D.; Çokgezme, Ö.F.; Çevik, M.; Engin, M.; İçier, F. Thermal Image Processing Technique for Determination of Temperature Distributions of Minced Beef Thawed by Ohmic and Conventional Methods. *Food Bioprocess Technol.* **2020**, *13*, 1878–1892. [CrossRef]
9. Badia-Melis, R.; Qian, J.P.; Fan, B.L.; Hoyos-Echevarria, P.; Ruiz-García, L.; Yang, X.T. Artificial Neural Networks and Thermal Image for Temperature Prediction in Apples. *Food Bioprocess Technol.* **2016**, *9*, 1089–1099. [CrossRef]
10. Wang, Y.; Zhang, M.; Mujumdar, A.S.; Mothibe, K.J. Microwave-Assisted Pulse-Spouted Bed Freeze-Drying of Stem Lettuce Slices—Effect on Product Quality. *Food Bioprocess Technol.* **2013**, *6*, 3530–3543. [CrossRef]

11. Xu, B.; Wei, B.; Ren, X.; Liu, Y.; Jiang, H.; Zhou, C.; Ma, H.; Chalamaiah, M.; Liang, Q.; Wang, Z. Dielectric Pretreatment of Rapeseed 1: Influence on the Drying Characteristics of the Seeds and Physico-Chemical Properties of Cold-Pressed Oil. *Food Bioprocess Technol.* **2018**, *11*, 1236–1247. [CrossRef]
12. Manickavasagan, A.; Jayas, D.S.; White, N.D.G.; Paliwal, J. Wheat Class Identification Using Thermal Imaging. *Food Bioprocess Technol.* **2010**, *3*, 450–460. [CrossRef]
13. Jiang, H.; Zhang, M.; Mujumdar, A.S.; Lim, R.-X. Analysis of Temperature Distribution and SEM Images of Microwave Freeze Drying Banana Chips. *Food Bioprocess Technol.* **2013**, *6*, 1144–1152. [CrossRef]
14. Ma, R.; Yu, S.; Tian, Y.; Wang, K.; Sun, C.; Li, X.; Zhang, J.; Chen, K.; Fang, J. Effect of Non-Thermal Plasma-Activated Water on Fruit Decay and Quality in Postharvest Chinese Bayberries. *Food Bioprocess Technol.* **2016**, *9*, 1825–1834. [CrossRef]
15. Mohd Ali, M.; Hashim, N.; Abd Aziz, S.; Lasekan, O. Quality Prediction of Different Pineapple (Ananas comosus) Varieties during Storage Using Infrared Thermal Imaging Technique. *Food Control* **2022**, *138*, 108988. [CrossRef]
16. Mohd Ali, M.; Hashim, N.; Abd Aziz, S.; Lasekan, O. Characterisation of Pineapple Cultivars under Different Storage Conditions Using Infrared Thermal Imaging Coupled with Machine Learning Algorithms. *Agriculture* **2022**, *12*, 1013. [CrossRef]
17. Guo, B.; Li, B.; Huang, Y.; Hao, F.; Xu, B.; Dong, Y. Bruise Detection and Classification of Strawberries Based on Thermal Images. *Food Bioprocess Technol.* **2022**, *15*, 1133–1141. [CrossRef]
18. Xie, W.; Wei, S.; Zheng, Z.; Jiang, Y.; Yang, D. Recognition of Defective Carrots Based on Deep Learning and Transfer Learning. *Food Bioprocess Technol.* **2021**, *14*, 1361–1374. [CrossRef]
19. Benmouna, B.; García, G.; Sajad, M.; Ruben, S.; Beltran, F. Convolutional Neural Networks for Estimating the Ripening State of Fuji Apples Using Visible and Near-Infrared Spectroscopy. *Food Bioprocess Technol.* **2022**, *15*, 2226–2236. [CrossRef]
20. Lin, P.; Li, X.L.; Chen, Y.M.; He, Y. A Deep Convolutional Neural Network Architecture for Boosting Image Discrimination Accuracy of Rice Species. *Food Bioprocess Technol.* **2018**, *11*, 765–773. [CrossRef]
21. Assadzadeh, S.; Walker, C.K.; Panozzo, J.F. Deep Learning Segmentation in Bulk Grain Images for Prediction of Grain Market Quality. *Food Bioprocess Technol.* **2022**, *15*, 1615–1628. [CrossRef]
22. Zhou, L.; Wang, X.; Zhang, C.; Zhao, N.; Taha, M.F.; He, Y.; Qiu, Z. Powdery Food Identification Using NIR Spectroscopy and Extensible Deep Learning Model. *Food Bioprocess Technol.* **2022**, *15*, 2354–2362. [CrossRef]
23. Kanezaki, A.; Kuga, R.; Sugano, Y.; Matsushita, Y. *Deep Learning for Multimodal Data Fusion*; Elsevier Inc.: Amsterdam, The Netherlands, 2019; ISBN 9780128173589.
24. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
25. Noreen, N.; Palaniappan, S.; Qayyum, A.; Ahmad, I.; Imran, M.; Shoaib, M. A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor. *IEEE Access* **2020**, *8*, 55135–55144. [CrossRef]
26. Zhou, J.; Li, J.; Wang, C.; Wu, H.; Zhao, C.; Teng, G. Crop Disease Identification and Interpretation Method Based on Multimodal Deep Learning. *Comput. Electron. Agric.* **2021**, *189*, 106408. [CrossRef]
27. Alves, T.S.; Pinto, M.A.; Ventura, P.; Neves, C.J.; Biron, D.G.; Junior, A.C.; De Paula Filho, P.L.; Rodrigues, P.J. Automatic Detection and Classification of Honey Bee Comb Cells Using Deep Learning. *Comput. Electron. Agric.* **2020**, *170*, 105244. [CrossRef]
28. Duong, L.T.; Nguyen, P.T.; Di Sipio, C.; Di Ruscio, D. Automated Fruit Recognition Using EfficientNet and MixNet. *Comput. Electron. Agric.* **2020**, *171*, 835. [CrossRef]
29. Villacrés, J.F.; Cheein, F.A. Detection and Characterization of Cherries: A Deep Learning Usability Case Study in Chile. *Agronomy* **2020**, *10*, 835. [CrossRef]
30. Ngugi, L.C.; Abelwahab, M.; Abo-Zahhad, M. Tomato Leaf Segmentation Algorithms for Mobile Phone Applications Using Deep Learning. *Comput. Electron. Agric.* **2020**, *178*, 105788. [CrossRef]
31. Hu, Z.; Tang, J.; Zhang, P.; Jiang, J. Deep Learning for the Identification of Bruised Apples by Fusing 3D Deep Features for Apple Grading Systems. *Mech. Syst. Signal Process.* **2020**, *145*, 106922. [CrossRef]
32. Koirala, A.; Walsh, K.B.; Wang, Z.; McCarthy, C. Deep Learning—Method Overview and Review of Use for Fruit Detection and Yield Estimation. *Comput. Electron. Agric.* **2019**, *162*, 219–234. [CrossRef]
33. Siow, L.-F.; Lee, K.-H. Determination of Physicochemical Properties of Osmo-Dehydrofrozen Pineapples. *Borneo Sci.* **2012**, *31*, 71–84.
34. Padrón-Mederos, M.; Rodríguez-Galdón, B.; Díaz-Romero, C.; Lobo-Rodrigo, M.G.; Rodríguez-Rodríguez, E.M. Quality Evaluation of Minimally Fresh-Cut Processed Pineapples. *LWT Food Sci. Technol.* **2020**, *129*, 109607. [CrossRef]
35. Katarzyna, R.; Paweł, M. A Vision-Based Method Utilizing Deep Convolutional Neural Networks for Fruit Variety Classification in Uncertainty Conditions of Retail Sales. *Appl. Sci.* **2019**, *9*, 3971. [CrossRef]
36. Khan, R.; Debnath, R. Multi Class Fruit Classification Using Efficient Object Detection and Recognition Techniques. *Int. J. Image Graph. Signal Process.* **2019**, *11*, 1. [CrossRef]
37. Sa, I.; Ge, Z.; Dayoub, F.; Upcroft, B.; Perez, T.; McCool, C. Deepfruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* **2016**, *16*, 1222. [CrossRef]
38. Ganesh, P.; Volle, K.; Burks, T.F.; Mehta, S.S. Deep Orange: Mask R-CNN Based Orange Detection and Segmentation. *IFAC-PapersOnLine* **2019**, *52*, 70–75. [CrossRef]
39. Zhang, W.; Zhang, Y.; Zhai, J.; Zhao, D.; Xu, L.; Zhou, J.; Li, Z.; Yang, S. Multi-Source Data Fusion Using Deep Learning for Smart Refrigerators. *Comput. Ind.* **2018**, *95*, 15–21. [CrossRef]

40. Sun, Y.; Lu, R.; Lu, Y.; Tu, K.; Pan, L. Detection of Early Decay in Peaches by Structured-Illumination Reflectance Imaging. *Postharvest Biol. Technol.* **2019**, *151*, 68–78. [CrossRef]
41. Liu, J.; Li, T.; Xie, P.; Du, S.; Teng, F.; Yang, X. Urban Big Data Fusion Based on Deep Learning: An Overview. *Inf. Fusion* **2020**, *53*, 123–133. [CrossRef]
42. Lu, T.; Yu, F.; Xue, C.; Han, B. Identification, Classification, and Quantification of Three Physical Mechanisms in Oil-in-Water Emulsions Using AlexNet with Transfer Learning. *J. Food Eng.* **2021**, *288*, 110220. [CrossRef]
43. Weng, S.; Tang, P.; Yuan, H.; Guo, B.; Yu, S.; Huang, L.; Xu, C. Hyperspectral Imaging for Accurate Determination of Rice Variety Using a Deep Learning Network with Multi-Feature Fusion. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *234*, 118237. [CrossRef]
44. Radu, V.; Tong, C.; Bhattacharya, S.; Lane, N.D.; Mascolo, C.; Marina, M.K.; Kawsar, F. Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* **2018**, *1*, 157. [CrossRef]
45. Huang, S.C.; Pareek, A.; Zamanian, R.; Banerjee, I.; Lungren, M.P. Multimodal Fusion with Deep Neural Networks for Leveraging CT Imaging and Electronic Health Record: A Case-Study in Pulmonary Embolism Detection. *Sci. Rep.* **2020**, *10*, 22147. [CrossRef] [PubMed]
46. Pandeya, Y.R.; Lee, J. Deep Learning-Based Late Fusion of Multimodal Information for Emotion Classification of Music Video. *Multimed. Tools Appl.* **2021**, *80*, 2887–2905. [CrossRef]