Small Convolutional Neural Network Trainer Designed through Transfer Learning in Dessert Classification

Hua-Ching Chen[1], Hsuan-Ming Feng[2]*

[1] School of Information Engineering, Xiamen Ocean Vocational College, China
[2] Department of Computer Science and Information Engineering, National Quemoy University, Taiwan
galaxy.km@gmail.com, hmfenghmfeng@gmail.com

Abstract

This paper established a Convolutional Neural Network (CNN) with the concept of transfer learning, and combined the main feature analysis calculation and clustering algorithm to further demonstrate the superiority of the proposed small CNN trainer in the identification of traditional Kinmen desserts. Food dessert identification methods directly use skin texture, color, shape, and other features as the basis. This paper effectively extracted image features of an object by the small CNN trainer and classified the featured dataset into the right food categories. It was able to complete classification quickly and also achieved high-precision classification results. Different types of Kinmen desserts were identified through a multi-layer training cycle. A total of 100 training images for each of the 10 food categories and each image size is converted into a smaller training data set by capturing the important features through the CNN trainer. Then, the main features were generated and the dimensions of each food image data were reduced again by using the t-Distributed Stochastic Neighbor Embedding (t-SNE) or Principal Component Analysis (PCA) methods. An individually K-means or k-nearest neighbors (KNN) algorithms efficiently completed the grouping results and in the classified image restoration. The experimental results compared the classifications after the learning cycle of different trainers and showed that the highest accuracy that the appropriated CNN trainer of the proposed methology obtained was up to 99% with a minimum executing time of about 99.37 seconds.

Keywords: Deep Learning, K-nearest neighbors (KNN), K-means, t-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA)

1 Introduction

The most primitive food in human history is a cake-shaped dessert made of flour, oil, and sugar. In addition to the quality of materials, the shape and color are essential to the identification of Kinmen's traditional dessert. China is very particular about traditional food, especially pastries. Therefore, China's food culture is one of the most recognized in the world. Its attractive charm has spread around the world, forming a special cultural consumption field. Among them, desserts often attract the attention of consumers of all ages. With the promotion of globalization and the Internet communication chain, consumers are now more eager to taste various local delicacies of different countries. With the movement of the globalized society and population, different eating habits and cultures have spread to other regions. The human pursuit of sweet taste because of the satisfaction it brings has a long history. Chinese-style dessert is different from those of the French and Japanese; unlike French desserts, which are delicate in appearance, rich in taste, and easy to identify, Chinese desserts are not easy to recognize at a glance. A tool should be developed to enable the public to obtain more convenience, joy, and satisfaction in finding Chinese food, especially desserts.

Through investigation, it was found that although the desserts launched by star-rated hotels or businesses in various countries follow traditional production, they have developed many different forms of distinctive desserts. With the different experiences and learning backgrounds of the makers, they have launched a variety of unique desserts to meet the needs of most consumers; however, in terms of the presentation of the desserts, the color, which is vital to the identity of the dessert, has been changed from the original. So, these desserts have become difficult to judge visually. The desserts uniquely designed by the dessert master have made it easier for consumers to identify them visually. Because of their visual appeal, these desserts have won international praise and recognition. The different methods of making local Kinmen desserts by each master demonstrate the increase in the market demand for food and drink culture in various countries. Secret image information becomes an important method to maintain the image integrity and information consistency while transfer desserts into the customer in the new era of the Internet of Things [1-3]. On the other hand, the advanced image classification technology of AI can help each customer quickly identify the type of food product and facilitate their selection and purchase.

Due to the rapid development of AI convolutional neural networks (CNN) in image processing, there have been a lot of applications in the classification of food images. For example, Ciocca [4] and others used CNN feature extraction and SVM object classification methods to establish a classification system that can distinguish food status and species. Uddin and other scholars [5] proposed a CNN image classification system with a transfer learning mechanism; it first uses the trained VGG16 model and then adds several simple fine-tuning stages to improve the accuracy of Bengali's traditional

*Corresponding Author: Hsuan-Ming Feng; E-mail: hmfenghmfeng@gmail.com
DOI: 10.53106/160792642023122407009

food category classification to about 98%. Tran and other scholars [6] completed the food recommendation system through the characteristics of the early Hog technology combined with different machine learning and deep learning model construction and logical regression and other analysis technologies. Yogaswara and Wibawa and other scholars proposed supervised learning algorithms such as MLP, CNN, etc., which can be applied to the identification of different objects such as food and non-food [7]. Meanwhile, Patil and Burkapalli [8] put forward the training model of ResNet and Inception V3 of transfer learning in depth learning, and completed the classification analysis of different types of food through residential learning. The general training method of CNN is to directly establish the recognition model through the processing of a multi-layer Convolution, Pooling, and Fully Connected Layer. Generally, the amount of computing data to be processed is very large, because the huge training data is necessary to obtain relatively high accuracy. Further, selecting appropriate and effective training to identify object feature values is also a very important factor. The contribution of this study in the field involves the use of the existing trained CNN model and inputting the image data of traditional Golden Gate pastry to capture its features. By analyzing and screening the main components of this feature data, the dimension of the overall training data can be reduced [9-10]. The receiver finds the cluster center of training data through a clustering algorithm and fuses the data set close to the cluster center into the same group. Through the tag of image data, the application of Chinese style western point type identification can be completed quickly. This method can reduce the training data, establish a small data training set for a specific Chinese-style western point, and compare the efficiency of common CNN, which can obtain a more accurate and efficient way to analyze traditional Chinese snacks. The research focus of this paper is to extract the features of the dataset by using several well-known CNN pre-training models such as AlexNet, GoogleNet, ResNet50, and MobileNet, and then combine the main feature extraction methods such as t-distributed stochastic neighbor embedding (t-SNE) or principal component analysis (PCA) to reduce the size of the eigenvalues and maintain the main characteristics of the original eigenvalues. Next, a simple and efficient K-means algorithm is used to analyze the feature distribution of each kind of food and classify the foods with similar feature values into the same group. Finally, the foods with the same group of feature values are classified into the same kind, completing the classification application of AI in traditional Chinese snacks.

2 Related Works

2.1 Image Feature Extraction

Image feature extraction is important to make the best description in food recognition applications. Since food is not always a grid object, its feature has an intrinsic high intra-class variability with different shapes, colors, and appearances even if the same food is being described. Automatic food recognition is an important technology for understanding the desired object categories, which are used in different applications. For example, a picture is taken using a smartphone and is then sent to an automatic food recognition system for identifying different foods and estimating the food quality. Another is when the total calorie of food needs to be calculated to control or monitor the weight of an individual for health reasons [11].

There are two main feature items, one is called Hand-Crafted (HC) features and the other is Learned Features (LF). The Hand-Crafted (HC) features are usually produced manually through the expert's knowledge [12-13]. Many researches employed fusion or separation methods to extract the appropriated Hand-Crafted (HC) features in food recognition applications [14].

The popular multiple kernel learning (MKL) has been applied in single food recognition applications. Many related food recognitions first extract the color histogram and Gabor texture features using the Histogram of Oriented Gradient (HOG) [15] and SIFT bag-of-features [16] respectively to approach the desired pre-trained model and make better classifications.

In SVM technology, the considered textons and the food images of vocabulary are used for classification applications [17]. In the same way, the selected local binary patterns and relationships between SIFT interest points to extract the local and spatial information [18].

The learned features data is always extracted by machine learning, especially CNN. CNNs are usually constructed through several layers, which work in linear and nonlinear operators. All parameters of CNN are regulated to reach an end-to-end manner for a particular task. The other considered case is that CNN is pre-trained individually to get the characteristics of data to prepare for the next tasks. Based on the powerful end-to-end feature extractions, many studies have taken in food recognition.

CNN is known as a very powerful deep-learning method for large-scale object recognition. However, a lot of training images are required to get the best accuracy for food applications. Therefore, this research prepared the pre-trained network by utilizing a large amount of image data and using the activation function to extract features from the CNN. Finally, it was programmed to regulate the fine-tuned networks for better approaching results [19].

2.2 Kinmen Dessert

The Kinmen pasty is a famous dessert in Taiwan. It usually resembles some part of its traditional culture [20]. People get their creative ideas of a country's food habits and food culture based on how the locals make their traditional food. The people of Taiwan are known for their hospitality, and this is shown in Kinmen's pasty. Kinmen's dessert is famous for its unique taste and how it showcases Taiwanese traditions. Sophisticated Kinmen pastries are still being made by the masses, and through several food events, they are promoted as part of cultural exchanges. Multimedia broadcasting used by gourmet enthusiasts is a powerful tool to influence and affect others' awareness and perception of food. In particular, almost everyone has a mobile phone that can take pictures of Kinmen pastries anytime and anywhere. The CNN model is known as a powerful and convenient tool used to recognize the shape, color, and texture of different

types of food, which could easily confuse an untrained eye such as tourists visiting Taiwan. Therefore, this paper employed a pre-trained CNN model to extract the original features of Kinmen pastries and select the primary kernel from the generated features of CNN to approximate the small but important core information for the efficient purpose in the next training stage.

2.3 Deep Learning (DL)

Machine learning has developed and progressed into the field of deep learning (DL) in the last decade. CNN is an important technology in the field of DL; through it, image classification tasks of various domains achieve better performance. Scholars have utilized the MLC-41 database with 41 food labels and 100 images for each label to perform the desired food classification system. ResNet deep feature sets are first collected to get better information gain and make a great SOM classifier with outstanding accuracy [21]. Meanwhile, one study employed Inception v3 as the transfer learning model in Indian food image recognization to minimize the training time and get better accuracy [22]. In another research, 3,960 Thai food images were collected as a training dataset and their fine-tuned strategies with Inception V3 were proposed to reach high accuracy [23]. Consequently, one research compared the Inception v3 model to SVM, Resnet18, neural network, and other related machine learning algorithms in the applications of FOOD-10; the former obtained the highest recognition accuracy of 83.97% [24]. The combination of the Inception v3 model and scratch CNN model was found to have better food classification results in the food-11 dataset [25]. This is an important finding because it only requires a small memory size, is more efficient in terms of computational time, and has better evaluation results. Due to its ability to realize powerful data, a pre-trained CNN model is likely to fill the domain knowledge in specific applications. This paper used the CNN feature extractor to fit the original data to create realistic features of Kinmen pasty. Convolution, pooling, and activation are complicated operations; thus, the original image is divided into several smaller visual feature pieces. Next, the suitable fully connected output layer is probed to collect the visual feature. After this stage, more essential characteristics are made available to be perfectly realized as the structure of the feature; however, the produced data is too large. Therefore, this study believes that a pre-trained CNN model with a transfer learning concept is a vital technology to get the feature in the discussed image set.

Nowadays, DL has been successfully applied to many application levels [26-28]. The success of the DL algorithm lies in its ability to learn characteristic values through a large number of data and extract those characteristics sequentially through unsupervised and semi-supervised learning methods. Therefore, data volume is an important topic in DL. Compared with traditional machine learning, DL is highly dependent on the amount of data because a large amount of data can fully understand the potential structure of the data. Generally speaking, the model size is linearly related to the resolution required by the data. The representation space of the model must be large enough to explore the patterns in the data and obtain the required features. This paper used the concept of transfer learning to obtain the required features in a relatively small amount of training data; then, the features of the training data set were extracted and the size of the data set was reduced through the method of main feature extraction. Finally, an individually K-means or KNN algorithm is to [approximate the point into the correct group and then rebuild images based on label's index.

3 Hybrid-Convolutional Neural Network (CNN) Feature Extraction and Reduction Dimension Classifier Design

3.1 Progressive Neural Networks based on Transfer Learning

In the field of image recognition, CNN has evolved to the extent that it can surpass the ability of the human eye and human brain processes, as the former can further refine and process part of the surrounding units. For image recognition processing, CNN is superior to other algorithms [29] as it has better operation results in image recognition. However, the training speed of the CNN model is dependent on the number of data sets. This makes the data collection training and calculation time too long. This paper designed a hybrid CNN combined with a pre-trained model, feature extraction, and dimension reduction classifier based on the transfer learning concept which is illustrated in Figure 1.

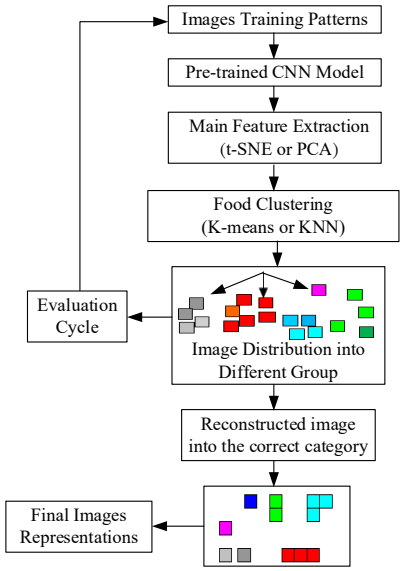

Figure 1. Small CNN design

This paper used the concept of transfer learning to complete the pre-trained model of the input image based on some well-known training models such as AlexNet, GoogleNet, ResNet50, and MobileNet, and employed it as a tool to extract preliminary characteristics through the filtering process from some CNN pre-trained model. Initially, the output layer of the original model network is deleted. Then, the image data samples of Kinmen dessert are fed into the selected pre-trained model network which is done at one time to obtain the main feature from the pre-trained model. The t-distributed stochastic neighbor embedding (t-SNE)

or Principal Component Analysis (PCA) are taken in the main feature extraction process to individually generate the important features from the distribution values of the pre-trained model. Next, the clustering algorithm is utilized to fuse data with high similarities into the same group. These are then correctly classified into 10 categories. Finally, the image restoration is based on fast and efficient food index analysis to match the desired image for presentation.

In this paper, the feature extraction of convolutional network using transfer learning concept is explained as follows: Firstly, the original food samples are selected and input to the selected pre-trained network model, and the CNN network obtains the feature vector of the food data set from the pre-trained model after one-time computation. The original image size of 1210 by 572 by 3 is transformed into a feature vector of size 1000 by 2, and the characteristics of each CNN model and the dynamic feature change experiments are described below.

In the AlexNet network, the image training data is computed by five convolutional and pooling layers and the overfitting is reduced by using dropout and data augmentation in the fully connected layers of layers 6 and 7. The feature vectors captured by the output of the fully connected layers of layer 8 can be input to the later feature data downscaling computation.

Figure 2 shows the dynamic feature variation within the AlexNet structural layers, where Figure 2(a) is the feature map corresponding to the first convolutional layer, Figure 2(b) shows the original input graph on the left, the result of the investigate the activations in specific Conv1 channels with AlexNet on the right, and Figure 2(c) shows the result of the investigate the activations in specific Conv1 channels with AlexNet on the fifth convolutional layer. Figure 2(c) shows the result of the feature map with the strongest values in the feature map corresponding to the fifth convolutional layer. Figure 2(d) shows the final FC8 fully-connected layer with the feature distribution generated by AlexNet and the feature data input in the next step dimensionality reduction.

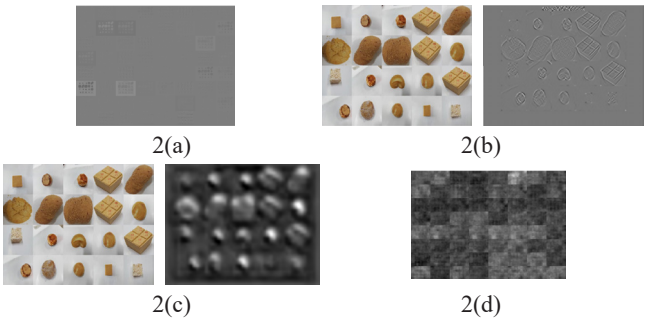

Figure 2. Feature variation within the AlexNet structural layer

The main concept of GoogleNet developed by Google is to concatenate 9 Inception modules with a wide structure and classification auxiliaries to build a deeper network structure with both width and depth, which is prone to the problem of gradient disappearance because of the deep network structure. The image training data will be first processed by convolutional operation and entered into the Inception module, and finally the feature vectors captured by the output of the loss3-classifier layer will be input to the later feature data downscaling operation.

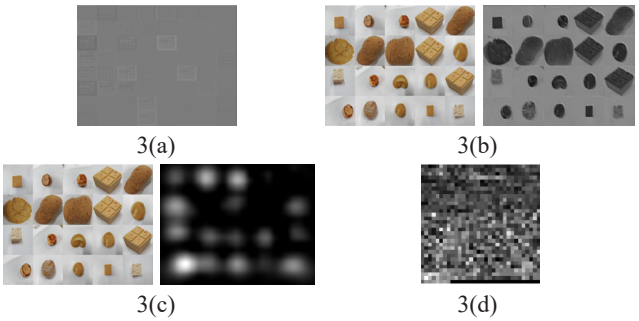

Figure 3. Feature variation within the GoogleNet structural layer

Figure 3 shows the feature variation within the GoogleNet structure layer, where Figure 3(a) is the feature map corresponding to the first convolutional Conv1-7X7-s2 layer, Figure 3(b) shows the original input graph on the left, and the result of the investigate the activations in specific Conv1-7X7-s2 channels with GoogleNet on the right is the result of the investigate the activations in specific Conv1-7X7-s2 channels with GoogleNet layer. Activations in specific Conv1-7X7-s2 channels with GoogleNet), Figure 3(c) is the result of labeling the points with the strongest values in the feature map corresponding to the Inception 5b module layer. Figure 3(d) is the result of the final loos3 This result is input to the feature data downscaling calculation.

MobileNet can be called the inception of lightweight networks, and its basic unit is depth-wise separable convolution (DSC). This method can double the time complexity and space complexity of the convolutional layer. It solves this problem by replacing the last layer of ReLU with a linear activation function. In this paper, the image training data is first computed by a convolutional operation and entered into 16 block modules including deep convolution and point-by-point convolution, and finally the feature vector captured by the output of the fully connected layer of logits is input to the later feature data dimensionality reduction operation.

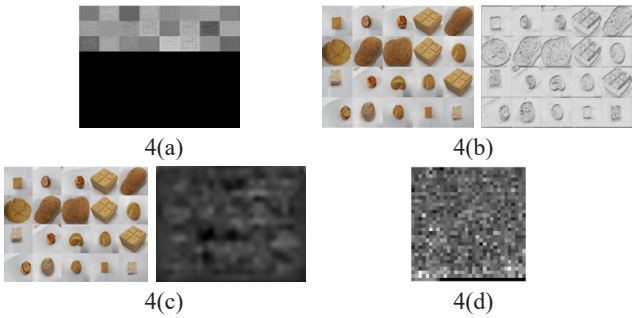

Figure 4. Variation of features within the MobileNet structure layer

Figure 4 shows the feature variation within the MobileNet structure layer, where Figure 4(a) is the feature map corresponding to the Block1-project, Figure 4(b) shows the original input graph on the left, and the result of the first Block1-project layer is presented by labeling the points of the activation function in the investigation map corresponding

to the first Block1-project layer on the right (Result of the investigate the activations in specific Block1-project channels with MobileNet), Figure 4(c) shows the result of the strongest value of the points in the feature map corresponding to the Block15-project layer, and Figure 4(d) is the result of the final Logits full connection layer. Figure 4(d) is the final Logits fully-connected layer, which is the feature distribution generated by MobileNet, and input this information to the feature data downscaling operation.

The idea of residual learning proposed by ResNet50 network can make the deep network easier to train. ResNet50 uses the extended residual block to become bottleneck block and spreads the overlapping action to complete a total of more than 170 network layers.

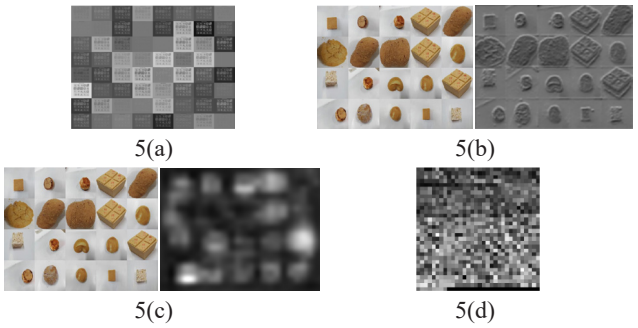

Figure 5. Variation of features within the RestNet50 structural layer

Figure 5 shows the feature variations within the RestNet50 structural layer, where Figure 5(a) is the feature map corresponding to Res2a-branch2c, Figure 5(b) shows the original input graph on the left, and on the right is the result of the first layer of the survey map corresponding to the Res2a-branch2c layer in which the points of the activation function are marked to present (result of the Investigate the activations in specific Res2a-branch2c channels with RestNet50), Figure 5(c) is the result presented by labeling the points with the strongest values in the feature map corresponding to the Res5c-branch2c layer, and Figure 5(d) is the result presented by the last Figure 5(d) is the final feature distribution of the Fc1000 fully connected layer after RestNet50 computation, and input this information to the feature data downscaling computation.

From Figure 2(c), Figure 3(c), Figure 4(c) to Figure 5(c), the feature maps obtained by the final convolutional operation in different CNN network models are shown. The feature maps of RestNet50 clearly highlight the image feature structures that are input and trained, so the final accuracy obtained by RestNet50 is the highest for the object recognition of dessert.

3.2 Main Feature Extraction Process

To reduce the size of the processed data, Gao [30] used the t-distributed stochastic neighbor embedding (t-SNE) in the CNN for hyperspectral images (HIS) applications. Its most prominent contribution is to extract high-quality features in different applications. According to Zhang, the DL algorithm is a powerful technology to learn and represent more meaningful features hierarchically [31]. It can easily project the HSI into low-dimensional spaces to find the natural structure of differently distributed data and can suppress the noise in images. The t-SNE is the popular manifold embedding type learning algorithm for nonlinear dimensionality reduction, but it inevitably consumes some crucial information. In a study by Hu, a new stacked structure of convolutional classifier was applied for denoising an environment, and a novel autoencoder mechanism was used to realize variations and similarities of the intra-class and filter the specific spatial information to improve efficiency in HSI classification application [32]. A multiscale feature fusion is another way to detect different spatial structures and extract textural features within the neighbor relationship [33]. However, the original image may contain noise. CNN does not only extract the required features to improve accuracy but it can also mess with noise information to drag the classification results. The dimension-reduced CNN [34] was proposed by combining the t-SNE method to reduce the data dimension and CNN to train the end-to-end feature.

Data features are probed from the CNN systems to be registered into the t-distributed stochastic neighbor embedding calculation. When X=[x1, x2, …, xN], ∈RCXN represents the vector denoting the matrix of the image data from the connection of the activation function=n layers of the pre-trained CNN model; in which N is the image vector's length, C is the vector size, and Y=[y1, y2, …, yN] is the output of converted vector X. The distribution of the conditional probabilities P and Q are illustrated below.

$$P\left(X_i \middle| Y_j\right) = \frac{S(X_i, X_j)}{\sum_{K \neq i}^{N} S(X_i, X_j)}, \tag{1}$$

$$Q\left(X_i \middle| Y_j\right) = \frac{S(Y_i, X_j)}{\sum_{k \neq i}^{N} S(Y_i, Y_j)}, \tag{2}$$

where:

S(·) is the selected Euclidean distance between two vectors of sample pixels.

The Kullback-Leibler (KL) divergence in (3) is to be approached as small as possible.

$$\sum_i \sum_j P(x_i, x_j) \log \frac{P(x_i, x_j)}{Q(y_i, y_j)}. \tag{3}$$

The t-SNE algorithm here is used to minimize the conditional probability between the original space and the embedded space by regulating the KL value. Since the pattern similarity from the multiple characteristics of data can be identified, t-SNE can find the local structure of the data. The t-SNE algorithm reconstructs hyperspectral data from the high-dimensional space to a low-dimensional space for catching the main feature. One of the advantages of t-SNE is its ability to eliminate the curse of dimensionality and crowding problems, but some parameters need to be defined first in advance.

Perplexity is considered as an effective size for the neighborhood sample in this t-SNE algorithm. Perplexity is normally set between 5 to 50. But, the bigger size dataset

always selects the larger perplexity value which is over 100. The lower perplexity divides the same category data into another group. The higher perplexity does not perfectly separate the different group data into the right category.

The principal component analysis (PCA) [35] is another way to extract discriminate features. It merges those with similar patterns and reduces the training data size. PCA is significantly able to map the original data feature X=[x1, x2,…,xN] ∈ RCXN into the small output vector Y=[y1,y2,… ,yN]. In the first step, it calculates the covariance of the original data sample using the following formula:

$$Cov = \frac{1}{C}\sum_{i}^{N}(x^{i})(x^{i}). \tag{4}$$

$$X^{(i)} = \left[x_1^i, x_2^i, ..., x_N^i \right]. \tag{5}$$

A singular value decom-position way is used to calculate the eigenvalues and eigenvectors of the covariance matrix through the following equation:

$$\left[U, S, V\right] = Svd(Cov), \tag{6}$$

where:

U and V represent two mutually orthogonal matrices.

U represents all eigenvectors of the computed covariance matrix

S represents a diagonal matrix consisting of eigenvalues.

A low-dimensional feature Z is then obtained based on the equation below.

$$Z = U \times X. \tag{7}$$

Z is regarded as the smaller size dataset by PCA algorithm to extract the main characteristics after the trained CNN module to reconstruct the primary feature.

3.3 Image Catalog Presentation through the K-means or KNN Algorithms

This article used a clustering algorithm to identify the output data distributions of the main feature extraction process which will then classify the different foods. The clustering algorithms divide the K groups from the training dataset to meet every data point containing the maximum similarity in the same clustering groups.

K-means algorithm is the most widely known method among clustering analysis algorithms. It has the advantages of simple operation, high efficiency for large data sets [36], and high scalability.

Assuming that c is the number of clusters, the K-means clustering method is performed through several steps to generate c clustering centers; then, the feature data belonging to the same cluster are classified into the same category and the food types are represented by the same cluster data. Since there are 10 types of desserts, we set the value of cluster C to 10. The training steps of K-means are as follows:

Step 1. Randomly select C initial points and regard them as the cluster centers.

Step 2. Calculate the distance between each data point x_i and the cluster center; then, find the nearest clustering center and add x_i to this group of clustering centers. If dataset is defined as $\{(xi, li)\}_{i=1}^{n}$, where xi are image point and li $\in\{1,2,\ldots,d\}$ are labels. Let C_j is jth cluster, and then, centroids are taken by minimizing the objection function with following step.

Step 3. Calculate the objective function

$$J(c_1, c_2, ..., c_d) = \sum_{j=1}^{d}\frac{1}{n}\sum_{i:x_{i\in C_j}}\left\|x_i - C_j\right\|^2 + \frac{1}{d}\sum_{j=1}^{d}G_j. \tag{8}$$

Where G is a Gini impurity index to isolate these clusters, which is detected by the following.

$$G_j = 1 - \sum_{j=1}^{d}\left(\frac{n_{k,j}}{n_j}\right)^2. \tag{9}$$

Where $n_{k,j} = \|\{x_t \in C_j : l_t = k\}$ and j=1,2,…,d.

If it remains unchanged, it means that the clustering results have converged into a stable state; thus, the clustering algorithm is finished.

Step 4. Calculate the average value of all data points in each same group as the new cluster center. Repeat the clustering process in Step 2 until all data do not need to be allocated to the belonging clustering group.

Step 5. The appropriate cluster centroid is categorized for each point, and the category of the cluster centroid with the highest relevance by its proximity is selected to identify the group to which each feature point belongs.

Step 6. Repeat to complete the categorization of all belonging image features and reconstruct the images. When using this method, it is necessary to master the number of clusters to facilitate data clustering and greatly save time for food category identification.

The advantage of K-means algorithm is that it does not need a large number of output nodes and it can greatly reduce the computational complexity.

Another algorithm is called KNN to assign labels based on the neighbors of each data point [37]. Firstly, we set the parameter k, this k is the number of neighbors to be considered when classifying. According to the definition of assignment, the category label that appears most times in the neighbors will be assigned to the data point. The classification algorithm of this KNN is described in the following steps.

Step 1. calculate the distance between all data points and the considered data points.

Step 2. The k data points with the lowest distance value are selected as the nearest neighbors.

Step 3. Retrieve the labels of all k neighbors.

Step 4. Among all the labels in step 3, the label that appears the most times is assigned to the data point under consideration.

4 Experiment Description and Result Analysis

In this study, NoteBook of Intel Core i7-10610U central processor with 16G RAM was used for image analysis.

The experimental sample data consisted of the ten most popular Kinmen desserts on the market, and the mobile phone camera function was used to collect images of the ten kinds of desserts. The data set contained a total of 1,000 Kinmen dessert images was used to expand the data set images four times. Finally, the data was divided into 10 categories. Each dessert used the autoAugment data to extend its amount, and the final dataset was divided into 70% training sets and 30% testing sets.

4.1 Experiment 1: AlexNet Model

In this experiment, we used the AlexNet model to construct the feature data set (1000 records) of the original food sample data. The PCA or t-SNE were used to extract the main features of the original feature data set. Figure 6 and Figure 7 show the dot plots of AlexNet model after the PCA or t-SNE data downscaling procedures, respectively. Finally, we used the K-means and KNN clustering algorithm to group the main feature data sets, classify the data of the same cluster into the same food, and finally reproduce the original image through the index value. Figure 8 and Figure 9 show the cake images recovered by t-SNE+K-means and PCA+K-means algorithms after searching the images according to their indices, respectively. The total number of tests was 5, and the average success rate of image recognition output t-SNE+K-means had an accuracy of 0.9013. The average success rate of PCA+K-means had an accuracy of 0.5171 and the average use time was 33.82 sec.

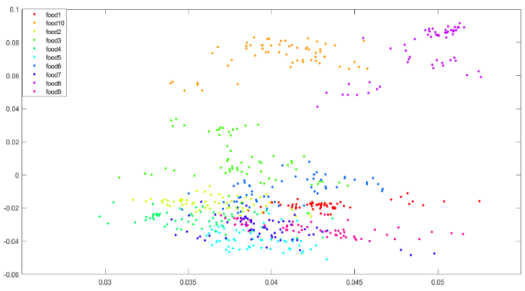

Figure 6. Distribution map of different food types in AlexNet model after PCA data capture

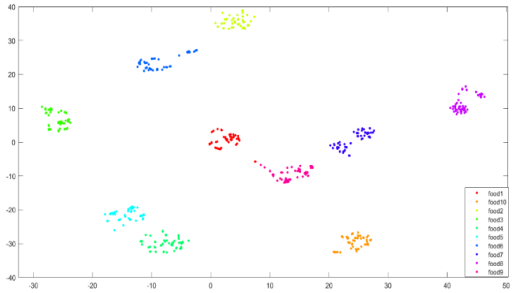

Figure 7. Distribution map of different food types in AlexNet model after t-SNE data capture

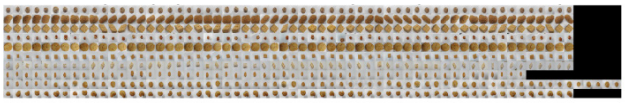

Figure 8. Diagram of AlexNet model restoration of different food types after t-SNE+K-means

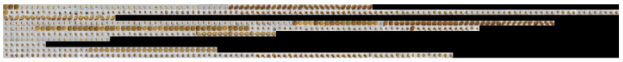

Figure 9. Diagram of AlexNet model restoration of different food types after PCA+K-means

4.2 Experiment 2: GoogleNet Model

This experiment used the GoogleNet Model to construct the feature data set (1000 records) of the original food sample data, and utilized t-SNE or PCA to extract the main features of the original feature data set. Figure 10 and Figure 11 show the dot plots of GoogleNet model after the PCA or t-SNE data downscaling procedures, respectively. Finally, K-means and KNN clustering algorithm was used to group the main feature data sets and classify the data of the same cluster into the same food. Finally, the original image was reproduced through the index value. Figure 12 and Figure 13 show the cake images recovered by t-SNE+K-means and PCA+K-means algorithms after searching the images according to their indices, respectively. The total number of tests was 5 and the average success rate of image recognition output t-SNE+K-means had an accuracy of 0.9135. Meanwhile, the average success rate of PCA+K-means had an accuracy of 0. 4123, and the average use time was 52.06 sec.

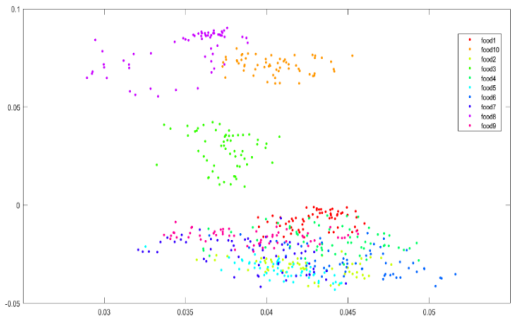

Figure 10. Distribution Map of Different Food Types after Grabbing PCA data with GoogleNet Model

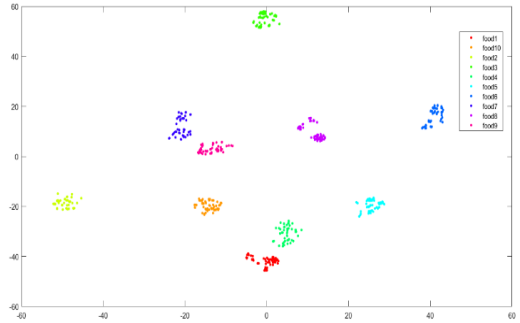

Figure 11. Distribution Map of Different Food Types after Grabbing t-SNE Data with GoogleNet Model

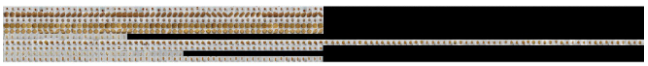

Figure 12. The restoration diagram of different food types of Googlenet model after t-SNE+K-means

Figure 13. The restoration diagram of different food types after PCA+K-means of GoogleNet model

4.3 Experiment 3: RestNet50 Model

This experiment used RestNet50 Model to construct the feature data set (1000 records) of the original food sample data. The t-SNE or PCA was employed to extract the main features of the original feature data set. Figure 14 and Figure 15 show the dot plots of RestNet50 model after the PCA or t-SNE data downscaling procedures, respectively. Finally, K-means and KNN clustering algorithm was used to group the main feature data sets and the data of the same cluster was classified into the same food. Figure 16 and Figure 17 show the cake images recovered by t-SNE+K-means and PCA+K-means algorithms after searching the images according to their indices, respectively. Finally, the original image was reproduced through the index value. The total number of tests was 5 and the average success rate of image recognition output t-SNE+K-means had an accuracy of 0.978. The average success rate of PCA+K-means had an accuracy of 0.5483 and the average use time was 99.58 sec.

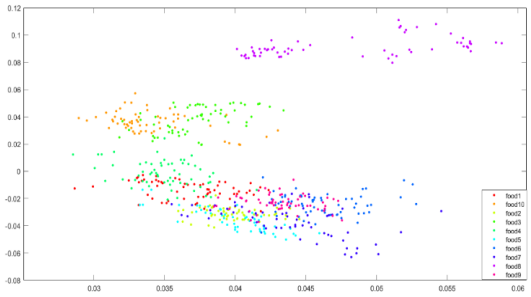

Figure 14. Distribution map of different food types of RestNet50 model after PCA data capture

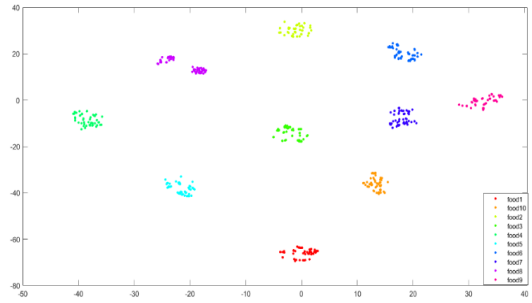

Figure 15. Distribution map of different food types in RestNet50 model after t-SNE data is captured

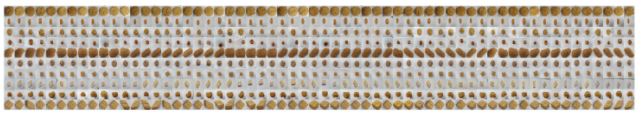

Figure 16. RestNet50 model restoration diagram of different food types after t-SNE+K-means

Figure 17. RestNet50 model restoration diagram of different food types after PCA+K-means

4.4 Experiment 4: MobileNet Model

In this experiment, we employed the MobileNet Model to construct the feature data set (1000 records) of the original food sample data and used the t-SNE and PCA to extract the main features of the original feature data set. Figure 18 and Figure 19 show the dot plots of MobileNet model after the PCA and t-SNE data downscaling procedures, respectively. Finally, we used the K-means and KNN clustering algorithm to group the main feature data sets and classify the data of the same cluster into the same food. Figure 20 and Figure 21 show the cake images recovered by t-SNE+K-means and PCA+K-means algorithms after searching the images according to their indices, respectively. Finally, we reproduced the original image through the index value. The total number of tests was 5 and the average successful identification rate of image recognition output t-SNE+K-means had an accuracy of 0.978. Meanwhile, the average successful identification rate of PCA+K-means had an accuracy of 0.5883 and the average computing time was 64.79 sec.

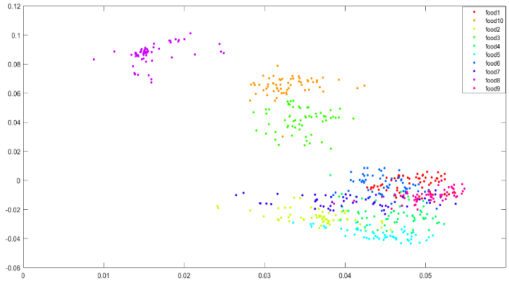

Figure 18. Distribution map of different food types in MobileNet model after PCA data capture

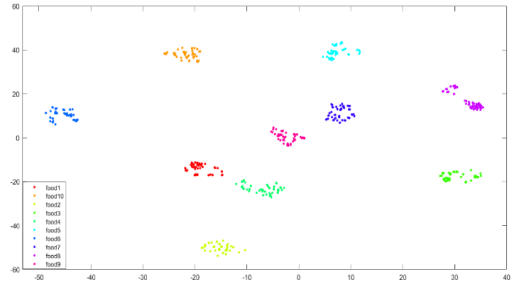

Figure 19. Distribution map of different food types in MobileNet model after t-SNE data capture

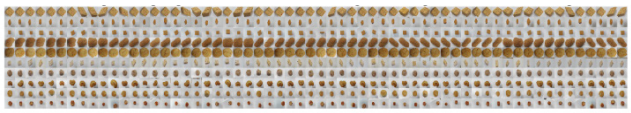

Figure 20. The reduction diagram of different food types of MobileNet model after t-SNE+K-means

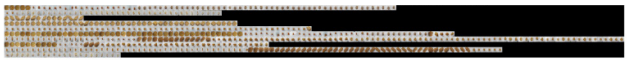

Figure 21. The reduction diagram of different food types of MobileNet model after PCA+K-means

4.5 Experiment 5: Performance Comparisons

The experiments showed in Table 1 that t-SNE with a nonlinear conditional probability and Gaussian distribution ability can perfectly identify the different small, medium, and big data spacing and separate them into a suitable group. For the cake image identification problem discussed in this paper, the original image data dimension has been significantly reduced by about 1,000 times by the feature capture procedure of CNN network model, and the t-SNE can distinguish the clusters of different cakes more clearly than the feature point map generated by PCA, so that the KNN and K-means algorithms can converge more quickly when building clusters. Therefore, although the time cost of t-SNE is slightly higher than that of PCA by about 3%, the accuracy of CNN+t-SNE+K-means can be significantly improved to more than 90%, which is much better than the 20% accuracy of CNN+PCA+K-means and about 75% accuracy of CNN+t-SNE+KNN. In terms of comparing different CNN models, RestNet50+ t-SNE+ k-means can achieve the best accuracy of 99.0% in the shortest time, while MobileNet can reach 97.8% and AlexNet and GoogleNet, about 90% accuracy. However, the time cost of RestNet50 > MobileNet > GoogleNet > AlexNet.

Table 1. Evaluation result in CNN type, Reduction and Clustering methods

CNN type	Reduction way	Clustering way	Accuracy	Cost time (s)
AlexNet	t-SNE	KNN	0.7870	29.63
AlexNet	PCA	KNN	0.2300	25.20
AlexNet	t-SNE	k-means	0.9013	33.82
AlexNet	PCA	k-means	0.5171	29.22
GoogleNet	t-SNE	KNN	0.7340	51.65
GoogleNet	PCA	KNN	0.2020	51.07
GoogleNet	t-SNE	k-means	0.9135	52.06
GoogleNet	PCA	k-means	04123	49.33
MobileNet	t-SNE	KNN	0.7530	63.95
MobileNet	PCA	KNN	0.3020	62.79
MobileNet	t-SNE	k-means	0.978	64.79
MobileNet	PCA	k-means	0.5883	60.43
RestNet50	t-SNE	KNN	0.7650	88.71
RestNet50	PCA	KNN	0.223	87.38
RestNet50	t-SNE	k-means	0.990	99.37
RestNet50	PCA	k-means	0.5483	93.22

All performance comparisons of t-SNE+K-means in the 4 types of models with the other methods are illustrated in Table 2. We used the RestNet50 model to capture the feature vectors of the Golden Gate cake and then used Linear regression classification (LRC) to complete the image classification experiment, and finally obtained 90% accuracy in 115.222307 seconds. We also use traditional RestNet50 model to make the We used the conventional RestNet50 model and set Mini Batch Size=30, Max Epochs=3, initial learn rate=0.0001 and optimizer using “sgdm”. Finally, 99% accuracy was achieved with a training time period of 683.960020 seconds.

Table 2. Evaluation result in different methods

Module ways	Accurancy	Computing time (seconds)
RestNet50 t-SNE+K-means	0.99	99.58
RestNet50 [8]	0.989	121.78
RestNet50+LRC [38]	0.9	115.222307
Traditional RestNet50	0.99	683.960020 seconds

This experiment showed that the proposed RestNet50 t-SNE+K-means approach both the best results 99% accuracy and the shortest training time within 99.58 seconds.

5 Conclusion and Future Works

This research constructed a hybrid-based CNN transfer learning machine with multiple main feature extraction, data clustering, and image reconstruction stages to appropriately extract the main feature of images, actually classify Kinmen desserts, and quickly reconstruct the image into the correct category within two minutes. The results showed that four types CNN structure with the connecting the suitable calculation layers (t-SNE+k-means) to efficiently and correctly classify the food category into ten types of Kinmen desserts.

In the small convolutional neural network trainer analysis, this AlexNet model only used a total of 22 sequential layers to realize the training dataset. GoogleNet module supported a total of 142 layers with 9 main blocks characteristic; each block contained 4 parallel layers. RestNet50 module used 174 layers which contained 16 double-layer networks. MobileNetv2 model contained 152 layers and its structure included 10 double block networks; however, the linking layer between the double block was too long. Based on the results, we found that RestNet50 module and MobileNet structures at the same t-SNE+ k-means sequential procedure for better main feature extraction and to approach the highest accuracy within the smaller cost time.

Future research will focus on simplifying the analysis by reducing the structure of the inner layers of each convolutional neural network and improving the performance of classification applications.

References

[1] H. Prasetyo, Winarno, C. H. Hsia, Intelligent SVD-based noise level estimation incorporating symbiotic organisms search, *Journal of Internet Technology*, Vol. 22, No. 1, pp. 61-69, January, 2021.

[2] H. Prasetyo, C. H. Hsia, J. Y. Deng, Multiple Secret Sharing with Simple Image Encryption, *Journal of*

Internet Technology, Vol. 21, No. 2, pp. 323-341, March, 2020.

[3] H. Prasetyo, C. H. Hsia, Improved multiple secret sharing using generalized chaotic image scrambling, *Multimedia Tools and Applications*, Vol. 78, No. 20, 29089-29120, October, 2019.

[4] G. Ciocca, G. Micali, P. Napoletano, State Recognition of Food Images Using Deep Features, *IEEE Access*, Vol. 8, pp. 32003-32017, February, 2020.

[5] A. M. Uddin, A. Al Miraj, M. Sen Sarma, A. Das, M. M. Gani, Traditional Bengali Food Classification Using Convolutional Neural Network, *2021 IEEE Region 10 Symposium (TENSYMP)*, Jeju, Korea, 2021, pp. 1-8.

[6] Q. L. Tran, G. H. Lam, Q. N. Le, T. H. Tran, T. H. Do, A Comparison of Several Approaches for Image Recognition used in Food Recommendation System, *2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, Purwokerto, Indonesia, 2021, pp. 284-289.

[7] R. D. Yogaswara, A. D. Wibawa, Comparison of Supervised Learning Image Classification Algorithms for Food and Non-Food Objects, *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, Surabaya, Indonesia, 2018, pp. 317-324.

[8] P. C. Patil, V. C. Burkapalli, Food Cuisine Classification by Convolutional Neural Network based Transfer Learning Approach, *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, Karnataka, India, 2021, pp. 1-5.

[9] S. N. Esfahani, V. Muthukumar, E. E. Regentova, K. Taghva, M. Trabia, Complex Food Recognition using Hyper-Spectral Imagery, *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2020, pp. 0662-0667.

[10] B. Arslan, S. Memiş, E. B. Sönmez, O. Z. Batur, Fine-Grained Food Classification Methods on the UEC FOOD-100 Database, *IEEE Transactions on Artificial Intelligence*, Vol. 3, No. 2, pp. 238-243, April, 2022.

[11] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, K. Murphy, Im2Calories: Towards an automated mobile vision food diary, *Proceedings of the IEEE international conference on computer vision (ICCV)*, Santiago, Chile, 2015, pp. 1233-1241.

[12] P. Napoletano, Visual descriptors for content-based retrieval of remote-sensing images, *International journal of remote sensing*, Vol. 39, No. 5, pp. 1343-1376, March, 2018.

[13] P. Napoletano, Hand-crafted vs learned descriptors for color texture classification, in: S. Bianco, R. Schettini, A. Trémeau, S. Tominaga (Eds.), *International Workshop on Computational Color Imaging*, Springer, Cham, 2017, pp. 259-271.

[14] A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, Columbus, OH, USA, 2014, pp. 806-813.

[15] O. L. Junior, D. Delgado, V. Gonçalves, U. Nunes, Trainable classifier-fusion schemes: An application to pedestrian detection, *2009 12th international IEEE conference on intelligent transportation systems*, St. Louis, MO, USA, 2009, pp. 1-6.

[16] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, San Jose, California, USA, 2010, pp. 270-279.

[17] G. M. Farinella, M. Moltisanti, S. Battiato, Classifying food images represented as bag of Textons, *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014, pp. 5212-5216.

[18] D. T. Nguyen, Z. Zong, P. O. Ogunbona, Y. Probst, W. Li, Food image classification using local appearance and global structural information, *Neurocomputing*, Vol. 140, pp. 242-251, September, 2014.

[19] K. Yanai, Y. Kawano, Food image recognition using deep convolutional network with pre-training and fine-tuning, *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Turin, Italy, 2015, pp. 1-6.

[20] J. C. Kwik, *Traditional food knowledge: Renewing culture and restoring health*, Master's thesis, University of Waterloo, Waterloo, Ontario, Canada, 2008.

[21] L. L. Pan, S. Pouyanfar, H. Chen, J. Qin, S. C. Chen, Deepfood: Automatic multi-class classification of food ingredients using deep learning, *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*, San Jose, CA, USA, 2017, pp. 181-189.

[22] J. Rajayogi, G. Manjunath, G. Shobha, Indian food image classification with transfer learning, *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, Bengaluru, India, 2019, Vol. 4, pp. 1-4.

[23] N. Hnoohom, S. Yuenyong, Thai fast food image classification using deep learning, *2018 International ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI-NCON)*, Chiang Rai, Thailand, 2018, pp. 116-119.

[24] D. J. Attokaren, I. G. Fernandes, A. Sriram, Y. S. Murthy, S. G. Koolagudi, Food classification from images using convolutional neural networks, *TENCON 2017-2017 IEEE Region 10 Conference*, Penang, Malaysia, 2017, pp. 2801-2806.

[25] M. T. Islam, B. N. K. Siddique, S. Rahman, T. Jabid, Food image classification with convolutional neural network, *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Bangkok, Thailand, 2018, vol. 3, pp. 257-262.

[26] H. Prasetyo, A. W. H. Prayuda, C. H. Hsia, M. A. Wisnu, Integrating Companding and Deep Learning on Bandwidth-Limited Image Transmission, *Journal of Internet Technology*, Vol. 23, No. 3, pp. 467-473, May, 2022.

[27] C.-H. Hsia, C.-H. Liu, New hierarchical finger-vein feature extraction method for iVehicles, *IEEE Sensors*

Journal, Vol. 22, No. 13, pp. 13612-13621, July, 2022.
[28] P. He, S. Ma, W. Li, Efficient Barrage Video Recommendation Algorithm Based on Convolutional and Recursive Neural Network, *Journal of Internet Technology*, Vol. 22, No. 6, pp. 1241-1251, November, 2021.
[29] L. Shi. S. Zhang, A. Arshad, Y. Hu, Y. He, Y. Yan, Thermo-physical properties prediction of carbon-based magnetic nanofluids based on an artificial neural network, *Renewable and Sustainable Energy Reviews*, Vol. 149, Article No. 111341, October, 2021.
[30] L. Gao, D. Gu, L. Zhuang, J. Ren, D. Yang, B. Zhang, Combining t-Distributed Stochastic Neighbor Embedding with Convolutional Neural Networks for Hyperspectral Image Classification, *IEEE Geoscience and Remote Sensing Letters*, Vol. 17, No. 8, pp. 1368-1372, August, 2020.
[31] L. Zhang, L. Zhang, B. Du, Deep learning for remote sensing data: A technical tutorial on the state of the Art, *IEEE Geoscience and remote sensing magazine*, Vol. 4, No. 2, pp. 22-40, Jun, 2016.
[32] W. Hu, Y. Huang, L. Wei, F. Zhang, H. Li, Deep convolutional neural networks for hyperspectral image classification, *Journal of Sensors*, Vol. 2015, Article No. 258619, July, 2015.
[33] Z. Li, L. Huang, D. Zhang, C. Liu, Y. Wang, X. Shi, A deep network based on multiscale spectral-spatial fusion for hyperspectral classification, in: W. Liu, F. Giunchiglia, B. Yang (Eds.), *International Conference on Knowledge Science, Engineering and Management*, Springer, Cham, 2018, pp. 283-290.
[34] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *Journal of machine learning research*, Vol. 9, No. 11, pp. 2579-2605, November, 2008.
[35] G. Yang, A. K. Gopalakrishnan, Network Traffic Threat Feature Recognition Based on a Convolutional Neural Network, *2019 11th International Conference on Knowledge and Smart Technology (KST)*, Phuket, Thailand, 2019, pp. 170-174.
[36] H. C. Chen, H. M. Feng, T. H. Lin, C. Y. Chen, Y. X. Zha, Adapt DB-PSO patterns clustering algorithms and its applications in image segmentation, *Multimedia Tools and Applications*, Vol. 75, No. 23, 15327-15339, December, 2016.
[37] H. I. Abdalla, A. A. Amer, Towards Highly-Efficient k-Nearest Neighbor Algorithm for Big Data Classification, *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*, Bandung, Indonesia, 2022, pp. 1-5.
[38] N. Aditama, R. Munir, Indonesian Street Food Calorie Estimation Using Mask R-CNN and Multiple Linear Regression, *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, Raipur, India, 2022, pp. 1-6.

Biographies

Hua-Ching Chen received the M.S. degree in CSIE from National Quemoy University, Taiwan, in 2010. He received Ph.D. degrees in EE from Xiamen University, China, in 2014. He is the lecturer with the School of Information Engineering, Xiamen Ocean Vocational College. His current research interests include wireless and Deep-learning system.

Hsuan-Ming Feng received M.S. and Ph.D. degrees in Computer Science and Information Engineering from Tamkang University, Taiwan, R.O.C., in 1994 and 2000, respectively. He is currently the full professor with the CSIE of National Quemoy University. His current research interests include wireless networks, machine learning, image processing and robot system.