Multi-optimized approach for classifying food images using faster RCNN

P Shekhar Reddy, Maseeha Naaz, V Naresh, G Rajesh, Prof. Ms. Shubhangi Mahule

Computer Science and Engineering,
ACE Engineering College,
Hyderabad, Telangana, India-501301

***Abstract*: For excellent health and nutrition, a balanced diet is necessary. It guards against a wide range of chronic noncommunicable diseases. Because health is such an important aspect of our lives, since people are now spending so much time and money on it, they are pursuing dietary programs to help them keep track of their caloric intake. It is impossible to be knowledgeable about all of the calorie values of food unless one is a calorie expert. The proposed method solves this problem by delivering the calorie value per 100gms of food, allowing the individual to ingest the appropriate amount. The suggested model selects a food image from a set of various food images in the data. Using faster RCNN, the image and name of the food, as well as the calorie value, will be reflected. In this proposed strategy, the accuracy of the pre-trained model has been increased.**

Keywords: RPN (Region Proposal Network), faster RCNN (Region-Based Convolutional Neural Network), CNN (Convolutional Neural Network), Deep Learning).

INTRODUCTION:

People are more mindful of their food and diet in today's world in order to avoid future or existing diseases or to maintain a healthy diet. People are fascinated by the digital era, therefore they rely solely on it, including for the provision of an application that automatically monitors an individual's food, which benefits them in a variety of ways. It also makes people more aware of their eating habits and diet. Over the last two decades, researchers have worked on automatically detecting food photos and their nutritional information from images taken using computer vision and machine learning techniques in order to ensure optimal dietary consumption. It is critical to accurately estimate the calorie value of food in order to appropriately measure dietary consumption. The majority of individuals nowadays are overweight and do not engage in sufficient physical activity. Given how busy and preoccupied individuals are these days, it's easy to forget or lose track of the food they consume on a daily basis, highlighting the need of correct food classification.

Current methods must be improved in order to improve accuracy and eliminate bias. A mobile cloud computing system, which uses smartphones to collect nutritional and calorie data, is one such potential solution. The next stage is to use the cloud's computational capacity to automatically analyze the dietary and calorie data for an objective evaluation. Users, on the other hand, must manually enter their data. In the field of visual-based dietary and calorie information analysis, a lot of research and development has been done during the previous few years. Furthermore, efficient information extraction from food photos and dietary attribute calculation remain a challenge. In this article, convolutional neural networks (CNNs) and regional proposal networks (RPNs) were used to identify food photos for future diet monitoring applications.

It is critical for an intelligent image classifier system to recognize images and classify them based on the learned input. Machine learning algorithms such as random forests, support vector machines, naive bayes, k means, and others can be used to train and test. Artificial neural networks, which imitate the neurons found in the human brain, are the most well-known and commonly used method in machine learning. Actually, these neurons are in 'n' numbers, which might be billions or trillions, as inspired by the nervous system of animals (of neurons). Dendrites connect each neuron to the rest of the nervous system. Dendrites assist transmit information from one neuron's axon to another. The artificial neural network, in a similar way, has a set of 'n' neurons that can be defined for a model.

In this paper, a quicker regional convolutional neural network with three convolutional layers is used. The main difference between a normal neural network and a convolutional neural network is that a convolutional layer connects one neuron from one layer to some other neurons in the next layer, whereas a normal neural network connects one neuron or more to all other neurons in the next layer, increasing the mathematical power of solving the equation. As a result, CNN is most commonly used in image classification models. The input is a food image that is delivered to the convolutional layer, which only accepts numerical data. As a result, RGB and black and white images must be transformed to pixel representations. These pixel values for a certain digit image are arranged in a matrix. This will now be forwarded to the convolutional layer.

For the probabilistic analysis of the output, the major topic is to reduce the matrix size from 380*280 to the smallest possible value. This is accomplished by the use of a convolution operation followed by an activation function. The size of the input matrix format image is decreased to the minimum achievable value after three convolutional procedures. This was then passed to the output layer's softmax layer (the final layer), where the softmax function analyzed the input image's probabilistic outcomes.

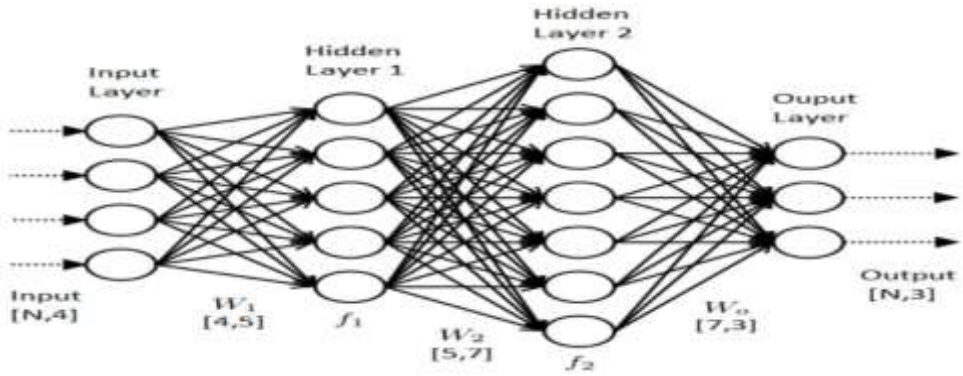


Fig. Artificial Neural Network

LITERATURE SURVEY:

Sl. No	Title of the Paper	Year, Author And Journal Name	Methodologies /Algorithms Used	Ideas for Adoption
1	Grab, Pay and Eat: Semantic Food Detection for Smart Restaurants	2018, Eduardo Aguilar, Beatriz Remeseiro, Marc Bolanos and Petia Radeva	Detection and Segmentation Algorithms	Downscaled the images to the same size for training and normalized.
2	A New Deep Learning-based Food Recognition System for Dietary Assessment on an Edge Computing Service Infrastructure	2017, Chang Liu, Yu Ca	Novel deep learning-based visual food recognition algorithms	Samples could be unrelated or wrong, leading to uncertain prior information
3	You Are What You Eat: Exploring Rich recipe information for Cross Region Food Analysis	2017, Weiqing Min, Bing-Kun Bao, Shuhuan Mei, Yaohui Zhu	Convolutional Neural Network	Do not consider label relation problem when there are three or more classes
4	Food Recognition: a new dataset, experiments and results	2016, Gianluigi Ciocca, Paolo Napoletano and Raimando Schettini	Segmentation algorithm JSEG	No better performance as a baseline model

RELATED WORKS:

Detection and categorization of images are the most common tasks of image classifier recognition systems. After the model has been successfully validated, picture detection can begin. Obesity appears to be a major issue in today's society. Obesity is caused by consuming more calories than we burn each day, which can have a negative impact on our quality of life. Pittsburgh Fast Food Picture Dataset (PFID) was created in 2009 using a food image and video dataset, and it contains 4545 still photographs of food from 101 different food items, such as "chicken fries" and "cheese pizza". The researcher used a Support Vector Machine (SVM) classifier on this dataset and got an overall classification accuracy of 11.2 percent using the colour histogram approach and 24.3 percent using the bag-of-SIFT (Scale-Invariant Feature Transform) features method. G. M. Farinella concentrated on this fundamental issue and devised a system for nutritional assessment that included item identification and quantity estimation. To portray food, they employed Gabor filters and colour characteristics.

To demonstrate that the novel procedure may successfully increase the performance of original SIFT and LBP feature descriptors, a multi-label SVM classifier method combined with a multi-class Ad boost algorithm is utilized. Around 50 categories of 100 food samples, such as soup, dumplings, and so on, were employed, with an accuracy of 68.3 percent. Y. Cao and Liu were inspired to develop and present a new prototype for food image categorization and dietary assessment using a smart phone and image processing and pattern recognition features. To understand the food photographs taken by the phone, the bag-of-words (BoW) model is employed. CNN is the most common and widely used deep learning architecture. Deep learning has significantly become a very effective tool for image identification techniques. On the ImageNet dataset, G. Ciocca used a deep convolutional neural network (DCNN) approach and got a 79 percent accuracy. For the identification and classification of food items, D. T. Nguyen used CNN on their own dataset. Traditional approaches such as support-vector-machine-based methods, which had an accuracy rate of 73.70 percent for recognition and 93.80 percent for detection, provided higher accuracy than CNN. On their own dataset of 574 food items, G. Ciocca, P. Napoletano, and R. Schettini created a 5-layer convolutional neural network to categorize food, with an overall accuracy rate of 84.8 percent.

The photos in this publication come from the Food-101 Dataset. Chen et al. introduced the PFID collection of food images, which is used to assess the accuracy of food image identification. The collection consists of 4,545 still photos classified into 101 categories of different food items using a typical computer vision approach. This is a dataset of food photos, each of which is labelled with a category of three instances. Images are taken in both a restaurant and a controlled lab setting for each meal item category. In a restaurant context, each meal type contains four still photographs, but in a laboratory setting, each food type has six still images. Food-101 introduces a challenging data set of 101 food categories, totaling over 101000 real-world photos. It contains a lot of

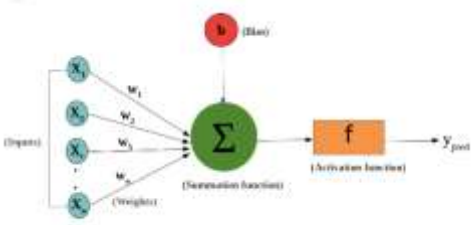
disunity, but when examined aesthetically, it seems as food classes, with 1000 photographs in each class, 250 of which are human reviewed test images and 750 of which are training images.

METHODOLOGY:

Food Image Detection and Classification:

The paper uses a quicker regional convolutional neural network structure for training and testing the model for detection and categorization of food photos. The network structure is comparable to that of the neurons found in a human brain. The technology is based on the artificial neuron network concept, which functions similarly to the human brain. Each neuron gets a specific set of inputs and passes them on to the next layer via a weighted sum to produce a bias function (sum of all input values * filter map weights). For training, testing, and deploying the model, photos related to food were obtained from a dataset called Food Images (FOOD-101).

Bias In Neural Network:

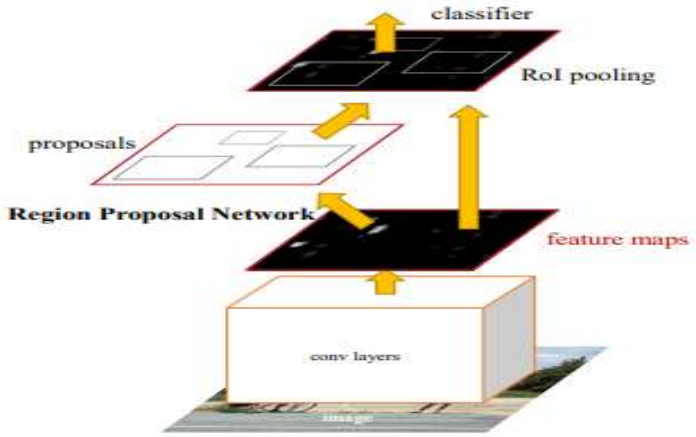


For processing and training, the neural network architecture model uses a set of input images. In neural networks, the bias function takes the set of input values from the food image's matrix format and assigns particular weights, with the sum of all weights and inputs being the bias function. The bias function is then used to create an activation function for a linear model of the graph's outputs.

=>Y= F(x) = xi * wi

=>Output=sum (weights * inputs) +Bias Function

The proposed model's architecture is as follows:



The system under the proposed architecture model is capable of changing the input food photos in terms of adjustment, size, form, and arrangement in order to train the model more effectively. Food photos are later converted to greyscale format for computer vision, which is referred to as the pre-processing stage of the image.

The greyscale image is then transformed to pixel values that range from 0 to 255.

The Regional Proposal Network is used to arrange pixel values based on the regions of interest. RPN concentrates on the food images' dense sections. A higher pixel value is assigned to a dense area, while a lower pixel value is assigned to a less dense area. A greyscale image's pixel values are represented as a matrix. The procedure was followed for all of the photographs that were sent in as input. The full set of pre-processed images is now fed into the convolutional layer of the network model, which is used to train and test the model for image output and classification.

A. DATASET

In terms of picture and voice recognition, the deep neural network has set a new bar. FOOD-101 dataset was used to train the system. Subsets of the entire food photos are included in the dataset in various amounts. The goal is to provide a more basic image analysis training set than CIFAR10 or GTSRB. To facilitate speedy trains and tests, the dataset contains significantly downscaled copies of the food images. For training and testing, the dataset is split into two sections.

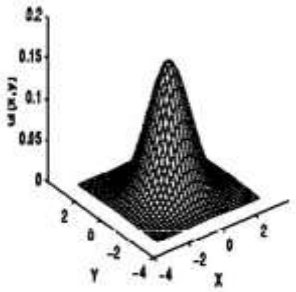
1. The training dataset contains 101 food image categories, each of which is represented by 1000 photos with a resolution of 384x384x3 (RGB,3 channels).
2. Similarly, the Testing dataset has 101 food picture categories, each of which is represented by 1000 photos with a resolution of 32x32x3 (RGB, 3 channels) (float32 from -1 to 1)

B. IMAGE PREPROCESSING:

The model's input image is supplied by the training set. The input image must be normalized and have pixel values ranging from 0 to 255 in order to train the neural network. By sharpening the edges and removing any further disturbances, such as high saturation, the input photos are altered and rearranged. The system uses the Gaussian Filter technique to filter the images, which helps reduce noise in the image's blur areas.

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}$$

A graphical representation of the 2D Gaussian distribution with mean(0,0) and σ = 1 is shown to the right.



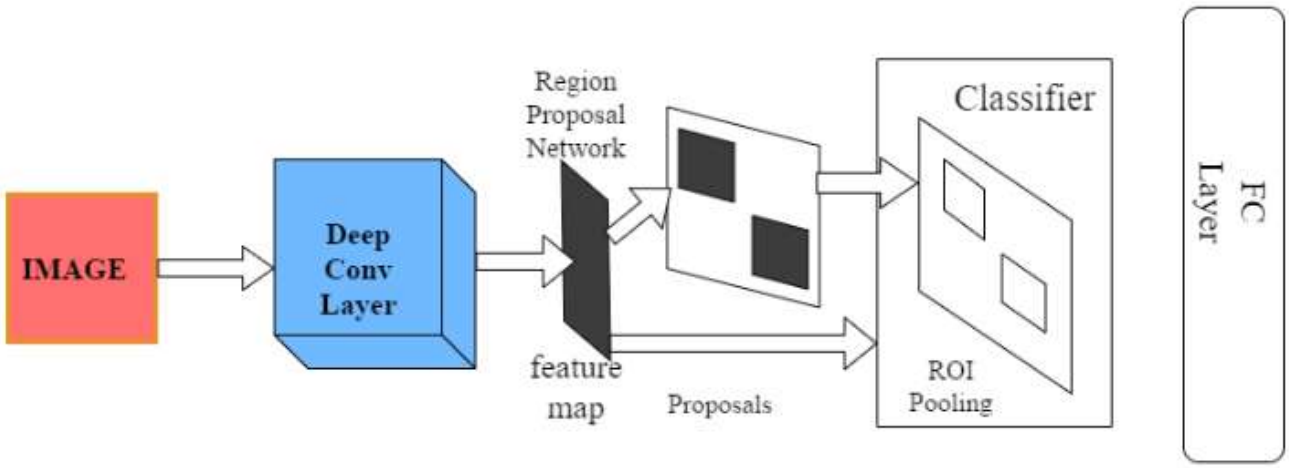
The grey world algorithm was used to remove any white balance that may have been present. The method normalizes the photos by converting them to a whiter balanced form. Simply said, it converts an RGB image to grayscale and minimizes noise and blur in the resulting grayscale image. The color, contrast, and saturation of the photos are also adjusted using the algorithm. In a matrix format ranging from 0 to 255, a greyscale image is assigned a set of pixel values. A matrix format is used to express the pixel values assigned to an image.

The outcome of the preprocessing can be displayed as

C. NETWORK STRUCTURE:

A quicker regional convolution neural network structure is used in the article. The network structure's convolution layer extracts features by selecting a region of interest from an input image. Also, by generating small squares on the input image of the matrix format, it extracts image features. Essentially, it is a mathematical procedure that accomplishes a task by taking into account two inputs: an image matrix and a filter or kernel.

RPN is responsible for assigning pixel values in the range of 0-255 to regions of interest in the given input food image.



=>Image matrix of dimension in terms of volume(h*w*d).
=>Kernal of filter fh *fs *fd.
=>Output (h-fh+1)*(w-fw+1)*1.

The mathematical act or merging of two functions to form a third function is referred to as convolution. It brings together two sets of data. The convolution process is done on the input data with the use of a filter map or kernel in the case of a CNN, which then produces a feature map that is passed to the next layer. The structure's next layers are the hidden layer and the output layer. The

study employs three convolution layers with a maximum pooling of 2*2. Max pooling is a technique for reducing the size of an image by taking into account the maximum number of pixel values in the grid. Max pooling reduces overfitting and gives the model a more general appearance. The output of the convolution layer was flattened before being transferred to the output layer, using a flatten layer. Finally, there is the softmax layer, which is in charge of the output's probability analysis.

D. FEATURE EXTRACTION AND TRAINING OF MODEL:

The model was trained using three convolution layers with a maximum pooling of 2*2. Max pooling is a technique for reducing the size of an image by taking into account the maximum number of pixel values in the grid. Max pooling reduces overfitting and gives the model a more general appearance. For better data accuracy, a flattened layer is connected between the two convolution layers. The completely connected layer was given a two-dimensional matrix layer as input, and the output was four-dimensional. The fully connected layers are moved to a softmax layer at the end of the process. The function of activation ELU is used to reduce the time it takes to merge the cost function to zero and produce more accurate test results. It differs from other activation functions in a number of ways. ELU is a positive number-defining constant. Except for negative inputs, ELU is quite similar to other activation functions such as RELU. For non-negative inputs, they both behave the same in identity function form.

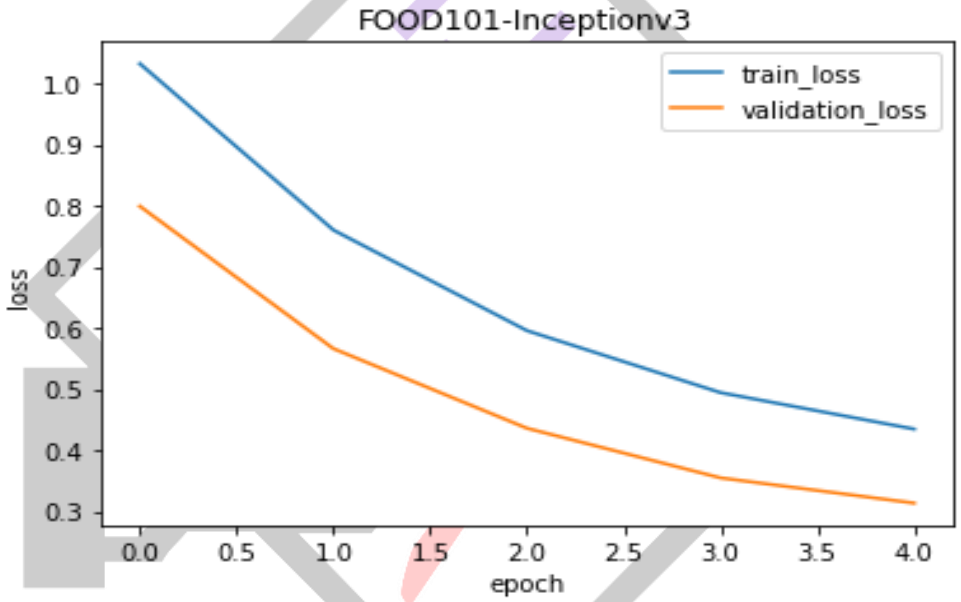
Expression in mathematics:

$R(z)=\{ z \; z>0$

$\alpha.(ez–1) \; z<=0\}$

Activation Function ELU:

Dropout, a type of regularization in which weights are kept standard with probablity values between 0 and 1, was used to improve the model's dependability and correctness. In the normalization stage of the neural network structure, the dropout function is utilized to avoid overfitting the data.

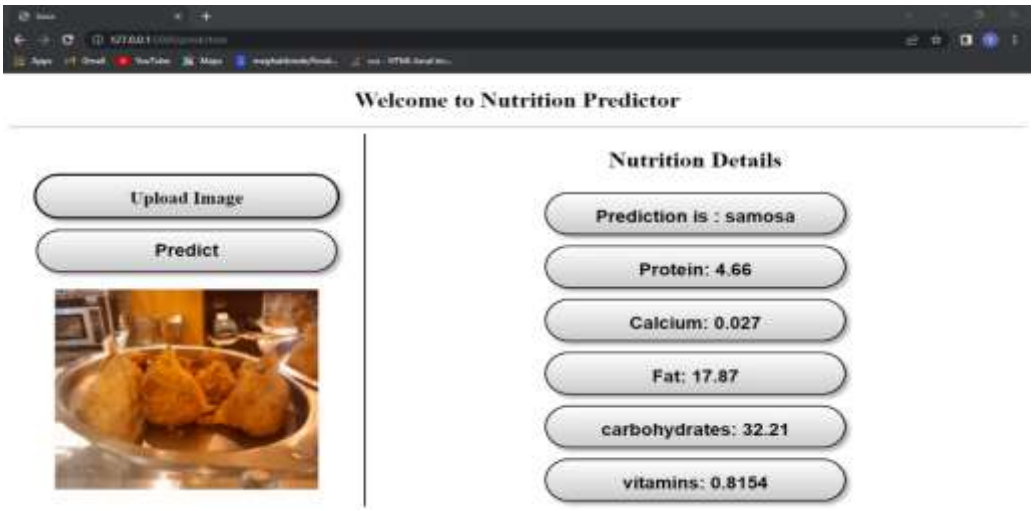


The complete network structure model, from the input layer to the output layer, is constructed using keras deep learning software in the backend and tensorflow as the frontend, with python as the neural network structure's base programming language. Keras is a deep learning software developed by Google that allows for rapid data prototyping.

Setting three convolutional layers, a max pooling layer, an activation elu layer, and finally attaching the output layer with a softmax function layer for probablity predictions of the output data completes the coding part (image). Pycharm was used to carry out the complete operation. The input tested image is searched over the food 101 database for calorie prediction, and the measured calorie value is returned from the database.

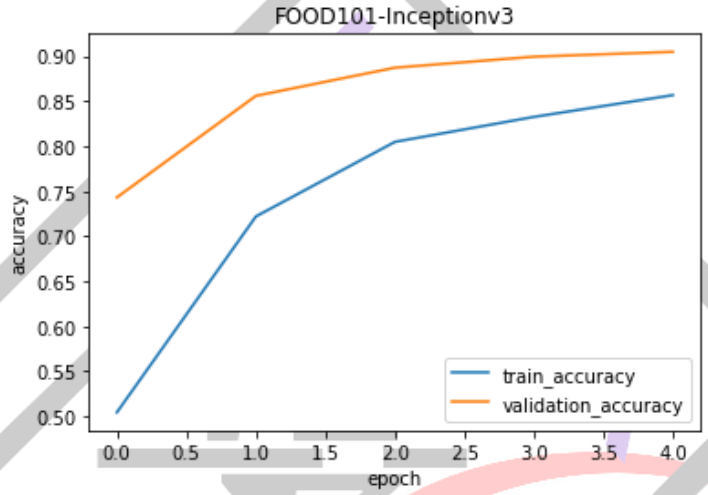
EXPERIMENTAL VERIFICATION:

Several changes were made to the data in order to train it. Since the photos for each class ranged from 2000 to 3000, the paper utilizes a batch size of 32. Precision accuracy was reported to be 96 percent after 30,000 iterations for an epoch of 52. For an epoch value of 90, the model can recognize the images with a 98 percent accuracy. The system may provide an accuracy rate of 92.1 percent when the epoch value is increased, although training time can take many hours or days, depending on the computer's performance.



RESULT:

Because the prediction accuracy was determined to be 92.9% after training the model over 50,000 iterations with an epoch of 100, which is a reasonable number for meal image recognition.



CONCLUSION:

Putting the emphasis on the need for individual healthcare. This research presents a system capable of recognizing meal images in the real world, which will aid in the development of dietary control software. The suggested model is a technique for recognizing food items in meal photographs of dining events. It's not easy to figure out how to recognize food items in a meal photograph automatically. As a result, we fully recognize that we will not be able to distinguish every food attribute. We are continuing our study to improve the system's accuracy and usability among individuals by refining and developing its capabilities.

ACKNOWLEDGEMENT:

We would like to thank our guide Ms Shubhangi Mahule and Soppari Kavitha for their continuous support and guidance. Due to their guidance, we completed our project successfully. Also, we are extremely grateful to Dr. M. V. VIJAYA SARADHI, Head of the Department of Computer Science and Engineering, Ace Engineering College for his support and invaluable time.

REFERENCES:

1. Q. Macedo, L. B. Marinho, and R. L. Santos, “Context-aware event recommendation in event-based social networks,” in Proceedings of the 9th ACM Conference on Recommender Systems, 2015, pp. 123–130.
2. D. T. Nguyen, Z. Zong, P. O. Ogunbona, Y. Probst, andW. Li, ``Food image classification using local appearance and global structural information," Neurocomputing, vol. 140, pp. 242-251, Sep. 2014.
3. W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, ``You are what you eat: Exploring rich recipe information for cross-region food analysis," IEEE Trans. Multimedia, vol. 20, no. 4, pp. 950-964, Apr. 2018.
4. N. Martinel, C. Piciarelli, C. Micheloni, and G. L. Foresti, ``A structured committee for food recognition," in Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW), Dec. 2015, pp. 92-100.
5. P. Yin, Q. He, X. Liu, and W.-C. Lee, “It takes two to tango: Exploring social tie development with both online and offline interactions,” Statistical Analysis and Data Mining: The ASA Data Science Journal, 2015.
6. Deep Food: Food Image Analysis and Dietary Assessment via Deep Model , LANDU JIANG (Member, IEEE), BOJIA QIU , XUE LIU , (Fellow, IEEE), CHENXI HUANG, AND KUNHUI LIN 1School of Informatics, Xiamen University, Xiamen 361005, China.

7. Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, ``Analysis of food images: Features and classification," in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2014, pp. 2744-2748.
8. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, M. Yunsheng, S. Chen, and P. Hou, ``A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure," IEEE Trans. Serv. Comput., vol. 11, no. 2, pp. 249- 261, Mar. 2018.
9. G. Ciocca, P. Napoletano, and R. Schettini, ``IVLFood-WS: Recognizing food in the wild using deep learning," in Proc. IEEE 8th Int. Conf. Consum. Electron.-Berlin (ICCEBerlin), Sep. 2018, pp. 1-6.
10. Z. Qiao, P. Zhang, Y. Cao, C. Zhou, L. Guo, and B. Fang, "Combining heterogenous social and geographical information for event recommendation," in Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
11. W. Zhang, J. Wang, and W. Feng, "Combining latent factor model with location features for event-based group recommendation," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 910–918.
12. G. M. Farinella, M. Moltisanti, and S. Battiato, ``Classifying food images represented as bag of Textons," in Proc. IEEE Int. Conf. Image Pro- cess. (ICIP), Oct. 2014, pp. 5212-5216.
13. G. Ciocca, P. Napoletano, and R. Schettini, ``Food recognition: A new dataset, experiments, and results," IEEE J. Biomed. Health Inform., vol. 21, no. 3, pp. 588-598, May 2017.
14. M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, ``A food recognition system for diabetic patients based on an optimized bag-of-features model," IEEE J. Biomed. Health Inform., vol. 18, no. 4, pp. 1261-1271, Jul. 2014.
15. R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and B. Guo, "Predicting activity attendance in event-based social networks: Content, context and social influence," in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2014, pp. 425–434.