Journal of King Saud University – Computer and Information Sciences 34 (2022) 1775–1784

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Image-based soft drink type classification and dietary assessment system using deep convolutional neural network with transfer learning

Rubaiya Hafiz [a,*], Mohammad Reduanul Haque [a], Aniruddha Rakshit [a], Mohammad Shorif Uddin [b]

[a] Dept. of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh
[b] Dept. of Computer Science & Engineering, Jahangirnagar University, Savar, Bangladesh

ARTICLE INFO

Article history:
Received 25 February 2020
Revised 4 August 2020
Accepted 29 August 2020
Available online 9 September 2020

Keywords:
Drinks classification
Noise reduction and contrast enhancement
Mean-shift segmentation
Bag-of-feature
Deep CNN models

ABSTRACT

Nowadays, people are taking soft drinks (carbonated nonalcoholic beverages) at an increasing rate. Health experts around the world have cautioned from time to time that these drinks lead to weight gain, raise the risk of non-communicable diseases, and so on. To develop consciousness among people, the present work describes an image-based tool to self-monitor the nutritional information of soft drinks by using a deep convolutional neural network (CNN) along with transfer learning. At first, a preprocessing function is done through noise reduction and contrast enhancement. Then the location of the drinks region is extracted through visual saliency and mean-shift segmentation technique. After removing backgrounds and segment out only the region of interest from the image a deep CNN-based transfer learning model is employed for the drink classification. Finally, the size of each drink bottle is estimated using the bag-of-feature (BoF) and distance ratio calculation to find the nutrition value from the nutrition fact table. To perform experimentation a dataset is built containing ten most consumed soft drinks in Bangladesh using images from the ImageNet dataset, internet sources and also self-capturing. The experiment confirms that our system can detect and recognize different types of drinks with an accuracy of 98.51%.

1. Introduction

High consumption of sugar-sweetened carbonated nonalcoholic beverages (SSB) such as soda, fruit and energy drinks containing the increased presence of sugar (glucose, fructose, sucrose), sweeteners, and some other additives is a common phenomenon among people. It is found that SSB can increase the risk of type 2 diabetes, heart disease, dental decay, bone loss, stomach and kidney problems, etc., (Schulze et al., 2004). Furthermore, higher consumption of sugary beverages has been linked with an increased risk of premature death (Malik et al., 2019), obesity and overweight related health problems.

The consumption of different types of beverages has rapidly increased in many parts of the world, especially in low- and middle-income countries, contributing to rising rates of many health-related problems. In spite of the fact that the situation is quite alarming, however, people are becoming more sensible at consuming foods as their weight and health are instantaneously impacted by the amount and type of foods and drinks they consume (Cawley et al., 2019). It is found that appropriate food and diet patterns work as a guard against heart disease, stroke, diabetes, and other complications. Consequently, a self-regulating system that can manifest the nutritional variety of different drinks can guide someone to determine a particular drink along with its quantity is beneficial for his/her well-being or not.

Despite having a large number of applications in real life, computer systems applied to the classification of SSB have not been widely studied. The widespread use of mobile devices such as digital cameras and smartphones can now be considered as data collection tools for dietitians (Martin et al., 2008). With the benefit of image processing techniques, some researchers proposed vision-based approaches to estimate the calories taken by the users. He et al. introduced k-nearest neighbors and vocabulary tree-based technique for food image analysis and classification (He et al.,

* Corresponding author.
E-mail addresses: rubaiya.cse@diu.edu.bd (R. Hafiz), reduan.cse@diu.edu.bd (M.R. Haque), aniruddha.cse@diu.edu.bd (A. Rakshit), shorifuddin@gmail.com (M.S. Uddin).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2020.08.015
1319-1578/

2014). They worked with 42 unique categories and improved the classification performance by 2%. For the recognition of food items, Yang et al. (2010) proposed a model that creates a feature vector in the discriminative classifier using a multi-dimensional histogram. They used 61 PFID (Pittsburgh Food Image Dataset) food categories as their dataset and organized these into 7 major groups. They achieved an accuracy of nearly 80%. Bosch et al. (2011) and Zhu et al. (2011) used both local and global features for the classification of different food images. Previously, we introduced a machine learning-based system that can detect and classify several types of drinks automatically from images (Hafiz et al., 2016). But that methodology cannot be able to show the nutrition value of each drink according to the size and actual quantity.

To reduce the problems discussed above, many researchers have been trying to develop a model that can estimate the calorie intake by an individual in an automatic or semi-automatic way. Recently, it is observed that deep convolutional neural network (DCNN) has shown remarkable progress in solving diverse image recognition and classification tasks (Rawat and Wang, 2017; Krizhevsky et al., 2012; Kölsch et al., 2017). Kagaya et al. applied CNN through parameter optimization for the detection and recognition of foods (Kagaya et al., 2014). When the baseline method was applied, they found an accuracy of almost 90% whereas for CNN it was nearly 94%. Later Kagaya and Aizawa proposed another model to classify food/non-food images using CNN with an accuracy of more than 95% (Kagaya and Aizawa, 2015). Mezgec et al. defined a new DCNN architecture called 'NutriNet' (Mezgec and Koroušić Seljak, 2017) and applied it on a dataset that contains 520 different types of food and drink images. They achieved a classification accuracy of around 87%. An improved methodology was proposed by Mezgec et al. where they have combined deep learning and natural language processing for the recognition of fake-food images (Mezgec et al., 2019). The deep learning model provides an accuracy of 92.18% on fake-food images.

Deep learning is well suited for different medical diagnosis problems, such as bone fracture classification (Tanzi et al., 2020), robot-assisted surgery (Gribaudo et al., 2019), radiotherapy (Shan et al., 2020), psychophysiological chronic disorders (Rubasinghe and Meedeniya, 2020) etc., as it can identify patterns in sparse and noisy data. As the diet-related information searching on the internet is rapidly increasing, so several applications are proposed to control overweight and obesity for both adults and adolescents (Hutchesson et al., 2015; Jensen et al., 2016). Moreover, machine learning and image processing-based apps (Spanakis et al., 2017) can help to warn people before a possible unhealthy eating event (Spanakis et al., 2017) and also monitor the physical activity in children to reduce overweight and obesity (Fergus et al., 2015). However, no research is done on drink-image classification and nutrition value assessment to warn people before consume.

Here, an automated system is developed based on deep CNN along with transfer learning for the recognition and classification of different types of drinks and assess nutrition values. The main contributions of this paper are as follows:

a. A neural network-based deep CNN with transfer learning is used for the recognition of different types of drinks.
b. A method for the estimation of nutrition value based on bottle size and contents by employing SURF (speeded up robust features) and color based bag-of-features along with bottle length and cap ratio.
c. Development of a dataset containing ten most consumed soft drinks in Bangladesh with different bottle sizes.

The forthcoming parts of this paper are fabricated as follows: Section 2 gives a general description of our suggested scheme including the pre-processing of the drink image, region of interest segmentation. It also includes the framework of learning and recognition of drinks followed by a nutrition value assessment process. Section 3 describes our developed dataset. The overall experimental results are presented in Section 4. Classification error analysis is discussed in Section 5. Eventually, the conclusion is drawn in Section 6.

2. Methodology

The overall approach of our proposed system is shown in Fig. 1. In addition to recognition and classification processes, we also have paid attention to precisely extract the region of interest (drink object) contained in each image. Since most of the images of our dataset contain a cluttered scene, we employed lots of pre-processing as shown in Fig. 2 to determine the location of our object of interest. A sample image passing through all steps is showed in Fig. 3. All the stages for the extraction of the drink region from an original image are described below.

2.1. Pre-processing for the detection of object of interest

2.1.1. Salient area detection

The graph-based method helps us to segment out our desired object from the images with a cluttered scene (Harel et al., 2007). Here, at first, the dissimilarity, denoted by d, of two location points, $M(i,j)$ and $M(x,y)$, is computed using Eq. (1).

$$d((i,j)||(x,y)) \triangleq \left|\log\frac{M(i,j)}{M(x,y)}\right| \tag{1}$$

Now, a fully connected directed graph is generated by joining each node with all other nodes. Each node are labeled with two indices $(x,y) \in [n]^2$ and a weight w_1 is assigned with each edge from node (i,j) to node (x,y) using Eqs. (2) and (3).

$$w_1((i,j),(x,y)) \triangleq d((i,j)||(x,y)).F(i-x,j-y) \tag{2}$$

Where

$$F(a,b) = exp\left(-\frac{a^2+b^2}{2\sigma^2}\right) \tag{3}$$

Here, σ is a free parameter and regions are connected via F. Finally, normalization is performed, and areas that contain 60% saliency are identified. Regions that have less than 60% saliency are considered as the background and these are removed. The resultant foreground image has used as the input of our next step.

2.1.2. Mean shift segmentation

Visual saliency can only detect the most salient area of the image rather than the complete salient object. For this reason, we have used a mean shift segmentation technique to divide the image into several regions in such a way that each region contains a set of pixels with the same characteristics such as color, texture, etc. (Suji et al., 2013). At first, the original image is altered to create a mean shift vector, $\overrightarrow{Mh}$ using a pixel P_a and a search window of radius h_s (Huang et al., 2013). This process is continued until $|\overrightarrow{Mh}|^2$ is less than a threshold ε. For any color image, the value of $\overrightarrow{Mh}$ can be 5, i.e., change of x coordinate ($\triangle x$), change of y coordinate ($\triangle y$), and the changes of intensities in RGB components ($\triangle R, \triangle G, \triangle B$).

If S denotes a window centered at P_a and s denotes a pixel within the window then $\triangle x$ is calculated using Eqs. (4) and (5).

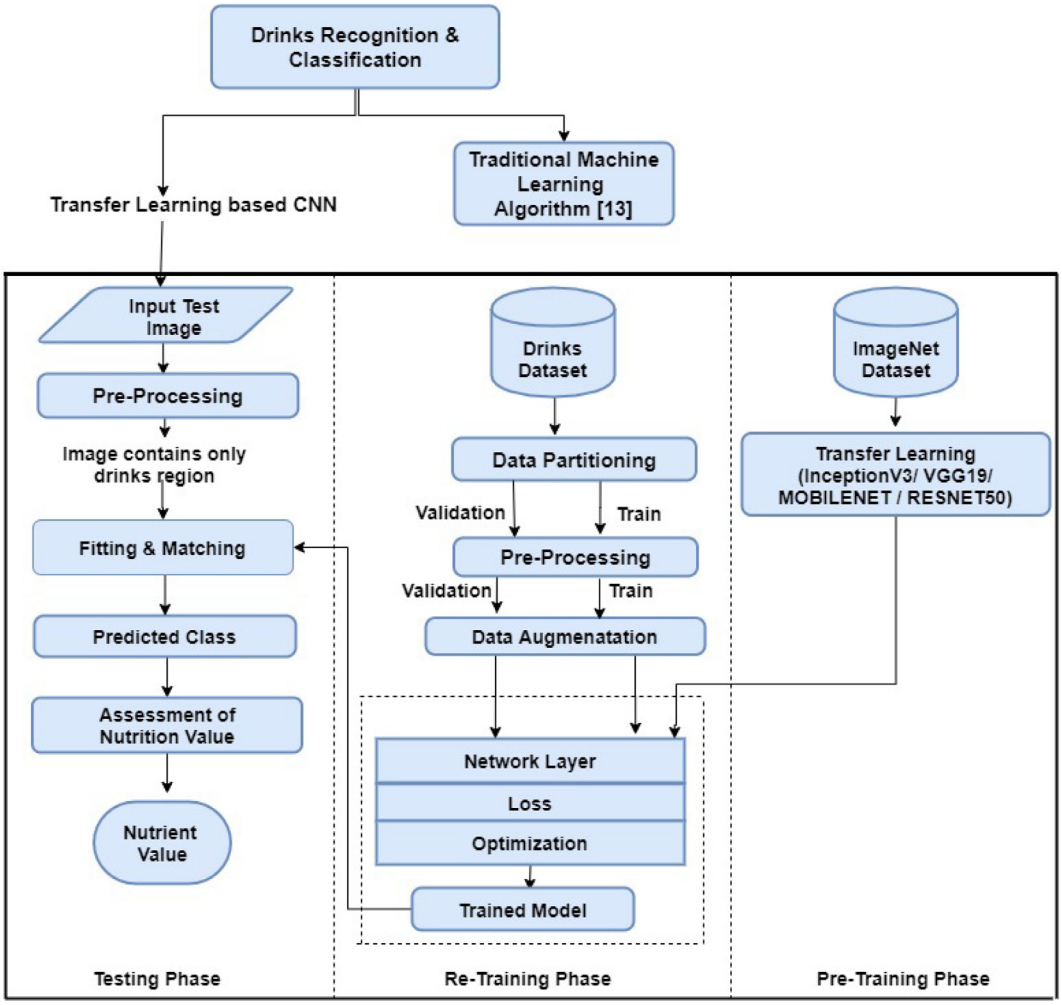

Fig. 1. Schematic diagram of the proposed deep CNN-based system for the drink type classification and dietary assessment.

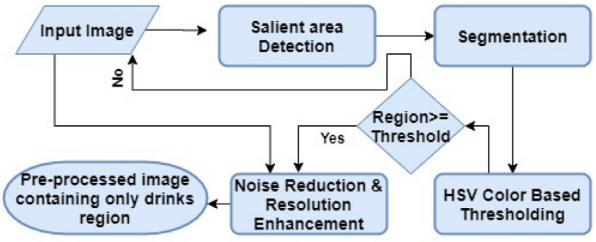

Fig. 2. Flow diagram of the pre-processing section.

Here, I_s, I_a are the intensities and X_s, W_s and h_r denote the x coordinate, the weight of pixel P_s and range bandwidth, respectively.

$$\triangle X = \sum_{s \in S} K\left(\frac{I_s - I_a}{h_r}\right) W_s . X_s \tag{4}$$

Where

$$K(x) = \begin{cases} 1, & if \ ||x|| < 1 \\ 0, & if \ ||x|| \geqslant 1 \end{cases} \tag{5}$$

Also, the new window centre P_b is calculated using Eq. (6) until window center no longer shifts.

$$\overrightarrow{P_b} \longleftarrow \overrightarrow{P_a} + \overrightarrow{Mh} \tag{6}$$

Lastly, three sub-steps (i.e. connecting regions, imposing transitive closure, and pruning spurious regions) are performed sequentially. As a result, the object of interest itself creates a separate region along with all other entities.

2.1.3. Color-based thresholding

Thresholding is a prominent way to decompose a given image into sub-regions like foreground and background based on the color (Kadouf and Mustafah, 2013). This method helps us to extract only the region of drinks from the image. We impose HSV color based thresholding because our segmented image may contain our desired object as well as a little portion of other objects too.

The combination of visual saliency, mean shift segmentation, and thresholding process enables our system to segment out the drink from the images. In some cases, the drink region is either partially visible or not clear (which is discussed in Section 5). Hence, our system may give us a pure black region as output after these steps which makes our classification accuracy lower. For this reason, after thresholding, we count the number of non-black pixels and if it is lower than a certain threshold value then we fed the raw image to the next step rather than the resultant one. We also improve the quality of the image by noise removal and resolution enhancement.

2.1.4. Noise reduction

Noise suppression is a very crucial factor since the performance of the classification technique depends on the quality of the image.

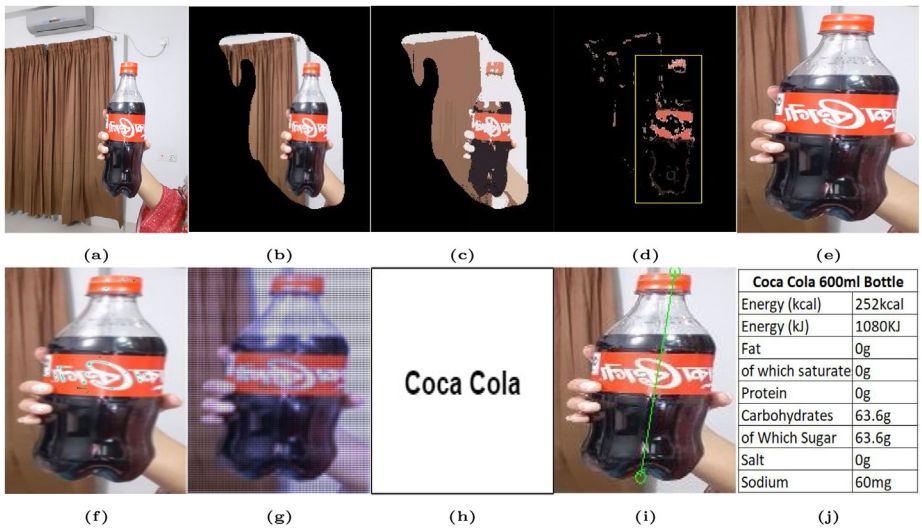

Fig. 3. (a) Input test image, (b) Image after employing saliency detection. The region containing more than 60% saliency is visible here and the remaining portion is removed as background, (c) Resultant image after the mean-shift segmentation, (d) Image after color based thresholding, (e) Resultant image of drink after removing the background by cropping, (f) Image after performing Gaussian noise reduction, (g) Resultant image after executing discrete wavelength transform-based resolution enhancement. After performing all these above-mentioned steps the resultant image is sent for fitting and matching along with the trained model. (h) This image is classified as 'Coca Cola', (i) To find the actual size of the drinks bottle from the image, we calculate the ratio from the length of the bottle and cap. Here we found the ratio as 7.3529. So we can say that it is a 600 ml bottle (discussed later in Section 4), (j) We have collected nutrition value for each drink with all available serving bottle size. Once we found that it is a 600 ml Coca Cola bottle, from the previously collected data, our method shows its nutrition value as the final result.

Here a fuzzy filter is imposed for reducing Gaussian noise while keeping the other features intact (Rahman et al., 2014). If the input image is denoted by f_p and f_{max} denotes the maximum intensity value among 8-neighboring pixels, then a function is calculated using Eq. (7).

$$F_p = \begin{cases} 1, & f_p = 0 \text{ or } 255 \\ exp\left(-\frac{(f_p - f_{max})^2}{2\times 8\times \sigma}\right), & otherwise \end{cases} \quad (7)$$

Where σ is the standard deviation of all intensity values. This filter is applied separately for each of the R, G, and B components of each color image. After filtering operation,we found our desired image by concatenating the three color components.

2.2. Resolution enhancement

The resolution of an image plays a crucial role in many image processing applications, especially in the feature extraction technique (Khaire and Shelkikar, 2013). To get an enhanced image, discrete wavelet transform (DWT) is used to disintegrate the input image into several sub-bands (Karunakar et al., 2013). After that, interpolation of the high-frequency sub-band images and the low-resolution input image is performed. Finally, combining of all these images is used to generate a resultant image using inverse DWT.

2.3. Data augmentation

To improve classification accuracy and reduce the overfitting problem we impose data augmentation on the training dataset. Random rotations, shifts, flips, and cropping techniques are applied while augmenting data for each category.

2.4. Fitting and matching

After finishing all of the previously discussed steps, the resultant training images are sent for retraining using transfer learning. All test images are also pre-processed separately and then fed to the system for prediction. In the case of transfer learning, knowledge from the previously trained similar model is transferred and leveraged to solve new problems. Because it is quite effortless and much faster to fine-tune a network than do the training from the scratch. The early layers of CNN contain generic features that can be reused while the final layers are more application-specific. Because of this property, the initial layers are well-preserved while the endmost ones are fine-tuned to train with the current dataset of interest (Albawi et al., 2017; Yosinski et al., 2014; Nogueira et al., 2017).

To attain high recognition accuracy, deep learning models need huge labeled data for training the classifier. However, it is quite challenging to get such a dataset for every domain since most of the deep learning models are extremely specialized to a distinct domain or even a specific task. For this reason, it is a suitable alternative to retrain a CNN (which is previously trained by a generalized label dataset) with only scarce data, initiated by Thrun (1996).

A study has conducted by Yosinski et al. (2014) to fine-tune the CNN based transfer learning model which was pre-trained on the ImageNet dataset. To transfer information from labeled data to unlabeled data, Tian et al. (2012) proposed a model for sparse transfer learning with a view to re-rank the video search. For medical image analysis, Tajbakhsh et al. (2016) investigated the performance of fully trained CNNs with pre-trained CNNs. The transfer learning technique also successfully utilized in video-based emotion recognition (Kaya et al., 2017), iris recognition (Nguyen et al., 2018), end-to-end airplane detection (Chen et al., 2018), and poverty mapping (Xie et al., 2016).

We retrain our drink dataset by popular 4 well-known deep CNNs: VGG19, InceptionV3, MobileNet, and ResNet50. A brief architecture of these deep CNNs is given in Table 1.

Once we recognize the drinks family, we can effortlessly extract its composition from our nutrient fact table.

2.5. Nutrition value assessment

After classification of any soft drink, we again impose color and SURF (Bay et al., 2006) based bag-of-feature (BoF) (O'Hara and Draper, 2011) technique to identify how it is served, i.e. in a glass, can, glass bottle or plastic bottle and the process is shown in Fig. 4.

Table 1
A brief architecture of 4 well-known deep CNNs investigated in this work.

	InceptionV3	VGG19	MobileNet	Resnet50
Input Size	227 × 227	224 × 224	224 × 224	224 × 224
Conv. Layers	21	19	28	34
Filter Size	1,3,5,7	3	1,3	1,3
Stride	1,2	1,2	2	2
Parameter	23 M	155 M	5855942	25.6 M
FC Layer	1	1	4	1
Size	92 MB	549 MB	17 MB	–
Depth	159	26	88	–

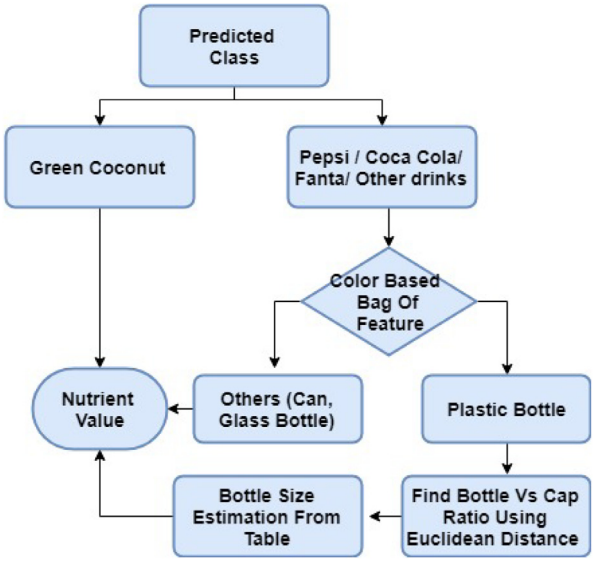

Fig. 4. Process of dietary assessment (nutrition value estimation) from image.

Since only SURF method extract features from grayscale images, we use color-based BoF to classify drinks according to their color information. As Coca Cola (or other soft drinks) can and glass bottle have a fixed size we can show their nutrition value easily. In case of a plastic bottle, for the proper calculation of nutrition value, it is necessary to find the actual size of the bottles from the image. We consider 5 different types (i.e. 250 ml, 400 ml, 600 ml, 1.25 liter, 2.25 liter) of Coca Cola bottles here. From the pre-processed image, we can easily get the top and bottom pixel value for both the cap as well as the whole bottle itself. We impose the Euclidean distance method to find the height of both cap and the full bottle from the image. The distance between two points in the image plane (Danielsson, 1980) with coordinates (x, y) and (a, b) is calculated by Eq. (8):

$$dist((x,y),(a,b)) = \sqrt{\left((x-a)^2 + (y-b)^2\right.} \tag{8}$$

Then the ratio is calculated by dividing the total length of the full bottle and the length of the cap and the equation is as follows:

$$ratio = \left(\frac{Length\ of\ full\ bottle}{Length\ of\ cap}\right) \tag{9}$$

3. Experimental dataset

Our experimental dataset consists of ten distinct categories of drinks and beverages. We have chosen these because they are the most common and available everywhere in Bangladesh. 1250 images of each category (total 1250 × 10 = 12,500 images) are used to construct our experimental dataset. Most of the images are collected from a hierarchical large scale database called ImageNet (Deng et al., 2009). Rests of the images for each category are collected from several internet sources as well as some are captured by us. The dataset is imparted into two subsets: 80% images (1000 for each category) are considered as a training set and the remaining 20% (250 for each category) as the testing set. In doing the experimentation, we used a 5-fold cross-validation with each category. Fig. 5 demonstrates some sample images from our training and testing dataset. Our developed dataset are available at: **https://github.com/reduan/Drink-Dataset.git**.

Since we have considered color and shape features for the drink identification from an image, the results may differ if the bottle shape and size is different from the trained dataset. If the manufacturer changes the shape and texture of any soft drink type then our system may not be able to classify it precisely.

4. Experimental results

The following section presents and discusses the experimental result of our system. For experimentation, at first, we partitioned our dataset according to the difficulty of classification into three parts, i. e. easy(images in which drinks are clearly visible and easily identifiable), medium(images containing a small amount of cluttered), hard (images containing the cluttered scene, smashed as well as top and/or partial view of drinks). The classification performance of each part is shown in Table 2. Fig. 7 shows the comparative performance of four popular transfer learning methods i.e. InceptionV3, VGG19, MobileNet, and ResNet50. Among these four CNNs, ResNet50 gives the highest performance. The accuracy and loss graphs for ResNet50 is shown in Fig. 6.

Fig. 7 and Table 3 show that the accuracy of our dataset using different methods such as InceptionV3, VGG19, MobileNet, and ResNet50 is 87.91%, 84.21%, 98.03%, and 98.51%, respectively. Similarly, the misclassification rate is 12.09%, 15.79% and 1.97% and 1.49%, respectively. ResNet50 performs comparatively better than other methods by giving a higher accuracy and a lower misclassification rate. Table 4 represents the resultant confusion matrices for ResNet50.

The outcome of the nutrition value estimation obtained using the techniques proposed in Section 2.5 is discussed here. At first, we have calculated the height of the bottle and cap from an image and then find the ratio. We have calculated the ratio for 500 test images (100 images of each category). Among them, one ratio of each category is shown in Fig. 8.

We have considered only the heights of the bottle and cap here. The heights can be changed if the distance, angle of the bottle are changed. To minimize this problem, we have used the height and cap ratio. Because, when the height of a bottle is changed due to distance or other factors in an image, the height of the cap in that image also changed accordingly.

As shown in Table 5, 250 ml Coca Cola bottle always gives a ratio between 4.0∼4.9, for 400 ml bottle it is 5.9∼ 6.9, for 600 ml

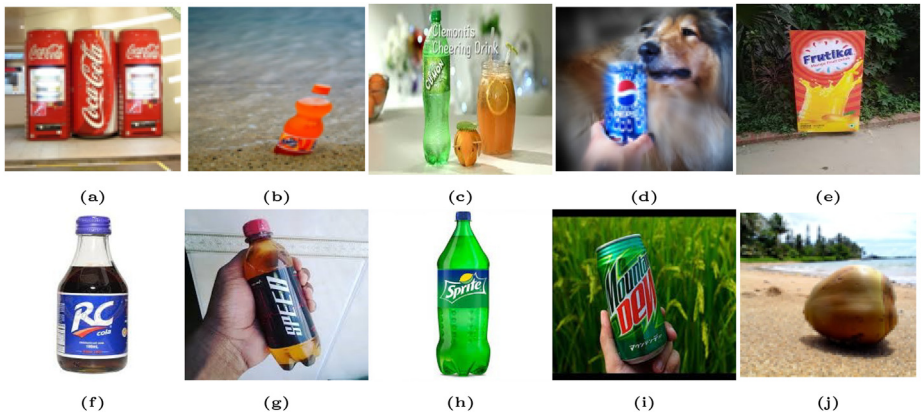

Fig. 5. Sample images from both test and training dataset. (a) Coca Cola, (b) Fanta, (c) Clemon, (d) Pepsi, (e) Frutika, (f) RC Cola, (g) Speed, (h) Sprite, (i) Mountain Dew (j) Green Coconut.

Table 2
Prediction accuracy of the drink classification task by different CNNs along with transfer learning. Here, images that contain clearly visible drinks are considered as 'easy', images with a small amount of cluttered are 'medium', and a cluttered, smashed and partial view of drink images are considered as 'hard'.

Method	Accuracy		
	Easy	Medium	Hard
VGG19	92%	87%	74%
InceptionV3	95%	90%	79%
MobileNet	99.48%	99.1%	96.05%
ResNet50	**99.88%**	**99.4%**	**96.25%**

it gives almost 7∼8, 1.25L gives 9.0∼9.9 and 2.25L gives 10.8∼11.8. The ratio for test image is also calculated and compared with the range shown in the Table 5. If the ratio is between 7.0 and 7.9, then we can say that this is a 600 ml Coca Cola bottle (as shown in Fig. 3) and we can show the nutrition value perfectly. Also, there are still some cases where the height may vary, hence, for bottle size estimation this method gives almost 80% accuracy as shown in Table 6.

For Green Coconut and other soft drinks that we have used here, we collected their nutrition fact in terms of Energy (kcal), Fat, Protein, Carbohydrates, Sugar, Salt, Sodium, etc., from internet hlcyan (Campos et al., 1996). After predicting an image as Green Coconut, its nutrition values are shown for 1 cup (240 g) fresh coconut water. For other soft drinks, once it is classified and its bottle size is estimated properly, then we show the nutrition facts from the previously collected data.

5. Error analysis

There is a little difference in the height of 400 ml and 600 ml bottle compared to the differences among other bottles. As we have calculated this ratio from several types of simple (image contains only Coca Cola bottle) as well as cluttered images due to the camera angle we got some overlapped values. Fig. 9 represents some of the images that can not classify correctly by our system.

This is because there are inter-class similarities among the drink bottles, poor resolution of the images, the single image con-

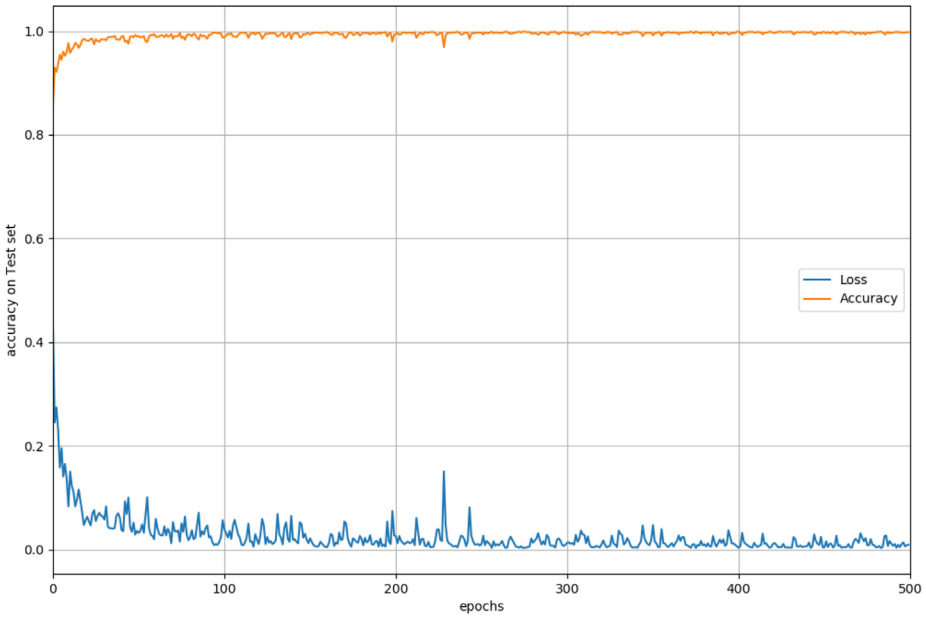

Fig. 6. Accuracy and loss function vs training epoch graph for ResNet50.

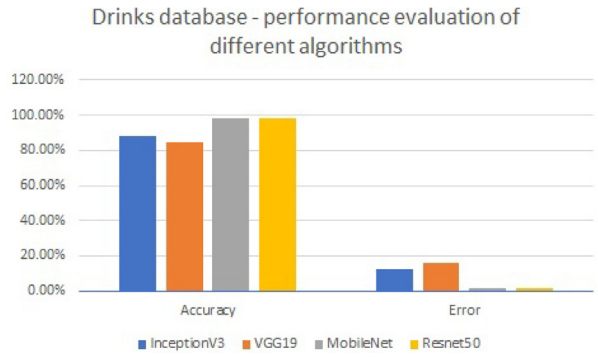

Fig. 7. Performance (Accuracy & Error) evaluation of drink dataset using different CNNs along with transfer learning algorithm.

Table 3
Overall identification accuracy using different CNNs along with transfer learning algorithm.

VGG19	InceptionV3	MobileNet	ResNet50
84.21%	87.91%	98.03%	**98.51%**

tains multiple classes, parts of the drinks are missing or very small part of the drink is visible, angular images, 2-D images are difficult to categorize for a 3-D object, etc.

6. Conclusion

Obesity or overweight becomes a significant nationwide health concern due to the fact that it has a great impact on numerous chronic diseases. Studies found that proper nutrition information is effective to control weight gain. With this view, in this work, we have developed a deep CNN-based transfer learning technique to recognize the type of beverages and then estimate its nutrition facts such as calories, energy, fat, sugar, etc. The method contains some pre-processing tasks before the final recognition and classification modules along with the estimation of dietary facts. For experimental purposes, we developed a dataset containing 10 most commonly consumed drinks in Bangladesh. After investigating the performances through experimentation of the four well-known deep CNNs such as VGG19, InceptionV3, MobileNet, and ResNet50 along with transfer learning it is being found that ResNet50 provides the highest recognition accuracy of 98.51%. We hope this method will help people in controlling overweight and make them health conscious.

Table 4
Confusion matrix for ResNet50.

	Predicted class										
Actual Class		0	1	2	3	4	5	6	7	8	9
	0	**98.33**	0.33	0	0.37	0.08	0	0.23	0.37	0.33	0
	1	0	**99.33**	0	0.1	0	0.33	0	0.23	0	0
	2	0	0	**99.33**	0	0.08	0	0.28	0	0.3	0
	3	0.33	0.21	0	**99**	0.12	0	0	0.33	0	0
	4	0	0	1.28	0	**90**	0	5.03	0	1.2	2.5
	5	0	0	0	0	0	**100**	0	0	0	0
	6	0	0	0	0	0	0	**97.5**	0	2.5	0
	7	0	2.53	0	2.48	0	0	0	**95**	0	0
	8	0	0	0	0	0	2.5	0	0	**97.5**	0
	9	0	0	0	0	0	0	0	0	0	**100**
	Accuracy ***98.51%***										

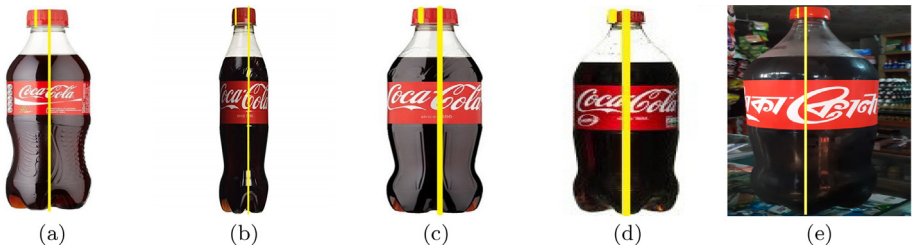

Fig. 8. Euclidean Distance Based Ratio (cap Vs whole bottle height) for some sample bottles. (a) For this 250 ml bottle image, we found the ratio is 4.7990, (b) 400 ml bottle gives a ratio of 6.2927, (c) For 600 ml, we found the ratio is 7.3327 (d) For 1.25 L the ratio is 9.5715, (e) For 2.25 L the ratio is 11.0513.

Table 5
Minimum, Maximum and Average ratio for 500 sample images (100 from each category). Due to variations in images, we did not get a fixed ratio. For 100 images of 250 ml Coca Cola bottle, we find the ratio between 4.023396 and 4.996035. So, we can say for any sample image, if the ratio is between this range, then it is a 250 ml bottle.

Different size of bottles	Min	Max	Average	Range
250 ml	4.023396	4.996035	4.512002	4.0∼4.9
400 ml	5.916175	6.997552	6.459844	6.0∼6.9
600 ml	6.890518	7.878597	7.353289	7.0∼7.9
1.25 L	9.009991	9.997666	9.527296	9.0∼9.9
2.25 L	10.80334	11.79863	11.26137	11.0∼11.9

Table 6
Accuracy of bottle size estimation process for nutrition value calculation. Our ratio calculation based size estimation method can identify a 250 ml Coca Cola bottle with an accuracy of 79.26%.

Coca Cola bottle size	250 ml	400 ml	600 ml	1.25 L	2.25 L
Accuracy	79.26%	67.83%	71.69%	78.12%	81.23%

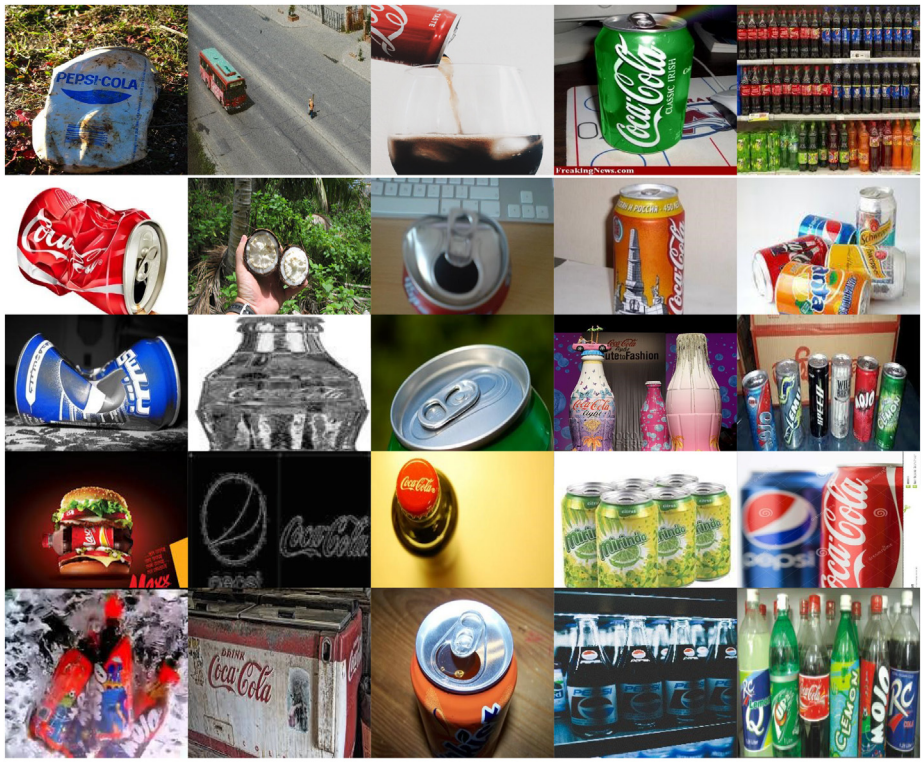

Fig. 9. Some sample error-prone images: (a) the drink can is smashed or unclear which causes imperfect identification, (b) Improper images found in the dataset, (c) Partially visible or top view; type of the drink is not clear, (d) Color of the can is not relevant with the original color of the class, (e) Images containing multiple classes of drinks.

7. Funding

No funding was received for this research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Confusion matrices of VGG19, Inception V3, and MobileNet are shown in the following three tables. See Tables 7–9.

Table 7
Confusion matrix for VGG19.

		Predicted class									
Actual Class		0	1	2	3	4	5	6	7	8	9
	0	**70.67**	3.67	5.67	1.5	0	3.84	0	5.57	2.43	6.67
	1	2.67	**86.67**	1.13	0	1.43	0	3.4	0.57	0.67	3.47
	2	0	3.2	**93**	0	0.8	0	1.67	1.33	0	0
	3	0.77	1.9	2.1	**88**	1.4	2.07	0.77	0	0	3
	4	9.75	0.75	0	0.75	**85.5**	0	1.4	0	1.85	0
	5	5.25	2.25	2.75	0	0	**82.5**	0	5	0	2.25
	6	1.1	0	0.85	2.9	5.9	0	**82.65**	0	4.25	2.5
	7	0.63	2.5	0	0	0.63	0	1.25	**95**	0	0
	8	1.29	0	0	1.21	0	0	0	0	**97.5**	0
	9	1.25	0	2.4	0	1.35	5	0	2. 1	0	**87.5**
						Accuracy ***84.21%***					

Table 8
Confusion matrix for InceptionV3.

		Predicted class									
Actual Class		0	1	2	3	4	5	6	7	8	9
	0	**75.33**	13.33	0	3.33	1.33	2	0	3.33	0	1.33
	1	1.33	**95.33**	1.33	0	0	0	1.33	0	0	0.66
	2	0	2	**90**	3.33	1.33	2	0	0	1.33	0
	3	3.33	3.33	2.66	**90**	0	0	0.66	0	0	0
	4	0	3.75	0	5	**87.5**	0	0	2.5	0	1.25
	5	5.01	4.83	0	0	0	**90.17**	0	0	0	0
	6	0	0	7.25	0	2.75	0	**85**	0	5	0
	7	0	8.5	0	6.75	0	0.25	0	**84.5**	0	0
	8	5	0	1.15	0	2.78	0	0.58	0	**90.5**	0
	9	0	0.775	0	0	0.7	0	2.8	0	0.75	**94.97**
						Accuracy ***87.91%***					

Table 9
Confusion matrix for MobileNet.

		Predicted class									
Actual Class		0	1	2	3	4	5	6	7	8	9
	0	**97.33**	0.5	0	0.27	0.33	0	0.63	0	0.93	0
	1	0.65	**98.68**	0	0.3	0	0	0	0.37	0	0
	2	0	0.75	**98**	0	0	0.33	0	0	0.33	0.58
	3	0.37	0	0.32	**98.5**	0.41	0	0	0.4	0	0
	4	0	0	0	0	**100**	0	0	0	0	0
	5	0.63	0	1.88	0	0	**97.5**	0	0	0	0
	6	0.5	0	0	2.4	0	0	**95**	0	2.1	0
	7	0	0	0	0	0	0	0	**100**	0	0
	8	0	0.75	0	0	0	1.75	0	0	**97.5**	0
	9	0	0	0	0	0	0	0	0	0	**100**
						Accuracy ***98.03%***					

References

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. In: European conference on computer vision. Springer, pp. 404–417.

Bosch, M., Zhu, F., Khanna, N., Boushey, C.J., Delp, E.J., 2011. Combining global and local features for food identification in dietary assessment. In: 2011 18th IEEE International Conference on Image Processing. IEEE, pp. 1789–1792.

Campos, C.F., Souza, P.E.A., Coelho, J.V., Glória, M.B.A., 1996. Chemical composition, enzyme activity and effect of enzyme inactivation on flavor quality of green coconut water 1. Journal of Food Processing and Preservation 20 (6), 487–500.

Cawley, J., Thow, A.M., Wen, K., Frisvold, D., 2019. The economics of taxes on sugar-sweetened beverages: a review of the effects on prices, sales, cross-border shopping, and consumption. Annual Review of Nutrition 39, 317–338.

Chen, Z., Zhang, T., Ouyang, C., 2018. End-to-end airplane detection using transfer learning in remote sensing images. Remote Sensing 10 (1), 139.

Danielsson, P.-E., 1980. Euclidean distance mapping. Computer Graphics and Image Processing 14 (3), 227–248.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Fergus, P., Hussain, A., Hearty, J., Fairclough, S., Boddy, L., Mackintosh, K., Stratton, G., Ridgers, N.D., Radi, N., 2015. A machine learning approach to measure and monitor physical activity in children to help fight overweight and obesity. In: International Conference on Intelligent Computing. Springer, pp. 676–688.

Gribaudo, M., Moos, S., Piazzolla, P., Porpiglia, F., Vezzetti, E., Violante, M.G., 2019. Enhancing spatial navigation in robot-assisted surgery: An application. In: International Conference on Design, Simulation, Manufacturing: The Innovation Exchange. Springer, pp. 95–105.

Hafiz, R., Islam, S., Khanom, R., Uddin, M.S., 2016. Image based drinks identification for dietary assessment. In: 2016 International Workshop on Computational Intelligence (IWCI). IEEE, pp. 192–197.

Harel, J., Koch, C., Perona, P., 2007. Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552.

He, Y., Xu, C., Khanna, N., Boushey, C.J., Delp, E.J., 2014. Analysis of food images: Features and classification. In: 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2744–2748.

Huang, M., Men, L., Lai, C., 2013. Accelerating mean shift segmentation algorithm on hybrid cpu/gpu platforms. In: Modern Accelerator Technologies for Geographic Information Science. Springer, pp. 157–166.

Hutchesson, M.J., Rollo, M.E., Krukowski, R., Ells, L., Harvey, J., Morgan, P.J., Callister, R., Plotnikoff, R., Collins, C.E., 2015. eh ealth interventions for the prevention and treatment of overweight and obesity in adults: a systematic review with meta-analysis. Obesity Reviews 16 (5), 376–392.

Jensen, C.D., Duncombe, K.M., Lott, M.A., Hunsaker, S.L., Duraccio, K.M., Woolford, S. J., 2016. An evaluation of a smartphone–assisted behavioral weight control intervention for adolescents: pilot study. JMIR mHealth and uHealth 4 (3), e102.

Kadouf, H.H.A., Mustafah, Y.M., 2013. Colour-based object detection and tracking for autonomous quadrotor uav. IOP Conference Series: Materials Science and Engineering, vol. 53. IOP Publishing, p. 012086.

Kagaya, H., Aizawa, K., 2015. Highly accurate food/non-food image classification based on a deep convolutional neural network. In: International Conference on Image Analysis and Processing. Springer, pp. 350–357.

Kagaya, H., Aizawa, K., Ogawa, M., 2014. Food detection and recognition using convolutional neural network. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM, pp. 1085–1088.

Karunakar, P., Praveen, V., Kumar, O.R., 2013. Discrete wavelet transform-based satellite image resolution enhancement. Advance in Electronic and Electric Engineering 3 (4), 405–412.

Kaya, H., Gürpınar, F., Salah, A.A., 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing 65, 66–75.

Khaire, M.G., Shelkikar, R., 2013. Resolution enhancement of images with interpolation and dwt–swt wavelet domain components. International Journal of Application or Innovation in Engineering and Management, vol. 2.

Kölsch, A., Afzal, M.Z., Ebbecke, M., Liwicki, M., 2017. Real-time document image classification using deep cnn and extreme learning machines. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, pp. 1318–1323.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Malik, V.S., Li, Y., Pan, A., De Koning, L., Schernhammer, E., Willett, W.C., Hu, F.B., 2019. Long-term consumption of sugar-sweetened and artificially sweetened beverages and risk of mortality in us adults. Circulation.

Martin, C.K., Han, H., Coulon, S.M., Allen, H.R., Champagne, C.M., Anton, S.D., 2008. A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method. British Journal of Nutrition 101 (3), 446–456.

Mezgec, S., Koroušić Seljak, B., 2017. Nutrinet: A deep learning food and drink image recognition system for dietary assessment. Nutrients 9 (7), 657.

Mezgec, S., Eftimov, T., Bucher, T., Seljak, B.K., 2019. Mixed deep learning and natural language processing method for fake-food image recognition and

standardization to help automated dietary assessment. Public Health Nutrition 22 (7), 1193–1202.
Nguyen, K., Fookes, C., Ross, A., Sridharan, S., 2018. Iris recognition with off-the-shelf cnn features: A deep learning perspective. IEEE Access 6, 18848–18855.
Nogueira, K., Penatti, O.A., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognition 61, 539–556.
O'Hara, S., Draper, B.A., 2011. Introduction to the bag of features paradigm for image classification and retrieval, arXiv preprint arXiv:1101.3354.
Rahman, T., Haque, M.R., Rozario, L.J., Uddin, M.S., 2014. Gaussian noise reduction in digital images using a modified fuzzy filter. In: 2014 17th International Conference on Computer and Information Technology (ICCIT), pp. 217–222.
Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural Computation 29 (9), 2352–2449.
Rubasinghe, I., Meedeniya, D., 2020. Automated neuroscience decision support framework. In: Deep Learning Techniques for Biomedical and Health Informatics. Elsevier, pp. 305–326.
Schulze, M.B., Manson, J.E., Ludwig, D.S., Colditz, G.A., Stampfer, M.J., Willett, W.C., Hu, F.B., 2004. Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women. Jama 292 (8), 927–934.
Shan, H., Jia, X., Yan, P., Li, Y., Paganetti, H., Wang, G., 2020. Synergizing medical imaging and radiotherapy with deep learning. Machine Learning: Science and Technology.
Spanakis, G., Weiss, G., Boh, B., Lemmens, L., Roefs, A., 2017. Machine learning techniques in eating behavior e-coaching. Personal and Ubiquitous Computing 21 (4), 645–659.
Suji, G.E., Lakshmi, Y., Jiji, G.W., 2013. Comparative study on image segmentation algorithms. International Journal of Advanced Computer Research 3 (3), 400–405.
Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging 35 (5), 1299–1312.
Tanzi, L., Vezzetti, E., Moreno, R., Moos, S., 2020. X-ray bone fracture classification using deep learning: A baseline for designing a reliable approach. Applied Sciences 10 (4), 1507.
Thrun, S., 1996. Is learning the n-th thing any easier than learning the first. In: Advances in Neural Information Processing Systems, pp. 640–646.
Tian, X., Tao, D., Rui, Y., 2012. Sparse transfer learning for interactive video search reranking. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 8 (3), 26.
Xie, M., Jean, N., Burke, M., Lobell, D., Ermon, S., 2016. Transfer learning from deep features for remote sensing and poverty mapping. In: Thirtieth AAAI Conference on Artificial Intelligence.
Yang, S., Chen, M., Pomerleau, D., Sukthankar, R., 2010. Food recognition using statistics of pairwise local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2249–2256.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328.
Zhu, F., Bosch, M., Khanna, N., Boushey, C.J., Delp, E.J., 2011. Multilevel segmentation for food classification in dietary assessment. In: 2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA). IEEE, pp. 337–342.