Image-based Soft Drink Type Classification and Dietary Assessment System Using Deep Convolutional Neural Network with Transfer Learning

[1]Rubaiya Hafiz, [1]Mohammad Reduanul Haque, [1]Aniruddha Rakshit, [2]Amina khatun and [2]Mohammad Shorif Uddin,
rubaiya.cse@diu.edu.bd reduan.cse@diu.edu.bd aniruddha.cse@diu.edu.bd amina_bashar@yahoo.com shorifuddin@juniv.edu
[1]Dept. of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh,
[2]Dept. of Computer Science & Engineering, Jahangirnagar University, Savar, Bangladesh

Summary
There is hardly any person in modern times who has not taken soft drinks instead of drinking water. The rate of people taking soft drinks being surprisingly high, researchers 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. Therefore, in this work an image-based tool is developed to monitor the nutritional information of soft drinks by using deep convolutional neural network with transfer learning. At first, visual saliency, mean shift segmentation, thresholding and noise reduction technique, collectively known as `pre-processing' are adopted to extract the location of drinks region. After removing backgrounds and segment out only the desired area from image, we impose Discrete Wavelength Transform (DWT) based resolution enhancement technique is applied to improve the quality of image. After that, transfer learning model is employed for the classification of drinks. Finally, nutrition value of each drink is estimated using Bag-of-Feature (BoF) based classification and Euclidean distance-based ratio calculation technique. To achieve this, a dataset is built with ten most consumed soft drinks in Bangladesh. These images were collected from imageNet dataset as well as internet and proposed method confirms that it has the ability to detect and recognize different types of drinks with an accuracy of 98.51%.

Keywords:
Drinks classification, Resolution enhancement, Gaussian noise removal, Mean-shift segmentation, Deep CNN, Bag-of-Feature, Euclidean distance.

1. Introduction

Sugar sweetened beverages (SSB) such as soda, fruit drinks and energy drinks that contains increased presence of sugar (glucose, fructose, sucrose), sweeteners and some other additives consumption become a highly visible and controversial public health issue nowadays. It is found that SSB can increase the risk of type 2 diabetes, heart disease, dental decay, bone density and bone loss, stomach problems, kidney problems etc. [1]. Furthermore, higher consumption of sugary beverages has been linked with an increased risk of premature death [2], obesity and overweight related health problems.

As stated in a survey report published in 2019, about 3.4 million people each year die from obesity [3]. Institute of Health Metrics and Evaluation claims that about 12% of the world's adults and 5% children are obese [3]. The consumption of sugar-sweetened beverages has been suggested as a contributory factor to the rising levels of childhood obesity in many countries worldwide [4]. Duncan Selbie, chief executive of Public Health England, said fruit juices, squash and fizzy drinks were a "major contributor" to one in three children aged 10 or 11 being overweight or obese [5]. Public health officials say children should have no more than six teaspoons which equal to 24g, of sugar a day, leaving no space for fizzy drinks or squashes in the daily diet. Medicare and Medicaid, a government program that pay for medical care for people over age 65, low-income families, and the disabled, now consume $1.42 trillion US dollar for the annual medical and economic costs of all obesity-related people [6].

Haque and Yunus [7] conducted a study among 445 Bangladeshi university going students and found that 35% of them felt addicted to soft drinks. Although this ratio is quite big, but most of the students do not have proper knowledge about the ingredients and the resultant complications of consuming soft drinks. However, 35% of them preferred to consume cola flavor (i.e. Coca-Cola, Pepsi), 24% lemon flavor (i.e. 7UP, Sprite), 23% orange avor (i.e. Fanta, Mirinda) soft drinks, 11% consumed juice (i.e. fruit juice, coconut water), 9% students consumed Mountain Dew and 9% students consumed energy drinks (i.e. speed). The consumption of this type 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 food as their weight

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.20

and health is instantaneously impacted by the amount and type of food and drinks they consume. It is found that appropriate food and diet patterns work as a guard against heart disease, stroke, diabetes and other well-being complications. Consequently, a self-regulating system that can manifest the nutritional variety of different drinks can guide someone to determine if a precise drink is beneficial for his/her well-being or not. A whole-of-government, whole-of-society approach is necessary to create environments for people and communities that are conducive to limiting consumption of sugar sweetened drinks.

Despite the huge impact on human body and also large number of applications in real life, computer systems applied to the classification of SSB have not been widely studied. Given the widespread use of mobile devices such as digital cameras and smartphones, these devices can now be considered as data collection tools for dietitians [8]. With the benefit of image processing techniques, some researchers proposed vision-based approach to identify the amount of calories taken by the users. There are some existing works that proposed the method which aimed at photo-based drink classification and recognition.

Ye He et al. have introduced a k-nearest neighbours and vocabulary tree based technique for food image analysis and classification [9]. They have worked with 42 unique categories and improved the classification performance almost 2%. For the recognition of food items S. Yang et al. [10] have proposed a model that creates a feature vector in discriminative classifier using multi-dimensional histogram. They have 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 and Zhu [11, 12] uses both local and global features for the classification of different food images. We introduced a machine learning based system that can detect and classify several types of drinks automatically from images previously [13]. But the proposed methodology cannot be able to show the nutrition value of each drink according to size.

To reduce the problems discussed above, many researchers have been trying to create a model that can deliver the amount of calories in automatic or semi-automatic way. Recently, it is observed that deep convolutional neural network (DCNN) has remarkable progression in different types of 2 image classification tasks [14, 15, 16]. H Kagaya et. al had applied CNN through parameter optimization for the detection and recognition of foods [17]. When the baseline method is applied, they have found an accuracy of almost 90% whereas for CNN it was nearly 94%. Later H Kagaya and K Aizawa have proposed another model to classify food/non-food images using CNN with an accuracy of more than 95% [18]. Mezgec S et al. had defined a new DCNN architecture called 'Nutrinet' [19] and have applied it on a dataset that contains 520 different types of food and drinks images. They achieved a classification accuracy of almost 87%. An improved methodology was proposed by Mezgec S et al. where they have combined deep learning and natural language processing for the recognition of fake-food images and food matching respectively [20]. The deep learning model provides an accuracy of 92.18% on fake-food images, whereas the overall classification accuracy is 93%. However, no research is done on drink classification.

In this paper, an automated system is developed based on deep CNN with transfer learning for the recognition and classification of different types of drinks (i.e. ten most consumed soft drinks in Bangladesh) and assess the nutrition values which has the following distinguished contributions:

a. A neural network based deep CNN with transfer learning is used for the recognition of different types of drinks.
b. Calculate the nutrition value on the basis of bottle size and contents by employing SURF and color based bag-of-features along with bottle length and cap ratio.
c. Developed a dataset containing ten types of drinks with different bottle sizes.

The forthcoming parts of this paper is fabricated as follows: step II gives the general approach of our suggested scheme. The first part of methodology is called 'pre-processing' which extracts the drink region from cluttered image and improves the quality of segmented region. Then we move on to our learning and recognition step followed by a nutrition value assessment section. Part III describes about our employed dataset and the overall experimental results are presented in section IV. The reasons of misclassification are discussed in section V and eventually, we have concluded in section VI.

2. Methodology

The overall approach of our proposed system is shown in Figure 1. To create a complete object recognition and classification system, it is not enough to concentrate on only the classification process. For proper prediction we should also pay attention to precisely estimate the regions of objects contained in each image. Since most of the images of our dataset contain cluttered scene, we

employed lots of pre-processing as shown in Figure 2, to determine the location of our object of interest. A sample image passing through all stapes is showed in Figure 3. All the stages for the extraction of the drink region from original image are given below.

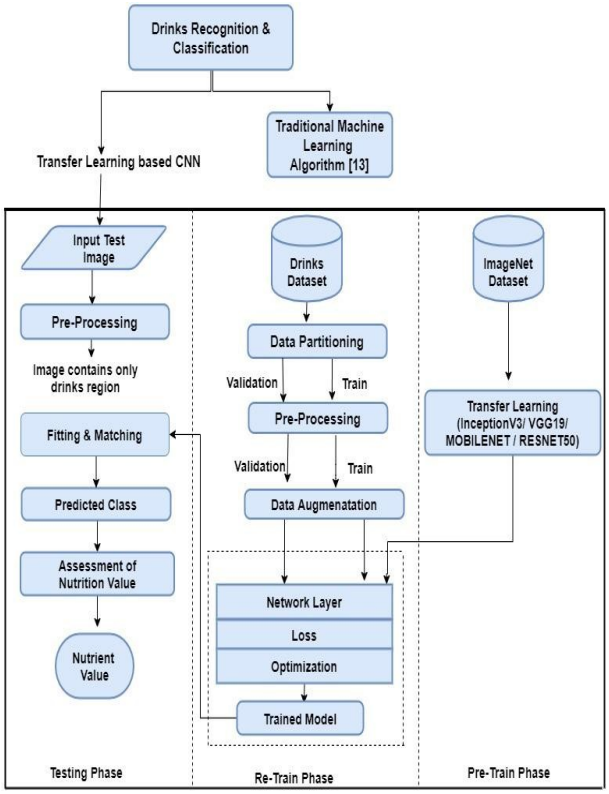

Fig. 1 General Structure of our proposed approach

2.1 Pre-processing for Detecting Object of Interest

2.1.1. Salient Area Detection

This graph-based method helps us to segment out only our desired object from images with cluttered scene [21]. Here, at first, feature vector is extracted from the image plane and an activation map is formed by using equation 1 and 2..

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

From this map, a fully connected directed graph, G_A, is generated by joining each and every node with all other nodes. Each node is labelled with two indices (p, q) ϵ $[n]^2$ and a weight value is assigned with each edge from node (p, q) to node $(x; y)$ using equation 2 and 3:

$$w1((i,j),(x,y)) \triangleq d((i,j)||(x,y).F(i-x,j-y) \quad (2)$$

Where

$$F(x,y) = \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (3)$$

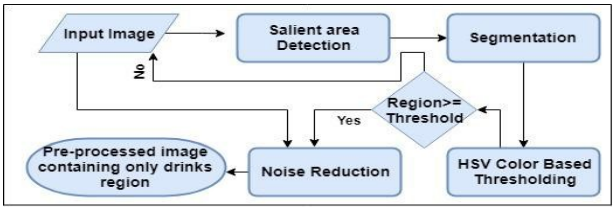

Fig. 2 Pre-processing methodology

Finally, normalization is performed and area that contains 60% saliency are identified. Regions that have less than 60% saliency are considered as the background and 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 mean shift segmentation technique to divide the image into several regions in such a way that each region contains a set of pixels with same characteristics such as color, texture etc. [22]. At first, the original image is filtered to create a vector, $\overrightarrow{Mh}$ using a pixel, P_a and a search window of radius h_s. This process is continued until $|\overrightarrow{Mh}|_2$ is less than a threshold ϵ. Changes are calculated by the following equation [23]:

$$\Delta X = \sum_{n=1} K\left(\frac{i_s - i_a}{hr}\right) W_s X_s \quad (4)$$

Where

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

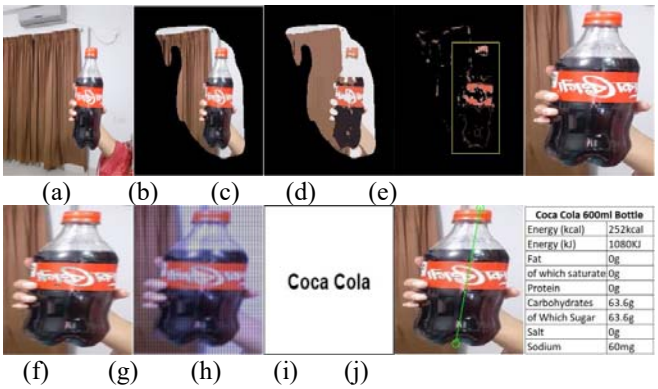

Fig. 3(a) Input image (b) partial image with 60% saliency,(c) resultant image after segmentation, (d)image after color based thresholding, (e) detected object of interest, (f) image after noise removal, (g) after resolution enhancement, (h) Classification result, (i) Ratio calculation from height, (j) after estimation resultant nutrition value.

Also, the new window center, P_b is calculated using the given equation:

$$\overrightarrow{P_b} \longleftarrow \overrightarrow{P_a} + \overrightarrow{Mh} \quad (6)$$

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

2.1.3. Color based Thresholding

Thresholding is a prominent way to decompose the given image into subregions like foreground and background based on the colour [24]. This method helps us to extract only the region of drinks from image. We impose HSV colour based thresholding because our segmented image may obtain 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 images. In some cases, drinks region is either partially visible or not clear (later discussed in section V). For these, our system gave us a pure black region as output after these steps which makes our classification accuracy lower. This is why, 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 version image to the next step rather than the resultant one. Then we improve the quality of image by noise removal and resolution enhancement.

2.1.4. Noise Reduction

Noise suppression is a very crucial factor since performance of classification technique depends on the quality of image also. Here a fuzzy filter is imposed for reducing Gaussian noise while keeping the other features intact [25]. If 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 equation 7.

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

Here, σ is the standard deviation of all intensity values. As we have color images, this filter is applied separately for eac h of the R, G, and B components of each image. At last, by c oncatenating these three components, we found our desired i mage.

2.2. Resolution Enhancement

Resolution of image plays a crucial role in many image processing applications especially in feature extraction technique [26]. To get a resolution enhanced image, Discrete Wavelet Transform (DWT) is used to disintegrate the input image into several sub bands [27]. Then, interpolation of the high frequency sub bands images and the low-resolution input image is performed. At last, combination of all these images are used to generate a resultant image using inverse DWT.

2.3. Data Augmentation

To improve classification accuracy and reduce over fitting 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 finished all previously discussed steps, the resultant train images are sent for retrain using transfer learning and used as test images for prediction. In case of transfer learning, knowledge from 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 scratch. The early layers of CNN contain generic features that can be re-used while the final layers are more application specific. Because of this property, the initial layers are well-preserved while the endmost ones are ne tuned to train with the current dataset of interest [28, 29].

To attain high recognition accuracy, deep learning models need huge labelled data to for training the classifier. However, it is quite challenging to get such dataset for each and 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 train CNNs with scarce using Previously trained dataset, initiated by Thrun [30].

A study has conducted by Yosinki et al. [28] to fine tune the CNN based transfer learning model which was pre-trained on ImageNet database. To transfer information from labelled data to unlabeled data, Tian et al. [31] proposed a model for sparse transfer learning with a view to re-rank the video search. For medical image analysis, Tajbakhsh et al. [32] investigated the performance of full trained CNNs with the pre-trained CNNs. The transfer learning technique also successfully utilized in video based emotion recognition [33], iris recognition [34], end-to-end airplane detection [35], and poverty mapping [36].

We retrain our drink dataset by popular 4 well known deep CNNs: VGG19, InceptionV3, MobileNet, and Resnet50. A brief architecture of these deep CNNs are given in Table I.

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

2.5. Nutrition Value Assessment

For any soft drinks, after classification, we again impose color and SURF [37] based Bag-of-Feature [38] technique to identify how it is served, i.e. in a glass, can, glass bottle or plastic bottle.

Table 1: A brief architecture of 4 well-known deep CNNs for transfer

	InceptionV3	VGG19	MobileNet	Resnet50
Input Size	227x227	224x224	224x224	224x224
Conv. Layer	21	19	28	34
Filter Size	1,3,5,7	3	1,3	1,3
Stride	1,2	1,2	2	2
Parameter	23M	155M	5855942	25.6M
Fc Layer	1	1	4	1
Size	92MB	549MB	17MB	-
Depth	159	26	88	-

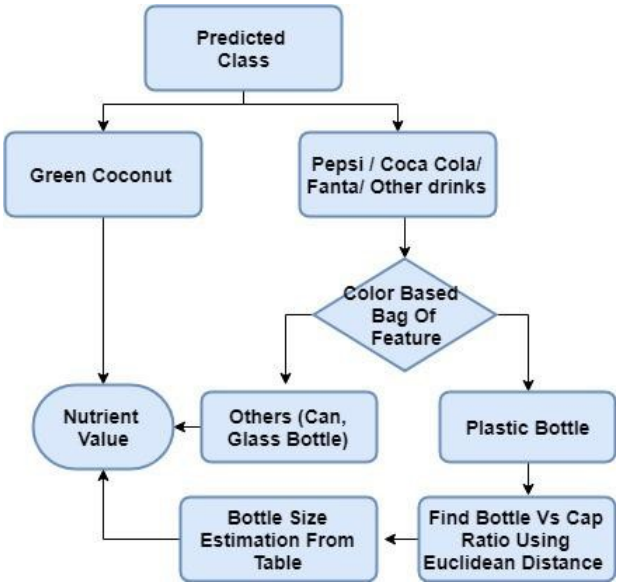

Fig. 4 Process of Nutrition value Estimation

We retrain our drink dataset by popular 4 well known deep CNNs: VGG19, InceptionV3, MobileNet, and Resnet50. A brief architecture of these deep CNNs are given in Table I.

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

$$dist((x, y), (a, b)) = \sqrt{((x - a)^2 + (y - b)^2} \quad (8)$$

Then the ratio is calculated as follows:

$$ratio=(Distance\ of\ full\ bottle\ /\ Distance\ of\ cap\ ) \quad (9)$$

3. Experimental Data

Our experimental data set consists of ten distinct categories of drinks and beverages. We have chosen these because they are mostly common and available everywhere in Bangladesh. 1250 images of each category (total 1250 x 10= 12,500) are used to construct our experimental data set. Most of the images are collected from a hierarchical large-scale database called ImageNet [39]. 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 training set and remaining 20\% (250 for each category) as testing set. Figure 5 demonstrate some sample images from our training and testing dataset.

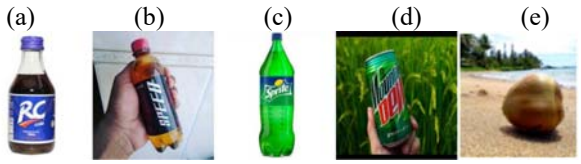

Figure 5: Sample images from our test and training dataset. (a) Coca Cola, (b) Fanta, (c) Clemon, (d) Pepsi, (e) Frutika, (j) RC Cola, (k) Speed, (m) Sprite, (n) Green Coconut.

4. Experimental Results

The following section present and discuss the exper imental result of our propound system. For the purpose of experimentation, at first, we partitioned our dataset accordi ng to the difficulty of classification into three parts, i. e. ea sy(images in which drinks are clearly visible and easily ide ntifiable), medium(images containing small amount of clut terness), hard(images containing cluttered scene, smashed as well as top and/or partial view of drinks). Classification performance of each part is shown in Table II. Figure 6 sh ows the comparative performance of four popular transfer l earning methods i.e. InceptionV3, VGG19, MobileNet and Resnet5
0. Here, MobileNet and Resnet50 performs better than othe r methods.

TABLE II: Predict Accuracy of Drink classification Task

	Accuracy		
Method	Easy	Medium	Hard
VGG19	92%	87%	74%
Inception V3	95%	90%	79%
MOBILENET	99.48%	99.1%	96.05%
RESNET50	**99.88%**	**99.4%**	**96.25%**

Figure 6 and Table III shows 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%. Resnet50 performs comparatively good than other methods by giving better accuracy and lower misclassification rate. Table IV represent the resultant confusion matrices for Resnet50.

The outcome of nutrition value estimation obtained using the techniques proposed in Section 2.4 is discussed here. We have calculated the ratio of Euclidean distances of full bottle length and bottle cap for total 500 test images (100 images of each category). Among them one ratio of each category is shown in Figure 7.

As shown in Table V, 250ml Coca Cola bottle always gives a ratio between 4.0~4.9, for 400ml bottle it is 5.9~6.9, for 600ml 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 matched with the range shown in the table. If the ratio is between 4.0 and 4.9, then we can

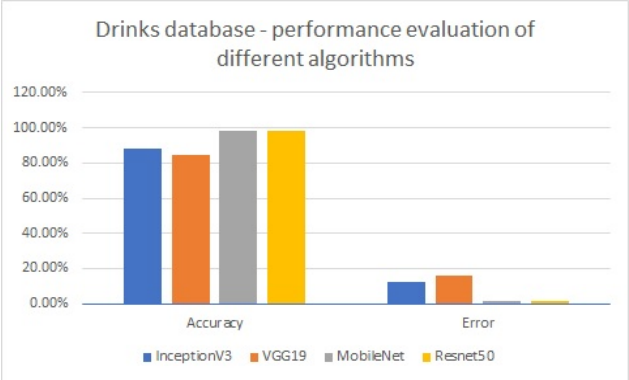

Figure 6: Performance evaluation of drink dataset using different transfer learning algorithms

TABLE III Overall identification accuracy using different transfer learning methods

VGG19	InceptionV3	MOBILENET	RESNET50
84.21%	87.91%	98.03%	98.51%

say that this is a 250ml Coca Cola bottle and we then can show the nutrition value perfectly.

There is a little difference in the height of 400ml and 600ml 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.

5. Error Analysis

Figure 8 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 resolut ion of the images, single image contains multiple classes, p arts of the drinks are missing very small part of the drinks are visible, angular images, 2-D images are difficult to cate gorize for 3-D object etc.

TABLE IV 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
	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%									

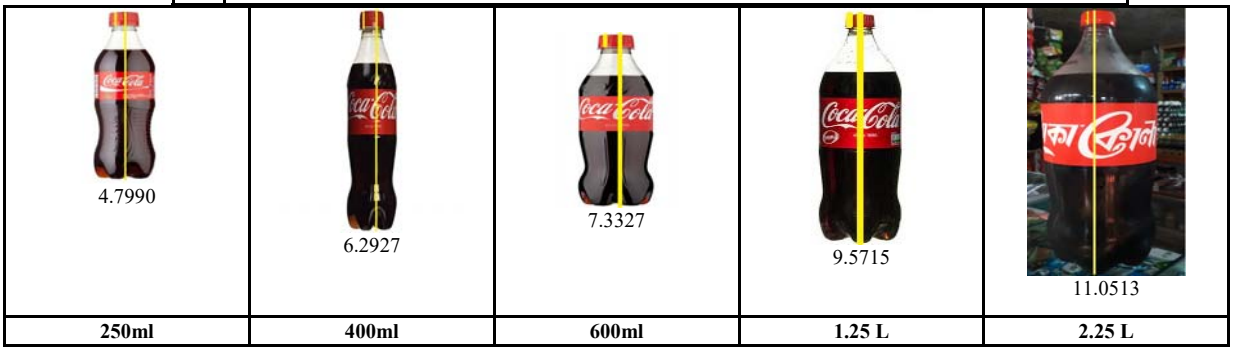

Figure 7: Euclidean Distance Based Ratio (cap Vs whole bottle height) for some sample bottles

Table V: Minimum, Maximum and Average ratio for 500 sample images (100 from each category).

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

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.20

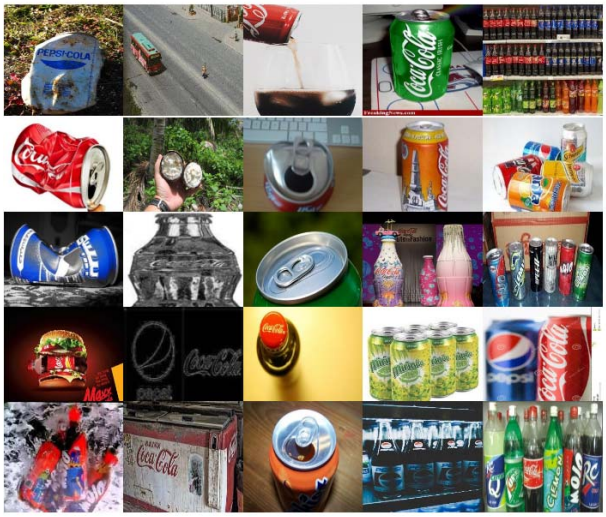

Figure 8: Sample error-prone images. (a) Can is smashed or unclear which causes imperfect identification, (b)Improper images found in the dataset, (c)Partially visible or top view; type of the drink is not clear, (d) Color of the can is not relevant with the original color of the class, (e) Images containing multiple classes of drinks.

6. Conclusion

Because of having a great impact on numerous chronic diseases, nowadays, obesity or overweight has become a significant nationwide health concern. Studies [27] found that proper nutrition information is effective to control weight gain. For this, a detailed research effort have been conducted in this area. In this work, we investigated popular transfer learning techniques to recognize the type of beverages and then estimate the amount of nutrition value such as calorie, energy, fat, sugar etc in it. Region of interest is segmented out using all the steps described in pre-processing part and the resultant image is then used to classify and recognize the drinks. Lastly, by employing some other techniques, our system could be able to identify the size of the bottle of drinks and thus it shows the nutrient information of that specific drink perfectly. After investigating the individual performance it is be observed that among all transfer learning models, RESNET50 provides the highest recognition accuracy of 98.51%.

References

[1] *Health Risks of Drinking Soft Drinks*, 2019 (Accessed June 10, 2019). http://www.historyofsoftdrinks.com/soft-drinks-facts/health-e ects-of-soft-drinks/.

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

[3] V. Shukla, *Top 10 Most Obese Countries In The World According To WHO And OECD*, January 14, 2019 (Accessed June 3, 2019). https://www.valuewalk.com/2019/01/top-10-most-obese-countries-oecd-who/.

[4] T. Lobstein, *Reducing consumption of sugar-sweetened beverages to reduce the risk of childhood overweight and obesity*, September 2014 (Accessed June 9, 2019). https://www.who.int/elena/titles/commentary/ssbschildhoodobesity=en/.

[5] D. Hyde, *Remove sugary drinks from children's diets, health officials say*, Jul 2015 (Accessed June 9, 2019). https://www.telegraph.co.uk/news/health/11745806/Remove-sugary-drinks-from-childrens-diets-health-o cials-say.html.

[6] D. Moza arian, *Food is medicine: How US policy is shifting toward nutrition for better health*, January 2019 (Accessed June 9, 2019). https://theconversation.com/food-is-medicine-how-us-policy-is-shifting-toward-nutrition-for-better-health-107650.

[7] M. Haque et al., *A Survey on Soft Drinks Intake Behaviour among University Going Students*, PhD thesis, East West University, 2018.

[8] C. K. Martin, H. Han, S. M. Coulon, H. R. Allen, C. M. Champagne, and S. D. Anton, "A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method," British Journal of Nutrition, vol. 101, no. 3, pp. 446-456, 2008.

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

[10] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2249-2256, IEEE, 2010.

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

[12] F. Zhu, M. Bosch, N. Khanna, C. J. Boushey, and E. J. Delp, "Multilevel segmentation for food classi cation in dietary assessment," in 2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 337-342, IEEE, 2011.

[13] R. Hafiz, S. Islam, R. Khanom, and M. S. Uddin, "Image based drinks identi cation for dietary assessment," in 2016 International Workshop on

Computational Intelligence (IWCI), pp. 192-197, IEEE, 2016.
[14] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classi cation: A comprehensive review," Neural computation, vol. 29, no. 9, pp. 2352-2449, 2017.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097-1105, 2012.
[16] A. Kolsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, "Real-time document image classification using deep CNN and extreme learning machines," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1318-1323, IEEE, 2017.
[17] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in Proceedings of the 22nd ACM international conference on Multimedia, pp. 1085-1088, ACM, 2014.
[18] H. Kagaya and K. Aizawa, "Highly accurate food/non-food image classi cation based on a deep convolutional neural network," in International Conference on Image Analysis and Processing, pp. 350-357, Springer, 2015.
[19] S. Mezgec and B. Korousic Seljak, "Nutrinet: A deep learning food and drink image recognition system for dietary assessment," Nutrients, vol. 9, no. 7, p. 657, 2017.
[20] S. Mezgec, T. Eftimov, T. Bucher, and B. K. Seljak, "Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment," Public health nutrition, vol. 22, no. 7, pp. 1193-1202, 2019.
[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in Advances in neural information processing systems, pp. 545-552, 2007.
[22] G. E. Suji, Y. Lakshmi, and G. W. Jiji, "Comparative study on image segmentation algo-rithms," International Journal of Advanced Computer Research, vol. 3, no. 3, pp. 400-405, 2013.
[23] M. Huang, L. Men, and C. Lai, "Accelerating mean shift segmentation algorithm on hybrid cpu/gpu platforms," in Modern Accelerator Technologies for Geographic Information Science, pp. 157-166, Springer, 2013.
[24] H. H. A. Kadouf and Y. M. Mustafah, "Colour-based object detection and tracking for autonomous quadrotor uav," in IOP Conference Series: Materials Science and Engineering, vol. 53, p. 012086, IOP Publishing, 2013.
[25] T. Rahman, M. R. Haque, L. J. Rozario, and M. S. Uddin, "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, IEEE, 2014.
[26] M. G. Khaire and R. Shelkikar, "Resolution enhancement of images with interpolation and dwt-swt wavelet domain components," International Journal of Application or Innovation in Engineering and Management, Vol2, 2013.
[27] P. Karunakar, V. Praveen, and O. R. Kumar, "Discrete wavelet transform-based satellite image resolution enhancement," Advance in Electronic and Electric Engineering, vol. 3, no. 4, pp. 405-412, 2013.
[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Advances in neural information processing systems, pp. 3320-3328, 2014.
[29] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," Pattern Recognition, vol. 61, pp. 539- 556, 2017.
[30] S. Thrun, "Is learning the n-th thing any easier than learning the first?," in Advances in neural information processing systems, pp. 640-646, 1996.
[31] X. Tian, D. Tao, and Y. Rui, "Sparse transfer learning for interactive video search reranking," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 8, no. 3, p. 26, 2012.
[32] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or ne tuning?," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1299-1312, 2016.
[33] H. Kaya, F. Gurp nar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," Image and Vision Computing, vol. 65, pp. 66-75, 2017.
[34] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris recognition with o -the-shelf cnn features: A deep learning perspective," IEEE Access, vol. 6, pp. 18848-18855, 2018.
[35] Z. Chen, T. Zhang, and C. Ouyang, "End-to-end airplane detection using transfer learning in remote sensing images," Remote Sensing, vol. 10, no. 1, p. 139, 2018.
[36] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[37] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in European conference on computer vision, pp. 404-417, Springer, 2006.
[38] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," arXiv preprint arXiv:1101.3354, 2011.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale

hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255, Ieee, 2009.

Appendix

TABLE VI. 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	0.82	98	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%										

TABLE VII. 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	6.25	0	0.25	0	85	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 VIII. 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	82.5	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	0	97.5
	9	1.25	0	2.4	0	1.35	5	0	2. 1	0	87.5
	Accuracy 84.21 %										

RUBAIYA HAFIZ received her Master of Science and Bachelor of Science degrees in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka. She is currently served as Senior Lecturer in the Department of Computer Science and Engineering at Daffodil International University, Savar, Dhaka, Bangladesh. Her research interests includes Computer Vision, Deep Learning, Image Processing, Artificial Intelligence etc.

MOHAMMAD REDUANUL HAQUE received his Master of Science and Bachelor of Science degrees in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka in 2012 and 2011, respectively. He is currently served as a Senior Lecturer in the Department of Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. His research interests includes Computer Vision, Deep Learning and Image Processing.

ANIRUDDHA RAKSHIT received his Master of Science and Bachelor of Science degrees in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka. He is currently served as a Senior Lecturer in the Department of Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. His research interests includes Computer Vision, Deep Learning and Image Processing.

AMINA KHATUN received her Master of Science and Bachelor of Science degrees in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka and North South University. She is currently served as an Assistant in the Department of Computer Science and Engineering at Jahangirnagar University, Savar, Dhaka,

Bangladesh. Her research interests includes Computer Vision, Deep Learning, Image Processing, Computer Architecture, Digital Systems and Software Engineering.

MOHAMMAD SHORIF UDDIN (M'13–SM'15) received his Doctor of Engineering degree in information Science from Kyoto Institute of Technology in 2002, Japan, Master of Technology Education degree from Shiga University, Japan in 1999, Bachelor of Electrical and Electronic Engineering degree from Bangladesh University of Engineering and Technology in 1991 and also MBA in from Jahangirnagar University in 2013. He began his teaching career as a Lecturer in 1991 at the Bangladesh Institute of Technology, Chittagong (Renamed as CUET). He joined in the Department of Computer Science and Engineering of Jahangirnagar University in 1992 and currently, he is a Professor of this department. He undertook postdoctoral researches at Bioinformatics Institute, Singapore, Toyota Technological Institute, Japan and Kyoto Institute of Technology, Japan, Chiba University, Japan, Bonn University, Germany, Institute of Automation, Chinese Academy of Sciences, China. His research is motivated by applications in the fields of imaging informatics and computer vision. Mohammad Uddin is an IEEE Senior Member and also a Fellow of Bangladesh Computer Society and The Institution of Engineers Bangladesh. He wrote more than 100 journal and received the Best Paper award in the International Conference on Informatics, Electronics & Vision (ICIEV2013), Dhaka, Bangladesh and Best Presenter Award from the International Conference on Computer Vision and Graphics (ICCVG 2004), Warsaw, Poland. He holds two patents for his scientific inventions. Currently, he is the Chair of IEEE CS Bangladesh Chapter and an Associate Editor of IEEE Access.