National College of Ireland

Food Image Recognition and Calorie Estimation Using Object Detection Algorithms

MSc Research Project
Programme Name

Manoj Kumar Yuganathan
Student ID: x20179189

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Manoj Kumar Yuganathan
Student ID:	x20179189
Programme:	Programme Name
Year:	2021-2022
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	31/01/2022
Project Title:	Food Image Recognition and Calorie Estimation Using Object Detection Algorithms
Word Count:	7284
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Manoj Kumar Yuganathan
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Food Image Recognition and Calorie Estimation Using Object Detection Algorithms

Manoj Kumar Yuganathan
x20179189

Abstract

Obesity has caused a significant increase in the number of people suffering from various health problems in recent years. Excessive calorie consumption is one of the most important factors contributing to obesity. This project proposes a solution for recognizing and identifying different food items, as well as tracking the estimated number of calories consumed for the various food items based on their corresponding food proportions. This study uses the convolutional neural networks and transfer learning algorithms such as VGG, InceptionNet, and ResNet to extract features to estimate calorie consumption. The novelty of this study stems from the fact that this implementation is done by combining data augmentation techniques with applying rapidly decreasing threshold learning rate schedulers to achieve high detection rates of multiple food items on a single food plate. This ensemble technique of performing feature selection, one-hot encoding, data augmentation, and applying the tuned dataset to the InceptionNet V3 model helped to achieve an increased accuracy of almost 87%. In future research, identifying the estimated calorie count using the bounding and applying calibration technique to accurately predict the number of calories consumed by an individual for their daily meal would be implemented.

Keywords - food recognition, deep learning, calorie consumption, InceptionNet v3, data augmentation, image classification.

1 Introduction

In the contemporary era, one common problem surging alarmingly in all countries irrespective of the urban or rural backgrounds is the huge amount of health ailments and health conditions recorded from consuming excessive or deficit number of calories and the underlying cause for this pattern stems from either being anemic or obese. Obesity and anemia can lead to an abundant number of health disorders and ramifications such as cardiovascular disease, musculoskeletal diseases, tuberculosis, and several unique variants of cancers. This issue affects all the countries unilaterally, and urban areas have reported higher level of impacts than the rural areas. Study of nutritional science has emerged as the primary focus for comprehending food composition constituents, identifying food patterns and researching the ideal amount of food quantities that should be consumed in every individual's meal on a day-to-day basis (Moorhead et al.; 2015).

Obesity is the root cause of a voluminous number of health issues and problems (Fernandes et al.; 2015). There has a been a huge raise in the number of health-ailment cases and the root cause for these reported cases is due to being obese and suffering

from overweight abnormalities. These kind of obesity-related health issues has increased alarmingly, particularly during the current pandemic. The most difficult challenge associated with this project is detecting the composition and category of food items in their daily meals, as well as monitoring the overall number of calories consumed by each individual, which becomes even more important in this epidemic situation where people are restricted to the comfort of their homes. This innovative solution allows nutritionists and doctors to track and instantly monitor what foods are being consumed, the proportions of the various food portions consumed, and the total calories count.

Based on all the previous researches carried out on food recognition using a wide range of deep learning and machine learning techniques, this study is aimed at establishing an ensemble modeling technique and the implementation is done by combining data augmentation techniques with applying rapidly decreasing threshold learning rate schedulers to achieve high detection rates of multiple food items on a single food plate. This study uses the convolutional neural networks and transfer learning algorithms such as VGG, InceptionNet, and ResNet to extract features to estimate calorie consumption. This ensemble technique of performing feature selection, one-hot encoding, data augmentation, and applying the tuned dataset to the InceptionNet V3 model helped to achieve an increased accuracy.

The results of this study can be used to solve the problems of identifying the food items that constitute our daily meal and these food proportions are estimated using the bounding box technique to precisely estimate the number of calories for each food portion and hence, the overall calorie count would be available to the individual to instantly help the obese individual of how optimal his food consumption has been. This will help the dieticians to take control and monitor the daily needs of their patients and at the same time, the resulting derivations of this research can be used to solve another subproblem of this domain of implementing a food recommendation system. The one most important research question that this project addresses to solve is stated as follows:

"How accurately can object detection algorithms recognize food and accurately estimate the calories of the food items to help people with obesity and health-ailments?"

The research objectives that are accomplished within this project of food recognition using object detection algorithms are stated as follows:

1. Critically researching and analyzing existing state-of-the-art models and thoroughly analyzing the strength, weaknesses, and limitations of all the relevant works carried out.

2. The ultimate aim of this study is to employ ensemble deep learning models that will lead to accurate recognition of the food items, classify the food constituents and precisely estimate the calories consumed for each food portion.

3. The convolutional deep learning models are trained using a dataset that is composed of 101,000 images and the images features are extracted, segmented and the images are classified/grouped into different labelled categories using one hot-encoding technique and subjecting the features as a set of vector variables instead of having one single feature.

4. The dataset is subjected to augmentation by setting zca whitening, rotation, width, height, horizontal and vertical flip, scaling, and zoom range and then modeled using

the InceptionNet v3 model to achieve higher detection accuracy.

5. Comparing the naïve model with another deep learning model which computes results by performing 10 different rotations namely upper left, lower right, lower left, upper right, center and finally arriving at the most common prediction out of the achieved 10 predictions.

6. Evaluating the developed models based on a certain set of evaluation metrics.

7. Comparing the results of the different models to pre-existing models to determine the most ideal model for solving the real-time problem at stake and providing the future scopes of research.

The novel contributions for this research are augmenting the image and the various attributes that are fixated for helping in transforming the image to the required dimensions are zca whitening, rotation, width, height setting, horizontal and vertical flip, scaling and zoom range. This data augmentation is then combined with feature extraction techniques and one-hot encoding to make the image dataset achieve higher accuracy when subjected to the numerous transfer learning algorithms within this study.

The overall contents of the research are structured as follows: The upcoming section, Section 2, focuses on the list of previous state-of-the-art research works that are relevant to this proposal. Section 3 narrates the methodologies and architecture required for this research. Section 4 highlights the system design and the overall structure of the project, 5 states the roadmap and implementation details of this project, 6 constitutes the various evaluation methods used and section 7 discusses the conclusions and future works of this research.

2 Related Work

The purpose of the related works section is to consolidate, discuss and critically analyze all the past works carried out that align with the theme of this project study. The state-of-the-art recognized works by various authors related to data augmentation, segmentation, transfer learning, object detection, and feature extraction works have been carefully scrutinized, reviewed, analyzed and corresponding feedbacks have been stated in this section. The research work's strengths, advantages, limitations, pitfalls, uncovered areas, and future works have also been addressed, and while this building this deep learning model, special care has been taken to ensure the same mistakes have not been repeated while persisting the highlights of the previous works.

2.1 Convolutional Neural Network, SVM and ANN Deep Learning Models

There is an ever-growing need for doctors, dieticians, and nutritionists to track and monitor the daily intake of their patients and to keep a check on the number of calories they have consumed and Rajayogi et al. (2019) presents a solution to the problem of how to monitor the patients in real-time. The visual interface is designated to be a mobile app and the application would be accessible to the end-users namely the patients and they would be obligated to upload the daily photos of the meal they would be consuming on a day-to-day basis. The deep learning model is based on CNN (Convolutional Neural

Network) and this model would identify the constituents of the food, the amount of calories that each food portion consists and it would notify the pieces of information to the health experts. The dataset is set on a scale of 10000 high definition images and the trained dataset returns an accuracy of 83%. The process of implementation involves subjecting the dataset to segmentation of graph cuts, texture, size, and color. The limitation of this study is that it works well with identifying images consisting of a single food variant and images having multiple food items with mixed food portions fared very poorly in this model. And the test accuracy was way less than the training accuracy indicating overfitting of the model.

Sengur et al. (2019) is similar to the Rajayogi et al. (2019) as it also caters to the necessities of the obese people who would want to keep a check on their diet and follow a healthy nutritional intake. The implementation is in the form mobile application where the built-in camera is used to take photos of the user's daily meal before the consumption of the food. The images are recorded, and the distortion is removed by a data cleansing algorithm and the images are calibrated based upon a unique geometric calculation requiring the images to be taken in a very specific manner. The SVM (Support Vector Machine) was used for implementing this data model and this machine learning classification detected single food item portions precisely. But, once again this model also behaved poorly when trying to classify multiple items with portions of mixed food and was not able to recognize the liquid food items. However, the data augmentation techniques employed in this paper helped in achieving significant accuracy of 86%.

M. and C. (2019) provides an in-depth critical analysis of previous research that has been performed in the domain of detection of foods, estimation of calories and determination of food composition on the basis of nutrients in the food. ANN and CNN are some of the Deep Learning methods that have been compared to conventional image processing techniques like spectral imaging and machine learning. The findings from previous research shows that CNN model performs the best among all traditional models in areas of food segregation and detection of food quality. Thus it can be inferred that there is a necessity to perform additional research in domain of food detection using complex deep learning techniques.

In Peddi et al. (2017), a mobile application that performs the collection of images related to food, categorisation the images and also calculation of resultant calorie consumption of every component in a meal has been suggested. The author proposes a solution that is based on the cloud environment that also enables auto-scaling and dynamic load balancing that ensures optimal and proper resource allocation as well as termination. Through the execution of parallel processing of cluster based images, the processing time has been found to be reduced by 45 percent. In the suggested naive approach, the accuracy of the classification is compromised but the process speed and response time are significantly reduced.

2.2 Color, Texture and Graph Cuts Segmentation

Dahiya et al. (2021) is an up-gradation built on top of the previous findings of the author which have been already discussed extensively in (Sengur et al.; 2019). Extensive adaptation of the Dahiya et al. (2021) Support Vector Machine Learning model was performed and it resulted in a significant improvement in accuracy and recorded higher efficiency in performing the data modeling as a result of pre-processing the images subject to segmentation of graph cuts. The segmentation of graph cuts has been used for ensuring smother

outlines, lesser distortions, to filter unwanted detail, and to segment the most important parts of the image located across various coordinates. This pre-processing step resulted in increased accuracy of 3% for food portions of single sizes and 5% for food portions constituting mixed food ingredients. One important disadvantage to still consider is the fact that the model's output with mixed food constituents was still very low. One highlight to pick out from this study is the improvement of the model's accuracy and efficiency when the images are dealt properly with adequate pre-processing stages hence ensuring better smoothening and color-filter undistorted images depicted in grayscale filtering unwanted noise.

The lack of varied and diverse food datasets is one of the most common challenges in the analysis and detailed study of food items. In Wei et al. (2016), this lack of proper dataset is addressed by the creation of datasets from different sources post collaboration, and this dataset will thus be useful for researchers to conduct productive food detection investigations. Different pre-processing steps such as color segmentation, analysis of texture and k-means clustering are performed on the source images in the dataset. The images in the dataset are also subjected to operations such as shrinking and compression to a fixed size of 970 X 720 pixels, which guarantees conistency across all images. The dataset used in this research consists of just 3000 images, which may be considered very small, but the dataset serves as a platform for experimentation using deep application of deep learning models and assess their suitabilty for the same. In Oliveira et al. (2014), the food components in everyday meals are recognised by the use of a mobile phone with a camera. In this research, adherent methods are suggested for recognising the food elements in an image taken by camera irrespective of the angles,lighting and cameras. Segmentation of these images are carried out by the break down into multiple regions that operate on different feature spaces. The classification performance is found to increase by 9%. The major limitation of this approach is that the system expects multiple images from different angles for proper functioning.

2.3 Feature Extraction and Object Detection

Feature extraction is the main feature that is discussed in this paper, here they identify the different shapes and patterns that are associated with a variety of foods. Deep learning is mainly used for image processing because this is used to perform hard identification that can be done in an image. In this paper Sun et al. (2014), they discussed the identification of hidden features in the face to perform face recognition. They performed this model by differentiating three scales and ten regions and integrated it into grayscale for clear analysis. In this identification, they get an accuracy of 97.45 %. The conversion of colored images to grayscale will reduce the image size while training the model and also this gray scaling will give a good view in model building. So we adopted this concept of Food identification in our model building. CNN is implemented in this research with some of the techniques such as Max Pooling, Convolution 2d, to extract more features in the image (Chen et al.; 2016). To avoid overfitting the model is trained with more data and also provides a saturated amount of epochs which will build a perfectly trained model. This approach is discussed for this optical evaluation with 3D Convolution Neural Network (CNN).

Object detection is also one of the important concepts that are used for image processing. Here we use this object detection for identifying different types of food. In this paper Lin et al. (2017), they extract all the features by joining the fast R-CNN

model to improve the different features in the image. The model building is done with a dataset of 2048 images because of this very low dataset for perfectly building a model. Primarily the object detection needs a large dataset for perfectly training a model. This object detection will provide high-level features in our image (Zhao et al.; 2019). The paper gives regression and classification-based CNN models like YOLO, SSD, and Faster-RCNN model. The findings in this paper give the method that can be implemented in a food identification in mixed food portion that can be adopted for our model building.

2.4 Faster-RCNN, YOLO and Transfer Learning Algorithms

Furthermore, because all suggestions need areas to be convoluted, region-based proposal algorithms such as RCNN are computationally costly. RPNs (Region Proposal Networks) use precisely specified areas to analyze and anticipate item detections at fast rates to address this constraint. A new suggestion was proposed to integrate RPNs with Fast-RNN models, and the resulting new build model was termed Faster-RCNN (S. Ren et al.; 2017). The newly presented approach, a groundbreaking invention, utilizes both region proposal techniques and regression/classification methodologies. The COCO dataset of 80k pictures was processed at 200 ms per image using our Faster-RCNN model. Fisheye cameras are employed in restaurants to track and photograph human motions. The pictures are utilized to calculate the average customer queue size, interarrival durations, and waiting periods for each order (Oner et al.; 2019). To process these metrics, faster-RCNN and YOLO models are applied, and the YOLO model beat its competition, obtaining a 15% better overall success detection rate than its peer model. In this experiment, both of these cutting-edge models, YOLO and Faster-RCNN, will be employed to execute food detection analyses.

The output of the optimum model is used to extract the patterns from the with the high phase with more computing time in this research paper. The already pertrained models are trained with a large dataset from ImageNet, and the output of the optimum model is used to extract the patterns from the with the high phase with more computing time. The model has been pre-trained to deal with photographs with bad lighting, irregular resolutions, and distorted images (Li et al.; 2018). In the medical field, these models are commonly employed. The 2D-DenseNet transfer learning algorithm is used in Li et al. (2018) to detect liver tumor patterns from tomographically scanned images, and 3D-ResNet transfer learning algorithms are used in Ragesh et al. (2019) to automate the billing of vegetables and fruits in a retail mart, eliminating the need for human intervention. The accuracy of the DenseNet and ResNet models was 98.7 and 70 percent, respectively. These solid transfer learning algorithms are always evolving, and there is still room for more research.

2.5 Research Highlights and Pitfalls

To conclude the literature review section, this study has studied and assimilated various data modelling research methodologies and data preparation techniques that have been implemented on various levels and for estimating the calories count, several papers relating to calibration techniques and bounding box techniques has also been taken into account. The strengths and flaws of all these discussed works have been identified and stated. Since this study has adopted the architecture of InceptionNet v3 based on augmented data processing, it had been ensured that there is no previous work that falls

under the same category. Table 1 depicts the key findings, advantages and limitations of the most pivotal works relevant to this study.

Table 1: Summary of Important Related Works

Authors	Deep Learning Model	Advantages	Limitations
Rajayogi et al. (2019)	CNN	Accuracy - 99%	Feature Selection not performed
Sengur et al. (2019)	SVM	Sensitivity - 86%	Hyperparameter tuning not considered
Sun et al. (2014)	3D-RCNN and Segmentation	Precision - 97.45%	Relatively less data and no augmentation
S. Ren et al. (2017)	Faster-RCNN	Accuracy - 93%	Evaluation parameters such as sensitivity, specificity not accounted for.
Oner et al. (2019)	YOLO	Better classification	Data pre-processing not stated and hence imbalance in dataset
Li et al. (2018)	DenseNet and Faster-RCNN	Sensitivity - 98.7%	Unrealistic accuracy rates and thus indicating model overfit
Ragesh et al. (2019)	ResNet and YOLO	Accuracy - 70%	Relatively lower sensitvity and accuracy and leading to failure in classifications

3 Methodology

This section outlines the various data analytics mining techniques used for extracting the dataset and the kind of methodologies and deep learning modelling techniques that are used within this study . Business understanding, understanding the data composition, preparation of data, data wrangling, data augmentation, feature extraction, and modelling the pre-processed images subject to various deep learning algorithms are detailed steps associated with research methodology. This section also encompasses the design specifications for the architecture of the project and the overall project workflow pipeline.

3.1 Business Understanding and Data Acquisition

The foremost before starting to work on the project implementation is to comprehend the numerous business aspects and financial intricacies of the project. Dietary control and monitoring is critical for humans in order to live a healthy life and have a longer life expectancy. The classification and identification of the ingredients used in our food items for each meal has become critical. Food segmentation and detection of portions of food aid in calculating the calorie representation of each food item and hence, helping to identify the overall calorie intake total. The dataset of food items based on 101 categories of food[1]

[1]Food images dataset: `https://www.kaggle.com/kmader/food41`

is extracted from Kaggle and made available as a public dataset for research purposes. The image classification dataset is made up of images that have been downscaled to a lower resolution to allow for faster processing. The dataset is made up of 101 categories of food, with each category containing 1000 images on its own, adding up to 101,000 (101 * 1000) images with a standard resolution of 384 x 384 x 3 and the images are denoted in RGB scale.

3.2 Tools and Equipment to be Used

Jupyter notebook has been used as the integrated development editor (IDE) for the implementation of cutting-edge deep learning models because it allows interactive code execution on the browser, and Python was the language opted for coding because it has a plethora of deep learning and random image generator libraries. The primary libraries used for performing the deep learning models and transfer learning algorithms are OpenCV, Tensorflow, and Keras, and they have extensive support out of the box helping for many open source transfer learning models and data augmentation methods which help in carrying out feature extraction and detecting objects in an image using object detection techniques.

3.3 Data Preparation and Pre-Processing

In order to ensure the dataset is fit enough for modelling deep learning algorithms, the dataset must go through a series of cleaning and preparation steps. Because transfer learning algorithms are tested and tuned to perform optimally with grayscale images, the images are encoded in a 3-channel RGB scale and then modified and converted to 2-channel grayscale images to achieve faster processing and better performance. The maximun threshold limit of pixels for transfer learning algorithms is 224 * 224 pixels, and image dimensions are standardised to this size.

The validation approach divides the dataset into train and test data, with 70% of the images collectively clustered into training data and the rest 30% is clustered and grouped into test data. Images are not associated with any labelled classes are removed, and the images are standardised by performing zca-whitening, setting width, height, zoom range, scaling, shearing, and flipping. Python and Jupyter notebook is used to carry out these steps.

3.4 Modelling Algorithms Used

As the overall implementation of the project has two sub-parts, there are two subsections under the proposed implementation section. Part one consists of the steps involved in executing extraction of features based on previously trained models, and the various image features are converted into vectors of binary patterns for each food item and the results of this prepared dataset is fed to the second step. Part two includes techniques for recognising food compositions of the dish and calculating their appropriate food proportion sizes based on bounding box and calibration techniques.

3.4.1 Transfer Learning and Feature Extraction Models

Transfer learning is a contemporary and powerful approach in which models that are already trained on a large number of existing image datasets serve as the starting point

for the implementation of the new model that is built. This method is used to extract features based on colour, co-ordinates, shape, size and for various food categories. This type of implementation reduces the execution time, consumption of memory overhead, and fastens the execution speed. The accuracy is also dramatically improved because the models are already trained on massive amounts of data, and the cutting-edge deep learning models that has been used in this study for the execution of the model are InceptionNet, VGG, and ResNet-50.

3.4.2 VGG Model

One of the evolutionary transfer learning models was the VGG and it was suggested by Liu and Deng (2015). The VGG has 2 variations of its architecture namely VGG-16 and VGG-19 and the numbers 16 and 19 represented the different pooling layers, flattening, and dropout layers associated with the model. Figure 1 showcases the architecture of a VGG model. This model achieved the top rank in the ILVSR classification of the ImageNet dataset and clocked the lowest error rate. The model is convoluted with 5 max-pooling layers thus ensuring lesser spatial resolution. The 5 max-pooling layers are further attached to 3 fully outer connected layers and the first 2 layers have a channel count of 4096, while the last layer has exactly 1000 channels.

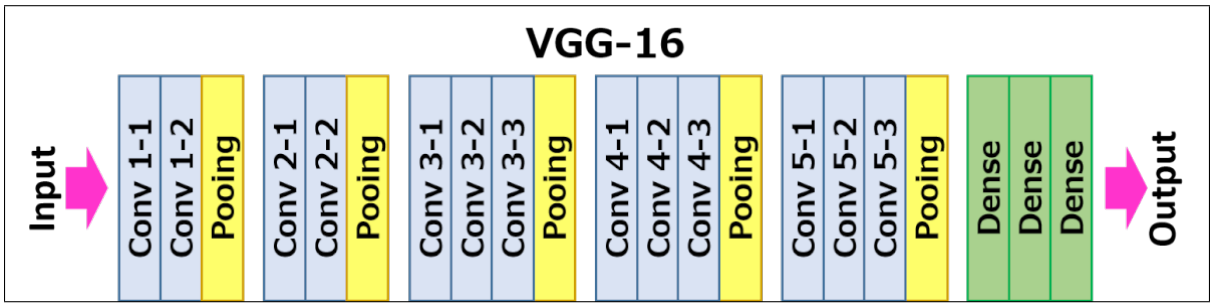

Figure 1: VGG Architecture

3.4.3 InceptionNet V3 Model

The second transfer learning model that is considered for this study is Google's Inception-Net v3 model and it was built by Szegedy et al. (2016) in 2015. The main breakthrough of this architecture was how it utilized less amount of computational resources to provide powerful models. This architecture constitutes 9 inception layers which are stacked in a linear fashion up to 22 layers and culminate into the final average global pooling layer at the end. The InceptionNet architecture is depicted in Figure 2 and it shows the auxiliary classifiers at the final phase to help in preventing the process from dying out as a result of batch normalization.

3.4.4 ResNet Model

Another convolution neural network architecture that turned out to be pretty successful was the ResNet 50 architecture. This model was devised in 2015 and built by He et al. (2016a) and this piece of architecture was based upon the concept of Residual Blocks and topped the COCO competition with a lowest top-5 error rate of 4.12%. The inclusion of

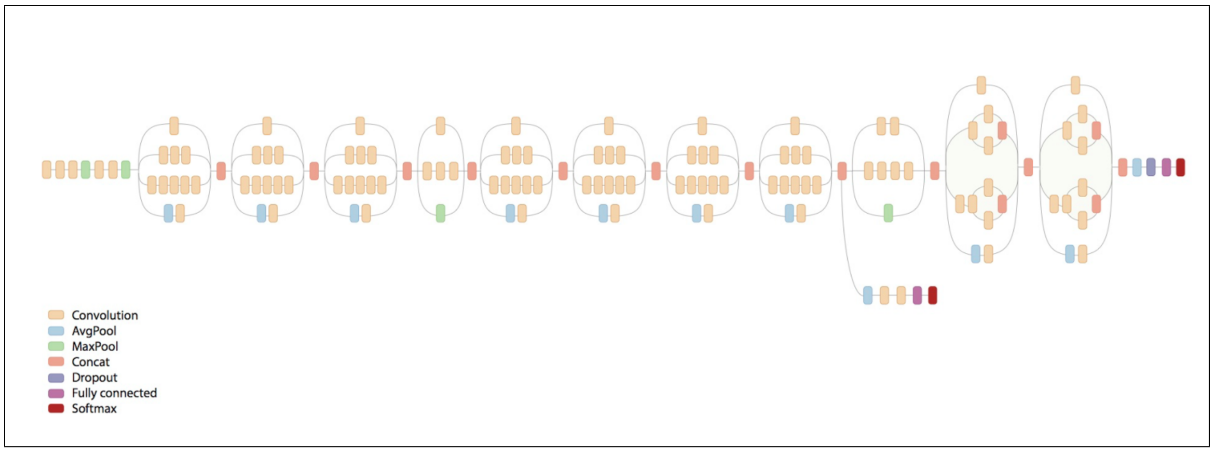

Figure 2: InceptionNet Architecture

residual mappings was aimed at eliminating the problem of vanishing gradient. Fig. 3 represents the architecture of ResNet 50 and it showcases the 50 deep layers associated with this architecture. This model accepts input of RGB nature with dimensions of 32 * 32 and has convolutional kernels of 7 * 7 and 3 * 3 followed by deep convolutional connected network layers. Similar to VGG, the last output layer in this architecture also has 1000 channels and this uses Softmax as the activation layer.

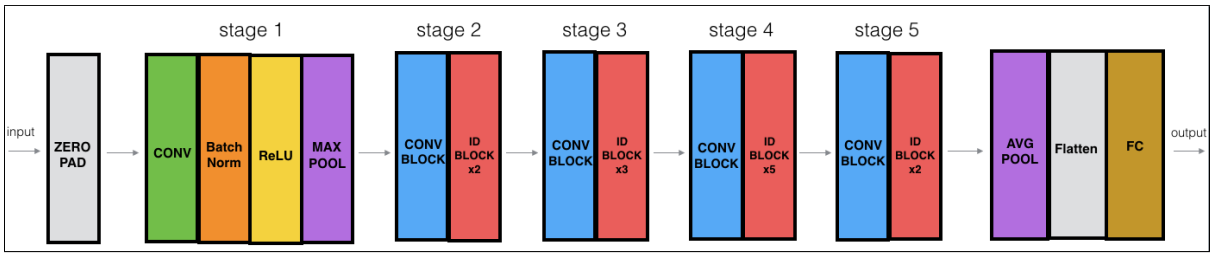

Figure 3: ResNet Architecture

4 Design Specification

The project architecture for this study is depicted in Figure 4 as a three-tier architecture. The food classification and calorie estimation measurement system is made up of three layers: the Presentation Layer, the Business Logic Layer, and the Application Layer, here the application layer is a web application X. Liu and Sha (2005). The logical layer includes the processes of acquiring and gathering data, preparing, pre-processing data, shearing, transforming, flipping, extracting features and splitting data into train and test. The images are then subjected to further segmentation and then fed to the database layer based on colour, texture, and graph cuts.

The database layer includes persisting the count of for each food ingredient in the database, and Google Drive or AWS S3 buckets is used for storing the images in the cloud based on the wide variety of food categories. The core implementation takes place in the application layer. Feature extraction of data is performed based on one-hot encoding

and getting the results as binary vectors of multiple features instead of one single feature and this suitable cleansed dataset is used for performing convolutional neural networking algorithms and transfer learning models such as ResNet, InceptionNet and VGG. The visual medium is the presentation layer, and in this scenario, a web application is deployed on the cloud and the users will be able to access the web URL on the browser and upload the photo of their intake meal. The uploaded image would be passed through the transfer learning data model and the resulting output would display the various constituents of the meal.

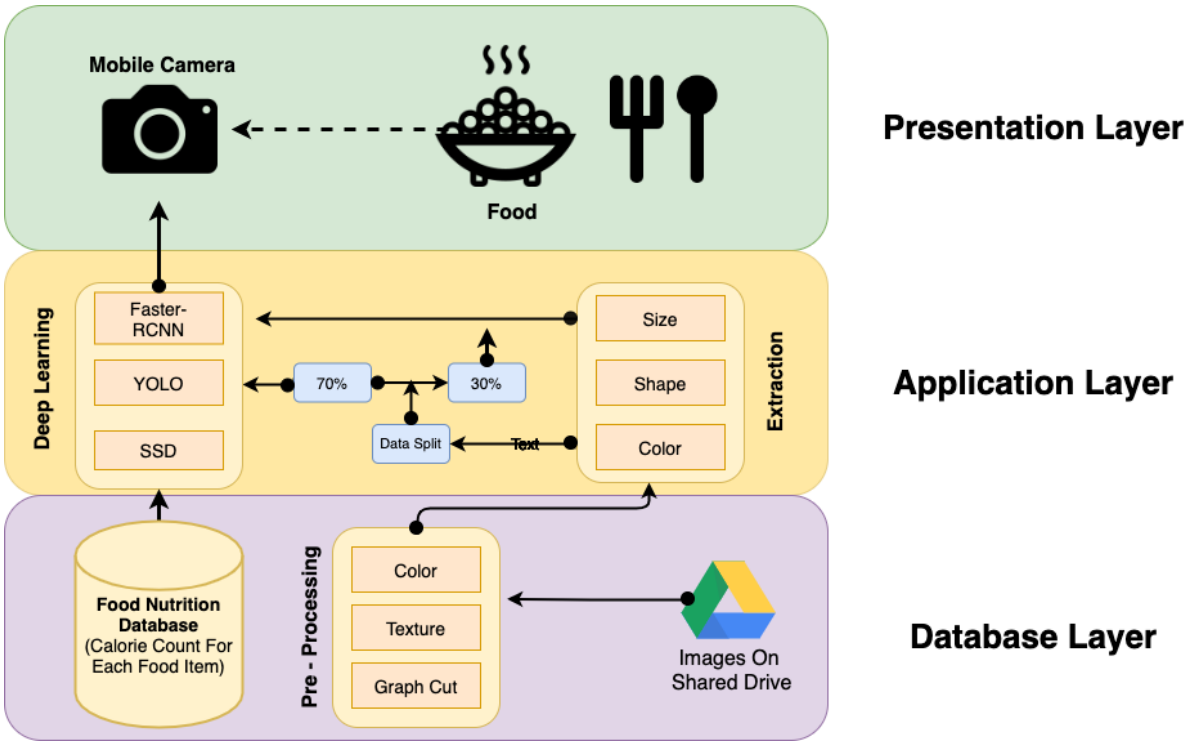

Figure 4: Project Architecture

4.1 Project Work-Flow Structure

The workflow pipeline of the process is depicted in Figure 5. The workflow has 3 separate sequential layers, the first step being the data preparation followed by segmentation of data and in this phase, the images are augmented, feature extraction is performed and then image segmentation takes place. In the final phase, the images are fed to the transfer learning model namely InceptionNet, VGG and ResNet-50 and the resulting output predicts the different constituents of food present in the image. These constituents are then calculated for their appropriate food portions and then these the calorie count of these food portions are retrieved from the calories database. The calorie counts of each food item is multiplied with their food portions and the overall count of calories is found by summing up all the respective food proportions and displayed on the visual medium by means of a web browser.

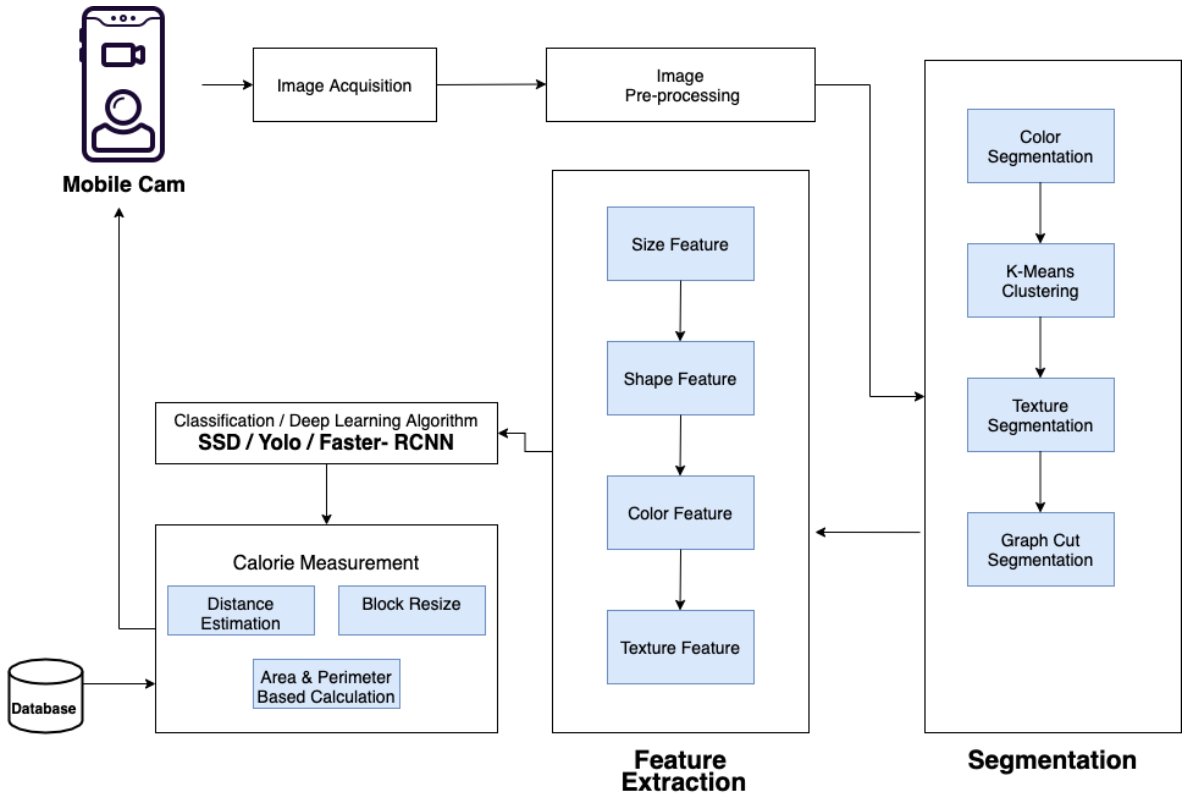

Figure 5: Project Workflow

5 Implementation

5.1 Development Environment

This project uses a wide range of deep learning tools for the execution of the best performing deep learning model to achieve food image classification. The implementation is carried out in a Jupyter notebook (.ipynb format) and uses Python version 3.9 for development and performing the sequence of pre-processing steps and data augmentation. Numpy[2] and Pandas[3] are used for analyzing the type of dataset involved and for performing data manipulation, wrangling, and data loading. Keras[4] and Tensorflow[5] are the deep learning libraries used and they have a wide variety of transfer learning models such as ResNet, DenseNet, VGG, and InceptionNet already pre-loaded within them. For exploratory data analysis, visualization tools and libraries such as Seaborn[6], Matplotlib[7], and Plotly were used and for performing feature selection and dimensionality reduction, SciKit[8] was used.

[2]Numpy: `https://numpy.org/`
[3]Pandas: `https://pandas.pydata.org/`
[4]Keras: `https://keras.io/api/applications/`
[5]Tensorflow: `https://www.tensorflow.org/`
[6]Seaborn: `https://seaborn.pydata.org/`
[7]Matplotlib: `https://matplotlib.org/`
[8]Sci-kit: `https://scikit-learn.org/stable/`

5.2 Data Augmentation and Feature Extraction

In data augmentation 6, the images in the dataset are subjected to a series of steps in order to eliminate all the unwanted and distorted detail in the image and to increase the speed of prediction of the deep learning algorithm by reducing the resolution of the image (Pereira et al.; 2016). The ImageData Generator which is available as part of the Tensorflow library is used for augmenting the image and the various attributes that are fixated for helping in transforming the image to the required dimensions are zca whitening, rotation, width, height setting, horizontal and vertical flip, scaling and zoom range. The validation approach that has been opted for splitting is the dataset is the train/test validation method and the ratio of the split is 80:20.

In feature extraction, the images in the food categories dataset are converted into vectors of binary features rather than having one single feature. The more the number of features available to the deep learning model, the better it is for predicting the most accurate result. The next step involves building multiple versions of the model by tuning and tweaking the hypertuning parameters (Canny; 1986).

5.3 Modelling Implementation

Once the images have been wrangled, augmented and the features have been extracted, Google's InceptionNet v3 which is a pre-trained and weighted transfer learning model based on the ImageNet dataset is used for model building. The hyper-tuning parameters are tweaked to accommodate for the average global spatial pooling layer, a fully connected second layer, dropout and flatten layers are added (De Guia et al.; 2019). The activation layer is set as Softmax as it is highly suited for these kind of images, the activation loss function is categorical entropy and the optimizer is Adam optimizer (Mehta et al.; 2019). This avoids any probability of over-fitting occurring in the model. On top of this base model, the stochastic gradient descent (SGD) is implemented with a learning rate of 0.01 and this quickly decreasing learning rate is used for achieving high accuracy.

The same base model built in the above step is used for building another deep learning model and this time the image is subjected to 10 different rotations namely upper right, lower right, upper left, lower left and center. This results in 10 different output predictions for the same image and most common prediction is considered as the final prediction outcome. This second model significantly outperformed the first model and the batch size for this model was set to 100 with 45 determined as the epoch cycle count.

5.4 User Interface Implementation

This is the final phase of the coding implementation of this research study. An intuitive web application is built using React.js which is a frontend framework for developing single page web applications (SPAs). Using this user interface, the end user would be able to select any test image and cross check the validity of the results by viewing the output of the predictions given by the InceptionNet v3 model and a snapshot of the figure7 is shown.

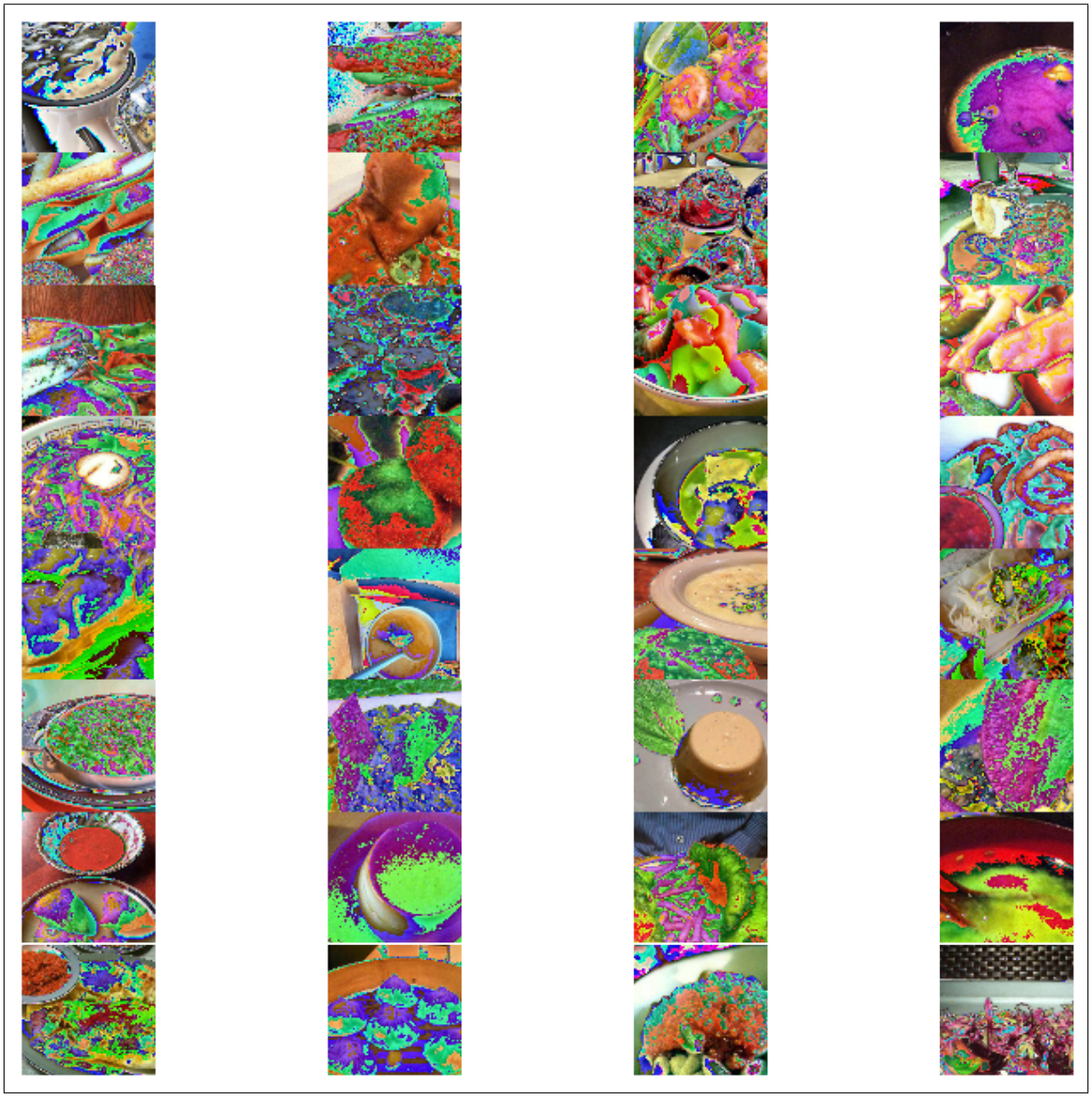

Figure 6: Augmented Image Data

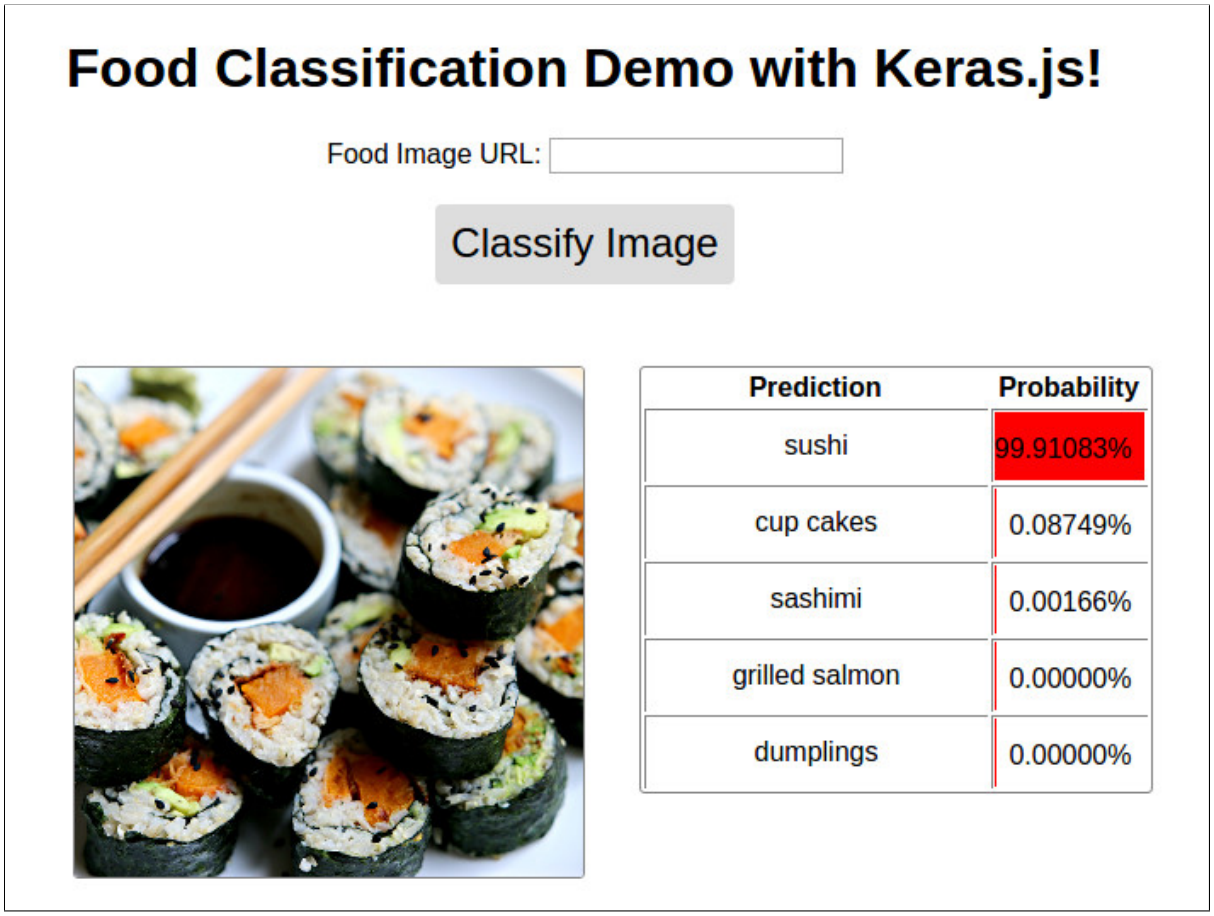

Figure 7: Web UI Interface

6 Evaluation

In the evaluation section, the comparison of the 2 InceptionNet v3 models is carried out and one model performs just a single prediction for the augmented data whereas the second model carries out the predictions on a set of 10 different image variations for each single image. Accuracy and validation loss are considered to be the 2 most pivotal evaluation metrics for establishing which model considerably performs better over the another (He et al.; 2016b). In the first InceptionNet model, an accuracy of 81.57% was achieved and the validation and training loss were 1.477 and 1.321 respectively. For the second variation of the InceptionNet model, a relatively high accuracy of 86.97% was achieved and 1.081 and 0.907 were the validation and training loss metrics (Liu et al.; 2017). From the stated evaluation figures, it is evident that the second model built is significantly better than the first and this model outperforms all the state-of-the-art existing deep learning models that have been researched in related works. The figures 10,11, 12 and 13 represent the accuracy plots and validation loss plots for both the models 1 and 2 respectively.

6.1 Experiment / Case Study 1

In the first adaptation of the InceptionNet v3 model, all the 101 categories of the food images are fed into the model. The batch size is set to 100 and the epoch cycles is set to 45 and this results in a computationally exhaustive algorithm and a learning rate scheduler was adopted to fine tune the results of the model. As visible from Fig. 10 and Fig. 11, an accuracy of 81.57% was achieved and the validation and training loss were 1.477 and 1.321. In Fig. 8, the class prediction accuracies for each of the 101 classes is depicted and it can be seen that certain classes have higher prediction rates while certain classes have lower prediction accuracy rates. Hence, the next model is iterated using higher number of features and higher number of images for balanced prediction accuracies across all the 101 class categories.

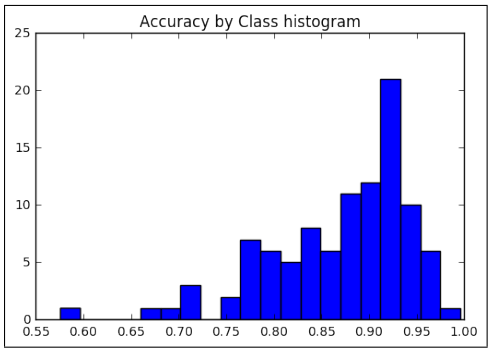

Figure 8: Histogram of class predictions

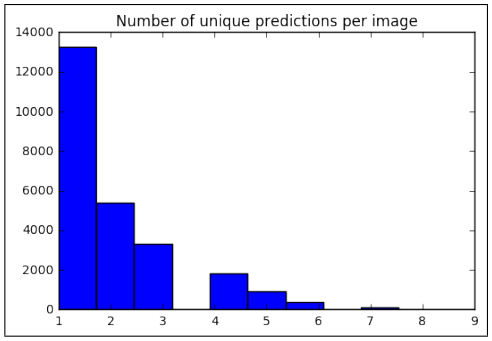

Figure 9: Unique predictions - model 2

6.2 Experiment / Case Study 2

In the second model, the InceptionNet v3 is still used as the base model, but in this step every image is subjected to a set of 10 rotations namely upper left, upper right, bottom

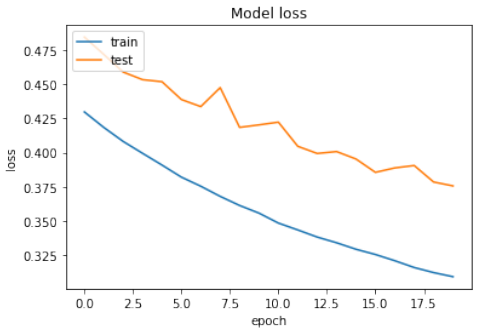

Figure 10: Accuracy plot - model 1

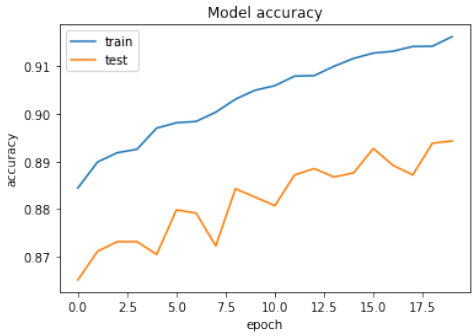

Figure 11: Loss plot - model 1

left, bottom right and center. These different illustrations of the same image help in detecting multiple food objects located different co-ordinates across the image. So, when a test image is uploaded for checking class prediction, there were 10 resulting predictions. And, the most repeating and common prediction out of the 10 predictions was taken as the final prediction. The number of unique predictions for each class is represented in Figure. 9. This second variation of the model helped achieve relatively high accuracy of 86.97% and 1.081 and 0.907 were the validation and training loss metrics. Hence, this is the best model achieved by subjecting the given images to a series of steps and the class prediction accuracies and loss thresh-holds was also evenly balanced across all the 101 categories in this model.

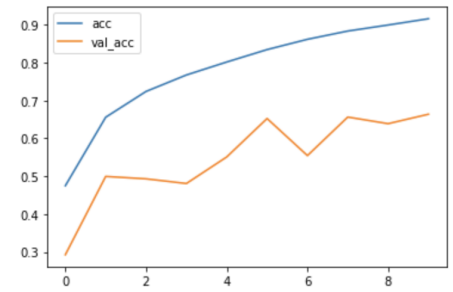

Figure 12: Accuracy plot - model 2

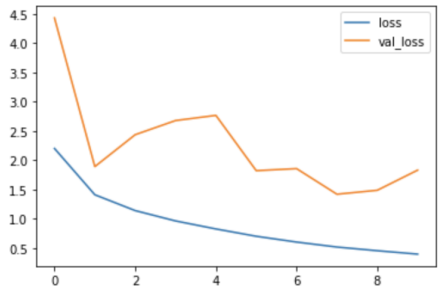

Figure 13: Loss plot - model 2

6.3 Discussion

In this research study, the emphasis has been on achieving the most balanced deep learning model that ensures there is no overfit with the training data as well as ensuring the balanced class prediction accuracies across all the 101 food category images. There have been 2 models built and each model exhibited different advantages and limitations. In the first model, the implementation was performed after subjecting the image to data segmentation and augmentation, resulting in a pretty decent accuracy of 81% but it failed to meet the validation loss metric. The validation loss was beyond tolerable range and this required the development of a second data model.

In the second model, the model was built by subjecting each image to a series of rotations and cropping each part in the image across different axes to arrive at numerous

predictions. The predictions were culminated and collectively summarized to decide the most optimal class category of the image. This model proved to perform very well and detected multiple food items in the images with fairly high accuracy of 87% and the validation losses were also minimal. However, 40 epoch cycles were used to arrive at this resulting model and hence, this also proves to be computationally expensive. Even though this model is much better than all the existing relevant works, there is still room for improvement as the hypertuning parameters can be revisited and revised to arrive at an even better model with lesser epoch cycles and faster execution rates.

7 Conclusion and Future Work

This purpose of this research paper was to address the important problem of recognizing food, identifying the food constituents, computing the calories and estimating the calories consumed by the individual in each daily meal. The proposed algorithm which is developed in InceptionNet v3 was able to achieve high accuracy rates and correctly classify the food images based upon the 101 categories of food images present in the dataset. This leads to answering the research question **"How accurately can object detection algorithms recognize food and accurately estimate the calories of the food items to help people with obesity and health-ailments?"**. The important evaluation metrics considered were accuracy and loss factors and the model was tweaked iteratively for significant gains in the prediction outcome. A web user interface was also developed for users to calculate their daily estimated calorie intake on a day-to-day basis. Hence, this study has thoroughly scrutinized, detailed and examined all the proposed approaches and the resulting developed model is fit for usage of dieticians/doctors and patients to solve the needs of obese patients and help in keeping the increasing health ailments and health conditions under check.

In future works, identifying the estimated calorie count using the bounding box technique will be implemented to accurately predict the number of calories consumed by an individual for their daily meal. In order to compute the calories of each food constituent, the areas of the food portions has to be effectively cropped and detected food portions need to be calibrated based on their distance. This requires intensive computation and highly efficient hardware, so this would be taken up in the next implementation phase where the calibration technique would be used to effectively calculate the estimated number of calories consumed and this will in turn lead to a more efficacious and self-sustaining system for future usage across the health sector.

Acknowledgement

I'm conveying my heartfelt thanks to my research supervisor and mentor Dr. Majid Laifi for guiding me through the research project submission. I would also extend my appreciation to all the NCI faculty members and my peer students for helping me and guiding me all along the way. I would like to thank my parents, friends and peers for their unwavering moral and financial support throughout my entire journey.

References

Canny, J. (1986). A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-8**(6): 679–698.

Chen, Y., Jiang, H., Li, C., Jia, X. and Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks, *IEEE Transactions on Geoscience and Remote Sensing* **54**(10): 6232–6251.

Dahiya, S., Puri, S. and Singh, S. (2021). Image segmentation techniques: A survey, *International Journal of Engineering and Applied Physics* **1**(2): 127–135.
URL: *https://ijeap.org/ijeap/article/view/26*

De Guia, J. D., Concepcion, R. S., Bandala, A. A. and Dadios, E. P. (2019). Performance comparison of classification algorithms for diagnosing chronic kidney disease, *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pp. 1–7.

Fernandes, F., Vicente, H., Abelha, A., Machado, J., Novais, P. and Neves, J. (2015). Artificial neural networks in diabetes control, *2015 Science and Information Conference (SAI), 2015* pp. 362–370.

He, K., Zhang, X., Ren, S. and Sun, J. (2016a). Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

He, K., Zhang, X., Ren, S. and Sun, J. (2016b). Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C. W. and Heng, P. A. (2018). H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes, *IEEE Transactions on Medical Imaging* **37**(12): 2663–2674.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017). Feature pyramid networks for object detection, *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **30**(11): 936–944.

Liu, S. and Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size, *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734.

Liu, X., Kumar, B. V. K. V., You, J. and Jia, P. (2017). Adaptive deep metric learning for identity-aware facial expression recognition, *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 522–531.

M., P. and C., Y. (2019). Food and therapy recommendation system for autistic syndrome using machine learning techniques, *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–6.

Mehta, S., Paunwala, C. and Vaidya, B. (2019). Cnn based traffic sign classification using adam optimizer, *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1293–1298.

Moorhead, A., Bond, R. and Zheng, H. (2015). Smart food: Crowdsourcing of experts in nutrition and non-experts in identifying calories of meals using smartphone as a potential tool contributing to obesity prevention and management, *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* pp. 1777–1779.

Oliveira, L., Costa, V., Neves, G., Oliveira, T., Jorge, E. and Lizarraga, M. (2014). A mobile, lightweight, poll-based food identification system, *Pattern Recognition* **5**: 1941–1952.

Oner, M. S. A., Guner, F. and Atakli, I. M. (2019). An activity recognition application based on markov decision process through fish eye camera, *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* pp. 251–258.

Peddi, S., Kuhad, P., Yassine, A., Pouladzadeh, P., Shirmohammadi, S. and Shirehjini, A. (2017). An intelligent cloud-based data processing broker for mobile e-health multimedia applications, *Future Generation Computer Systems* **66**: 71–86.

Pereira, S., Pinto, A., Alves, V. and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in mri images, *IEEE Transactions on Medical Imaging* **35**(5): 1240–1251.

Ragesh, N., Giridhar, B., Lingeshwaran, D., Siddharth, P. and Peeyush, K. P. (2019). Deep learning based automated billing cart, *2019 International Conference on Communication and Signal Processing (ICCSP)* pp. 0779–0782.

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

S. Ren, K. H., Girshick, R. and Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6): 1137–1149.

Sengur, A., Akbulut, Y. and Budak, U. (2019). Food image classification with deep features, *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–6.

Sun, Y., Wang, X. and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* p. 1891.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016). Rethinking the inception architecture for computer vision, pp. 2818–2826.

Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y. and Yan, S. (2016). Learning to segment with image-level annotations, *Pattern Recognition* **59**: 234–244. Compositional Models and Structured Learning for Visual Recognition.
URL: *https://www.sciencedirect.com/science/article/pii/S0031320316000364*

X. Liu, J. H. and Sha, L. (2005). Modeling 3-tiered web applications, *13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems* pp. 307–310.

Zhao, Z., Zheng, P., Xu, S. and Wu, X. (2019). Object detection with deep learning: A review, *IEEE Transactions on Neural Networks and Learning Systems* **30**(11): 3212–3232.