Deep FoodRec: Food Recognition Technology using Computer Vision and Deep Learning for Health Monitoring

Submitted by

Mazhar Hussain

Supervised by

Prof. Sebastiano Battiato
Prof. Alessandro Ortis

PHD Thesis, Academic Years 2020-2023
(XXXV Cycle)

UNIVERSITY OF CATANIA

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF CATANIA

CATANIA, ITALY.

April 2023

Author's Declaration

I, Mazhar Hussain, hereby state that my PHD thesis titled *Deep FoodRec: Food Recognition Technology using Computer Vision and Deep Learning for Health Monitoring* is my own work and it has not been previously submitted by me for taking partial or full credit for the award of any degree at this University or anywhere else in the world.

_____

Mazhar Hussain

Date: _____

Abstract

The food recognition project (FoodRec) aims to define an automatic framework using computer vision and deep learning techniques to recognize diverse foods from images. The goal of food recognition is to extract and infer semantic information from the food images and to classify different foods present in the image. The developed system acquires images of the food eaten by the user or subject over time, which will then be processed by food recognition algorithms to extract and infer semantic information from the images containing food. The extracted information will be exploited to track and monitor the dietary habits of people involved in a smoke-quitting protocol.

Food recognition is an active research area due to its wide range of potential real-world applications. For example, it would allow people to track their food intake of what they consume by simply taking a picture, to increase the awareness in their daily diet by monitoring their food habits, the kind and amount of taken food, how much time the user spends eating during the day, how many and what times the user has a meal, analysis on user's habits changes, bad habits and other inferences related to user's behavior. It can help a doctor to have a better opinion with respect to the patient's behavior, quitting treatment response, and hence his health needs.

This project involves several stages from image acquisition to recognition. In particular, the efforts are devoted to the development of new segmentation and recognition algorithms to perform the food recognition task accurately. This PhD thesis presents semantic food segmentation and recognition using deep learning techniques. The proposed approaches have been developed in the context of the FoodRec project, which aims to define an automatic framework for the monitoring of people's health and habits, during their smoke-quitting program. The aim is to extract and infer semantic information from the food images to analyze diverse foods present in the image.

We introduce a new FoodRec-50 dataset with 50 food categories collected by the iOS and Android smartphone applications, taken by 164 users during their smoking cessation therapy. Data preprocessing, data annotations, and data augmentation with different transformations are performed for further processing after the data has been collected by the application. For food recognition, we propose a Deep Convolutional Neural Network able to recognize food items of specific users and monitor their habits. It consists of a food branch to learn visual representation for the input food items and a user branch to take into account the specific user's eating habits. Experimental results show that the proposed food recognition method outperforms the baseline model results on the FoodRec-50 dataset. We also performed an ablation study, which demonstrated that the proposed architecture is able to tune the prediction based on the users' eating habits. For food segmentation, we propose a novel Convolutional Deconvolutional Pyramid Network (CDPN) for food segmentation to understand the semantic

information of an image at a pixel-level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. Furthermore, the proposed method demonstrated significant improvements on two well-known public benchmark food segmentation datasets. We propose another Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images at a pixel-level to recognize different food items present in an image. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns, while the proposed CDPN also uses interpolation for features upsampling along with the deconvolution layers. Our proposed network demonstrated significant improvements in the results on the benchmark food dataset as compared to the state-of-the-art methods. Additionally, we also conducted a cross-data qualitative analysis of our proposed segmentation method to assess its generalization capabilities on our FoodRec dataset.

The research outcomes of the food recognition include 2 journals and 3 conference papers. This project is a research grant where I collaborated to develop food recognition algorithms. This research was sponsored by ECLAT srl, a spin-off of the University of Catania, with the help of a grant from the Foundation for a Smoke-Free World Inc., a US nonprofit 501(c)(3) private foundation with a mission to end smoking in this generation. My contributions in this project include data preprocessing, data annotations, and data augmentation for further processing after the data has been collected by the application. Then, to study, develop, and evaluate algorithms using computer vision and deep learning techniques to track and monitor the dietary habits of people.

The additional research work is collaborated with the Department of Drug and Health Science, University of Catania. The aim of the work was to find a correlation between well-defined and selected parameters such as the type of nanocarrier, the particle size and the surface charge, and the targeting efficiency indexes %DTE and %DTP. We performed nose-to-brain drug delivery data cleaning, conversion, standardization, and classification using state-of-the-art machine learning algorithms. We are working to publish a journal from this work with the collaboration of the Department of Drug and Health Science.

Acknowledgements

First and foremost, I would like to thank the ALLAH Almighty for reasons too numerous to mention and beyond words to describe. He has given me the power to stay motivated all the time and pursue my objective. I could never have done this without the faith I have in the ALLAH Almighty.

I would like to express my sincere gratitude to my advisors Prof. Sebastiano Battiato, and Prof. Alessandro Ortis, University of Catania, Department of Mathematics and Computer Science, for providing me the encouragement, logistic support, advice, enduring patience, constant support, continuous guidance and valuable suggestions.

I am also thankful to all of our teachers and friends who had been guiding me throughout my work. Their knowledge, guidance enabled me to complete the work efficiently. There are also so many other people that I cannot thank the explicitly. God bless them.

Dedication

This thesis is dedicated to my parents, who support me in every stage of my life, to my teachers who made me able to achieve milestones which I wish for and to my friends.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

FoodRec Food Recognition
CDPN Convolutional Deconvolutional Pyramid Network
CNN Convolutional Neural Network
DNN Deep Neural Network
FCN Fully Convolutional Network
FPN Feature Pyramid Network
EDFN Encoder Decoder Food Network
ResNet Residual Network
PSPNet Pyramid Scene Parsing Network
KNN K-Nearest Neighbor
SVM Support Vector Machine
RCNN Region-Based Convolutional Neural Network
PSPNet Pyramid Scene Parsing Network
HoG histogram of oriented gradients
LBP Local Binary Patterns
BoF Bag of Features
SIFT Scale Invariant Feature Transform
RF Random Forest
MKL Multiple Kernel Learning
SE Squeeze and Excitation
GAN Generative Adversarial Network
PRENet Progressive Region Enhancement Network
FSFW Foundation for a Smoke Free World
IoU Intersection over Union
mIoU Mean Intersection over Union

Chapter 1

Introduction

1.1 Overview

Food recognition is an area of research that deals with the development of computer vision systems that can automatically identify and classify food items in images and videos. With the rise of digital cameras and smartphones, food images have become ubiquitous on the internet, making food recognition an essential and practical problem for many applications such as calorie counting, food delivery, and recipe recognition. This technology aims to provide a more accurate, efficient, and automated way to identify food items by analyzing food images and extracting relevant information. This field of research is highly interdisciplinary, drawing on expertise from computer vision, image processing, machine learning, deep learning, and nutrition science. This research aims to explore the current state of the art in food recognition including traditional food recognition techniques, deep learning-based methods, and to develop new algorithms and techniques that can improve the accuracy and robustness of food recognition systems.

Traditional food recognition techniques rely on a combination of feature extraction and classification techniques. These techniques are based on the idea to extract certain visual characteristics from food images such as color, texture, and shape, that can be used to distinguish different food items. The process of food recognition typically involves extracting these features from food images and then using them to train a machine learning model that can classify new images. During feature extraction, different visual features such as color, shape, and texture are extracted (Anthimopoulos et al., 2014). Color histograms (Kawano and Yanai, 2013) feature extraction technique is used in food recognition that involves dividing an image into small regions and extracting the color histogram of each region. Color histograms represent the distribution of colors in an image and can be used to capture the color characteristics of

food items. Feature extraction using texture (Hoashi et al., 2010) analysis involves extracting texture features from images, such as the distribution of intensity values and the spatial distribution of texture patterns. Texture analysis can be used to capture the texture characteristics of food items, such as the roughness of a surface or the presence of patterns. Shape analysis (Phetphoung et al., 2014) is also a feature extraction technique used in food recognition that involves extracting shape features from images, such as the size, shape, and position of objects. Shape analysis can be used to capture the shape characteristics of food items, such as the presence of round or rectangular shapes. To achieve an optimal feature extraction process, it is essential to extract informative visual data from food images. This data can be obtained through descriptors that collect various basic features, such as color, texture, and shape. Examples of such descriptors include histogram of oriented gradients (HOG) (Kawano and Yanai, 2014), local binary patterns (LBP) (Chen et al., 2012), bag-of-features (BoF) (Hoashi et al., 2010), scale-invariant feature transform (SIFT) (Anthimopoulos et al., 2014), and Gabor filter (Pouladzadeh et al., 2015) which can be applied individually to capture image features. Once the features have been extracted, they can be used to train a machine learning model. There are many traditional machine learning techniques have been applied for food recognition (Joutou and Yanai, 2009; Vivek et al., 2018; Kumar et al., 2021; Jiménez-Carvelo et al., 2019; Munira Shifat et al., 2022; Giovany et al., 2017) and popular machine learning algorithms have been used for food recognition, such as k-Nearest Neighbors (KNN) (He et al., 2014; Hemamalini et al., 2022), Random Forests (RF) (Bossard et al., 2014), Support Vector Machines (SVMs) (Bosch et al., 2011; Yang et al., 2010; Sharma et al., 2022), and Multiple Kernel Learning (MKL) (Matsuda et al., 2012; Hoashi et al., 2010). These algorithms can learn to recognize food items based on the features that have been extracted from food images.

Then paradigm shift and change in perspective brought by deep learning (Stevens et al., 2020) is shown in Figure 1.1, where we can observe two distinct approaches for feature engineering in learning a model. On the left, a practitioner is manually defining engineering features and then feeding them into a learning algorithm. The effectiveness of this approach is heavily reliant on the quality of the features that the practitioner engineers can achieve. In other words, the performance of the learning algorithm will be limited by the practitioner's ability to select informative features. On the right side of the Figure 1.1, we see an alternative method where deep learning algorithms automatically extract hierarchical features from raw data. This process involves optimizing the performance of the algorithm on the task at hand without any explicit feature engineering. The accuracy of this approach is heavily reliant on the practitioner's ability to drive the algorithm towards its objective. This involves selecting appropriate training data, tuning hyperparameters, and designing the architecture of the model.

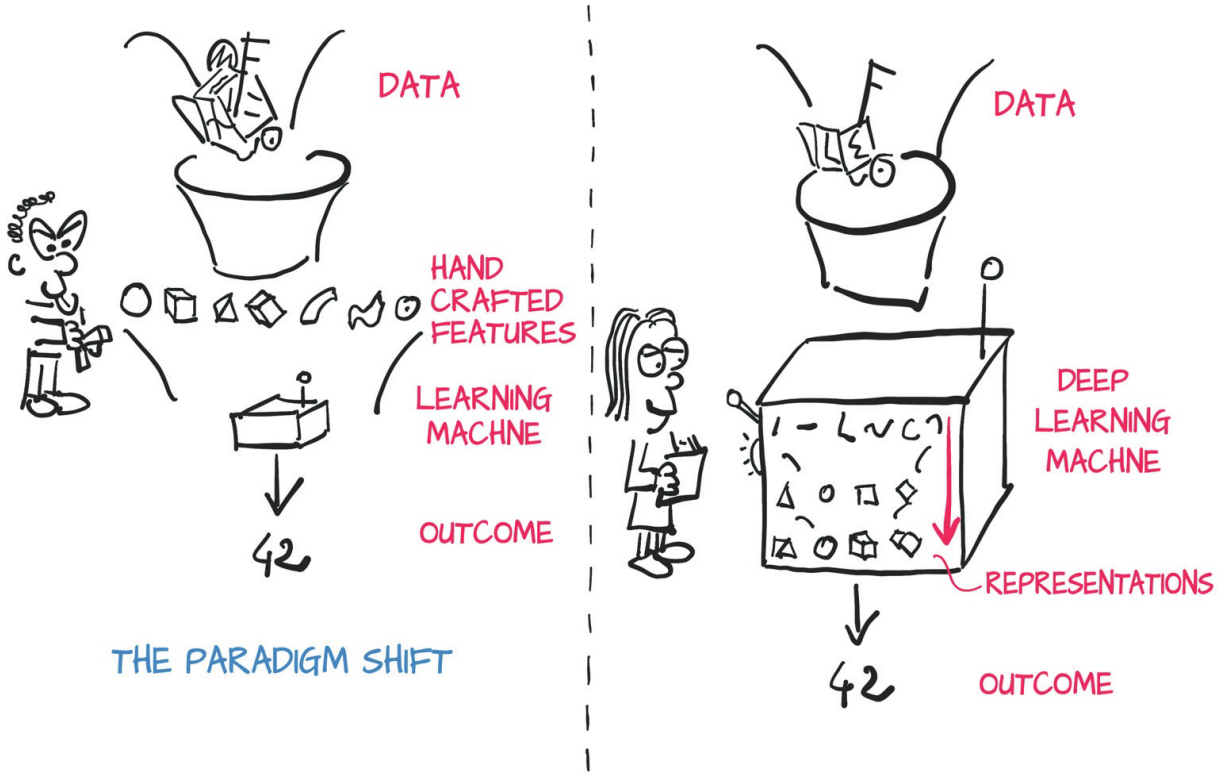


Figure 1.1: The paradigm shift from machine learning to deep learning and change in perspective brought by deep learning (Stevens et al., 2020)

Deep learning has emerged as a powerful tool for solving complex problems in computer vision, including the food recognition. In recent years, deep learning-based methods have achieved state-of-the-art performance for food image analysis on a wide range of food recognition tasks, including food classification, food detection, and food segmentation. Deep learning uses a built-in mechanism to automatically extract features through a series of connected layers and a fully connected layer for final classification. It has gained popularity due to its exceptional performance and outstanding classification abilities compared to traditional machine learning methods. Convolutional neural network (CNN) (LeCun et al., 2015) is a prominent deep learning technique widely preferred in computer vision applications, particularly for classification of large-scale image data. CNN has outperformed traditional methods by a large margin and is commonly used in food recognition (Rahmat and Kutty, 2021; Razali et al., 2021; Zahisham et al., 2020; Hussain et al., 2019; Teng et al., 2019; Zhou et al., 2019; Kagaya et al., 2014; Termritthikun et al., 2017; Pandey et al., 2017; Hassannejad et al., 2016) and dietary assessment (Dalakleidi et al., 2022; Tahir and Loo, 2021a; Liu et al., 2017a, 2016; Christodoulidis et al., 2015; Kong and Tan, 2012) research. Food recognition is a

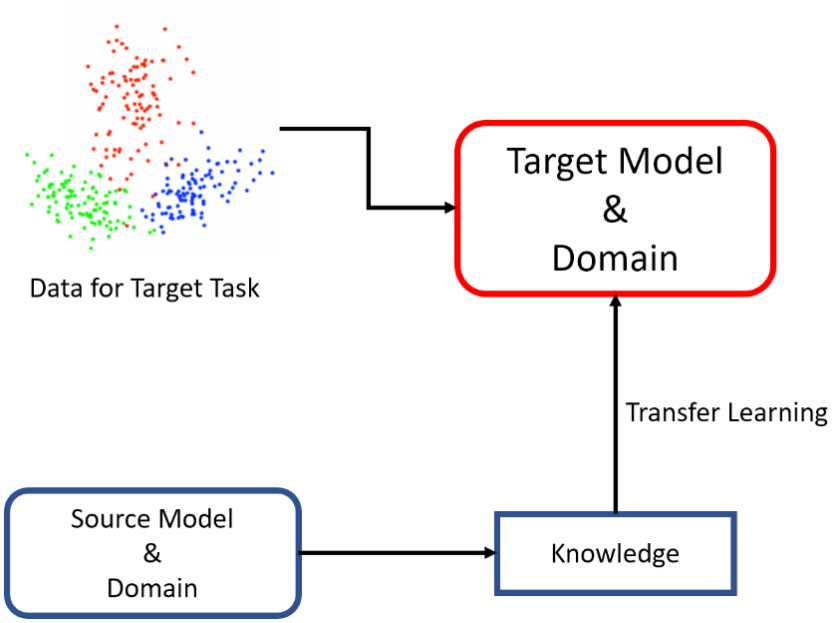


Figure 1.2: Concept of transfer learning to reuse a model or knowledge for another related task (Sarkar et al., 2018).

challenging task due to the large variability in food appearance. Deep learning models, such as convolutional neural networks (CNNs), have been shown to be effective for food recognition. However, these models require a large amount of labeled data to train. In many cases, collecting and annotating a large dataset of food images is not feasible.

Transfer learning provides a solution to this problem by allowing the knowledge learned by a model on a large dataset to be transferred to a new task with limited data. Transfer learning is a powerful technique that allows a model trained on one task to be adapted to another related task with minimal training data. The concept of transfer learning is shown in Figure 1.2 to demonstrate how transfer learning enables reusing existing knowledge for new related tasks (Sarkar et al., 2018). In recent years, transfer learning has been widely used in computer vision tasks, including food recognition (Tai et al., 2022; Rajesh et al., 2022; Murugaiyan et al., 2021; Merchant and Pande, 2019; Yu et al., 2016; Pehlic et al., 2019; Basrur et al., 2022; Temdee and Uttama, 2017; Xie et al., 2021; Tasci, 2020). The idea behind transfer learning is to leverage the knowledge learned by a model on a large dataset and use it to improve the performance of a model on a new task with limited data. The transfer learning to food recognition can be applied using a pre-trained convolutional neural network, such as ResNet (He et al., 2016), AlexNet (Krizhevsky et al., 2017), GoogLeNet (Szegedy et al., 2015), Inception (Szegedy et al., 2016) and VGG net (Simonyan and Zisserman, 2014), etc., as a fixed

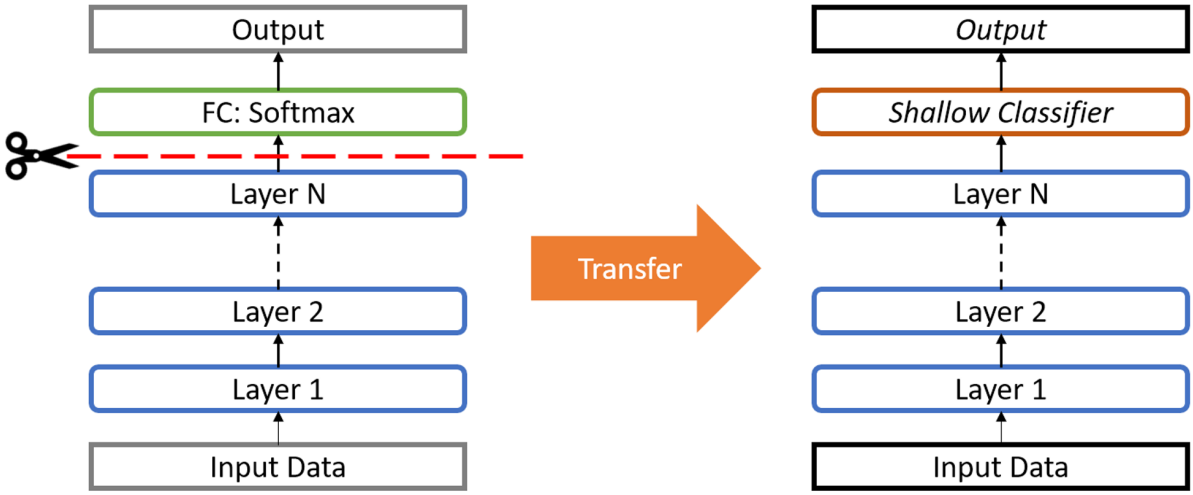


Figure 1.3: Deep transfer based on feature extraction by removing the last fully connected layer of the pretrained model and using the remaining layers to extract features (Sarkar et al., 2018).

feature extractor or fine-tuning a pre-trained model. A pre-trained model as a fixed feature extractor involves removing the last fully connected layer of the pre-trained model and using the remaining layers to extract features from food images, as shown in Figure 1.3. These features are then used to train a new classifier on the target dataset. Another approach is to fine-tune a pre-trained model on a dataset of food images. This involves training the last fully connected layer of the pre-trained model on the target dataset while keeping the remaining layers fixed.

This thesis presents a system for semantic food recognition using computer vision and deep learning techniques for health monitoring. The presented approaches have been developed in the context of the FoodRec project (Battiato et al., 2021), which aims to define an automatic framework for the monitoring of people's health and habits during their smoke-quitting program. The aim is to extract and infer semantic information from the food images to analyze diverse foods present in the image. This project involves several stages from image acquisition to recognition. In particular, the efforts are devoted to the development of new segmentation and recognition algorithms to perform the food recognition task accurately. Further, we aim to estimate the quantities of each food item detected within an image.

1.2 Motivation

Food recognition technology is an exciting and rapidly growing field that has the potential to revolutionize the way we interact with food. With the proliferation of digital cameras and smartphones, food images have become ubiquitous on the internet, making food recognition an important and practical problem for a wide range of applications, from calorie counting and dietary tracking to food delivery and recipe recognition. Food recognition technology has the power to make our lives easier, healthier, and more enjoyable. Additionally, food recognition technology has the potential to have a significant impact on public health and nutrition by providing more accurate and automated ways to identify food items and track dietary intake. With the advancement of computer vision and deep learning techniques, food recognition technology is becoming more accurate and robust, making it a viable solution for many applications.

Food recognition technology is an active research field and has a wide range of potential applications in many areas from culture (Giampiccoli and Kalis, 2012; Sajadmanesh et al., 2017), food science (Killgore and Yurgelun-Todd, 2005; Ofli et al., 2017), agriculture (Lu et al., 2017; Chatnuntawech et al., 2018), medicine to biology (Min et al., 2019), etc., but not limited to. This thesis focuses on food recognition applications in health monitoring during smoke-quitting process to monitor the dietary habits of people with the potential applications as follows:

Track and Monitor the Dietary Habits of People

Food recognition technology is used to identify and classify food items in images taken by individuals in order to track and monitor their dietary habits over time. This can be done by using a smartphone app that utilizes food recognition technology to automatically identify and classify food items in images taken by the user. The app can then store this information and provide the user with detailed information about their dietary intake, including calorie and nutrient information, and make personalized recommendations based on their dietary habits. The technology can also be used in the healthcare sector, for example, in hospitals, where it can be used to monitor the dietary habits of patients, especially those with chronic diseases, and to track their progress over time. Furthermore, it can be used in research studies that aim to understand the dietary habits of different population groups and identify potential health risks.

Increase the Awareness in People Daily Diet

Food recognition technology can be used to help individuals become more aware of what they are eating by tracking and analyzing their food intake over time using a

smartphone application. By monitoring their food habits over time, individuals can become more aware of the foods they eat, how much they eat, and how often they eat them. This can help them to identify patterns in their eating habits that may be contributing to weight gain or other health issues. It can also help them to identify foods that they might want to eat more or less in order to achieve specific dietary goals.

Kind and Amount of Taken Food

Food recognition technology can be used to identify and classify food items in images taken by individuals and then use that information to determine the types and quantity of food that they have consumed. This technology can be used to help individuals to understand the nutritional value of the food they are consuming and monitor the balance of their diet over time. This technology can be used to help people make healthier food choices, identify consumption of certain nutrients or calories intake, and monitor progress toward weight loss or other dietary goals. This technology can also be used to monitor the food consumption of individuals with specific health conditions, such as diabetes, and to make personalized recommendations based on their dietary habits.

How Many and What Times the User has a Meal

Food recognition technology can be used to track and monitor the frequency and timing of meals consumed by an individual. A smartphone application can be used to store the dietary information extracted from user meals and provide the user with detailed information about the number of meals they have consumed and the times at which they were consumed. This helps to track and monitor the frequency and timing of meals consumed by an individual to understand their eating patterns and make more informed decisions about when to eat and how often to eat.

Nutrition analysis, Calorie Counting and Dietary Tracking Apps

Food recognition technology can be used to identify the food items in images and provide detailed nutrition information about the meal. Food recognition technology is used to accurately identify food items in images taken by the user, allowing them to keep track of their calorie and nutrient intake.

Food recognition technology can be used to identify the ingredients in a dish and suggest recipes that include those ingredients. Food recognition technology can be used to identify food items in images and track consumer preferences, allowing for

more targeted marketing and advertising strategies. Food recognition technology can be used to identify the food items in images of meals, allowing for faster and more accurate delivery and restaurant recommendations.

Food Recognition for Healthcare and Medications

Food recognition technology can be used for healthcare and medications to study the effects of different diets on various health conditions. In a clinical setting, food recognition technology can also be used to monitor a patient's dietary intake, for example by analyzing images of their meals to ensure compliance with a prescribed diet or to track nutrient intake for patients with specific health conditions. One of the uses is the diagnosis of food allergies or sensitivities through analysis of blood and dietary habits. It can also be used for monitoring nutrient levels in patients with chronic conditions such as diabetes or heart disease, which can help in the development of personalized nutrition plans for patients based on their genetic and health history. Additionally, food recognition technology can aid in the detection of harmful contaminants or pathogens in food products, making the food supply safer. Furthermore, it can also help in the implementation of dietary interventions for the management of specific medical conditions such as inflammatory bowel disease or celiac disease, and it can track nutrient intake for patients on specialized diets, such as a low-sodium or low-carb diet. It can also identify potential drug-nutrient interactions to prevent adverse reactions in patients taking multiple medications. Finally, monitoring of food and nutrient intake in clinical trials to assess the effectiveness of new drugs or therapies can also be achieved with food recognition technology.

1.3 Problem statement

Studies have shown a strong correlation between dietary habits' changes of individuals and the smoking cessation process (Morabia et al., 1999). Abstinence from smoking is associated with several negative effects, including the gain of weight, eating disorders, mood changes, and irritability during the initial period of smoke quitting (Pisinger and Jorgensen, 2007; MacLean et al., 2018). The objective of our FoodRec project is to study, develop, and evaluate an automatic framework able to monitor the dietary habits of people involved in a smoke-quitting protocol. The system will periodically acquire images of the food consumed by the users, which will be analyzed by modern food recognition algorithms able to extract and infer semantic information from food images. Such information will be exploited to perform advanced inferences and to make correlations between eating habits and smoke quitting process steps, providing specific information to the clinicians about the response to the quitting protocol that is

directly related to observable changes in eating habits. It can help a doctor to have a better opinion with respect to the patient's behavior, mood changes over time, quitting treatment response, and hence his health needs.

The aim of the project is to achieve high semantic level inferences using computer vision and deep learning techniques applied to food image recognition to make such technology available to smoking cessation program patients through a smart and intuitive smartphone application named FoodRec (Battiato et al., 2021). In particular, efforts are required to the development of new segmentation and recognition algorithms to perform the food recognition task accurately. We believe that more attention is needed to improve the image recognition algorithms on real world food images and let the system to learn from the user, his specific eating habits. This will allow to build a reliable dietary monitoring system able to automatically infer the quality of the patients' diet, as well as habits changes.

1.4 Thesis contribution

The major contributions of this thesis are listed:

- We introduce a new FoodRec-50 dataset collected by 164 users during their smoking cessation therapy to monitor their dietary habits using FoodRec (Battiato et al., 2021) iOS and Android smartphone applications. Data is annotated manually by embedding the user ID with the image label as it is fed simultaneously to the network. Data augmentation is applied to increase the food image variability and to compensate the problem of some of the underrepresented food classes, as they have a relatively imbalanced number of samples per class. So, different transformation techniques have been applied, such as image rotation, random crop, flip, etc.

- We propose a Deep Convolutional Neural Network (Hussain et al., 2022) able to recognize food items of specific users and monitor their eating habits. Our proposed study differs from the recognition that happens in the development of general purposes food recognition systems as the proposed approach considers the specific user as well to learn and monitor its eating habits. It is composed of a food branch and a user branch. The food branch is learning visual representation of the input food items like the traditional food recognition algorithm. On the other hand, the user branch takes into account the specific user's eating habits by learning the user's eating weight matrix. As we change the user bias input, the result in the prediction is being changed according to the dietary habits of that user. The proposed network with one-hot user vector is effective because

it learns specific user eating habits with the food as compared to the traditional food recognition systems with competitive results.

- We propose a novel Convolutional Deconvolutional Pyramid Network (CDPN) (Hussain et al., 2023) for food segmentation to understand the semantic information of an image at a pixel level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. Furthermore, the proposed method demonstrated significant improvements on two well-known public benchmark food segmentation datasets.

- We propose another Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images at a pixel level to recognize different food items present in an image. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns, while the proposed CDPN also uses interpolation for features upsampling along with the deconvolution layers. Our proposed network demonstrated significant improvements in the results on the benchmark food dataset as compared to the state-of-the-art methods. In addition to evaluating the performance of our proposed segmentation method on the MyFood dataset, we also performed cross-data experiments to assess its generalization capabilities on our FoodRec dataset. By conducting cross-data experiments on the FoodRec dataset, we were able to determine that our method could effectively make accurate predictions in different contexts. This qualitative evaluation served as an important complement to our evaluation on the FoodRec dataset, and helped to strengthen our confidence in the effectiveness of our method.

- We provide a comprehensive overview of the literature regarding not only to food recognition technology but also to computer vision, traditional machine learning, and deep learning techniques used for food recognition and in general.

This thesis is based on the following research papers that were written during the course of my Ph.D. program. Here is the list of published and accepted research publications.

Journals

1. Mazhar Hussain, Alessandro Ortis, Riccardo Polosa, and Sebastiano Battiato. "Semantic Food Segmentation using Convolutional Deconvolutional Pyramid Network for Health Monitoring".

Accepted in International Journal of Computer Theory and Engineering IJCTE Journal, 2022.

2. Mazhar Hussain, Alessandro Ortis, Riccardo Polosa, and Sebastiano Battiato. "Deep FoodRec: Food Segmentation and Recognition using Deep Learning for Health Monitoring".

   Preprint 2023.

Conferences

1. Sebastiano Battiato, Pasquale Caponnetto, Oliver Giudice, Mazhar Hussain, Roberto Leotta, Alessandro Ortis, and Riccardo Polosa. "Food Recognition for Dietary Monitoring during Smoke Quitting." (Battiato et al., 2021)

   The paper (Battiato et al., 2021) won the Best Poster Award Certificate.

   In Proceedings of the International Conference on Image Processing and Vision Engineering (IMPROVE 2021), pp. 160-165. 2021.

   $DOI : 10.5220/0010492701600165$

2. Mazhar Hussain, Alessandro Ortis, Riccardo Polosa, and Sebastiano Battiato. "User-Biased Food Recognition for Health Monitoring." (Hussain et al., 2022)

   In Proceedings of the 21st International Conference on Image Analysis and Processing (ICIAP 2022), pp. 98-108. Cham: Springer International Publishing, 2022.

   $DOI : 10.1007/978 - 3 - 031 - 06433 - 3_9$

3. Mazhar Hussain, Alessandro Ortis, Riccardo Polosa, and Sebastiano Battiato. "Semantic Food Segmentation for Health Monitoring". (Hussain et al., 2023)

   In Proceedings of the 15th International Conference on Machine Vision (ICMV 2022), pp. 106-113. SPIE, 2023.

   $DOI : 10.1117/12.2679721$

1.5 Thesis Organization

In Chapter 2, we discuss related works and contributions to the field and explore the current state-of-the-art in food recognition technology. Food recognition technology is a rapidly growing field that leverages advancements in computer vision and deep learning. Researchers and industry experts have been actively working on developing

and improving food recognition technology, and there is a growing body of literature on this topic.

Chapter 3 presents our ongoing project, which aims to study, develop, and evaluate an automatic framework able to track and monitor the dietary habits of people involved in a smoke-quitting protocol. The system will periodically acquire images of the food consumed by the users, which will be analyzed by modern food recognition algorithms able to extract and infer semantic information from food images. The extracted information, together with other contextual data, will be exploited to perform advanced inferences and to make correlations between eating habits and smoke quitting process steps, providing specific information to the clinicians about the response to the quitting protocol that is directly related to observable changes in eating habits.

Chapter 4 presents a user-biased food recognition system. The presented approach has been developed in the context of the FoodRec project, which aims to define an automatic framework for the monitoring of people's health and habits during their smoke quitting program. The goal of food recognition is to extract and infer semantic information from the food images to classify diverse foods present in the image. We propose a novel Deep Convolutional Neural Network able to recognize food items of specific users and monitor their habits. It consists of a food branch to learn visual representation for the input food items and a user branch to take into account the specific user's eating habits. Furthermore, we introduce a new FoodRec-50 dataset with 2000 images and 50 food categories collected by the iOS and Android smartphone applications, taken by 164 users during their smoking cessation therapy. The information inferred from the users' eating habits is then exploited to track and monitor the dietary habits of people involved in a smoke-quitting protocol. Experimental results show that the proposed food recognition method outperforms the baseline model results on the FoodRec-50 dataset. We also performed an ablation study, which demonstrated that the proposed architecture is able to tune the prediction based on the users' eating habits.

Chapter 5 presents semantic food segmentation to detect individual food items in an image. The presented approach has been developed in the context of the FoodRec project, which aims to study and develop an automatic framework to track and monitor the dietary habits of people during their smoke quitting protocol. The goal of food segmentation is to train a model that can look at the images of food items and infer semantic information to recognize individual food items present in an image. In this contribution, we propose a novel Convolutional Deconvolutional Pyramid Network (CDPN) for food segmentation to understand the semantic information of an image at a pixel level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. Furthermore, the proposed method

demonstrated significant improvements on two well-known public benchmark food segmentation datasets.

Chapter 6 describes thesis conclusions and future works.

The appendix presents additional research work with the collaboration of the Department of Drug and Health Science, University of Catania, done during my Ph.D. but not directly related to this thesis. The aim of the work was to find a correlation between well-defined and selected parameters such as the type of nanocarrier, the particle size and the surface charge, and the targeting efficiency indexes %DTE and %DTP. The DTP is used by researchers to define the trigeminal and olfactory involvement instead of the systemic pathway when a nanomedicine is intranasally administrated. In addition, the possible influence of the molecular weight of the conveyed drug was also considered in the correlation studies. We performed nose-to-brain drug delivery data cleaning, data conversion, data standardization, and data classification using state-of-the-art machine learning algorithms. We are working to publish a journal from this work with the collaboration of the Department of Drug and Health Science.

Chapter 2

Literature Review

Recently, computer vision and deep learning techniques have gained a lot of attention due to the high level of performance in various research fields and applications, as well as in food recognition technology. Computer vision research devoted to the analysis of food images including previous works on food detection, classification, and segmentation. In this section, we present the related research works in the field of food recognition technology.

Allegra et al. (2020) presented a review on food recognition technology and its applications, particularly in the health sector for monitoring dietary and calories intake. Several computer vision techniques have been explored for food understanding in the areas such as automatic food detection and recognition for automatic harvesting, food quality assessment for industrial purposes, dietary management, tracking food consumption, classification, and retrieving food using publicly available datasets such as Recipe1M+ (Marin et al., 2018), UPMC Food-101 (Wang et al., 2015), UNICT-FD889 (Farinella et al., 2015), FRIDa (Foroni et al., 2013), UPMC Food-101 (Wang et al., 2015), ETHZ Food-101 (Bossard et al., 2014), UEC FOOD 100 (Matsuda and Yanai, 2012), etc. However, achieving automatic food recognition in healthcare applications requires to meet rigorous medical protocol standards and to collect annotated datasets that contain information about the type of food, its quantity, location, and calories of each food item in an image.

Min et al. (2019) covered food computing, which involves a range of tasks such as acquisition, analysis, perception, recognition, retrieval, recommendation, and prediction. The article examines how this field of study has been applied in various domains such as health, culture, agriculture, medicine, and biology. The research explores different computer vision and machine learning techniques such as GoogLeNet (Szegedy et al., 2015), Inception V3 (Szegedy et al., 2016), VGG (Simonyan and Zisserman, 2014), and CaffeNet (Jia et al., 2014), which have been employed for various food recognition

tasks including single-label, multi-label, sensor-based, food portion estimation, and personalized food recognition. Additionally, it provides insights into the benchmark food datasets used for experimentation and evaluation purposes.

Amugongo et al. (2023) reviewed mobile computer vision-based solutions for food recognition, volume estimate, and calorific estimation. This systematic review also assessed the level of explanation offered by these applications to help users understand their classification and prediction results. According to their research, 90.9% of applications do not distinguish between food and non-food items. In addition, only one study offered a mobile computer vision-based nutritional intake application that sought to explain the contributing features to classification (Tahir and Loo, 2021b). The use of mobile computer vision-based applications in healthcare is gaining interest due to their potential in managing chronic diseases like diabetes. These applications have the capacity to promote healthy eating habits and mitigate the complications associated with unhealthy food.

Min et al. (2023) proposed a novel deep learning architecture, called progressive region enhancement network (PRENet), for food recognition. The network employs progressive local feature learning and region feature enhancement techniques to extract discriminative features from food images. The progressive local feature learning strategy learns complementary multi-scale finer local features by progressively training the network. On the other hand, the region feature enhancement technique uses self-attention to capture richer context information at multiple scales to further improve local feature representation. This paper also introduced a new Food2K dataset that is larger than the ImageNet dataset (Yanai and Kawano, 2015). The proposed PRENet (Min et al., 2023) architecture consists of two sub-networks, one for global and the other for local feature extraction. The global features representation is useful for recognizing food images belonging to different superclasses with clear visual differences. The PRENet uses global average pooling to extract global features after the last convolution layer of the network. On the other hand, the local features representation is useful for learning fine-grained features of food in different sub-classes under the same superclass, where there is high inter-class similarity. To extract local features, the paper employs a progressive training strategy that trains the network from low to high stages, starting with a small receptive field and gradually increasing it to encompass a larger field surrounding the local region. The local feature extraction process involves convolutional blocks and a global maximum pooling layer. To further enhance the local feature representation, the paper employs a self-attention mechanism that captures the relationship between different local features by identifying the co-occurring food features in the feature map. The local features are first extracted, then enhanced using self-attention, and finally combined with the global feature representations.

Farinella et al. (2016) presented a survey of the studies in the context of food image processing from the perspective of computer vision to the current state-of-the-art

methods. This paper introduced a new UNICT-FD1200 dataset that consists of 4754 food images of 1200 distinct dishes acquired during real meals. The UNICT-FD1200 dataset has been annotated manually with eight categories. A new representation has been proposed based on the perceptual concept of Anti-Textons that performs better as compared to the existing representations in the context of food retrieval and classification.

Natesan et al. (2023) built a convolutional neural network to classify food products using deep learning principles. The efficacy of convolutional neural networks to recognize food items is attributed to their ability to exclude distracting characteristics present in the images. The proposed architecture has various layers to convert the three-dimensional input into an output volume and to filter input to a higher level of abstraction. Its implementation can aid individuals in identifying healthy and unhealthy food items, potentially preventing disease. The study employed a combination of image datasets obtained from social media, Kaggle, and Google.

Ciocca et al. (2016) designed an automatic framework for tray food analysis to find the region of interest of the input image and then predict the food class for each region. To accomplish this task, a range of visual descriptors were employed, including opponent Gabor features, chromaticity moments, color histogram, local color contrast, Gabor features, complex wavelet features, and convolutional features. The food classification process was performed using two classifiers, namely the k nearest neighbor (KNN) and support vector machine (SVM), which were applied to a novel LTC food dataset.

Fakhrou et al. (2021) proposed a smartphone application that utilizes a deep convolutional neural network (CNN) model, trained for food and fruit recognition, to aid children with visual impairments. This paper introduces a novel food dataset comprising both food dishes and fruit varieties to address the food classification problem. Furthermore, the proposed approach leverages ensemble learning, combining multiple deep CNN architectures on a customized food dataset, using the soft voting method to aggregate the results of multiple models. The performance achieved by the ensemble model in various food datasets validates its efficacy in addressing food classification problems.

Sapna et al. (2023) developed a six-layer convolutional neural network architecture to classify images and extract their characteristics. The study introduced a calorie estimation system where users can upload pictures of food items, which are then analyzed to determine the approximate number of calories. This system also provides users with weekly updates on their calorie consumption and recommends the number of calories they should consume to avoid obesity-related diseases such as heart attack and cancer. To enable the system to recognize complex images, the study compiled a database of food images that contains 20 categories, each with 500 images. With the help of this

image-based calorie estimation system, users and healthcare professionals can identify dietary patterns and food choices that can improve their health.

A semisupervised generative adversarial network (GAN) is used for food recognition (Mandal et al., 2018) using partially labeled data. Network architecture consists of a generator and discriminator. The generator produces dataset fake samples, and the discriminator learns the nature of the problem and further recognizes different food items with partially labeled training data. The author claimed that outperformed results on the ETH Food-101 datasets and Indian food dataset as compared to the AlexNet, GoogleNet, and Ensemble Net.

The ResNet deep residual learning architecture (He et al., 2016) is proposed for image recognition with powerful representational capability for learning discriminative features from complex scenes. ResNet network architecture designed for the classification task, trained on the ImageNet dataset of natural scenes that consists of 1000 classes. Evaluation has been performed with the residual network with depths of 18-layers, 34-layers, 50-layers, 101-layers, and 152-layers. The ResNeXt (Xie et al., 2017) architecture is a combination of ResNet and InceptionNet. It is composed of a series of residual blocks that have the same topology and follow a split-transform-merge structure within each block. This design introduces a novel dimension called cardinality, which represents the size of the set of transformations used in the network. The addition of this cardinality dimension allows for greater model capacity and improves the model's performance on a variety of image classification tasks. and further Hu et al. (2018) introduced a new architectural unit called the "squeeze-and-excitation" (SE) block. The SE block is comprised of two main operations: squeeze and excitation. First, the squeeze operation aggregates channel-wise feature responses to produce a channel descriptor. Then, the excitation operation models the channel descriptor to produce a set of weights, which are multiplied by the original features to produce the final output. The SE block is designed to improve the quality of representations learned by the network, leading to better performance on a variety of tasks.

Subhi et al. (2019) provided a comprehensive overview of the state-of-the-art vision-based approaches for food recognition and dietary assessment. The paper focused to the importance of dietary assessment in healthcare and the details of vision-based approaches for automatic food recognition including food detection, food segmentation, food classification, volume, and weight estimation. It also presents a comprehensive review of the existing datasets and benchmarks for food recognition and dietary assessment. It compares the performance of various state-of-the-art approaches on these datasets.

Chopra and Purwar (2022) provided a comprehensive review of segmentation techniques and their applicability to food image segmentation. There are several techniques to segment data including threshold segmentation (Kawano and Yanai, 2015; Muthukr-

ishnan and Radha, 2011; Bhargavi and Jyothi, 2014), clustering segmentation (Maione et al., 2019; Kapoor and Singhal, 2017; Dehariya et al., 2010; e Silva et al., 2018), edge detection segmentation (Pouladzadeh et al., 2014; He et al., 2015; Ganesan and Sajiv, 2017; Ansari et al., 2017), region growing segmentation (Dehais et al., 2016; Anthimopoulos et al., 2013; Merzougui and El Allaoui, 2019), watershed segmentation (Liu et al., 2017b; Kornilov and Safonov, 2018; Vincent and Soille, 1991), graph partitioning segmentation (Moussawi et al., 2020), grab cut segmentation (Li et al., 2018) and also semantic segmentation (Thoma, 2016; Sun and Wang, 2018; Zhang et al., 2018).

Lu et al. (2020) proposed a system to effectively estimate nutrient intake by using RGB depth image pairs that are captured before and after meal consumption. The goal is to reduce disease-related malnutrition among hospitalized patients. The system incorporates a novel multi-task contextual network for food item segmentation, classification with few-shot learning-based algorithms (Snell et al., 2017) for food recognition, and 3D food surface extraction. This allows for automatic estimation of nutrient intake by sequentially segmenting, recognizing, and estimating the consumed food volume for each meal. The paper also describes a new database containing food images, recipes, and nutrient information collected from real hospital scenarios. The experimental results show that the estimated nutrient intake is highly correlated with ground truth values and outperforms existing techniques for nutrient intake assessment. Maintaining good nutritional status is crucial for patients, and the healthcare system, as malnutrition can lead to increased risk of hospital infections, higher mortality and morbidity rates, longer hospital stays, and greater healthcare expenses. The proposed artificial intelligence algorithms for food recognition rely on limited training data, overcoming the challenge of sophisticated annotation requirements that limit the quality and size of food image databases for nutrient intake assessment.

Freitas et al. (2020b) introduced a system for automatic monitoring of user diet and nutritional intake by classifying and segmenting food presented in images. This paper compares the performance of state-of-the-art algorithms for food recognition using a dataset composed of nine classes of the most consumed Brazilian food types. Additionally, the study proposes an integrated system into a mobile application that automatically recognizes and estimates the nutrients in a meal to assist people in better nutritional monitoring.

Joshua et al. (2023) developed a "smart plate health to eat" system that assists patients and users in identifying the type of food, its weight, and nutrient contents. The research involves 50 food categories using the YOLOv5 (Jocher et al., 2021) algorithm to evaluate food identification, weight measurement, and nutritional value with the help of a Chenbo load cell weight sensor, weighing A/D module pressure sensor, and camera module. By implementing the YOLOv5s approach and loadcell sensor readings, the system can compute the quantity of food calories. The loadcell sensor's results indicated its capability to operate with high accuracy while providing accurate

nutritional information.

Zhao et al. (2017) proposed a pyramid scene parsing (PSPNet)network for semantic segmentation. It came first in the ImageNet scene parsing challenge 2016 on PASCAL VOC 2012 benchmark and cityscapes benchmark datasets segmentation. It consists of a pyramid pooling module to exploit the local and global context information. First, a feature map is extracted from the last convolutional layer of the convolution neural network that is fed to the pyramid pooling module in order to harvest different pooling pyramids. Further, upsampling and concatenation layers are used for the final features representation. Finally, the convolution layer is applied to obtain the pixel-wise prediction.

Lin et al. (2017) developed a feature pyramid network for segmentation and object detection. Feature pyramids play an important role in recognition tasks. An architecture with skip connections is designed to extract high-level semantic feature maps that involve a bottom-up pathway, which computes a feature hierarchy, and a top-down pathway, which computes semantically stronger feature maps from higher pyramid levels enhanced with features from the bottom-up pathway. Feature pyramid network is used for land segmentation with Resnet encoder pre-trained on ImageNet dataset in the bottom-up pathway by (Seferbekov et al., 2018).

Raju et al. (2023) presented a novel technique for food imaging that employs two types of imaging sensors: color and thermal. They also offer a multi-modal four-dimensional (RGB-T) image segmentation technique that use a k-means clustering algorithm to locate areas of similar-looking food items in combinations of hot, cold, and room temperature foods. Six combinations of two food items were captured using both RGB and infrared sensors to collect data. The resulting RGB and thermal data were combined to create an RGB-T image, and three sets of data were examined (RGB, thermal, and RGB-T). A bootstrapped optimization of within-cluster sum of squares was used to estimate the ideal number of clusters for each example. When compared to the RGB-T data alone, the combined RGB-T data produced superior outcomes.

Pfisterer et al. (2019) developed an automatic semantic food segmentation method using multi-scale encoder-decoder network architecture for food intake tracking and estimation in long-term care homes. A deep convolutional neural network macroarchitecture has been proposed for pixel-level classification of food that consists of a residual encoder and decoder microarchitecture and per-pixel food/no-food classification layer. For the encoder, ResNet architecture trained on the ImageNet dataset is used because of its discriminative feature learning ability. For the decoder, a pyramid scene parsing network is chosen. The proposed method achieved comparable results to semi-automatic graph cuts.

Ronneberger et al. (2015) introduced UNet architectures for semantic segmentation for biomedical image segmentation. It follows an encoder-decoder structure with

skip connections between corresponding layers of the encoder and decoder. The architecture comprises a contracting path, which is similar as the conventional convolutional network, and an expansive path that facilitates the upsampling of feature maps. UNet++ (Zhou et al., 2018) is an extension of UNet that was proposed to further improve biomedical image segmentation performance. UNet++ architecture is more powerful than U-Net in which sub-networks encoder and decoder are connected through a series of nested and dense skip pathways to perform the image segmentation.

Sharma et al. (2021) proposed GourmetNet, a network for food segmentation that integrates spatial and channel attention using the waterfall atrous spatial pooling module. The network is based on a segmentation approach (Peng and Ma, 2020)that uses stride spatial pyramid pooling to obtain multi-scale semantic information and a dual attention decoder with a channel attention branch and a spatial attention branch to capture semantic feature map representation.

Mask RCNN (He et al., 2018) is an advanced version of faster RCNN (Ren et al., 2015) introduced to generate segmentation masks for each detected object. It was developed by the Facebook research group to perform instance segmentation to identify every instance of an object in an image and determining the pixels that correspond to the specific classes within that image. While faster RCNN focuses on detecting object bounding boxes, Mask RCNN extends this capability by including a separate object mask prediction branch. By adding this mask prediction branch, Mask RCNN can accurately outline the object's shape and boundaries.

Harshitha et al. (2023) proposed a system using Otsu's method to detect the contour of each food item and estimate its calories using data trained with faster RCNN. To detect the actual caloric quantity of a meal involves considering factors such as the food item's region, size, and weight. Utilizing deep learning algorithms, the object can be identified, and the calories can be estimated based on object detection and volume estimation methods. The proposed scheme involves three stages, including image segmentation to determine the contour of each food item, image recognition utilizing faster RCNN, and estimation of the food's weight and caloric content.

Long et al. (2015) proposed a fully convolutional network (FCN) for image semantic segmentation. FCN employs skip architecture to combine information from multiple layers of the network from a deep, coarse layer with appearance information from a shallow, fine layer to generate accurate segmentation results. In particular, several state-of-the-art classification networks have been transformed into fully convolutional networks, including AlexNet (Krizhevsky et al., 2017), GoogLeNet (Szegedy et al., 2015), and VGG net (Simonyan and Zisserman, 2014) by leveraging their learned representations to the segmentation task through fine-tuning.

Badrinarayanan et al. (2017) designed SegNet a convolutional neural network for semantic image segmentation. It consists of an encoder network, a decoder network,

and a pixel-wise classification layer. The encoder network mirrors the topology of VGG16's 13 convolutional layers, while the decoder network upsamples low-resolution encoder feature maps to full input resolution feature. The decoder performs non-linear upsampling by using pooling indices computed during the corresponding encoder's max-pooling step.

Paszke et al. (2016) proposed a deep neural network created specifically for real-time segmentation on embedded platforms. Its design is heavily influenced by Inception architecture (Szegedy et al., 2016) and optimized for efficient large-scale computations to achieve better performance evaluation of the proposed architecture on embedded systems.

Aslan et al. (2018) utilized DeepLabV2 (Chen et al., 2017) for semantic food segmentation for two tasks related to food such as simultaneously performing food segmentation and recognizing food types in food images and detecting food regions in a given image by performing food and non-food segmentation. Chen et al. (2018) proposed a more advanced variant DeepLabv3+ model that builds upon the existing DeepLabv3 (Chen et al., 2014) by introducing a decoder module that enhances the segmentation outcomes, particularly at object boundaries. To improve the performance of the encoder-decoder network, the xception model is considered to use depthwise separable convolution for both the atrous spatial pyramid pooling and decoder modules to refine the segmentation results.

Related studies described above present traditional food image recognition without taking into account user habits. Our proposed study differs from the recognition that happens in the development of general purpose food recognition systems as the proposed approach considers the specific user that uploaded the food image to learn and monitor its eating habits. Our proposed architecture (Hussain et al., 2022) consists of two branches, the food branch and the user branch to extract concatenated feature map from two branches to recognize the food items. The food branch is learning visual representation of the input food items like the traditional food recognition algorithm. On the other hand, the user branch takes into account the specific user's eating habits by learning the users eating weight matrix. The user branch takes one hot vector user input and learns the users eating weight matrix. As we change the user bias input, the result in the prediction is being changed according to the dietary habits of that user. The proposed network with one-hot user vector is effective because it learns specific user eating habits with the food as compared to the traditional food recognition systems with competitive results. To accomplish this task, we introduced a new FoodRec-50 dataset with 50 food categories collected by FoodRec (Battiato et al., 2021) smartphone applications, taken by 164 users during their smoking cessation therapy. This dataset is specific to the users who are involved in the smoke-quitting process to monitor their dietary habits. The dataset is produced to study the correlation between eating information with smoking habits. In the future, such data will be used to find

correlations with respect to the smoking activity of the subjects during the period of observation. To collect the food data, the iOS and Android smartphone applications are used. Users can upload a meal intake image by taking a picture of what they eat and can assign labels to the food image what it contains. Furthermore, we proposed architecture for semantic food segmentation able to produce a rich segmentation map of the input food image that would further allow people to estimate the volume and, hence, the quantities of each food item to the assessment of nutrient intake and dietary analysis. A novel Convolutional Deconvolutional Pyramid Network (CDPN) (Hussain et al., 2023) is proposed for food segmentation to understand the semantic information of an image at a pixel level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. Furthermore, the proposed method demonstrated significant improvements on two well-known public benchmark food segmentation datasets. We proposed another Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images at a pixel level to recognize different food items present in an image. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns, while the proposed CDPN also uses interpolation for features upsampling along with the deconvolution layers. Our proposed network demonstrated significant improvements in the results on the benchmark food dataset as compared to the state-of-the-art methods. Additionally, we also conducted a cross-data qualitative analysis of our proposed segmentation method to assess its generalization capabilities on our FoodRec dataset.

Chapter 3

FoodRec Project

FoodRec Project aims to study, develop, and evaluate an automatic framework able to track and monitor the dietary habits of people involved in a smoke-quitting protocol. The system will periodically acquire images of the food consumed by the users, which will be analyzed by modern food recognition algorithms able to extract and infer semantic information from food images. The extracted information, together with other contextual data, will be exploited to perform advanced inferences and to make correlations between eating habits and smoke quitting process steps, providing specific information to the clinicians about the response to the quitting protocol that is directly related to observable changes in eating habits.

Food recognition from digital images for the analysis of dietary habits has become an important aspect in health monitoring applications in different domains. On the other hand, food monitoring is a crucial part of human life since the health is strictly affected by diet (Nishida et al., 2004). The impact of food in people's lives led research efforts to develop new methods for automatic food intake monitoring and food logging (Kitamura et al., 2010). We present the state of the FoodRec project founded by ECLAT, which objective is the study, development, and evaluation of state-of-the-art digital technologies to define a framework able to track the dietary habits of an observed person and make correlations with the smoking cessation process that the subject is performing. The system will periodically acquire images of the food eaten by the patient over time, that will then be processed by food recognition algorithms able to detect and extract semantic information from the images containing food. The extracted data will be exploited to infer the dietary habits, the kind and amount of taken food, how much time the user spends eating during the day, how many and what times the user has a meal, etc. Inferences performed on different days can be compared and further processed to perform analysis on user's habits changes and other inferences related to user's behaviour, such as increase of junk food intake and mood changes over time. The recording and semantic organization of daily habits can

help a doctor to have a better opinion with respect to the patient's behaviour,quitting treatment response, and hence his health needs. So far, many efforts have been spent in the application of technology on smoke monitoring (Ortis et al., 2020) and food recognition (Allegra et al., 2020), this project represents the first attempt of the application of Artificial Intelligence (AI) and multidisciplinary competences for the definition of a framework able to drive and support people who are trying to stop smoking, by acting on multiple aspects simultaneously. The Food Recognition project (FoodRec) is granted by the Foundation for a SmokeFree World (FSFW)

3.1 FoodRec - Research Plan

The project involves several phases, which are sketched by the chart shown in Figure 3.1. The diagram shows five main phases as follows:

- Initial procedures
- Preliminary investigation
- Research planning
- Software development
- Food recognition research
- Deploy

The first ones are related to preliminary studies and research, whereas the last ones regard the development of algorithms and software toward the final deployment of the obtained solutions. With respect to the diagram in Figure 3.1, we can group the project's phases into two main macro tasks, which are detailed in the following paragraphs such as:

- State of the art evaluation
- Applied research and development

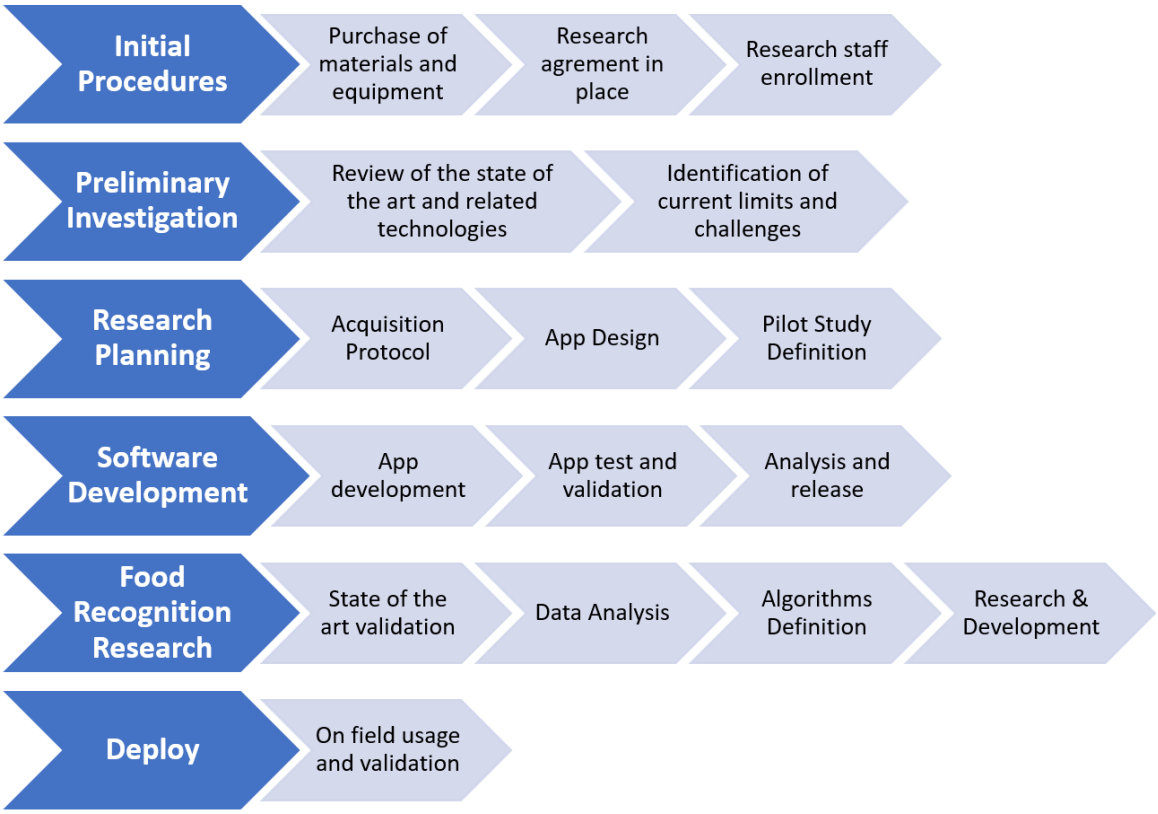


Figure 3.1: FoodREC project's phases.

3.1.1 State of the art Evaluation

After initial procedures (see Figure 3.1) have been completed, we focused on the preliminary investigation of the tasks and research problems related to the project purposes. This included the study of the state of the art related to food recognition and dietary monitoring technologies. As a result, a report concerning the existing products and approaches in terms of algorithms and smartphone apps has been produced and published in (Allegra et al., 2020), detailing the features and performances of each evaluated solution. The results of the study in (Allegra et al., 2020), revealed that modern food recognition techniques can support the traditional self-reporting approaches for eating diary, however, more efforts should be devoted to the definition of large-scale labeled image datasets. The new dataset design should focus on the quality of annotations related to the type of food, areas, quantities, and calories of each food item depicted in an image. So far, state-of-the-art focused on specific tasks performed in controlled conditions. The extreme variability of food appearance makes this task

challenging. Especially for ingredients inference and, hence, for nutritional values estimation. The study concludes that food recognition for dietary monitoring is still an in-progress technology and more efforts are needed to reach standards for reliable medical protocols, such as smoke-quitting programmes.

3.1.2 Applied Research and Development

After the study of the state-of-the-art and consequent analysis and definition of current limits and challenges, the research moved to the applied research and development phase. This phase has a dual objective. One is related to the development of the technological aspects of the framework, and the other one is related to the development of analysis algorithms. The iOS/Android FoodRec smartphone app for image acquisition and analysis and dietary monitoring has been released. The mobile app FoodRec has been designed with the objective of providing a smart and accessible system for the daily eating habits monitoring of the users, with the definition of a dietary diary.

The innovation that characterizes the FoodRec app is the automatisms related to the food analysis and associated inferences. Indeed, the user just uploads a picture on the system, then all inferences are performed automatically, by means of Computer Vision and Artificial Intelligence technologies. Figure 3.2 shows the main interface screens of the FoodRec app. First, a meal over four possibilities is chosen (a), then the app requires to state the mood associated to the meal (b), then the picture is taken (c) and uploaded (d). The app automatically learns the daytimes associated to food intake, and sends a notification to the user if the meal has not been inserted yet at the expected time. After the image is uploaded to the server, the recognition algorithms are applied, and the resulting inferences are shown in the app interface, as in the example shown in Figure 3.3. At this step, the user can edit the results (if needed) and confirm the new record for the eating diary. The information about user corrections are exploited for the further improvements of the algorithms, as well as their specialization with respect to the specific user habits.

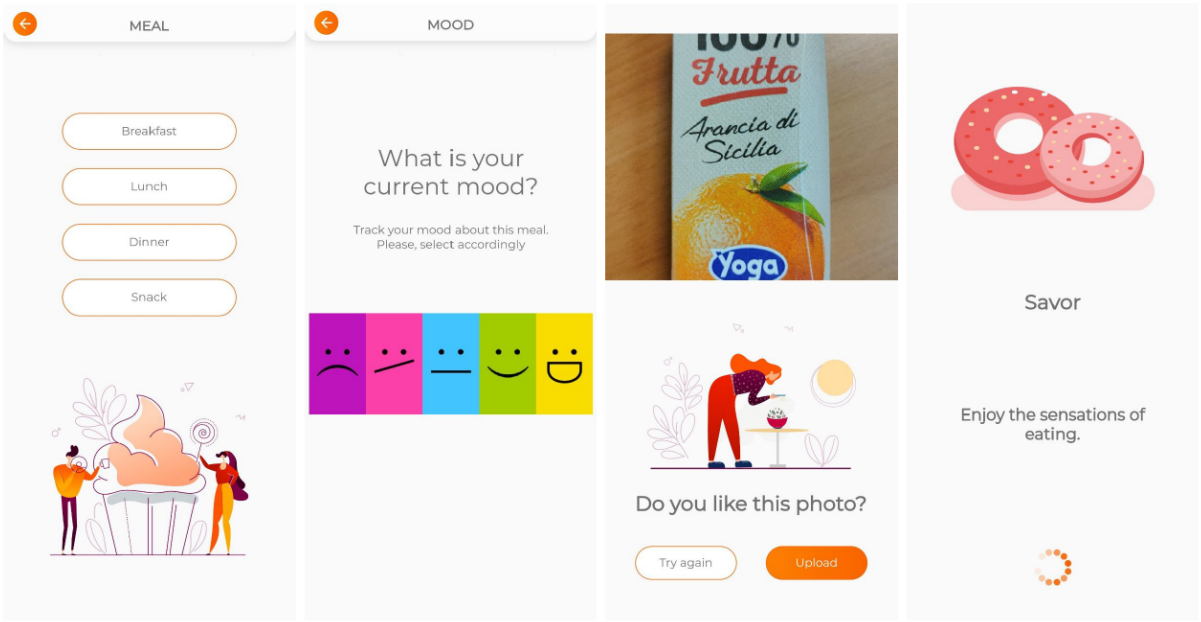


Figure 3.2: FoodRec example screens. Meal selection (a), mood associated to the meal (b), picture upload, motivational sentence (d).

FoodRec developed features also include water intake and weight tracker. Moreover, the user can inspect the statistics related to his/her eating habits, including the dominant food categories, ingredients, as well as temporal visualizations of specific parameters (see Figure 3.4).

The analysis algorithms will comprise several steps, including image normalization, registration, feature extraction, food detection, and classification. The research team is currently evaluating new methods and techniques for the improve of the performances of the food recognition algorithms exploited by the system. In particular, the efforts are devoted to three main tasks:

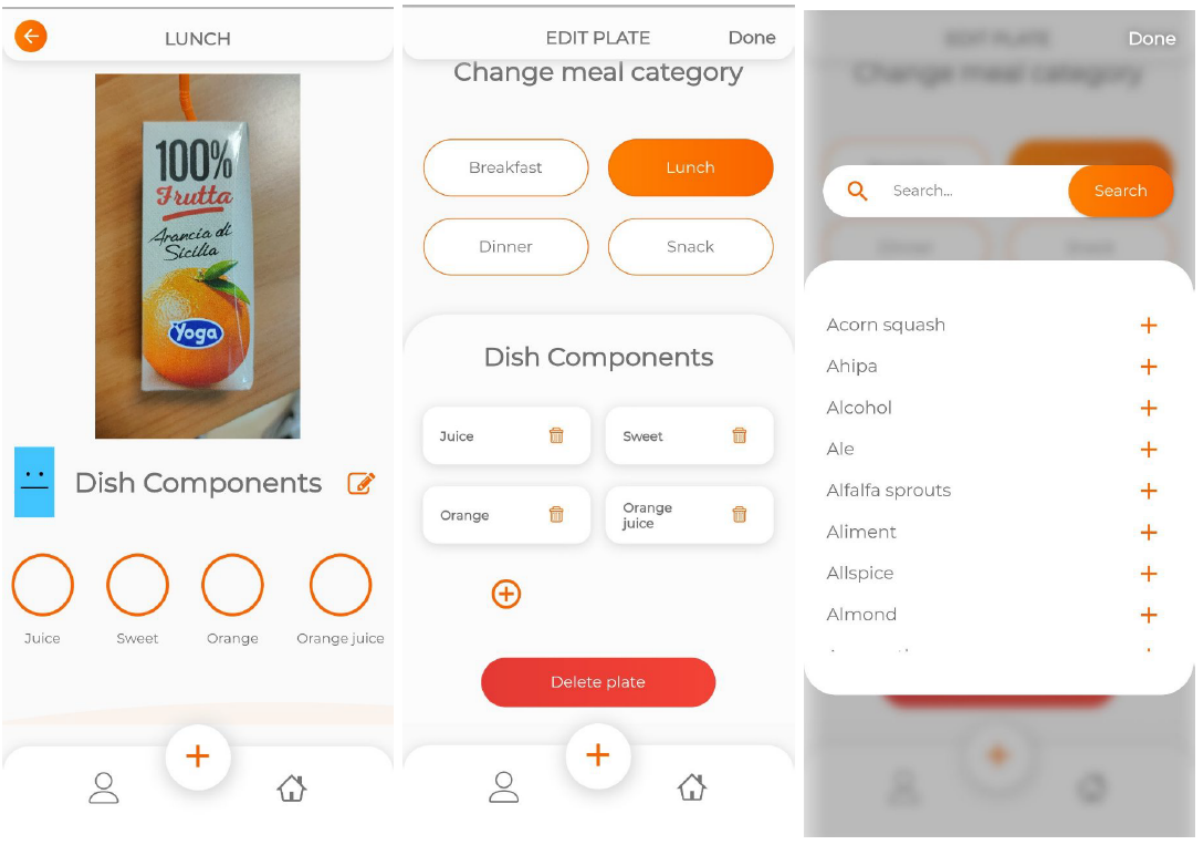
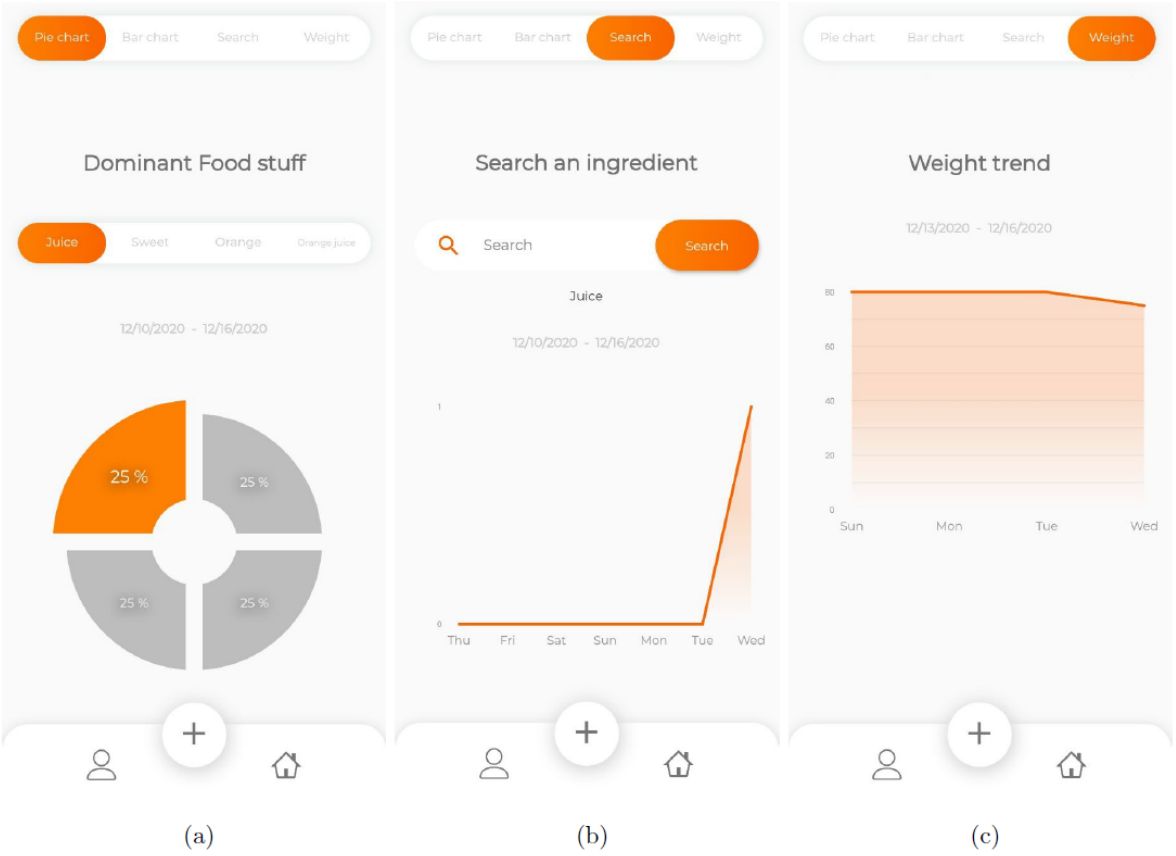


Figure 3.3: FoodRec interface showing the results of the food recognition system.

Food Segmentation: the aim of this task is the segmentation of the multiple food items that are depicted in a meal picture. This will output the areas of the pixels associated to each food item.

Food Classification: this classic task combined with the food segmentation output will provide a semantic segmentation of the input image, which details at pixel level the parts of the image related to specific food categories.

Volume Estimation: this task represents one of the most difficult aimed achievements. Indeed, the objective of this task is to estimate the volume of each food item. This task results very challenging because it involves the estimation of 3D information from monocular vision, at very small scale detail.



(a) (b) (c)

Figure 3.4: FoodRec in-app statistics.

At this current stage, research methods related to the above-mentioned tasks have been applied only on images available from the state-of-the-art in food recognition and image segmentation. However, we plan to specialize such algorithms on the data coming from the FoodRec app, which is specific with respect to our purposes. The proposed system aims to recognize food items of specific users and monitor their habits. This task significantly differs from the recognition of any food instance depicted by a picture, such as happens in the development of general purposes food recognition systems.

3.2 Expected Outcomes

Abstinence from smoking is associated with several negative effects, including irritability, gain of weight, and eating disorders, especially in the first period of abstinence. All

these effects are connected to each other. The output of the food recognition system will provide indications about the user's dietary habits and anomalies at different times during the smoke-quitting progress. The evaluation strategy will leverage well-known statistical methodologies for assessing correlations between the observed data and known information about the smoking quitting treatment.

3.3 Conclusion

At this stage, the initial procedures, preliminary investigation, and research planning phases of the project (see Figure 3.1) have been completed. The other phases, except the final deployment, are currently being carried out. Furthermore, the FoodRec app has been tested and evaluated with a small controlled group of test users. The tests were carried out for a period of about four months that began on 12 August 2020 and ended on 02 December 2020, with the participation of 164 people aged between 19 and 60. The Table 3.1 summarizes the mainly statistics and activities performed by the users in the aforementioned tests. In particular, the Table 3.1 shows the number of interactions there were among the users and the main features of FoodRec (i.e., the upload of a meal's photo or the update of the drunk water), instead the Table 3.2 reports the distribution of the uploaded photos among the following categories: breakfast, lunch, snack, and dinner. Once the tests have been ended, the users participated to a survey panel, reporting the feedback with respect to the app usage, which will be exploited to further improve the app features. The next step will be the evaluation on a larger audience of users in real-case scenarios (i.e., not controlled users). Such "on the wild" evaluation will produce a large set of real-case images from real users of the system, which will be exploited to develop novel algorithms and inference methods for the specific purposes of the project.

Table 3.1: Usage frequencies

Usage Statistics	**Counts**
Participants	164
Meals upload	1657
Drinks update	721

Table 3.2: Meals type frequencies

Meal Type	**Counts**
Breakfast	396
Lunch	553
Snack	305
Dinner	403

The developed dietary monitoring system could be extended to work with videos recorded by a fixed camera system, considering a set of cameras recording the scene from different fixed points of view. The collected data about the mood associated to food images can be combined with approaches related to sentiment analysis based on images (Ortis et al., 2020). Such approaches can be investigated in order to automatically infer the mood of the user (e.g., depression, happiness, etc.) based on dietary monitoring, avoiding to ask the user about his/her mood.

Chapter 4

Food Recognition

We present a user-biased food recognition system. The presented approach has been developed in the context of the FoodRec project, which aims to define an automatic framework for the monitoring of people's health and habits, during their smoke-quitting program. The goal of food recognition is to extract and infer semantic information from the food images to classify diverse foods present in the image. We propose a novel Deep Convolutional Neural Network able to recognize food items of specific users and monitor their habits. It consists of a food branch to learn visual representation for the input food items and a user branch to take into account the specific user's eating habits. Furthermore, we introduce a new FoodRec-50 dataset with 2000 images and 50 food categories collected by the iOS and Android smartphone applications, taken by 164 users during their smoking cessation therapy. The information inferred from the users' eating habits is then exploited to track and monitor the dietary habits of people involved in a smoke-quitting protocol. Experimental results show that the proposed food recognition method outperforms the baseline model 101 results on the FoodRec-50 dataset. We also performed an ablation study, which demonstrated that the proposed architecture is able to tune the prediction based on the users' eating habits.

Recognizing food from images is an extremely useful task for a variety of use cases. For example, it would allow people to track their food intake of what they consume by simply taking a picture, to increase awareness of their daily diet by monitoring their eating habits, the kind and amount of taken food, how much time the user spends eating during the day, how many and what times the user has a meal, analysis on user's habits changes, bad habits, and other inferences related to user's behavior and mood (Ortis et al., 2020). It can help a doctor to have a better opinion with respect to the patient's behaviour, in the applications on quitting treatment response, smoking detection and quitting technologies (Ortis et al., 2020), dietary monitoring during smoke quitting (Battiato et al., 2021) and smoking cessation system (Maguire et al.,

2021). Food monitoring plays a vital role in human health that is directly affected by diet (Nishida et al., 2004). Humans life is strictly affected by the food, this encourages researchers to introduce new methods for food logging and automatic food dietary monitoring (Kitamura et al., 2010), food retrieval and classification (Farinella et al., 2016). We present a novel food recognition method that takes into account the specific user to systematically analyze and infer his/her eating habits. The idea is to introduce a bias related to the user in the food classification pipeline. In particular, inspired by deep learning approaches applied on text representation learning (Le and Mikolov, 2014), the proposed architecture learns a user's eating habits feature representation space. We also collected a new FoodRec-50 dataset that will be used for evaluation of the food recognition technology for dietary monitoring during smoke quitting.

The main steps involved in our food recognition pipeline are as follows:

1. Food data acquisition
2. Food data annotation
3. Data augmentation and normalization
4. Food recognition

In this chapter, we will further present in detail the proposed architecture that consists of food branch and user branch, user data annotation to embed the user information with image label, data augmentation for the classification task, different users eating habits, and finally proposed FoodRec model results comparison with the baseline 101 model. We have conducted the classification experiments on our FoodRec-50 data, as we have extracted 50 categories/ classes from our FoodRec Data collected by the App. ResNet model for classification is fine-tuned on the 50 classes of our FoodRec-50 data. The ResNet model will act as the baseline model for the classification task. Experimental results show that the proposed FoodRec model performs better as compared to the baseline results on the FoodRec data.

4.1 Food Data Acquisition

The objective is to build a new unique robust dataset useful for the food recognition technologies development and evaluation stages. Our dataset is specific to the users who are involved in the smoke-quitting process to monitor their dietary habits. The dataset is produced to study the correlation between eating information with smoking habits. In the future, such data will be used to find correlations with respect to the

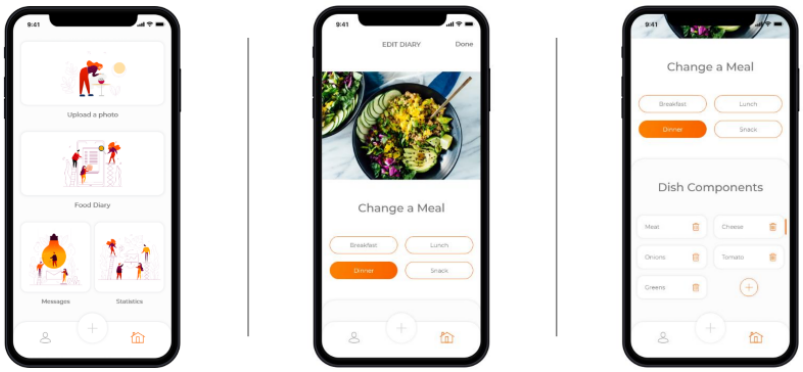


Figure 4.1: Food Data Acquisition using Smartphone app

smoking activity of the subjects during the period of observation. To collect the food data, the iOS and Android smartphone apps are used, as shown in Figure 4.1. Users can upload a meal intake image by taking a picture of what they eat and can assign labels to the food image what it contains.

4.2 Food Data Annotation

Annotations are required during the supervised training of the network and also to test the examples during the evaluation phase for the food recognition method. To perform the experiments, we have first extracted food images for 50 classes from the FoodRec data. However, some of the classes (beans, breadstick, carrot, chickpeas, corn, popcorn, grape, peas, zucchini, etc.) have few images, so data augmentation is performed to compensate the problem with underrepresented classes. Further, data is annotated manually for training and evaluation of the model, which contains around 1100 images.

4.3 Proposed FoodRec Architecture

We proposed FoodRec architecture for data coming from the FoodRec app, which is specific with respect to our purposes. The proposed system aims to recognize food

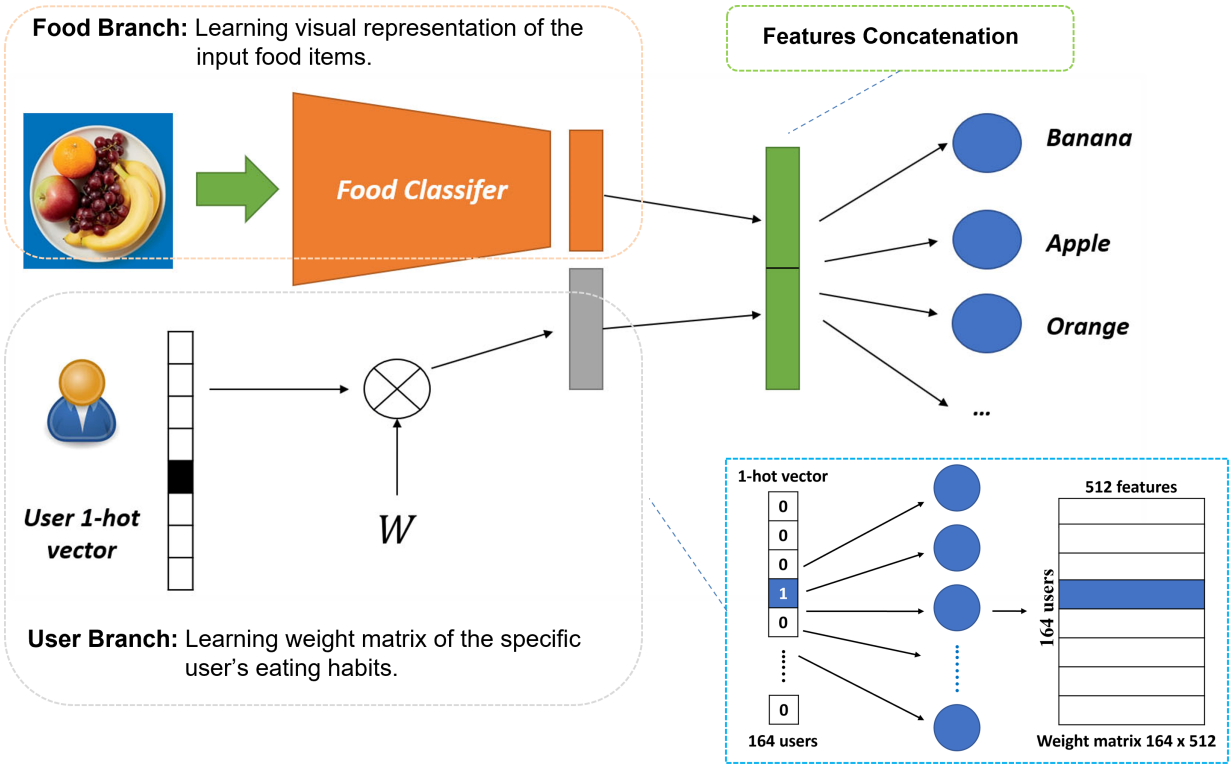


Figure 4.2: Food recognition proposed architecture.

items of specific users and monitor their habits. This task significantly differs from the recognition of any food instance depicted by a picture, such as happens in the development of general purposes food recognition systems. Figure 4.2 shows the proposed FoodRec architecture. In particular, a common multi-label food classifier is composed by a Convolutional Neural Network, which defines a meaningful feature representation for the input images, based on the training task. Then, the representation is fed to multiple logistic units (i.e., blue circles in the Figure), which are activated if the associated food item is present in the picture. The proposed architecture will take into account the specific user that uploaded the picture. Indeed, since the proposed system is aimed to systematically analyze and infer user habits, our objective is to add to the food classification pipeline a bias related to the user. As a consequence, the individual logistic activations will be fed with a feature that is obtained by concatenating the image and user feature. The latter one, is represented by the weight matrix W in Figure 5, which will be learned from the users' habits during the training stage.

Our proposed FoodRec Model consists of two branches, the food branch and the user branch to extract concatenated feature map from two branches to recognize the food items. The food branch is learning visual representation of the input food items

like the traditional food recognition algorithm. On the other hand, the user branch takes into account the specific user's eating habits by learning the users eating weight matrix. The user branch takes one hot vector user input and learns users eating weight matrix. The food branch extracts feature map of the food image using ResNet101, and further a fully connected layer is applied to get a feature vector. The user branch extracts the feature vector from the user bias using a fully connected layer. Finally, features from food branch and user branch are concatenated to further perform the final prediction. User one-hot vector just selects the weights corresponding to the user eating habits learned during the training with the food branch. As we change the user bias input, the result in the prediction is being changed according to the dietary habits of that user. The proposed network with one-hot user vector is effective because it learns specific user eating habits with the food as compared to the traditional food recognition systems with competitive results.

4.4 User Data Annotation

As previously, we have extracted data for 50 classes but with only food items annotations. Proposed FoodRec architecture requires users' annotation as well along with the food items eaten by them to train and test the model. Hence, data is annotated to embed the user information with image label so that we can feed the data simultaneously to the network. Now, FoodRec data image names contain user ID and image label. The idea is to extract the user ID from the image name while reading the images so that we can implement the dataloader with the simultaneous user input and corresponding image data for both branches for the experiments. Data loading is the initial stage in constructing a deep learning pipeline or training a model. When the data becomes more complicated, this task becomes more difficult. DataLoader class in PyTorch provides a helpful solution to this challenge through its dataLoader class. PyTorch's DataLoader class is important for efficiently loading and iterating over elements in a dataset. It lets you determine how the data is loaded, including batch size, shuffling, etc. Furthermore, it can be used to load data from a variety of sources, such as images, text, and audio.

4.5 Data Augmentation

Data augmentation is a commonly used technique in many state-of-the-art deep learning and computer vision applications including image recognition, object detection, and semantic segmentation. These methods are applied to increase the amount of

training data available to deep neural networks, which helps enhance their performance and avoid overfitting. However, obtaining sufficient data for training can be challenging due to various reasons. For example, it could be hard to collect the required data due to restrictions and costs involved in this process. One of the reasons that makes it difficult to create a sufficient dataset is to label the data with the appropriate category after collecting the image data. For instance, image classification requires to assign the correct class labels to each image. Therefore, labeling the data is necessary for tasks such as image classification, object detection, and semantic segmentation. This process is manual, and labeling the data can be very expensive. Data augmentation is one way to overcome these limitations when you have not enough data to feed to the deep neural network. The goal of image augmentation is to produce new and diverse data samples from the existing data. Data augmentation is a technique that is used to transform data in order to improve a deep learning model's ability to recognize different variations of an image. Consequently, this improves the amount of information available to the model to learn the feature representation, which can enhance its quality and performance.

We used Albumentations by (Buslaev et al., 2020) for the image data transformations. Albumentations provides pixel-level transformations and spatial-level transformations to augment the image data. Pixel-level transformation alters the input image only, which includes Blur, GaussNoise, Equalize, HistogramMatching, InvertImg, RandomBrightnessContrast, RandomShadow, Sharpen, ToGray, ImageCompression, etc. Spatial-level transformation alters the input image with the corresponding mask or bounding box as well, which includes CenterCrop, HorizontalFlip, PadIfNeeded, RandomCrop, RandomRotate90, RandomSizedBBoxSafeCrop, ShiftScaleRotate, Affine, Resize, Transpose, etc.

Although, FoodRec data consists of 1100 food images for 50 classes but some of the classes like beans, breadstick, carrot, chickpeas, corn, popcorn, grape, peas, zucchini, etc. still have very few images. Here, we go with the data augmentation technique to deal with the lack of data. We have selected the top 20 users with the highest eating frequencies for all the food items. So, we have augmented the data for these users to produce many altered and transformed versions of the same image. Image augmentation increases the training data as we don't have enough data with some food categories containing less food images and makes a classifier more robust with a wide variety of transformed images. Different transformations are applied to the data as given below:

1. Image resize
2. Image random crop
3. Image horizontal and vertical flip

4. Image Random rotate
5. Image Motion Blur
6. Image Optical Distortion
7. Image Gaussian Noise
8. Random Brightness and Contrast
9. CLAHE Adaptive Histogram Equalization
10. Hue and Saturation Value

Three images have been taken from each food category and augmented for the the top 20 users with the highest eating frequencies for all the food items. For example, figures 4.3 and 4.4 show three different juice augmentation. FoodRec data consists of around 2000 images after data augmentation.

Figure 4.3: Juice Category Three Original Images Subjected to Transformation

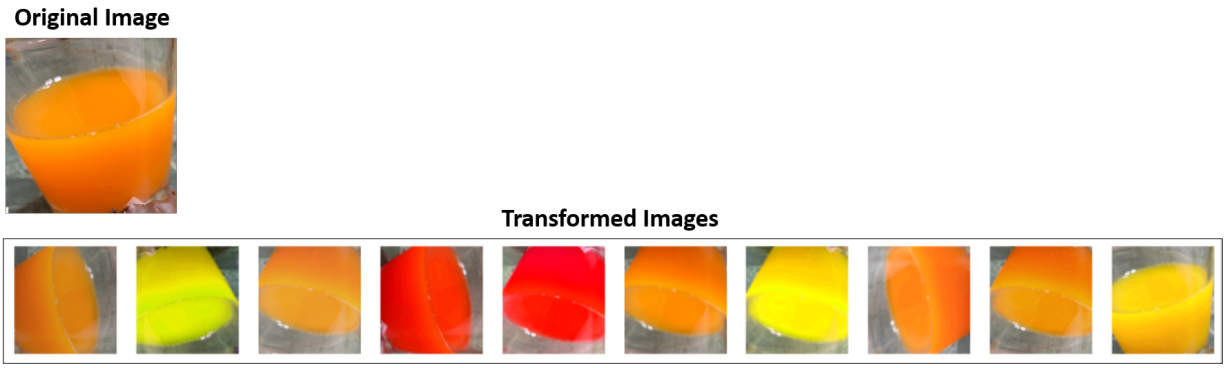


Figure 4.4: Three different juices original images with transformed images

4.6 Users and Their Eating Habits

At this stage, FoodRec data consists of 164 users eating different food items. Eating habits only for users 87, 109, 55, 27, and 117 for each food item is listed below in Table 4.1 with individual food items and their eating frequency for that user. These users are chosen intelligently using the Euclidean Distance function to observe the difference in decision-making. We defined an ”eating matrix” U x N, where U is the number of users (164 rows) and N is the number of considered food items (50 columns). Each row corresponds to the eating frequencies for each food item of a specific user. Then, we computed the distance for each user to find two users (87, 109) with the maximum distance between their eating vectors. Further, we calculated the sum of eating frequencies for each user and selected one user (55) with average and two users (27, 117) with the lowest sum of eating frequencies.

Table 4.1: Users and their eating frequencies

Food Items	Users Eating Frequencies					Food Items	Users Eating Frequencies				
	User 109	User 87	User 55	User 27	User 117		User 109	User 87	User 55	User 27	User 117
Almond	1	1	1	0	0	Green Tea	2	1	1	0	0
Apple	2	6	2	0	0	Jam	3	8	1	0	0
Arugula	3	1	4	0	0	Juice	3	1	2	0	0
Banana	1	3	1	0	0	Lentil	2	3	2	0	2
Bean	1	1	1	0	0	Meat	6	1	1	0	0
Biscuit	5	1	3	0	0	Milk	8	1	1	0	0
Blueberry	1	1	1	0	0	Mushroom	1	1	1	0	0
Bread	1	8	3	1	0	Orange	1	2	2	0	0
BreadStick	1	1	1	0	0	Pasta	6	1	3	1	0
Cake	2	3	2	0	0	Peas	1	3	1	0	0
Carrot	1	1	1	0	2	Pizza	5	2	2	0	1
Cereal	3	1	1	0	1	Popcorn	1	1	3	0	0
Cheese	2	2	2	0	0	Pork	2	1	1	1	0
Chicken	1	1	2	1	0	Potato	1	1	2	0	0
Chickpeas	1	1	2	0	0	Rice	2	2	1	0	0
Chips	2	1	1	0	0	Salad	1	2	1	0	0
Chocolate	1	2	2	1	0	Soup	2	1	1	0	0
Coffee	2	2	11	2	0	Spaghetti	2	2	1	0	0
Corn	1	1	1	2	0	Strawberry	1	1	1	0	0
Cracker	1	1	3	0	2	Tea	2	2	3	1	1
Croissant	2	1	1	0	0	Tomato	2	1	1	0	1
Doughnut	1	1	1	0	0	Tortellini	2	1	1	0	0
Egg	2	1	3	0	1	Vegetable	1	1	1	0	0
Fish	2	3	1	0	0	Yogurt	2	1	1	0	0
Grape	1	3	1	0	0	Zucchini	1	1	1	0	0

4.7 Experimental Results

Our proposed food recognition method results are compared with the baseline ResNet101 to evaluate the performance. Figures 4.6, 4.7, 4.8 and 4.9 show the results comparison. For example, Figure 4.6 contains a test food image with two food items, coffee and biscuits. The baseline represents the ResNet model trained only with food images, and results are shown next to the test image in the figures. Then, the proposed FoodRec Model results with five different users are shown in the Figure. The baseline model consists of a pre-trained ResNet101 (He et al., 2016) model trained on ImageNet that is fine-tuned to extract a 1024-dimensional features vector to perform the food items classification. This model is trained using only the food images like the traditional classification algorithm without taking into account the user bias into the final decision-making to classify the food items.

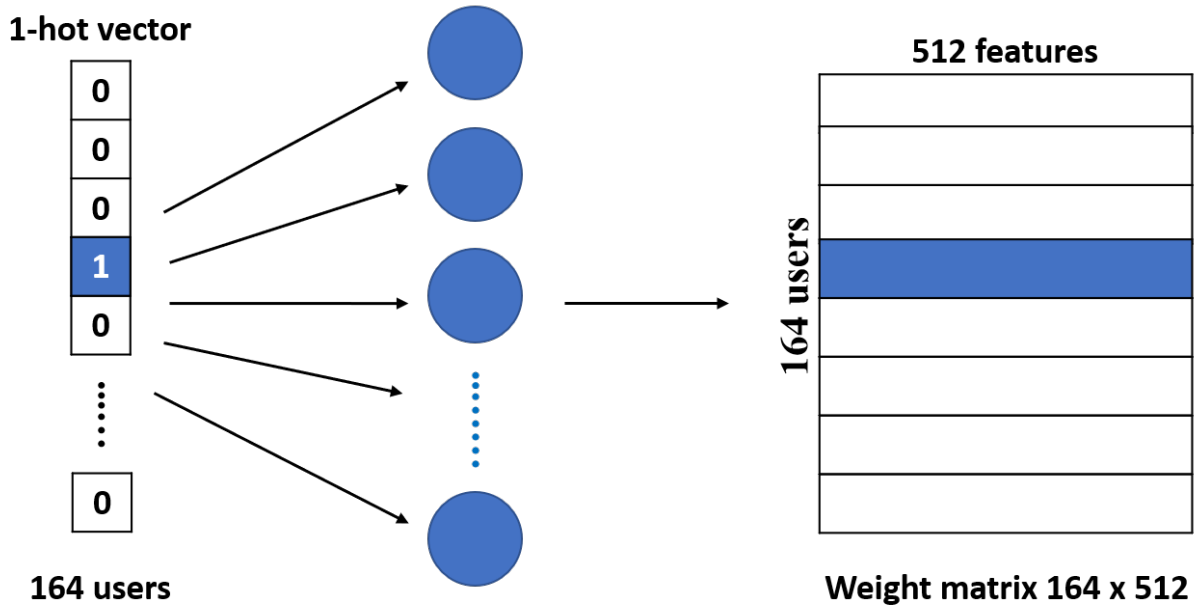


Figure 4.5: User branch of the proposed network to learn specific user eating habits.

The proposed food recognition model consists of two branches, the food branch, and the user branch to extract a 1024-dimensional concatenated feature map from these branches to recognize the food items. The food branch extracts 1024-dimensional feature map of food image using ResNet101 architecture with transferred weights from ImageNet dataset containing 1000 image categories, and further averaging pooling layer, flatten layer, and fully connected layer are applied to get a 512-dimensional feature vector. The user branch extracts a 512-dimensional feature vector from the user bias using a fully connected layer. Finally, the output 1024-dimensional feature vector is obtained by concatenating the features extracted from both branches.

The user branch of the proposed network is learning 164 user eating weight vectors with 512 features. So, the user weight matrix can be represented with 164 rows (one for each user) and 512 columns. as shown in Figure 4.5. If you multiply a 1 x 164 one-hot vector by a 164 x 512 matrix, it will just select the weight matrix row corresponding to the user eating habits learned during the training with the food branch.

The proposed network with one-hot user vector is effective because it learns specific user eating habits with the food image features as compared to the traditional food recognition systems. This approach is inspired by the document representation approach known as doc2vec presented by Le et al. (Le and Mikolov, 2014). Indeed, the model presented in (Le and Mikolov, 2014) implements a document representation architecture in which the word/sentence features are affected by the document from

which they have been extracted. In this way, the same word/sentence is represented differently depending on the source document, which acts as a context for the encoded words and is represented as a one-hot input vector. The document branch is combined with another network branch devoted to represent single words, as we do combining the branch representing the food image with the one representing the user will act as a bias for the image representation.

The proposed and baseline networks are trained using the same settings. Networks are trained using the Adam optimizer and the cross-entropy loss function. The learning rate is set to 0.001, the batch size is set to 32, and networks are trained for 200 epochs. The FoodRec-50 data consists of 164 users eating different food items. Eating habits only for users 87, 109, 55, 27, and 117 for each food item are listed below in Table 4.1 with an individual food item and its eating frequency for that user. These users are chosen using the Euclidean Distance function to observe the difference in decision-making. The distance between user eating frequencies tells how much one user eating habits are different from the other. Therefore, the distances between user eating frequencies have been used to select the users with different habits and perform specific tests aimed to assess the efficacy of our approach and its capability to encode the user eating habits. The effect can be observed in Figures 4.6 and 4.7 showing the results. While the baseline method finds difficulties in the recognition of multiple food items in cluttered scenarios, the proposed method shows better performances, especially for users that have high frequencies for the food items present in the test image. We defined an "eating matrix" U x N, where U is the number of users (164 rows) and N is the number of considered food items (50 columns). Each row corresponds to the eating frequencies for each food item of a specific user. Then, we computed the distance for each user to find two users (87, 109) with a maximum distance between their eating vectors. Further, we calculated the sum of eating frequencies for each user and selected one user (55) with average and two users (27, 117) with the lowest sum of eating frequencies.

The proposed FoodRec model results are compared with the baseline ResNet model. Figures 4.6 and 4.7 show multi-label food classification results comparison. Food and user concatenated representation fed to the logistics units containing the sigmoid function to produce the independent probabilities for specific food classes. In particular, the output of each sigmoid is the probability that the input belongs to one specific food item. In other words, each sigmoid outputs $P(class = Banana|x), P(class = Bread|x)$, etc. Therefore, the score shown in Figures 4.6 and 4.7 for the food types is the percentage of the sigmoid output probabilities for each food item.

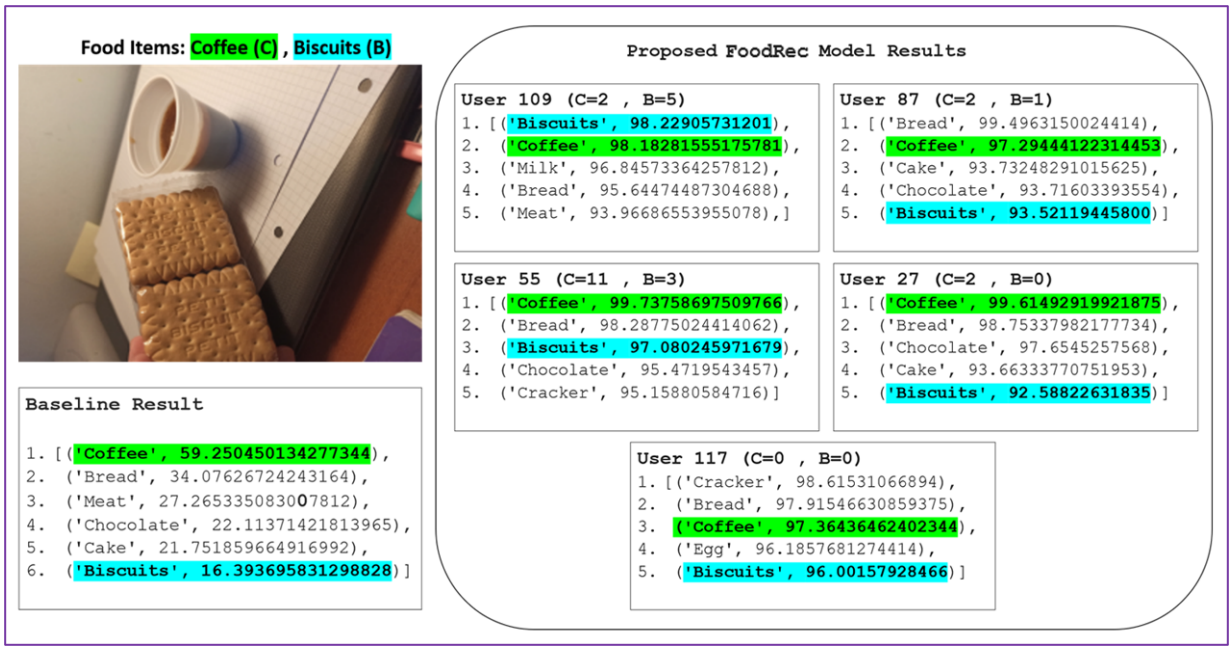


Figure 4.6: Results comparison of test image containing Coffee and Biscuits.

For example, Figure 4.6 contains a test food image with two food items coffee and biscuits with multi-label predictions. The baseline represents the ResNet model trained only with food images, and results are shown below the test image in the figures. Then, the proposed FoodRec model results with five different users are shown next to the test image in the figures. We can observe from Figure 4.6 that the baseline model recognizes the coffee with a score 59.25 at the very first place, but biscuits with a score 16.39 occur at the 6th place in the prediction order. On the other hand, the proposed FoodRec recognizes the same two food items with improved score and occur in the top five predictions for all five users. For user 109, the top two predictions are biscuits and coffee with scores 98.22 and 98.18, respectively. This happens because the user with ID 109 has a relatively higher number of instances for these kinds of food in the eating matrix. For user 117, although it did not drink coffee (C=0) or eat biscuits (B=0) but the model recognizes these food items at 3rd and 5th places respectively because the model learns both image and user features. Similarly, you can also observe the difference between the FoodRec and the baseline models in Figure 4.7 with another test food image.

Figure 4.7: Results comparison of test image containing Apple and Juice.

The FoodRec model improves the score as well as learns the user's dietary habits because the model is learning weight matrix for the users. As we change the user bias input, the result in the prediction is being changed according to the dietary habits for that user, as you can observe in the given figures. So, adding a bias related to the user to the food classification pipeline is effective to systematically analyze and infer user's habits. Users and their eating habits can be observed in Table 4.1. Moreover, the proposed model improves the general food recognition task with respect to the baseline model, as shown in Table 4.2.

Table 4.2: Results comparison

Method	User-biased	Top-5 Accuracy (%)
Baseline Model	No	59.6
Proposed Model	Yes	71.1

Food Items: Bread (B) , Potato (P) , Cheese(C) , Tomato (T)

Baseline Result

```
1. [('Potato', 19.795429229736328),
2.  ('Pasta', 12.008089065551758),
3.  ('Egg', 8.844963073730469),
4.  ('Bread', 6.170302391052246),
5.  ('Cake', 4.763753890991211),
6.  ('Chicken', 4.683365345001221),
7.  ('Fish', 3.6393024921417236),
8.  ('Mushroom', 3.4651060104370117),
9.  ('Meat', 3.197643995285034),
10. ('Pork', 2.52731990814209)]
```

Proposed FoodRec Model Results

User 109 (B=1 , P=1 , C=2 , T=2)

```
1. [('Pasta', 97.41432189941406),
2.  ('Potato', 96.74486541748047),
3.  ('Egg', 94.59818267822266),
4.  ('Meat', 93.17474365234375),
5.  ('Fish', 93.00230407714844),
6.  ('Cheese', 92.45442199707031),
7.  ('Rice', 87.18347930908203),
8.  ('Pork', 85.9800033569336),
9.  ('Yogurt', 83.33782196044922),
10. ('Bread', 83.22250366210938),
```

User 87 (B=8 , P=1 , C=2 , T=1)

```
1. [('Bread', 97.80791473388672),
2.  ('Potato', 97.17204284667969),
3.  ('Fish', 95.9795150756836),
4.  ('Egg', 92.3043212890625),
5.  ('Pasta', 92.19780731201172),
6.  ('Cheese', 91.2406234741211),
7.  ('Rice', 87.45726013183594)]
```

User 55 (B=3 , P=2 , C=2 , T=1)

```
1. [('Potato', 97.69432830810547),
2.  ('Egg', 95.74720001220703),
3.  ('Pasta', 94.9932632446289),
4.  ('Bread', 92.83966827392578),
5.  ('Cheese', 90.92951202392578),
6.  ('Chicken', 90.1448974609375)]
```

User 27 (B=1 , P=0 , C=0 , T=0)

```
1. [('Potato', 98.07691955566406),
2.  ('Pasta', 97.83070373535156),
3.  ('Chicken', 95.5448226928711),
4.  ('Bread', 94.70703887939453),
5.  ('Egg', 93.17799377441406),
6.  ('Fish', 92.02178955078125),
7.  ('Pork', 91.51859283447266),
8.  ('Rice', 87.1675033569336),
9.  ('Cheese', 86.87367248535156)]
```

User 117 (B=0 , P=0 , C=0 , T=1)

```
1. [('Egg', 97.92064666748047),
2.  ('Potato', 96.99077606201172),
3.  ('Pasta', 92.27947998046875),
4.  ('Fish', 91.95279693603516),
5.  ('Bread', 91.38661193847656),
6.  ('Cheese', 91.19795227050781),
7.  ('Rice', 88.22586059570312)]
```

Figure 4.8: Proposed FoodRec Model Results Comparison with Baseline Model

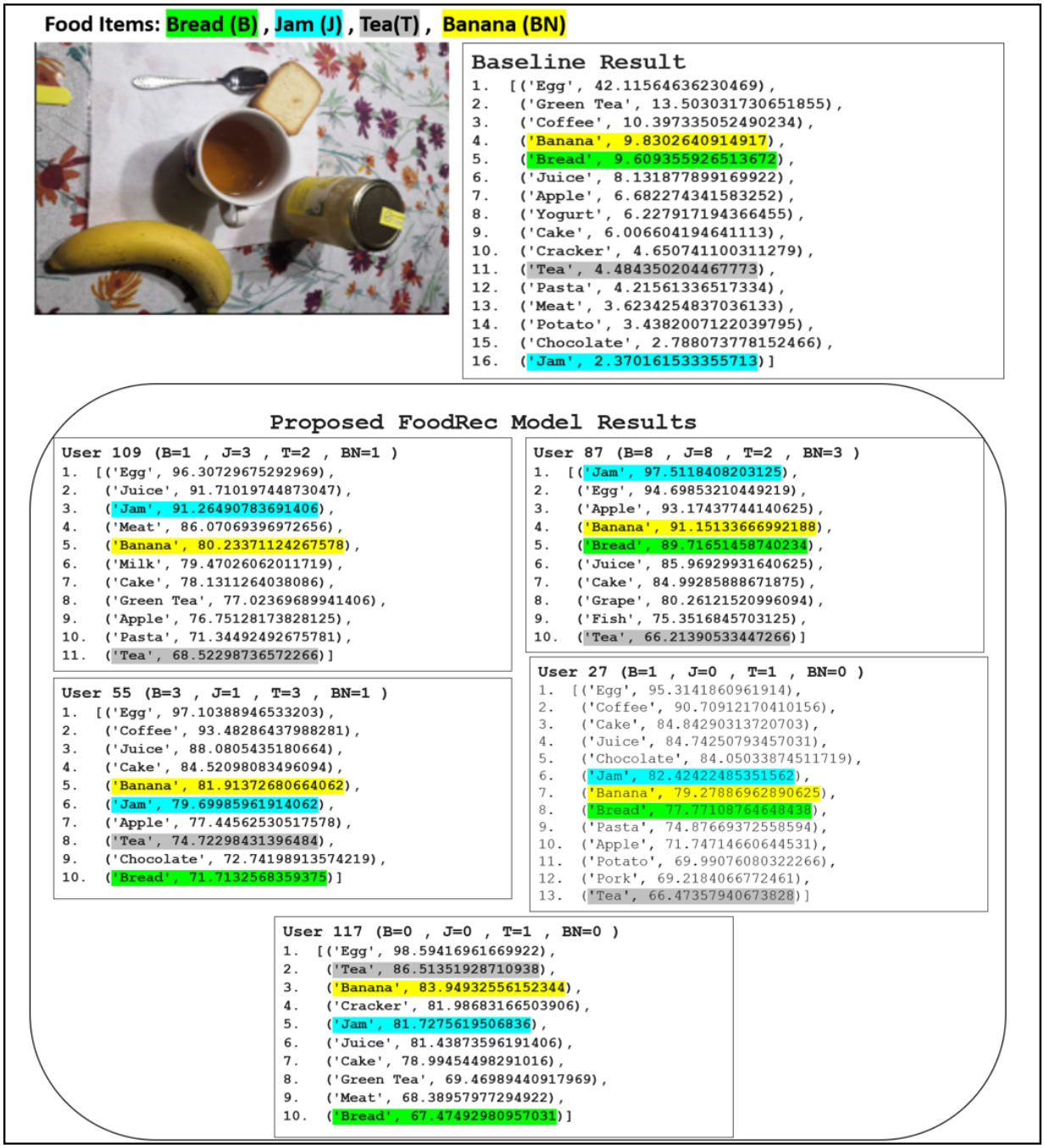


Figure 4.9: Proposed FoodRec Model Results Comparison with Baseline Model

4.8 Conclusion

We developed an automatic framework for food recognition using computer vision and deep learning techniques that plays a significant role to the health and food intake of people. The developed system acquires images of the food eaten by the user or subject over time, which will then be processed by the proposed food recognition model to extract and infer semantic information from the food images. We propose a novel Deep Convolutional Neural Network able to recognize food items of specific users and monitor their eating habits. It is composed of a food branch and a user branch. The food branch employs Convolutional Neural Network for learning highly descriptive feature representations of the food items. The user branch takes into account the specific user's eating habits by learning the users eating weight matrix. User one-hot vector just selects the weight matrix row corresponding to the user's eating habits learned during the training with the food branch. As we change the user bias input, the result in the prediction is being changed according to the dietary habits. The proposed network with one-hot user vector is effective because it learns specific user eating habits with the food image features as compared to the traditional food recognition systems. Experiments show that the proposed user-biased food recognition is effective and achieves higher results as compared to the baseline. The proposed model is hence able to influence the prediction by encoding the user bias as a result it also improves the task performance.

Chapter 5

Food Segmentation

We present semantic food segmentation to detect individual food items in an image. The proposed approach has been developed in the context of the FoodRec project, which aims to study and develop an automatic framework to track and monitor the dietary habits of people, during their smoke-quitting protocol. The goal of food segmentation is to train a model that can look at the images of food items and infer semantic information to recognize individual food items present in an image. In this contribution, we propose a novel Convolutional Deconvolutional Pyramid Network (CDPN) for food segmentation to understand the semantic information of an image at a pixel level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. Furthermore, the proposed method demonstrated significant improvements on two well-known public benchmark food segmentation datasets.

Food segmentation plays a key role in the context of food recognition technology for dietary monitoring (Battiato et al., 2021) to predict and detect multiple food items present in an image. The output of a food segmentation system is a set of image regions associated to each detected food item to provide a semantic segmentation map of the input image. This accurate segmentation of food regions can be used to estimate the volume and, hence the quantities of each food item detected within an image. This would allow people to estimate the assessment of calories and nutrients to track their food intake of what they consume to increase awareness of their daily diet by monitoring their eating habits, the type and amount of food, how often and what times the user eats a meal, how much time he spends eating in a day, advanced inferences performed can be compared to make correlations between eating habits and quitting process steps, bad habits, user's behavior and mood changes (Ortis et al., 2020) over time. The semantic organization of daily habits can help a doctor to have a better

opinion with respect to the patient's behaviour and habits changes, in the applications on quitting treatment response and health needs, smoke monitoring technology (Ortis et al., 2020), dietary monitoring during smoking cessation and smoke quitting program (Maguire et al., 2021). Food plays a crucial role in human life that is strongly affected by diet (Nishida et al., 2004). Then, food recognition technology and its applications especially in the health department (Allegra et al., 2020) for dietary and calorific monitoring motivated computer vision specialists to develop new methods in the areas such as food logging and automatic food dietary monitoring (Kitamura et al., 2010), food retrieval and classification (Farinella et al., 2016), food recognition to monitor users' eating habits (Hussain et al., 2022), and segmentation for food understanding and analysis. We present a novel Convolutional Deconvolutional Pyramid Network for semantic food segmentation. Experiments are conducted on the tray food dataset that reveals significant improvements in the results.

The food segmentation is performed on the benchmark food datasets such as MyFood and TrayDataset. This enables to make comparisons and to measure the performance of the proposed method with state-of-the-art food segmentation techniques. The FoodRec dataset is useful for the segmentation but not annotated for the segmentation task at this moment. In fact, cross-data experiments are conducted by training the model on the MyFood dataset and testing on the FoodRec dataset to evaluate the performance of our proposed segmentation method. Both MyFood and our FoodRec datasets contain common food classes, such as apple, beans, egg, spaghetti, chicken, rice, and salad. The qualitative results of our proposed method are evaluated on a subset of our FoodRec dataset. By performing the cross-data experiments, we aimed to simulate a scenario where the model is trained on a MyFood dataset and is tested to perform segmentation on the MyFood dataset. The results of our experiments demonstrate that our proposed method is capable of generalizing well to new and unseen FoodRec dataset as well.

5.1 Proposed Convolutional Deconvolutional Pyramid Network

The food segmentation aims to develop a model which is able to extract and infer semantic information from the food images at pixel-level to recognize different food items present in an image. This would further allow people to estimate the assessment of calories to track their food intake and to increase awareness of daily diet by monitoring their eating habits. In this context, a novel Convolutional Deconvolutional Pyramid Network (CDPN) is proposed for image semantic segmentation, which takes food image as input and outputs a segmentation map of the individual food items

detected, as described in Fig.1. A pre-trained convolution neural network is used to harvest meaningful feature representation. Then, a feature pyramid is built with multi-scale feature maps representation. The proposed network employs convolution and deconvolution layers to generate a feature pyramid and achieves high-level semantic feature map representation. The deconvolution develops upsampling of the input features using learnable parameters to produce generalized upsampling of the feature representation. As a result, the proposed segmentation network generates a dense and precise segmentation map of the input food image.

Initially, we applied deep ResNet architecture with pre-activations due to its capability for learning highly descriptive feature representation by downsampling spatial resolution for complex scenes. The ResNet architecture with transferred weights from the ImageNet dataset extracts two feature set representations from the input RGB food image. More precisely, the proposed Convolutional Deconvolutional Pyramid Network obtains two discriminative feature sets of 512 channels with a downsampled spatial resolution (h/8, w/8), and 2048 channels with a downsampled spatial resolution (h/32, w/32) from an input image (height(h), width(w)) using Resnet-101 as the backbone network. The convolution coupled with upsampling layers is used to the low-resolution spatial information to produce high-resolution semantically strong segmentation map. Then, the convolution layer and deconvolution layer are employed to each feature set obtained by the ResNet to generate four high-level feature maps pyramid. The deconvolution layers densify the feature map with learned filters to output upsampled and rich feature map representation. As a result, a feature pyramid is produced with four multi-scale feature maps of sizes h/4 x w/4 x 256 channels, h/8 x w/8 x 256 channels, h/16 x w/16 x 256 channels, and h/32 x w/32 x 256 channels by applying convolution layer with 1x1 kernel size and deconvolution layer with 2x2 kernel size on each feature set obtained by the ResNet. Feature maps of each level are fused and concatenated through the convolution layer with 3x3 kernel size, group normalization, ReLU, and upsampling layer to resize the feature map using interpolation. Further, a deconvolutional layer is applied with 4x4 kernel size and output channels equal to the number of food categories followed by a softmax layer for the food segmentation pixel-wise predictions. The final output of the proposed CDPN network is a segmentation map of the same size as the input image that represents the probability of each pixel belongs to one of the food classes.

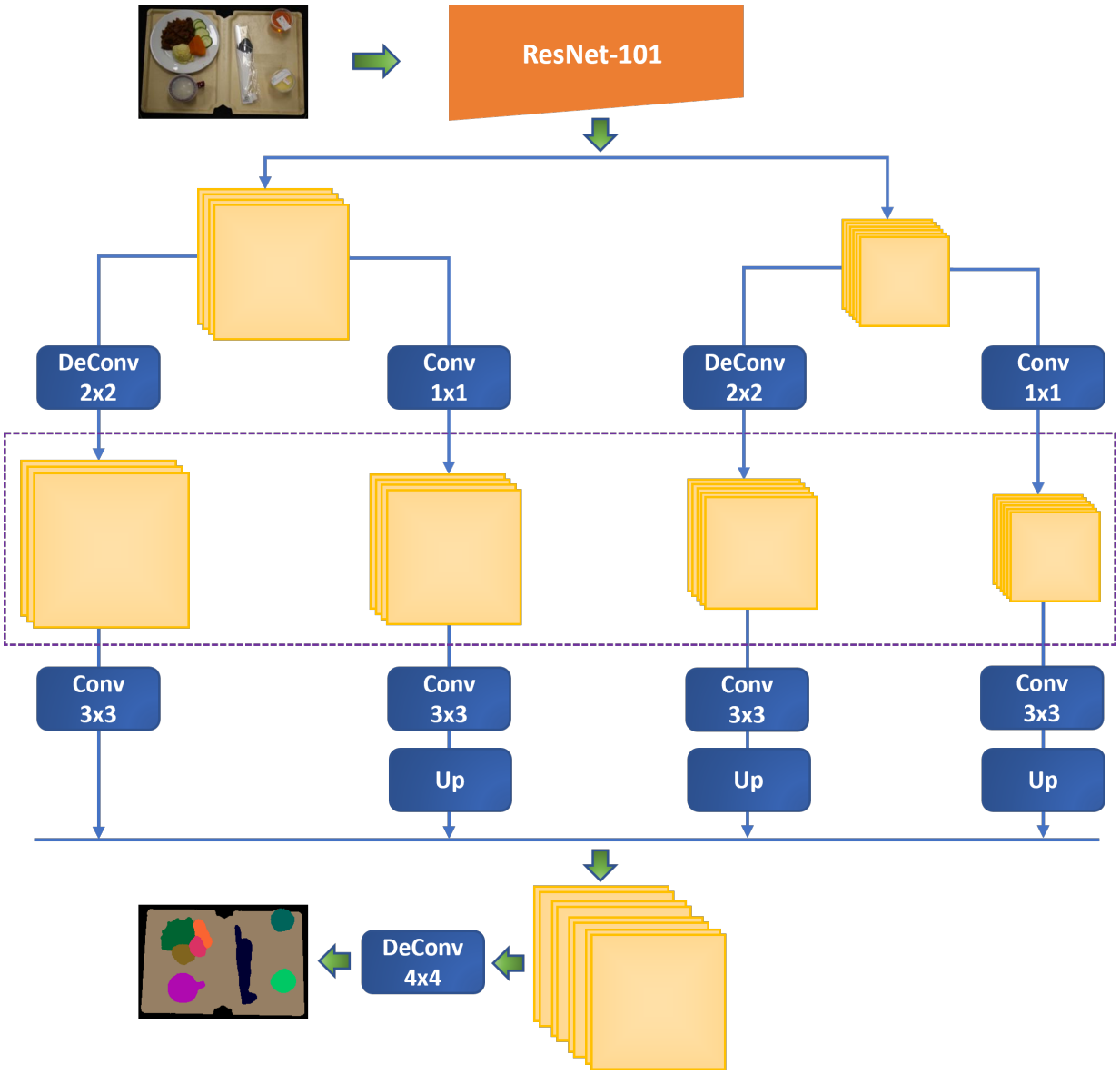


Figure 5.1: The proposed Convolutional Deconvolutional Pyramid Network for semantic food segmentation. This network takes food image as input and outputs a segmentation map of the individual food items present in the image. In the architecture, "Conv" represents the convolutional layer, "DeConv" represents the deconvolutional layer, and "Up" represents an upsampling layer. All convolutional layers with 3x3 kernel size are followed by group normalization layers and ReLU activation layers.

5.1.1 Performance Metrics

We used common performance metrics for segmentation to compare quantitative performance of the proposed method with others. To evaluate the comparative performance, we use intersection over union (IoU) which is a widely used metric in image segmentation tasks to measure the quantitative performance. This metric measures the degree of overlap between the target and predicted segmentation by comparing the number of common pixels. The IoU in Equation 5.1 is calculated by dividing the number of common pixels (ground truth $\cap$ predicted) by the total number of pixels (ground truth $\cup$ predicted) present in both masks. The intersection between ground truth and predicted refers to the set of pixels that are present in both the ground truth mask and the predicted mask. The union of the two masks is simply the set of pixels that are present in either the predicted mask or the ground truth mask. A higher IoU value indicates a better overlap between the target and predicted segmentation, meaning that the segmentation algorithm has performed well in identifying the object of interest. So, we evaluated the semantic food segmentation results using performance measures such as intersection over union (IOU) in Equation 5.1 for each food category, mean intersection over union (mIOU) in Equation 5.2 is calculated by taking the average of the IoU scores across all classes to provide an overall measure of how well the segmentation algorithm performs across all classes present in the dataset, and pixel-level accuracy in Equation 5.3 that represents the percentage of correctly classified pixels where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

$$IoU = \frac{target \cap predicted}{target \cup predicted} \tag{5.1}$$

$$mIoU = \frac{1}{n}\sum_{i=1}^{n} IoU_i \tag{5.2}$$

$$PixelAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.3}$$

5.1.2 Datasets

Tray food dataset: TrayDataset (url) is a food segmentation dataset comprised of 43 food classes. This database contains a total of 1241 food images with 17 unique trays where images are rotated, wrapped, and flipped versions of the unique trays. The dataset is composed of distinctive food classes, e.g., bread, ham, custard, margarine, pumpkin, zucchini, milk, baked fish, creamed potato, orange juice, soup, carrot, vanilla yogurt, cucumber, broccoli, beef, etc. TrayDataset database is a well-defined

publicly available with all food images and their respective ground truth segmentation masks on Kaggle at the following link: https://www.kaggle.com/datasets/thezaza102/tray-food-segmentation.

MyFood dataset: MyFood dataset (Freitas et al., 2020a) is a well-defined publicly available database for food image segmentation. This dataset consists of the most consumed food types by the Brazilian population containing 1250 total images. The dataset is composed of nine food classes such as spaghetti, apple, beans, boiled egg, chicken breast, rice, salad, steak, and fried egg, with an average of 125 food images in each class. For research and evaluation experiments, the dataset is divided into 60% for training, 20% for validation, and 20% for testing. MyFood database is publicly available with all food images and their respective segmentation masks on the Zenodo website with training, validation, and testing folder structure. It can be downloaded at the following link: http://doi.org/10.5281/zenodo.4041488.

5.1.3 Experimental Results

We thoroughly tested our proposed segmentation scheme using two publicly available benchmark food segmentation datasets which are TrayDataset and MyFood dataset. Further details are provided on these datasets. We extensively evaluated the performance of our proposed method compared to the existing state-of-the-art approaches.

Evaluation on Tray food Dataset

Experiments are conducted using TrayDataset that consists of 43 distinctive food classes. The proposed Convolutional Deconvolutional Pyramid Network (CDPN) results are compared with other methods such as feature pyramid network (FPN) (Lin et al., 2017), and encoder-decoder food network (EDFN) (Pfisterer et al., 2019) using intersection over union and pixel accuracy. The EDFN (Pfisterer et al., 2019) architecture is used for automatic semantic segmentation of tracking food and fluid Intake in long-term care homes. It is a deep convolutional encoder-decoder architecture for food pixel-wise segmentation. It employs ResNet architecture as an encoder to get 256 feature maps of the input image and a pyramid scene parsing is used as decoder microarchitecture to decode the feature maps from the encoder. FPN (Lin et al., 2017) is designed for image segmentation and multi-scale object detection. This architecture is developed to extract semantic feature maps that involve a bottom-up pathway which computes a feature hierarchy, and a top-down pathway which computes stronger feature maps from higher pyramid levels. The proposed CDPN, FPN, and EDFN are trained using the same hyperparameters and settings for the comparative evaluation. All the networks are trained using the Adam optimizer and the standard Dice loss function. The

learning rate is set to 0.0001. The batch size is set as 8 and networks are trained for 250 epochs. The networks are trained using ResNet-101 as the backbone network for the evaluation experiments.

On the TrayDataset, the experimental results of our proposed CDPN are compared with FPN and EDFN. The experimental results comparison of the networks using class-wise intersection over union is shown in Table 5.1 where the top IOU is represented in bold for each food category. Dataset also has the background class with IOU of 99.49% for the proposed CDPN, 99.28% for EDFN, and 98.52% for FPN. For most of the food classes, our proposed CDPN approach achieved higher class-wise IOU results as compared to others. The lowest IOU scores for the proposed CDPN, FPN, and EDFN are 75.08% for the bakedfish class, 72% for the creamedpotato class, and 70% for the jelly class, respectively. From the experimental results described in Table 5.1, our proposed CDPN method achieved a competitive IOU score as compared to the FPN and EDFN.

The visual representation is presented in Figure 5.2 of the original input image, ground truth, and output segmentation maps of the proposed network and baseline networks on TrayDataset. For example, consider the original input image (1) and its output segmentation maps generated by models, EDFN confuses beefmexicanmeatballs with pumpkin. The proposed CDPN and FPN predict correctly. Now, consider the original input image (2) and its output segmentation maps generated by models, EDFN mispredicts the vanillayogurt region by confusing most of the part of it with zucchini. The EDFN does not detect the margarine region as well. The FPN misclassifies the vanillayogurt region by confusing it with margarine and zucchini. The proposed CDPN detects vanillayogurt accurately but confuses margarine with vanillayogurt.

Table 5.1: Clalss-wise Intersection Over Union (IOU) of food items. In the table, BeefTC means BeefTomatoCasserole, "BeefMM" means beefmexicanmeatballs, "SpinachPR" means SpinachandPumpkinRisotto.

Food Items	Class-wise IOU (%)			Food Items	Class-wise IOU (%)		
	Proposed CDPN	EDFN	FPN		Proposed CDPN	EDFN	FPN
Tray	**96.49**	83.95	89.28	Pumpkin	**88.17**	72.02	85.99
Cutlery	**87.96**	82.06	85.08	Celery	100.0	100.0	100.0
Bread	**93.67**	80.00	80.00	Sandwich	**90.59**	88.25	82.62
Straw	100.0	100.0	100.0	SideSalad	**92.86**	90.00	90.00
Custard	**93.09**	78.45	80.58	TartareSauce	85.00	85.00	85.00
Beef	100.0	100.0	100.0	JacketPotato	85.00	**91.00**	90.00
Roastlamb	85.00	85.00	85.00	CreamedPotato	**89.51**	72.00	72.00
BeefTC	83.74	75.35	**98.34**	Form	100.0	100.0	100.0
Ham	**98.64**	90.00	90.00	Margarine	80.00	78.00	**90.00**
Bean	90.00	90.00	90.00	Soup	**97.71**	83.61	92.45
Cucumber	**90.00**	**90.00**	86.79	Apple	100.0	100.0	100.0
Leaf	**97.31**	91.20	95.00	CannedFruit	82.85	**90.00**	**90.00**
Tomato	90.00	90.00	90.00	Milk	84.68	**90.00**	77.91
Boiledrice	**81.37**	80.00	80.00	VanillaYogurt	**88.23**	82.00	79.00
BeefMM	**99.47**	79.00	89.49	Jelly	**88.80**	70.00	73.95
SpinachPR	79.29	**90.00**	79.25	Meatball	100.0	100.0	100.0
BakedFish	75.08	**85.00**	**85.00**	LemonSponge	**99.34**	95.00	95.00
Gravy	89.83	86.78	**94.42**	Juice	100.0	100.0	100.0
Broccoli	100.0	100.0	100.0	AppleJuice	77.71	**90.00**	**90.00**
Carrot	100.0	100.0	100.0	OrangeJuice	**94.28**	90.00	76.00
Zucchini	**93.82**	75.50	88.11	Water	**97.14**	86.77	85.00

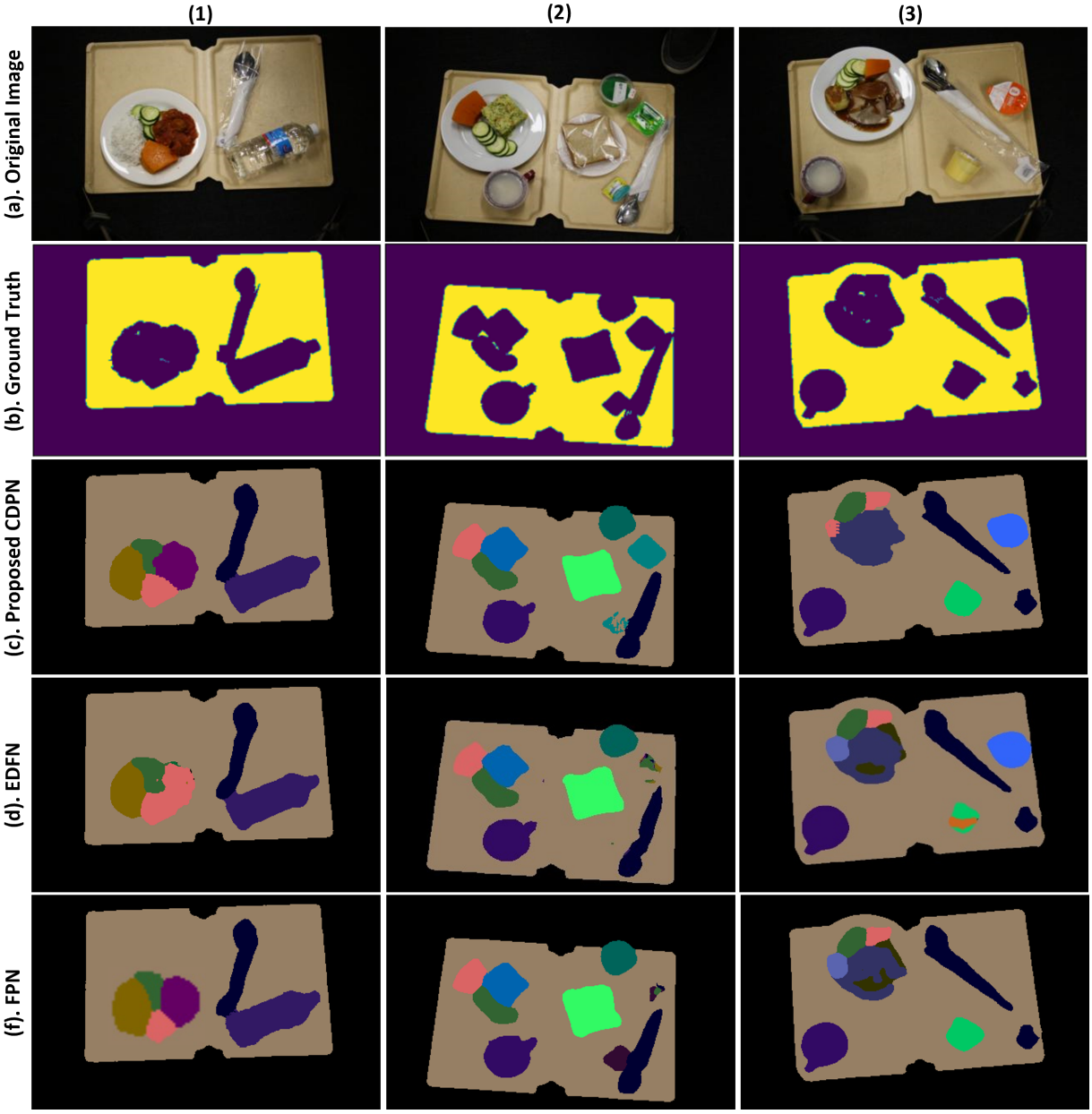


Figure 5.2: Food segmentation results visualization of the proposed CDPN, FPN, and EDFN on TrayDataset. (a) represents original images, (b) represents ground truths, (c) represents the proposed CDPN output segmentation maps, (d) represents EDFN output segmentation maps, and (f) represents FPN output segmentation maps.

The results comparison of the proposed method CDPN with FPN and EDFN networks using mean intersection over union and global pixel level accuracy is presented in Table 5.2. From the experimental results presented in Table 5.1, 5.2, and Figure 5.2, our proposed CDPN method achieved competitive results. These experimental results show that the proposed Convolutional Deconvolutional Pyramid Network outperformed both EDFN and FPN.

Table 5.2: The proposed CDPN results comparison with EDFN and FPN

Method	**Backbone**	**Mean IOU(%)**	**Pixel Accuracy (%)**
Proposed CDPN	ResNet-101	91.77	98.90
FPN (Lin et al., 2017)	ResNet-101	89.30	98.49
EDFN (Pfisterer et al., 2019)	ResNet-101	88.02	97.93

Split the image into pixel-level segments

We divided the original image into individual food pixel-level segments by setting the background to black on the basis of the segmented image map for further food analysis such as food annotation, classification, volume estimation, etc. In this process, the pixels from each segmented food item in the network-detected segmented image map are utilized to extract the corresponding pixels from the original image, and any residual pixels are set to zero. Finally, we obtained individual food segments of the original image with a black background for each food segment detected in the segmented image map. Figure 5.3 displays the original image of TrayFood together with its segmented image map and each food segment that was extracted from the original image based on the segmented image map at the pixel level.

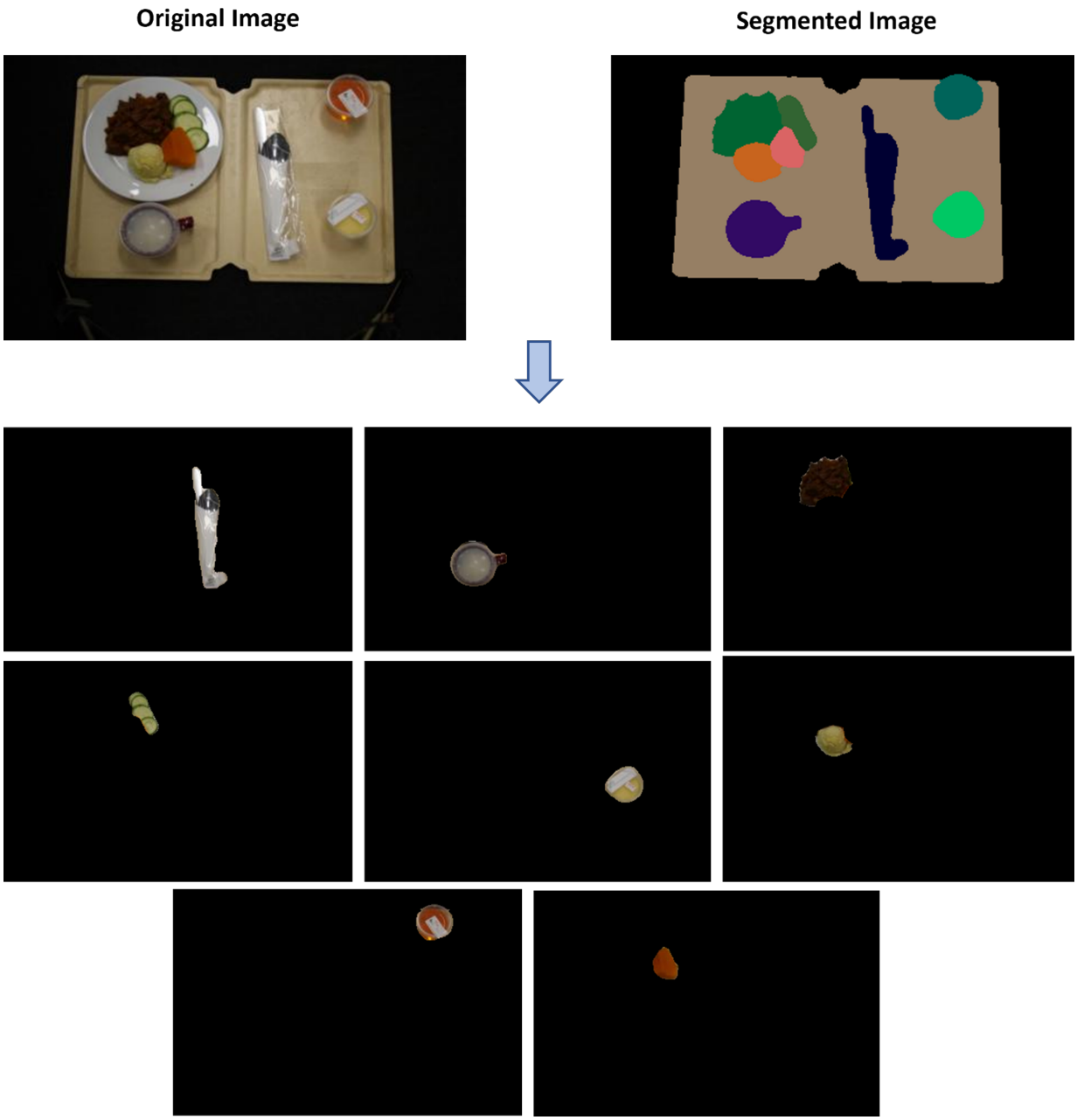


Figure 5.3: Individual food pixel-level segments of the original image for each food segment detected in the segmented image map. Each food segment was extracted from the original image based on the segmented image map.

Evaluation on MyFood Dataset

On the MyFood (Freitas et al., 2020a) dataset, we compared our approach with Segnet (Badrinarayanan et al., 2017), Mask R-CNN (He et al., 2018), FCN (Long et al., 2015), UNet++ (Zhou et al., 2018), Enet (Paszke et al., 2016), and DeepLabV3+ (Chen et al., 2018). The Segnet (Badrinarayanan et al., 2017) is a deep convolutional encoder-decoder architecture for semantic pixel-wise segmentation. The encoder architecture is topologically identical to the 13 convolutional layers of the VGG16 architecture. To perform non-linear upsampling, the decoder utilizes pooling indices computed during the max-pooling step of the encoder. The mask R-CNN (He et al., 2018) architecture is an end-to-end convolutional neural network introduced by the Facebook AI research group with an accurate detection effect when it comes to targeting object instance segmentation. The mask R-CNN (He et al., 2018) is the extended improvement of Faster R-CNN with an object mask prediction branch in parallel with the existing bounding box detection branch. FCN (Long et al., 2015) is a fully convolutional network for segmentation with skip architecture that combines layers of the feature hierarchy to produce refined segmentation. The classification networks GoogLeNet, VGG net, and AlexNet are extended to fully convolutional networks by transferring their learning for the segmentation. The UNet++ (Zhou et al., 2018) is the extension of U-Net architecture in which sub-networks encoder and decoder are connected through a series of nested and dense skip pathways. It was proposed for biomedical image segmentation that is more powerful as compared to the U-Net. Enet (Paszke et al., 2016) is a deep neural network for real-time segmentation performance on embedded platforms. It is heavily inspired by the ResNet and Inception architectures with the aim to perform large-scale computations efficiently. The DeepLabV3+ (Chen et al., 2018) is the extended version of DeepLabv3 developed for semantic segmentation with the concept of atrous separable convolution. It employs an encoder-decoder structure with atrous convolution comprised of a deep convolution and a clockwise convolution. The encoder is used to rich the contextual information, and the effective decoder is used to refine the segmentation results.

Table 5.3: The proposed CDPN results comparison with EDFN and FPN

Method	Optimizer	Learning Rate	Decay	Batch Size
Proposed CDPN	Adam	1E-4	1E-5	8
UNet++	Adam	1E-4	1E-5	8
Enet	Adam	5E-4	-	10
DeepLabV3+	SGD	1E-2	-	32
Mask R-CNN	SGD	1E-3	1E-4	2
FCN	SGD	1E-2	-	32
Segnet	SGD	1E-2	-	32

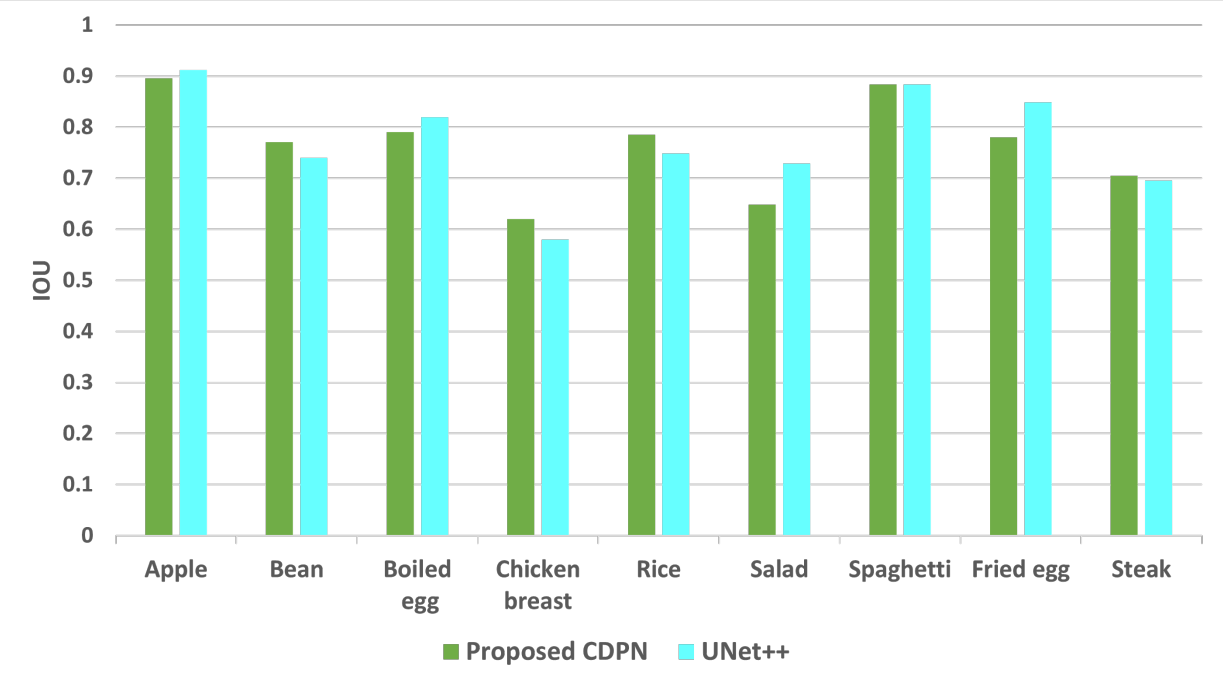


Figure 5.4: The proposed CDPN method and UNet++ class-wise intersection over union (IOU) results comparison on the MyFood segmentation dataset.

The hyperparameters used are given in Table 5.3 where all the networks are trained for 100 epochs for the comparative evaluation. The parameters for Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+ are described in the research (Freitas et al., 2020b) using the MyFood segmentation dataset. The proposed CDPN and UNet++ are trained using the same hyperparameters with the Adam optimizer, the standard Dice loss function, the learning rate is set to 0.0001, and the batch size is set as 8. The parameters used for training the proposed CDPN and other methods are listed in Table 5.3 for experiments evaluation on the MyFood dataset.

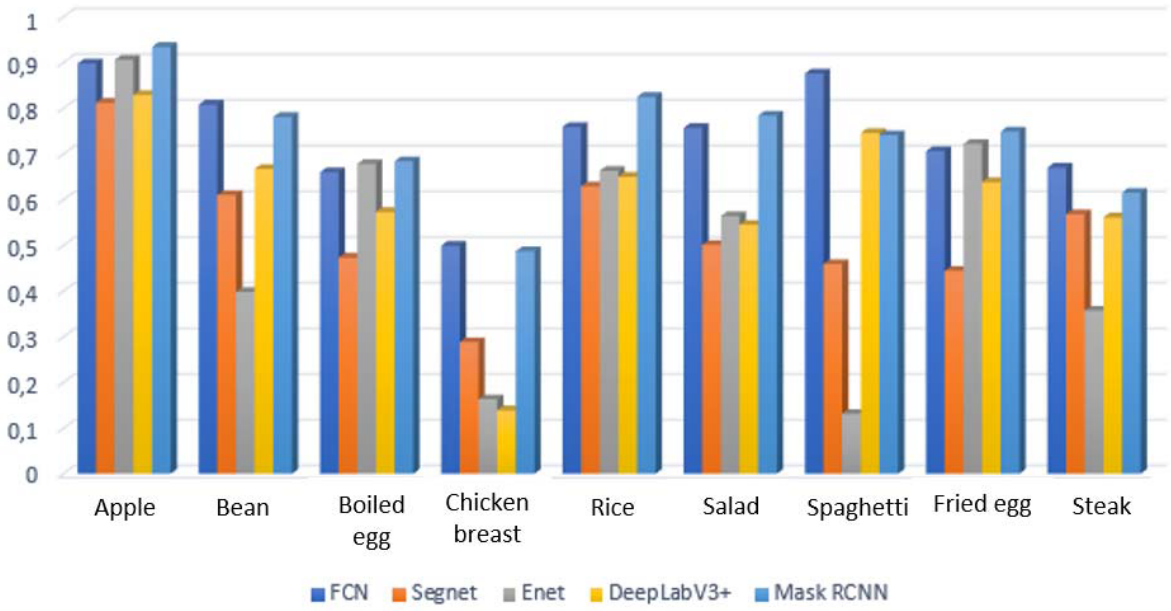


Figure 5.5: Class-wise intersection over union (IOU) on MyFood dataset. This figure is adopted from the research (Freitas et al., 2020b) that shows the comparison of the class-wise results for Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+.

The experimental results comparison of the networks is shown in Figure 5.4 and Figure 5.5 using class-wise intersection over union. The results achieved by the proposed CDPN and UNet++ are presented in Figure 5.4 . For the class-wise IOU comparative evaluation with the proposed CDPN and UNet++, the results (Freitas et al., 2020b) are shown in 5.5, which shows IOU for Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+. Dataset also has the background class with IOU of 0.90 for the proposed CDPN, and 0.92 for UNet++. Our proposed CDPN approach provided a competitive class-wise intersection over union score in comparison with other methods. According to Figure 5.4 and Figure 5.5, the chicken breast class had the lowest IOU score, with the proposed CDPN obtaining the highest IOU score of 0.62, UNet++ having an IOU score of 0.58, and Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+ having an IOU score less than 0.50. However, the proposed CDPN, Unet++, FCN, and Mask R-CNN produced results with comparatively high class-wise IOU scores.

The segmentation results on the MyFood dataset are presented in Figure 5.6 with the visual representation of the input image and output segmentation maps of the proposed CDPN, Segnet, Mask R-CNN, FCN, UNet++, Enet, and DeepLabV3+. In the case of a single food in an image, we can notice that most methods performed well to produce the output segmentation maps of the input image. On the other hand, the proposed CDPN, UNet++, and FCN results are comparable when multiple food items

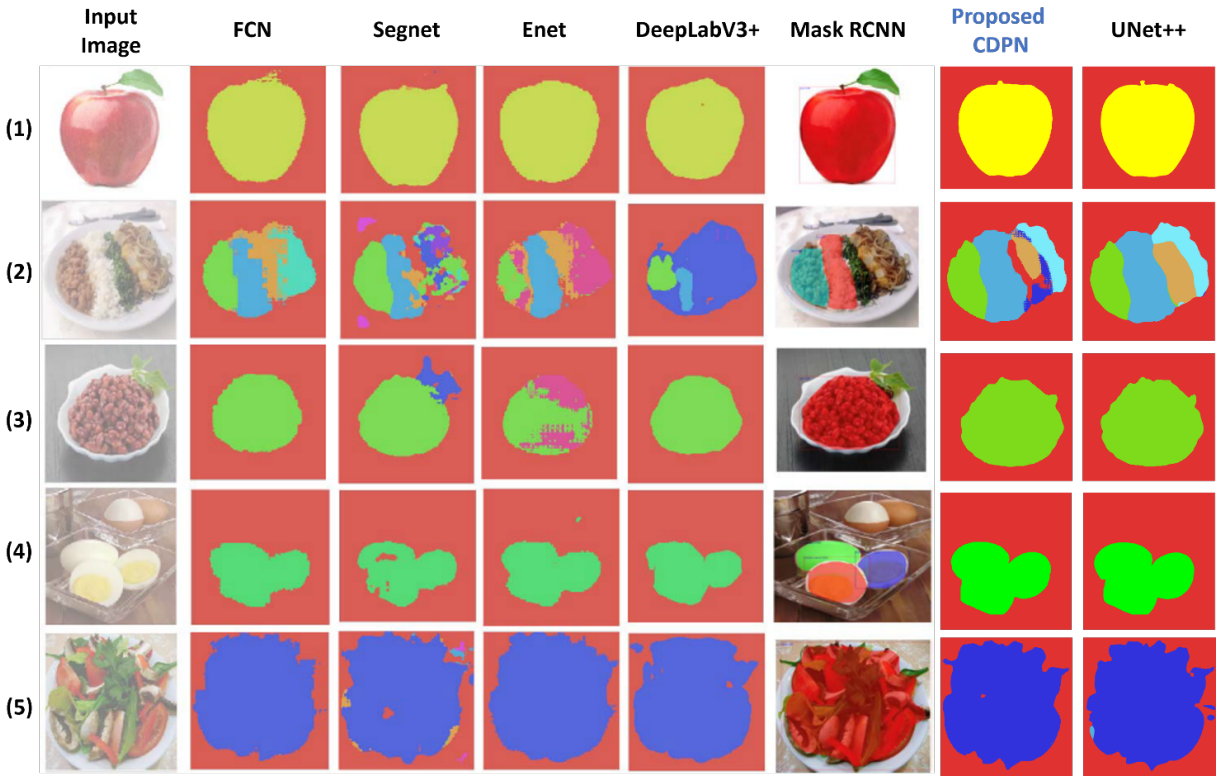


Figure 5.6: Food segmentation results visualization of the proposed CDPN approach with other methods on the MyFood dataset. For instance, the input image (1) is an apple, and its output segmentation maps generated by each network are presented here. We added the output segmentation maps of proposed CDPN and UNet++ together with the visualization of food image segmentation results described in research (Freitas et al., 2020b) for comparative evaluation with Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+.

are present in an image. In both cases, the proposed CDPN achieved better segmentation results compared to the Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+ as shown for input image (2) in Figure 5.6. However, UNet++ produced better results than the proposed CDPN method.

On the MyFood dataset, the segmentation results of our proposed CDPN method are outstanding compared to the state-of-the-art approaches. The UNet++ obtained higher results with 0.79 mean IOU, but there was very little marginal difference in mean IOU when compared to our proposed CDPN with 0.77 mean IOU. From the detailed experimental analysis presented in Figures 5.4, 5.5, 5.6, and Table 5.4, our proposed CDPN approach provided competitive results. These experimental results show that the proposed CDPN outperformed other approaches such as Segnet, Mask

R-CNN, FCN, Enet, and DeepLabV3+ on the MyFood segmentation dataset.

Table 5.4: The proposed CDPN method results in comparison with other methods on the MyFood dataset.

Method	Backbone	IOU
UNet++ (Zhou et al., 2018)	ResNet-101	0.79 (0.11)
Proposed CDPN	ResNet-101	0.77 (0.09)
Mask R-CNN (He et al., 2018)	ResNet-101	0.70 (0.2)
FCN (Long et al., 2015)	VGG16	0.70 (0.2)
Segnet (Badrinarayanan et al., 2017)	-	0.52 (0.2)
Enet (Paszke et al., 2016)	-	0.51 (0.3)
DeepLabV3+ (Chen et al., 2018)	MobileNet	0.50 (0.3)

5.2 Food Convolutional Deconvolutional Network (FCDN)

We propose another food segmentation architecture to understand the semantic information of an image at a pixel level. The proposed FCDN employs only learnable features upsampling using deconvolution as compared to the CDPN. The segmentation output is a set of image regions, and accurate segmentation of food regions can be used to estimate the quantities of each food item detected within an image. This would enable people to estimate calories and nutrient intake in order to track their food intake and become more aware of their daily diet. We propose a Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images at a pixel level to recognize different food items present in an image. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns, while the proposed CDPN also uses interpolation for features upsampling along with the deconvolution layers.

The proposed FCDN network employs encoder-decoder architecture for pixel-wise segmentation. The encoder utilizes ResNet architecture with transferred weights from the ImageNet to capture highly descriptive feature representations from the input RGB image. The novelty of the proposed network lies in the way the decoder architecture upsamples and densifies lower-resolution input features from the encoder to high-resolution feature maps using learnable filters. In particular, the decoder employs convolution and deconvolution layers to build multi-scale feature maps representation and to further produce rich and concise segmentation map of the input image.

The performance of our proposed segmentation approach was evaluated through experiments conducted on the MyFood dataset. The results of our experiments indicate

noteworthy enhancements in the outcomes achieved, as compared to previous methodologies. In addition to evaluating the performance of our method on the MyFood dataset, we also performed cross-data experiments to assess its generalization capabilities on our FoodRec dataset. By conducting cross-data experiments on the FoodRec dataset, we were able to determine that our method could effectively make accurate predictions in different contexts. This qualitative evaluation served as an important complement to our evaluation on the FoodRec dataset, and helped to strengthen our confidence in the effectiveness of our method.

Proposed Segmentation Architecture

We develop a food segmentation model capable of extracting and inferring semantic information from food images at a pixel level to accurately recognize and identify the various types of food items present within an image. In this context, a novel Food Convolutional Deconvolutional Network (FCDN) is proposed for image semantic segmentation to analyse food images and to generate a precise segmentation map of individual food items, as illustrated in Figure 5.7. The proposed network is an encoder-decoder architecture for semantic food segmentation.

The encoder consists of a Convolutional Neural Network that harvests meaningful feature map representation from an input image. As the encoder processes the image, it gradually increases the number of channels in each step, adding more depth to the feature maps. Additionally, the encoder downsamples the image's spatial resolution to produce high-level information by reducing image height and width. We used the transfer learning concept for the encoder part and exploited the pre-trained encoder with the transfer knowledge from the ImageNet dataset. The decoder architecture consists of convolution layers coupled with upsampling layers to produce a high-resolution segmentation feature map representation from the low-resolution feature map of the encoder. To create the segmentation map, the decoder applies upsampling layers that gradually increase the resolution of the feature maps. The decoder also applies convolutional layers to reduce the number of channels in the feature maps and outputs a segmentation map, which is the region of pixels of the same size as the input image.

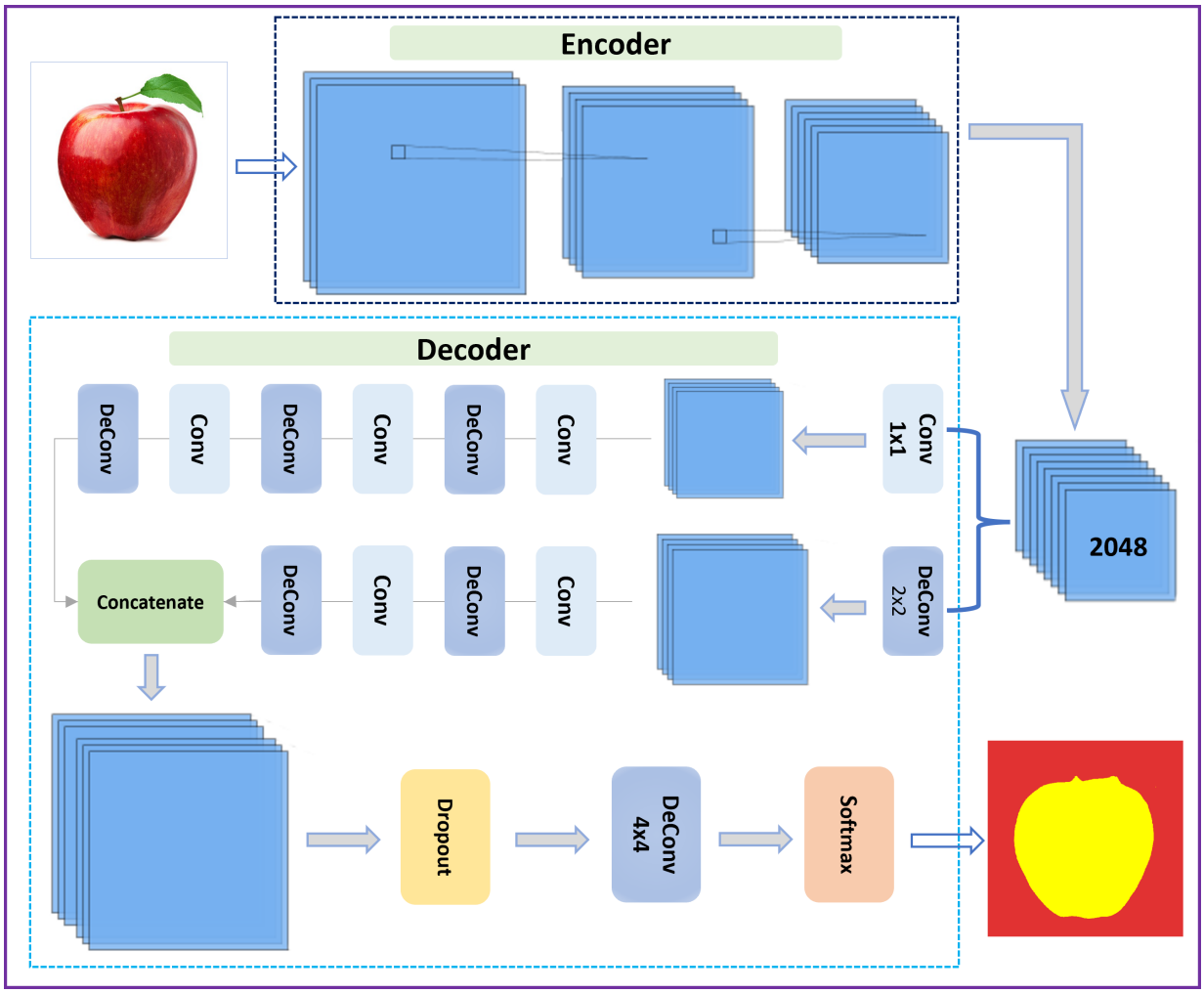


Figure 5.7: Our proposed Food Convolutional Deconvolutional Network architecture for pixel-wise food segmentation. The network takes food image as input to extract and infer semantic information from it and outputs a segmentation map of the individual food items present in the image. The ”Conv” represents the convolutional layer, and ”DeConv” represents the deconvolutional layer. All convolutional layers with a kernel size of 3x3 are followed by group normalization and ReLU activation layers.

To effectively capture highly descriptive feature representations for complex food images, we employed the deep ResNet architecture with pre-activations as an encoder. By leveraging the power of transfer learning, we utilized ResNet architecture with pre-trained weights from the ImageNet dataset to extract feature representation from the input RGB food image. The ResNet-101 architecture which is a variant of the ResNet architecture was chosen as the backbone network due to its capability in learning relevant feature representation. In particular, we obtain a highly descriptive feature set of 2048 channels from the input image with a downsampled spatial resolution (h/32, w/32) using Resnet-101 as the backbone network. Our proposed decoder architecture employs the convolution layer and deconvolution layer to feature map obtained by the ResNet to generate two high-level feature maps. The use of convolution and deconvolution layers enables us to capture important features at different scales and resolutions. By employing these layers, we are able to generate two distinct multi-scale feature maps. The first feature map has a size of h/16 x w/16 x 512 channels, which is obtained by applying the deconvolution layer with a 2x2 kernel size. The deconvolution densifies the input features using learnable parameters to produce generalized upsampling of the feature representation. The second feature map has a size of h/32 x w/32 x 512 channels which is produced by applying a convolution layer with a 1x1 kernel size. Next, a series of convolutional layers with a kernel size of 3x3 and deconvolution layers with a kernel size of 2x2 are employed. Each convolutional layer in this sequence is followed by a group normalization and ReLU activation function, which help to improve the performance. After the upsampling of the feature maps from each level, the resulting feature maps are fused and concatenated. As a result, we obtain two feature maps of the same size w/4 x h/4 x 256 after they have been upsampled and then merged to produce a high-dimensional representation. The Dropout layer is then applied to avoid overfitting and improve the generalization performance of the model. Finally, the feature maps are passed through a deconvolutional layer with a kernel size of 4x4 and output channels equal to the number of food categories followed by a Softmax activation function to produce the final segmentation map. The output segmentation map has the same spatial dimensions as the input image where each pixel is assigned a class label corresponding to the region of the image that it belongs to.

5.2.1 Experimental Results

We present the results of our proposed segmentation approach on a publicly available benchmark food segmentation dataset. In addition, we also provide a qualitative analysis of the results by visually comparing the segmentation outputs of our model with the existing methods. We also perform cross-data experiments on our FoodRec data to evaluate the qualitative performance of our method. We thoroughly tested to

demonstrate the performance of our proposed method compared to the state-of-the-art approaches.

Experiments Evaluation on MyFood Dataset

We conducted a comprehensive evaluation to assess the performance of our proposed approach by comparing it with state-of-the-art methods such as UNet++ (Zhou et al., 2018), Mask R-CNN (He et al., 2018), FCN (Long et al., 2015), Segnet (Badrinarayanan et al., 2017), Enet (Paszke et al., 2016), and DeepLabV3+ (Chen et al., 2018) on MyFood dataset (Freitas et al., 2020a). The hyperparameters were carefully selected to achieve optimal performance and are presented in Table 5.5. To facilitate comparative evaluation, all networks were trained for 100 epochs. The parameters used for training DeepLabV3+, FCN, Enet, Mask R-CNN, and Segnet were described in research (Freitas et al., 2020b) using the MyFood segmentation dataset. The proposed FCDN and UNet++ (Zhou et al., 2018) were trained using the same hyperparameters, which included the use of the Adam optimizer, the standard Dice loss function, a learning rate of 0.0001 which is decayed to 0.00001 after 30 epochs, and a batch size of 8. For the experiments conducted on the MyFood dataset, the hyperparameters used for training the proposed FCDN and other methods can be found in Table 5.5. It is worth noting that the selection of appropriate hyperparameters is essential for achieving optimal performance in deep learning-based image segmentation tasks, and our study emphasizes this critical aspect by providing detailed information on the hyperparameters employed in our experiments.

Table 5.5: Hyperparameters employed in training each segmentation network on MyFood Dataset.

Method	Optimizer	Learning Rate	Decay	Batch Size
Proposed FCDN	Adam	1E-4	1E-5	8
UNet++	Adam	1E-4	1E-5	8
Enet	Adam	5E-4	-	10
DeepLabV3+	SGD	1E-2	-	32
Mask R-CNN	SGD	1E-3	1E-4	2
FCN	SGD	1E-2	-	32
Segnet	SGD	1E-2	-	32

The performance of different image segmentation models has been evaluated on the MyFood dataset using the intersection over union (IoU) metric to measure the similarity between the predicted segmentation and the ground truth segmentation. In this study, we compared and evaluated the performance using IoU score of our proposed method with different state-of-the-art segmentation models, including UNet++ (Zhou

et al., 2018), Mask R-CNN (He et al., 2018), FCN (Long et al., 2015), Segnet (Badrinarayanan et al., 2017), Enet (Paszke et al., 2016), and DeepLabV3+ (Chen et al., 2018). Our results as described in Table 5.6 indicate that UNet++ and the proposed model achieved the highest IoU scores as compared to the other methods. The UNet++ achieved the highest mean IoU score of 0.79 (standard deviation of 0.11), followed closely by the proposed model with a mean IoU score of 0.78 (standard deviation of 0.09). Mask R-CNN and FCN both achieved mean IOU scores of 0.70 (standard deviation of 0.2), which is relatively high but not as accurate as the top-performing models. Segnet and Enet, on the other hand, exhibited mean IoU scores of 0.52 (standard deviation of 0.2) and 0.51 (standard deviation of 0.3), respectively. Finally, DeepLabV3+ achieved a mean IoU score of 0.50 (standard deviation of 0.3), which is the lowest among all evaluated models. Overall, our results demonstrate that the proposed model achieved competitive segmentation results and outperform several existing segmentation methods.

Table 5.6: The proposed FCDN method results in comparison with other methods on the MyFood dataset.

Method	**Backbone**	**Inersection over Union (IoU**
UNet++ (Zhou et al., 2018)	ResNet-101	0.79 (0.11)
Proposed FCDN	ResNet-101	0.78 (0.09)
Mask R-CNN (He et al., 2018)	ResNet-101	0.70 (0.2)
FCN (Long et al., 2015)	VGG16	0.70 (0.2)
Segnet (Badrinarayanan et al., 2017)	-	0.52 (0.2)
Enet (Paszke et al., 2016)	-	0.51 (0.3)
DeepLabV3+ (Chen et al., 2018)	MobileNet	0.50 (0.3)

In Figure 5.8, we present the IoU scores for each class, which provide insights into the models' performance in detecting foods of different classes. The analysis of the IoU scores for each class provides valuable insights into the strengths and weaknesses of different segmentation models. The results suggest that while some classes are easier to detect than others, the proposed FCDN, UNet++, FCN, and Mask RCNN models demonstrate higher performance for most classes for semantic segmentation in the given dataset. The results indicate that the apple class yielded the highest IoU scores where all the models produced IoU scores more or equal to 0.8, suggesting that it is the easiest class to detect. In contrast, the chicken breast class achieved the lowest IoU scores among all the classes suggesting that it is the most challenging class to detect accurately. The proposed FCDN method achieved the highest intersection over union (IoU) score of 0.64 for the chicken breast class, outperforming all other evaluated models. The UNet++ model achieved an IoU score of 0.58 for the chicken breast class, which was the second-best performance among the models. On the other hand, the

Mask R-CNN, FCN, DeepLabV3+, Segnet, and Enet models achieved IoU scores less than 0.50 for the chicken breast class, indicating comparatively lower performance.

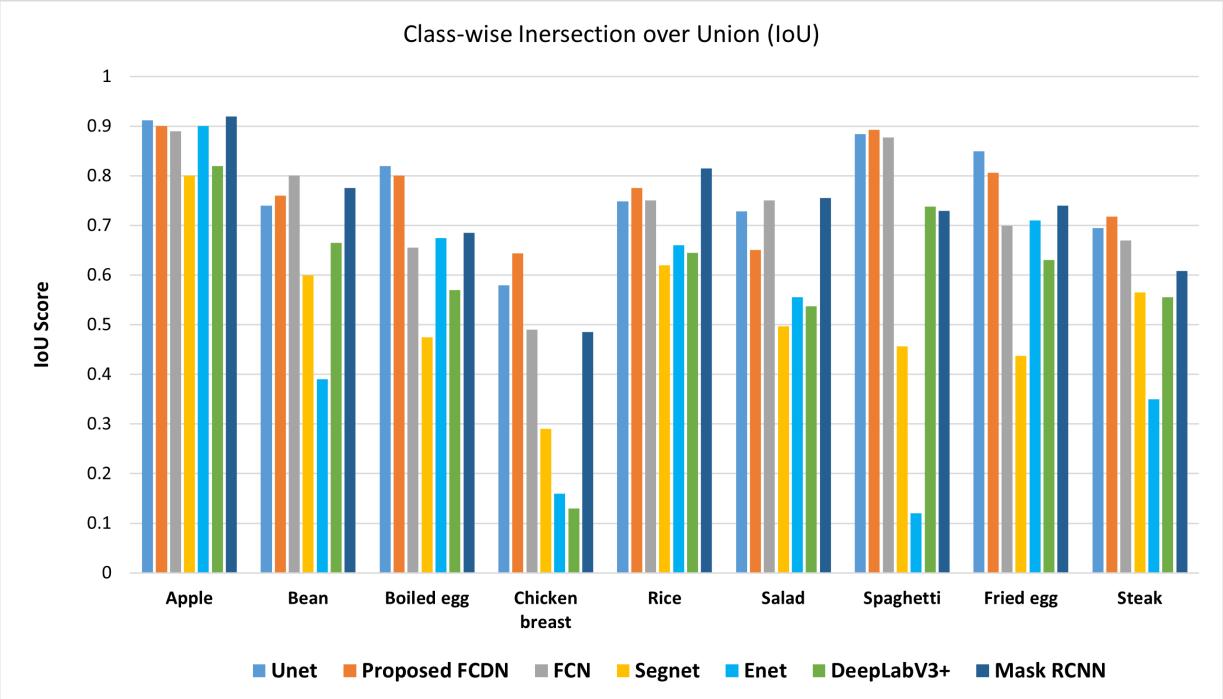


Figure 5.8: The proposed FCDN method class-wise intersection over union (IoU) results comparison with other methods on the MyFood segmentation dataset. The x-axis represents the food classes in the dataset, and the y-axis shows the intersection over union (IoU) score obtained by each network for the food classes

In addition to the food classes, the dataset used in this study also includes a background class. The intersection over union (IoU) scores for the background class was evaluated for the proposed method and the UNet++ model. The results indicate that the proposed method achieved an IoU score of 0.91 for the background class, while UNet++ achieved an IoU score of 0.92. The class-wise evaluation results described in Figure 5.8 show that our proposed FCDN approach provided a competitive class-wise intersection over union score when compared to other state-of-the-art methods for segmentation.

We present the qualitative results of the proposed method with other state-of-the-art methods for comparative analysis of the food segmentation. Figure 5.9 provides visualizations of the segmentation outputs of the proposed method with other methods for five sample images from the MyFood dataset. The visual representation of the input image, ground truth mask, and output segmentation maps of the proposed FCDN, UNet++, FCN, Segnet, Enet, DeepLabV3+, and Mask R-CNN are presented.

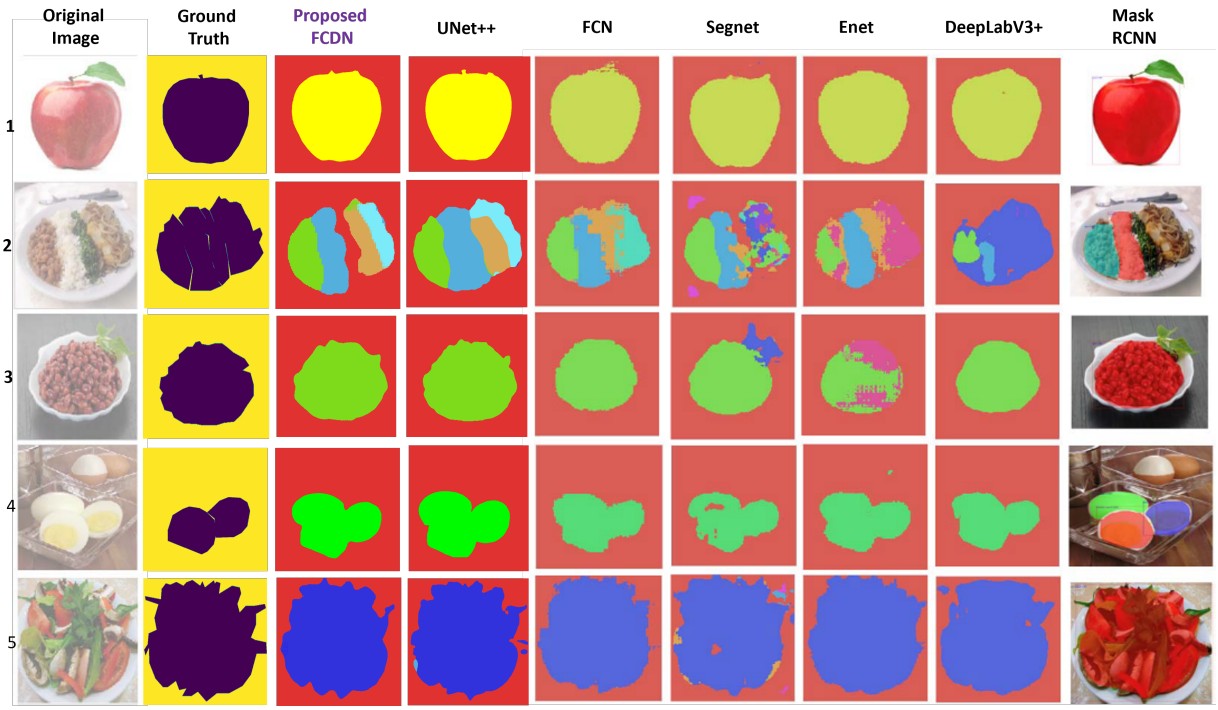


Figure 5.9: Visualization of qualitative segmentation results of the proposed FCDN approach with other methods on the MyFood dataset. For example, the input image 3 represents beans and its output segmentation maps generated by each network are presented. Qualitative results of FCN, Segnet, Enet, DeepLabV3+, and Mask R-CNN are also described in (Freitas et al., 2020b).

As an example, consider the first input image which depicts an apple and its output segmentation maps generated by each network. As observed from Figure 5.9, most methods performed better to generate the output segmentation map when there is only one food item present in the image. However, when the image contains multiple food items, the proposed FCDN and Unet++ performed well segmentation as compared to others. Moreover, the proposed FCDN method has produced better output segmentation maps as compared to FCN, Segnet, Enet, DeepLabV3+, and Mask R-CNN for both single and multiple food items present in the image. For example, consider the second input image shown in Figure 5.9 which depicts multiple food items and its output segmentation maps generated by each network.

Based on the quantitative and qualitative results presented in Table 5.6, Figures 5.8 and 5.9, we conclude that our proposed approach achieved comparatively better results, with a mean IoU of 0.78 on MyFood dataset. These experimental results demonstrate that the proposed method outperforms state-of-the-art methods including FCN, Segnet, Enet, DeepLabV3+, and Mask R-CNN on MyFood (Freitas et al., 2020a)

segmentation dataset. However, UNet++ produced higher results than our proposed method with a marginal difference in the mean IoU score.

Qualitative Evaluation on FoodRec Dataset

We finally conducted cross-data experiments to evaluate the performance of our proposed FCDN segmentation method. The qualitative results of our proposed method are evaluated on a subset of our FoodRec dataset. To test the generalization capabilities of our method, we trained our model on the MyFood dataset (Freitas et al., 2020a) , and evaluated its qualitative performance on our FoodRec dataset described in chapter 4 with food classes presented in Table 4.1. Both MyFood and our FoodRec datasets contain common food classes, such as apple, beans, egg, spaghetti, chicken, rice, and salad. The cross-data experiments enabled us to assess how well our proposed method could perform on new and unseen data with different context collected from the 164 real users during their smoking cessation therapy using smartphone application. By performing the cross-data experiments, we aimed to simulate a scenario where the model is trained on a limited dataset and is expected to perform well on a larger and more diverse dataset. Further, we extracted the detected region of the input food items in an image using the output segmentation map of that image. This can be used for food classification, assessment of quantities of each food item, food volume estimation, food annotation, and calories estimation. To obtain detected food segments, we utilize the output segmentation map to identify the location of each food segment. In this step, the pixels in the output segmentation map are used to extract the matching pixels from the original image. We then use this information to extract the corresponding regions from the original image while replacing the remaining parts with a black background. Figure 5.10 presents the qualitative results where the food region has been extracted from the original image based on the output segmentation map. For instance, input image 4 represents rice, its output segmentation map generated by the proposed approach, and extracted food region from the input image based on the segmentation map. These results of our experiments demonstrate that our proposed method is capable of generalizing well to new and unseen FoodRec data.

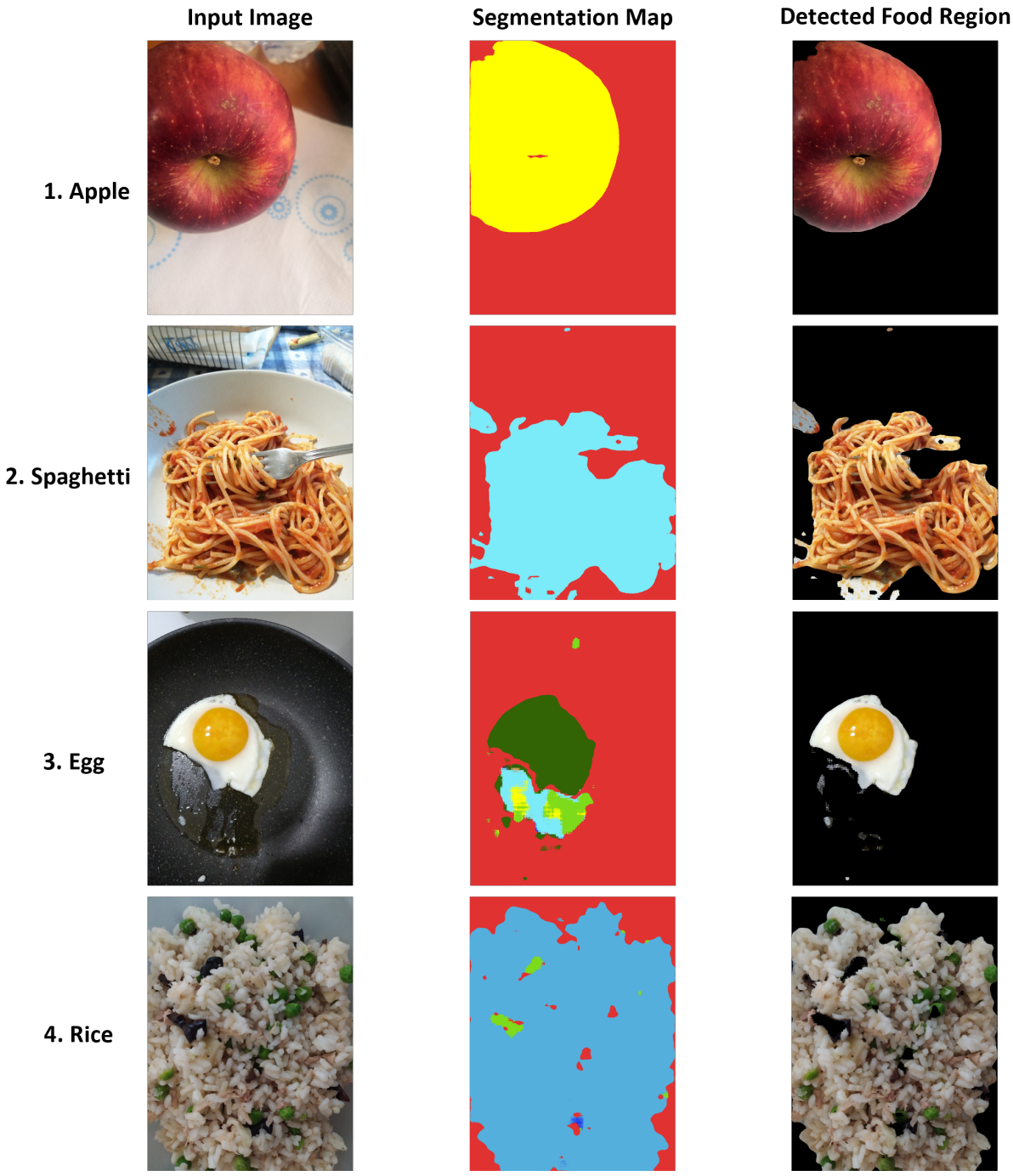


Figure 5.10: Visualization of qualitative segmentation results of the proposed FCDN approach on FoodRec dataset. For example, input image 2 represents spaghetti, its output segmentation map generated by the proposed approach, and extracted food region from the input image based on the segmentation map.

5.3 Discussion and Comparison between CDPN and FCDN methods

The proposed Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images at a pixel level is the extended version of Convolutional Deconvolutional Pyramid Network (CDPN) as we are trying to improve it. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns while proposed Convolutional Deconvolutional Pyramid Network (CDPN) also uses interpolation for features upsampling along with the deconvolution layers.

The proposed CDPN obtains two discriminative feature sets of 512 channels with a downsampled spatial resolution (h/8, w/8), and 2048 channels with a downsampled spatial resolution (h/32, w/32) from an input image (height(h), width(w)) using Resnet-101 as the backbone network. In FCDN, a descriptive feature set of 2048 channels from the input image with a downsampled spatial resolution (h/32, w/32) is obtained using Resnet-101 as the backbone network. Then, these features are further upsampled using only deconvolutional layer instead of bilinear interpolation as compared to the CDPN to generate the final output segmentation map. Deconvolution is achieved using ConvTranspose2d that is a learnable upsampling technique that uses transposed convolutional layers to increase the spatial resolution of the feature maps. It has the advantage of being able to learn complex patterns and features specific to the given task, however, it is computationally expensive. Bilinear interpolation is a non-learnable upsampling technique that is based on a weighted average of the nearest neighboring pixels in the input image. It is simple and computationally efficient, and can be used as a baseline for comparison with more complex methods. However, it may not capture complex patterns and features in the input image, and can result in blurry output. In conclusion, if the goal is to achieve high accuracy and capture complex patterns in the upsampling process, ConvTranspose2d is a better choice. If the goal is to achieve a balance between computational efficiency and reasonable quality, bilinear interpolation is a more suitable option. Ultimately, the choice of upsampling technique should be made based on the specific needs and requirements of the task at hand. The comparison of the results of both techniques, together with those from the state of the art is already explained in in Table 5.4, Table 5.6, Figure 5.5, Figure 5.6, Figure 5.8 and Figure 5.9.

5.4 Conclusion

Food Image analysis has become an increasingly important task in recent years, with the growing interest in healthy eating and nutrition. With the proliferation of smartphones and digital cameras, food images have become ubiquitous on the internet, making food recognition technology an important and practical problem to have more accurate ways to identify food items and track dietary intake. The ability to automatically recognize food items from images has a wide range of potential applications. For example, it would allow people o easily track their food intake and monitor their daily diet by simply taking a picture of their meals, to increase awareness of their daily diet by monitoring their eating habits, kind and amount of taken food, how much time the user spends eating during the day, how many and what times the user has a meal, analysis on user's habits changes, bad habits, and other inferences related to user's behavior and mood changes over time. It can help a doctor to have a better opinion with respect to the patient's behaviour and habits changes, in the applications on quitting treatment response, smoke monitoring technology, dietary monitoring during smoke quitting, user evaluation on smoking detection and quitting, and smoking cessation system. Food monitoring plays a vital role in human health that is directly affected by diet. Humans life is strictly affected by the food, this encourages computer vision and deep learning researchers to introduce new methods for food logging and automatic food dietary monitoring, food retrieval and classification, food recognition to monitor users' eating habits that can help individuals make healthier food choices and monitor their dietary intake over time, and food segmentation to understand and analyse food images at pixel level. Moreover, food recognition technology has the potential to significantly improve the health and well-being of individuals by providing them with valuable insights into their eating behaviors. By tracking their food intake, people can better understand their dietary patterns and make informed decisions to improve their overall health and quality of life.

Semantic segmentation is an important task in the field of computer vision and getting a lot of attention due to deep learning techniques providing a high-level of accuracy for image analysis. The aim is to develop an automatic framework for food image analysis using deep learning to track and monitor the health and food intake of people. The developed system acquires images of the food eaten by the user or subject over time which will then be processed by the proposed food recognition model to extract and infer semantic information from the food images. We proposed a new approach in the context of our FoodRec project towards the challenging task of semantic food segmentation in order to develop a system capable of producing state-of-the-art results. We proposed a Convolutional Deconvolutional Pyramid Network for food segmentation to infer semantic information from the food images at the pixel level and to recognize individual food items in the image. The network employs convolution and

deconvolution layers to build a feature pyramid that generates a semantically strong and rich segmentation map of the input food image. Moreover, the detailed results were demonstrated on two benchmark food datasets for food segmentation performance evaluation and comparison of the proposed CDPN with the existing methods. Our proposed approach produced comparatively higher results with 91.77% mean IOU On TrayDataset and 77% mean IOU on MyFood dataset. Our proposed CDPN method achieved very competitive results as compared to the state-of-the-art approaches.

We propose another Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images at a pixel level to recognize different food items present in an image. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns while proposed CDPN also uses interpolation for features upsampling along with the deconvolution layers. Our proposed network demonstrated significant improvements in the results on the benchmark food dataset as compared to the state-of-the-art methods. Additionally, we also conducted a cross-data qualitative analysis of our proposed segmentation method to assess its generalization capabilities on our FoodRec dataset. By conducting cross-data experiments on the FoodRec dataset, we were able to determine that our method could effectively make accurate predictions in different contexts. This qualitative evaluation served as an important complement to our evaluation on the FoodRec dataset, and helped to strengthen our confidence in the effectiveness of our method.

In the future, we plan to deploy our algorithms to smartphone application to track and monitor the food intake of people participating in a smoking cessation program where they can upload food images of what they eat for dietary monitoring. Because, a significant correlation between smoking cessation and diet exists, resulting in adverse effects such as reduced appetite, weight loss, and other related outcomes. Our research would further allow people to estimate the volume and hence the quantities of each food item, nutrient intake assessment, and calorie estimation for health monitoring to raise awareness of their diet. We plan to extend our food segmentation research to a such applications to track the food intake of the people by simply taking a picture of what they consume to increase awareness of their daily diet. We believe that this work will contribute to the advancement of food analysis for health monitoring, and further applied to other related problems.

Chapter 6

Thesis Conclusions and Future Works

Food recognition is a dynamic interdisciplinary field focused on employing computational methodologies to acquire and analyze diverse food-related data from various sources. With the increasing availability of extensive datasets related to food, a range of computational techniques specialized in food recognition, spanning disciplines such as computer vision and machine learning, are either being widely adopted or swiftly developed to drive progress in the realm of food recognition. Given its interdisciplinary nature, food recognition holds relevance across numerous domains, including health, culture, agriculture, medicine, and biology. Food is not only vital for human survival but also deeply ingrained in the human journey. Research on food holds the potential to facilitate diverse applications, including shaping behavior, enhancing health, and unraveling culinary traditions. The proliferation of social networks, mobile connectivity, and the Internet of Things has resulted in the widespread sharing and recording of food-related content, resulting in an abundance of extensive food data. This data carries significant insights into food and its broader societal implications to cope with key human-centric challenges.

The food recognition project (FoodRec) aims to define an automatic framework using computer vision and deep learning techniques to recognize diverse foods from the images. The goal of food recognition is to extract and infer semantic information from the food images and to classify different foods present in the image. The developed system acquires images of the food eaten by the user or subject over time, which will then be processed by food recognition algorithms to extract and infer semantic information from the images containing food. In this context, we propose a novel user-biased Deep Convolutional Neural Network able to recognize food items of specific users and monitor their habits. It consists of a food branch to learn visual representation for the input food items and a user branch to take into account the specific user's eating habits. The proposed method is learning visual features of food items as traditional classification methods, in addition to learning the user who eats that food

item. The proposed method predicts the food item even if the user does not eat, but if the user eats that food item, then the model predicts with greater accuracy. The user branch in the classification architecture is aimed to improve the classification of difficult images and also to increase the confidence of the model. The proposed user-biased model is better compared to the traditional method as it predicts the food item even if the user does not eat, but if the user has that food item in his habit, then the model predicts with improved accuracy. Furthermore, we introduce a new FoodRec-50 dataset with 2000 images and 50 food categories collected by the iOS and Android smartphone applications, taken by 164 users during their smoking cessation therapy. Data preprocessing, data annotations, and data augmentation with different transformations are performed for further processing after the data has been collected by the application.

In the future, we aim to include more users habits and increase in food dataset categories. Our proposed user-biased food recognition method will perform even better with the increase in users' eating frequencies. During the current phase of our investigation, our dataset comprises a modest collection of instances with notably low occurrence rates of food items for individual users. This is reflective of the initial stages where individual consumption patterns are yet to be fully captured due to the limited data available. However, the trajectory of our research anticipates a noteworthy evolution. As the dataset becomes mature with time, a comprehensive archive of user-specific consumption data will be amassed, encompassing a diverse spectrum of food items ingested by each individual. This temporal evolution of the dataset constitutes a pivotal catalyst in fostering the performance and efficacy of our model. Hence, with the influx of abundant training data for each food item consumed by specific users, the model performance will be outstanding as it will learn the user-specific habits with more training data for each food item for the specific user.

For food segmentation, we propose a novel Convolutional Deconvolutional Pyramid Network (CDPN) for food segmentation to understand the semantic information of an image at a pixel level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. We propose another Food Convolutional Deconvolutional Network (FCDN) for semantic segmentation to extract and infer semantic information from the food images and to recognize different food items present in an image. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns, while the proposed CDPN also uses interpolation for features upsampling along with the deconvolution layers. Our proposed networks demonstrated significant improvements in the results on the benchmark food dataset as compared to the state-of-the-art methods.

The proposed FCDN for semantic segmentation is the extended version of CDPN to extract and infer semantic information from the food images at a pixel level. The proposed FCDN employs only learnable features upsampling using deconvolution layers to increase the spatial resolution of the feature maps and to learn the complex patterns, while the proposed Convolutional Deconvolutional Pyramid Network (CDPN) also uses interpolation for features upsampling along with the deconvolution layers. Deconvolution is achieved using ConvTranspose2d, that is a learnable upsampling technique that uses transposed convolutional layers to increase the spatial resolution of the feature maps. It has the advantage of being able to learn complex patterns and features specific to the given task, however, it is computationally expensive. Bilinear interpolation is a non-learnable upsampling technique that is based on a weighted average of the nearest neighboring pixels in the input image. It is simple and computationally efficient and can be used as a baseline for comparison with more complex methods. However, it may not capture complex patterns and features in the input image and can result in blurry output. In conclusion, if the goal is to achieve high accuracy and capture complex patterns in the upsampling process, ConvTranspose2d is a better choice. If the goal is to achieve a balance between computational efficiency and reasonable quality, bilinear interpolation is a more suitable option. Ultimately, the choice of upsampling technique should be made based on the specific needs and requirements of the task at hand.

The food segmentation is performed on the benchmark food datasets such as MyFood and TryDataset. This enables to make comparisons and to measure the performance of the proposed method with state-of-the-art food segmentation techniques. The FoodRec dataset is useful for the segmentation but not annotated for the segmentation task at this moment. In fact, cross-data experiments are conducted by training the model on the MyFood dataset and testing on the FoodRec dataset to evaluate the performance of our proposed segmentation method. Both MyFood and our FoodRec datasets contain common food classes, such as apple, beans, egg, spaghetti, chicken, rice, and salad. The qualitative results of our proposed method are evaluated on a subset of our FoodRec dataset. So, In addition to evaluating the performance of our proposed segmentation method on the MyFood dataset, we also performed cross-data experiments to assess its generalization capabilities on our FoodRec dataset. By conducting cross-data experiments on the FoodRec dataset, we were able to determine that our method could effectively make accurate predictions in different contexts. This qualitative evaluation served as an important complement to our evaluation on the FoodRec dataset and helped to strengthen our confidence in the effectiveness of our method. By performing the cross-data experiments, we aimed to simulate a scenario where the model is trained on a MyFood dataset and is tested to perform segmentation on the MyFood dataset. The results of our experiments demonstrate that our proposed method is capable of generalizing well to new and unseen FoodRec dataset as well.

In the future, we aim to estimate the weight (i.e., quantities) of each food item detected within an image. This task results very challenging because it involves the estimation of 3D information at very small scale detail. Precisely determining the weight of food items constitutes a pivotal element within the realms of clinical assessments and research investigations centered on dietary patterns. The weight is estimated to calculate the food nutrient content after the food image has been segmented and recognized. Broadly, there are two primary methodologies to assess the weight of food items. The first approach involves estimating the volume of the food and subsequently utilizing density information specific to that particular food category to infer its weight (Kelkar et al., 2011). A food specific shape template method to reconstruct a 3D model of the food item is implemented to estimate the food volume. On the other hand, the second method entails a direct estimation of the food's weight by considering its area in conjunction with training data.

In the future, we plan to deploy our algorithms to smartphone applications to track and monitor the food intake of people participating in a smoking cessation program where they can upload food images of what they eat for dietary monitoring. Because a significant correlation between smoking cessation and diet exists, resulting in adverse effects such as reduced appetite, weight loss, and other related outcomes. Our research would further allow people to estimate the volume and, hence, the quantities of each food item, nutrient intake assessment, and calorie estimation for health monitoring to raise awareness of their diet. Food recognition technology can be used to identify the food items in images and provide detailed nutrition information about the meal. Food recognition technology is used to accurately identify food items in images taken by the user, allowing them to keep track of their calorie and nutrient intake. Food recognition technology can be used to identify the ingredients in a dish and suggest recipes that include those ingredients. Food recognition technology can be used to identify food items in images and track consumer preferences, allowing for more targeted marketing and advertising strategies. Food recognition technology can be used to identify the food items in images of meals, allowing for faster and more accurate delivery and restaurant recommendations. We plan to extend our food recognition research to such applications to track the food intake of people by simply taking a picture of what they consume to increase awareness of their daily diet. We believe that this work will contribute to the advancement of food analysis for health monitoring and further applied to other related problems.

The developed dietary monitoring system could be extended to work with videos recorded by a fixed camera system, considering a set of cameras recording the scene from different fixed points of view. The collected data about the mood associated to food images can be combined with approaches related to sentiment analysis based on images (Ortis et al., 2020). Such approaches can be investigated in order to automatically infer the mood of the user (e.g., depression, happiness, etc.) based on dietary

monitoring, avoiding to ask the user about his/her mood.

Food recognition holds the potential for a multitude of promising applications within various specialized domains. One prominent instance is its capacity to foster diverse applications within the realm of smart homes, encompassing areas like the smart kitchen and personalized nutrition tracking. Within smart-home ecosystems, food recognition methodologies can gather invaluable insights into users' preferences, nutritional intake, and health metrics, leveraging techniques such as food recognition and comprehension of cooking videos. Illustratively, prior research as evidenced by (Kojima et al., 2015) has leveraged textual information to enhance audio-visual scene understanding for culinary support robots. Anticipating the future trajectory, the evolution of smart kitchen robots necessitates augmented functionality, more sophisticated multimodal interactions, and enriched dialogue capabilities. Here, the synergy among food recognition, recipe recommendation, and food-related text processing is envisioned to be instrumental in realizing these ambitious objectives.

As the field of food recognition continues its trajectory of development, its influence and application are poised to transcend traditional boundaries, empowering a diverse array of specialized sectors to harness its capabilities for enhanced efficiency, safety, and user experience.

Appendix: Nose to Brain Drug Delivery Data Classification

(Note: In this section, we present additional research work with the collaboration of the Department of Drug and Health Science, University of Catania, done during the Ph.D. but not directly related to this thesis.)

Today efforts are being made to exploit nanomedicine to directly target the brain through intranasal administration, which utilizes the olfactory neurons and trigeminal nerves (maxillary and ophthalmic branches) and reducing systemic involvement. Using devices that facilitate drug delivery to the olfactory region and the brain is critical. Conducting this type of research involves assessing the efficacy of intranasal drug administration in comparison to other routes of delivery and identifying the neural pathways involved if the researcher claims direct access to the brain. The efficiency evaluation should encompass quantitative analysis of drug levels in brain tissue and blood as well as behavioral studies in animal models. By carefully evaluating drug delivery methods and identifying the pathways that facilitate brain access, researchers can develop more effective methods.

After demonstrating the efficacy of intranasal administration, it is important to determine whether the drug has reached the brain directly or has been absorbed systemically. Nanoparticles can serve as vectors to transport drugs through axonal nerves to the brain, where they are released. Alternatively, drugs can be released in the nose, where they take the pathway that leads directly to the brain through the olfactory or trigeminal nerves (Ahmad et al., 2017). The fate of drugs or drug-loaded nanocarriers in the brain depends on the chemical and physical characteristics of the nanomedicine under study.

Two important indices, namely %DTE and %DTP that allow for an assessment of the effectiveness of a formulation in reaching the brain via direct access pathways, such

as the olfactory or trigeminal pathways, following intranasal administration. These parameters are integral for obtaining a comprehensive understanding of the formulation data, which enables a quantitative evaluation of whether the drug has utilized the direct nose-to-brain pathways to reach the brain.

%DTE refers to drug targeting efficiency. This targeting index is defined differently by researchers. According to Pokharkar (Pokharkar et al., 2020) "%DTE represents the time average partitioning ratio of the drug", between the blood and the brain. According to Kozlovskaya (Kozlovskaya et al., 2014) "%DTE indicates the relative exposure of the brain to the drug following intranasal administration vs. systemic administration". It's the partitioning ratio between brain tissue and plasma through the IN route vs. the IV route (Mao et al., 2019). The percentage of medication that enters the brain after intranasal delivery following the direct channels (trigeminal and/or olfactory neural pathways) is referred to as the nose-to-brain direct transport percentage (%DTP). High DTP levels suggest a high percentage of direct transport from the nose to the brain.

The aim of this study was to find a correlation between well-defined and selected parameters such as the particle size, the surface charge zeta potential, the type of nanocarrier, and the targeting efficiency indexes %DTE and %DTP. When a nanomedicine is administered intranasally, researchers employ the DTP to characterize trigeminal and olfactory participation rather than the systemic pathway. Furthermore, the correlation studies took into consideration the possible impact of the molecular weight of the conveyed drug.

This research study aimed to provide a comprehensive analysis of the literature data on PubMed regarding scientific research on nanomedicine for Nose-to-brain drug delivery in the last ten years, from 2010 to February 2021. The selected research papers were studied to gain insights into the developments in this field over the last decade. To ensure accurate analysis of the data, a database was constructed, taking great care to allow for the creation of mathematical models. A bibliographic research was conducted on "nanomedicine and nose-to-brain drug delivery" to gather relevant data. The ultimate goal was to evaluate the progress made in this field of research over the past decade. The selection criteria for the articles were carefully chosen to include only those of actual interest to the project's objectives. After collecting the relevant articles, a quantitative analysis was conducted, paying particular attention to identifying the AUC values necessary for calculating the targeting indexes, such as AUC brain and AUC blood relative to nasal and systemic administration. If the examined articles did not report the DTE and DTP targeting indexes, they were calculated using the AUC values. To ensure data homogeneity, any values of these parameters that were already calculated in the selected articles were recalculated and reported in this work by rounding to the second decimal number and neglecting the standard deviation. Therefore, some values may be slightly different from those reported in the original research articles. Overall, this study provides a detailed analysis of the literature data

on nanomedicine for nose-to-brain drug delivery in the last decade, allowing for the evaluation of progress made in this field of research and the creation of a database that can be used for further analysis.

7.1 Research Methodology

This study involves several phases, which are the following:

- Data Collection
- Data Cleanup
- Data Conversion
- Data Standardization
- Data Classification

7.1.1 Data Collection

In order to identify scientific literature reporting studies on nose-to-brain drug delivery, a search was conducted in the PubMed database using specific keywords. The search date was February 18, 2021, and this date is referred to as "the search date" in this thesis. During the bibliographic research process, several criteria were established to screen the publications. The selected articles were those that met the following criteria: 1) publication date no earlier than January 2010, 2) reporting of particle size and zeta potential, 3) in vivo and/or in vitro studies are conducted in each article, 4) biodistribution studies of the drug conducted via both intranasal and intravenous or oral routes, and 5) reporting of the AUC values for both the brain and blood concentration versus time curve of the drug for both the intranasal and parenteral routes. Only the research papers that fulfilled these criteria were included in the data analysis.

Once the relevant articles were collected, a quantitative analysis was performed. Special attention was given to identifying the AUC values necessary for calculating the targeting indexes. AUC values for both the brain and blood were reported in the tables or graphics of each article for both nasal and systemic administration. If these values were not reported in the articles, they were calculated using the DTE and DTP targeting indexes. In order to ensure homogeneity of the collected data, any previously calculated values of these parameters in the selected articles were recalculated and reported in this study, with standard deviation neglected and rounded to the second

decimal place. As a result, some values may differ slightly from those reported in the original research articles.

To conduct a thorough analysis of the data, we started to work on the construction of a database that would facilitate the creation of mathematical models. Before proceeding with the IT analysis, we conducted extensive groundwork to gather the necessary data. Specifically, we undertook a comprehensive bibliographic search on ”Nanomedicine and nose-to-brain drug delivery” to evaluate the latest advancements in this field over the past decade. Our search criteria enabled us to select only the most relevant articles for the project. We then used the data gathered from these selected articles to construct the ”NANOSE” database, which was further refined to create a second database called ”NANOSE2B” for computer-based study. The collected data contains several attributes that include pharmacological category, particle size, zeta potential, molecular mass of the active ingredient, name of the active ingredient, type of nanocarrier used, DTE, and DTP. With this rich data, we can develop a comprehensive understanding of the various aspects of nanomedicine and nose-to-brain drug delivery, thereby enabling us to assess the progress made in this field over the past decade.

7.1.2 Data Cleanup

Data cleanup activity has been performed to make it useful for the analysis. A data cleanup activity was first carried out starting from the initial database. All the data collected should be used in the analysis, each row was corrected in order to refer to a single value. Since particle size and zeta potential columns values are in range, we extracted both extreme and mean values for particle size and zeta potential to perform the further experiments. Missing values are adjusted by taking mean based on the same type/subtype or pharmacological categories.

7.1.3 Data Conversion

It was necessary to perform a conversion from categorical to numerical data for some columns. Data conversion from categorical values to numerical is needed for data analysis and training a model because most machine learning algorithms and models are designed to work with numerical data. Categorical data, on the other hand, represents discrete values and cannot be used directly in most machine learning algorithms. The following columns or attributes are converted from categorical to numerical pharmacological categories, active agent, and type/subtype.

7.1.4 Data Standardization

In order to perform a correct data analysis, the standardization process must be carried out. During this phase, the aim is to illustrate the study about the analysis, properly treating and modeling the data. Standardization is a process of data transformation consisting of a change of scale, in order to be able to directly compare the data between each other. In this way, we define a new range where these values will be inserted. The z-scoring strategy for data standardization is applied to perform a correct data analysis. The z-scoring standardization reports the data in a new range where the mean of the data is equal to 0 and the standard deviation is equal to 1.

The standardization equation 7.1 is used to transform a given variable to a standard normal distribution. The equation is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{7.1}$$

Where z is the standardized value, x is the original value of the variable, μ is the mean of the data, and σ is the standard deviation of the data. The equation shows that to standardize a variable, you need to subtract the mean from the variable and then divide the result by the standard deviation. The resulting value z will have a mean of 0 and a standard deviation of 1, making it easier to compare with other standardized variables.

7.1.5 Data Classification

Classification experiments have been conducted using different state-of-the-art machine learning algorithms such as Support Vector Machine (SVC), NuSVC, RandomForestClassifier, ExtraTreesClassifier, BaggingClassifier, DecisionTreeClassifier, GradientBoostingClassifier, RidgeClassifierCV, LinearDiscriminantAnalysis, BernoulliNB, AdaBoostClassifier, LinearSVC, GaussianProcessClassifier, KNeighborsClassifier, GaussianNB, PassiveAggressiveClassifier, and SGDClassifier.

Machine Learning Algorithm	10-fold Accuracy (mean)	10-fold Accuracy (std)
NuSVC	0.74	0.19
RandomForestClassifier	0.65	0.18
ExtraTreesClassifier	0.70	0.20
BaggingClassifier	0.63	0.19
DecisionTreeClassifier	0.65	0.20
GradientBoostingClassifier	0.63	0.17
RidgeClassifierCV	0.65	0.22
LinearDiscriminantAnalysis	0.65	0.22
BernoulliNB	0.66	0.24
AdaBoostClassifier	0.63	0.17
LinearSVC	0.65	0.22
GaussianProcessClassifier	0.63	0.15
KNeighborsClassifier	0.68	0.17
SVC	0.64	0.15
GaussianNB	0.60	0.16
PassiveAggressiveClassifier	0.60	0.18
SGDClassifier	0.45	0.27

Figure 7.1: 10-fold accuracy Comparison using particle size and zeta potential mean values.

Results using particle size and zeta potential Mean Values

The Figure 7.1 presents the results of different machine learning algorithms on the dataset using 10-fold cross-validation. The mean accuracy and standard deviation of accuracy across the 10 folds are reported for each algorithm. The results suggest that NuSVC, ExtraTreesClassifier, and KNeighborsClassifier are the top-performing algorithms, while SGDClassifier performs the lowest. We have also reported results for all the classifiers using 20% data as test split using only two attributes particle size and zeta potential mean values as shown in Figure 7.2.

Algorithm used	Precission	Recall	F-score	Accuracy
AdaBoostClassifier	0.750000	0.6	0.666667	0.70
RandomForestClassifier	0.714286	0.5	0.588235	0.65
ExtraTreesClassifier	0.714286	0.5	0.588235	0.65
NuSVC	0.800000	0.4	0.533333	0.65
BaggingClassifier	0.666667	0.4	0.500000	0.60
GradientBoostingClassifier	0.625000	0.5	0.555556	0.60
DecisionTreeClassifier	0.625000	0.5	0.555556	0.60
KNeighborsClassifier	0.625000	0.5	0.555556	0.60
GaussianProcessClassifier	0.600000	0.3	0.400000	0.55
SVC	0.600000	0.3	0.400000	0.55
RidgeClassifierCV	0.555556	0.5	0.526316	0.55
LinearDiscriminantAnalysis	0.555556	0.5	0.526316	0.55
LinearSVC	0.555556	0.5	0.526316	0.55
SGDClassifier	0.555556	0.5	0.526316	0.55
BernoulliNB	0.500000	0.5	0.500000	0.50
GaussianNB	0.500000	0.4	0.444444	0.50
PassiveAggressiveClassifier	0.500000	0.5	0.500000	0.50

Figure 7.2: Results for all the classifiers using 20% data as test split using only two attributes particle size and zeta potential mean values.

We have reported classification results using particle size and zeta potential mean/middle values where type of DTE/DTP is the target attribute. In this experiment, 10-fold mean accuracy is obtained for all the classifiers using only two attributes particle size and zeta potential mean values. The comparison of different algorithms is given in the table with 10-fold accuracy mean and standard deviation.

Among all the algorithms, NuSVC has the highest mean accuracy of 0.74, followed by ExtraTreesClassifier with 0.70 and KNeighborsClassifier with 0.68 mean accuracy. On the other hand, SGDClassifier has the lowest mean accuracy of 0.45. Box plot comparison for 10-fold is shown in the Figure 7.3.

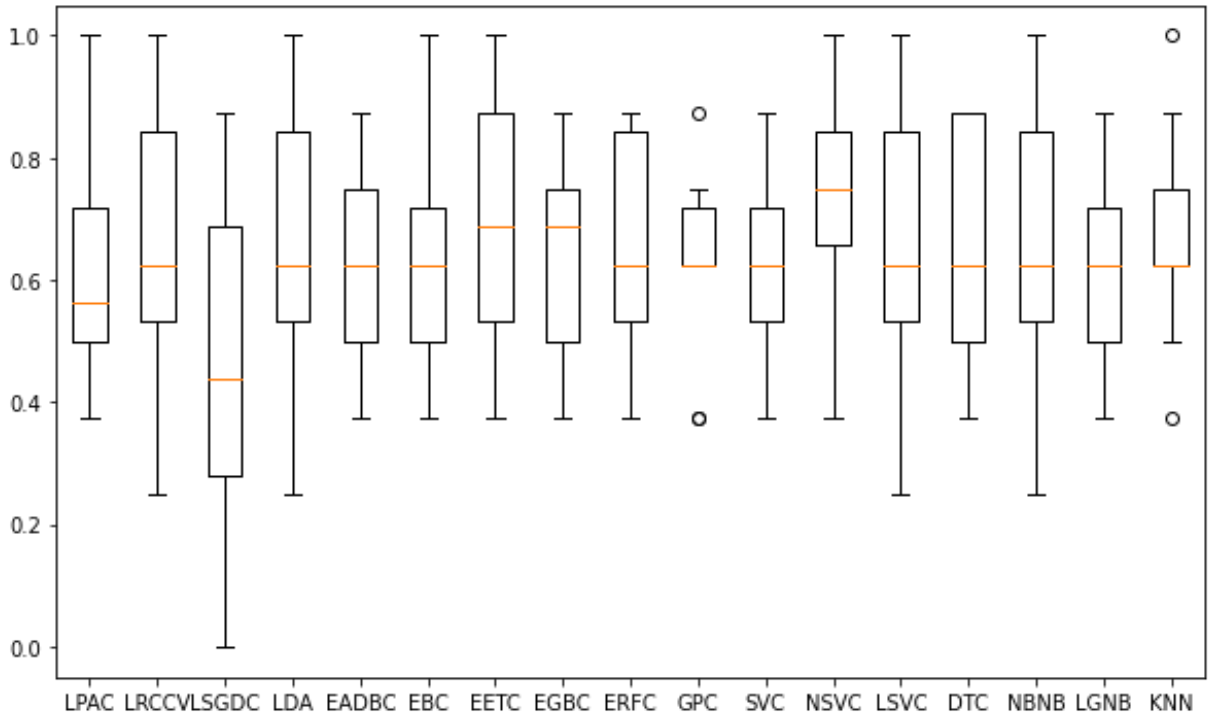


Figure 7.3: Box plot 10-fold comparison using particle size and zeta potential mean values.

Results using particle size and zeta potential Extreme Values

We have performed classification results using particle size and zeta potential extreme values where type of DTE/DTP is the target attribute. In this experiment, 10-fold mean accuracy is obtained for all the classifiers using only two attributes particle size and zeta potential extreme values. The comparison of different algorithms is given in

the Figure 7.4 with 10-fold accuracy mean and standard deviation. The table shows the performance of different machine learning algorithms in terms of their 10-fold accuracy mean and standard deviation. Among the algorithms listed, NuSVC and RandomForestClassifier have the highest accuracy mean of 0.70 and 0.69, respectively, while the SGDClassifier has the lowest mean accuracy of 0.45. We have also reported results for all the classifiers using 20% data as test split using only two attributes particle size and zeta potential mean values, as shown in Figure 7.5. Box plot comparison for 10-fold is shown in Figure 7.6.

Machine Learning Algorithm	10-fold Accuracy (mean)	10-fold Accuracy (std)
NuSVC	0.70	0.19
RandomForestClassifier	0.69	0.19
ExtraTreesClassifier	0.69	0.17
BaggingClassifier	0.66	0.16
DecisionTreeClassifier	0.66	0.19
GradientBoostingClassifier	0.66	0.17
RidgeClassifierCV	0.66	0.22
LinearDiscriminantAnalysis	0.66	0.22
BernoulliNB	0.66	0.24
AdaBoostClassifier	0.65	0.16
LinearSVC	0.65	0.22
GaussianProcessClassifier	0.65	0.13
KNeighborsClassifier	0.63	0.19
SVC	0.63	0.14
GaussianNB	0.63	0.15
PassiveAggressiveClassifier	0.58	0.13
SGDClassifier	0.45	0.26

Figure 7.4: 10-fold accuracy comparison using particle size and zeta potential extreme values.

Algorithm used	Precission	Recall	F-score	Accuracy
RandomForestClassifier	0.833333	0.5	0.625000	0.70
DecisionTreeClassifier	0.714286	0.5	0.588235	0.65
AdaBoostClassifier	0.714286	0.5	0.588235	0.65
BaggingClassifier	0.750000	0.3	0.428571	0.60
NuSVC	0.666667	0.4	0.500000	0.60
ExtraTreesClassifier	0.625000	0.5	0.555556	0.60
GradientBoostingClassifier	0.625000	0.5	0.555556	0.60
LinearSVC	0.555556	0.5	0.526316	0.55
KNeighborsClassifier	0.571429	0.4	0.470588	0.55
RidgeClassifierCV	0.555556	0.5	0.526316	0.55
LinearDiscriminantAnalysis	0.555556	0.5	0.526316	0.55
SGDClassifier	0.500000	0.5	0.500000	0.50
BernoulliNB	0.500000	0.5	0.500000	0.50
PassiveAggressiveClassifier	0.500000	0.1	0.166667	0.50
GaussianProcessClassifier	0.500000	0.3	0.375000	0.50
SVC	0.500000	0.3	0.375000	0.50
GaussianNB	0.444444	0.4	0.421053	0.45

Figure 7.5: Results for all the classifiers using 20% data as test split using only two attributes particle size and zeta potential extreme values.

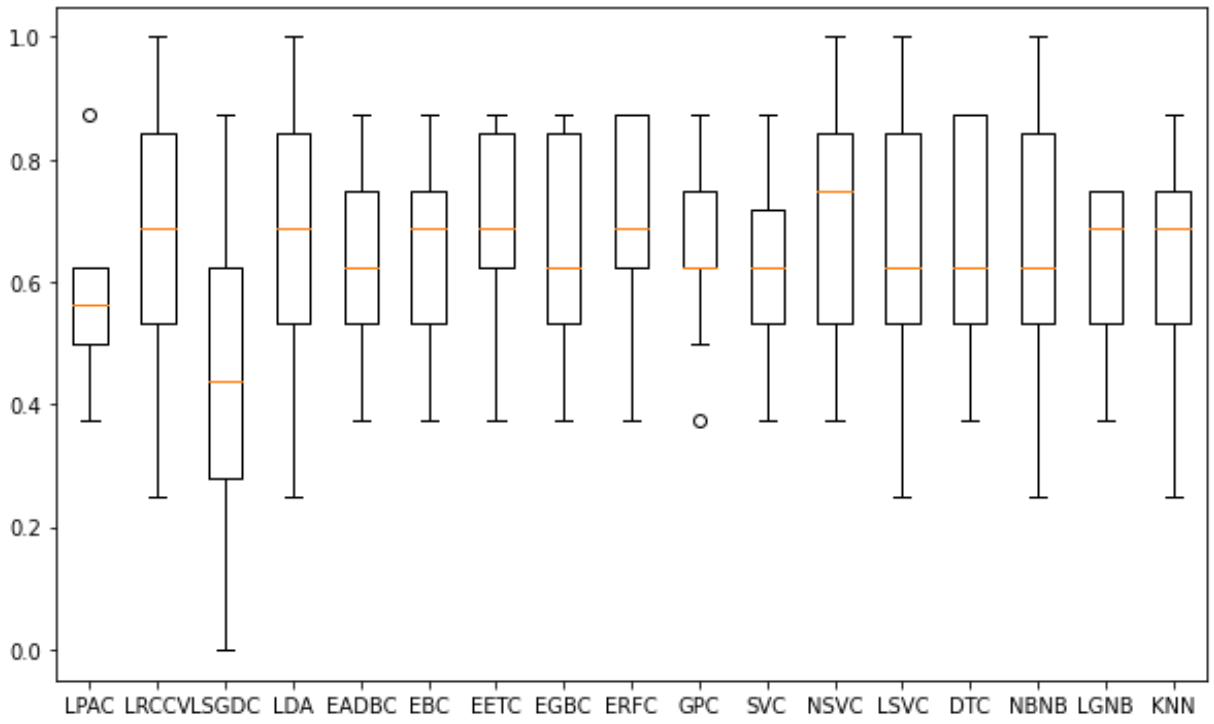


Figure 7.6: Box plot 10-fold comparison using particle size and zeta potential extreme values.

7.1.6 Results Evaluation

Objective: To understand if specific ranges of zeta potential and/or particle size are correlated with good predictions.

1. Zeta Potential

We have created a histogram for the classifier's success percentage in a specific range of zeta potential where the x-axis shows the classifiers and the y-axis shows the percentage of each classifier success in each bin range as shown in Figure 7.7. The data is divided into 5 bins, where the first bin starting from minimum value and last bin ending with maximum value of zeta potential as you can see the bins in Figure 7.8. For example, the first bin range is (-47.5, -25.5] and this bin contains n=9 instances or elements in that range. Then, plot is created only for the percentage of true values for each classifier in that range. True value means the classifier gets the right target result using zeta potential values. For example, if the zeta potential value is -47.5 and target

for type of DTE/DTP is NCvsDRUG SOLUTION, then the classifier gets success only if the result is NCvsDRUG SOLUTION.

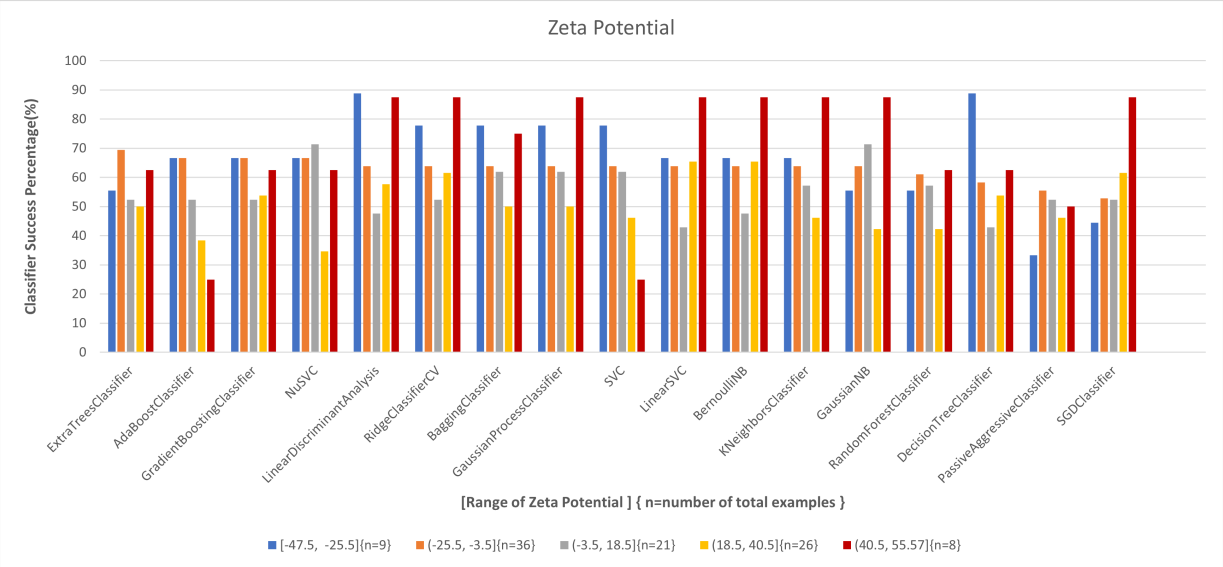


Figure 7.7: Classifier success percentage using zeta potential attribute.

If we observe the graph, the ranges A and E (blue and red bins in the plot) have good percentage results, but these ranges contain a limited number of elements (9 and 8 respectively), making this statement not generalizable from a statistical point of view. Regarding the remaining ranges, in general, good results are associated with range B (orange bins), followed by C and D. In some cases, elements in C have comparable results with those in B such as for the nuSVC, Bagging, Gaussian Process or SVC classifiers.

Ranges	No. of Examples
A. [-47.5, -25.5]	9
B. (-25.5, -3.5]	36
C. (-3.5, 18.5]	21
D. (18.5, 40.5]	26
E. (40.5, 55.57]	8

Figure 7.8: Zeta potential bin ranges and their number of elements.

2. Particle Size

We have plotted the classifier success percentage as shown in Figure 7.9 in a specific range of particle size where the data is divided into 5 bins, as you can see the bins in Table 7.10. .

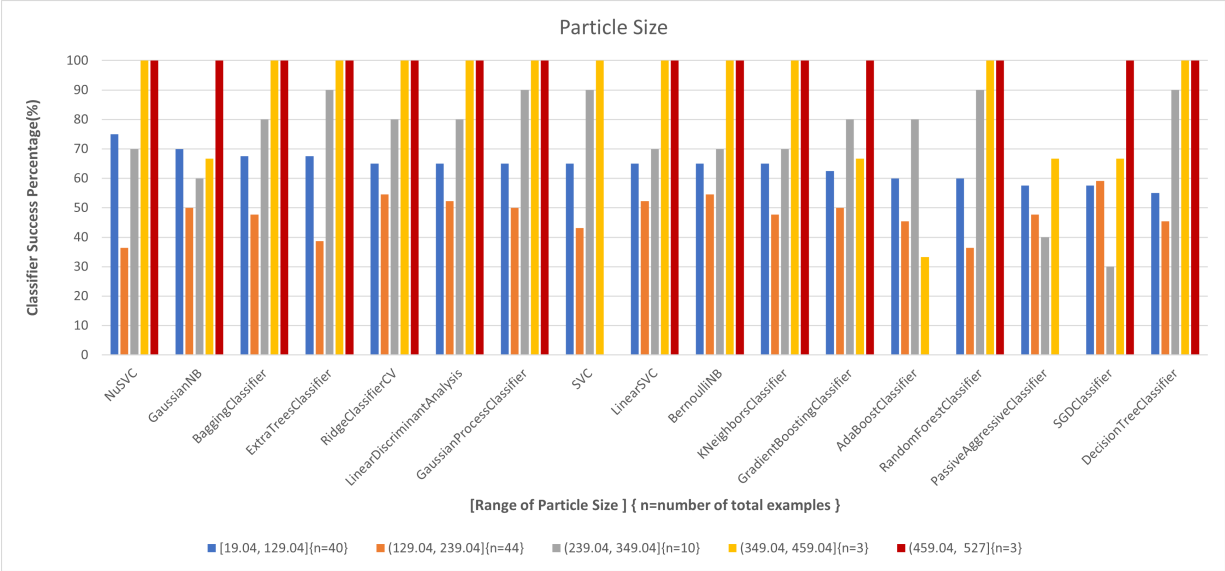


Figure 7.9: Classifier success percentage using particle size attribute.

As done previously, although results associated with ranges D and E are perfect (i.e., 100%), since the number of elements in this range is limited to only 3 examples each, we can't state any statistically evident statement. The same applies to range C, which has a slightly higher number of elements (10) and good success classification percentages between 70% and 90%. For the remaining ranges, we can observe that all classifiers have a better success rate with elements in range A with respect to inputs from range B. In particular, success rates associated to range A span from about 55% to 75%, whereas results associated to range B are in the range 35% - 55%.

Ranges	No. of Examples
A. [19.04, 129.04]	40
B. (129.04, 239.04]	44
C. (239.04, 349.04]	10
D. (349.04, 459.04]	3
E. (459.04, 527]	3

Figure 7.10: Particle size bin ranges and their number of elements.

7.2 Conclusion

We performed nose-to-brain drug delivery data preprocessing, and data classification using state-of-the-art machine learning algorithms. To collect the dataset required for the analysis, a comprehensive search was performed on the PubMed database to identify research articles on "nanomedicine and nose-to-brain drug delivery," which were published between January 2010 and February 2021. The search criteria were carefully defined to ensure the selection of relevant articles. Particular attention was given to the presence of biodistribution studies, as this information was essential for the calculation of the targeting indices %DTE and %DTP. After the data collection, we performed data prepossessing such as data cleaning, data conversion, and data standardization to correctly analyze and process the data for training the machine learning models. Further, classification experiments have been conducted using different state-of-the-art machine learning algorithms using particle size and zeta potential mean values and extreme values. Results show that NuSVC achieved the highest 10-fold mean accuracy as compared to other classifiers. We also reported the results to understand if specific ranges of zeta potential and/or particle size are correlated with good predictions.

Funding

This research was funded by ECLAT srl, a spin-off of the University of Catania, with the help of a grant from the Foundation for a Smoke-Free World Inc., a US nonprofit 501(c)(3) private foundation with a mission to end smoking in this generation. The contents, selection, and presentation of facts, as well as any opinions expressed herein are the sole responsibility of the authors and under no circumstances shall be regarded as reflecting the positions of the Foundation for a Smoke-Free World, Inc. ECLAT srl. is a research based company from the University of Catania that delivers solutions to global health problems with special emphasis on harm minimization and technological innovation

References

Ahmad, E., Y. Feng, J. Qi, W. Fan, Y. Ma, H. He, F. Xia, X. Dong, W. Zhao, Y. Lu, et al. (2017). Evidence of nose-to-brain delivery of nanoemulsions: cargoes but not vehicles. *Nanoscale 9*(3), 1174–1183.

Allegra, D., S. Battiato, A. Ortis, S. Urso, and R. Polosa (2020). A review on food recognition technology for health applications. *Health Psychology Research 8*(3).

Amugongo, L. M., A. Kriebitz, A. Boch, and C. Lütge (2023). Mobile computer vision-based applications for food recognition and volume and calorific estimation: A systematic review. In *Healthcare*, Volume 11, pp. 59. Multidisciplinary Digital Publishing Institute.

Ansari, M. A., D. Kurchaniya, and M. Dixit (2017). A comprehensive analysis of image edge detection techniques. *International Journal of Multimedia and Ubiquitous Engineering 12*(11), 1–12.

Anthimopoulos, M., J. Dehais, P. Diem, and S. Mougiakakou (2013). Segmentation and recognition of multi-food meal images for carbohydrate counting. In *13th IEEE international conference on bioinformatics and bioengineering*, pp. 1–4. IEEE.

Anthimopoulos, M. M., L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou (2014). A food recognition system for diabetic patients based on an optimized bag-of-features model. *IEEE journal of biomedical and health informatics 18*(4), 1261–1271.

Aslan, S., G. Ciocca, and R. Schettini (2018). Semantic food segmentation for automatic dietary monitoring. In *2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, pp. 1–6. IEEE.

Badrinarayanan, V., A. Kendall, and R. Cipolla (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence 39*(12), 2481–2495.

Basrur, A., D. Mehta, and A. R. Joshi (2022). Food recognition using transfer learning. In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1–5. IEEE.

Battiato, S., P. Caponnetto, O. Giudice, M. Hussain, R. Leotta, A. Ortis, and R. Polosa (2021). Food recognition for dietary monitoring during smoke quitting. In *IMPROVE*, pp. 160–165.

Battiato, S., P. Caponnetto, R. Leotta, G. Marotta, A. Midolo, A. Ortis, and R. Polosa (2023). Development and user evaluation of a food-recognition app (foodrec): Experimental data and qualitative analysis. *Health Psychology Research 11*.

Bhargavi, K. and S. Jyothi (2014). A survey on threshold based segmentation technique in image processing. *International Journal of Innovative Research and Development 3*(12), 234–239.

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

Bossard, L., M. Guillaumin, and L. Van Gool (2014). Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer.

Buslaev, A., V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin (2020). Albumentations: fast and flexible image augmentations. *Information 11*(2), 125.

Chatnuntawech, I., K. Tantisantisom, P. Khanchaitit, T. Boonkoom, B. Bilgic, and E. Chuangsuwanich (2018). Rice classification using spatio-spectral deep convolutional neural network. *arXiv preprint arXiv:1805.11491*.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence 40*(4), 834–848.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.

Chen, M.-Y., Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung (2012). Automatic chinese food identification and quantity estimation. In *SIGGRAPH Asia 2012 Technical Briefs*, pp. 1–4.

Chopra, M. and A. Purwar (2022). Recent studies on segmentation techniques for food recognition: A survey. *Archives of Computational Methods in Engineering 29*(2), 865–878.

Christodoulidis, S., M. Anthimopoulos, and S. Mougiakakou (2015). Food recognition for dietary assessment using deep convolutional neural networks. In *New Trends in Image Analysis and Processing–ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proceedings 18*, pp. 458–465. Springer.

Ciocca, G., P. Napoletano, and R. Schettini (2016). Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics 21*(3), 588–598.

Dalakleidi, K. V., M. Papadelli, I. Kapolos, and K. Papadimitriou (2022). Applying image-based food-recognition systems on dietary assessment: A systematic review. *Advances in Nutrition 13*(6), 2590–2619.

Dehais, J., M. Anthimopoulos, and S. Mougiakakou (2016). Food image segmentation for dietary assessment. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management*, pp. 23–28.

Dehariya, V. K., S. K. Shrivastava, and R. Jain (2010). Clustering of image data set using k-means and fuzzy k-means algorithms. In *2010 International conference on computational intelligence and communication networks*, pp. 386–391. IEEE.

e Silva, B. V. R., M. G. Rad, J. Cui, M. McCabe, and K. Pan (2018). A mobile-based diet monitoring system for obesity management. *Journal of health & medical informatics 9*(2).

Fakhrou, A., J. Kunhoth, and S. Al Maadeed (2021). Smartphone-based food recognition system using multiple deep cnn models. *Multimedia Tools and Applications 80*(21), 33011–33032.

Farinella, G. M., D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato (2016). Retrieval and classification of food images. *Computers in biology and medicine 77*, 23–39.

Farinella, G. M., M. Moltisanti, and S. Battiato (2015). Food recognition using consensus vocabularies. In *New Trends in Image Analysis and Processing–ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proceedings 18*, pp. 384–392. Springer.

Foroni, F., G. Pergola, G. Argiris, and R. I. Rumiati (2013). The foodcast research image database (frida). *Frontiers in human neuroscience 7*, 51.

Freitas, C., F. Cordeiro, and V. Macario (2020a, September). Myfood dataset. https://doi.org/10.5281/zenodo.4041488.

Freitas, C. N., F. R. Cordeiro, and V. Macario (2020b). Myfood: A food segmentation and classification system to aid nutritional monitoring. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 234–239. IEEE.

Ganesan, P. and G. Sajiv (2017). A comprehensive study of edge detection for image processing applications. In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)*, pp. 1–6. IEEE.

Giampiccoli, A. and J. H. Kalis (2012). Tourism, food, and culture: Community-based tourism, local food, and community development in m pondoland. *Culture, Agriculture, Food and Environment 34*(2), 101–123.

Giovany, S., A. Putra, A. S. Hariawan, and L. A. Wulandhari (2017). Machine learning and sift approach for indonesian food image recognition. *Procedia computer science 116*, 612–620.

Harshitha, I., T. Sunil Kumar, and P. Venkateswara Rao (2023). Estimation of dietary calories using image processing. In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2*, pp. 357–372. Springer.

Hassannejad, H., G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni (2016). Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management*, pp. 41–49.

He, H., F. Kong, and J. Tan (2015). Dietcam: multiview food recognition using a multikernel svm. *IEEE journal of biomedical and health informatics 20*(3), 848–855.

He, K., G. Gkioxari, P. Dollar, and R. Girshick (2018). Mask r-cnn. facebook ai research (fair). *arXiv preprint arXiv:1703.06870*.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Y., C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp (2013). Food image analysis: Segmentation, identification and weight estimation. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE.

He, Y., C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp (2014). Analysis of food images: Features and classification. In *2014 IEEE international conference on image processing (ICIP)*, pp. 2744–2748. IEEE.

Hemamalini, V., S. Rajarajeswari, S. Nachiyappan, M. Sambath, T. Devi, B. K. Singh, and A. Raghuvanshi (2022). Food quality inspection and grading using efficient image segmentation and machine learning-based system. *Journal of Food Quality 2022*, 1–6.

Hoashi, H., T. Joutou, and K. Yanai (2010). Image recognition of 85 food categories by feature fusion. In *2010 IEEE International Symposium on Multimedia*, pp. 296–301. IEEE.

Hu, J., L. Shen, and G. Sun (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Hussain, G., M. K. Maheshwari, M. L. Memon, M. S. Jabbar, and K. Javed (2019). A cnn based automated activity and food recognition using wearable sensor for preventive healthcare. *Electronics 8*(12), 1425.

Hussain, M., A. Ortis, R. Polosa, and S. Battiato (2022). User-biased food recognition for health monitoring. In *International Conference on Image Analysis and Processing*, pp. 98–108. Springer.

Hussain, M., A. Ortis, R. Polosa, and S. Battiato (2023). Semantic food segmentation for health monitoring. In *Fifteenth International Conference on Machine Vision (ICMV 2022)*, Volume 12701, pp. 106–113. SPIE.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675–678.

Jiménez-Carvelo, A. M., A. González-Casado, M. G. Bagur-González, and L. Cuadros-Rodríguez (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity–a review. *Food research international 122*, 25–39.

Jocher, G., A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, A. Laughing, A. Hogan, J. Hajek, L. Diaconu, Y. Marc, et al. (2021). ultralytics/yolov5: v5. 0-yolov5-p6 1280 models aws supervise. ly and youtube integrations. *Zenodo 11*.

Joshua, S. R., S. Shin, J.-H. Lee, and S. K. Kim (2023). Health to eat: A smart plate with food recognition, classification, and weight measurement for type-2 diabetic mellitus patients' nutrition control. *Sensors 23*(3), 1656.

Joutou, T. and K. Yanai (2009). A food image recognition system with multiple kernel learning. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 285–288. IEEE.

Kagaya, H., K. Aizawa, and M. Ogawa (2014). Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1085–1088.

Kapoor, A. and A. Singhal (2017). A comparative study of k-means, k-means++ and fuzzy c-means clustering algorithms. In *2017 3rd international conference on computational intelligence & communication technology (CICT)*, pp. 1–6. IEEE.

Kawano, Y. and K. Yanai (2013). Real-time mobile food recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–7.

Kawano, Y. and K. Yanai (2014). Foodcam-256: a large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 761–762.

Kawano, Y. and K. Yanai (2015). Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications 74*, 5263–5287.

Kelkar, S., S. Stella, C. Boushey, and M. Okos (2011). Developing novel 3d measurement techniques and prediction method for food density determination. *Procedia Food Science 1*, 483–491.

Killgore, W. D. and D. A. Yurgelun-Todd (2005). Body mass predicts orbitofrontal activity during visual presentations of high-calorie foods. *Neuroreport 16*(8), 859–863.

Kitamura, K., C. De Silva, T. Yamasaki, and K. Aizawa (2010). Image processing based approach to food balance analysis for personal food logging. In *2010 IEEE International Conference on Multimedia and Expo*, pp. 625–630. IEEE.

Kojima, R., O. Sugiyama, and K. Nakadai (2015). Audio-visual scene understanding utilizing text information for a cooking support robot. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4210–4215. IEEE.

Kong, F. and J. Tan (2012). Dietcam: Automatic dietary assessment with mobile camera phones. *Pervasive and Mobile Computing 8*(1), 147–163.

Kornilov, A. S. and I. V. Safonov (2018). An overview of watershed algorithm implementations in open source libraries. *Journal of Imaging 4*(10), 123.

Kozlovskaya, L., M. Abou-Kaoud, and D. Stepensky (2014). Quantitative analysis of drug delivery to the brain via nasal route. *Journal of controlled release 189*, 133–140.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60*(6), 84–90.

Kumar, V. A. et al. (2021). Survey on food recognition system using machine learning. *Smart Intelligent Computing and Communication Technology 38*, 134.

Le, Q. and T. Mikolov (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR.

LeCun, Y., Y. Bengio, G. Hinton, et al. (2015). Deep learning. nature, 521 (7553), 436-444. *Google Scholar Google Scholar Cross Ref Cross Ref*, 25.

Li, Y., J. Zhang, P. Gao, L. Jiang, and M. Chen (2018). Grab cut image segmentation based on image region. In *2018 IEEE 3rd international conference on image, vision and computing (ICIVC)*, pp. 311–315. IEEE.

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125.

Liu, C., Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma (2016). Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *Inclusive Smart Cities and Digital Health: 14th International Conference on Smart Homes and Health Telematics, ICOST 2016, Wuhan, China, May 25-27, 2016. Proceedings 14*, pp. 37–48. Springer.

Liu, C., Y. Cao, Y. Luo, G. Chen, V. Vokkarane, M. Yunsheng, S. Chen, and P. Hou (2017a). A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure. *IEEE Transactions on Services Computing 11*(2), 249–261.

Liu, C., Y. Cao, Y. Luo, G. Chen, V. Vokkarane, M. Yunsheng, S. Chen, and P. Hou (2017b). A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure. *IEEE Transactions on Services Computing 11*(2), 249–261.

Long, J., E. Shelhamer, and T. Darrell (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Lu, Y., Y. Huang, and R. Lu (2017). Innovative hyperspectral imaging-based techniques for quality evaluation of fruits and vegetables: A review. *Applied Sciences 7*(2), 189.

Lu, Y., T. Stathopoulou, M. F. Vasiloglou, S. Christodoulidis, Z. Stanga, and S. Mougiakakou (2020). An artificial intelligence-based system to assess nutrient intake for hospitalised patients. *IEEE transactions on multimedia*.

MacLean, R. R., A. Cowan, and J. A. Vernarelli (2018). More to gain: dietary energy density is related to smoking status in us adults. *BMC Public Health 18*(1), 1–7.

Maguire, G., H. Chen, R. Schnall, W. Xu, and M.-C. Huang (2021). Smoking cessation system for preemptive smoking detection. *IEEE Internet of Things Journal 9*(5), 3204–3214.

Maione, C., D. R. Nelson, and R. M. Barbosa (2019). Research on social data by means of cluster analysis. *Applied Computing and Informatics 15*(2), 153–162.

Mandal, B., N. B. Puhan, and A. Verma (2018). Deep convolutional generative adversarial network-based food recognition using partially labeled data. *IEEE Sensors Letters 3*(2), 1–4.

Mao, D., F. Li, Q. Ma, M. Dai, H. Zhang, L. Bai, and N. He (2019). Intraocular administration of tetramethylpyrazine hydrochloride to rats: a direct delivery pathway for brain targeting? *Drug Delivery 26*(1), 841–848.

Marin, J., A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba (2018). Recipe1m+: a dataset for learning cross-modal embeddings for cooking recipes and food images. *arXiv preprint arXiv:1810.06553*.

Matsuda, Y., H. Hoashi, and K. Yanai (2012). Recognition of multiple-food images by detecting candidate regions. In *2012 IEEE International Conference on Multimedia and Expo*, pp. 25–30. IEEE.

Matsuda, Y. and K. Yanai (2012). Multiple-food recognition considering co-occurrence employing manifold ranking. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 2017–2020. IEEE.

Merchant, K. and Y. Pande (2019). Convfood: a cnn-based food recognition mobile application for obese and diabetic patients. In *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, Volume 1*, pp. 493–502. Springer.

Merzougui, M. and A. El Allaoui (2019). Region growing segmentation optimized by evolutionary approach and maximum entropy. *Procedia Computer Science 151*, 1046–1051.

Min, W., S. Jiang, L. Liu, Y. Rui, and R. Jain (2019). A survey on food computing. *ACM Computing Surveys (CSUR) 52*(5), 1–36.

Min, W., Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang (2023). Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Morabia, A., F. Curtin, and M. S. Bernstein (1999). Effects of smoking and smoking cessation on dietary habits of a swiss urban population. *European journal of clinical nutrition 53*(3), 239–243.

Moussawi, A. E., N. B. Seghouani, and F. Bugiotti (2020). A graph partitioning algorithm for edge or vertex balance. In *Database and Expert Systems Applications: 31st International Conference, DEXA 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings, Part I 31*, pp. 23–37. Springer.

Munira Shifat, S., T. Parthib, S. Talukder Pyaasa, N. Maitra Chaity, N. Kumar, and M. Kishor Morol (2022). A real-time junk food recognition system based on machine learning. *arXiv e-prints*, arXiv–2203.

Murugaiyan, J. S., M. Palaniappan, T. Durairaj, and V. Muthukumar (2021). Fish species recognition using transfer learning techniques. *International Journal of Advances in Intelligent Informatics 7*(2), 188–197.

Muthukrishnan, R. and M. Radha (2011). Edge detection techniques for image segmentation. *International Journal of Computer Science & Information Technology 3*(6), 259.

Natesan, P., T. Selvan, M. Shrivarshini, B. Dhanya, and S. Kalaiselvi (2023). Prediction of healthy and unhealthy food items using deep learning. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 233–239. IEEE.

Nishida, C., R. Uauy, S. Kumanyika, and P. Shetty (2004). The joint who/fao expert consultation on diet, nutrition and the prevention of chronic diseases: process, product and policy implications. *Public health nutrition 7*(1a), 245–250.

Noh, H., S. Hong, and B. Han (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528.

Ofli, F., Y. Aytar, I. Weber, R. Al Hammouri, and A. Torralba (2017). Is saki# delicious? the food perception gap on instagram and its relation to health. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 509–518.

Ortis, A., P. Caponnetto, R. Polosa, S. Urso, and S. Battiato (2020). A report on smoking detection and quitting technologies. *International journal of environmental research and public health 17*(7), 2614.

Ortis, A., G. M. Farinella, and S. Battiato (2020). Survey on visual sentiment analysis. *IET Image Processing 14*(8), 1440–1456.

Pandey, P., A. Deepthi, B. Mandal, and N. B. Puhan (2017). Foodnet: Recognizing foods using ensemble of deep networks. *IEEE Signal Processing Letters 24*(12), 1758–1762.

Paszke, A., A. Chaurasia, S. Kim, and E. Culurciello (2016). Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

Pehlic, A., A. Abd Almisreb, M. Kunovac, E. Skopljak, and M. Begovic (2019). Deep transfer learning for food recognition. *Southeast Europe Journal of Soft Computing 8*(2).

Peng, C. and J. Ma (2020). Semantic segmentation using stride spatial pyramid pooling and dual attention decoder. *Pattern Recognition 107*, 107498.

Pfisterer, K. J., R. Amelard, A. G. Chung, B. Syrnyk, A. MacLean, and A. Wong (2019). Fully-automatic semantic segmentation for food intake tracking in long-term care homes. *arXiv preprint arXiv:1910.11250*.

Phetphoung, W., N. Kittimeteeworakul, and R. Waranusast (2014). Automatic sushi classification from images using color histograms and shape properties. In *2014 Third ICT International Student Project Conference (ICT-ISPC)*, pp. 83–86. IEEE.

Pisinger, C. and T. Jorgensen (2007). Waist circumference and weight following smoking cessation in a general population:: The inter99 study. *Preventive medicine 44*(4), 290–295.

Pokharkar, V., S. Suryawanshi, and V. Dhapte-Pawar (2020). Exploring micellar-based polymeric systems for effective nose-to-brain drug delivery as potential neurotherapeutics. *Drug Delivery and Translational Research 10*, 1019–1031.

Pouladzadeh, P., S. Shirmohammadi, and R. Al-Maghrabi (2014). Measuring calorie and nutrition from food image. *IEEE Transactions on Instrumentation and Measurement 63*(8), 1947–1956.

Pouladzadeh, P., S. Shirmohammadi, A. Bakirov, A. Bulut, and A. Yassine (2015). Cloud-based svm for food categorization. *Multimedia Tools and Applications 74*, 5243–5260.

Rahmat, R. A. and S. B. Kutty (2021). Malaysian food recognition using alexnet cnn and transfer learning. In *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 59–64. IEEE.

Rajesh, A. A., M. Raghu, and J. Sangeetha (2022). Fast food image recognition using transfer learning. In *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, pp. 1–10. IEEE.

Raju, V. B., M. H. Imtiaz, and E. Sazonov (2023). Food image segmentation using multi-modal imaging sensors with color and thermal data. *Sensors 23*(2), 560.

Razali, M. N., E. G. Moung, F. Yahya, C. J. Hou, R. Hanapi, R. Mohamed, and I. A. T. Hashem (2021). Indigenous food recognition model based on various convolutional neural network architectures for gastronomic tourism business analytics. *Information 12*(8), 322.

Ren, S., K. He, R. Girshick, and J. Sun (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems 28*.

Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer.

Sajadmanesh, S., S. Jafarzadeh, S. A. Ossia, H. R. Rabiee, H. Haddadi, Y. Mejova, M. Musolesi, E. D. Cristofaro, and G. Stringhini (2017). Kissing cuisines: Exploring worldwide culinary habits on the web. In *Proceedings of the 26th international conference on world wide web companion*, pp. 1013–1021.

Sapna, S., K. Yaswanth, P. Kumar, et al. (2023). Calorie estimation of food and beverages using deep learning. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 324–329. IEEE.

Sarkar, D., R. Bali, and T. Ghosh (2018). *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing Ltd.

Seferbekov, S., V. Iglovikov, A. Buslaev, and A. Shvets (2018). Feature pyramid network for multi-class land segmentation. 2018 ieee. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Volume 32, pp. 324–325.

Sharma, P., A. SHARMA, et al. (2022). Hybrid approach for food recognition using various filters. *International Journal of Advanced Computer Technology 11*(1), 1–5.

Sharma, U., B. Artacho, and A. Savakis (2021). Gourmetnet: Food segmentation using multi-scale waterfall features with spatial and channel attention. *Sensors 21*(22), 7504.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Snell, J., K. Swersky, and R. Zemel (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems 30*.

Stevens, E., L. Antiga, and T. Viehmann (2020). *Deep learning with PyTorch*. Manning Publications.

Subhi, M. A., S. H. Ali, and M. A. Mohammed (2019). Vision-based approaches for automatic food recognition and dietary assessment: A survey. *IEEE Access 7*, 35370–35381.

Sun, W. and R. Wang (2018). Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm. *IEEE Geoscience and Remote Sensing Letters 15*(3), 474–478.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Tahir, G. A. and C. K. Loo (2021a). A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment. In *Healthcare*, Volume 9, pp. 1676. Multidisciplinary Digital Publishing Institute.

Tahir, G. A. and C. K. Loo (2021b). Explainable deep learning ensemble for food image analysis on edge devices. *Computers in Biology and Medicine 139*, 104972.

Tai, T. T., D. N. H. Thanh, and N. Q. Hung (2022). A dish recognition framework using transfer learning. *IEEE Access 10*, 7793–7799.

Tasci, E. (2020). Voting combinations-based ensemble of fine-tuned convolutional neural networks for food image recognition. *Multimedia Tools and Applications 79*(41-42), 30397–30418.

Temdee, P. and S. Uttama (2017). Food recognition on smartphone using transfer learning of convolution neural network. In *2017 Global Wireless Summit (GWS)*, pp. 132–135. IEEE.

Teng, J., D. Zhang, D.-J. Lee, and Y. Chou (2019). Recognition of chinese food using convolutional neural network. *Multimedia Tools and Applications 78*, 11155–11172.

Termritthikun, C., P. Muneesawang, and S. Kanprachar (2017). Nu-innet: Thai food image recognition using convolutional neural networks on smartphone. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9*(2-6), 63–67.

Thoma, M. (2016). A survey of semantic segmentation. arxivpreprint. *arXiv preprint arXiv:1602.06541*.

TrayDataset (url). Trayfood segmentation dataset available on kaggle:. https://www.kaggle.com/datasets/thezaza102/tray-food-segmentation. accessed: 2023-04-02.

Vincent, L. and P. Soille (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence 13*(06), 583–598.

Vivek, M., N. Manju, and M. Vijay (2018). Machine learning based food recipe recommendation system. In *Proceedings of International Conference on Cognition and Recognition: ICCR 2016*, pp. 11–19. Springer.

Wang, X., D. Kumar, N. Thome, M. Cord, and F. Precioso (2015). Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE.

Xie, S., R. Girshick, P. Dollár, Z. Tu, and K. He (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.

Xie, W., S. Wei, Z. Zheng, Y. Jiang, and D. Yang (2021). Recognition of defective carrots based on deep learning and transfer learning. *Food and Bioprocess Technology 14*(7), 1361–1374.

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

Yang, S., M. Chen, D. Pomerleau, and R. Sukthankar (2010). Food recognition using statistics of pairwise local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2249–2256. IEEE.

Yu, Q., D. Mao, and J. Wang (2016). Deep learning based food recognition. *Technical report, Stanford University*.

Zahisham, Z., C. P. Lee, and K. M. Lim (2020). Food recognition with resnet-50. In *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1–5. IEEE.

Zhang, Y., Z. Qiu, T. Yao, D. Liu, and T. Mei (2018). Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6810–6818.

Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890.

Zhou, L., C. Zhang, F. Liu, Z. Qiu, and Y. He (2019). Application of deep learning in food: a review. *Comprehensive reviews in food science and food safety 18*(6), 1793–1811.

Zhou, Z., M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Springer.