Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2026-08

Machine-vision-based Bangladeshi street food recognition

Siza, Tasnin

Independent University, Bangladesh (IUB)

https://ar.iub.edu.bd/handle/11348/1578

Downloaded from IUB Academic Repository

Machine-Vision-Based Bangladeshi Street Food Recognition

August 2026

Prepared by

Tasnin Siza

ID:2130480

Shahida Yesmin Sumi

ID: 2110727

Mst Shahida Araby Tisha

ID: 2220789

Department of Computer Science and Engineering

Independent University, Bangladesh

Supervised By

Dr. Md. Tarek Habib

Associate Professor

Department of Computer Science and Engineering

Independent University, Bangladesh

Attestation

We are aware of the fact that plagiarism is strictly prohibited, and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards proper guidelines.

Author Name:

Tasnin Siza

Signature:

. . . . . .

Author Name:

Shahida Yesmin Sumi

Signature:

. . . . . .

Author Name:

Mst Shahida Araby Tisha

Signature:

. . . . . .

Evaluation Committee

Supervision Panel

.....

Supervisor
Dr. Md. Tarek Habib, CSE, IUB.

Examiners

.....

Internal Examiner 1
Md. Asif Bin Khaled, CSE, IUB.

.....

Internal Examiner 2
Dr. Asif Mahmood, CSE, IUB.

.....

External Examiner
Mr. Faisal Ahmed, Director, Digital Platforms & Engineering,
Banglalink Digital Communications Ltd

Office Use

.....

Dr. Saadia Binte Alam
Program Coordinator

.....

Dr. AKM Mahbubur Rahman
Head of the Department

.....

Prof. M Arshad Momen
Director, Graduate Studies, Research and Industry Relations,

.....

Md. Shahajada Masud Anowarul Haque
Librarian

Acknowledgement

We would like to thank Almighty God who enabled us to carry out this research successfully.

We are extremely grateful to our highly esteemed supervisor, Assistant Professor Md. Tarek Habib, PhD, Independent University, Bangladesh, for his on-going guidance, insightful inputs, constructive feedback and generous support throughout this project. This research was greatly influenced and completed by his knowledge, encouragement and care.

We also would like to express our gratitude to the Department of Computer Science & Engineering, especially to the Head of the Department, Dr. Saadia Binte Alam, for creating an enabling atmosphere in the college and for encouraging us during our undergraduate period. We are highly grateful to Assistant Professor Md. Sadekur Rahman, Daffodil International University for his valuable motivation, encouragement & guidance throughout the process of this work.

We would also like to thank the Fab Lab for their great help, resources and ongoing support. The positive atmosphere and the collaborative setup of Fab Lab was of a great help in developing and implementation of this project.

We are truly grateful to Shahriar Sultan Ramit for his guidance, motivation and support during the project. He gave us many suggestions and encouragement on our way to do the research and when facing with difficulties.

Lastly, we would like to say a special thank you to our friends and family for their ongoing encouragement, understanding and support. Their encouragement and motivation helped us to stay motivated during difficult times of this research. We would also like to express our gratitude to all the members of the PLP who gave of their time and expertise, either directly or indirectly, in making this research possible.

Letter of Transmittal

22 August 2026

To,

Dr. Md. Tarek Habib
Associate Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

Subject: Submission of Final Project Report

Dear Sir,

It is with great pleasure that we present our final project report titled “Machine-Vision-Based Bangladeshi Street Food Recognition” as a partial fulfillment of the requirements for the Bachelor of Science (B.Sc) in Computer Science and Engineering Degree from Independent University, Bangladesh.

In your guidance this research introduces a deep learning based solution to recognize 10 classes of street foods in Bangladesh. The study consists of data set preparation, comparison to pretrained CNN models and a two-stage MobileNetV2 transfer learning method to get better classification result. Explainable AI techniques were also applied to analyze the model's predictions.

Thanks for your precious guidance, feedback, encouragement and support to complete this project.

Sincerely,

Tasnin Siza , ID: 2130480
Shahida Yesmin Sumi , ID: 2110727
Mst Shahida Araby Tisha, ID: 222078

Abstract

Bangladeshi street food undoubtedly forms an important aspect of culinary culture, however, its automatic identification has been largely unstudied because of the lack of appropriate datasets and visual similarities among food categories that already exist. This paper offers a framework for recognizing Bangladeshi street food based on machine vision and deep learning. The contribution of the study lies in the fact that it offers a custom dataset of 4,000 original images and then augmented it to 16,000 images. Five pre-trained CNN models (InceptionV3, DenseNet121, MobileNetV2, NASNetMobile, Xception) were evaluated using the standard metrics. The baseline model of MobileNetV2 demonstrated the accuracy of 93.75%. Further research was based on two-stage transfer learning. The model obtained 99% accuracy, precision, recall, F_1-score , which is much more than that obtained by the MobileNetV2 model. Grad-CAM and LIME technologies were used for better interpretability of the learning model.

Table of Contents

Attestation ii
Evaluation Committee iii
Acknowledgement iv
Letter of Transmittal v
Abstract vi
Table of Contents vii
List of Figures vii
1 Introduction 1
1.1 Background 1
1.2 Problem Statement 2
1.3 Objectives of the Research 2
1.4 Significance of the Research 3
1.5 Scope of the Research 3
2 Literature Review 4
2.1 Prior Studies 4
2.2 Research Gap and Motivation 8
3 Computer Vision for Bangladeshi Street Food Recognition 9
3.1 Deep Learning 10
3.2 Convolutional Neural Networks (CNN): 10
3.4 ImageNet 12
4 Methodology 13
4.1 Proposed Workflow 13
Dataset preparation 14
Data Augmentation and Preprocessing 14
Training Configuration 15
Model Evaluation 15
4.2 Dataset Collection 15
4.2.1 Data Acquisition 15
4.2.2 Street Food Categories 16
4.2.3 Dataset Size 18
4.2.4 Real-World Image Variability 18
4.2.5 Challenges During Data Collection 19
4.3.1 Dataset Composition 20
4.3.2 Characteristics of the Dataset 20
4.3.3 Dataset Diversity and Generalization 20
4.3.4 Dataset Quality 21
4.3.5 Importance of the Dataset 21

4.4 Dataset Organization..... 21
4.4.1 Class-wise Folder Organization..... 22
4.4.2 Dataset Splitting..... 22
4.4.3 Label Assignment..... 23
4.4.4 Label Encoding..... 23
4.4.5 Advantages of the Dataset Organization..... 24
4.5 Image Augmentation..... 24
4.5.1 Purpose of Image Augmentation..... 25
4.5.2 Augmentation Techniques..... 25
4.5.3 Dataset Expansion..... 26
4.5.4 Representative Augmented Images..... 27
4.6 Image Preprocessing..... 29
4.6.1 Image Resizing..... 29
4.6.2 Pixel Value Normalization..... 30
4.6.3 Data Loading..... 30
4.6.4 Label Encoding..... 30
4.6.5 Batch Generation..... 31
4.6.6 Benefits of Image Preprocessing..... 31
4.7 Transfer Learning..... 31
4.7.1 Motivation for Using Transfer Learning..... 32
4.7.2 Transfer Learning Process..... 32
4.7.4 Advantages of Transfer Learning..... 34
4.8 MobileNetV2 Fine-Tuning..... 35
4.8.1 Proposed MobileNetV2 Architecture..... 35
4.8.2 Phase 1: Feature Extraction..... 36
4.8.3 Phase 2: Fine-Tuning..... 37
4.8.4 Training Strategy..... 38
4.8.5 Advantages of the Two-Stage Fine-Tuning Strategy..... 39
4.8.6 Performance of the Proposed Model..... 39
4.9 Other Compared Models..... 39
4.9.1 InceptionV3..... 40
4.9.3 DenseNet121..... 41
4.9.4 NASNetMobile..... 42
4.9.5 Xception..... 42
4.9.9 Comparative Evaluation..... 43
4.10 Performance Metrics..... 43
4.10.1 Confusion Matrix..... 44
4.10.2 Accuracy..... 44
4.10.3 Precision..... 44
4.10.4 Recall..... 45
4.10.5 F_1-score..... 45

4.10.6 Classification Report....45
4.10.7 Importance of Multiple Evaluation Metrics....46
5 System Implementation....47
5.1 System Overview....47
5.2 Development Environment....49
5.2.1 Hardware Configuration....49
5.2.2 Software Environment....49
5.2.3 Training Configuration....50
5.3 Model Implementation....51
5.3.1 Baseline CNN Model Implementation....52
5.3.2 Proposed Two-Stage MobileNetV2 Implementation....52
5.4 Model Inference and Classification....53
6 Performance Evaluation of the Compared CNN Models....55
6.1 Proposed Two-Stage Transfer Learning Model....63
6.1.1 Performance of the Proposed Two-Stage Transfer Learning Model....64
6.1.2 Training and Validation Curves....65
6.1.3 Confusion Matrix Analysis....67
6.2 Comparative Performance Analysis....69
6.3 Discussion of Experimental Results....70
7 Explainable AI Analysis....71
7.1 Gradient-weighted Class Activation Mapping (Grad-CAM)....71
7.2 Local Interpretable Model-Agnostic Explanations (LIME)....72
8 Comparative Analysis....74
8.1 Baseline Model Performance and MobileNetV2 Selection....74
8.2 Comparison of the Proposed Dataset with Existing Food Datasets....74
8.3 MobileNetV2 as the Best-Performing Model....75
8.3.1 Strengths....76
8.3.2 Weaknesses....76
9 Limitations....77
9.1 Dataset scope and diversity....77
9.2 Real world environmental variations....77
10 Conclusion & Future Work....78
10.1 Conclusion....78
10.2 Future Work....78
Bibliography....80

List of Figures

Figure 4.1: Overall Workflow of the Proposed Machine-Vision-Based Bangladeshi Street Food Recognition System 19

Figure 4.2: Sample Images of the Ten Bangladeshi Street Foods 24

Figure 4.3: Directory Structure of the Bangladeshi Street Food Dataset 30

Figure 4.5: Transfer Learning Workflow of the Bangladeshi Street Food Recognition Research 43

Figure 4.6: Proposed MobileNetV2 Custom Head Classification 46

Figure 5.1: Overall System Architecture of the Implemented Bangladeshi Street Food Recognition System 62

Figure 6.11: Training and Validation Loss and Accuracy Curves of the Proposed Two-Stage MobileNetV2 Model 85

Figure 6.12: Confusion Matrix of the Proposed Two-Stage MobileNetV2 Model on the Test Dataset 87

Figure 6.13: Comparative Test Accuracy of the Evaluated Transfer Learning Models 89

1 Introduction

Bangladeshi street food is an important part of the country's culinary culture, with popular foods such as Velpuri, Fuska, Jhalmuri, Chotpoti, Boot Makha, Pakora, and Papor Vaja being widely consumed across different regions. These foods differ in their ingredients, texture, shape, color, and presentation, while several categories also share similar visual characteristics, making their recognition challenging from images. Variations in lighting, viewpoint, background, serving style, and preparation further increase the difficulty of automated classification. Traditionally, food identification relies primarily on human visual recognition, which can be subjective and time-consuming. Recent advances in artificial intelligence, computer vision, and deep learning provide new opportunities for developing automated food recognition systems capable of learning visual characteristics directly from images.

Although deep learning has been successfully applied to general food image classification, comparatively limited research has focused specifically on Bangladeshi street foods. The absence of a dedicated benchmark dataset and the visual similarity among local food categories create additional challenges for developing reliable recognition systems. Therefore, this research develops a dedicated Bangladeshi street food dataset and investigates multiple pretrained CNN architectures using transfer learning, followed by a two-stage MobileNetV2 fine-tuning approach to improve classification performance.

1.1 Background

Food is an important component of cultural identity, and Bangladesh has a diverse street food culture consisting of numerous traditional and locally popular foods. Street foods such as Fuska, Velpuri, Jhalmuri, Chotpoti, Pakora, and Papor Vaja are widely consumed across different regions of the country. Although these foods are visually recognizable to people familiar with them, automatic recognition is challenging because their appearance can vary considerably depending on preparation method, ingredients, serving style, lighting conditions, camera viewpoint, and surrounding environment.

Recent advances in computer vision and deep learning have significantly improved the ability of machines to recognize and classify objects from images. Convolutional neural networks (CNNs), in particular, have demonstrated strong performance in image classification because they can automatically learn hierarchical visual representations from image data. However, training deep CNNs from scratch generally requires large, diverse, and carefully labeled datasets. For specialized domains such as Bangladeshi street food, large publicly available datasets are limited or unavailable.

Transfer learning provides an effective solution to this problem. Instead of learning all visual features from the beginning, a model pretrained on a large-scale dataset such as ImageNet can be adapted to a new classification task. The pretrained network already contains general visual representations that can be further specialized for the target domain. This approach can reduce training requirements while achieving strong classification performance.

In this research, a dedicated Bangladeshi street food image dataset was constructed to address the lack of a domain-specific dataset. The dataset initially contained 4,000 images across ten food categories. Image augmentation was subsequently used to increase the dataset to 16,000 images and introduce additional variation in image orientation, scale, position, and appearance. This expanded dataset was used to investigate the effectiveness of different pretrained CNN architectures for street food recognition.

1.2 Problem Statement

Although there have been technical progress in food identification using deep learning techniques, research into Bangladeshi street food remains limited. Datasets available are insufficient for representing local food classes and classification is made difficult because of similarity in their colors, textures, ingredients, and presentation. Furthermore, it is vital to choose the model that has a good trade-off concerning accuracy and computational efficiency. As a result, the study proposes a dedicated dataset, tests multiple pretrained convolutional neural networks, and suggests a two-step MobileNetV2 transfer learning framework to enhance the recognition performance.

1.3 Objectives of the Research

The primary goal of this work is to create an effective deep learning system which can recognize street food images in Bangladesh.

Specific objectives are to:

1. Create a personalized image database of 10 categories of food items in street food from Bangladesh.
2. Use the right image augmentations to expand data sets
3. Pre-processing and transfer learning methods for automatic street food classification.
4. Evaluate different pre-trained CNNs.
5. Design an effective architecture with respect to classification accuracy and computational efficiency.
6. Design a two-stage MobileNetV2 transfer learning model using selective fine-tuning.

7. Check the proposed model using standard criteria and visualization methods for classification.
8. Interpret the predictions of the model by using Explainable AI techniques such as LIME and Grad-CAM.

1.4 Significance of the Research

The research improves the process of developing a computer vision solution. The created data set provides images relevant to the local context and can become a basis for research in the area of identifying foods and related to Machine Learning technology.

The offered two-stage model based on MobileNetV2 shows that a Light CNN can effectively classify objects with the help of transfer learning and fine-tuning techniques. This is especially useful for applications where resources may be limited. Besides, the use of Explainable AI gives more details on what visual features were used for the classification process.

1.5 Scope of the Research

The research emphasizes image classification in ten popular types of street food in Bangladesh. The methodology applies food images in RGB color space and uses deep learning approaches to identify those foods. Specifically, the study evaluates several pretrained CNN models, and also enhances MobileNetV2 through the implementation of a two-stage finetuning process.

This study focuses on identifying individual food classes, instead of determining nutrient contents, recognizing multiple foods co-existing, or doing live video recognition tasks. Thus, the proposed framework enables implementation of those tasks later in the future.

2 Literature Review

Machine vision technology and deep learning technique have advanced considerably in terms of food identification while gaining popularity in the area of automatic categorization of food types. Early methods were based primarily on manually measured characteristics of food items like color, structure, shape etc., that are susceptible to changes in illumination, background and presentation. In order to overcome the drawbacks of early techniques, more recent studies have shown a growing tendency to use methods based on deep learning technology, especially Convolutional Neural Networks (CNNs), transfer learning and lightweight systems to achieve successful food identification.

2.1 Prior Studies

Using transfer learning, pre-trained CNNs will be evaluated as a way of finding a fast, accurate, and practical system for deployment [1].Researchers have used both Lightweight architectures like MobileNetV2 and MobileNetV3, and have achieved very good accuracies up to 93-95% given the relatively low computational requirements; hence are well adapted to real-time applications and mobile use cases [1] [2] [3]. Also, due to the variability of food within a class and the similarity of many foods visually, it can be difficult to classify foods accurately [5].Recent review articles state that CNN is the architecture that has remained predominant in food recognition due to its capability of automatically extracting hierarchical image features. However, recent architectures have shown further advancements in visual food category recognition, such as Vision Transformers (ViTs), CNN-Transformer hybrids, attention models, and ensemble learning methods [6]. In order to improve the performance of the models while also reducing the cost of training the models, research-ers have used transfer learning with pre-trained models including MobileNet, ResNet, and DenseNet. Researchers have begun collecting Large Scale data sets to improve generalization of the models for example, Food-11, Food-101, Chinese FoodNet. However, many of the cases have a limited cultural diversity of food and under represent foods that are specific to certain regions [6].Studies focusing on food safety also highlight the public health importance of street food. In developing countries, lack of personal hygiene, food handling practices, and correct regulations still play their role in causing food-borne diseases [7].Developing countries, particularly those with street food, present further challenges when it comes to developing a machine vision system. Street food is prepared informally, presents food in various ways, and typically has poor hygiene, all of which add to the complexity of street food's visual qualities and public health risk [7] [9]. Research done on the knowledge, attitudes and practices of street food vendors shows that while many street food vendors have basic food safety knowledge, there is a lot of deficiency in food safety practices

and sanitation infrastructure [9]. There has recently been significant effort to create localized databases of food information. Examples are Indonesian and African food data sets to provide culturally relevant data used to improve the accuracy of recognition [8] [10] .In addition to food classification, the application of machine vision technology has been a key technology in food quality inspection and safety assurance. Advanced imaging technologies like hyperspectral imaging, thermal imaging, X-ray imaging, magnetic resonance imaging (MRI), RGB-D cameras, and stereo vision systems have been combined with deep learning algorithms to evaluate food quality, estimate food volume, detect bruises, identify contaminants, and assess food freshness [11]. As a result, there is no comprehensive way of developing a fast and accurate machine vision system for culturally appropriate street foods. The objective of this study is to develop a Bangladesh street food recognition system using real world datasets collected from cooks who prepared food in a variety of environ-ments. nutritional interventions and improved public health monitoring. Traditional methods of collecting dietary data are limited by their reliance on manual food records, where users visually assess the components of their meals and then estimate the nutrition associated with each meal based on this visual assessment, a process that can introduce considerable error. Classifying food images, in general, has been a strong suit for Convolutional Neural Networks (CNNs), which are a form of deep neural networks commonly used for image classifi-cation. CNNs have also performed well with many other complex dataset types [12] .A few works have investigated how to detect objects and to recognize multiple food items in one image for realistic dining scenarios where multiple food items are present in a single image. Pouladzadeh and Shirmohammadi came up with a mobile food recognition system assisted by cloud computing that utilizes selective search, CNN feature extraction, and support vector machine classification to recognize multiple foods simultaneously with an overall accuracy of over 94% [13].The mobile food recognition is playing an increasingly important role in research due to the accessibility of mobile devices for dietary monitoring and nutrition management. To enable automatic dietary assessment using smartphone apps, NutriNet developed a deep CNN model that was trained on more than 225,000 images of food and beverages [14]. FoodAI also successfully deployed transfer learning to the Healthy 365 app that handles around 400,000 images from 756 food categories in Singapore [15]. In spite of the advances outlined above, perfor-mance in real world implementation is still negatively affected by changes in lighting, occluded food, and issues surrounding image quality [15]. Most food recognition systems currently available have not been trained on the conditions found at street food outlets and rely on controlled datasets, which are either expensive or difficult to obtain in low resource environments such as Bangladeshi street food [16]. In addition, the methods of calorie estimation currently available are not always practical in natural environments, and require controlled imaging environments and special hardware or fixed assumptions about the portion sizes [16]. In recent years, the object detection methods are more developed, using the YOLO model to make the food object detected in real time. YOLOv3 was successfully used in junk food detection, achieving a high accuracy rate of around 98% and real-time processing speed [17]. Similarly, a YOLOv4 model was developed to recognize food items in fast food restaurants with a mAP of

93.29% and to provide an estimation of calorie information and record the food intake using an Android application [18].

The development of machine vision has also been seen to be a fast one in automated fish species recognition for the conservation of biodiversity and the management of fisheries. CNN-based architectures such as ResNet, EfficientNet, and hybrid deep learning models have also demonstrated high recognition accuracy for freshwater fish species; and transfer learning and object detection based on YOLO have also demonstrated the possibility of recognizing the species efficiently and in real time, with lower computational costs. Moreover, models based on the vision transformer (ViT) and attention-based models have been more effective in the task of differentiating visually similar fish species under challenging environmental conditions. Most of the current systems are based on a relatively small or controlled dataset, and are not often easily transferable to real-world environments for freshwater due to domain shift, class imbalance, water turbidity, occlusion, and differences in illumination. These studies highlight the need for creating larger, representative datasets as well as lightweight deep learning architectures that are deployable on mobile/edge devices, which are also pertinent to food recognition systems [22].The importance of street food recognition for cultural tourism, nutrition and public health has generated several studies focusing on this topic. The mobile assistant application developed for pedestrians who consume street food in Bandung used the CNN image classification method to identify street foods with local names, and then displayed the nutritional information of street food using an Android application [23]. The developments show that mobile and edge computing is gaining in significance for real-world food recognition applications.

The study results showed a high accuracy (around 80%), but highlighted the importance of increasing the number of samples and the accuracy of the recognition method for the identification of similar street foods found in the pictures [23].There are numerous CNN architectures (DenseNet121, Incep-tionV3, MobileNetV2, and Xception) used in food imaging applications and many of them have shown signif-icant performance on multiclass item detection..Culturally-specific datasets have also been developed to recognize Indonesian traditional food [27], African food and Indian food and to create indigenous food datasets for tourism applications to reduce the geographical bias found in the existing food recognition systems.FoodCam developed an on-device food recognition system that can be used to recognize food without relying on cloud processing which alleviates communication latency and privacy concerns of the users [28].

Edge-computing-based food recognition systems were also developed, such as a combination of mobile and cloud computing, realizing both low response time and high recognition accuracy with low energy consumption, more recently [29].

Similarly, the ensemble CNN models created for visually impaired users yielded the recognition rates over 95% which showed that the combination of various deep learning models was proved to be effective [30], [31].Multi-level Attention Feature Fusion Network (MAF-Net) is an example that integrates multi-level CNN features, self-attention mechanism and subclass center

loss for better than 90% Top-1 accuracy on multiple benchmark datasets and better fine-grained food recognition [32], [33].

Deep Learning, specifically CNN has made significant strides in recognition of food images and has been able to improve on traditional methods of machine learning by automatically extracting hierarchical visual features [5] [12]. Currently, there are also New Models including YOLO and Mask R-CNN which provide multi food detection and segmentation functionality for possible applications in estimating calories and moni-toring dietary intake [4] [16] [17] [18].

In order to boost the robustness of the models, researchers often use image rotation, image flipping, brightness adjustment, zooming, translation, noise injection, and geometric transformations. These methods have the ability to diversify the dataset, decrease over-fitting and boost classification performance, especially if training data is scarce [3], [5], [24]. The results of survey studies and have always shown that the use of the transfer learning technique along with the use of effective data augmentation techniques has been proven to yield better results than training CNN models from scratch.[5],[12],[24].

The above studies illustrate the potential of object detection frameworks in real-world applications of dietary assessment systems.

Food analysis using CNN based machine vision systems is always better than conventional inspection methods in terms of objective, automatic, and non-destructive analysis of food [5], [11], [24].

Despite great advances in food recognition research, there are some issues that have yet to be resolved. Due to high intra-class variation and inter-class similarity, occlusion, different illumination, complex background and varying food presentation, it becomes difficult to make an accurate classification of food images [5], [6], [12], [24].

These data sets also are very skewed towards common international dishes, and there is a significant lack of regional and traditional dishes [8], [10], [19], [27].

Therefore, researchers are interested in creating more culturally diverse datasets, creating lightweight deep learning architectures, creating multimodal learning frameworks that incorporate recipes and ingredient information, few-shot learning approaches, and transformer-based models that can recognize more recipes with greater accuracy and efficiency of deployment [6], [24], [26].

In the broader context, the literature highlights that deep learning has made substantial progress in facilitating automatic food recognition, with CNN, transfer learning, attention mechanism, object detection models showing strong performance compared to traditional machine learning models.

2.2 Research Gap and Motivation

As shown in the reviewed studies, the use of deep learning and transfer learning for food image recognition is effective. But there are some gaps in the literature. The existing food recognition databases are mostly on the global food or on the widely studied food categories and Bangladesh traditional, regional foods have been given less attention. The absence of dedicated datasets makes it challenging to build and test recognition systems with the visual features of the Bangladeshi street foods.

Another significant challenge is the visual similarity of food categories. The color, texture, ingredients, shape and presentation of street food in Bangladesh can be similar. Depending on the light source, perspective, background and presentation of the food, the classification accuracy can change again. Previous research has also demonstrated intra-class variation and inter-class similarity to affect food recognition in the real world. The choice of an appropriate CNN architecture is also important. Deeper, more powerful networks can offer good feature extraction capabilities, but they're not necessarily appropriate for applications with limited resources. Alternatively, architectures like MobileNetV2 provide a balance between computational efficiency and classification ability, which are lightweight architectures. Our previous work has shown the effectiveness of MobileNetV2 for food recognition, leading to its attention for food classification in the region.

Considering these gaps, this research is driven by the desire for a specific Bangladeshi street food dataset and an accurate and efficient recognition system which can be computed.

3 Computer Vision for Bangladeshi Street Food Recognition

In this research, the basic technology to recognize street food in Bangladesh is computer vision. The proposed system uses computer vision techniques to analyze the images of the food products and classify the food into predefined categories without manual feature extraction. This method could serve as an effective solution for street food recognition in real-world scenarios and facilitate automated dietary assessment, which alleviates some difficulties arising from the manual recording of food.

The research targets Bangladeshi street food, as there are limited food recognition datasets for street food items in Bangladesh. To overcome this gap, a Bangladeshi Street Food Dataset (BSFD) has been created, which includes ten popular street food items: Badam, Boot Makha, Chanachur Makha, Chotpoti, Fuska, Hawai Mithai, Jhalmuri, Pakora, Papor Vaja, and Velpuri. Images were taken from different locations and subjects under varying environmental conditions such as lighting, angle of view, background and food presentation. This diversity allows the proposed recognition system to learn a representative visual features of each food category.

In the proposed framework, the images collected are processed using computer vision techniques prior to providing them to the deep learning models. Images are preprocessed and augmented for quality and diversity of the data set. These processed images are then fed to the pretrained convolutional neural network (CNN) models for food classification. To select the best model for Bangladeshi street food recognition, five CNN models: InceptionV3, DenseNet121, MobileNetV2, NASNetMobile, and Xception are analyzed.

The research results are shown on the experiments, which prove that the food recognition using computer vision can be used even if the images contain complex background and environmental changes. The proposed framework is able to successfully learn discriminative visual features needed to accurately classify images through the use of image pre-processing, image augmentation, transfer learning, and fine-tuning. MobileNetV2 model was found to be the best model to assess and it was optimized using a two-stage approach to fine-tuning.

In general, computer vision is a basis of the proposed street food recognition system in Bangladesh. It helps in automatically analysing the street food images, helps in efficient feature

extraction with the help of deep learning models and helps as the foundation to recognize the street food categories accurately in the real world when there is street food in Bangladesh.

3.1 Deep Learning

In this research, deep learning is the most common method to develop the proposed Bangladeshi street food recognition system. It can automatically extract representative visual features from the street food images of Bangladesh and classify them into the pre-defined food classes. The proposed system directly learns features from these images collected, and do not require manually designed feature extraction methods, which makes the approach effective to recognize the street food automatically.

The study uses deep learning algorithms to identify the images of street food in Bangladesh captured in real-world, unconstrained settings. In these images, there are differences in lighting, background, camera angle and food arrangement which makes classification more difficult. To face these challenges, some of the popular and pre-trained Convolutional Neural Network (CNN) Architecture were used to check which one is well suited for recognition of Bangladeshi street foods. The models evaluated are: InceptionV3, DenseNet121, MobileNetV2, NASNetMobile and Xception.

In addition, transfer learning is performed with pretrained weights from ImageNet and image augmentation to boost learning performance. The augmented dataset enables the deep learning models to learn more representative visual features and improves their ability to classify previously unseen street food images. In addition, the study introduces a two-stage fine-tuning method for the top-performing model, MobileNetV2, using a pretrained model to make it more adaptable to the Bangladeshi Street Food Dataset.

The experimental results show that deep learning is an effective solution for recognition of street food in Bangladesh. In the different CNN architectures evaluated, MobileNetV2 showed the best baseline performance, and following fine-tuning, the proposed architecture achieved an overall classification accuracy of 99%. This study reveals that deep learning can learn discriminative visual features for accurate recognition of the images of Bangladeshi street foods under various real-world settings.

In general, deep learning is used as the basis of the proposed recognition system. With transfer learning, image augmentation and fine-tuning, the proposed system can correctly and consistently classify images of street food in Bangladesh.

3.2 Convolutional Neural Networks (CNN):

In this research, the mostly adopted deep learning models for automatic recognition of Bangladeshi street food images are Convolutional Neural Networks (CNNs). The proposed recognition system is based on CNN transfer learning models to automatically extract the visual

features of food images and classify the food images into a predefined street food category. Conventional image classification methods rely on manually engineered features to represent the images, but CNNs learn the representative visual patterns directly from the training images. This will allow the models to understand complex features like colour, texture, shape and structural appearance which differentiate one category of food from another.

The CNN is chosen for its ability in image classification and food recognition task. The images of Bangladeshi street foods have been captured in real world scenarios, which introduces a lot of variations in terms of lighting, camera angle, background, object positioning, and presentation. The experimental comparison revealed that the performance of each CNN architecture in terms of feature extraction and classification were different. The research also suggests that certain food categories, such as Fuska and Velpuri, were challenging to classify as the food was visually similar in terms of color, texture, ingredients, and presentation style. However, the CNN models learned representative visual features for most food categories and augmentation of images significantly helped to boost their overall classification accuracy. The confusion matrices in the experimental results exemplify the fact that most of the predictions were correctly classified, while there were a few predictions classified incorrectly between classes that were visually similar.

In conclusion, Convolutional Neural Networks are the backbone of the proposed Bangladeshi street food recognition system for feature extraction and classification. They have the capability to automatically learn hierarchical visual representations from images, thus achieving accurate category recognition for a wide range of categories of street food collected under real-world conditions.

3.3 Fine-Tuning

The fine tuning is an essential part of the suggested Bangladeshi street food recognition system, which further enhances the accuracy of the pretrained CNN model after transfer learning. The fine-tuning technique was implemented on the best-performing CNN architecture (MobileNetV2) from the models tested in this research. The goal of the fine-tuning process was to fine-tune the pretrained features from ImageNet more effectively for the specific nature of images of street foods in Bangladesh, resulting in higher classification accuracy and less inter-class confusion between similar street food images.

The proposed fine-tuning framework is based on a two-step approach.The experimental results demonstrated in the study showed that the proposed fine-tuning strategy improved the overall classification performance to a great extent. The fine-tuned model outperformed the baseline model (MobileNetV2) by a significant margin and had fewer cross-categorization misclassifications. The training and validation curves also show a good stability of the model convergence, without excessive overfitting, which demonstrates the efficacy of the two-step optimization process in improving the model learning capacity.

Overall, fine-tuning plays a crucial role in the proposed Bangladeshi street food recognition system. The proposed framework is capable of transferring the generalized visual information

into the target dataset with a customized classification head and an optimized pre-trained MobileNetV2 network. This allows the model to attain the best classification results, which is one of the main contributions of this research.

3.4 ImageNet

In this research, all convolutional neural network (CNN) models make use of the pretrained knowledge learned from ImageNet. To avoid random initialization of the networks, the proposed Bangladeshi street food recognition system uses models pre-trained on the ImageNet data set. These learned weights are those that are pre-trained on the models, instead of receiving them, they will be used for figuring out common image properties before being fine-tuned for the Bangladeshi street food classification task. This approach is the basis for the transfer learning strategy that was used in this study.

The five CNN architectures that were tested are the InceptionV3, DenseNet121, MobileNetV2, NASNetMobile and Xception, all pretrained with ImageNet weights using the TensorFlow Keras Applications library. The proposed framework leverages pretrained models that provide generic feature extraction without the need to train the networks with the image of basic visual patterns. As a result, the models can focus on learning the unique characteristics of Bangladeshi street food images during the training process.

Prior to training, all the images were pre-processed to match the format of the pre-trained ImageNet models. The images were scaled to the input size needed for each CNN architecture, and then normalized before fed into the CNN. This pre-processing step ensured that the input data was formatted in the same way as the pretrained models and also made the feature extraction process efficient during training and testing.

The experimental results have shown that the contribution of the pretrained weights of the ImageNet for the CNN models has played a crucial role in the overall recognition performance. The pre-trained feature representations helped the models to learn the features of Bangladeshi street foods with less training.

All in all, ImageNet is a crucial resource assisting the proposed street food recognition system for Bangladesh, providing feature representations for all the CNN models tested. The application of ImageNet weights helps to achieve efficient transfer learning, to get the model converged quickly and thus to the high classification performance of the proposed recognition system.

4 Methodology

The methodology employed for creating the suggested machine vision-based recognition system for street food in Bangladesh is introduced in this chapter. Moreover, the process of dataset accumulation and preparation is explained with reference to image augmentation and preprocessing methods, techniques of transfer learning, and systems architecture, as well as the MobileNetV2 two-stage tuning method used in this research. The experimental approach is developed to make sure that data are prepared according to the same principles to guarantee an unbiased testing of the analysed deep-learning methods.

4.1 Proposed Workflow

The method that was used in this research has been developed for reaching successful and fast results in determining images of Bangladeshi street food. The method includes the numerous steps starting with preparing data and finishing with evaluation of the output of the trained model . The key components help maintaining the quality of data, increasing the generalizability of the model and ensuring successful classification outcomes when working with real data . The flow chart of the method is presented in figure 4.1.

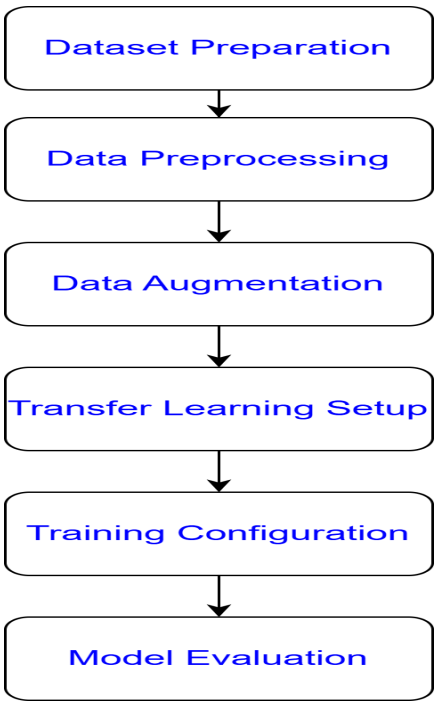


Figure 4.1: Overall Workflow of the Proposed Machine-Vision-Based Bangladeshi Street Food Recognition System.

The process consists of six main steps including preparation of the dataset, data preprocessing, augmentation of data, transfer learning setup, training configuration, and performance evaluation. After obtaining the evaluation results of the model, it was subject to the next stage involving Explainable AI (XAI) technologies such as LIME and Grad-CAM aimed to explain the results of the classification. The stage of XAI is discussed apart from the evaluation stage since its aim is different from the objectives of the training process. Hence, it can be said that every stage contributes to making the system more efficient. The details of each stage are discussed below,

Dataset preparation

The first step in the development of the model consisted of the creation of the dataset of images of Bangladeshi street food. Due to the absence of any benchmark dataset publicly available in this area, images were collected in the real world with the help of smartphone cameras. Moreover, some additional authentic images were downloaded from popular social media platforms like Facebook and Instagram to ensure a larger variety of the dataset. The images that were collected are representative of ten popular Bangladeshi street food types that have been taken in various lighting conditions, angles, backgrounds, and serving methods, which allowed the model to learn features that are very similar to the way it is used in real life.

Data Augmentation and Preprocessing

Data was then augmented to increase the variety of data used for training the model and mitigate overfitting using image augmentation techniques after data collection. The augmentation process created several variations of each of the original images by horizontally flipping, rotating, zooming, shifting the width, shifting the height, and shearing the original images. This resulted in the number of images in the dataset, 16,000 images.

All images were preprocessed prior to the training of the model after augmentation. All images were resized to 224 × 224 pixels to comply with the input size of the chosen CNN models. All pixel values were normalised to the range [-1, 1] and the labels were also one-hot encoded. 80% of the dataset was used for training, 10% was used for validation and the remaining 10% was used for testing.

Transfer Learning Using Pre-trained CNN Models

The dataset that underwent preprocessing was applied to train a variety of advanced CNN models via transfer learning. Pretrained weights learned from the ImageNet dataset were used in place of creating deep networks from scratch in order to increase speed of convergence and performance on the classification tasks. The selection of CNN architectures was based on their popularity, the presence of different architectural designs and complexity levels, as well as their

ability to extract features. The chosen CNN architectures were InceptionV3, DenseNet121, MobileNetV2, NASNetMobile, and Xception.

Training Configuration

Once the framework for transfer learning was created, the training began to ensure that no model except for the one being evaluated had received any prior training. The dataset was split into parts, where 80% was allocated for training, while both validation and testing received 10% each. Each model was trained with the use of mini-batch gradient descent with the batch size of 32. The input image underwent resizing and normalization before being fed into the network. The training was provided by the Adam optimizer, with categorical cross-entropy being implemented as the loss function. To stabilize the process of training and prevent overfitting, such callbacks techniques as ModelCheckpoint, EarlyStopping, and ReduceLROnPlateau were used. This way of doing things allowed the models to converge efficiently.

Model Evaluation

Standard classification metrics were utilized to evaluate the performance of the trained models. The model's performance was measured in terms of accuracy, precision, recall, and F_1-score, involved during the evaluation of its performance, which made the evaluation very comprehensive. Also, confusion matrices were generated to evaluate the precision of predictions per each class and check whether there are some remaining misclassified classes based on the similarity between different food categorizations. The curves showing training/validation accuracy and corresponding loss were analyzed so that to check the convergence processes and detect possible instances of overfitting during the training before the model deployment.

In conclusion, this proposed workflow combines systematic data preparation, image pre-processing/augmentation, transfer learning, model training, performance evaluation, and Explainable Artificial Intelligence in an effort to develop a single recognition system.

4.2 Dataset Collection

The performance of image classification systems based on deep learning is heavily dependent on the quality, diversity and representativeness of the training set. As there is no image dataset for Bangladeshi street food recognition, a custom image dataset was created for this research. The aim was to produce a good variety of real and authentic street food images to accurately depict the look of popular street foods from Bangladesh in the real world.

4.2.1 Data Acquisition

The data set was mainly assembled through the use of camera of the devotees of different street food stalls in diverse public areas of Bangladesh. The photos were taken in a mostly horizontal format, which is the most common format for mobile photos. Some images were also taken on the vertical axis. To further expand the dataset, however, a handful of images were also obtained from public social media outlets, such as Facebook and Instagram. Images were only used for the dataset if they clearly represented the targeted food categories.

The images in this study were deliberately taken in natural settings, unlike many other food recognition datasets which are collected in controlled laboratory environments. It allows the trained model to gain strong visual representations more representative of real deployment conditions, enhancing it for better generalization to unseen data.

4.2.2 Street Food Categories

The dataset includes ten street food types that are frequently found in different regions in Bangladesh. These foods were chosen due to their relevance to the local culture, popularity among the general population, and visual variety. Each food type includes certain number of components, features, colors, and its own style.

The selected foods are represented in Table 4.1.

Table 4.1: Bangladeshi Street Food Categories

Class Label	**Street Food**
C1	Velpuri
C2	Papor Vaja
C3	Fuska
C4	Jhalmuri
C5	Boot Makha
C6	Chanachur Makha
C7	Chotpoti
C8	Pakora
C9	Hawai Mithai
C10	Badam

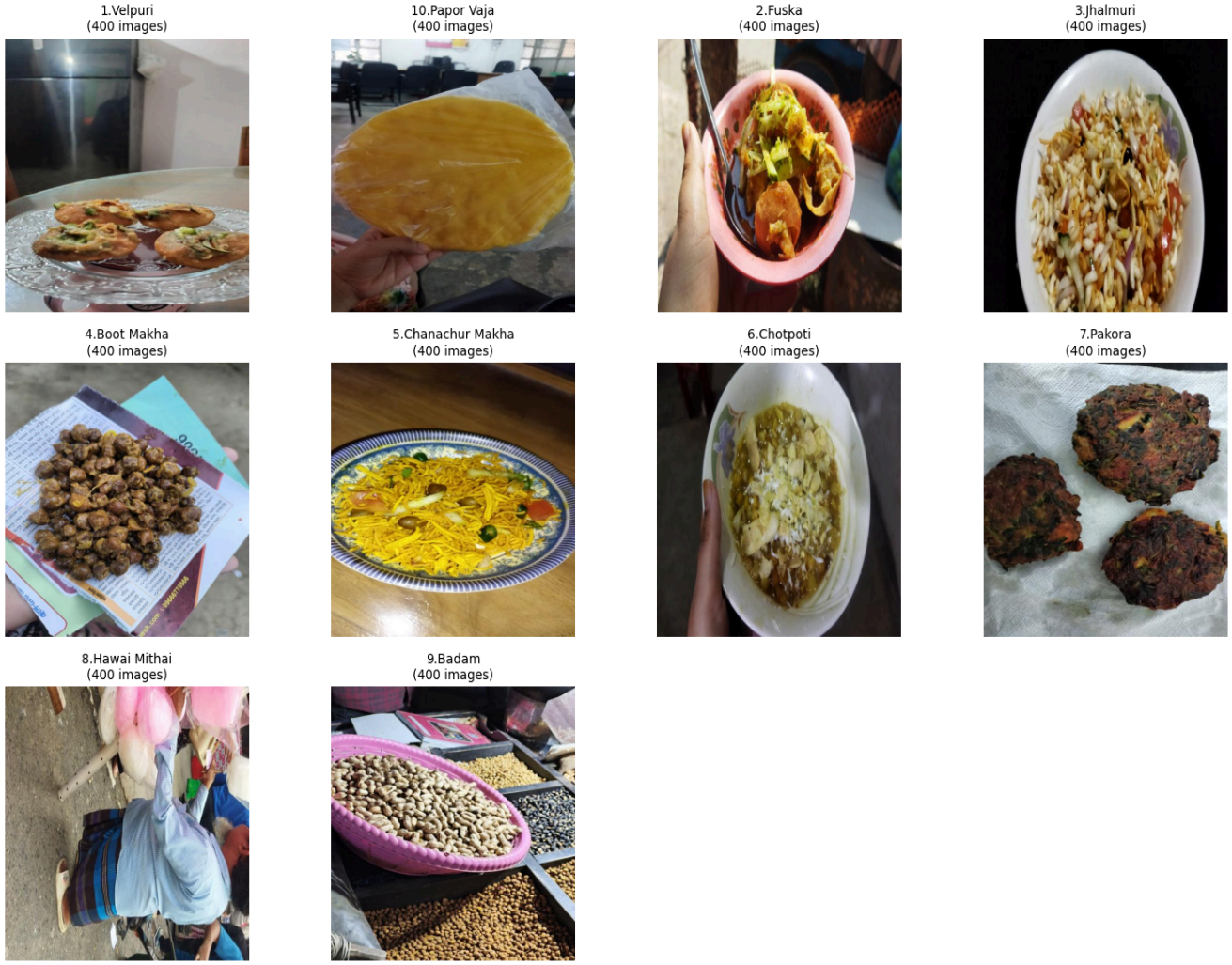


Figure 4.2: Sample images of the ten Bangladeshi streetfood.

Figure 4.2 shows pictures of various types of street food included in the database. These images enable to analyze varieties in visual appearance, creation, presentation techniques, lighting conditions, and background environments.

In some cases, these street foods have similar appearance. For example, “velpuri” and “fuska” as well as “chotpoti” and “boot makha” include similar components and garnishing styles, which can lead to difficulties in distinguishing them only by their looks.

4.2.3 Dataset Size

In the beginning, 400 original images were amassed for each category of the ten street food types, leading to a total of 4,000 images. In order to increase the size and quality of the training data, the dataset was augmented, leading to a total of 1,600 images for each category and yielding an overall dataset of 16,000 images.

The distribution of images for each category before and after the augmentation process is displayed in Table 4.2.

Table 4.2: Dataset Distribution Before and After Augmentation.

Class Label	Street Food Type	Number Of Images Before Augmentation	Number Of Images After Augmentation
C1	Velpuri	400	1600
C2	Fuska	400	1600
C3	Jhalmuri	400	1600
C4	Boot Makha	400	1600
C5	Chanachur Makha	400	1600
C6	Chotpoti	400	1600
C7	Pakora	400	1600
C8	Hawai Mithai	400	1600
C9	Badam	400	1600
C10	Papor Vaja	400	1600
Total		4,000	16,000

Maintaining an equal number of images for every class helps reduce class imbalance and enables the models to learn representative features from all food categories equally, leading to a fair and reliable evaluation.

4.2.4 Real-World Image Variability

The main goal of the dataset gathering process was to capture the variations that occur in the actual street food environment. Contrary to controlled picture datasets, the collected images contain huge differences regarding environmental and photographic conditions.
The dataset comprises images taken in:

- Different types of light, such as natural daylight, evening time, and artificial lighting.
- Varied angles and views.

- Different kinds of backgrounds including street-side stalls, outdoor markets, and crowded streets.
- Differences in the scale of images and depth.
- Differences in the serving dishes, pin plates, newspapers, and disposable container.
- Different styles of garnishing and composition of ingredients.

These differences cause extra complexity in the classification problem but allow for the creation of efficient representations which develop algorithms that are less sensitive to any changes in the environment.

4.2.5 Challenges During Data Collection

Several practical problems were faced while gathering the complete data of street foods from Bangladesh. Street food is usually made and eaten outdoors, often under unsteady lighting conditions, in shade, when some parts of the image are obscured, or when there is background noise.

One problem that occurred was the need to keep a decent mix in each class. This same food can be presented in many different ways and may have different ingredients and preparation styles depending on the vendor. Hence, images were taken from several locations and from various vendors to obtain these natural variations as well as to increase the representativeness of the dataset.

The dataset created in the study is a custom dataset that contains a wide range of street food images from the real world, which can be used for the classification of street food images using deep learning algorithms. The photographs consist of images taken from the street and images selected from public online photos to ensure a wide variety of food appearance in the street, while preserving the class balance of the photographs.

4.3 Dataset Description

Performance of deep learning models is greatly influenced by the quality and variety of a dataset. A huge street food image database comprising 10 common street food categories in Bangladesh was created to build an accurate and strong street food recognition system. The dataset aimed to capture the natural variation in the appearance of street food, thus allowing the trained models to function well in actual real-world environments.
This is an unique dataset created for this particular application because as compared to many available public datasets, the data here represents authentic Bangladeshi street foods which is not usually found in western or international cuisine. Images are varied in ingredients, presentation

style, containers of serving, light conditions, viewing angle, and background, resulting in a realistic measure for assessing image classification algorithms.

4.3.1 Dataset Composition

The dataset has 16,000 images. 1,600 in each class. The ratio of the pictures enables the total number of images to be constant for each type and thus reduces class imbalance.
The full dataset composition is presented in Table 4.2. So, the balanced number of images for each type helps to have an unbiased learning experience without a favor for the classes with greater quantities of pictures for training.

4.3.2 Characteristics of the Dataset

Images collected from the database are highly varied and display distinct visual characteristics. Since images were taken from real-world situations, we can see a significant amount of intra-class variation within classes where at the same time classes differ from each other.
The dataset shows varying and changing:

- Types and colors of food.
- Styles of decoration.
- Amount of ingredients.
- Utilization of platters, bowls, newspapers, and disposable containers.
- Camera distance and angles.
- Lighting conditions and shadows.
- The backgrounds both inside and outside.
- Extensions of hiding from covers from cutlery and other objects around.

The above variations make the dataset more practical in terms of real-world street food recognition and also increase the works of deep learning models.

4.3.3 Dataset Diversity and Generalization

An important goal of dataset creation was to guarantee that the models trained would function in real life, not just in laboratory settings. For this reason, the dataset was purposely made to contain photos taken in different environmental circumstances.

- Some instances of different conditions in the dataset are:
- Bright sunlight and very small illumination.
- Different smartphone cameras used.
- Many angles of sight.
- Different volumes and servings of food.

- Cluttered images.
- Natural shadows and reflections.

All the mentioned variations should make the models capable of learning useful visual representations no matter how the situation changes in the real world around them.

4.3.4 Dataset Quality

All the images collected were manually checked for overall quality and correctness prior to training. Photos that were not in focus enough, duplicated, corrupted or misidentified were not included. All images were manually verified to ensure trustworthy class labels for each image and its corresponding street food category.

Moreover, the number of images was kept constant on each category resulting in a balanced dataset to enable fair model training and unbiased performance evaluation.

4.3.5 Importance of the Dataset

The proposed dataset is the bedrock of this study and caters to the absence of publicly available datasets focused on the recognition of street food in Bangladesh. The developed food dataset is unique in that it represents the visual attributes of street foods from Bangladesh from authentic environments, compared to existing international food datasets that mostly represent food from western countries or China.

This dataset will not only enable the creation of precise deep learning systems but also serve as a resource for future studies in food recognition and dietary evaluation, nutritional monitoring, and computer vision studies on Bangladeshi cuisine. It is balanced, highly visual and has realistic environmental conditions and is suitable to test transfer learning methods and build realistic machine vision applications.

4.4 Dataset Organization

Efficient model training, validation and testing require proper organization of the data set. A well-structured dataset makes the data management easier also reduces labelling mistakes, and can be used with deep learning frameworks like TensorFlow and Keras. Before the street food images were collected, they were classified and divided into different food categories on a systematic basis, and then preprocessed and trained in the model in this research.

4.4.1 Class-wise Folder Organization

The obtained photographs have been sorted into ten individual folders, each of which corresponds to one of the Bangladeshi street food categories. The folders become class labels in the course of the model training. The filenames of the pictures were not changed in folders, because the image label was assigned automatically according to the parent folder.

The dataset directory structure is shown below.

The directory structure of the dataset is illustrated below.

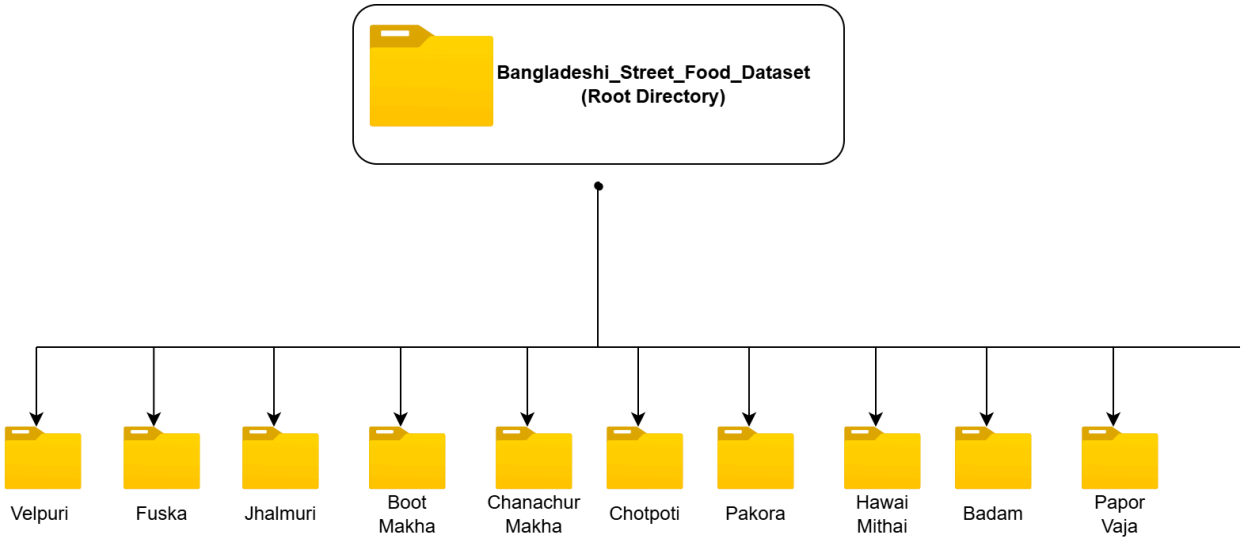


Figure 4.3 : Directory structure of the Bangladeshi street food dataset.

4.4.2 Dataset Splitting

Once the dataset was organized, the image filename and class labels were documented into a Pandas DataFrame. The data set was then divided into three sections Using the train_test_split() function in scikit-learn package.The data set was then divided into three subdata sets. Using train_test_split() from the Scikit-learn package. The dataset was split into 80% training, 10% validation and 10% testing.

The data used for training the deep learning models was the training data set, and the remainder of the data was used for testing. validation dataset was used to monitor the efficiency of the models, as well as The training dataset was optimized, while the testing dataset was used for fundamental Evaluation of models performance on images not encountered.

The final repartitioning of the dataset is shown in Table 4.3.

Table 4.3: Dataset Distribution

Dataset Subset	Percentage	Number of Images
Training	80%	12,800
Validation	10%	1,600
Testing	10%	1,600
Total	**100%**	**16,000**

4.4.3 Label Assignment

During the process of data loading, the data generator of TensorFlow/Keras automatically identified the respective class labels from the name of the respective folder. Hence, it was not necessary to alter the name of every single image. The folder names worked as the categorical labels which were suitable for one-hot encoding for multi-class classification in the preprocessing phase.

4.4.4 Label Encoding

The process of data loading consisted of different categories of street food that were assigned a specific class label depending upon the parent directory name with the help of the TensorFlow/Keras data generator. The class labels are then converted into one-hot encoded format class_mode='categorical' for multi-class classification. One-hot encoding indicates that each class is represented in the form of a binary vector which helps the Softmax output layer to give output in the form of a probability for each of ten categories of street food. This method of

encoding is beneficial from the standpoint of deep learning as it helps in computing categorical cross-entropy loss.

4.4.5 Advantages of the Dataset Organization

The adopted dataset organization offers several benefits during model development process:

- Arranges photos based on their food associations.
- Simplifies automatic loading of data from TensorFlow/Keras data generators.
- Avoids having to name each image file.
- Aids in effective batch generation while training the model.
- Supports the automatic splitting of the dataset into training, validation and testing sets. During the pre-processing, a testing sets is set.
- Supports multiple deep learning models in training and evaluating them in a consistent data structure.

The performance of all the evaluated models could be compared fairly in identical experimental conditions, as all of these models were trained with the same pipeline for organizing and pre-processing the data.

The Bangladeshi street food data set was carefully categorized into individual folders by each food class so that during data loading, the class labels can be automatically recognized by TensorFlow/Keras. The data was kept in this class wise folder structure even after data preparation and the training, validation, and testing subsets were created programmatically in the pre-processing stage, with a random data split. This organization made the management of datasets easier, ensured efficient model training, and established a common platform for assessing various deep learning models.

4.5 Image Augmentation

One of the major preprocessing for deep learning is 'image augmentation' which simply means augmenting dataset with created data such as varying versions of existing images. It provides more powerful and generalized feature representations for convolutional neural networks (CNNs) through the process without needing to collect extra images. To enhance the performance of deep learning models as classifiers, image augmentation was used as the technique for this research.

The data set was well balanced, but to increase diversity of the available images, images were augmented separately prior to the preprocessing step, to diversify the dataset and minimize overfitting. Thus the number of images was 16000, which gives 1600 images per class.

4.5.1 Purpose of Image Augmentation

The major goals of Image Augmentation in this study were to:

- Make more training points, using more memory but no extra images.
- Make the model more robust to generalize to unseen images.
- Remove overfitting due to small data sets.
- Represent real world changes in orientation, position and scale of images.
- Strengthen the performance of their trained models to cope with environmental fluctuations.
- Exposing the models to multiple variations of the same image, the network is able to learn the general visual features of each food type instead of just memorizing the image pattern.

4.5.2 Augmentation Techniques

Every original photo was augmented by six different techniques resulting in similar but new training examples. The augmentation techniques were implemented in a way ensuring that the semantic content of the photos was preserved while realistic differences usually encountered in practical image acquisition are introduced. Description of the augmentation techniques used in the present study can be found in Table 4.6.

Table 4.4: Image Augmentation Techniques

Technique	Parameter	Purpose
Horizontal Flip	Applied	Simulates opposite viewing direction
Rotation	40°	Introduces orientation variation
Zoom	0.2	Simulates different camera distances
Width Shift	0.2	Simulates horizontal object displacement
Height Shift	0.2	Simulates vertical object displacement
Shear	0.2	Introduces geometric distortion similar to perspective changes

The augmentation parameters were chosen so as to provide enough variability of images while retaining the visual features of every food category. These transformations made it possible for the models to learn invariant features.

4.5.3 Dataset Expansion

The technique used for image augmentation involved applying the selected techniques to all original images, thus creating additional image variants without modifying the original class labels. As a result, the file size of the image dataset has increased from 4,000 to 16,000 images. The augmentation process is summarized in Table 4.5.

Table 4.5: Dataset Size Before and After Augmentation

Description	**Number of Images**
Original Dataset	4,000
Images per Class (Original)	400
Augmented Dataset	16,000
Images per Class (After Augmentation)	1,600

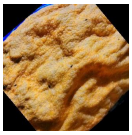
The equal distribution of augmented images across different classes ensures that all food classes contributed equally to the learning process, thus eliminating class imbalance.

4.5.4 Representative Augmented Images

As an example of the process of augmentation, Table 4.7 illustrates some samples of the images produced. In the figure we can see how various augmentation methods affect the image whereas the identity of the food remains the same.

Table 4.7: Data Augmentation effects on each image class

Street Food Classes	Original Image	Augmented Images					
		Horizontal flip	Rotation :40 degrees	Zoom: 0.2	Width shift: 0.2	Height shift: 0.2	Shear: 0.2
Velpuri							
Fuska							
Jhalmuri							
Boot Makha							
Chanachur Makha							
Chotpoti							
Pakora							
Hawai Mithai							

Badam							
Papor Vaja							

In Table 4.7 one can see the source image and its augmented images created with horizontal flip, rotation (40°), zoom (0.2), shifting in the direction of width and height (0.2), and shearing (0.2). The variations in the images will lead to improvement of the quality of the data and make the deep learning systems stronger.

4.6 Image Preprocessing

Image preprocessing, being a basic step of image classification, is needed to convert the original image to a unified service form, which can be used to train the model. The images with which you may have come across can be of different dimensions, resolution, brightness and quality. If not properly preprocessed, such variations can have an adverse influence on the learning process and further decrease the classification efficiency of deep learning models. Therefore, a series of preprocessing operations were applied to all images before they were used for training and evaluation. The Image preprocessing pipeline is shown in Figure 4.4.

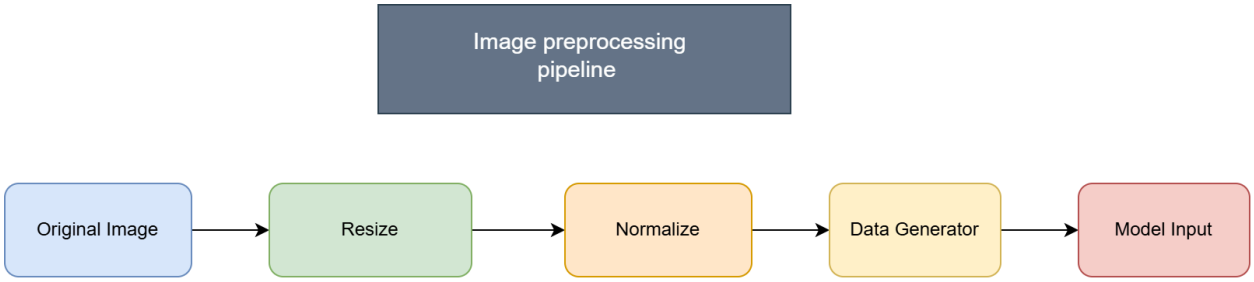


Figure 4.4: Image preprocessing pipeline for the Bangladeshi street food dataset.

4.6.1 Image Resizing

The images in the data set were of varying resolution based on the camera device and image source. All images were converted to a common format, 224 × 224, prior to supplying the

images to the neural networks to ensure uniformity amongst images for all the deep learning architectures.

The chosen image dimensions are commonly found in Transfer Learning Networks like MobileNetV2, MobileNetV3-Large, ResNet50, DenseNet121, EfficientNetB0, ConvNeXt-Tiny, Xception, NASNetMobile and InceptionV3. Also, by making the images of the same size, the amount of computation becomes lower and processing of several images can be performed in a batch.

4.6.2 Pixel Value Normalization

Typically, deep learning models have improved performance with normalized input features. In this study, we have normalized pixel values using the following pre-processing formula:
Normalized Pixel Value = (x/127.5) – 1
where “x” corresponds to the original pixel intensity in the range from 0 to 255. Thus, pixel values were converted in the range of [-1, 1]. Such a normalization contributes to numerical stability during optimization as well while meeting pre-processing requirements of MobileNetV2 and other pretrained convolutional neural networks.

4.6.3 Data Loading

The ImageDataGenerator class was utilized in the TensorFlow/Keras framework to load the images after preprocessing. By using a Pandas DataFrame to generate the images, both the paths to the image files and the respective class labels were read by the framework. In this way, different batch images of training, testing, and validation were created automatically.The batch size for this training was determined at 32 which is the same for the validation and testing purposes.

4.6.4 Label Encoding

Due to the process of image classification involving images of street food into 10 categories, categorical class labels automatically got converted into one-hot encoding using the parameter 'class_mode=categorical' of Keras’ data generator.

The one-hot encoding technique makes it possible for the Softmax layer of output to generate probabilities of the images belonging to each of the ten food classes and therefore allows for the use of the categorical cross-entropy loss function that is employed while training the model.

4.6.5 Batch Generation

The ImageDataGenerator did not load images all at once in memory, but produced the images in batches during the time of training. While usefully saving memory and processing power, this method has helped the model in processing enough data at the same time, as the image data contained 16,000 samples. The advantage of producing batches is using them while keeping the images in the same pipeline for both training/validation and testing.

4.6.6 Benefits of Image Preprocessing

The study shows that there are several advantages of the operations of pre-processing. These are:

1. Gives the same dimension of the image for deep learning.
2. Normalizes the pixel intensity to stabilize learning.
3. Provides support for pre-trained convolutional neural network models.
4. Helps with using batch training mode.
5. Decreases the complexity of calculation processes.
6. Uses one data format for all the analyzed models.

The importance of pre-processing algorithms lies in their impact on convergence, generalization, and correctness of classification.The dataset was pre-processed by resizing the images to 224×224 pixels and normalizing them to the range from -1 to 1.After that, one hot encoding of the class labels was carried out.With the help of TensorFlow/Keras, the batches of images were formed which provided the models with uniform data and improved the training process of deep learning models with the goal of ensuring the accuracy and correctness of street food classification.

4.7 Transfer Learning

Transfer learning is a technique in the field of deep learning that is used to leverage a pre-trained model for a different and related task. Transfer learning leverages pre-learned "what is going on in the visual world," such as edges, textures, shapes, and what an object is. This process reduces the size of training set and computation power needed and enhances its classification performance.

In the present research, Transfer Learning has been employed for the development of an automated street food recognition system on the streets in Bangladesh. Although the data used in

this research is extensive, they are still small scale data that can be compared to large volumes of image data. Thus, if the classification of images of objects with known names had a large number of categories, such as those found in the ImageNet dataset [22] of millions of images of various objects, the models could leverage a pre-trained representation learned from this dataset. Thus the pretrained transfer learning models have proven to be efficient in solving Solve the complex visual tasks and achieve good results in various applications based on the computer vision.

4.7.1 Motivation for Using Transfer Learning

Deep Convolutional Neural Networks (DCNN) are trained using a large number of labeled samples, which is impractical due to the high computational costs and time requirements. It's time consuming to collect millions of labeled images, and in many real world applications, including Bangladeshi street food, expensive can be a deal breaker. recognition. To overcome this, the field of 'transfer learning' has emerged, which involves using a pre-trained model to retain knowledge acquired by a previous classification task and transfer it to a new one.

Some reasons for implementing transfer learning in this work are:

1. Minimizing the number of training instances needed.
2. Decreasing training time and computational cost.
3. Enhancing the extraction ability of features.
4. Reducing the risk of overfitting.
5. Achieving higher classification accuracy.

These benefits make transfer learning a very promising option for image classification tasks with a moderate-sized dataset.

4.7.2 Transfer Learning Process

The transfer learning method used in this research used several stages in sequence. To start the process, the pretrained CNN model trained on the set of the ImageNet is used in the feature extractor used was selected as such. The older classification part was removed from the trained model and swapped with a new classification head made for the ten categories of the Bangladeshi street food. At the next step, the new classification layers were trained on the prepared Bangladeshi street food dataset with the help of the pretrained network that was utilized as a feature extractor. In addition, MobileNetV2 architecture also included further fine-tuning stage after the first training.

The whole transfer learning scheme used for the research can be seen on Figure 4.5.



Figure 4.5: Transfer Learning Workflow of the research of Bangladeshi street food recognition.

The process of transfer learning applied for this research can be seen in Figure 4.5. It includes the use of the pretrained model on the ImageNet dataset, the removal of the classification part, as well as the training on the Bangladeshi street food dataset.

4.7.3 Pretrained Models

For the purpose of assessing how effective transfer learning is when it comes to recognition of Bangladeshi street food, a number of advanced and widely recognized CNN architectures were researched. The CNN architectures were selected because they have been reported to be effective for image recognition tasks and have the capacity to obtain the features for classification from complex images. The following are some of the pretrained models used in the present study,

1. InceptionV3
2. DenseNet121
3. MobileNetV2
4. NASNetMobile
5. Xception

Table 4.8 Summary of Selected Deep Learning Models.

Model	Pretrained Dataset	Depth	Parameters (M)	Key Features
InceptionV3	ImageNet	159	23.9	Efficient modules, reduced computation
DenseNet121	ImageNet	242	8.1	Feature reuse, smaller size
MobileNetV2	ImageNet	105	3.5	Lightweight, mobile-friendly
NASNetMobile	ImageNet	88	5.3	Auto-designed, opti-mized for mobile
Xception	ImageNet	71	22.9	Depthwise separable, efficient design

All models used were initialized with ImageNet pretrained weights and adjusted so that they could classify the 10 types of street food of Bangladesh. Following that, the performances of all models were compared using the same dataset as well as preprocessing techniques and evaluation technique.

4.7.4 Advantages of Transfer Learning

There are several interesting benefits for transfer learning over training deep neural networks from scratch.Existing representations of features are used to significantly minimize the number of training samples and improve efficiency and accuracy of the models in training.

The following are the major advantages of transfer learning:

- Speedy convergence of models during training.
- Lower complexity of computations.
- Excellent acquisition of features from small data sets.
- Good generalisation skills.
- There is a very low risk of overfitting.
- Very high rates of classification accuracy.

Thus, transfer learning is a good alternative when it is difficult or even unfeasible to collect huge sets of data for image classification problems.

The proposed Bangladeshi street food recognition system was based on the transfer learning method. One of the key elements of the approach was the use of pretrained CNN networks trained on the ImageNet dataset that could be used to get useful visual representations with some minimal time loss.

4.8 MobileNetV2 Fine-Tuning

The MobileNetV2 architecture has been chosen as the basis of the proposed classification framework since it is lightweight and computationally efficient while achieving superior results on image classification. To enhance the classification performance of the model for Bangladeshi street food recognition, a two-stage transfer learning technique was applied. The model was first able to learn the features for a specific task through a custom classification head, after which it was able to refine the pretrained feature plane through selective fine-tuning of deeper layers.

4.8.1 Proposed MobileNetV2 Architecture

MobileNetV2 is a compact convolutional neural network, which is very efficient in terms of image classification. Applies depthwise separable convolution, and Inverted residual blocks, and linear bottlenecks. The aforementioned techniques reduce the overall amount of parameters and model complexity, especially when Compared with traditional convolutional networks.A pre-trained model of MobileNetV2, with weights pre-trained on ImageNet dataset, was used in this study. The classification head in the original model was replaced with an original classification model designed to identify street food in Bangladesh.

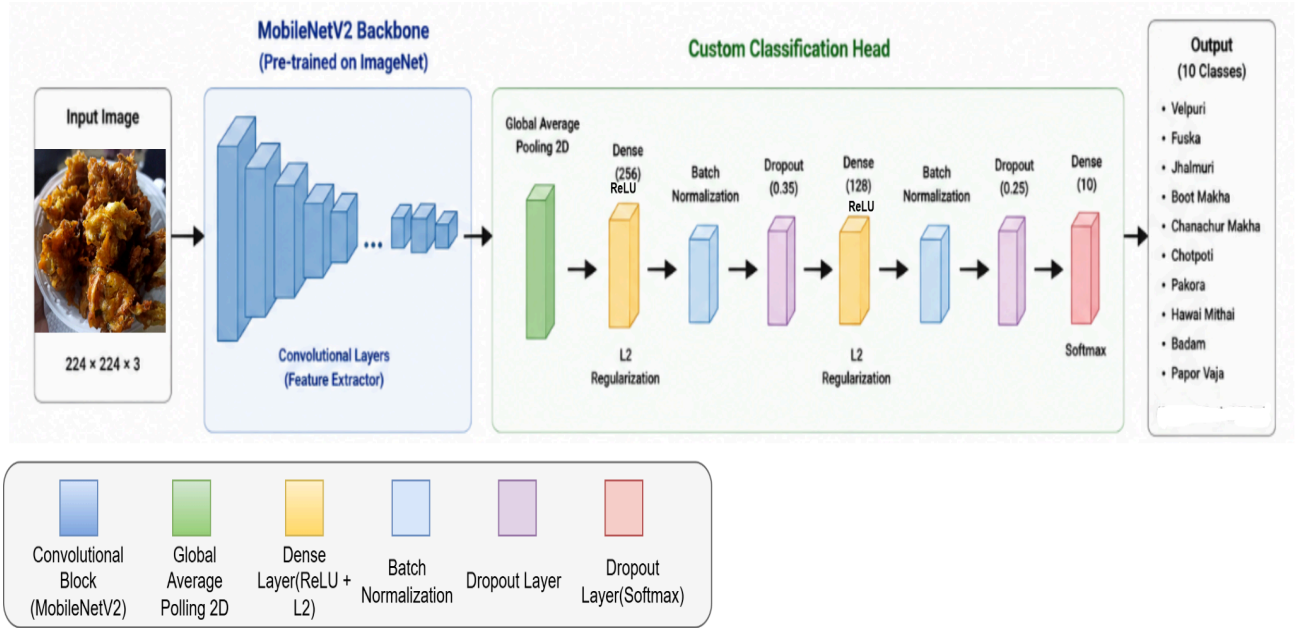


Figure 4.6: Proposed MobileNetV2 Custom head classification.

4.8.2 Phase 1: Feature Extraction

In the initial training phase, the pre-trained MobileNetV2 backbone was not tweaked in any way, thus preserving the gains it previously made in feature representation. During this phase, only the newly formed classification layers were trained on the prepared street food dataset.

The classification part consisted of the following components:

- Global Average Pooling layer.
- Fully connected layer with 256 neurons and ReLU activation function.
- Batch Normalization layer.
- Dropout layer, which had the dropout ratio of 0.35.
- Also, there was a fully connected layer with 128 neurons and ReLU activation.
- Batch Normalization layer.
- Dropout layer, which had the dropout ratio of 0.25.

As a result, the Softmax output was produced, which contained 10 neurons that were intended for the ten classes of street foods. Then L2 regularization was used in order to avoid the problem of overfitting, which is peculiar to models of that type. Therefore, the training of the model was carried out through the use of the Adam optimizer with learning rate of 0.001 and loss function

categorical cross-entropy. The overall architecture of the proposed Phase 1 model is illustrated in **Figure 4.7**.

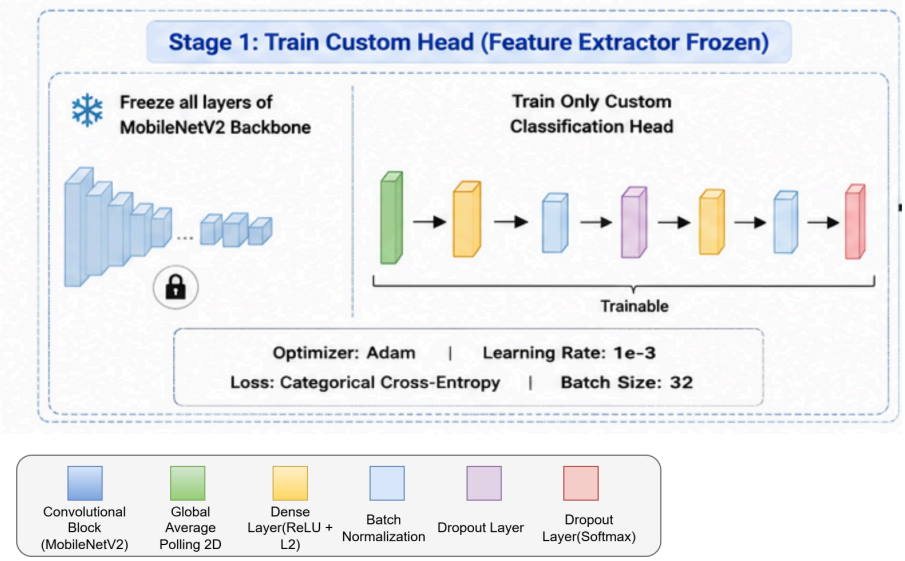


Figure 4.7: Phase 1 of proposed MobileNetV2 Architecture with Custom Classification Head.

4.8.3 Phase 2: Fine-Tuning

After the custom classification head is trained, the next step is to fine-tune the pretrained MobileNetV2 model. Only the last part of the pre-trained model was unfrozen and the previous part was frozen. This method enabled the model to learn general visual attributes from ImageNet while fine-tuning the upper layers of the model to learn about Bangladeshi street food images.To prevent that the performance of the already learnt features would be affected by large updates to weights, the learning rate was set relatively low (0.00001) for this phase. Also used was the Adam optimizer and categorical cross-entropy loss function.

This two-stage training technique led to an increase in the performance of the model in terms of differentiating visually similar street food categories and reducing overfitting.

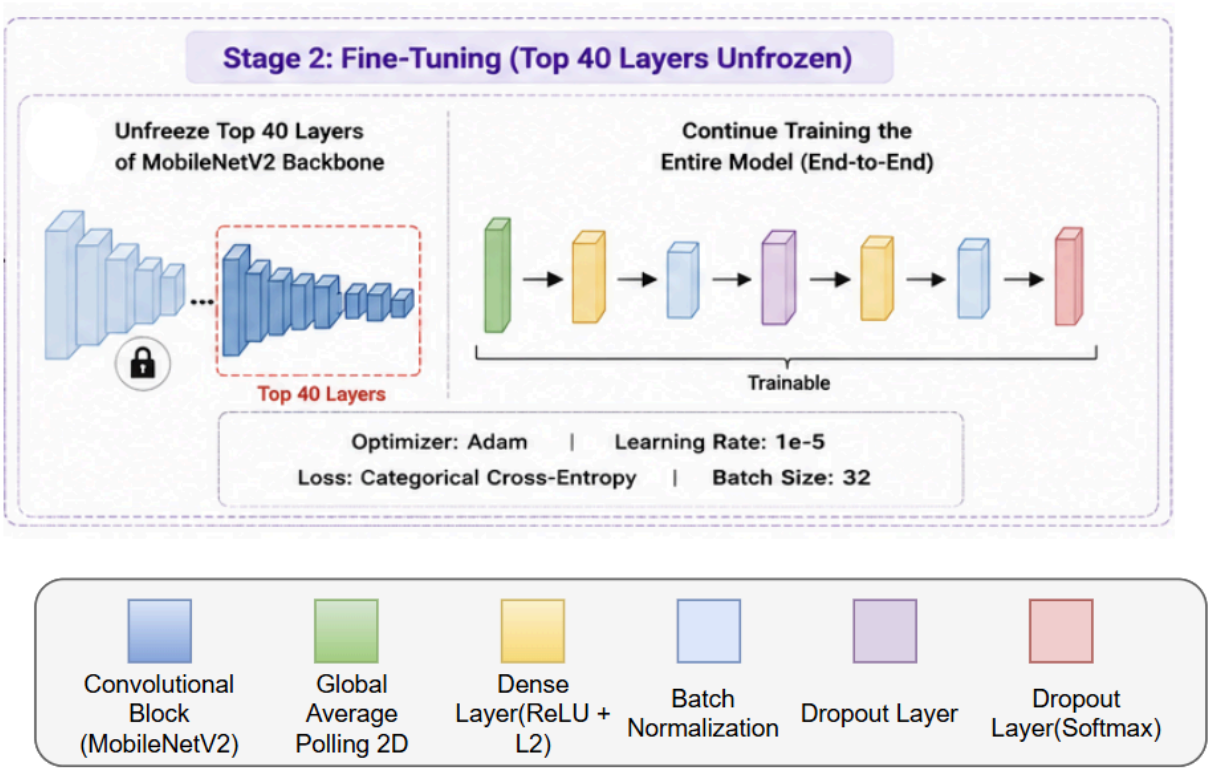


Figure 4.8 : Phase 2 focused on fine-tuning the pretrained MobileNetV2 model.

4.8.4 Training Strategy

Several techniques were incorporated during model training to improve learning stability and prevent overfitting.

- **Early Stopping** was used to terminate training automatically when the validation performance stopped improving.
- **Model Checkpoint** saved the best-performing model during training based on validation accuracy.
- **Reduce Learning Rate on Plateau** automatically decreased the learning rate when validation loss ceased to improve, allowing finer optimization during later stages of training.
- **Batch Normalization** improved training stability by normalizing intermediate feature distributions.
- **Dropout** randomly deactivated neurons during training to reduce overfitting and improve generalization.

These techniques collectively contributed to improved convergence and robust model performance.

4.8.5 Advantages of the Two-Stage Fine-Tuning Strategy

The proposed two-step learning has several advantages compared to standard transfer learning. These are,

- It preserves the interesting pre-trained ImageNet features.
- Gets features for tasks using a custom classification head.
- Minimizes the chances of overfitting.
- Boosts convergence steadiness.
- Enhances the discrimination between closely similar classes of street foods.
- Increases the accuracy of classifications and keeps costs down.

The method of gradual adapting of the pretrained model can provide the ability to successfully transfer generic knowledge while at the same time learning features particular to Bangladeshi street foods.

4.8.6 Performance of the Proposed Model

The proposed MobileNetV2 fine-tuning framework outperforms all the tested transfer learning models. The feature extraction, selective fine-tuning, batch normalization, dropout and L2 regularization are sufficient to learn highly discriminative visual representations while achieving excellent generalization performance.
The results of our proposed method suggest that the integration of lightweight deep learning architectures and optimized two-stage transfer learning method in the field of Bangladeshi street food recognition is effective.

4.9 Other Compared Models

In order to compare the efficacy of the proposed MobileNetV2 fine-tuning model, multiple well-known state of the art CNN architectures have been included in the process. Although these architectures are different in terms of their designs, they have been shown to be useful for the image classification applications. Different models will provide a comprehensive assessment of the results obtained in this field.

In order to compare the models fairly, all those models were pre-trained using ImageNet weights and were trained on the same data and using the same data processing and evaluation techniques.

4.9.1 InceptionV3

InceptionV3 can be considered as a deep convolutional network which employs Inception modules making it more effective for extraction of visual features at multiple scales. The unique aspects of the model related to the use of factorized convolution, as well as auxiliary classifiers improves efficiency and stability of its training process.

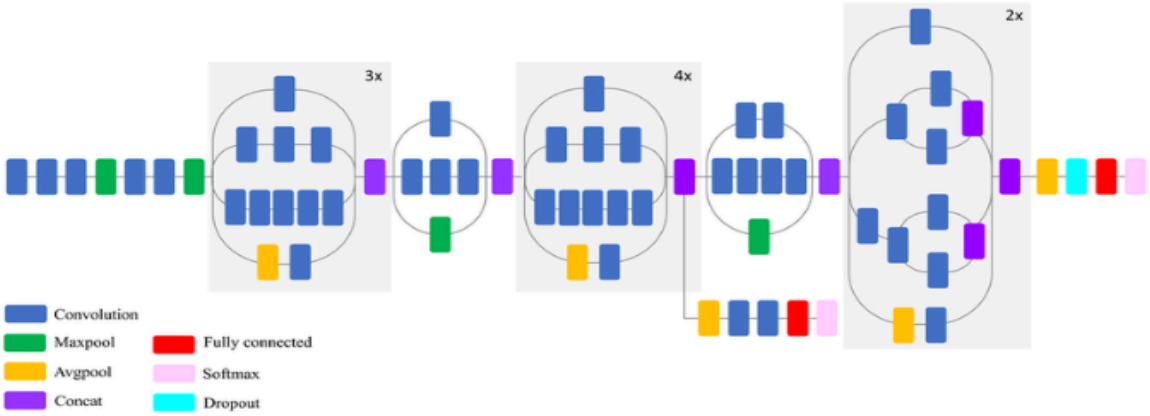


Figure 4.9: Architecture of the InceptionV3 transfer learning model used in this study.

4.9.2 MobileNetV2

To achieve high classification accuracy with low computational complexity, MobileNetV2 is a small convolutional neural network. The architecture adopts the concept of inverted residual blocks and linear bottlenecks to render efficient feature extraction by cutting down the number of trainable parameters and computation steps while preserving performance to an acceptable level. The design is well suited for resource-limited applications like embedded systems and smartphones, and the resulting architecture, MobileNetV2. The design is ideal for mobile devices with limited resources, like smartphones and embedded systems, and the resulting architecture. Because of its efficient architecture and good transfer learning ability, MobileNetV2 has been adopted widely for many computer vision applications such as image classification, object detection and medical image analysis.

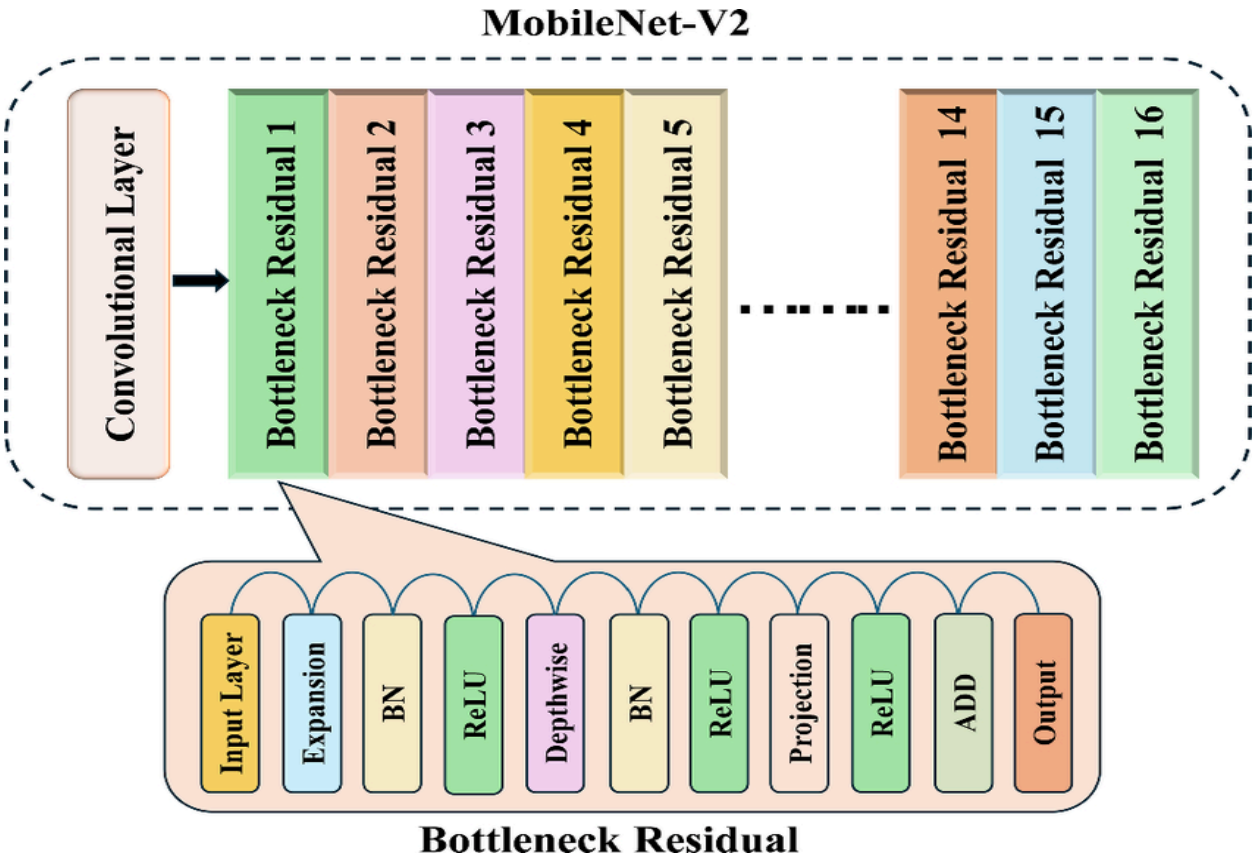


Figure 4.10: Baseline MobileNetV2 transfer learning architecture.

4.9.3 DenseNet121

DenseNet121 is a dense network, where the features of each layer are all linked to each other. This design can be reused, boost gradient flow and lower the number of parameters to be trained. The highly compact DenseNet121 network has proven effective in image classification tasks with good parameter utilization.

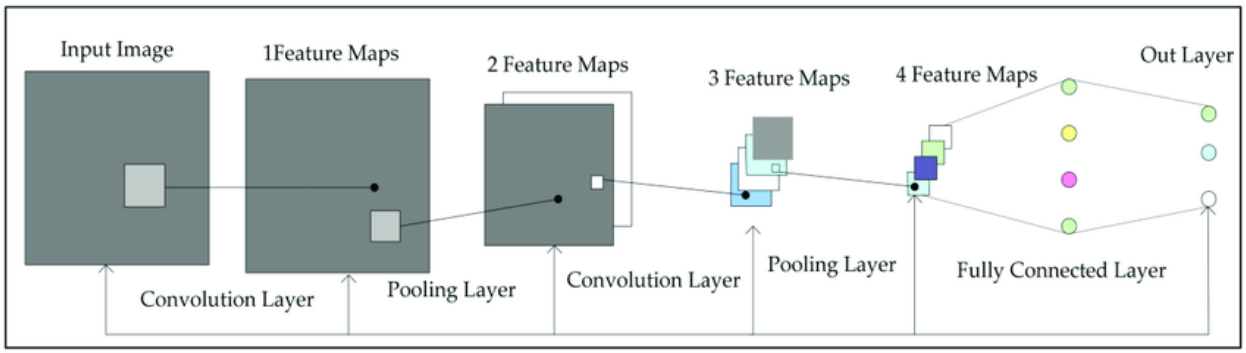


Figure 4.11: DenseNet121 transfer learning architecture**.**

4.9.4 NASNetMobile

NASNetMobile is a neural architecture search (NAS)-based model whose architecture was automatically optimized through machine learning techniques. It is specifically designed to deliver strong classification performance while low computational demands that allow it to be used in demanding environments like the resource-constrained environments.

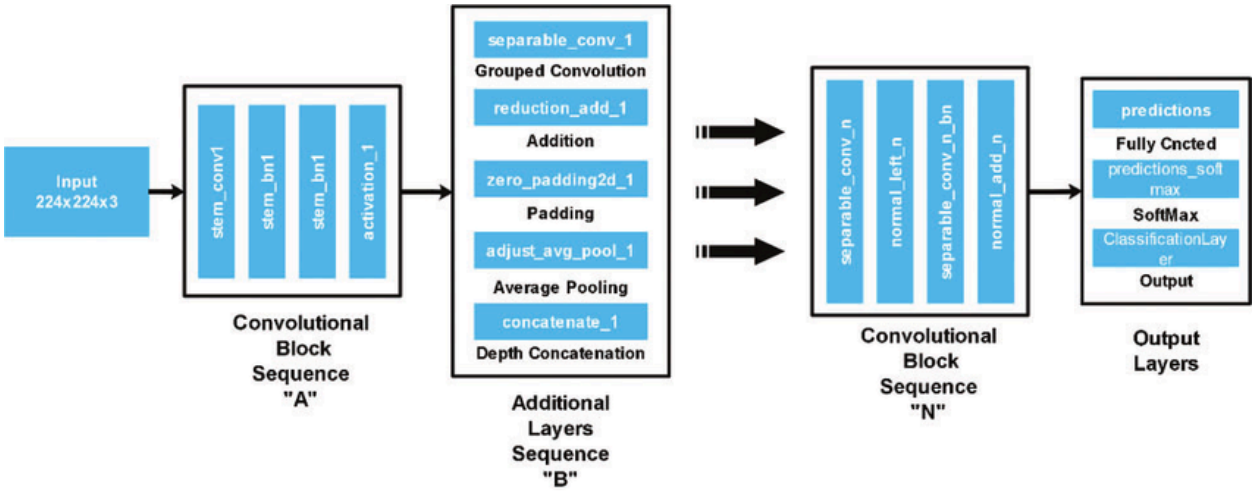


Figure 4.12: NASNetMobile transfer learning architecture.

4.9.5 Xception

Xception involves the upgrade of depthwise separable convolutions by substituting the typical convolutional process with more effective methods of feature extraction. As a result, this architecture is distinguished by its effectiveness in terms of computation and quality of classification results, which make it ideally suitable for transfer learning problems which include dealing with challenging datasets.Xception involves the upgrade of depthwise separable convolutions by substituting the typical convolutional process with more effective methods of feature extraction. As a result, this architecture is distinguished by its effectiveness in terms of computation and quality of classification results, which make it ideally suitable for transfer learning problems which include dealing with challenging datasets.

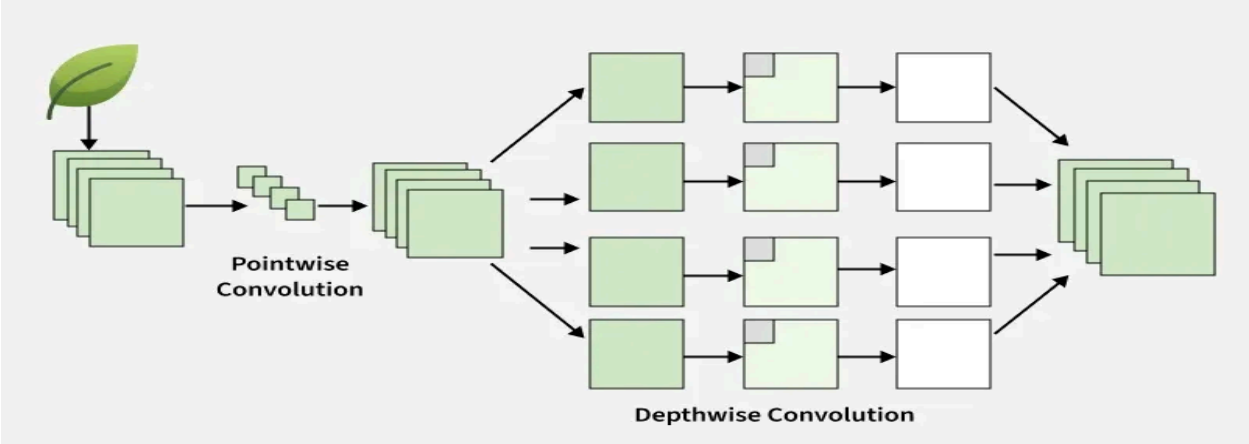


Figure 4.13: Xception transfer learning architecture.

4.9.9 Comparative Evaluation

The pre-trained network is processed, trained and evaluated with the same street food dataset, preprocesses and evaluation metrics from Bangladesh. This has guaranteed that the differences in performances are due to the architectures themselves rather than the data processing or experimental setup. The comparison was carried out in this manner, which enabled a precise comparison of the pros and cons of each model in the area of street food image recognition. A comparison of the conducted five pre-trained CNN architectures helped to evaluate the proposed street food recognition system on the four mentioned CNN architectures.

Although the model we examined is very helpful to the use of transfer learning in the application, it seems that the model we provided MobileNetV2 fine-tuned is the best framework in terms of processing efficiency and classification performance.

4.10 Performance Metrics

The performance of the newly designed Bangladeshi street food recognition system was evaluated using a number of standardized classification performance metrics. These metrics provide an in-depth assessment of the ability of the system to classify by looking at various elements of class performance. The same rules for calculation of the pre-trained CNN models were applied to calculate results of different models for comparison purposes.

4.10.1 Confusion Matrix

The confusion matrix helps to present the classification results based on the comparison of the predicted and actual classes. The numbers in each of the columns of the matrix corresponds to the actual class while each row refers to the predicted class. The correct classifications are represented on the principal diagonal line of the matrix while the values in the other locations of the matrix correspond to the incorrect predictions. The confusion matrix gives insight into the classes correctly recognized and which classes are more likely to be confused because of visual similarities.

4.10.2 Accuracy

Accuracy is the percentage of correctly classified images to all test images. One of the most widely used evaluation measures for multi-class classification problems.

The accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

Owing to the higher accuracy, the overall classification performance is improved

4.10.3 Precision

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

High precision indicates that the classifier produces fewer false-positive predictions.

4.10.4 Recall

Recall assesses the fraction of true positive results among all positive predictions of the model.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

When recall is at a high level, it means that the model is able to recognize most of the cases of the particular class type.

4.10.5 F_1-score

The F_1-score is a single score that incorporates both precision and recall to produce a harmonic mean.

$$F_1\text{-score} = \frac{2\times Precision \times Recall}{Precision + Recall} \tag{4}$$

This measure gives a well balanced evaluation of the categorization accuracy, especially when comparing between different classes.

4.10.6 Classification Report

The precision, recall, F_1-score and support for each single class is summarized in a classification report. It gives in depth details about the pros and cons of the classifier in each street food category and enables comparing of different CNN models.

Table 4.10: Summary of the evaluation metrics and their purpose.

Metric	**Purpose**
Accuracy	Overall classification performance
Precision	Reliability of positive predictions
Recall	Ability to detect actual positive samples
F_1-score	Balance between precision and recall
Confusion Matrix	Visualization of classification results
Classification Report	Per-class performance analysis

4.10.7 Importance of Multiple Evaluation Metrics

While classification accuracy is used as an overall measure of the performance of a model, using it alone may not fully reflect the strengths and weaknesses of a multi-classifier model. This is because precision, recall, F_1-score, confusion matrices and classification reports were also used to provide a complete assessment of each model. This application of several evaluation metrics makes the proposed system evaluated from various perspectives, and also the various evaluation metrics provide comparison of the evaluated transfer learning models more reliable.

Confusion matrices, accuracy, precision, recall, F_1-score and classification report The proposed street food recognition system was tested with the following in Bangladesh. Through these complementary metrics, a detailed evaluation of the classification ability of each model was carried out, which also allowed for a fair comparison of the CNN architectures evaluated. The results obtained from these metrics are presented and discussed in the following chapter.

5 System Implementation

The implementation of proposed machine vision based Bangladeshi street food recognition system is addressed in this chapter. Although the overall methodology is presented in Chapter 4, this chapter is dedicated to the development and implementation of the proposed system with the use of deep learning techniques. It covers the pipeline of implementation of a model from data preparation to preprocessing, model development, model prediction, to explainable artificial intelligence (XAI).

Kaggle Notebook was used to implement it using Python programming language and the TensorFlow and Keras deep learning APIs. A custom dataset was created that includes 10 classes of Bangladeshi street food items and used for the training and testing of various pretrained Convolutional Neural Network (CNN) architectures using transfer learning. The dataset was expanded and processed with image augmentation techniques and a standard preprocessing pipeline prior to modeling to enhance recognition performance.

The implementation of the baseline CNN models and the proposed two-stage MobileNetV2 transfer learning framework is also described in this chapter. It also outlines the process of which the trained model is used to make the image classification and how the Explainable Artificial Intelligence (XAI) techniques like LIME and Grad-CAM were incorporated to explain the prediction results. The information and explanations given in this chapter will help to understand the complete picture of the implementation of the proposed recognition system for the classification of street food in Bangladesh.

5.1 System Overview

In this chapter the implementation of proposed machine vision based street food recognition system for identification of Bangladeshi street foods is presented. The system has been created with transfer learning of pre-trained Convolutional neural network (CNN) architectures for automatic image classification of 10 street food categories from the Bangladeshi regions. The implementation combines the preparation of the datasets, pre-processing of the images, training of the model, prediction and Explainable Artificial Intelligence (XAI) techniques into a single stream. Experiments were run in the Kaggle cloud environment and the models were developed using Python, TensorFlow and Keras.

The first step in the implementation is to prepare the enhanced dataset and all the images are resized to 224 × 224 pixel, normalized, and split into training, validation, and test sets. To provide baseline performance, five pretrained CNN models, namely InceptionV3, DenseNet121, MobileNetV2, NASNetMobile and Xception, were implemented by transfer learning. In a comparative evaluation, the MobileNetV2 network was then improved with a proposed two-stage transfer learning strategy with progressive fine-tuning to enhance the classification accuracy without compromising the computational efficiency.

Once the model is trained, the optimized model can predict the class label of an unseen street food image by predicting the class probabilities using the Softmax output layer. For better interpretability and transparency of the recognition process, Grad-CAM and LIME were integrated to visualize the regions of the image and the superpixels which contributed most to the model's predictions. The rest of the paper explains each step of the system implementation in detail: dataset processing, development environment, model implementation, prediction workflow and XAI analysis. A general design of the implemented Bangladeshi street food. recognition system is shown in figure 5.1. In the implementation, first an input image goes through some preprocessing such as resizing and normalizing before it is fed to the evaluated pretrained CNN models. The baseline models were compared and MobileNetV2 was chosen for further optimization with the proposed two-stage transfer learning strategy. The optimized model is then used to make the final prediction about the food category, and afterwards, Grad-CAM and LIME are used to produce a visual explanation of the prediction to make the food system easier to interpret.

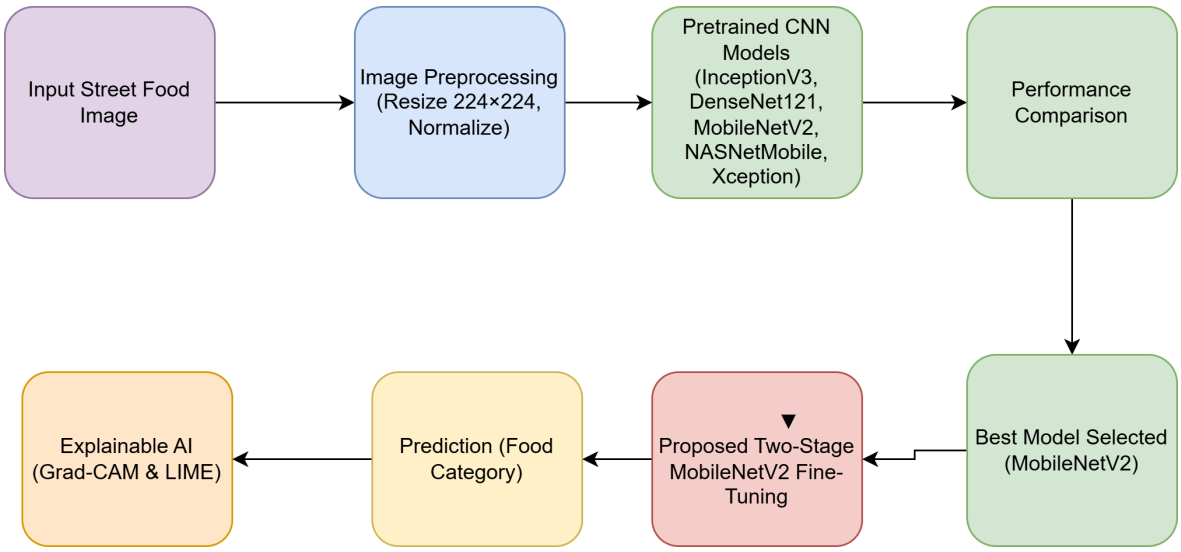


Figure 5.1: Overall System Architecture of the Implemented Bangladeshi Street Food Recognition System .

5.2 Development Environment

5.2.1 Hardware Configuration

The experiments were conducted with the hardware resources provided by the Kaggle Notebook environment. GPU acceleration was used to minimize the overall time required for training deep learning models.

Table 5.1: Hardware Configuration

Component	**Specification**
Computing Platform	Kaggle Notebook
Processing Unit	GPU-enabled cloud environment
Graphics Processing Unit (GPU)	NVIDIA Tesla GPU*
System Memory	Kaggle cloud environment
Storage	Kaggle cloud storage

5.2.2 Software Environment

The proposed framework was developed utilizing the Python language and TensorFlow/Keras deep learning framework. Others libraries were also employed for data manipulation and performance evaluation.

Table 5.2: Software Environment

Software / Library	Purpose
Python	Programming language
TensorFlow	Deep learning framework
Keras	Model development and training
Scikit-learn	Dataset splitting and performance evaluation
NumPy	Numerical computations
Pandas	Data organization and analysis
Matplotlib	Performance visualization
OpenCV	Image processing
LIME	Explainable AI
Seaborn	Confusion matrix visualization

5.2.3 Training Configuration

The experiments were conducted with the hardware resources provided by the Kaggle Notebook environment. GPU acceleration was used to minimize the overall time required for training deep learning models. The proposed framework was developed utilizing the Python language and TensorFlow/Keras deep learning framework. Others libraries were also employed for data manipulation and performance evaluation. It must be noted that all CNNs used for the tests were trained with the same partitions and preparatory procedures to guarantee the same conditions during the process. The images were resized to 224 x 224 pixels and normalized prior to being passed to the networks. Adam optimizer and categorical cross-entropy loss function were used for the optimization of the models.

As for the proposed MobileNetV2 framework, the training process was divided into two stages. During the first stage the backbone was frozen and only the classification head was trained. After that, in the second stage, some layers of the trained backbon were fine tuned in order to improve classification.

The important parameters of training are provided in Table 5.3.

Table 5.3: Training Parameters

Parameter	**Value**
Input Image Size	224 × 224 pixels
Batch Size	32
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Output Activation	Softmax
Initial Learning Rate (Phase 1)	0.001
Fine-Tuning Learning Rate (Phase 2)	0.00001
Epochs (Phase 1) (Proposed two staged transfer learning model)	10
Epochs (Phase 2) (Proposed two staged transfer learning model)	20
Weight Initialization	ImageNet
Callbacks	EarlyStopping, ReduceLROnPlateau, ModelCheckpoint, CSVLogger

5.3 Model Implementation

This part deals with the use of the convolutional neural network models in order to recognize Bangladeshi street food. The convolutional networks were built on the foundation of TensorFlow and Keras platforms. The models were trained in the Kaggle Notebook environment using

Python programming. The programming team used the transfer learning method for all models, applying the transfer learning technology with the use of ImageNet pre-trained weights, at the same time improving the efficiency of the neural network training process.

InTrainings Five models of Convolutional neural networks were applied as the reference. The team assessed the models using the same set of techniques, methods of processing the data and the assessment procedure. All pre-trained networks were based on the ImageNet pre-trained weights passed to the new model from the ImageNet.

Out of all reference models evaluated, MobileNetV2 seems to be the best in terms of the performance and effectiveness of classification. As a result, the MobileNetV2 model was improved and enhanced with the help of two-phase transfer learning approach. The details of both models are provided in the next chapter.

5.3.1 Baseline CNN Model Implementation

The baseline CNN architecture was created by deploying pre-trained models which can be found in the TensorFlow Keras Applications library. Every model was initialized with pre-trained weights on ImageNet, while the most recent fully connected classification layer was omitted (include_top = False). This way, the pre-trained convolutional backbone was used rather as a general feature extractor for street food images from Bangladesh.
The authors followed the convolutional backbone by a global average pooling layer. that transformed the feature maps derived from the above into a compact feature vector.Then, the Softmax classification layer was applied. It was fully connected and had ten output neurons that performed multi-class classification suited to ten street food categories. During baseline training, the convolutional backbone did not change, suggesting that only the new classification layer could be optimized. The Adam optimizer and categorical cross-entropy loss function were used during the entire process, while EarlyStopping was introduced to avoid unnecessary training when the validation results stop improving.
This method of implementation allowed getting the same transfer learning framework for each of the architecture being studied and assures that any differences in results occur not due to dissimilarities in training strategies but because of distinct feature extraction capabilities of the pretrained CNN models.

5.3.2 Proposed Two-Stage MobileNetV2 Implementation

This section explains the practical implementation of the proposed MobileNetV2 framework. All implementations are conducted on the Kaggle notebook environment using TensorFlow and Keras libraries. The implemented framework strictly follows the architecture presented in Methodology and has two different training procedures described below.

During the testing phase, the preexisting MobileNetV2 framework (trained on ImageNet data) will be used to initialized weights of MobileNetV2 backbone and our designed classification header. Finally we construct a end-to-end model for the recognition problem. The complete framework is separated into different procedures in the following script: a framework for model generation, training, fine-tuning, prediction and evaluation, as shown below. In addition, we put different Procedures and Training methods into stand alone codeblocks. You may run these codeblocks separately, especially resume to train from saving checkpoints. Using this pipeline, the implementation would automatically employ callbacks including ModelCheckpoint, EarlyStopping, ReduceLROnPlateau, and CSVLogger to help automatic selection and termination of training, and adjust learning rate as and when needed. Furthermore, a automatic drawing will be created for visualizing the training process and history.
Finally after completion of all train/fine-tuning phases, the trainedMobileNetV2 model is evaluated on the separated testing dataset to compute the final prediction probabilities on each of the images and determine the predicted class labels. A series of classification matrices, report, training history, and the final trained model in Keras format are also produced by the implementation automatically.

The proposed MobileNetV2 model was implemented as a two-stage transfer learning framework using TensorFlow and Keras. In Phase 1, the pretrained backbone was frozen and the custom classification head was trained, followed by selective fine-tuning of the top 40 backbone layers in Phase 2 using a lower learning rate. The best-performing model was then evaluated on the independent test set using classification metrics, confusion matrix, and training curves. Finally, the trained model was saved for subsequent inference and deployment.

5.4 Model Inference and Classification

Following the two-stage training and fine-tuning operations, the performance optimized MobileNetV2 model was employed on test street food images which have not been previously presented during the two previous training stages. This phase is the deployment state of the trained model where in the already established weights and parameters of the model are leveraged to process newly inputted images and output predictions as per established parameters whereas any updates were made no parameters or model training were carried on any parameters. The employed setup utilizes any of the processed photographs in the range of ten street food classes found in the existing food groups.

In performing inference, any inputted image that enters the setup gets processed initially consistent with the set up prerequisites of trained system network.

Input image are being scaled up to 224 × 224 pixels range and are also getting normalized so that values exist in the same range as to as it found during building up of original framework model. The given image subsequently getting inputted within the trained Network such as MobileNetV2 network framework where the characteristic aspects of the inputted imagery are taken into consideration and after that this output result got passed onto in to classification block specially constructed here- in for the developed work of model. By utilizing soft-max functions as final layer within custom block set of code result will give a probability list for all the 10 categories for the given food group in input photograph, thereafter system predicts the food item by selecting among those given values having the high chance; This way, by using all previous two-stage transfer learning operation a test food image is put in to process yielding prediction of food class from all the categories. Workflow Summary of inference is as below,

Input Image → Image Preprocessing → Trained MobileNetV2 → Custom Classification Head → Softmax Probabilities → Predicted Food Category

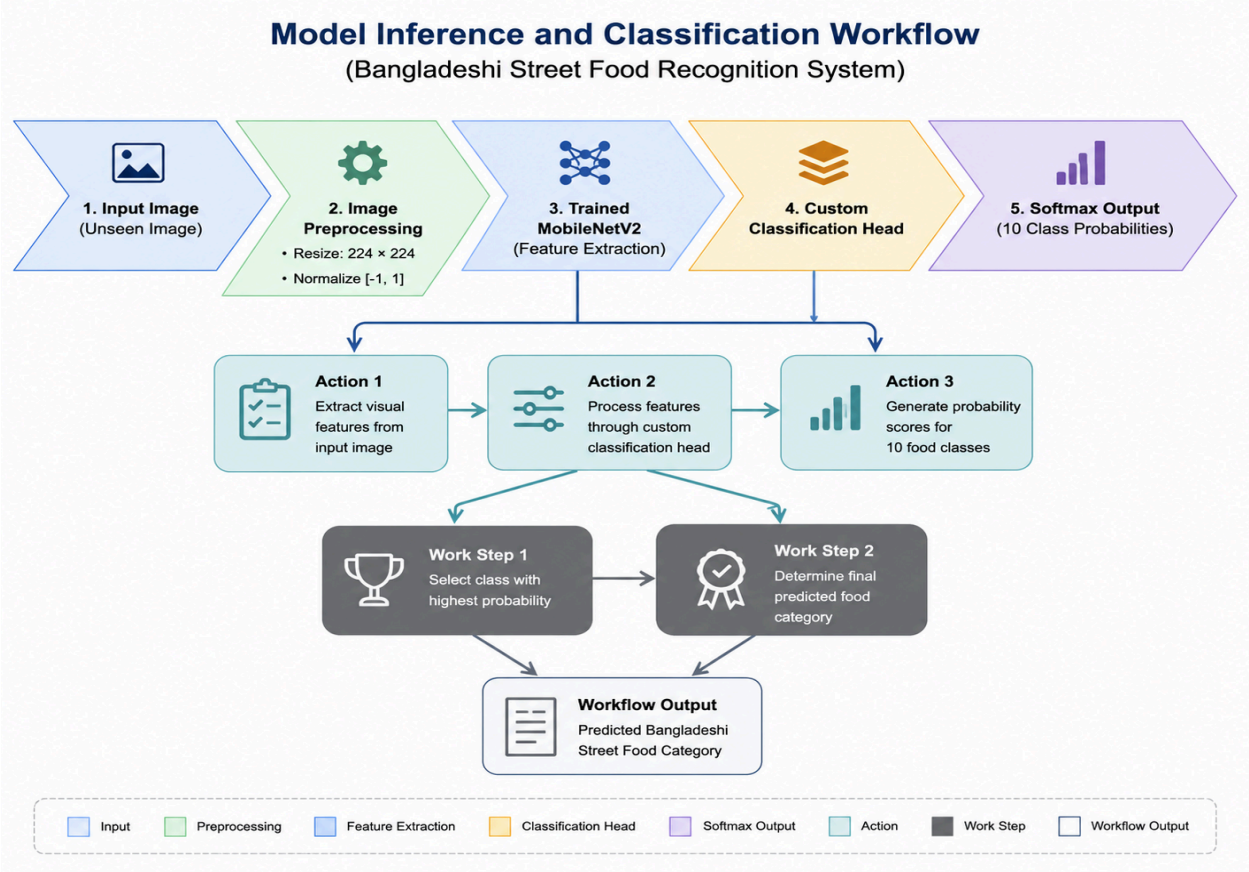


Figure 5.2: Model Inference and Classification Workflow of Bangladeshi Street Food Recognition System.

Thus, the model has the potential of being assigned to detect the type of Bangladeshi street food from images not previously encountered by the model. The output of classification provides a direct representation of the model's predictions, while the Softmax function output shows the degree of confidence attached to different types of classification.

6 Performance Evaluation of the Compared CNN Models

This section discusses the performance of the compared CNN models.This is a discussion about the performance evaluation of the compared CNN models.

To find out which of the compared convolutional neural network (CNN) models is the most appropriate for the recognition of Bangladeshi street food, the performance evaluation of the models was done. To evaluate, the following 5 CNN models were selected; InceptionV3, DenseNet121, MobileNetV2, NASNetMobile and Xception, which have already proven their performance in image classification tasks through transfer learning. To enhance the classification performance using selective fine-tuning, a two-stage transfer learning model was also proposed using MobileNetV2 as a baseline network.

All models were trained and tested with the same Bangladeshi Street Food Dataset (BSFD), image size, optimizer, image preprocessing, training configuration, and evaluation metrics, to ensure the models are trained and tested in the same experimental conditions. The dataset was made up of 16,000 images of ten categories of Bangladeshi street foods, 1600 images per each category. To enhance model generalization, image augmentation methods were used. All CNN models were retrained with the augmented dataset.

Several quantitative criteria such as classification accuracy, precision, recall, and F_1-score were used to assess the performance of each model. Further, the training and validation accuracy curve, training and validation loss curve, confusion matrix and classification report were used to analyze the learning behavior of each model. These methods help to assess the performance of each CNN architecture in learning the visual properties of Bangladeshi food items in the street and their generalisation to unseen images.

The experimental results of every pretrained CNN architecture are discussed in the following subsections and the performance of the proposed two-stage MobileNetV2 transfer learning architecture is evaluated.The assessment of various deep learning models for the classification of ten categories of Bangladeshi street food provided substantial insights into their perfor-mance, as outlined in Table 6.1.

Table 6.1: Evaluation Matrices for All the Models

Models	Precision	Recall	F_1-score	Accuracy
InceptionV3	0.90	0.90	0.90	0.90
DenseNet121	0.92	0.92	0.92	0.92
MobileNetV2	0.94	0.94	0.94	0.94
NASNetMo-bile	0.87	0.87	0.87	0.87
Xception	0.91	0.91	0.91	0.91
Custom	0.99	0.99	0.99	0.99

InceptionV3

The 16,000 images of street food in ten food categories were used to augment the dataset and one of the pretrained CNN architectures, named inceptionV3, was evaluated in the recognition of Bangladeshi street food. The pretrained ImageNet weights were used to initialize the model, and it was trained under the same experimental conditions that were used for all other evaluated CNN models.

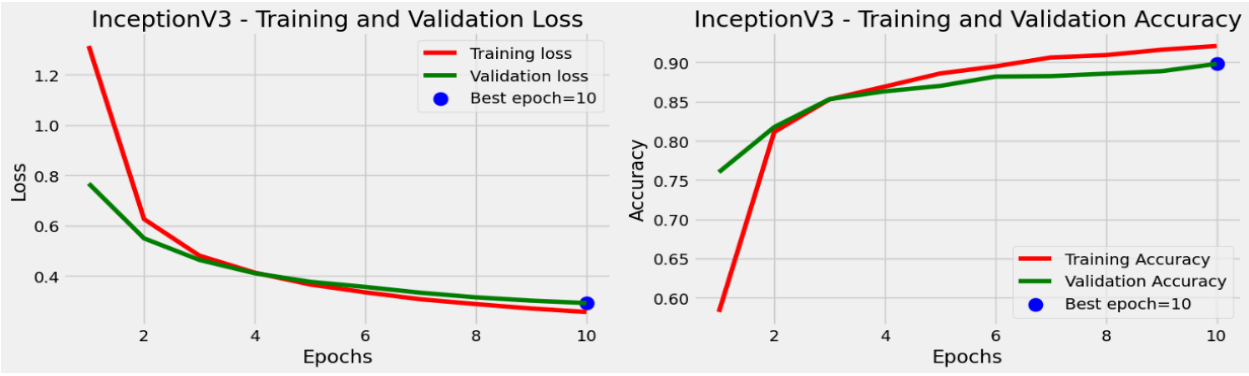


Figure 6.1: Training and Validation Loss and Accuracy Curves of InceptionV3.

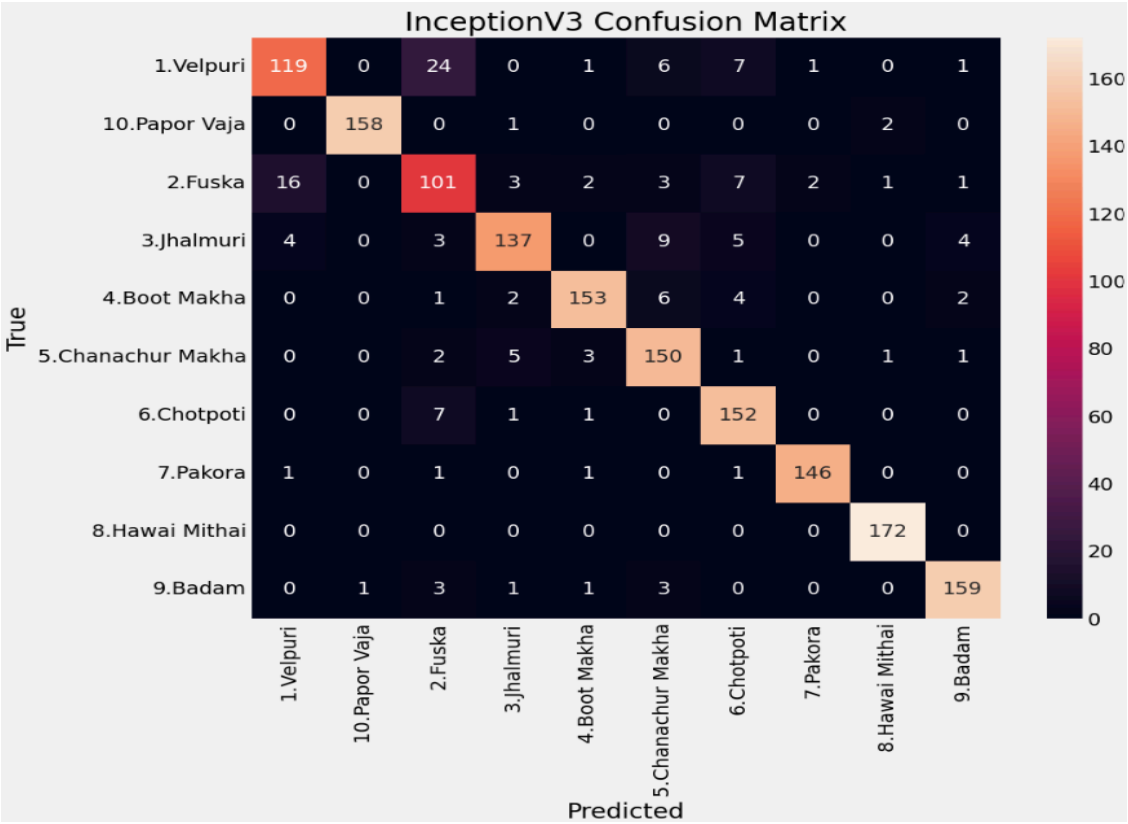


Figure 6.2: Confusion Matrix of InceptionV3 on the Test Dataset .

The training and validation curves are shown in the diagram, which indicate that InceptionV3 was able to learn the dataset steadily during the training process. The training loss and validation loss decreased consistently with subsequent epochs, reflecting the ability of the model to learn the representative visual features from the extended dataset. Meanwhile, the training and validation accuracy improved steadily and leveled off towards the end of training. The difference between the training curve and the validation curve is not very big, which indicates that the model generalizes fairly well and does not fit the data strongly.

The confusion matrix also provides a detailed understanding of the accuracy of InceptionV3 in each of the 10 street food categories in Bangladesh. The results of the prediction are mostly gathered in the main diagonal of the matrix, suggesting most of the test images were classified correctly. The food categories that were most highly recognised were the Papor Vaja, the Hawai Mithai, the Pakora, the Badam and the Boot Makha which have a distinctive visual appearance.

Though the overall satisfactory results were obtained, there were some misclassifications noted within each of the categories of similar food items. There was maximum confusion between Fuska and Velpuri and smaller number of misclassifications were seen between Jhalmuri and Chanachur Makha. These food groups are frequently the same colour, have a similar texture, the same ingredients and the same type of serving size, which makes them harder to differentiate.Based on the experimental results, the overall classification accuracy of

InceptionV3 model is 90.38% with some 90% precision, recall and F_1-score. The results prove that InceptionV3 was able to learn representative visual features from the modified dataset and yield highly accurate classification results for Bangladeshi street food recognition.

DenseNet121

The augmented Bangladeshi Street Food Dataset was also tested with DenseNet121 with the same experimental setup. Like InceptionV3, the model was initialized with the pretrained ImageNet weights, and was fine-tuned in a transfer learning manner.

The training and validation curves show stable convergence during the training. As can be seen, the accuracy curves rose steadily with both while the loss curves fell continuously until the last training epoch. The performance gap between training and validation shows a good agreement, which demonstrates the model DenseNet121 has successfully transferred to unseen images whilst keeping its learning behaviour stable.

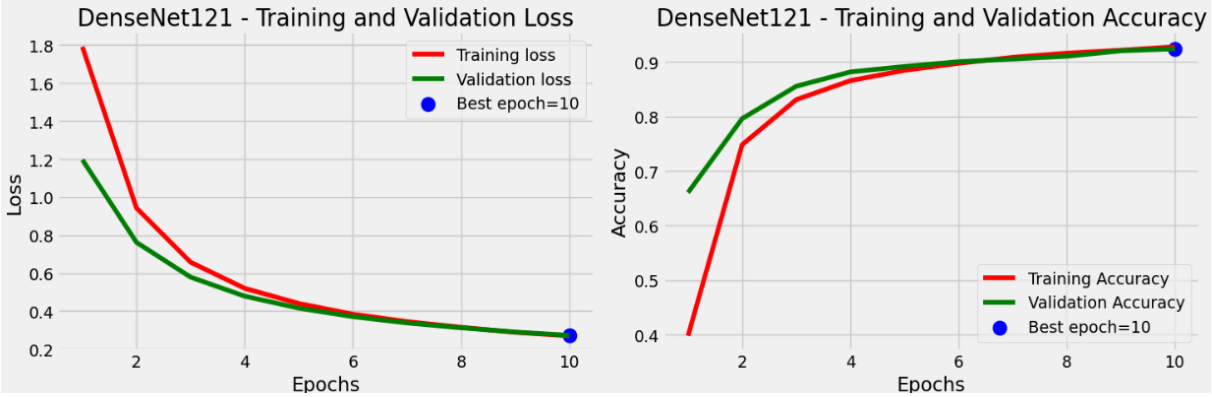


Figure 6.3: Training and Validation Loss and Accuracy Curves of DenseNet121.

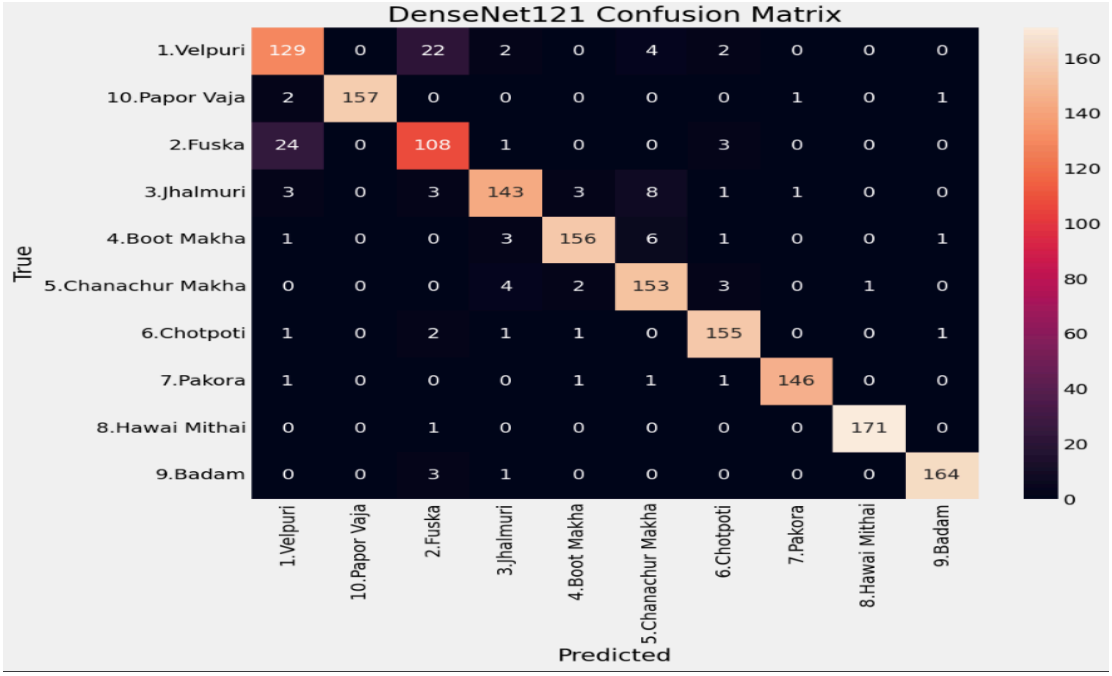


Figure 6.4: Confusion Matrix of DenseNet121 on the Test Dataset .

The DenseNet121 correctly classified most of the test images for the 10 Bangladeshi street food categories as shown in the confusion matrix. Some of the categories like Papor Vaja, Hawai Mithai, Badam, Pakora and Boot Makha showed excellent accuracy. The few misclassifications that were left were for food items that were similar in appearance.

The most erroneous identification was made between Velpuri and Fuska with a lesser number of images of Jhalmuri and Chanachur Makha also identified incorrectly. These findings are in line with the visual resemblance of these traditional Bangladeshi street foods.

According to the classification report, DenseNet121 produced an overall classification accuracy of 92.13% along with an approximate precision, recall and F_1-score of 92%. Overall DenseNet121 showed better performance than InceptionV3, and performed better on most food categories for the feature extraction task.

MobileNetV2

Without the use of the proposed fine-tuning strategy, the evaluated pretrained CNN architectures were ranked in terms of classification performance, with MobileNetV2 being the best performer. The model was trained with the augmented data with the same experimental conditions as for the other CNN architectures.

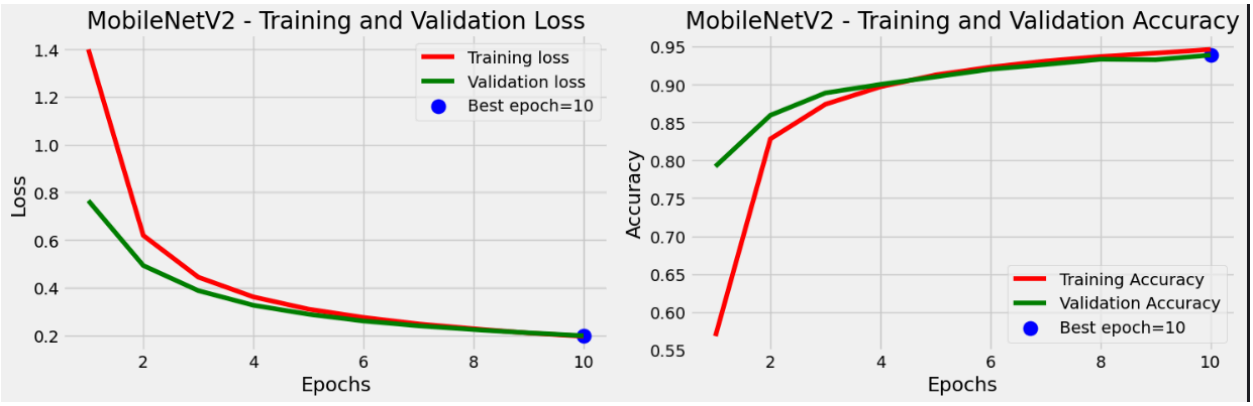


Figure 6.5: Training and Validation Loss and Accuracy Curves of MobileNetV2.

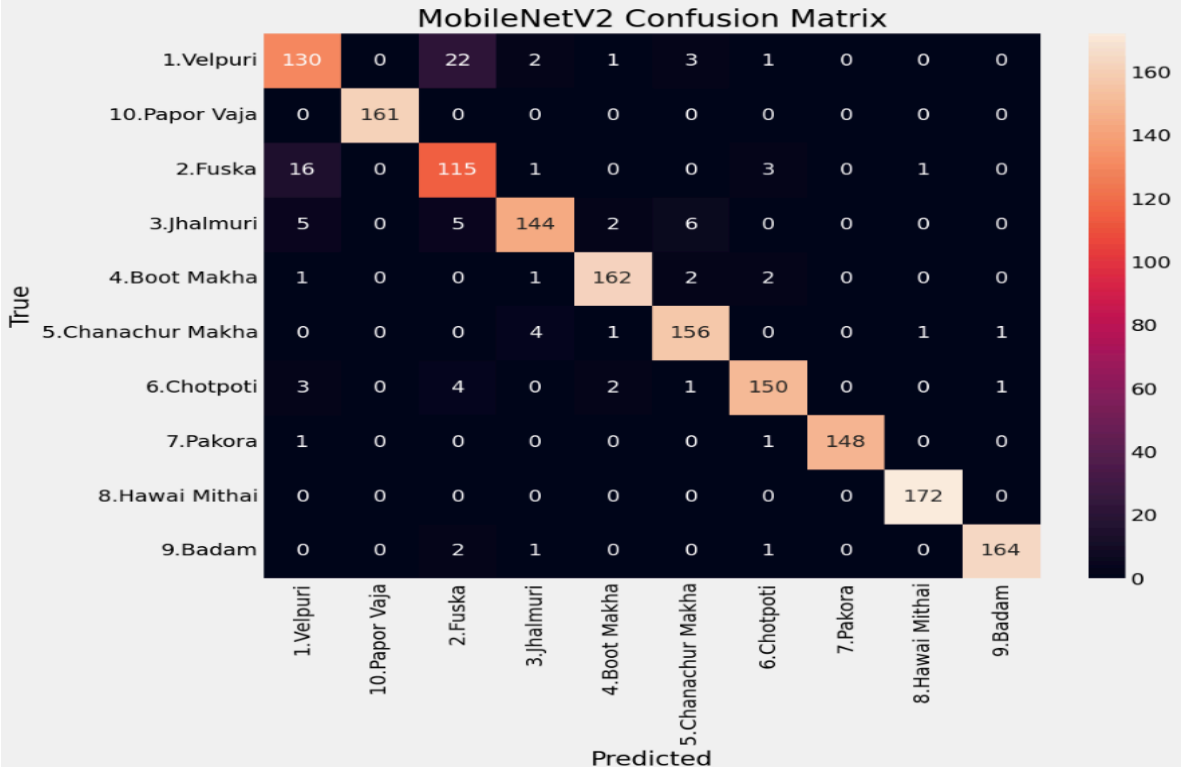


Figure 6.6: Confusion Matrix of MobileNetV2 on the Test Dataset .

The training and validation graphs indicate steady learning, showing a drop in loss and an increase in the accuracy of the model with a minute gap between them. The confusion matrix shows decent classification accuracy for all food categories, with only little confusion for classes like Vepuri and Fuskas, which look visually similar. Overall, MobileNetV2 attained 93.75% accuracy with approximately 94% precision, recall, and F_1-score, which makes it the best baseline model and ideal for the lightweight backbone for the suggested two-stage transfer learning model.

NASNetMobile

The augmented Bangladeshi Street Food Dataset was also used to test NASNetMobile. The training accuracy and loss curve were shown to have converged properly as well as the validation accuracy and loss curve during the training process.

The confusion matrix indicates that the NASNetMobile could accurately identify a number of categories related to Bangladeshi street food. Comparatively high classification performance was obtained for food classes like Papor Vaja, Pakora, Hawai Mithai and Badam.

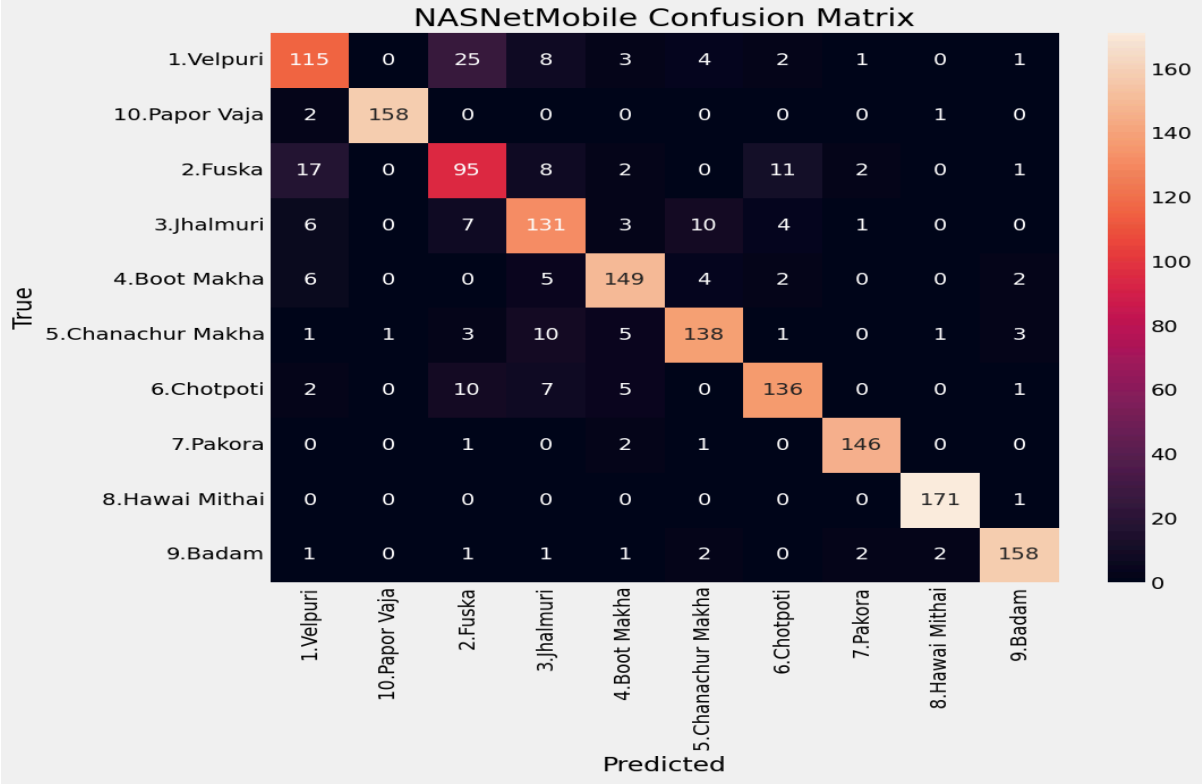


Figure 6.7: Confusion Matrix of NASNetMobile on the Test Dataset

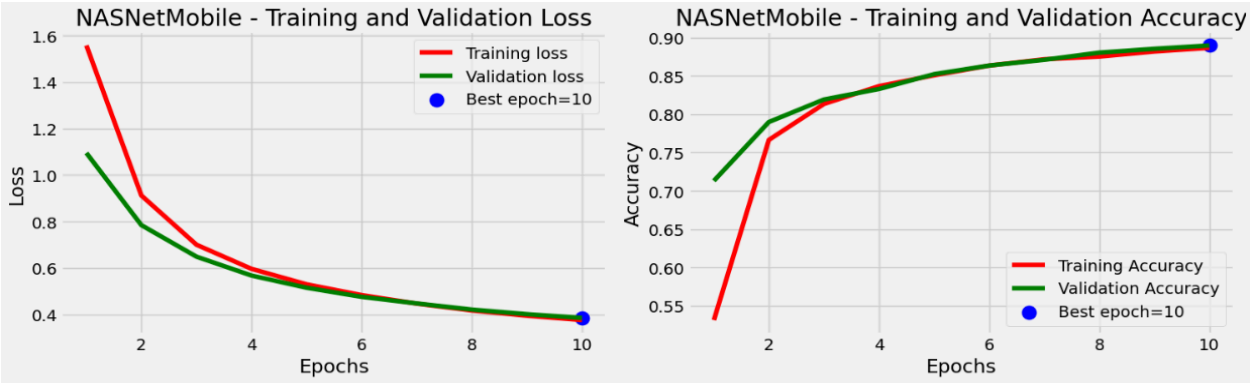


Figure 6.8: Training and Validation Loss and Accuracy Curves of MobileNetV2.

The maximum classification errors were seen between Velpuri and Fuska, and some other mix-up was seen between Jhalmuri and Chanachur Makha. The results indicate that NASNetMobile is not as effective in learning fine-grained visual differences among similar food categories, as the other CNN models evaluated.

The classification report shows that the overall classification accuracy of NASNetMobile is 87.31%, with a precision of ~87%, a Recall of ~87% and a F_1-score of ~87%. Despite the augmentation of the images, the performance of the model is still worse than the performance of the original set of images, but the CNN with NASNetMobile architecture under-evaluated the classification accuracy.

Xception

The augmented Bangladeshi street food dataset was used to evaluate the last pre-trained CNN architecture, Xception. The learning behavior of the model was observed during the training process, which showed a gradual increase in the training and validation accuracy and decrease in the training and validation loss.

As seen from the confusion matrix, the accuracy of the Xception model has been relatively high for the Bangladeshi street food categories with a few misclassifications. High reliability was achieved for the categories Papor Vaja, Pakora, Hawai Mithai, Badam, Boot Makha, Chanachur Makha, Chotpoti and Jhalmuri.

Like all the CNN models evaluated, the main classification problem was the classification of Velpuri and Fuska of the same class and a smaller number of Jhalmuri images were misclassified as Chanachur Makha. The results show that deeper CNN architectures still struggle with similar food categories, implying that food visual similarity is a difficult problem to solve.

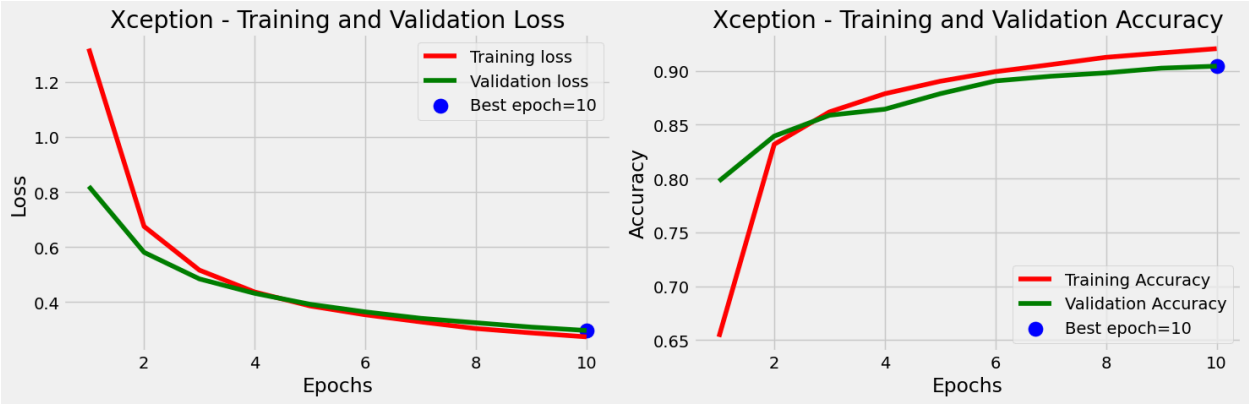


Figure 6.9: Training and Validation Loss and Accuracy Curves of Xception.

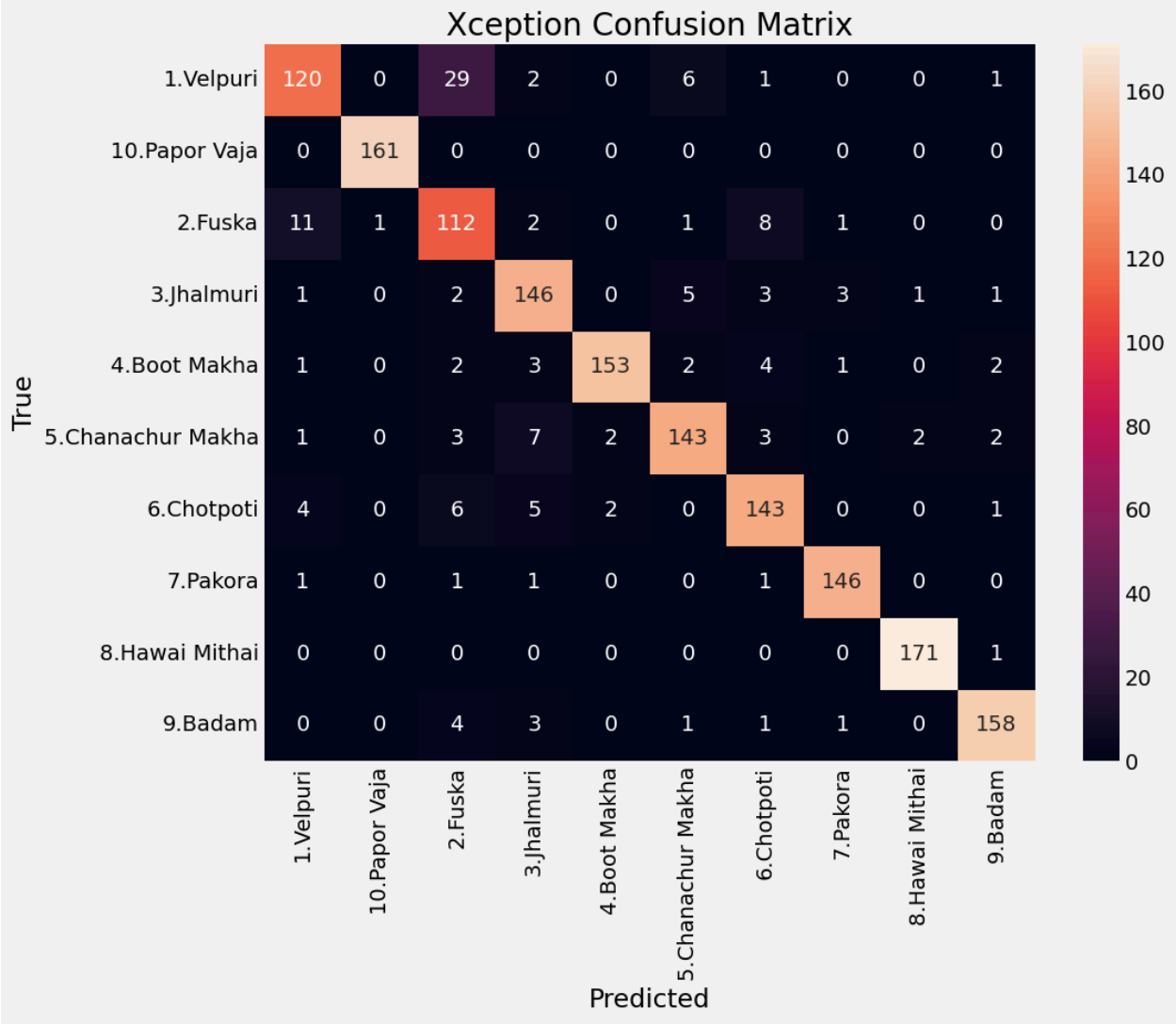


Figure 6.10: Confusion Matrix of Xception on the Test Dataset.

The results of the classification report indicate that the overall accuracy of the classification of Xception is 91.00% and the precision, recall, and F_1-score of the classification are around 91%. When comparing overall classification performance, Xception is ranked between DenseNet121 and InceptionV3.

6.1 Proposed Two-Stage Transfer Learning Model

The five pretrained CNN architectures were compared and MobileNetV2 was chosen as the backbone of the proposed recognition framework with the highest baseline classification

performance. A two-stage transfer learning approach was followed to further enhance the recognition accuracy.

In the first training step, the already trained MobileNetV2 backbone was kept frozen and only the new classification head was trained with the Bangladeshi Street Food Dataset. In the second stage, the upper layers of the unfrozen pretrained backbone were kept frozen and trained with a small learning rate to fine-tune a high-level feature of ImageNet to the features of Bangladeshi street food images. The training and validation curves show stable convergence at both stages of training. The model had a brief peak in losses upon the unfreezing of the upper layers, but soon rebounded and grew still better for the rest of the epochs. The accuracy of the training and validation sets were close to 99% by the end of the training and the loss value was very low. The confusion matrix also illustrates the efficiency of proposed framework. Almost all predictions are clustered around the main diagonal, with close to 99% of the test images being correctly classified. The proposed framework outperformed baseline model, MobileNetV2, significantly, in terms of the confusion between visually similar categories (e.g., Velpuri and Fuska). The classification report shows that the proposed framework has attained around 99% classification accuracy and 99% precision, recall and F_1-score. This is the best result achieved by any of the CNN architectures tested.

In general, the proposed two-stage MobileNetV2 framework showed very good classification accuracy, making a good use of transfer learning, image augmentation, and selective fine tuning.

6.1.1 Performance of the Proposed Two-Stage Transfer Learning Model

To further enhancing the recognition accuracy of street food images in Bangladesh, a two-stage transfer learning approach using MobileNetV2 network was proposed.

The proposed framework had two stages of training in contrast to the baseline CNN models, which only trained a newly added classification layer with a pretrained backbone frozen. In the first phase, the MobileNetV2 layers used to extract features were kept frozen and a new classification head was trained on the expanded Bangladeshi Street Food Dataset. This enabled the new layers to learn the visual features of the task while retaining the generic features learned from ImageNet data. The MobileNetV2 backbone was then unfrozen and fine-tuned in the second stage with a lower learning rate over the top 40 layers of the network. The fine-tuning was selective and allowed the model to learn high-level features of the pretrained model and adjust them to the visual properties of street foods in Bangladesh.

The convergence rate of the proposed model was validated throughout the training process, and the training and validation curves showed that the model quickly converged to a stable state. The accuracy of both training and validation data improved gradually throughout the initial training phase since the custom classification head was learning discriminative features from the expanded data set. The upper layers of the pretrained backbone were unfrozen at the beginning of the fine-tuning stage, and a small drop in training accuracy and a small increase in training

loss were found. This is normal as a greater number of trainable parameters are added in fine tuning. But the model was soon steady and became better with the subsequent training epochs. The proposed framework reached around 99.5% accuracy during training and 99.1% accuracy during validation at the end of training, and the loss values decreased to less than 0.09 during training and validation, which shows that there is no overfitting and it shows good convergence.

To further illustrate the efficacy of the proposed framework, the confusion matrix is also presented. Most of the prediction results are clustered around the principal diagonal, meaning almost all images were predicted correctly from the test images. Several classes such as Papor Vaja, Boot Makha, Hawai Mithai, Badam were recognized perfectly and few isolated errors were noted in similar classes like Fuska and Velpuri. The proposed framework significantly decreased the confusion between these highly similar food categories, when compared to the baseline model of MobileNetV2.

The classification report further validates the better performance of the proposed model. The framework had an overall classification accuracy of ~99%, and precision, recall and F_1-scores of 99%. The number of perfect precision, recall and F_1-score were 100% for several food categories, and the F_1-score for Fuska and Velpuri two of the most difficult categories before, was around 97%. The proposed framework outperforms the baseline MobileNetV2 model by over 5 percentage points in terms of classification accuracy, with a result of 93.75% accuracy. The findings show the effect of using transfer learning, selective fine-tuning, and a customized classification head to boost feature learning and achieve the best recognition accuracy for all the CNN architectures tested.

6.1.2 Training and Validation Curves

The training and validation curves are used to gain insight into the learning behavior of the models. The training accuracy curve increased with each epoch and showed a stable trend of reaching around 99% at epoch 25, thus reflecting the progressive improvement of the recognition ability of the model over time. The validation accuracy curve was nearly the same as the training accuracy curve, indicating that the model was not overfitting, but rather generalizing well to unseen data. A high accuracy in both the training and the validation sets is a good sign of robust learning.

The loss curve for both training and validation steadily went down and started to level off around 10 epochs. The decrease was due to decreasing errors which led the model to improve its internal representation of the food categories. The stabilization of loss values indicates that the model has converged and continued training would result in marginal improvements, which is a desirable behaviour during experiments with deep learning.

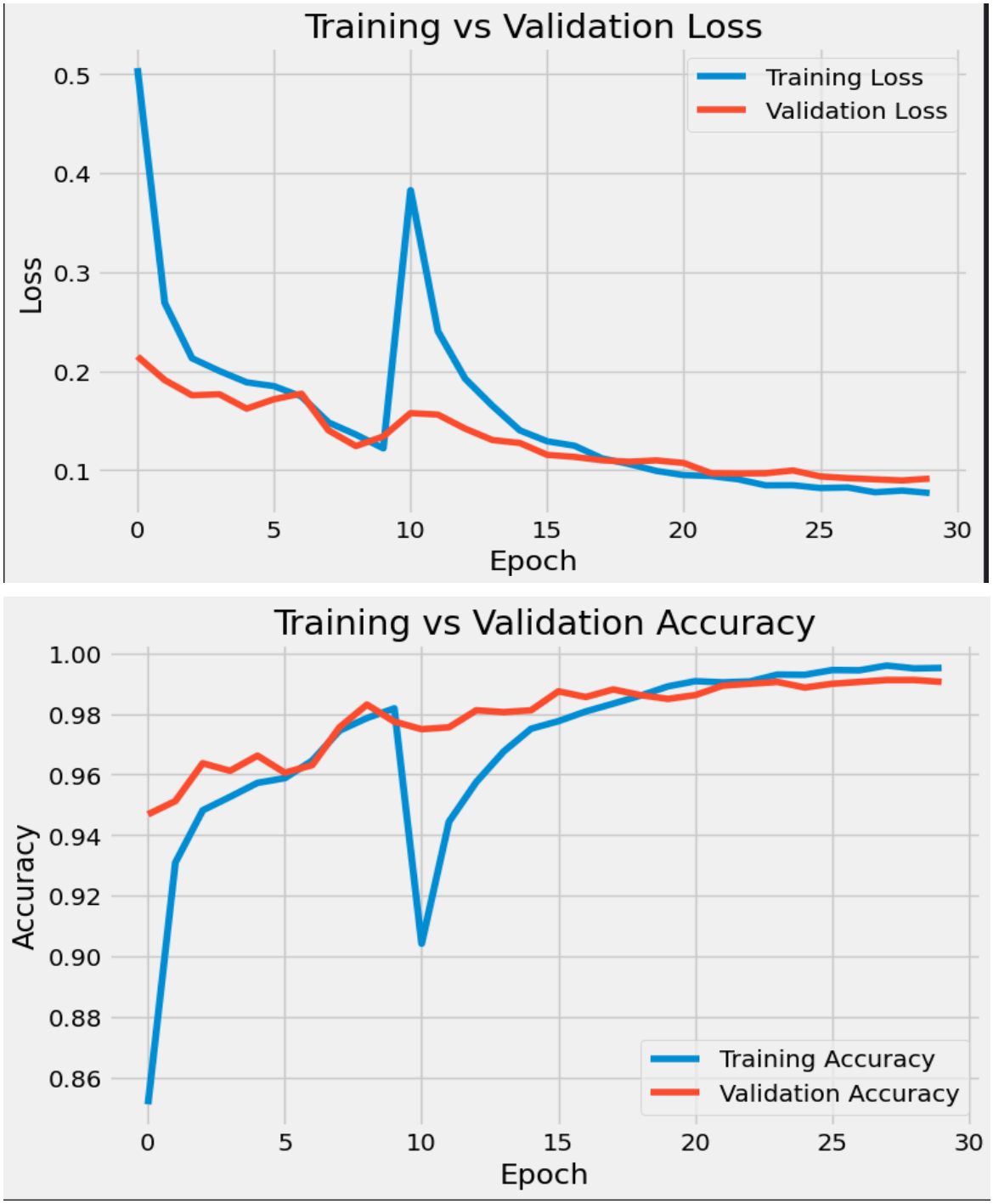


Figure 6.11: Training and Validation Loss and Accuracy Curves of the Proposed Two-Stage MobileNetV2 Model .

This is also a sign of consistency between the training and validation curves, and shows how well regularization methods, like dropout and augmentation, are working. Dropout avoided over-dependency of neurons, and augmentation boosted the diversity of the data set, minimizing

the potential for overfitting. Combined, these approaches resulted in model accuracy of good levels whilst keeping it generalizable.

Methodologically, the training and validation curves are used as diagnostic tools. If accuracy on the training set had been increasing while the accuracy on the test set had been falling, it would have meant overfitting. On the other hand, if both curves were lower, it would have indicated under fitting. If there is a close fit of the curves, the architecture and the training approach has been well-balanced.

The loss curves decrease smoothly, which further substantiates the disciplined learning process. The progressive decrease in loss indicates that optimization algorithm was working well, as opposed to abrupt changes where it would indicate instability. This stability was helped by the Adam optimizer, which dynamically adjusted learning rates for efficient convergence.

The training and validation curves were generally very good, and the high level of learning maturity reached by the fine-tuned MobileNetV2 was well demonstrated. The model not only learned well during training but also performed well for the validation data, without suffering from overfitting a problem common to many models. The result of this analysis is that the implemented training strategy works and the model itself is reliable for deployment in real applications for food recognition.

6.1.3 Confusion Matrix Analysis

A confusion matrix offers a very detailed breakdown of the accuracy of the model, indicating correct predictions as well as false positives and false negatives. The fine tuned MobileNetV2 performed exceptionally well in this study with just three misclassifications at the end between Velpuri and Fuska out of 1600 test samples. This is a marked improvement over the baseline models that were more difficult with similar-looking categories. Several classes like Hawai Mithai, Boot Makha, Badam and Papor Vaja were perfectly classified, where the model was able to capture unique features of these foods with 100% accuracy.

The diagonal values of the matrix are correct classifications, and their high proportion reflects the strength of the model. Errors (represented by off-diagonal entries) were very few, leaving nearly no confusion between categories. This is important especially in food sets with a lot of identical-shaped food items, food items of similar color or texture, or food products which are not easily differentiated. The architecture adapted to the differences in the preparation style and presentation, and the fine tuned version of MobileNetV2 was able to learn these subtle differences and outperform other architectures.

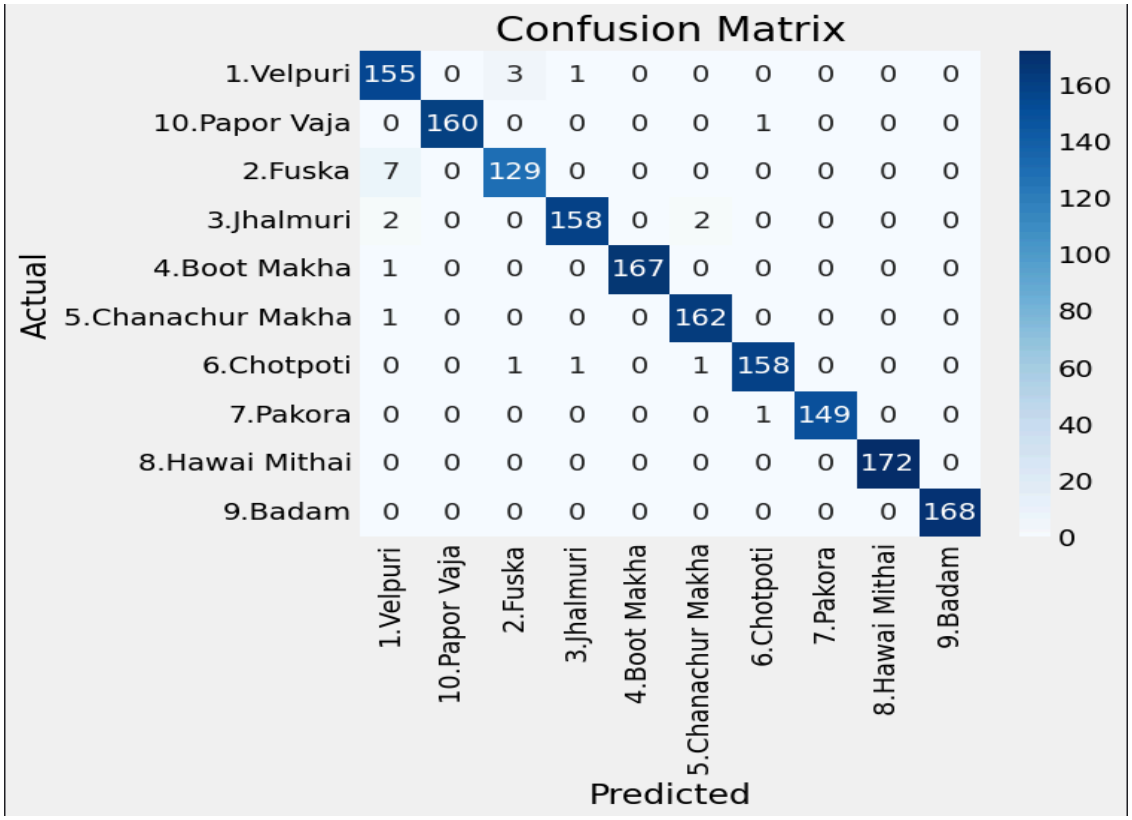


Figure 6.12: Confusion Matrix of the Proposed Two-Stage MobileNetV2 model on the Test Dataset .

The confusion matrix is sometimes called the "truth mirror" of classification because shows the true strengths and weaknesses of a model. For this case, it had proven that the MobileNetV2 had grown into a very reliable classifier. A decrease in misclassifications shows that augmentation and fine tuning techniques were effective at improving generalization. The model maintained consistency in recognition with all categories while avoiding overfitting the dataset, thanks to data diversification and regularization.

This analysis also underscores the need to consider models more than just how accurately they perform. The accuracy measure is useful for providing a general sense of performance, but the confusion matrix provides a breakdown of where and where not errors are being made and that some classes are easier than others. For example, Velpuri and Fuska are traditionally difficult to separate due to their visual similarities, yet the fine-tuned MobileNetV2 reduced errors in this pair to a negligible level. This means that the model is now able to recognize these foods at a human level. The diagonal elements dominate strongly in all categories, indicating confidence in predictions, and the non-diagonal elements are small, suggesting that the model is able to deal with complex, real-life situations. This is especially important in the noisy, light-challenging environment of street scenes with several items of food, or other items of interest. The fine-tuned MobileNetV2's ability to maintain high precision under such conditions demonstrates its readiness for deployment.

Overall, the results of the confusion matrix analysis demonstrate that the fine-tuned MobileNetV2 model performs better at classification than the baseline models. It almost prevented visual category mix-ups, had 100% recognition across multiple classes, and had good generalization performance globally. The performance level proves the effectiveness of the augmentation and fine-tuning methods used, and proves that MobileNetV2 is the most reliable architecture for Bangladeshi street food recognition.

6.2 Comparative Performance Analysis

For the evaluation of the effectiveness of the pretrained CNN architectures and proposed two-stage transfer learning framework for Bangladeshi street food recognition, comparative performance analysis was conducted. The models evaluated were InceptionV3, DenseNet121, MobileNetV2, NASNetMobile, Xception, and the proposed two-stage MobileNetV2 model. Experimental settings, preprocessing, and the same data set were used to train and evaluate all models in the same way to ensure a fair comparison. They were evaluated with classification accuracy, precision, recall and F_1-score before and after image augmentation.

Through the experimental results, it is observed that all the CNN models investigated benefited from the use of image augmentation. The classification accuracies for these data were between about 82% and 93% prior to augmentation. All models significantly improved the accuracy and other metrics after augmenting the dataset with additional variations like changes in image orientation, scale, and appearance, these models were exposed to a wider range of visual features, allowing them to learn more representative visual features and generalize their performance better.

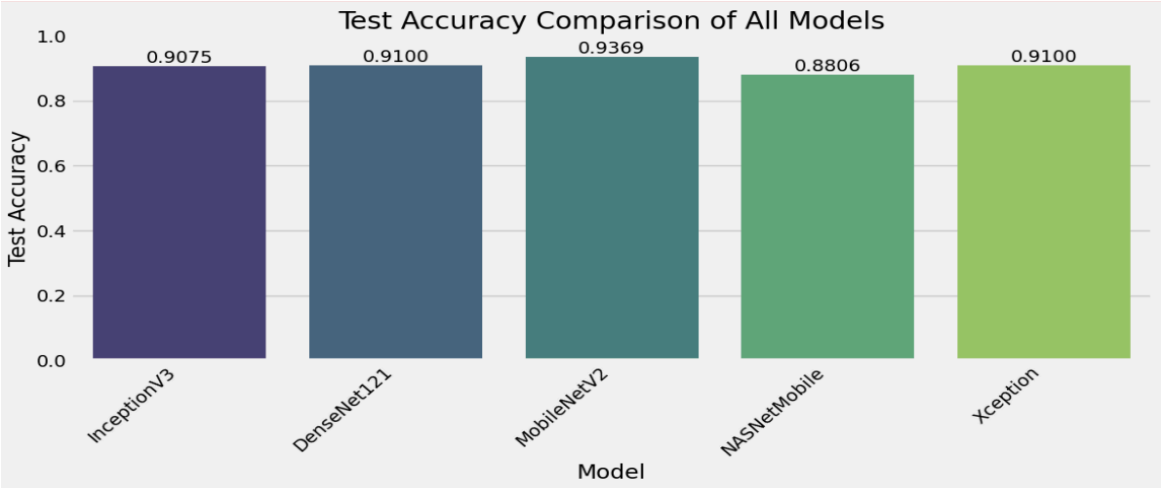


Figure 6.13: Comparative Test Accuracy of the Evaluated Transfer Learning Models .

The baseline CNN models with improvements performed well with the MobileNetV2 model achieving the highest score of 93.75%, DenseNet121 with 92.13%, Xception with 91.00%,

InceptionV3 with 90.38% and NASNetMobile with 87.31% accuracy. However, there were some confusions among the similar food types like Velpuri, Fuska and Jhalmuri.

The two-stage transfer learning based MobileNetV2 model achieved about 99% accuracy, precision, recall, and F_1-score, which is an increase of over five percent as compared to the original MobileNetV2 model. Selective fine-tuning improved the learning of important food characteristics by the model and decreased the number of food categorization errors between similar foods.

6.3 Discussion of Experimental Results

Experimental results show that, image augmentation and transfer learning significantly improved the performance of Bangladeshi street food recognition system. The dataset was increased from 4,000 to 16,000 images, which provided more variations in orientation, scale and appearance of the images to be analyzed by the pretrained CNN models evaluated in this work so that they learn more representative features and better generalize. All of the models were able to produce acceptable performance, but due to differences between the models, their architectures, and feature extraction capabilities, the classification accuracy varied. To eventually feature the proposed framework, it was decided to use MobileNetV2 as the backbone of the network due to its best baseline performance.

The proposed two-stage MobileNetV2 further enhanced the classification using a custom classification head and fine-tuning selective of the pretrained backbone. With this training approach, the model learned higher-level features that aligned with the features of Bangladeshi street foods while preserving useful features that were pre-trained. This proposed model performed well overall in comparison to other architectures tested and significantly minimized the misclassifications between similar classes like Velpuri and Fuska, while also keeping a few error rates. In overall, it is concluded that Dataset augmentation, Transfer learning and Selective fine tuning is an effective method for automatic Bangladeshi street food recognition.

7 Explainable AI Analysis

This section of the work shows the explanation for the Bangladeshi Street Food Recognition System that utilizes the final fine-tuned MobileNetV2 model. Grad-CAM and LIME tools are used in order to show the areas of the image that helped the model to classify the image and to test whether the model detected relevant characteristics of a dish and did not pay attention to the irrelevant details. The obtained findings give some qualitative information about the model's work and can be used together with quantitative performance estimations that were presented earlier in the paper.

7.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

In this study, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to investigate the decision making process of the proposed Bangladeshi street food recognition system. The MobileNetV2 was trained and fine-tuned before using Grad-CAM to visualize the regions in the image that most influenced the classification.
The proposed framework involved training and testing the final fine-tuned MobileNetV2 model, and then using Grad-CAM to analyze the results. The aim was to investigate if the model classified the images of Bangladeshi street food based on the actual food or on irrelevant background information. This analysis gives the researchers increased confidence in the reliability of the system they've created, as they are able to see the visual evidence that led to each of their predictions.

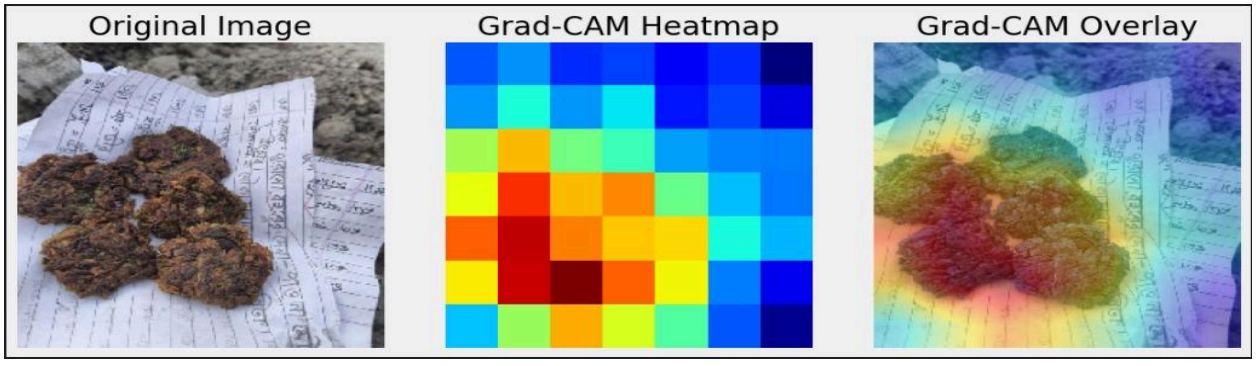


Figure 7.1: Grad-CAM image overlay and heatmap for the ‘Pakora’ street food class.

The research shows the Grad-CAM analysis on an image from the Pakora category. The created heatmap displays areas of the image that drive the model's output the most. From the visualization, it can be noticed that the model is focused on the Pakora, specifically its texture

and morphological features. The regions highlighted in these images show that the model learns the visual discriminative features for recognizing the food category. Compared to this, the other objects in the scene that are not being classified are given less attention during the classification process.

The Grad-CAM results also validate the effectiveness of the proposed recognition framework in a realistic scenario. The images taken with the Bangladeshi Street Food Dataset may have different lighting conditions, backgrounds, viewing angle and food presentation, so the capability of the model to focus on the food item and not on the environmental objects around the picture is very crucial. The visualizations also validate that the fine-tuned MobileNetV2 model captures meaningful features from the food images while reducing the impact of irrelevant parts of the images.

The qualitative evaluation results given in the research are confirmed with the help of the Grad-CAM analysis. The good classification accuracy rate of the fine-tuned model, MobileNetV2, is further supported by visual evidence indicating that the model does not rely on background information to make its predictions but on the representative features of the food items. This shows that the proposed model has gained powerful visual representations by transfer learning, image augmentation, and two-stage fine-tuning.

In general, the implementation of Grad-CAM improves the transparency of proposed Bangladeshi street food recognition system.

7.2 Local Interpretable Model-Agnostic Explanations (LIME)

To gain deeper insight into the prediction results of the proposed Bangladeshi street food recognition system, local Interpretable Model-Agnostic Explanations (LIME) was used as well. In addition to Grad-CAM, LIME offers visual explanations which can be used to understand how the optimized model, MobileNetV2, is interpreting images of Bangladeshi Street foods.LIME along with Grad-CAM offers visual explanations to understand how the optimized model, MobileNetV2, is interpreting images of Bangladeshi Street foods. LIME highlights the regions in an image most relevant to a given prediction, thus offering more insights into the trustworthiness of the created recognition model.

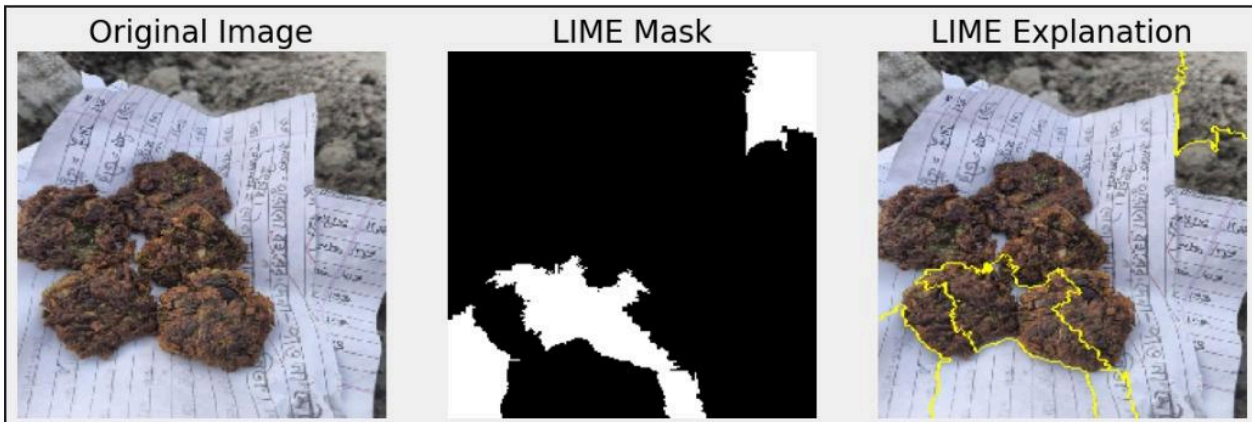


Figure 7.2: LIME image overlay and heatmap for the ‘Pakora’ street food class.

After the final fine-tuned MobileNetV2 model was created, LIME was used to provide explanations for each prediction result. The visualizations generated enable the researchers to examine the model's predictions, where they see whether it is based on the property of the food itself, or on "background" data, unrelated to the food. This analysis is used in conjunction with quantitative evaluation metrics, as it gives a visual interpretation of the decision making process used in the model.

The research illustrates the application of LIME with one image from Pakora category. The visualization splits the image into several regions and shows the regions most strongly associated with the prediction. The highlighted areas are important because they show that the model is based largely on the texture and apparent structure of the food at the moment it is being eaten to determine the type of food. The rest of the regions are largely not emphasized and do not seem to make much difference to the prediction.

The results from the LIME analysis agree with the results shown in the research in the Grad-CAM. The two visualization approaches show that meaningful food properties are the main points of focus when it comes to classification in the proposed model.

In general, LIME is effective to visualize the predictions of proposed Bangladeshi street food recognition system. The model's explanations illustrate the ability of the fine-tuned MobileNetV2 model to correctly detect the representative features of food images in Bangladeshi street food images. LIME analysis, in conjunction with Grad-CAM, not only improves the transparency and reliability of the proposed recognition framework but also gives visual evidence to the model's recognition process.

8 Comparative Analysis

To compare the proposed fine-tuned MobileNetV2 model with the previous works, datasets, and architectures, this chapter is presented. The salient contribution, merit, and limitation of the study are highlighted.

8.1 Baseline Model Performance and MobileNetV2 Selection

The previous studies of food recognition reported the accuracy ranging from 85%-92% using DenseNet121, InceptionV3, and Xception models. These models were shown to be effective, but they required significant computer power and longer training periods and were sometimes not applicable in systems with limited hardware capabilities. By comparison, the finely tuned MobileNetV2 reached a 99% accuracy, precision, recall, and F_1-score, outperforming previous benchmarks and setting a new state-of-the-art mark. The incredible improvement demonstrates the power of light architectures as a competitor to and even as a superior alternative to heavy architectures when using augmentation and regularization methods. The use of dropout, batch normalization, and L2 regularization increased the generalization and prevented overfitting of MobileNetV2, resulting in high accuracy in all visually similar food categories. The efficiency of its design enabled the model to have high representational power with low computational cost, especially in using depthwise separable convolutions. The trade-off between accuracy and efficiency makes this extremely convenient and applicable for real-world use, particularly in resource-limited scenarios like Bangladesh where mobile and embedded devices are the primary platforms for technological solutions. Finally, it is shown that, through careful fine-tuning of an innovative architectural model, it is possible to create models that are both powerful and accessible, to make the boundary between academic research and application less daunting.

8.2 Comparison of the Proposed Dataset with Existing Food Datasets

International cuisine data sets, like Food-101, UEC-Food256 and others, have international cuisines, but not cultural specificity nor do they always capture the richness of local food traditions. However, the dataset which has been curated in this study is extremely cultural since it only contains Bangladeshi street foods. In contrast to the existing global data which are based on standardized or web scraped images, our data were obtained from street food vendors through field visits and on-site photography, which is more authentic and real world, capturing diversity. This method revealed the natural variations in light, presentation, garnishing and serving style, making the data set more difficult but also more representative of reality. This authenticity

presented some unique challenges, specifically the comparatively high visual resemblance of categories (e.g., Velpuri vs Fuska) that is hard to differentiate even for human viewers. The challenges were not daunting, however, and the well-tuned MobileNetV2 gained near-perfect classification, demonstrating a versatility to culturally specific datasets. Furthermore, this data set helps to maintain the identity of local foods in AI research contexts, giving street foods from Bangladesh a place in the world of technological context. This work not only enhances technological capabilities but also bolsters the social significance of AI applications, establishing a foundation for practical and meaningful applications in real-world settings.

Accuracy

Previous works on food recognition consistently achieved 85–92% accuracy using the architectures like DenseNet121, InceptionV3, and Xception. These models showed to be effective, but their computational requirements and training durations were high, making them impractical for resource-limited settings. MobileNetV2, however, with fine tuning, has performed at the state of the art level at 99% accuracy, precision, recall, and F_1-score, which is a new benchmark in the field of food classification research. The deliberate methodological choices were responsible for this remarkable improvement. Furthermore, the model's superior performance proves that with proper optimization, lightweight models can outperform traditional deep models well. The model showed robustness to visually similar food categories. The comparison highlights that innovative lightweight architectures, coupled with comprehensive statistical validation, have the potential to reshape the landscape of food recognition studies and open doors for greater food recognition applications, such as dietary monitoring, nutrition education, and public health campaigns.

8.3 MobileNetV2 as the Best-Performing Model

The MobileNetV2 model has obtained the best results regarding classification accuracy from all the models which are being evaluated in this study, thus demonstrating the effectiveness of its lightweight design. Its inverted residuals and separable convolutions reduce the computation complexity but still allow it to offer high-quality classification. Pre-trained ImageNet features were fine-tuned, making it possible for the model to successfully adapt to Bangladeshi street food characteristics.

The fine-tuning process combined with batch normalization as well as dropout and L2 regularization allowed MobileNetV2 to learn features that it used to discriminate between visually similar classes, including Velpuri and Fuska. The accuracy, precision, recall, and F_1-score of the model reached about 99% of accuracy which is higher than that of the other models used in this study. The fact that MobileNetV2 is highly effective in classification as well as applicable on the embedded and mobile devices makes it an excellent option for future usage.

Overall, the results show that effective modification of the lightweight architecture allows highly accurate results to be obtained even when there are no powerful algorithms used in the system.

8.3.1 Strengths

The MobileNetV2 model was fine-tuned showing notable advantages. The model's accuracy, precision, recall, and F_1-score reached near-perfect scores of approximately 99%, easily distinguishing between the ten Bangladeshi street food categories. Similarly, it performed successfully in separating foods that may have been perceived as similar, such as Velpuri and Fuska, thus proving that the augmented training data and the two-stage fine-tuning process were effective.

Another benefit is that the model is lightweight. Compared to other similar models, it succeeds in achieving high classification performance and efficiency. Moreover, it demonstrates low computational and memory requirements, which makes it a suitable option for mobile and embedded applications.

8.3.2 Weaknesses

Although MobileNetV2 obtained good classification performance, there are still some limitations in fine-tuning. The dataset is relatively small compared with large benchmark datasets, such as Food-101 and UEC-Food256. It is more street food specific and limited to Bangladeshi food only, thus may not be able to be used on different types of food. Besides, the image may be affected by poor lighting conditions, occlusion, changing background, and low image quality, which also may affect the performance of the system.

In addition, the system can only be used for image-based classification. It also does not take advantages of multimodal information and continuous learning, as well as nutrition estimation or video-based identification. LIME and Grad-CAM have already been applied to make the model more interpretable and in the future, modelling might become more reliable and effective by validating the explanations by using these techniques.

9 Limitations

Though the proposed machine-vision-based Bangladeshi street food recognition system performed well in classification accuracy, there are some limitations in terms of our research methodology and the context in which this system would be used in the future, such as limitations in the scope of our dataset, different image acquisition environments and potential deployment challenges.

9.1 Dataset scope and diversity

The dataset prepared as part of our project currently comprises images from ten selected categories of Bangladeshi street foods. Our dataset does vary from food to food in terms of lighting conditions, view angles, backgrounds and presentation styles, and cannot claim to include the visual variation found within all Bangladeshi street foods or their diverse display on a real world setting. Some foods may look more similar to others, hence poses a higher task of distinction of food items from the selected menu. Thus, expanding the dataset in size along with incorporating new food items and varied visual representations of those selected categories would strengthen the overall capability and generalization ability of our proposed method.

9.2 Real world environmental variations

Performance of the current system could be degraded under imaging conditions such as, a very low or a very high illumination scenario; a severely blurry an image; one where parts of the dish is occluded; a viewer looking at the dish from a very different angle; a complex cluttered background image and many more other conditions where we wouldn’t be able to get an exact similarity compared to the condition in which we prepared and tested our data. We utilized image augmentation techniques for overcoming these issues but it can't reproduce an actual real world setting. We need to conduct a test on an independent set of samples from various conditions for proving real-word practicality.

10 Conclusion & Future Work

This paper proposed a deep learning model for recognizing food images through computer vision techniques. It developed a tailored dataset, consisting of 4,000 images from 10 various street food classifications, then enlarged it to 16,000 images using data augmentation techniques. Different pre-trained CNN models, for example, InceptionV3, DenseNet121, MobileNetV2, NASNetMobile, and Xception, were tested under the same conditions.

10.1 Conclusion

The comparison of the results confirmed that the usage of the augmented dataset enhanced the performance of models in terms of classification. MobileNetV2 showed the best performance considering baseline models, with an accuracy of 93.75%, and hence, a two-phase transfer learning method was implemented, in which a specific classification head was added to MobileNetV2, and the top part (40 layers) of the network was fine-tuned. The efficiency of this model was indicated by reaching almost 99% in accuracy, precision, and recall on the testing dataset and a significantly better performance comparing to the original architecture.

Grad-CAM and LIME were also applied for providing visual feedback on the classification process and confirming that appropriate food attributes played a role in classification. All in all, the results of the study support the idea that the light-weight architecture of MobileNetV2, combined with transfer learning, image augmentation and selective fine-tuning can provide an accurate and effective solution for Bangladeshi street food recognition.

10.2 Future Work

The proposed system produced reliable classification results, but there remain several options for future work. The dataset could be enlarged to include more classifications of food products, larger amounts of the original images, and more pictures taken in different conditions. Future research could also be dedicated to aspects of real-time identification with the help of mobile or edge devices, especially because the compact version of MobileNetV2 is particularly useful in this case.

The future work could also cover extensions of the proposed system beyond the classification of images and its applying in real-life situations of the detection of food, assessing nutrition facts and recognizing street food in videos. It would also be beneficial to investigate more explainability approaches that would offer deeper understanding of the decisions the model makes. Finally, successful deployment of the model in mobile/web applications would help turn the proposed system from a laboratory prototype into an actual system for recognizing street food in Bangladesh.

11 Publication History

The paper titled “Machine-Vision-Based Bangladeshi Street Food Recognition” prepared for submission to the 6th International Conference on Emerging Trends and Technologies on Intelligent Systems (ETTIS-2026) and passed the peer-review and revision process accordingly. The study was accepted for presentation and publication in the framework of the conference which took place from 01 to 03 July 2026. The presentation of the paper at this event was given in collaboration with various international academic and research institutions such as IIT Roorkee, NSUT Delhi, Petroleum-Gas University of Ploiesti (Romania), Université de Haute-Alsace (France), Northeastern University (USA). The acceptance and presentation at the conference demonstrates the significance and value of the study in the implementation of machine vision and deep learning methods in the sphere of recognition of street food in Bangladesh.

Bibliography

[1] S. Rokhva, B. Teimourpour, and A. H. Soltani, “Computer vision in the food industry: Accurate, real-time, and automatic food recognition with pretrained MobileNetV2” Food and Humanity, vol. 3, p. 100378, 2024, doi: 10.1016/j.foo-hum.2024.100378.

[2] C. Kiourt, G. Pavlidis, and S. Markantonatou, “Deep Learning Approaches in Food Recognition,” Learning and Analytics in Intelligent Systems, 2020, doi: 10.1007/978-3-030-49724-8_4.

[3] J. Patel and K. Modi, “Indian food image classification and recognition with transfer learning technique using MobileNetV3 and data augmentation,” Eng. Proc., vol. 56, p. 197, 2023, doi: 10.3390/asec2023-15341.

[4] S. P. Mohanty et al., “The Food Recognition Benchmark: Using deep learning to recognize food in images,” Frontiers in Nutrition, vol. 9, p. 875143, 2022, doi: 10.3389/fnut.2022.875143.

[5] Y. Zhang et al., “Deep learning in food category recognition,” Information Fusion, vol. 98,p. 101859, 2023, doi: 10.1016/j.inffus.2023.101859.

[6] D. Liu, E. Zuo, D. Wang, L. He, L. Dong, and X. Lu, “Deep learning in food image recognition: A comprehensive review”, Applied Sciences, vol. 15, no. 14, p. 7626, 2025, doi: 10.3390/app15147626.

[7] B. A. Alimi, “Risk factors in street food practices in developing countries: A review," Food Science and Human Wellness, vol. 5, no. 3, pp. 141–148, 2016, doi: 10.1016/j.fshw.2016.05.001.

[8] A. Wibisono, H. A. Wisesa, Z. P. Rahmadhani, P. K. Fahira, P. Mursanto, and W. Jatmiko, “Traditional food knowledge of Indonesia: a new high-quality food dataset and automatic recognition system,” Journal of Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00342-5.

[9] M. M. Meher, M. Afrin, A. K. Talukder, and Md. G. Haider, “Knowledge, attitudes and practices (KAP) of street food vendors on food safety in selected areas of Bangladesh,” Heliyon, vol. 8, no. 12, p. e12166, 2022, doi: 10.1016/j.heliyon.2022.e12166.

[10] G. Ataguba, R. Ezekiel, J. Daniel, E. Ogbuju, and R. Orji, “African foods for deep learning-based food recognition systems dataset,” Data in Brief, vol. 53, p. 110092, 2024, doi: 10.1016/j.dib.2024.110092.

[11] Z. Xiao, J. Wang, L. Han, S. Guo, and Q. Cui, “Application of machine vision system in food detection,” Frontiers in Nutrition, vol. 9, p. 888245, 2022, doi: 10.3389/fnut.2022.888245.

[12] N. O. M. Salim, S. R. M. Zeebaree, M. A. M. Sadeeq, A. H. Radie, H. M. Shukur, and Z. N. Rashid, “Study for food recognition system using deep learning,” J. Phys.: Conf. Ser., vol. 1963, no. 1, p. 012014, 2021, doi: 10.1088/1742-6596/1963/1/012014.

[13] P. Pouladzadeh and S. Shirmohammadi, “Mobile multi-food recognition using deep learning,” ACM Transactions on Multimedia Computing Communications and Applications, vol. 13, no. 3s, pp. 1–21, 2017, doi: 10.1145/3063592.

[14] S. Mezgec and B. K. Seljak, “NutriNet: a deep learning food and drink image recognition system for dietary assessment,” Nutrients, vol. 9, no. 7, p. 657, 2017, doi: 10.3390/nu9070657.

[15] D. Sahoo et al., “FoodAI,” in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2019, pp. 2260–2268, doi: 10.1145/3292500.3330734.

[16] A. Dhar, M. T. Hossain, and P. Barua, “Bangladeshi street food calorie estimation using improved YOLOV8 and regression model,” arXiv preprint arXiv:2509.01415, 2025, doi: 10.48550/arxiv.2509.01415.

[17] S. M. Shifat, T. Parthib, S. T. Pyaasa, N. M. Chaity, N. Kumar, and M. K. Morol, “A Real-Time Junk Food Recognition System Based on Machine Learning,” Communications in Computer and Information Science, 2022, doi: 10.1007/978-3-031-17181-98.

[18] Y.-C. Chen and H.-C. Chiang, “Deep learning-based automatic food identification with numeric label,” Multimedia Tools and Applications, 2025, doi: 10.1007/s11042-025-20648-x.

[19] M. N. Razali *et al.*, “Indigenous food recognition model based on various convolutional neural network architectures for gastronomic tourism business analytics,” *Information*, vol. 12, no. 8, p. 322, Aug. 2021, doi: 10.3390/info12080322.

[20] X. Chen, Y. Zhu, H. Zhou, L. Diao, and D. Wang, “ChineseFoodNet: A large-scale image dataset for Chinese food recognition,” *arXiv (Cornell University)*, May 2017, doi: 10.48550/arxiv.1705.02743.

[21] M. A. Moni, A. S. Chaitee, K. M. Tanvir, M. A. Rahman, S. Banik, and M. T. Habib, “Machine vision-based pepper breed classification using YOLOv10 and transfer learning models,” in *Proc. 6th Int. Conf. Image Processing and Capsule Networks (ICIPCN)*, Dhulikhel, Nepal, 2026, pp. 544–549, doi: 10.1109/ICIPCN67432.2026.11438307.

[22] A. B. Roy, S. G. Nil, A. Sultana, and Md. T. Habib, “Machine-vision-based extinct fish recognition: Bangladeshi Freshwater case,” *Lecture Notes in Networks and Systems*, pp. 235–245, Jan. 2026, doi: 10.1007/978-3-032-23515-2_23.

[23] J. A. S. Wicaksono, K. D. Kurniawan, and A. A. Gozali, “Mobile assistant application for street food consumers in Bandung,” *Journal of Dinda Data Science Information Technology and Data Analytics*, vol. 4, no. 2, pp. 82–90, Aug. 2024, doi: 10.20895/dinda.v4i2.1470.

[24] Y. Zhang *et al.*, “Deep learning in food category recognition,” *Information Fusion*, vol. 98, p. 101859, May 2023, doi: 10.1016/j.inffus.2023.101859.

[25] G. Ciocca, P. Napoletano, and R. Schettini, “Food recognition: a new dataset, experiments, and results,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 588–598, Dec. 2016, doi: 10.1109/jbhi.2016.2636441.

[26] W. Min *et al.*, “Large scale visual food recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9932–9949, Jan. 2023, doi: 10.1109/tpami.2023.3237871.

[27] M. Chun, H. Jeong, H. Lee, T. Yoo, and H. Jung, “Development of Korean food image classification model using public food image dataset and deep learning methods,” *IEEE Access*, vol. 10, pp. 128732–128741, Jan. 2022, doi: 10.1109/access.2022.3227796.

[28] G. Suddul and J. F. L. Seguin, “A comparative study of deep learning methods for food classification with images,” *Food and Humanity*, vol. 1, pp. 800–808, Aug. 2023, doi: 10.1016/j.foohum.2023.07.018.

[29] A. Wibisono, H. A. Wisesa, Z. P. Rahmadhani, P. K. Fahira, P. Mursanto, and W. Jatmiko, “Traditional food knowledge of Indonesia: a new high-quality food dataset and automatic recognition system,” *Journal of Big Data*, vol. 7, no. 1, Aug. 2020, doi: 10.1186/s40537-020-00342-5.

[30] K. A. Nfor, T. P. T. Armand, K. P. Ismaylovna, M.-I. Joo, and H.-C. Kim, “An explainable CNN and vision transformer-based approach for real-time food recognition,” *Nutrients*, vol. 17, no. 2, p. 362, Jan. 2025, doi: 10.3390/nu17020362.

[31] Y. Kawano and K. Yanai, “FoodCam: A real-time food recognition system on a smartphone,” *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5263–5287, Apr. 2014, doi: 10.1007/s11042-014-2000-8.

[32] C. Liu *et al.*, “A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure,” *IEEE Transactions on Services Computing*, vol. 11, no. 2, pp. 249–261, Jan. 2017, doi: 10.1109/tsc.2017.2662008.

[33] A. Fakhrou, J. Kunhoth, and S. A. Maadeed, “Smartphone-based food recognition system using multiple deep CNN models,” *Multimedia Tools and Applications*, vol. 80, nos. 21–23, pp. 33011–33032, Aug. 2021, doi: 10.1007/s11042-021-11329-6.

[34] Z. Chen, J. Wang, and Y. Wang, “Enhancing food image recognition by multi-level fusion and the attention mechanism,” *Foods*, vol. 14, no. 3, p. 461, Jan. 2025, doi: 10.3390/foods14030461.