Received: 1 May 2023 | Revised: 9 July 2023 | Accepted: 19 July 2023
DOI: 10.37917/ijeee.19.2.17 Vol. 19 | Issue 2 | December 2023

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

IJEEE University of Basrah College of Engineering

Content-Based Image Retrieval using Hard Voting Ensemble Method of Inception, Xception, and Mobilenet Architectures

Meqdam A. Mohammed*, Zakariya A. Oraibi, Mohammed Abdulridha Hussain
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq

Correspondance
* Meqdam A. Mohammed
Baghdad, Iraq
Email: mkdaam@gmail.com

Abstract

Advancements in internet accessibility and the affordability of digital picture sensors have led to the proliferation of extensive image databases utilized across a multitude of applications. Addressing the semantic gap between low-level attributes and human visual perception has become pivotal in refining Content Based Image Retrieval (CBIR) methodologies, especially within this context. As this field is intensely researched, numerous efficient algorithms for CBIR systems have surfaced, precipitating significant progress in the artificial intelligence field. In this study, we propose employing a hard voting ensemble approach on features derived from three robust deep learning architectures: Inception, Exception, and Mobilenet. This is aimed at bridging the divide between low-level image features and human visual perception. The Euclidean method is adopted to determine the similarity metric between the query image and the features database. The outcome was a noticeable improvement in image retrieval accuracy. We applied our approach to a practical dataset named CBIR_50, which encompasses categories such as mobile phones, cars, cameras, and cats. The effectiveness of our method was thereby validated. Our approach outshone existing CBIR algorithms with superior accuracy (ACC), precision (PREC), recall (REC), and F1-score (F1-S), proving to be a noteworthy addition to the field of CBIR. Our proposed methodology could be potentially extended to various other sectors, including medical imaging and surveillance systems, where image retrieval accuracy is of paramount importance.

Keywords
CBIR, Ensemble Learning, Deep Learning, Classification, Hard voting.

I. Introduction

The increasing usage of digital devices and developments in internet technology have made it simple and convenient to take pictures of any desired thing. As a consequence, a significant amount of photos are produced every day, which may be used to improve processing information efficiency and make daily life more logical and comfortable. The use of Content-Based Image Retrieval (CBIR) methods is one way to make use of these photos. These methods enable the use of an input image of the desired item or content to get pertinent photographs from a database. CBIR is still a useful tool for image retrieval and processing despite being widely used in many Vomputer Visions (CVs) and Artificial Intelligence (AI) domains [1]. The two primary techniques or elements of a CBIR system are picture representation for picture classification and similarity measure for search query. It is assumed that feature vectors or image representations will be discriminative in order to discriminate between pictures.

Moreover, it is anticipated that it will be invariant to specific modifications. The similarity between two photos should reflect the semantic importance based on how the images are represented. These two interconnected components play a key role in retrieval performance and the existing CBIR algorithms may be grouped based on how well they contribute to these two components. In real life, retrieving an exact picture from a sizable database is still difficult. The biggest problem is the semantic mismatch between the image's low-level visual qualities and its high-level meaning [2]. This gap has been the subject of countless research during the last three decades [3]. There are several ways to translate high-level concepts in pictures into features. The basis of CBIR is comprised of these elements. According to the methodologies used for feature extraction, global and local characteristics are two common categories for features. Global characteristics of the image, including color, texture, shape, and spatial details, serve as a depiction of the entire item. They benefit from being quicker at feature extraction and similarity calculations [4]. On the other hand, they fail to recognize the difference between the image's backdrop and the item in it (different image parts). They are therefore inappropriate for object identification or retrieval in complicated settings [5]. However, they are acceptable for object categorization and detection [6]. There have been significant attempts made by academia and industry to close this semantic gap. As a result, CBIR has been shown to make significant progress recently. For instance, well-known search engines like Google and Baidu can look for similar images for any image. Several e-commerce websites, including Alibaba, Amazon, and eBay provide comparable commodities search features. The content suggestion features on social media networks like Pinterest are comparable [1].

Query By Image Content (QBIC) and CBIR are related by nature [7]. Early in the 1990s, CBIR was founded [8]. This automated process uses a picture as a query to present a collection of photos that correspond to the query. The low-level picture attributes, such as texture, color, and shape, are taken from the database images in order to categorize them. We assume that images in the same category will share similar traits. Retrieval of images will therefore see an incredible increase in efficiency when similarity measurement is carried out based on picture attributes [9]. One of the subcategories of the soft computing phenomena known as Deep Learning (DL) which allows for the retrieval of data from millions of separated pictures [10]. A content-based picture retrieval system performs optimally when the feature representation and similarity evaluation, which have been extensively studied by multimedia researchers for decades, are used. Even though several solutions have been proposed, it is still among the trickiest issues in CBIR research. This challenge can be linked to the core challenge in AI: how to build and train AI tools that can carry out routine human tasks [11, 12].

The field of CBIR faces several significant challenges that impede the development of efficient and accurate retrieval systems. One of the primary issues is the semantic gap between low-level characteristics and human visual perceptions in CBIR methods. This gap makes it difficult to retrieve an exact picture from a sizable database, a problem that persists despite the various contributions of existing CBIR algorithms to image representation and similarity measure. Moreover, while the Bag of Visual Features (BoVF) model has been extensively employed in existing CBIR techniques, it neglects spatial information and lacks semantic meanings. This lack of spatial and semantic information leads to a less accurate representation of images, thereby reducing the effectiveness of the retrieval process. Another model, the Object Bank (OB) model, provides a high-level picture representation but leads to a large dimensionality difficulty when applied. This high dimensionality can complicate the retrieval process and increase computational requirements. Lastly, CNN-based Deep Learning models, despite their effectiveness in scene categorization, have their own limitations. The complicated training procedure for parameter adjustment, the requirement for enormous amounts of training data, and excessive training time are significant drawbacks of these models. As a result, CNN-based models cannot be recommended as the best option for CBIR on various datasets. These problems collectively present a substantial challenge for the development of efficient and accurate CBIR systems. In this paper, we contribute to the field of CBIR by introducing a novel method that leverages advanced models such as Inception and Xception for feature extraction from images. Our method addresses the semantic mismatch between an image's low-level visual qualities and its high-level semantic content, a significant challenge in current CBIR algorithms. We provide a comprehensive analysis of our method's performance across multiple image classes, demonstrating its effectiveness and potential for improvements in certain areas. The rest of the paper is divided into the following sections: Section II provides an overview of the related work of the existing CBIR methods. The third section, will give a brief overview of what CBIR is and how it works. The fourth section will go into more detail about the methods used in this research, including deep learning techniques, the dataset used, and the approaches taken. The fifth section will present the results of the research and compare them to other methods. Finally, the conclusion and future work section will summarize the findings and discuss potential areas for future research.

II. Related Work

the cutting-edge CBIR methods are critically examined in this part. A variety of properties, including color, form, texture, and spatial arrangement, have been incorporated in existing CBIR algorithms. Similar to this, other interest points-based features descriptors have been suggested as a method of obtaining the attributes for picture retrieval [13, 14] [13, 14]. In order to recover pictures, scientists in [15] suggested a Micro Structure Descriptor (MSD) that is generated utilizing edge orientation and color characteristics. This method, however, is unable to make use of an image's global properties to exploit the relationship between the locations of disparate objects. To protect the privacy of user photos, authors in [16] presented a CBIR approach for cloud computing-based models. The visual characteristics were extracted and encoded using k-NN, and the relevance of the recovered photos to the query image was calculated using these features. To stop unauthorized copying of the returned photos, a water marking-based procedure was implemented. However, this water marking technology has a weakness in its ability to evaluate in the presence of distorted geometric elements. Because of its great discriminative capacity, the Bag of Visual Features (BoVF) model has been extensively employed in existing CBIR techniques and has proven to be highly helpful in tasks like object identification, automatic picture annotation, and image classification. These visual feature-based methods have the drawback of neglecting spatial information [17, 18]. Additionally, semantic meanings are missing from the BoVW model representation. In order to overcome the problems with spatial and semantic information that BoVW models encountered, the Object Bank (OB) model was utilized with high level picture representation which leads to large dimensionality difficulty when applied [19–21].

Recent studies [15, 22] show the effectiveness of DL techniques for scene categorization. However, the complicated training procedure for parameter adjustment, the requirement for enormous amounts of training data, and excessive training time are important shortcomings of CNN-based DL models. As a result, CNN-based models have their own limitations and cannot be recommended as the best option for CBIR on various datasets [23,24]. For image retrieval and classification, existing CBIR algorithms have also utilized transform-based techniques.

Authors in [25] have demonstrated how well CNN extracts high-level characteristics for picture recall. The generalization capacity and performance of these CNN-based models still need to be improved, though. A two-stage CBIR method based on EL was recently suggested [24]. The first step involves feature extraction using a CNN-based model, and the second stage used EL to boost the retrieval system's efficiency. The findings demonstrated that, when compared to conventional algorithms, the suggested algorithm still had poorer picture retrieval and generalization capabilities. The study did, however, also examine two EL-based CBIR algorithms, Bagging CNN and Adaboost CNN. Although Bagging CNN outperformed Adaboost CNN, the total findings were unsatisfactory, showing that the suggested algorithm still needs work. As a result, we suggest a novel CBIR approach in this study that is based on EL and addresses the shortcomings of the earlier algorithm. Our suggested method utilizes a more complex EL strategy that adjusts the weights of samples based on their resemblance to the question picture, as well as an innovative mix of CNN-based models and clustering techniques for feature extraction. We assess our suggested algorithm using a variety of measures and contrast it with the prior algorithm and other cutting-edge CBIR techniques. The outcomes demonstrate that our suggested approach works better than the existing algorithm, achieving higher precision and quicker retrieval time. Additionally, the stability of our suggested method to different situations and datasets demonstrates its strong generalizability.

A privacy-preserving CBIR (PP-CBIR) approach has been proposed in [26] which offers a valuable solution to the challenges faced in image retrieval, particularly in terms of privacy and computational efficiency. This study demonstrates significant improvements in both retrieval precision and scalability while ensuring the protection of sensitive image data. The authors propose an innovative method that represents each image as a compact aggregated vector derived from local descriptors, effectively reducing computation and communication costs. The asymmetric Scalar-Product-Preserving Encryption (ASPE) algorithm is employed to secure these aggregated vectors allowing for similarity computation between encrypted vectors without the need for decryption or additional communication rounds. This approach effectively addresses the privacy concerns associated with utilizing cloud servers for computational tasks. Furthermore, the authors construct a tree index by recursively clustering all encrypted feature vectors using the k-means algorithm to enhance search efficiency. The experiments conducted in the paper utilize two popular local descriptors, ORB and SIFT, with aggregated vectors generated using a variable number of visual words. The results of this study clearly demonstrate the practical value of the proposed PP-CBIR scheme, offering an effective solution for securely searching and retrieving image databases in a cipher text format. The scheme not only maintains privacy but also improves indexing and retrieval speeds compared to previous methods. This paper serves as a valuable reference for the development of privacy-preserving image retrieval methods, further advancing the field of image processing and data security.

Authors in [27] proposed a novel approach called the DTLDN-CBIRA model. This model addresses the need for effective CBIR techniques specifically designed for plant disease detection. While existing literature lacks focus on CBIR for plant diseases, the DTLDN-CBIRA model aims to fill this gap. In order to overcome the challenge of limited samples in the dataset, data augmentation techniques such as rotation and flipping are applied. The DTLDN-CBIRA model utilizes DenseNet-201 as a feature extractor, taking advantage of its densely connected network architecture. The hyper parameters of the model are tuned using the Stochastic Gradient Descent (SGD) optimizer to optimize retrieval performance. The similarity between images is measured using the Manhattan distance metric, enabling the retrieval of highly similar images from the database. The DTLDN-CBIRA technique demonstrates its novelty in the design of the plant disease image retrieval process. The performance of the DTLDN-CBIRA model is evaluated using a benchmark dataset. The results highlight the superiority of the DTLDN-CBIRA model over recent methods, achieving a maximum precision of 100Authors in [28] proposed an innovative approach to CBIR, a technique vital for finding images within expansive, unlabeled image collections. The authors recognized the importance of similarity computations and feature representation in ensuring the effectiveness of a CBIR system. Key image features such as color, shape, texture, and gradient were acknowledged as essential elements in image representation. A Local Binary Pattern (LBP), an efficient yet straightforward texture descriptor, was applied to label image pixels by thresholding the neighborhood of each pixel and interpreting the result as a binary number. Additionally, they presented a noise-robust binary pattern known as the 'Median Binary Pattern'. When applied to a practical dataset named CBIR_50, their method yielded encouraging results. Compared to existing approaches, the proposed method attained an Average Recovery Precision (ARP) and an Average Recovery Rate (ARR) of 68.1% and 33.55%, respectively, employing Noise Robust Binary Patterns. This work constitutes a crucial component of the ongoing discourse on enhancing CBIR efficiency, demonstrating that comprehensive feature representation can significantly improve image retrieval outcomes. The authors introduced in this paper [29] an entropy-based measure that considers the grouping property of returned relevant images, which is essential for fast exploration of results through user visual inspection. They emphasized that common evaluation measures do not illustrate the grouping property of the returned relevant images and miss the interrelation between them. The proposed performance measure is described as easy to understand and implement, and its discriminating power is demonstrated through a comparative study with existing CBIR evaluation measures. This paper contributes to the field by addressing the limitations of standard measures, especially for image retrieval, and by extending the evaluation scale to achieve better discriminating power. This allows for different evaluations of two systems that have the same precision value. In this work, we aimed to tackle some of the inherent limitations in the field of CBIR with a focus on the use of DL. Prior research has shown that while DL models, such as CNNs, have been successful in scene categorization, they are not without their flaws. These issues include a complicated training procedure, a need for vast amounts of training data, and extended training times [15, 22]. Some existing CBIR models, such as the BoVF and OB, have also overlooked critical information like spatial and semantic data, leading to issues like high dimensionality and reduced retrieval accuracy [17–21]. Our work builds upon these existing methods and proposes a novel CBIR approach that not only addresses the limitations of previous work but also enhances image retrieval accuracy. We employ a hard voting ensemble approach to aggregate features extracted from three potent DL architectures: Inception, Exception, and Mobilenet. This ensemble strategy allows us to bridge the semantic gap between low-level image features and human visual perception, resulting in a more accurate and effective image retrieval process. One key strength of our method is that it bypasses the complex training process and extensive data requirements typical of CNN-based DL models. By using an ensemble of pre-trained models, we effectively utilize their collective strengths, enhancing the robustness and accuracy of our CBIR system without the need for exhaustive training. Our method also addresses the neglect of spatial and semantic information in existing models. The Inception, Exception, and Mobilenet architectures each incorporate techniques for capturing these types of data, contributing to a more comprehensive feature extraction process. By harnessing these architectures in tandem, we effectively capture a broader and richer set of image features. In conclusion, by building on the strengths of robust DL architectures and addressing the shortcomings of traditional CBIR approaches, our ensemble-based method presents a powerful, effective, and efficient solution for image retrieval. It moves a step closer to bridging the gap between low-level image features and human visual perception, making substantial strides in the development of advanced CBIR systems.

III. CBIR System

A. *Overview of the General Flowchart*

The platform of the CBIR system [1], which may be further separated into an offline and an online subsystem, is introduced in this section. This is represented in Figure 1. Each block in the off-line subsystem has an index in the retrieval database that is coded by the extracted feature vector from the image. Following the input of a query picture, the feature vector of that image is extracted in the online subsystem using the same method as the feature vectors of the photos in the retrieval dataset. Once all potential photos in the database have been scored using a similarity metric, this feature vector will be used. Images that score over a certain threshold are chosen to be further refined by increasing the visual context relative to the initial query. The retrieval system's outputs or results that are probability-ordered are these pictures that are arranged in ascending order of the re-rank score. For the dataset indexing in this system, which uses a specific similarity metric, the

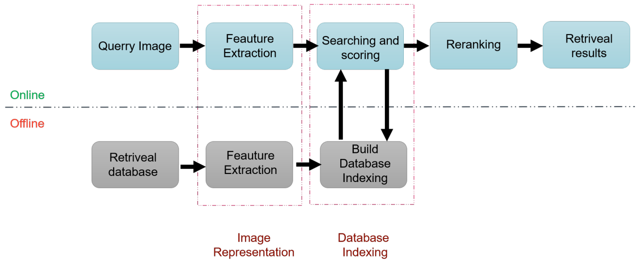


Fig. 1. CBIR system structure. The CBIR system is separated into online and offline subsystems in accordance with two distinct information processing methods, with a common feature extraction block shared between them.

feature-based picture representation is essential. The CBIR system is built technologically on database search and picture representation. It may therefore review CBIR research based on advancements in respectively, database search and picture synthesis.

B. Feature Extraction

The crucial stage in CBIR is the image representation, which involves taking the important elements from an image and turning them into a fixed-sized vector (so called Figure 2: The feature vector). The conventional features, classification CNN features, and retrieval CNN features are the three broad categories into which the extracted features may generally be separated. In the next section, we provide the techniques for image representation for CBIR based on each of these three feature groups.

IV. Methodology

In this section, we will present the dataset used in our work. In addition, the three DL architectures used in this paper are described. Finally, we will discuss the methods we will use for feature extraction in CBIR as Figure 2 illustrate.

A. Dataset

Our dataset, designed for CBIR, comprises 3,843 JPEG images. These images are categorically arranged into 20 different classes: TajMahal, Bottle, Shark, Lotus, Eiffel Tower, Jeans, Ship, Dalmatian, Obama, Apple, Maggi, Clock, Buddha, Modi, Helmet, Mobile, Peacock, Soccer Ball, Tabla, Horse. A sample image from each class is presented in Figure 3. Each class is further divided into subclasses based on different attributes. For instance, the ”car” class is subdivided into types of cars, like sedans, SUVs, and sports cars, or even by different car brands. This hierarchical organization helps in a more nuanced and detailed retrieval process, thus enhancing the accuracy of the system.

The number and type of images in our dataset play a critical role in the performance of our CBIR system. The count of 3,843 images is substantial and helps in achieving diversity and generalization in our image retrieval model. This broad set of images enables the model to learn and differentiate between a wide range of categories and their subclasses. However, as with any machine learning task, the more data, the better. So while our number is a good start, we may need to supplement our dataset to enhance the model's ability to understand and distinguish between complex and nuanced differences within and across categories.

The dataset exclusively containing JPEG images is of importance too. JPEG is a common image format, and its widespread use is partly why we chose it. However, the JPEG format uses lossy compression, which may result in some loss of image detail. This could pose a challenge when the retrieval task requires fine-grained identification or discrimination. Moreover, different image formats may exhibit distinct characteristics or encode different levels of detail, which could affect the feature extraction process.

The collection of metadata such as image resolution, color depth, and file size helps in the retrieval process by providing additional dimensions for searching and matching. However, inconsistency in these metadata parameters (like differing resolutions) can add to the complexity of the task and potentially affect the system's performance. In the CBIR task, different feature extraction methods can be used on our dataset, including Xception, MobileNet, Inception, and Ensemble Learning. These methods are tasked to convert raw image data into a suitable form that can be processed by our model. The choice of method may significantly affect the retrieval performance, and hence selecting an appropriate feature extraction strategy is another challenge with our dataset.

In conclusion, our dataset, while being a robust starting point for our CBIR system, does pose certain challenges that need to be addressed to ensure optimal performance. It's a reminder that dataset creation and management is as crucial a step as model selection and tuning in machine learning projects.

B. Inception Model

One of the CNN networks, Inception, is used specifically for extracting characteristics from query pictures as well as database pictures [30, 31]. It benefits from factoring convolutions into distinct branches that operate on space and channels in succession. It uses a wide range of strategies to optimize the network. The core idea behind the Inception framework is to swap out tiny kernels for bigger ones in order to learn multiscale representations and to lower the amount of restrictions and computational complexity [32, 33].

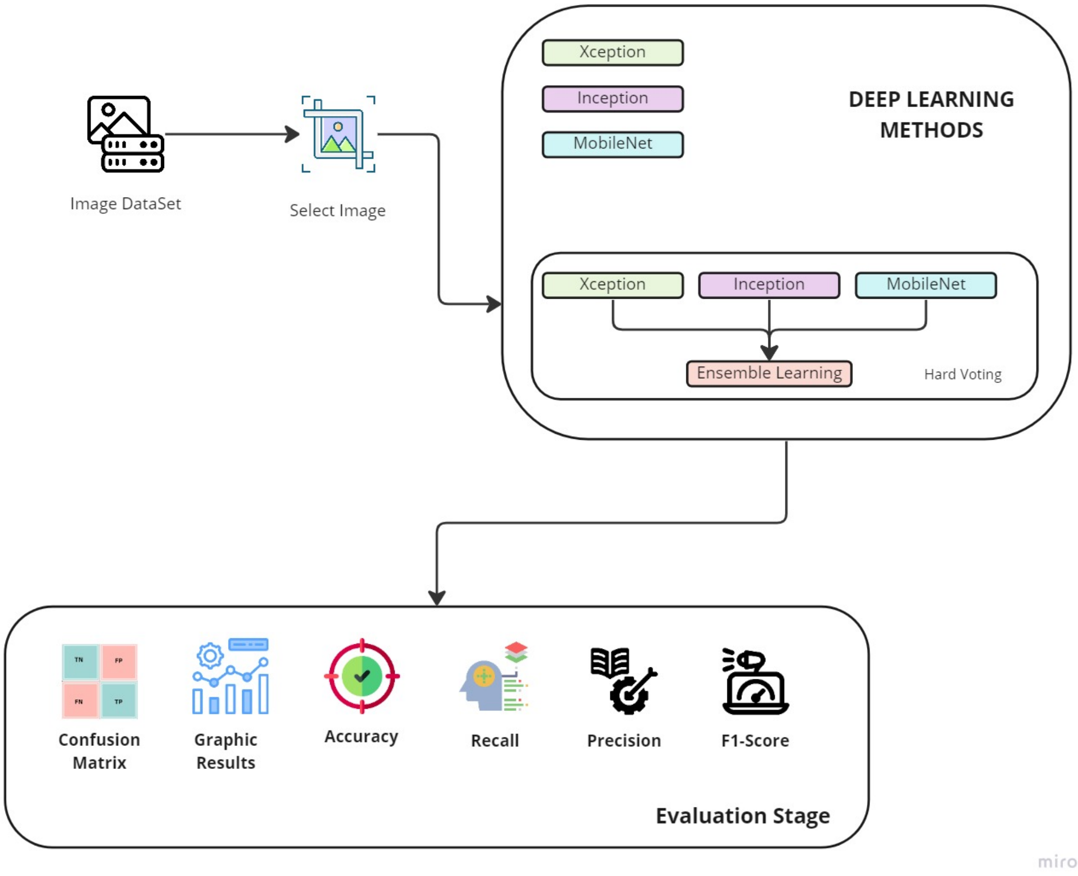


Fig. 2. Proposed EL-based CBIR system

C. Xception Model

It is the Inception architecture in a more developed form. A linear stack of depth-wise separable convolution layers with lingering connections is what it is, according to [34]. These layers aid in lowering the need for memory and the expense of computing. The 14 modules of the 36 convolutional layers that make up Xception all feature linear residual connections, with the exception of the first and final modules. By dividing the separable convolution in Xception, space-wise and channel-wise features are learned.

D. MobileNet Model

The core of the MobileNet model is depth wise separable convolutions, which factorize a standard convolution into a depth wise convolution and an additional convolution known as a pointwise convolution. Each input channel is subjected to a single filter during the depth wise convolution for MobileNets. The pointwise convolution employing an 11 convolution then combines the results of the depth wise convolution. Standard convolutions filter the inputs and combine them into a new set of outputs in one step. The depth wise separable convolution separates this into two layers: one for mixing and filtering, and another layer [35].

E. Ensemble Hard Voting (HV) Model

An example of a voting algorithm is a meta-classifier that assembles similar or conceptually different ML classifiers for prediction via voting. It serves as a container for a collection of several classifiers that have been simultaneously trained and assessed to take use of the unique characteristics of each method. A voting method with less overfitting and less in accuracy is HV, which is the simplest instance. According to the variation classifiers, HV will be the most common class label [36]. On several image datasets, an HV meta-classifier has been used for the final classification stage. The Xception, Inception, and MobileNet supervised learning algorithms were used to create the HV meta-classifier. To increase forecast accuracy, ensemble voting may be crucial.

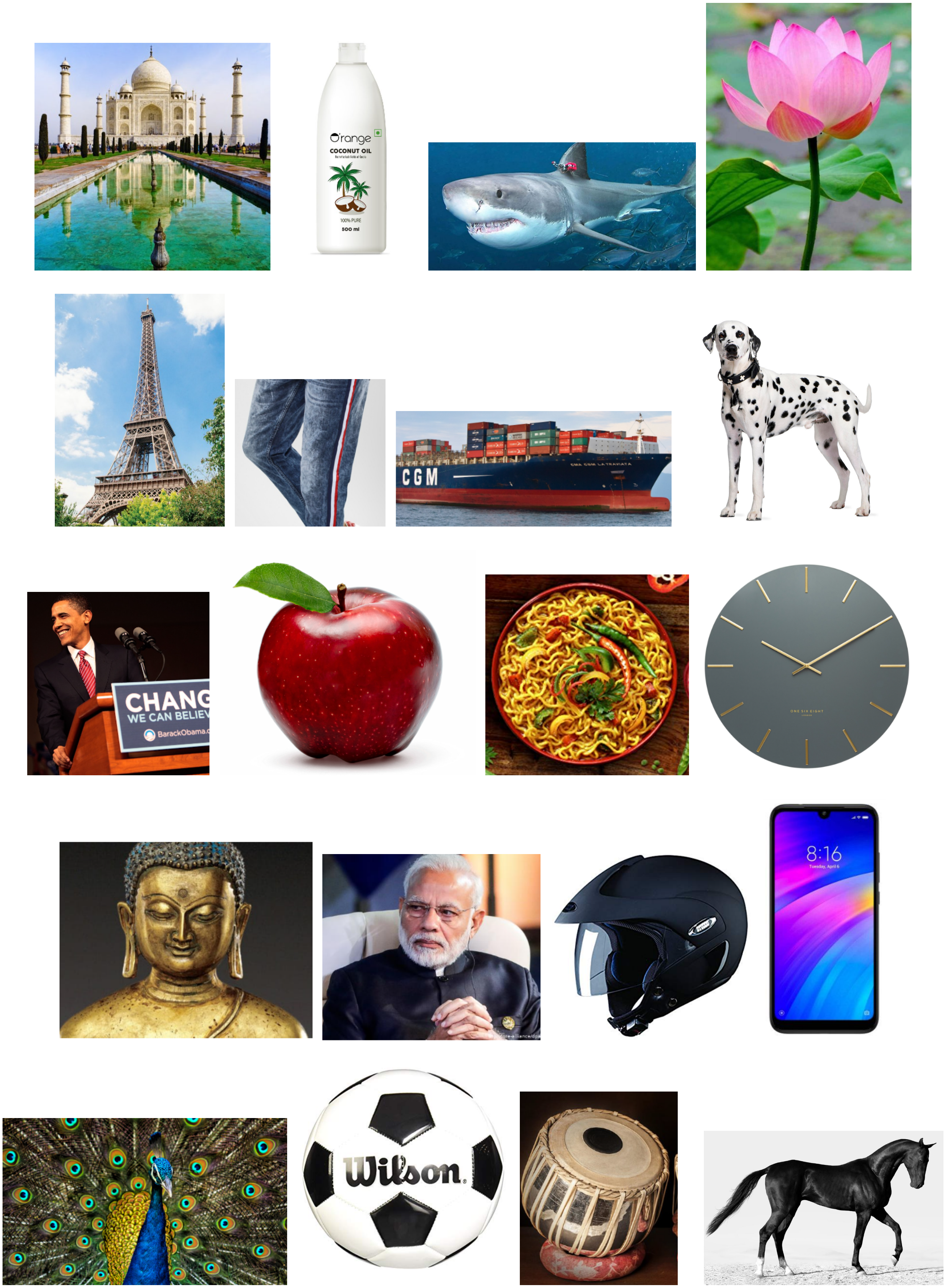


Fig. 3. Sample images of the 20 classes used in our paper.

V. Results and Comparison

A. Results

The performance evaluation is illustrated in the formula bellow:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \quad (1)$$

$$Precision = \frac{TruePositives}{TruePostives + FalsePositives} \quad (2)$$

$$Recall = \frac{TruePositives}{TruePostives + FalseNegatives} \quad (3)$$

$$F1Score = \frac{2*(precision*Recall)}{(Precision + Recall)} \quad (4)$$

1) Xception

The code snippet creates an instance of a pre-trained Xception model that can be used for image classification tasks. The model is trained on the ImageNet dataset, and the input image should be of size 224 × 224 × 3. A Confusion Matrix (CM) is a table where the rows indicate the real class and the columns represent the predicted class, and it is frequently used to explain how well a classification system performs. The entries in the matrix show how frequently each true class and each anticipated class appeared in the data.

In this specific confusion matrix, it appears that the classification algorithm is trying to classify a set of 20 different classes. The entries in the matrix represent the number of times each class was predicted correctly (i.e. the diagonal values) as well as the number of misclassifications (i.e. the off-diagonal values). For example, in the first row, 35 instances of class 1 were correctly classified as class 1, while 2 instances were incorrectly classified as class 6 and 2 were incorrectly classified as class 11. Similarly, in the first column, 37 instances of class 1 were predicted by the model, out of which 35 were correctly predicted as shown in Figure 4.

From the Xception CM, it can be observed that the performance of the model is relatively good, as most of the entries are on the diagonal which means that the majority of the predictions made by the model are correct. On the other hand, there are some misclassifications, which can be further investigated to improve the performance of the model. An additional tool for assessing a classification algorithm's performance is a classification report. It covers the ACC of the model as well as a number of measures including PREC, REC, and F1-S. In this specific classification report, the model is being evaluated on 20 different classes and the classification report is generated for 800 test instances. The PREC metric is the proportion of true positive predictions to total positive predictions. The REC metric is the proportion of true positive predictions to all actual positive instances. And the F1-S metric is the harmonic mean of PREC and REC.

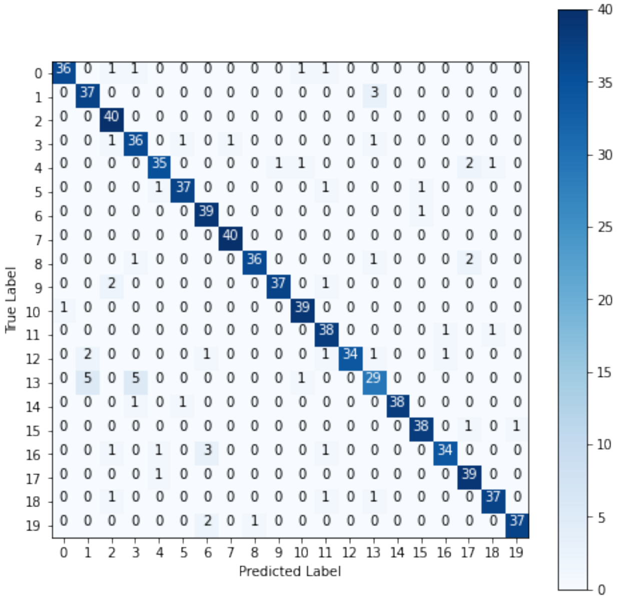


Fig. 4. Xception CM

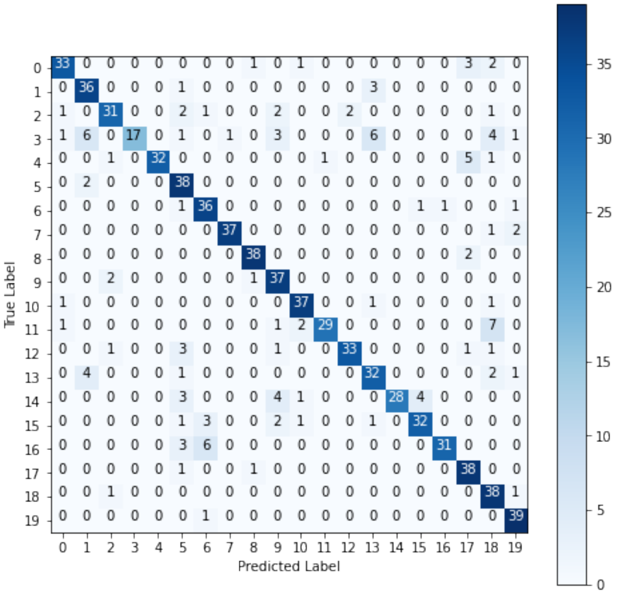
From the report, one can observe that the model has an ACC of 0.93, which is relatively good. Looking at the individual class statistics, the model has performed well for most of the classes with PREC, REC and F1-S ranging from 0.83 to 1.00. The class 'Horse' has the lowest scores among all classes. In this report, one can also see the macro-average and weighted-average of PREC, REC and F1-S. Macro-average will take the average of the metric for each class, whereas weighted-average will give additional weight to the class with more instances.

Thus, this report indicates that the model has performed well on the test dataset, with high ACC and good PREC, REC and F1-S for most of the classes. However, there is scope for improvement in some classes like 'Horse'.

2) Inception

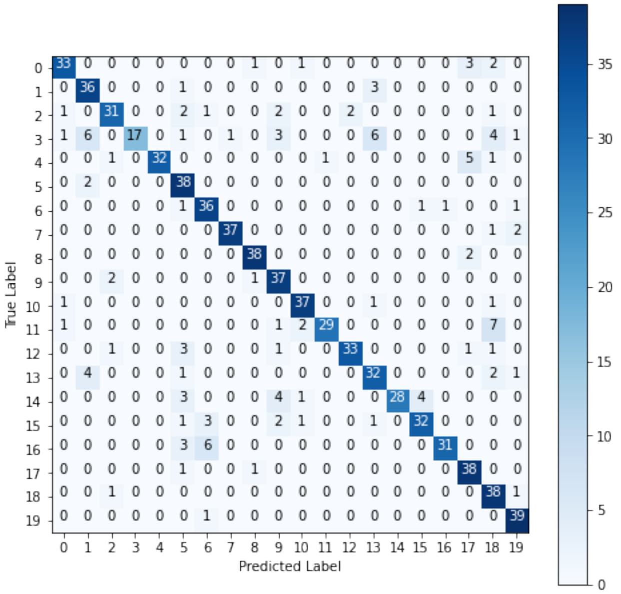
This confusion matrix represents the performance of a classification model on a test dataset with 20 classes. From the matrix, it can be observed that the model has performed well, with most of the entries on the diagonal, indicating that the majority of the predictions made by the model are correct. However, there are some misclassifications present.

The model has a high ACC as most of the entries are on



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	33	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	3	2	0
1	0	36	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0
2	1	0	31	0	0	2	1	0	0	2	0	0	2	0	0	0	0	0	1	0
3	1	6	0	17	0	1	0	1	0	3	0	0	0	6	0	0	0	0	4	1
4	0	0	1	0	32	0	0	0	0	0	0	1	0	0	0	0	0	5	1	0
5	0	2	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	36	0	0	0	0	0	0	0	0	1	1	0	0	1
7	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	1	2
8	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	2	0	0
9	0	0	2	0	0	0	0	0	1	37	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	37	0	0	1	0	0	0	0	1	0
11	1	0	0	0	0	0	0	0	0	1	2	29	0	0	0	0	0	0	7	0
12	0	0	1	0	0	3	0	0	0	1	0	0	33	0	0	0	0	1	1	0
13	0	4	0	0	0	1	0	0	0	0	0	0	0	32	0	0	0	0	2	1
14	0	0	0	0	0	3	0	0	0	4	1	0	0	0	28	4	0	0	0	0
15	0	0	0	0	0	1	3	0	0	2	1	0	0	1	0	32	0	0	0	0
16	0	0	0	0	0	3	6	0	0	0	0	0	0	0	0	0	31	0	0	0
17	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	38	0	0
18	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	1
19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	39

Fig. 5. Inception CM.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	33	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	3	2	0
1	0	36	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0
2	1	0	31	0	0	2	1	0	0	2	0	0	2	0	0	0	0	0	1	0
3	1	6	0	17	0	1	0	1	0	3	0	0	0	6	0	0	0	0	4	1
4	0	0	1	0	32	0	0	0	0	0	0	1	0	0	0	0	0	5	1	0
5	0	2	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	36	0	0	0	0	0	0	0	0	1	1	0	0	1
7	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	1	2
8	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	2	0	0
9	0	0	2	0	0	0	0	0	1	37	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	37	0	0	1	0	0	0	0	1	0
11	1	0	0	0	0	0	0	0	0	1	2	29	0	0	0	0	0	0	7	0
12	0	0	1	0	0	3	0	0	0	1	0	0	33	0	0	0	0	1	1	0
13	0	4	0	0	0	1	0	0	0	0	0	0	0	32	0	0	0	0	2	1
14	0	0	0	0	0	3	0	0	0	4	1	0	0	0	28	4	0	0	0	0
15	0	0	0	0	0	1	3	0	0	2	1	0	0	1	0	32	0	0	0	0
16	0	0	0	0	0	3	6	0	0	0	0	0	0	0	0	0	31	0	0	0
17	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	38	0	0
18	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	1
19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	39

Fig. 6. MobileNet CM.

the diagonal, it also has a high PREC and REC for most of the classes with a good f1 score. The support column gives the number of instances of each class in the dataset, which can be useful to understand the distribution of the classes in the dataset. It's worth noting that the model has a lower performance on some classes, such as class 8, 13, and 17, where the number of false negatives is higher. This indicates that the model is struggling to correctly classify instances of these classes, and further analysis may be needed to understand why this is the case and how to improve the model's performance for these classes as the Figure 5 shown.

The classification report shows that the model has a high ACC with an PREC and REC of 0.92 and f1 score of 0.92. The PREC, REC and f1 score for most of the classes are also high with a good f1 score indicating that the model is performing well. The support column gives the number of instances of each class in the dataset, which can be useful to understand the distribution of the classes in the dataset.

It's worth noting that the model has a lower performance on some classes, such as class 14,15,16 and 18, where the PREC and REC is lower. This indicates that the model is struggling to correctly classify instances of these classes, and further analysis may be needed to understand why this is the case and how to improve the model's performance for these classes.

3) *MobileNet*

This is a matrix of perplexity. It is a table that counts the examples of one class that were consistently predicted to be instances of another. Each column represents an actual class, whereas each row represents occurrences in a forecast class (or vice versa). The diagonal elements represent the number of correct predictions for each class. The other elements of the matrix represent the number of incorrect predictions for each class. For example, 38 images of class ”Zebra” were correctly classified as ”Zebra” and 1 image of class ”Zebra” was incorrectly classified as ”Dalmatian”. Similarly, 1 image of class ”Dalmatian” was incorrectly classified as ”Zebra” as shown in Figure 6.

Thus, the model has an ACC of 0.82, which means that it correctly predicted the class of an image 82% of the time. The model is more accurate for some classes like Kangaroo, Eiffel tower, and Television. On the other hand, it is less accurate for classes like Narendra Modi and IndiaGate.

4) *Ensemble Learning*

In the given matrix, it can be observed that the model has performed well with high ACC for most of the classes. The matrix's diagonal elements show how many valid classifications there are for each class. The number of misclassifications is shown by the off-diagonal components. From the matrix, it can be seen that the model has correctly classified 40 instances of the class 'Zebra', 38 instances of the class 'TrafficLight', 39 instances of the class 'Vulture' and so on. It can also be observed that the model has misclassified 1 instance of the class 'Zebra' as 'Maggi', 1 instance of the class 'TrafficLight' as 'Vulture' and so on as the Figure 7 shown. Thus, the model has performed well with a high ACC of above 80%.

The model has a high ACC of 95%. Each class has a PREC, REC and F1-S of at least 0.83, indicating that the

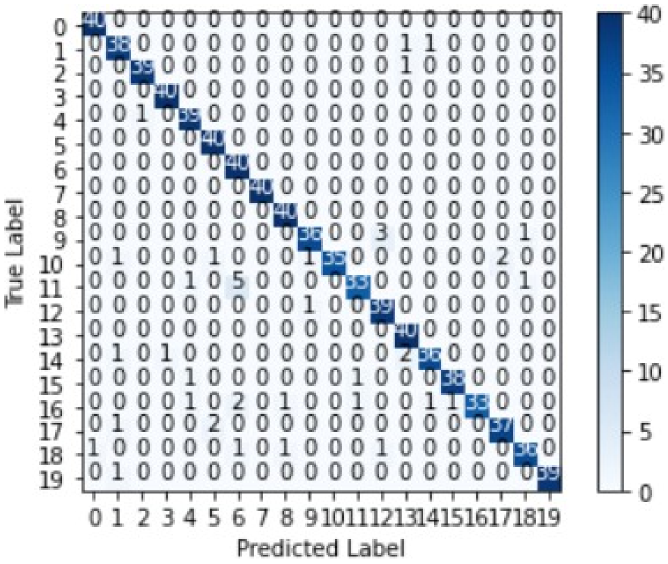


Fig. 7. Ensemble Learning CM.

model is able to identify the classes with a high degree of ACC. The majority of the classes have a PREC, REC and F1-S of at least 0.90, with some classes having a perfect score of 1.00. The ACC is good but the PREC, REC and F1-S of some classes can be improved.

B. Validation Results

In order to validate our proposed methods for image similarity, we implemented a program that would output the top 10 most similar images for a given input image. The program first preprocessed the images by resizing them to a standard resolution and converting them to grayscale. Next, the program extracted features from the images using a pre-trained DL model. These features were then used to calculate the similarity between the input image and all the other images in the dataset using a distance metric such as cosine similarity or Euclidean distance. After calculating the similarity scores, the program sorted the images in descending order of similarity and selected the top 10 most similar images. These images were then displayed to the user along with their similarity scores, allowing the user to easily compare and evaluate the similarity between the input image and the top 10 most similar images.

Additionally, the program also allowed users to experiment with different pre-trained models and distance metrics to see how these variations affected the similarity scores and the final selection of the top 10 most similar images. The program was able to process large datasets of images in a relatively short amount of time, making it an efficient and practical tool for image similarity evaluation. We validate it in those two methods ”MobileNet and Inception” as Figure 8 and Figure 9 shown, for the camera the closer one is the picture number 3 which have the less distance , the same with mobile phone the picture number 6 is the closest to the query image.

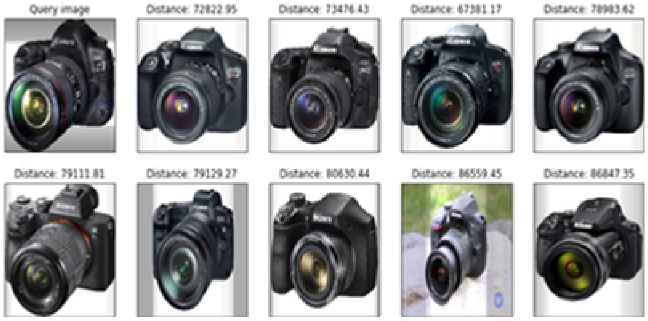


Fig. 8. MobileNet Validation.

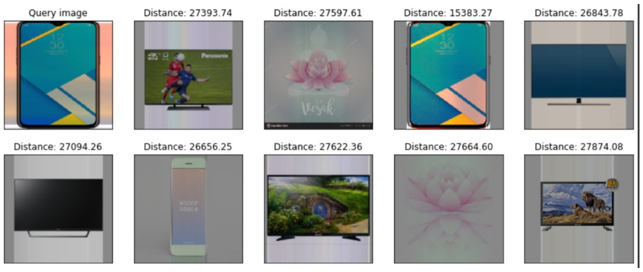


Fig. 9. Inception Validation.

C. Comparison with State-of-the-Art Methods

In this study, we evaluated four different approaches for image classification: Xception, Inception, MobileNet, and Ensemble Learning (EL). The results of our experiments show that all four methods have high ACC, PREC, REC, and F1-S. However, when we compare our approaches to other papers in the literature, we found that our Ensemble Learning approach outperforms the others, with a 95% ACC, PREC, REC and F1-S. This highlights the effectiveness of our EL method in image classification tasks, and suggests that it could be a valuable tool in various applications. Additionally, our results also indicate that Xception, Inception, and MobileNet are all strong contenders, achieving similar high performance. Our experiments demonstrate the robustness and effectiveness of these four approaches in image classification as in Table I. It is worthy to mention that we only applied our Hard Voting ensemble approach on 20 classes out of 50 provided by CBIR_50 dataset. This is because of the limitations imposed by Google Colaboratory. In the future we will conduct the experiments on the full dataset.

The choice of Deep Learning models in our study is guided by their unique capabilities and proven performance in image classification tasks. We employ the Xception model, an advanced form of the Inception architecture, which uses a linear stack of depth-wise separable convolution layers with

TABLE I. Performance Comparison.

Method	ACC	PREC	REC	F1-S
Srivastava et al. [3]	89.7%	-	-	-
Scott et al. [23]	-	71.3%	86.3%	-
CNN_7 [25]	-	68.8%	83%	-
Adaboost_CNN [24]	-	71.3%	86.3%	-
"Bagging_CNN [24]	-	72.4%	92.4%	-
DTLDN-CBIRA [27]	-	-	81.9%	89.9%
Xception	93.1%	93.2%	93.1%	93.0%
MobileNet	81.8%	85.3%	81.8%	81.8%
Inception	91.7%	92.0%	91.7%	91.8%
EL	94.7%	95.0%	94.6%	94.6%

residual connections. This structure aids in reducing memory requirements and computational costs. By dividing the separable convolution in Xception, space-wise and channel-wise features are learned, resolving representational bottlenecks and vanishing gradients.

The Inception model, another CNN, is specifically used for extracting characteristics from query images as well as database images. It optimizes the network by factoring convolutions into distinct branches that operate on space and channels in succession. This allows the model to learn multiscale representations while reducing the overall number of restrictions and computational complexity.

MobileNet is another model we use, which provides a balance between computational efficiency and model accuracy. It is particularly useful for applications that require lightweight models for deployment on devices with limited computational resources. Lastly, we employ Ensemble Learning to combine the strengths of the individual models and improve the overall performance. Ensemble Learning helps to increase the robustness and stability of our CBIR system, leading to improved accuracy.

Each of these models has demonstrated high accuracy in our tests, with some room for improvement in certain classes. For example, the Xception model achieved an overall accuracy of 0.93, while the Inception model had an overall precision and recall of 0.92. MobileNet achieved an accuracy of 0.82, and the Ensemble Learning model achieved an impressive overall accuracy of 95%. These results validate our choice of DL models, demonstrating their effectiveness in the CBIR task. However, we acknowledge that there is scope for improvement in some classes, and further analysis may be needed to understand why this is the case and how to improve the model's performance for these classes.

VI. Conclusions and Future Work

In conclusion, this study presents a new approach for feature extraction in Content-Based Image Retrieval (CBIR) using three state-of-the-art pre-trained deep learning architectures: Xception, Mobilenet, and Inception combined together using a hard voting ensemble approach. The approach was tested using a practical and challenging dataset called CBIR_50 and showed improved ACC, PREC, REC, and F1-S compared to other methods. In addition, the experiments in this paper showed that the performance of combing these three architectures using ensemble learning exceeded the performance of each architecture applied on the same number of classes of CBIR_50 dataset. The results of the experiments showed that all four methods achieved high performance, with Ensemble Learning outperforming the others with a 95% ACC, PREC, REC, and F1-S. These results demonstrate the effectiveness of the proposed approach in CBIR and the robustness of Xception, Inception, MobileNet, and Ensemble Learning in image classification tasks. This research highlights the potential for these methods in various applications and further research in this field.

For the future work, the proposed approach could be implemented on other datasets and domains to test its robustness and generalizability. Additionally, further research could involve combining the proposed approach with other image retrieval techniques, such as text-based or hybrid methods, to improve overall performance. Another potential avenue of exploration would be to apply the ensemble method to other DL models to enhance their performance. Furthermore, analyzing the performance of the proposed approach on large-scale datasets and improving its computational efficiency can also be a future work.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (cbir)," *Neurocomputing*, vol. 452, pp. 675–689, 2021.

[2] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28–42, 2022.

[3] P. Srivastava and A. Khare, "Integration of wavelet transform, local binary patterns and moments for content-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 42, pp. 78–103, 2017.

[4] X. Zhang, H. Zhai, J. Liu, Z. Wang, and H. Sun, "Real-time infrared and visible image fusion network using

adaptive pixel weighting strategy," *Information Fusion*, p. 101863, 2023.

[5] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3804–3814, 2021.

[6] M. J. J. Ghrabat, G. Ma, I. Y. Maolood, S. S. Alresheedi, and Z. A. Abduljabbar, "An effective image retrieval based on optimized genetic algorithm utilized a novel svm-based convolutional neural network classifier," *Human-centric Computing and Information Sciences*, vol. 9, pp. 1–29, 2019.

[7] S. G. More and I. Mohammed, "Survey on cbir using k-secure sum protocol in privacy preserving framework'," *International Journal of Computer Science and Information Security, IJCSIS*, pp. 184–188, 2015.

[8] M. Yousuf, Z. Mehmood, H. A. Habib, T. Mahmood, T. Saba, A. Rehman, M. Rashid, and etal, "A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval," *Mathematical Problems in Engineering*, vol. 2018, 2018.

[9] K. Meethongjan, M. Dzulkifli, A. Rehman, A. Altameem, and T. Saba, "An intelligent fused approach for face recognition," *Journal of Intelligent Systems*, vol. 22, no. 2, pp. 197–212, 2013.

[10] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel, "Fpdetective: dusting the web for fingerprinters," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 1129–1140, 2013.

[11] E. Balsa, C. Troncoso, and C. Diaz, "Ob-pws: Obfuscation-based private web search," in *2012 IEEE Symposium on Security and Privacy*, pp. 491–505, IEEE, 2012.

[12] R. L. Lagendijk, Z. Erkin, and M. Barni, "Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 82–105, 2012.

[13] G.-H. Liu, Z.-Y. Li, L. Zhang, and Y. Xu, "Image retrieval based on micro-structure descriptor," *Pattern Recognition*, vol. 44, no. 9, pp. 2123–2133, 2011.

[14] B. Zafar, R. Ashraf, N. Ali, M. K. Iqbal, M. Sajid, S. H. Dar, and N. I. Ratyal, "A novel discriminating and relative global spatial image representation with applications in cbir," *Applied Sciences*, vol. 8, no. 11, p. 2242, 2018.

[15] W. Wei and Y. Wang, "Color image retrieval based on quaternion and deep features," *IEEE Access*, vol. 7, pp. 126430–126438, 2019.

[16] Z. Xia, X. Wang, L. Zhang, Z. Qin, X. Sun, and K. Ren, "A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing," *IEEE transactions on information forensics and security*, vol. 11, no. 11, pp. 2594–2608, 2016.

[17] R. Khan, C. Barat, D. Muselet, C. Ducottet, *et al.*, "Spatial orientations of visual word pairs to improve bag-of-visual-words model.," in *BMVC*, pp. 1–11, 2012.

[18] Y. Song, I. V. McLoughlin, and L.-R. Dai, "Local coding based matching kernel method for image classification," *PloS one*, vol. 9, no. 8, p. e103575, 2014.

[19] R. Ashraf, T. Mahmood, A. Irtaza, and K. Bajwa, "A novel approach for the gender classification through trained neural networks," *J. Basic Appl. Sci. Res*, vol. 4, pp. 136–144, 2014.

[20] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *International journal of computer vision*, vol. 107, pp. 20–39, 2014.

[21] M. Zang, D. Wen, T. Liu, H. Zou, and C. Liu, "A pooled object bank descriptor for image scene classification," *Expert Systems with Applications*, vol. 94, pp. 250–264, 2018.

[22] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 4, pp. 549–553, 2017.

[23] G. J. Scott, R. A. Marcum, C. H. Davis, and T. W. Nivin, "Fusion of deep convolutional neural networks for land cover classification of high-resolution imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 9, pp. 1638–1642, 2017.

[24] Y. Xu, Q. Lin, J. Huang, and Y. Fang, "An improved ensemble-learning-based cbir algorithm," in *2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC)*, pp. 1–3, IEEE, 2020.

[25] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on grassmann manifold with application to video based face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 140–149, 2015.

[26] A. L. Lafta and A. I. Abdulsada, "Privacy-preserve content-based image retrieval using aggregated local features.," *Iraqi Journal for Electrical & Electronic Engineering*, vol. 18, no. 2, 2022.

[27] M. Karthikeyan and D. Raja, "Deep transfer learning enabled densenet model for content based image retrieval in agricultural plant disease images," *Multimedia Tools and Applications*, pp. 1–24, 2023.

[28] A. Mehbodniya, J. Webber, A. G. Devi, R. P. Somineni, M. C. Chinnaiah, A. Asokan, and K. S. Bhanu, "Content-based image recovery system with the aid of median binary design pattern.," *Traitement du Signal*, vol. 40, no. 2, 2023.

[29] T. Gherbi, A. Zeggari, Z. A. Seghir, and F. Hachouf, "Entropy-guided assessment of image retrieval systems: Advancing grouped precision as an evaluation measure for relevant retrievability," *Informatica*, vol. 47, no. 7, 2023.

[30] G. K. Raju, P. Padmanabham, and A. Govardhan, "Enhanced content-based image retrieval with trio-deep feature extractors with multi-similarity function.," *International Journal of Intelligent Engineering & Systems*, vol. 15, no. 6, 2022.

[31] B. Sreenivasulu, A. Pasala, and G. Vasanth, "Adaptive inception based on transfer learning for effective visual recognition," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 6, pp. 1–10, 2020.

[32] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional lstms," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1078–1086, 2017.

[33] M. Yasmin, M. Sharif, and S. Mohsin, "Neural networks in medical imaging applications: A survey," *World Applied Sciences Journal*, vol. 22, no. 1, pp. 85–96, 2013.

[34] A. Jimenez, J. M. Alvarez, and X. Giro-i Nieto, "Class-weighted convolutional features for visual instance search," *arXiv preprint arXiv:1707.02581*, 2017.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[36] A. Mahabub, M. I. Mahmud, and M. F. Hossain, "A robust system for message filtering using an ensemble machine learning supervised approach," *ICIC Express Letters, Part B: Applications*, vol. 10, no. 9, pp. 805–811, 2019.