# Perception of Natural Scenes: Objects Detection and Segmentations using Saliency Map with AlexNet

Muhammad Waqas Ahmed
Faculty of Computing and AI
Air University, Pakistan
222633@students.au.edu.pk

Abdulwahab Alazeb
Department of Computer Science
Najran University, Saudi Arabia
afalzeb@nu.edu.sa

Naif Al Mudawi
Department of Computer Science,
Najran University, Saudi Arabia
nalmudaw@nu.edu.sa

Touseef Sadiq
Department of Information and Communication Technology
University of Agder, Norway
touseef.sadiq@uia.no

Bayan Alabdullah
Department of Information Systems,
Princess Nourah bint Abdulrahman University, Saudi Arabia
bialabdulah@pnu.edu.sa

Hammed ur Rehman
Faculty of Computing and AI
Air University, Pakistan
hameed.rahman@mail.au.edu.pk

Asaad Algarni
Department of Computer Sciences
Northern Border University, Saudi Arabia
Asad.Algarni@nbu.edu.sa

Ahmad Jalal
Faculty of Computing and AI
Air University, Pakistan
ahmadjalal@mail.au.edu.pk

**Abstract:** *Object detection and classification play a crucial role in accurately tracking objects in complex environments. In recent years, there has been a significant increase in interest among researchers towards object analysis, fueled by the necessity to address challenges and explore opportunities across diverse technological domains. This study introduces a methodologically novel method for image classification through a custom-designed architecture inspired by AlexNet, tailored to process feature vectors for improved pattern recognition. The methodology incorporates Density-Based Spatial Clustering of Applications with Noise (DBSCAN) segmentation to partition images into meaningful regions, showcasing computational efficiency. Additionally, saliency mapping highlights visually significant areas within these segmented images. Various feature extraction methods, including Maximally Stable Extremal Regions (MSER), Binary Robust Invariant Scalable Keypoints (BRISK), and Wavelet transform, are employed to capture unique structures within the images. These features are then fused and optimized using the Fish Swarm Algorithm (FSA), a nature-inspired optimization technique. The refined features, enhanced through the FSA process, are input into a modified AlexNet architecture, enhancing image classification accuracy. The evaluation metrics used include accuracy, precision, recall, and F1-score, providing a comprehensive assessment of performance. The proposed model achieved a classification accuracy of 95.65% on the VOC 2012 dataset, outperforming contemporary methods by a margin of 2-5%, and 93.66% and 92.71% on Caltech-101 and Microsoft Common Objects in Context (MS COCO) datasets, respectively. This innovative blend of techniques harnesses the strengths of FSA and deep learning, yielding precise and robust classification outcomes, outperforming many contemporary methods on datasets like VOC 2012, Caltech 101, and MS COCO*.

**Keywords:** *Pattern recognition, alexNet, fish swarm algorithm, object detection*.

*Received July 11, 2024; accepted March 12, 2025*
*https://doi.org/10.34028/iajit/22/3/4*

## 1. Introduction

A few of the fundamental challenges of computer vision in visual recognition are related to image classification [6, 7], object detection [15, 16], and segmentation [12]. Image classification deals with the recognition of the semantic classes of the objects that are existing in the image. To the contrary, object detection is about finding the objects in an image. It does so by drawing the bounding boxes around these objects in the images [1, 2, 3]. Primarily, segmentation is implemented to predict pixel-wise classifiers that assign a class character for each pixel to add more depth of understanding to a certain image. These activities together make up the foundation of visual comprehension in computer vision, making possible applications in different areas [21, 27].

In its early days, object detection functions via proposal generation, feature vector extraction, and region classification. Traditional methods were dedicated to designing feature descriptors and hence, producing embedding for the regions of interest, which demonstrated comparative outcomes on different datasets. Deep learning (DL) has a few categories that suit well for object detection. These DL models have multiple hidden layers. Features are extracted at each

layer, which refer to the fault characteristics, and then these features are used as inputs in the next layer [47]. DL also has the advantage of automatic transformation of low-dimensional features into high-dimensional representations, taking into account at the same time the non-linear relationship between input and output. DL can be utilized to build features without any prior information as well [11].

While compared to other types of Neural Network (NN) architectures, Convolutional Neural Networks (CNNs) are especially strong. This is due to the fact that CNNs can capture more informative, higher-level features and take advantage of the deep pixel-level correlation in input images [13]. After the successful application of Deep Convolutional Neural Networks (DCNN) for image classification, the development of object detection also accelerated as a result of deep learning methods. DCNNs by nature produce hierarchical characteristics by that convert raw pixels into high-level semantic information, autonomously learning from training data, and demonstrate improved discriminative prowess in complex situations. This led to object detection algorithms based on deep convolutional neural networks with end-to-end optimization and more powerful feature representation [45].

Nevertheless, the traditional detectors encountered issues such as overwhelming redundant proposals, large hand-crafted window scales, dependence on manually designed feature descriptors, and individual optimization for each detection at each and every one of these stages. Task challenges come from object appearance inherent variability, diversity of appearance environments, and the requirement for robustness in dealing with complex visual contexts. Thus, the creation of high-quality object detection and classification techniques is an important effort in the improvement of the capabilities of computer vision systems. In this paper, a new approach is presented that unites the fused segmentation and saliency mapping methodology, employing feature fusion methods and using the AlexNet architecture for efficient classification. The main contributions of this research are as follows.

1. Introduction of a novel image analysis pipeline that combines Density-Based Spatial Clustering of Applications with Noise (DBSCAN) segmentation with saliency mapping to focus on visually important regions within the segmented images, facilitating more efficient and effective feature extraction from the most relevant areas of interest.
2. Integration of Maximally Stable Extremal Regions (MSER), Binary Robust Invariant Scalable Key points (BRISK), and Wavelet Ttransform (WT) for robust feature extraction capture the distinctive structures within images through a fusion of complementary feature descriptors.
3. Proposition of a novel optimization framework that combines the Fish Swarm Algorithm (FSA) with deep learning techniques, specifically the ALEXNet architecture, for enhancing the performance of image classification by leveraging the strengths of both bio-inspired optimization and deep neural networks

The remaining sections of the paper are structured as follows: Section 2 explores previous related research works. Section 3 presents an overview of the proposed methodology, encompassing segmentation, saliency maps, feature extraction, and fusion techniques. Section 4 delves into the details of the datasets employed, the experimental setup, and the obtained results. Finally, section 5 concludes the study by summarizing the key findings and contributions.

## 2. Related Work

### 2.1. Object Segmentation

Image segmentation is the process of converting an image into a set of pixel regions denoted by a mask or labels over an image. This segmentation of an image into a pixel grid allows the processing of key portions [32]. There are many different methods introduced for the image segmentation process. Guo *et al*. [14] introduces a novel approach is based on k-means clustering, which focuses on the enhancement of segmentation pixels using color features. In order to find the mean value of distribution the gray value components are calculated for R, G, and B distributions. Kuan *et al*. [19] in their approach, utilizes region mean pooling for object detection, gathering contextual data from surrounding regions rather than focusing solely on individual objects. However, the computational complexity increases as large images are divided into a grid of cells, and a fixed context size proves unsuitable for all images. Song *et al*. [38] introduce a novel approach for addressing image segmentation and color separation challenges using Fuzzy C-Means (FCM) and graph cut algorithms. A distance-based model combining relative spatial information with visual distance and angles is introduced, enhancing object segmentation by facilitating quick identification and separation in complex environments. Yu *et al*. [53] proposes a fusion of the Region-based Convolutional Neural Network (RCNN) and region regression methods, employing an enhanced RCNN network to detect and classify diverse objects within images. While the suggested methodology exhibits favorable accuracy, there is a need for the appropriate replacement of candidate region sizes.

### 2.2. Object Detection and Classification

The object classification process tends to have many challenges for researchers, such as locating individual objects, analyzing and describing interactions between them, recognizing occluded objects, and grouping them effectively to achieve meaningful results [48]. Srikar and Malathi [40] introduces a novel approach to improving

object localization in web images by incorporating a rotation-invariant Histogram of Oriented Gradients (HOG) feature within a top-down searching technique. While the traditional HOG descriptor is effective, its lack of rotation invariance is addressed in this proposed method. The aim is to enhance performance in web image databases by mitigating issues related to rotation and scale dependencies associated with the HOG feature. Wei *et al*. [46] introduce a new approach for object detection framework that works around a contour shape descriptor and it provides high accuracy for detecting the objects even in cluttered image scenarios. Muralidharan and Chandrasekar [25] uses Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) for images classification purposes. SVM provides robust classification boundaries, while KNN enhances local decision-making, creating a powerful hybrid model for object recognition tasks. Ouadiay *et al*. [30] introduces a thorough methodology for identifying objects and approximating their postures. This involves creating bounding boxes around the specified objects during the training phase and, importantly, precisely locating each object within images during subsequent testing phases. The principal accomplishment of this study is the precise generation of bounding boxes throughout the training process, enhancing the accuracy of object localization in images during testing. Pramanik *et al*. [32] novel model is based on granulated region-based convolutional neural network units to detect multiple objects from a single image. The proposed multi-object detection approach consisted of two main

phases, object detection and class recognition. During the Granulated RCNN (G-RCNN) model execution, the most critical operation was defining the RoIs based on the granulation technique applied to the object regions. RoI extraction was performed in an unsupervised mode with granulation and spatio-temporal data as inputs It's the fact that this model considered only the positive ROIs during the training phase and this selective approach helped to enhance the model's effectiveness. Bo and Sminchisescu [8] introduces a unique class based approach for image classification. It makes use of the Gaussian Mixture Model (GMM) to define the characteristics of images within each category. Afterwards, they calculated the Euclidean distance between the features of the image and the GMM models and used these distances as representations of each image. In order to do so, they pooled the representations across all classes. This uniting characteristic aspects of the class and the visual approach allowed them to characterize an image by attributes, thus, capturing both semantic and visual information.

## 3. Methods

The methodological approach presented in this paper is an advanced image processing and classification system using a set of highly sophisticated techniques. During the preprocessing stage, an adoptive mean filter is applied, resulting in increased image quality by dynamically adjusting filter parameters based on local properties. Next, two segmentation methods are used the GMM and DBSCAN, in order to divide the image into meaningful regions. However, DBSCAN was computationally effective, and with better segmentation results we move forward with DBSCAN segmented images. Saliency mapping is subsequently used to emphasize visually important regions in the DBSCAN-segmented images, which simplifies further analysis. Feature extraction is done with the help of MSER, BRISK and Wavelet transform in order to capture the distinctive structures within the images. These feature points are then fused using the concatenation method and are optimized with the help of the Fish Swarm Algorithm for further processing. Ultimately, the features that are optimized are inserted into the ALEXNet architecture for image classification, combining the advantages of FSA and deep learning in order to obtain precise and robust classification results. The overview of the proposed model is shown in Figure 1.

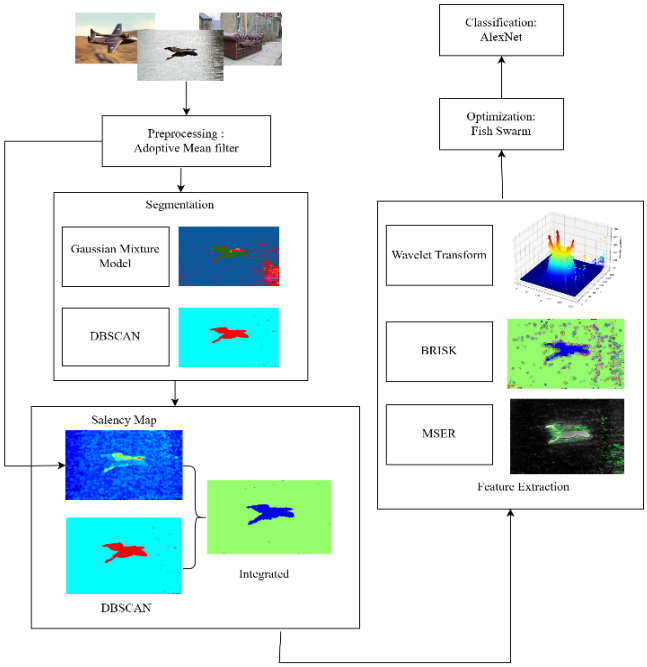



Figure 1. Structural diagram of our novel proposed model.

### 3.1. Preprocessing: Adaptive Mean Filter

First of all, we applied an adoptive mean filter on to the images as part of the preprocessing. It involves utilizing a filtering technique that dynamically adjusts its parameters based on the local characteristics of the image. Usually mean filters employ a fixed window size for smoothing; however, in the adoptive mean filter the window size is adopted according to the local content of the image [4, 17]. This adaptability enables the filter to better preserve edges and fine details while effectively reducing noise and enhancing overall image quality [54]. Mathematically, the adoptive mean filter can be

represented by Equation (1).

$$I(x,y) = 1/N\,(x,y) \sum_{i=x-w}^{x+w} \sum_{j=y-w}^{y=w} I(i,j) \quad (1)$$

where $I(x, y)$. is the pixel value at position $(x, y)$ in the filtered image, $I(i, j)$ represents the pixel value at position $(i, J)$ in the original image, $N(x, y)$ is a normalization factor, and w is the dynamically adjusted window size based on the local characteristics of the image. The adaptive mean filter is considered a beneficial technique in image preprocessing because it allows the balance between noise removal and detail retention. Figure 2 depicts the few of the preprocessed images.

a) Original image. b) Original image. c) Original image.

d) Preprocessed image. e) Preprocessed image. f) Preprocessed image.

Figure 2. Preprocessed images using adoptive mean filter.

## 3.2. Segmentation

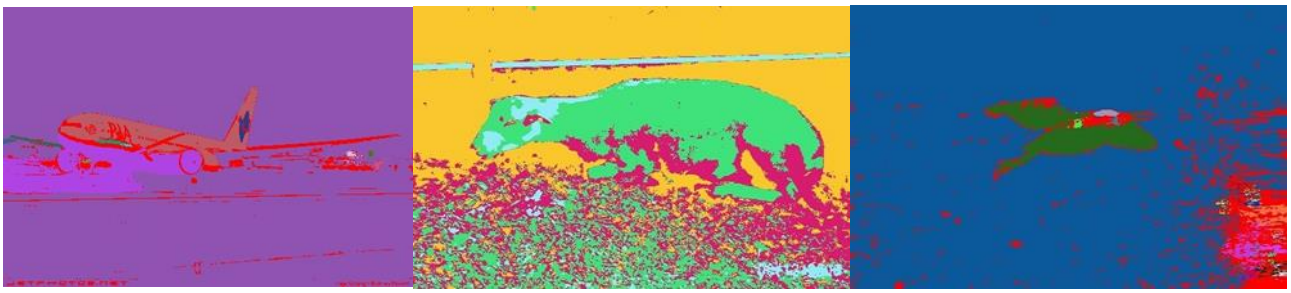
Next, we have applied segmentation to preprocesses images. Segmentation is the action of dividing an image into a series of meaningful and semantic homogeneous region. The objective is to divide an image into segments based on some criterion like color, grayness, texture, and other visual attributes, as it is one of the essential aspect of image detection and classification. So for the segmentation process, we have employed the GMM and DBSCAN. Our criteria for evaluation include computational time and segmentation accuracy. Through careful evaluation of these factors, we strive to identify a suitable approach that will serve as the basis for further analysis and improvement of the selected segmentation method.

### 3.2.1. Gaussian Mixture Model Segmentation

Gaussian Mixture model segmentation is a probabilistic model that is commonly used in computer vision for image segmentation. It assumes that the pixel intensities in an image can be approximated as a combination of several Gaussian distributions, each of which represents a distinct region [27, 34]. GMM segmentation is powerful in a way that it can work for complex and non-uniform intensity distributions within an image. It involves three main steps, including initialization, Expectation-Maximization (EM) algorithm, and classification. During the first steps the parameters including mean, covariance and weights of the Gaussian component are initialized. Next these parameters are adjusted by the EM algorithm in an iterative way in order to maximize the likelihood of the data [50] as shown in Equation (2).

$$P(x) = \sum_{i=1}^{K} \pi_{iN}(x|\mu i, \Sigma i) \quad (2)$$

where $P(x)$ is the probability density function, $\pi$ represents the weight of the *ith* Gaussian component, $N(x|\mu i,\Sigma i)$. is the Gaussian distribution with mean $\mu i$ and covariance $\sigma i$, and k is the number of components. In the last step of the GMM method each pixel is assigned to the Gaussian on the highest posterior probability. This operation provides an efficient separation of the image into regions based on the identified Gaussian distributions. GMM segmentation is insensitive to the Gaussian on the highest posterior probability. This operation provides an efficient separation of the image into regions based on the identified Gaussian distributions. GMM segmentation is insensitive to highly variable intensity patterns, and therefore it is applicable to images with various textures and structures. The performance of GMM segmentation mainly depends on the right initialization and on the number of components k. Moreover, the computational efficiency of GMM segmentation tends to be problematic in large datasets because of its iterative EM algorithm [29, 30]. Few of the segmented images are displayed in Figure 3.



a) GMM segmented image. b) GMM segmented image. c) GMM segmented image.

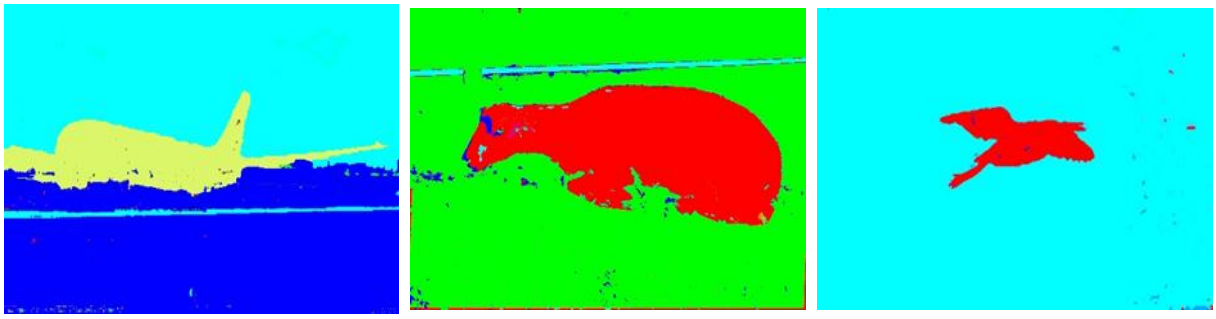
Figure 3. Few of the Gaussian mixture model segmented images.

### 3.2.2. Density-Based Spatial Clustering of Applications with Noise

DBSCAN is a clustering algorithm commonly used in image processing and computer vision applications to segment datasets by the density of data points. Compared to the traditional clustering algorithms, DBSCAN does not need a priori information about the number of clusters in the dataset, which is why it proves especially convenient for the image segmentation in cases when number of objects are not known. It works by classifying data points into three categories: core points, border points, and noise points. Core points are the ones that have a designated number of neighbors within a radius, forming the core of a cluster. Border points have less number of neighbors than qualifying them as core points, but it falls within the radius of a core point. Outlier points are data points that are not part of any cluster [36, 43] as depicted in (3).

$$R = \{q \in D \mid dist(p,q) \leq \varepsilon\} \tag{3}$$

where $D$ is the dataset, $p$ is a core point, $q$ is a data point in the region, and *dist* is the distance function. DBSCAN is competitive for its ability to find clusters of any form, its insensitivity to noise, and with its flexibility in terms of cluster densities. Nevertheless, it may have difficulties with groups of very different densities and is very sensitive to the choice of parameters; therefore, careful tuning of the parameters is needed. Few of the DBSCAN segmented images are depicted in Figure 4.



a) DBSCAN segmented image. b) DBSCAN segmented image. c) DBSCAN segmented image.

Figure 4. Few of the segmented images using DBSCAN.

The effectiveness of a segmentation method is assessed using the Intersection over Union (IoU) metric. *IoU* is a commonly utilized measure for determining the precision of segmentation models. It evaluates accuracy by calculating the overlap between the predicted segmentation and the ground truth. This score is obtained by dividing the area of their intersection by the area of their union, as expressed in (4).

$$IOU = \frac{Area\ of\ intersection}{Area\ of\ Union} \tag{4}$$

After careful evaluation it was found that DBSCAN, besides being very efficient in computations, does not sacrifice the quality of segmentation results as shown in Table 1. The rapid speed and the algorithm capacity to effectively work with the datasets of changing density and irregularly shaped clusters make it a better choice for the next stages of our model.

Table 1. Evaluation of computational time and segmentation accuracy.

| Datasets | Comp. Time (GMM) | Comp. Time (DBSCAN) | IOU (GMM) | IOU (DBSCAN) |
|---|---|---|---|---|
| **VOC 2012** | 171.13s | 152.3s | 82.9% | 89.7% |
| **Caltech 101** | 163.32s | 147.7s | 84.5% | 88.9% |
| **MS COCO** | 1605.29s | 150.1s | 83.9% | 89.1% |

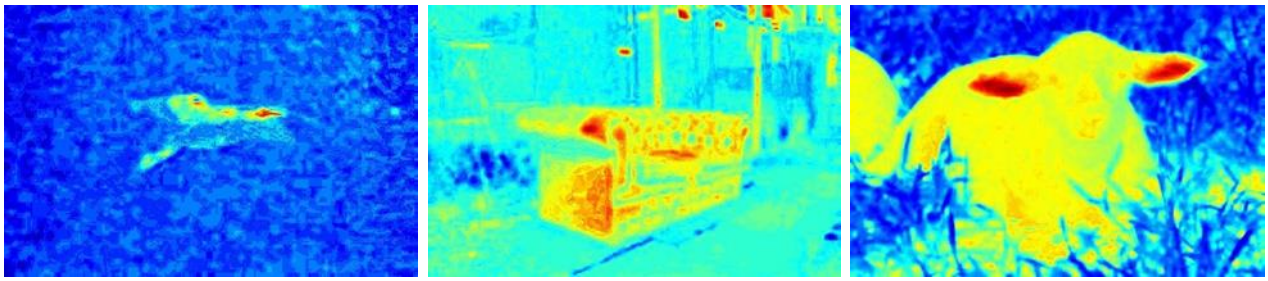
## 3.3. Saliency Mapping

Although DBSCAN is effective in detecting clusters in the data with different densities, it sometimes fails to cluster objects with complex structures, also objects with visual differences and the algorithm sometime can merge them into only one cluster. Saliency mapping was used to overcome the drawbacks of DBSCAN segmentation. Saliency maps point out the most salient areas in an image, which facilitates clearer recognition of important objects. The purpose is to maximize the accuracy of the segmentation results by focusing on the parts that are likely to be ignored or wrongly associated

by the initial clustering algorithm [5]. The Saliency Map is resulting from computational models that mimic the possible human visual attention mechanism. It scores each picture in terms of pixel importance and based on the factors such as color variability, brightness, and spatial impact [24, 39]. The saliency score $s(x,y)$ for pixel coordinates $(x, y)$ can be calculated using Equation (5).

$$s(x,y) = f(I(x,y), intensity, spatial\ frequencies \quad (5)$$

where, $s(x, y)$ represents the intensity value of the pixel, and f is a function that encapsulates the combination of intensity and spatial features contributing to the saliency score as displayed in Figure 5.



a) Saliency map. b) Saliency map. c) Saliency map.

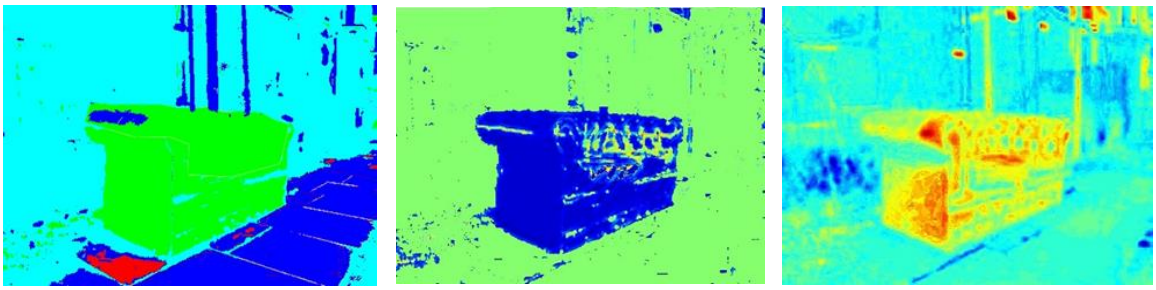
Figure 5. Saliency map applied on the preprocessed images.

### 3.4. Image Integration

Saliency Map is now integrated with the DBSCAN-segmented image. This integration is achieved through a pixel to pixel multiplication. The corresponding pixels from Saliency Map and the DBSCAN-segmented image, determine the resulting image, where at each position, the pixels are the product of the corresponding pixels in Saliency Map and the DBSCAN-segmented image as shown in Equation (6).

$$Ii(x,y) = Is(x,y) \times Id(x,y) \quad (6)$$

where, $Ii(x, y)$ and $Is(x, y)$ represent the pixel values at coordinates $(x, y)$ in the Saliency Map and the DBSCAN-segmented image, respectively. The result of integration is displayed in Figure 6.



a) DBSCAN segmented image. b) Integrated image. c) Saliency map Image.

Figure 6. Image integration process on the top left DBSCAN image, on the top right is saliency map and at bottom middle is integrated image.

### 3.5. Feature Extraction

For the feature extraction process, we used three distinct feature extraction methods MSER, BRISK, and Wavelet transform. The resulting features from these complementary descriptors were subsequently fused, facilitating a robust and comprehensive representation of the image characteristics.

#### 3.5.1. MSER Feature Extraction

Adaptation of MSER for feature extraction on saliency-mapped images increases the stability by searching for relatively stable regions in images over a range of scales. MSER stands out is its ability to retain the information of unique structures, objects and borders. MSER also identifies local areas that remain unchanging even though a number of threshold levels are varied showing parts of the image with the same degree of intensity. The criterion of stability guarantees that the regions, discovered in images, maintain their stability over varying intensity levels. Also, the parameters of MSER allow for choosing the level of detail of the detected regions. The minimum and maximum stability thresholds limit the parameters fulfilled during region extraction [23, 42] as shown in Equation (7).

$$R(t) = \{(x,y) \in I \mid I(x,y) \geq t\} \quad (7)$$

Where, $I$ is the integrated image and $R(t)$ represent the region at threshold $t$.

#### 3.5.2. BRISK Feature Extraction

We next applied BRISK which is well-known to be a fast, scale and rotation invariant image feature extraction algorithm. It determines the key points of the image which is used as a feature descriptor to get reliable feature matching and recognition [22]. The integration of BRISK into the feature extraction process of conjoined images enhances its ability to locate individual distinctive points in the visually significant portions of

the image and, consequently, construct a more comprehensive representation of the important features that are within an image and can be calculated using Equation (8).

$$Di = sign(I(x1, y1) - I(Pi))sign(I(x2, y2) - I(Pi)) \dots sign(I(xn, yn) - I(Pi)) \quad (8)$$

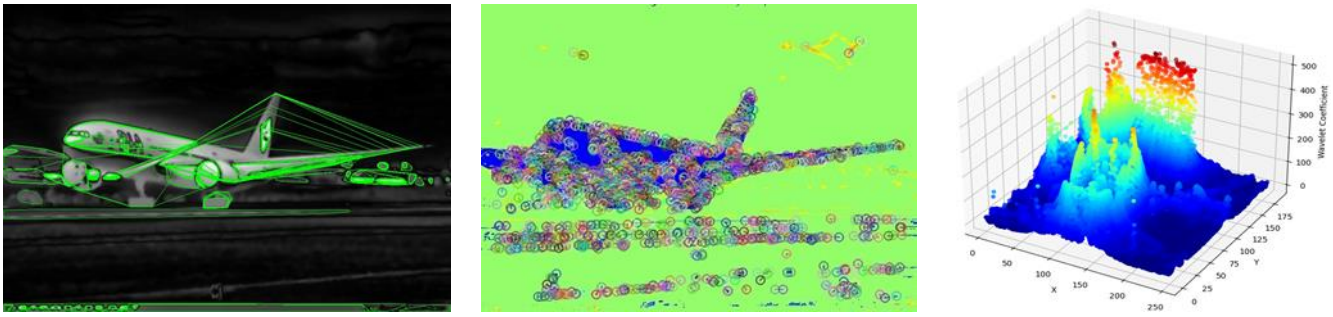
where *I* is the integrated image, *Pi* represents a pixel in the image. The BRISK *Di* is computed by comparing the intensity differences at various sampled points (*xj, yj*) within a circular region around *Pi*.

### 3.5.3. Wavelet Transform Feature Extraction

Next we applied Wavelet Transform for feature extraction to the integrated images which helps to improve the capability of capturing both the global and local information at different frequency scales. It separates an image into various frequency components, permitting a multi-resolution analysis of its content as shown in Equation (9).

$$W(a,b) = \sum_m \sum_n s(m,n)\psi a, b(m,n) \quad (9)$$

Here, *W(a,b)* signifies the wavelet coefficients, *s(m,n)* represents the pixel values of the saliency-mapped image, and *ψa, b(m,n)* denotes the 2D wavelet function, which varies based on the scale a and position b. This equation encapsulates the transformation of the integrated image into its wavelet domain, where the resulting coefficients preserve both global and local features [52]. The extracted features from all three methods are displayed in Figure 7.




a) MSER feature extraction. b) BRISK feature extraction. c) Wavelet transform.

Figure 7. Feature extraction using MSER, BRISK and wavelet transform.

## 3.6. Feature Fusion

Our feature fusion involves combining the extracted features from BRISK, MSER, and Wavelet Transform, into a single feature vector using the concatenation method. The resulting feature vector becomes a holistic representation of the image, incorporating both local key point descriptors, stable extremal region information, and multi-resolution spectral details and can be calculated using Equation (10).

$$Fconcat = [FBRISK, FMSER, F\,Wavelet] \quad (10)$$

where, *Fconcat* represents the concatenated features, *FBRISK*, *FMSER* and *Fwavelet* represent the feature vectors obtained from BRISK, MSER, and Wavelet Transform, respectively. Next on this feature vector normalization is carried out so that all the components of the feature contribute equally to the overall analysis, which minimizes the difference in scale and magnitude between features [24].

## 3.7. Optimization and Categorization: Fish Swarm Algorithm

Fish Swarm Algorithm is a bioinspired optimization algorithm that mimics the collective behavior of fish schools. Simulating the interactions among fish, FSA endeavors to effectively search and utilize the search space for optimal solutions. In the context of feature optimization, FSA is used to optimize feature vector derived from the previous processing [10, 31, 41]. Initially, the algorithm sets the positions of the swarm of fish, as a representation of potential feature configurations in the concatenated and normalized vector. At the same time, initial velocities are assigned which include an inertia weight (*w*) and random acceleration coefficients ($c_1$) and ($c_2$). They thereby introduce a critical tradeoff between exploration and exploitation, and sustain wide searching throughout the solution space [51]. The iterative update of fish positions *Xi*(*t*+1) is defined by using Equation (11).

$$xi(t+1) = xi(t) + vi(t+1) \quad (11)$$

where *Xi*(*t*) denotes the position of fish and *Xi(t +1)* is the velocity vector determined by the velocity update as shown in Equation (12).

$$vi(t+1) = w \cdot vi(t) + c1 \cdot r1 \cdot (pi(t) - xi(t)) + c2 \cdot r2 \cdot (g(t) - xi(t)) \quad (12)$$

Here, *r1*, *r2* are random values between 0 and 1, *Pi(t)* represents the personal best position of fish *i*, and *g(t)* signifies the global best position among all fish in the swarm. Moreover, in order to make sure the convergence and control computational resources of the FSA several termination criteria have been implemented. Firstly, the algorithm may cease after a predefined number of iterations. Additionally, a convergence check is incorporated, monitoring changes in the fitness function of the swarm. Termination also occurs if there is negligible improvement over a set number of iterations,

indicative of convergence. Furthermore, an objective function threshold minimum objective value is defined which leads to termination when the fitness of the feature configuration falls below determined threshold [39]. The object categorization of different classed is displayed in Figure 8.

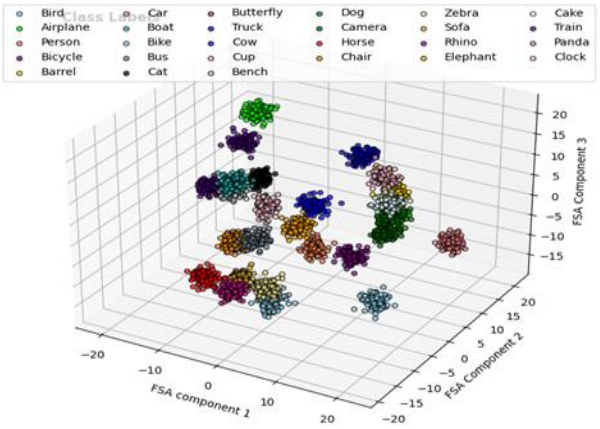



Figure 8. Object categorization using fish swarm algorithm.

### 3.8. AlexNet for Image Classification

For our image classification task, we propose an AlexNet-inspired deep convolutional neural network architecture tailored to learn patterns from feature vectors and classify images accurately. Instead of directly processing raw image data, our model will take feature vectors as input, enabling it to leverage the rich information extracted through prior feature engineering techniques [29, 49]. The model contains five convolutional layers and three fully connected layers. An optimized feature vector of 500x375 is fed to the input layer of AlexNet instead of conventional image input. The first convolutional layer, Conv1, contains 32 kernels of size 5x5, with a stride of 1 and padding of 2 is applied to the input feature vector. A ReLU activation layer is applied next, which introduces non-linearity by converting all negative values to zero while retaining the positive ones. This helps the network capture non-linear patterns in the data. Following this, a max-pooling layer with a 3x3 kernel and a stride of 2 is used, which helps in reducing the spatial dimensions of the feature maps while retaining the most salient features by selecting the maximum value within each pooling window. Subsequently, the Conv2 layer comprises 64 convolutional filters, each with dimensions of 5x5. These filters operate with a stride of 1, and padding of 2 is applied. Following the convolutional operation, a Rectified Linear Unit (ReLU) activation function is employed, succeeded by a max-pooling layer. This max-pooling layer utilizes a 3x3 kernel size and a stride of 2. The next layer, Conv3, encompasses 128 convolutional filters of size 3x3. These filters are applied with a stride of 1 and padding of 1. After the convolution, a ReLU activation function is employed, followed by another max-pooling layer with a 3x3 kernel size and a stride of 2. The fourth and fifth convolutional layers, Conv4 and Conv5, feature an increased number of filters, with 256 and 512 filters, respectively. These filters share the same dimensions of 3x3, and they are applied with a stride of 1 and padding of 1. These layers will be interspersed with ReLU activations and max-pooling layers, as in the original AlexNet structure. After the convolutional layers, our network will employ three fully connected layers (FC6, FC7, and FC8). The FC6 layer will have 4096 neurons, followed by a ReLU activation and dropout regularization. The FC7 layer will also have 4096 neurons with ReLU activation and dropout. The final layer which is FC8, contains 10 neurons which represent the 10 classes in our dataset, and it will utilize a softmax activation function for classification. Our model is optimized for handling input feature vectors of size 500x375 and is efficient for pattern learning, resulting in precise predictions across the 10 object classes under consideration.

## 4. Result and Analysis

We evaluated the ability of our model using three datasets: Caltech-101, PASCAL VOC2012, and MS COCO. Each dataset was divided into 70:30 ratios for training and testing purposes. Ten object classes were selected from each dataset for training and testing our model. The experiments were performed on a computing system equipped with an Intel Core i3 processor running at 1.7 GHz and Windows 10. The implementation and analysis of our approach were done using the Python and its relevant libraries.

### 4.1. Dataset Description

#### 4.1.1. MS COCO Dataset

The MS COCO [44] dataset is a well-known and widely used for detecting and classify objects. The dataset contains a total of 328,000 images, which depict 91 different object classes. For the purpose of performance validation in this study, 10 specific object classes were selected from the MS COCO dataset, namely: person, car, bus, truck, bench, bird, zebra, cup, cake, and clock.

#### 4.1.2. VOC 2012 Dataset

Another commonly used dataset for object detection and classification tasks is the PASCAL Visual Object Classes dataset [28], which includes 21,738 images across twenty classes. These classes cover a wide range of objects, including persons, vehicles, animals, and indoor objects. For the purpose of validating object detection performance in this study, a subset of 10 object classes was chosen from the PASCAL VOC dataset. These classes include cat, cow, dog, horse, sheep, bicycle, boat, train, bird, and sofa, representing a diverse range of objects encountered in various real-world scenarios.

### 4.1.3. Caltech 101 Dataset

The Caltech-101 dataset [36] is a widely recognized benchmark in computer vision for object recognition tasks. It encompasses 101 diverse object categories, ranging from animals and vehicles to everyday objects. In this study we used 10 classes form the dataset airplane, Barrel, Bike, Butterfly, Cup, Camera, Chair, Rhino, Elephant and Pandas.

## 4.2. Performance Evaluation

The confusion matrix results, which are presented in Tables 2, 3, and 4, were obtained by comparing the actual and predicted values using accuracy metric. Accuracy, along with precision, recall, and F1-score [20, 35], was employed to comprehensively evaluate the performance of the proposed model. Accuracy measures the proportion of correctly classified samples among all samples and is mathematically defined by Equation (13).

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (13)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Precision, which quantifies the ratio of true positive predictions among all predicted positives, is calculated using Equation (14).

$$Precision = TP/(TP + FP) \quad (14)$$

Recall, on the other hand, represents the ratio of true positive predictions among all actual positives as can be calculated using Equation (15).

$$Recall = TP/(TP + FN) \quad (15)$$

Finally, the F1-score, which provides a harmonic mean of precision and recall, can be calculated using Equation (16).

$$F1 - Score = 2 \times ((precision * recall))))/((precision + recall)) \quad (16)$$

The results depict that the model performed very well for classifying objects across all three datasets. The core of our model's effectiveness lies in the systematic integration of sophisticated methods across the processing pipeline. The results of the model depict the effect of the feature extraction techniques such as MSER, BRISK, and Wavelet transform. The resultant features fusion, fine-tuned by the fish swarm algorithm, improves the model's ability learn complex patterns in the images. This shows that our model is robust, can find and learn complex patterns from the images and can achieve great accuracy for classifying the objects.

Table 2. Confusion matrix for object classification on VOC 2012 dataset.

| Classes | Ct | Cw | Dg | He | Sp | Be | Bt | Tn | Bd | Sa |
|---|---|---|---|---|---|---|---|---|---|---|
| **Cat (Ct)** | 96.6 | 0.00 | 2.00 | 0.00 | 1.4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Cow (CW)** | 0.00 | 96.2 | 1.5 | 1.4 | 0.90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Dog (DG)** | 4.5 | 0.00 | 91.4 | 1.30 | 2.8 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Horse (He)** | 1.60 | 3.2 | 1.5 | 93.7 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Sheep (Sp)** | 01.00 | 2.20 | 2.00 | 0.60 | 94.2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Bicycle (Be)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 91.2 | 6.2 | 2.60 | 0.00 | 0.00 |
| **Boat (Bt)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 91.0 | 2.9 | 2.70 | 3.4 |
| **Train (Tn)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.60 | 92.7 | 0.00 | 3.70 |
| **Bird (Bd)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.70 | 0.00 | 93.1 | 4.20 |
| **Sofa (Sa)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.50 | 1.00 | 0.00 | 96.5 |

Table 3. Confusion matrix for object classification on MS COCO dataset.

| Classes | Pn | Cr | Bs | Tk | Bh | d | Za | Cp | Ce | Ck |
|---|---|---|---|---|---|---|---|---|---|---|
| **Person (pn)** | 96.72 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.28 | 0.00 | 0.00 | 0.00 |
| **Car (cr)** | 0.00 | 96.84 | 1.74 | 1.42 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Bus (Bs)** | 0.00 | 1.60 | 96.54 | 1.86 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Truck (Tk)** | 0.00 | 2.04 | 3.48 | 94.48 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Bench (Bh)** | 0.00 | 0.00 | 1.08 | 0.86 | 98.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Bird (Bd)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 97.75 | 0.00 | 1.75 | 0.50 | 0.00 |
| **Zebra (Za)** | 1.65 | 0.00 | 0.00 | 0.00 | 0.78 | 0.00 | 97.57 | 0.00 | 0.00 | 0.00 |
| **Cup (Cp)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 97.33 | 0.89 | 1.78 |
| **Cake (Ce)** | 0.00 | 0.00 | 0.00 | 0.00 | 4.61 | 6.48 | 0.00 | 7.80 | 72.66 | 8.45 |
| **Clock (Ck)** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 | 1.74 | 97.33 |

Table 4. Confusion matrix for object classification on Caltech-101 dataset.

| Classes | Ae | Bl | Bk | By | Cp | Ca | Cr | Ro | Et | Pa |
|---|---|---|---|---|---|---|---|---|---|---|
| **Airplane (Ae)** | 93.5 | 0.0 | 2.4 | 4.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Barrel (Bl)** | 0 | 94.7 | 0.0 | 0.0 | 3.2 | 2.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Bike (Bk)** | 4.9 | 0.0 | 92.2 | 0.0 | 0.0 | 2.9 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Butterfly (By)** | 4.2 | 0.0 | 2.5 | 93.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Cup (Cp)** | 0.0 | 6.9 | 0.0 | 0.0 | 89.9 | 4.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Camera (Ca)** | 0.0 | 0.0 | 2.0 | 0.0 | 3.6 | 93.7 | 0.0 | 0.7 | 0.0 | 0.0 |
| **Chair (Cr)** | 0.0 | 9.1 | 0.0 | 0.0 | 0.0 | 1.2 | 89.7 | 0.0 | 0.0 | 0.0 |
| **Rhino (Ro)** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 90.8 | 6.7 | 2.5 |
| **Elephant (Et)** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.6 | 92.2 | 1.2 |
| **Panda (Pa)** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.2 | 1.7 | 97.1 |

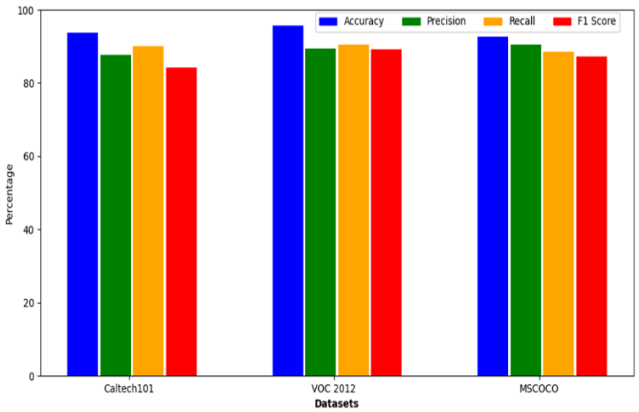



Figure 9. Evaluation matrix results on all three datasets.

Table 5. Object detection accuracy for VOC, Caltech101 and MS COCO datasets.

| VOC 2012 | Acc. | Caltech101 | Acc. | MS COCO | Acc. |
|---|---|---|---|---|---|
| **Bird** | 89.7 | **Airplane** | 90.1 | **Person** | 83.2 |
| **Bicycle** | 86.4 | **Barrel** | 93.3 | **Car** | 89.7 |
| **Boat** | 88.1 | **Bike** | 89.3 | **Bus** | 90.2 |
| **Cat** | 91.5 | **Butterfly** | 87.8 | **Truck** | 90.6 |
| **Cow** | 92.3 | **Cup** | 95.2 | **Bench** | 87.7 |
| **Dog** | 91.6 | **Camera** | 94.7 | **Bird** | 86.1 |
| **Horse** | 88.1 | **Chair** | 89.3 | **Zebra** | 87.9 |
| **Sofa** | 90.2 | **Rhino** | 87.1 | **Cup** | 93.9 |
| **Sheep** | 91.9 | **Elephant** | 89.4 | **Cake** | 90.4 |
| **Train** | 87.3 | **Panda** | 88.3 | **Clock** | 89.1 |
| **Mean** | **89.7** | **Mean** | **90.4** | **Mean** | **88.8** |

Figure 9 presents the accuracy, precision, recall, and F1 scores for our proposed object recognition model evaluated on three widely-used datasets: The datasets are Caltech 101, VOC 2012, and MS COCO. On Voc 2012 dataset our model outperformed the SOTA model with accuracy of 95.65, a precision of 89.30, a recall of 90.51, and an F1 score of 89.17, which means that the model effectively finds most of the objects in the dataset. Our model on Caltech 101 has shown excellent performance with 93.66 accuracy, 87.76 precision, 90.12 recall and 84.1 F1 score which means a good trade-off between precise object detection and localization. Moreover, our model was able to achieve high scores of 92.71 accuracy, 90.38 precision, 88.45 recall, and 87.23 F1 score when tested on the difficult MS COCO dataset with diverse object classes and complex scenes, hence validating its capability to detect objects with minimal false positives while comprehensively capturing the objects. Overall, the evaluation findings highlight the strength and effectiveness of our proposed object recognition framework on different datasets. The combination of DBSCAN segmentation, salecny map, feature fusion, and AlexNet based classification has been found to be reliable in the very accurate detection and classification of objects. The object detection accuracy of our model across all three datasets is shown in Table 5. The mean detection accuracy across VOC 2012, Caltech101 and MS COCO is 89.7%, 90.4%, and 88.8% respectively. Among all objects Cup class achieved the highest accuracy of 95%. This result emphasizes the strength of our feature selection methods, and then the impact that the fish swarm algorithm made on feature optimization. The uniform high accuracy among diverse object classes confirms the performance improvement provided by our method for the object detection in diverse datasets.

Table 6 shows the comparison of AlexNet with other well-known classifiers as it can be seen that AlexNet, consistently outperforms alternative classifiers across all three datasets VOC 2012, Caltech-101, and MSCOCO. The high accuracy of AlexNet with rates of 95.65%, 93.66%, and 92.71% respectively shows its efficiency in image classification tasks. The results for random forest, CNN, and DBN classifiers were reproduced by the authors using the same experimental settings and datasets described in the methodology section to ensure a fair and consistent comparison. This outstanding performance exceeds that of random forest, CNN and DBN classifiers, proving the power of deep learning architectures in capturing complex patterns and characteristics for higher classification accuracy.

Table 6. Comparison with other well-known classifiers using the same experimental settings.

| Dataset | Alex Net | Random forest | CNN | DBN |
|---|---|---|---|---|
| **VOC 2012** | 95.65 | 85.52 | 91.27 | 89.31 |
| **Caltech-101** | 93.66 | 83.12 | 90.19 | 87.54 |
| **MSCOCO** | 92.71 | 81.23 | 89.76 | 89.15 |

In comparison to other SOTA model as displayed in Table 7 and it can be observed that our model performed better than most of the existing models in almost all three datasets VOC 2012, Caltech101, and MS COCO. Our model scored the best accuracy on VOC 2012 dataset of 95.65 and 93.66 on caltech 101 among all of the compared models and just lacking behind H-CNN in MS Coco dataset which clearly shows the robustness of our model on different datasets. The major aspect of the proposed model is integrating the DBSCAN segmented images with Saliency mapped images which helped in better feature extraction also the combination of feature extraction has contributed big time in achieving the result as shown in ablation study.

Table 7. Comparison of mean accuracies for different models on various datasets with SOTA methods.

| Models | VOC 2012 | Caltech101 | MS COCO |
|---|---|---|---|
| CNN [37] | 85.24 | – | – |
| MCLR [2] | – | 85.75 | – |
| ODSR [33] | 87.67 | 88.60 | – |
| Double Attn [55] | 91.1 | – | 77.5 |
| FCMME [18] | 93.53 | 89.26 | – |
| H-CNN [9] | 95.04 | – | 94.53 |
| **Proposed model** | **95.65** | **93.66** | **92.71** |

## 4.3. Ablation Study

We experimented with various feature extraction techniques, including MSER, BRISK, and Wavelet Transform, to evaluate their individual and combined contributions to object classification tasks across three datasets Caltech 101, VOC 2012, and MS COCO as shown in Figure 10. At first, we evaluated the results of each feature extraction method separately. On the

Caltech 101 dataset, MSER reached 69.75% accuracy, BRISK achieved 72.31%, and WT scored 71.11%. In the VOC 2012 dataset the individual accuracies were: 67.56% MSER, 71.63% BRISK, and 69.79% WT. On the other hand, on MS COCO dataset, the performance of MSER, BRISK, and WT were 70.12%, 72.23%, and 71.11%, respectively. In order to capitalize the complementary features of these feature descriptors, the feature fusion strategies were discussed. Combining MSER and BRISK features, the accuracies were 77.67% (Caltech 101), 75.12% (VOC 2012), and 78.12% (MS COCO). The combination of BRISK and WT characteristics increased the accuracies to 85.67% (Caltech 101), 83.21% (VOC 2012), and 87.54% (MS COCO). On a similar note, combining MSER and WT features achieved the accuracy of 82.12% (Caltech 101), 80.36% (VOC 2012) and 84.14% (MS COCO). Inspired by the remarkable feature fusion results, we have concatenated three feature descriptors (MSER, BRISK, and WT) into one feature set. With this all-encompassing feature fusion strategy, substantial improvements were achieved with accuracies of 93.66%, (Caltech 101), 95.65% (VOC 2012), and 92.71% (MS COCO). The results demonstrate the necessity of feature fusion in representing the diverse and complementary features of objects. While individual feature descriptors provide valuable information, their combination through fusion strategies effectively harnesses their strengths and, consequently, enhances performance in object classification tasks across different datasets.

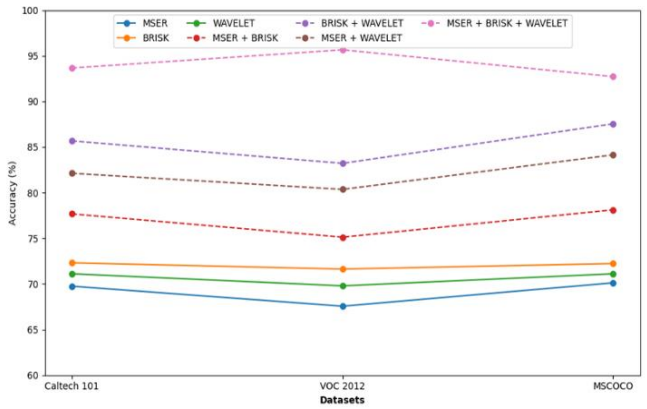



Figure 10. Classification accuracy using different combination of features on all three datasets.

This paper presents a new practical implementation of object detection and classification that is based on AlexNet. Adaptive mean filtering is used for noise reduction in the image pre-processing before segmentation with DBSCAN. Saliency maps are used to address the inconsistency of segmentation. The segment and saliency map images are fused, and features are extracted from the fused images using wavelet transformation, BRISK, and MSER methods. Fusion of features is carried out by concatenation, and feature optimization is done by the Fish Swarm Algorithm (FSA). The proposed system is evaluated on various datasets including PASCAL VOC2012, Caltech 101 and MS COCO, datasets achieving the classification accuracy of 95.65, 93.66 and 92.71, respectively. The further work covers to apply the model on the aerial image dataset. The emphasis will be placed on re configuring the model to deal with the specific difficulties of aerial images, changing object scales, occlusions and viewpoints.

## Acknowledgment

The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the Research Group Funding program grant code (NU/RG/SERC/13/40). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBUFFR-2024-231-07."

## References

[1] Ahmad M., Shabbir S., Roy S., Hong D., Wu X., and Yao J., "Hyperspectral Image Classification-Traditional to Deep Models: A Survey for Future Prospects," *IEEE Journal Selected Topics Applied Earth Observations Remote Sensing*, vol. 15, pp. 968-999, 2022. DOI:10.1109/JSTARS.2021.3133021

[2] Ahmed A., Jalal A., and Kim K., "A Novel Statistical Method for Scene Classification Based On Multi-Object Categorization and Logistic Regression," *Sensors*, vol. 20, no. 14, pp. 3871, 2020. https://doi.org/10.3390/s20143871

[3] Ahmed M., Almujally N., Alazeb A., Algarni A., and Park J., "Enhanced Object Detection and Classification via Multi-Method Fusion," *Computers, Materials and Continua*, vol. 79, no. 2, pp. 3315-3331, 2024. https://doi.org/10.32604/cmc.2024.046501

[4] Ahmed M. and Jalal A., "Dynamic Adoptive Gaussian Mixture Model for Multi-Object Detection Over Natural Scenes," *in Proceedings of the 5th International Conference on Advancements in Computational* Sciences, Lahore, pp. 1-8, 2024. DOI:10.1109/ICACS60934.2024.10473231

[5] Ahmed M. and Jalal A., "Robust Object Recognition with Genetic Algorithm and Composite Saliency Map," *in Proceedings of the 5th International Conference on Advancements in Computational Sciences*, Lahore, pp. 1-7, 2024. DOI:10.1109/ICACS60934.2024.10473285

[6] Alkhatib M., Al-Saad M., Aburaed N., Almansoori S., Zabalza J., Marshall S., and Al-Ahmad H., "Tri-CNN: A Three Branch Model for Hyperspectral

Image Classification," *Remote. Sens*, vol. 15, no. 2, pp. 316, 2023. https://doi.org/10.3390/rs15020316

[7] Bharadiya J., "Convolutional Neural Networks for Image Classification," *International Journal of innovative Science and Research Technology*, vol. 8, no. 5, pp. 673-677, 2023. https://doi.org/10.5281/zenodo.8020781

[8] Bo L. and Sminchisescu C., "Efficient Match Kernel Between Sets of Features for Visual Recognition," *in Proceedings of the 23rd International Conference on Neural Information Processing Systems*, Vancouver, pp. 135-143, 2009.

[9] Borade J. and Lakshmi M., "Multi-Class Object Detection System Using Hybrid Convolutional Neural Network Architecture," *Multimedia Tools and Applications*, vol. 81, pp. 31727-31751, 2022. https://doi.org/10.1007/s11042-022-13007-7

[10] Cheng B., Girshick R., Dollar P., Berg A., and Kirillov A., "Boundary IoU: Improving Object-Centric Image Segmentation Evaluation," *in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Nashville, pp. 15334-15342, 2021. DOI:10.1109/CVPR46437.2021.01508

[11] Dai Y., Gieseke F., Oehmcke S., Wu Y., and Barnard K., "Attentional Feature Fusion," *in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, pp. 3560-3569, 2021. DOI:10.1109/WACV48630.2021.00360

[12] Dimitrovski I., Kitanovski I., Kocev D., and Simidjievski N., "Current Trends in Deep Learning for Earth Observation: An Open-Source Benchmark Arena for Image Classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 18-35, 2023. https://doi.org/10.1016/j.isprsjprs.2023.01.014

[13] Gokulalakshmi A., Karthik S., Karthikeyan N., and Kavitha M., "ICM-BTD: Improved Classification Model for Brain Tumor Diagnosis Using Discrete Wavelet Transform-Based Feature Extraction and SVM Classifier," *Soft Computing*, vol. 24, pp. 18599-18609, 2020. https://doi.org/10.1007/s00500-020-05096-z

[14] Guo Y. and Sengur A., "A Novel Color Image Segmentation Approach Based on Neutrosophic Set and Modified Fuzzy C-Means," *Circuits, Systems, and Signal Processing*, vol. 32, pp. 1699-1723, 2013. https://doi.org/10.1007/s00034-012-9531-x

[15] Hosny K., Kassem M., and Fouad M., "Classification of Skin Lesions into Seven Classes Using Transfer Learning with AlexNet," *Journal of Digital Imaging*, vol. 33, pp. 1325-1334, 2020. https://doi.org/10.1007/s10278-020-00371-9

[16] Ibrahim R., Abualigah L., Ewees A., Al-qaness M., Yousri D., Alshathri S., and Abd Elaziz M., "An Electric Fish-Based Arithmetic Optimization Algorithm for Feature Selection," *Entropy*, vol. 23, no. 9, pp. 1189, 2021. https://doi.org/10.3390/e23091189

[17] Iqball T. and Wani M., "Weighted Ensemble Model for Image Classification," *International Journal of Information Technology*, vol. 15, pp. 557-564, 2023. https://doi.org/10.1007/s41870-022-01149-8

[18] Jalal A., Ahmed A., Rafique A., and Kim K., "Scene Semantic Recognition Based on Modified Fuzzy C-Mean and Maximum Entropy Using Object-To-Object Relations," *IEEE Access*, vol. 9, pp. 27758-27772, 2021. DOI:10.1109/ACCESS.2021.3058986

[19] Kuan K., Manek G., Lin J., Fang Y., and Chandrasekhar V., "Region Average Pooling for Context-Aware Object Detection," *in Proceedings of the IEEE International Conference on Image Processing*, Beijing, pp. 1347-1351, 2017. DOI:10.1109/ICIP.2017.8296501

[20] Li S., Wang L., Li J., and Yao Y., "Image Classification Algorithm Based on Improved AlexNet," *Journal of Physics: Conference Series*, vol. 1813, no. 012051, pp. 1-9, 2021. DOI:10.1088/1742-6596/1813/1/012051

[21] Liu Y., Zhang H., Guo H., and Xiong N., "A Fast-Brisk Feature Detector with Depth Information," *Sensors*, vol. 18, no. 11, pp. 3908, 2018. https://doi.org/10.3390/s18113908

[22] Madhukar B., Bharathi S., and Ashwin M., "Classification of Breast Cancer using Ensemble Filter Feature Selection with Triplet Attention Based Efficient Net Classifier," *The International Arab Journal of Information Technology*, vol. 21, no. 1, pp. 17-31, 2024. 23 https://doi.org/10.34028/iajit/21/1/2

[23] Martins P., Carvalho P., and Gatta C., "On the Completeness of Feature-Driven Maximally Stable Extremal Regions," *Pattern Recognition Letters*, vol. 74, pp. 9-16, 2016. https://doi.org/10.1016/j.patrec.2016.01.003

[24] Mukherjee P., Lall B., and Shah A., "Saliency Map Based Improved Segmentation," *in Proceedings of the IEEE International Conference on Image Processing*, Quebec City, pp. 1290-1294, 2015. DOI:10.1109/ICIP.2015.7351008

[25] Muralidharan R. and Chandrasekar C., "Object Recognition Using SVM-KNN based on Geometric Moment Invariant," *International Journal of Emerging Trends and Technology in Computer Science*, vol. 1, no. 3, pp. 215-220, 2011.

[26] Narayanan L., Krishnan S., and Robinson H., "A Hybrid Deep Learning Based Assist System for Detection and Classification of Breast Cancer from Mammogram Images," *The International Arab Journal of Information Technology*, vol. 19, no. 6, pp. 965-974, 2022. 22

https://doi.org/10.34028/iajit/19/6/15

[27] Naseer A., Alzahrani H., Almujally N., Al-Nowaiser K., Al-Mudawi N., Algarni A., and Park J., "Efficient Multi-Object Recognition Using GMM Segmentation Feature Fusion Approach," *IEEE Access*, vol. 12, pp. 37165-37178, 2024. DOI:10.1109/ACCESS.2024.3372190

[28] Naseer A., Almujally N., Alotaibi S., Alazeb A., and Park J., "Efficient Object Segmentation and Recognition Using Multi-Layer Perceptron Networks," *Computers, Materials and Continua*, vol. 78, no. 1, pp. 1381-1398, 2024. https://doi.org/10.32604/cmc.2023.042963

[29] Othman G. and Zeebaree D., "The Applications of Discrete Wavelet Transform in Image Processing: A Review," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 31-43, 2020.

[30] Ouadiay F., Bouftaih H., Bouyakhf E., and Himmi M., "Simultaneous Object Detection and Localization Using Convolutional Neural Networks," *in Proceedings of the International Conference on Intelligent Systems and Computer Vision*, Fez, pp. 1-8, 2018. DOI:10.1109/ISACV.2018.8354045

[31] Pourpanah F., Wang R., Lim C., Wang X., and Yazdani D., "A Review of Artificial Fish Swarm Algorithms: Recent Advances and Applications," *Artificial Intelligence Review*, vol. 56, pp. 1867-1903, 2023. https://doi.org/10.1007/s10462-022-10214-4

[32] Pramanik A., Pal S., Maiti J., and Mitra P., "Granulated RCNN and Multi-Class Deep Sort for Multi-Object Detection and Tracking," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 1, pp. 171-181, 2022. DOI:10.1109/TETCI.2020.3041019

[33] Rafique A., Gochoo M., Jalal A., and Kim K., "Maximum Entropy Scaled Super Pixels Segmentation for Multi-Object Detection and Scene Recognition via Deep Belief Network," *Multimedia Tools and Applications*, vol. 82, pp. 13401-13430, 2023. https://doi.org/10.1007/s11042-022-13717-y

[34] Riaz F., Rehman S., Ajmal M., Hafiz R., Hassan A., and Aljohani N., "Gaussian Mixture Model Based Probabilistic Modeling of Images for Medical Image Segmentation," *IEEE Access*, vol. 8, pp. 16846-16856, 2020. DOI:10.1109/ACCESS.2020.2967676

[35] Sengupta D., Ali S., Bhattacharya A., Mustafi J., Mukhopadhyay A., and Sengupta K., "A Deep Hybrid Learning Pipeline for Accurate Diagnosis of Ovarian Cancer Based on Nuclear Morphology," PLoS One, vol. 17, no. 1, pp. 1-20, 2022. DOI:10.1371/journal.pone.0261181

[36] Shen J., Hao X., Liang Z., Liu U., Wang W., and Shao L., "Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5933-5942, 2016. DOI:10.1109/TIP.2016.2616302

[37] Shetty S., "Application of Convolutional Neural Network for Image Classification on Pascal Voc Challenge 2012 Dataset," *arXiv ٠reprint*, vol. arXiv:1607.03785, pp. 1-6, 2016. https://doi.org/10.48550/arXiv.1607.03785

[38] Song L., Gao M., Wang S., and Wang S., "An Image Segmentation Method by combining Fuzzy C-Means Clustering and Graph Cuts Optimization for Multiphase Level Set Algorithms," *in Proceedings of the 2nd International Conference on Information Science and Control Engineering*, Shanghai, pp. 611-615, 2015. DOI:10.1109/ICISCE.2015.141

[39] Song S., Jia Z., Yang J., and Kasabov N., "A Fast Image Segmentation Algorithm Based on Saliency Map and Neutrosophic Set Theory," *IEEE Photonics Journal*, vol. 12, no. 5, pp. 1-16, 2020, DOI:10.1109/JPHOT.2020.3026973

[40] Srikar M. and Malathi K., "A Supervised Stable Object Detection with Image Feature Extraction Using Image Segmentation by Comparing Histogram of Oriented Gradients (HOG) Algorithm Over Scale Invariant Feature Transform (SIFT) Algorithm Model," *Journal of Pharmaceutical Negative Results*, vol. 13, no. 4, pp. 1708-1714, 2022. https://doi.org/10.47750/pnr.2022.13.S04.205

[41] Sun K., Sun L., Zhao Y., Chen Y., Hao X., Liu H., Liu X., and Chen J., "XGBG: A Novel Method for Identifying Ovarian Carcinoma Susceptible Genes Based on Deep Learning," *Frontiers in Oncology*, vol. 12, pp. 1-7, 2022. DOI:10.3389/fonc.2022.89750.

[42] Al-Mudawi N., Tayyab M., Ahmed M., and Jalal A., "Machine learning Based on Body Points Estimation for Sports Event Recognition," *in Proceedings of the IEEE International Conference on Autonomous Robot Systems and Competitions*, Paredes de Coura, pp. 120-125, 2024. DOI:10.1109/ICARSC61747.2024.10535954

[43] Wang C., Ji M., Wang J., Wen W., Li T., and Sun Y., "An Improved DBSCAN Method for Lidar Data Segmentation with Automatic Eps Estimation," *Sensors*, vol. 19, no. 1, pp. 172, 2019. https://doi.org/10.3390/s19010172

[44] Wangsaputra D., Anam C., Adi K., and Naufal A., "Impact of Adaptive Mean Filter as the Preprocessing Stage of Histopathological Image Classification of Breast Tumor Using Transfer Learning VGG16 for Various Magnifications," *International Journal of Scientific Research in Science and Technology*, vol. 10, no. 2, pp. 274-280, 2023. https://doi.org//10.32628/IJSRST52310239

[45] Ahmed M., Alshahrani A., Almjally A., Al-

Mudawi N., Algarni A., and Al-Nowaiser K., "Remote Sensing Image Interpretation: Deep Belief Networks for Multi-Object Analysis," *IEEE Access*, vol. 12, pp. 142360-142379, 2024. DOI:10.1109/ACCESS.2024.3466220

[46] Wei H., Yang C., and Yu Q., "Contour Segment Grouping for Object Detection," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 292-309, 2017. https://doi.org/10.1016/j.jvcir.2017.07.003

[47] Wu X., Sahoo D., and Hoi S., "Recent Advances in Deep Learning for Object Detection," *Neurocomputing*, vol. 396, pp. 39-64, 2020. https://doi.org/10.1016/j.neucom.2020.01.085

[48] Xia Y., Tian Z., Yu J., Zhang Y., Liu S., Du S., and Lan X., "A Review of Object Detection Based on Deep Learning," *Multimedia Tools and Applications*, vol. 79, pp. 23729-23791, 2020. https://doi.org/10.1007/s11042-020-08976-6

[49] Yan S. and Dong Y., "GMM Based Simultaneous Reconstruction and Segmentation in X-Ray CT Application," *in Proceedings of the International Conference on Scale Space and Variational Methods in Computer Vision*, Cabourg, pp. 503-515, 2021. https://doi.org/10.1007/978-3-030-75549-2_40

[50] Yang H., Tian J., and Yang J., "New Medical Image Segmentation Algorithm Based on Gaussian-Mixture Model," *Biomedical Photonics and Optoelectronic Imaging*, vol. 4224, pp. 40-44 2000. https://doi.org/10.1117/12.403921

[51] Yang R. and Li D., "Adaptive Wavelet Transform Based on Artificial Fish Swarm Optimization and Fuzzy C-Means Method for Noisy Image Segmentation," *Computer Science and Information Systems*, vol. 19, no. 3, pp. 1389-1408, 2022. https://doi.org/10.2298/CSIS220321039Y

[52] Yao R., Lin, G., Xia S., Zhao J., and Zhou Y., "Video Object Segmentation and Tracking: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 4, pp. 1-47, 2020. https://doi.org/10.1145/3391743

[53] Yu L., Chen X., and Zhou S., "Research of Image Main Objects Detection Algorithm Based on Deep Learning," *in Proceedings of the IEEE 3rd International Conference on Image, Vision and Computing*, Chongqing, pp. 70-75, 2018. DOI:10.1109/ICIVC.2018.8492803

[54] Yugander P., Tejaswini C., Meenakshi J., Kumar K., Varma B., and Jagannath M., "MR Image Enhancement Using Adaptive Weighted Mean Filtering and Homomorphic Filtering," *Procedia Computer Science*, vol. 167, pp. 677-685, 2020. https://doi.org/10.1016/j.procs.2020.03.334

[55] Zhao H., Zhou W., Hou X., and Zhu H., "Double Attention for Multi-Label Image Classification," *IEEE Access*, vol. 8, pp. 225539-225550, 2020. DOI:10.1109/ACCESS.2020.3044446

**Muhammad Waqas Ahmed** received his MS degree in Computer Sciences from COMSATS. He is currently pursuing his Ph.D. in computer science from Air University, Islamabad, Pakistan. His research interests include Artificial Intelligence, Computer Vision, Machine Learning Algorithms, Deep Learning, Image, Video Processing, and Intelligent Systems.

**Abdulwahab Alazeb** is currently Assistant Professor at Department of Computer Science and Information system, Najran University. He received the B.S. degree in computer science from King Khalid University, Abha, Saudi Arabia, in 2007, and the M.S. degree in computer science from the Department of Computer Science, University of Colorado Denver, USA in 2014. He holds a PhD degree as well as a Graduate Certificate in cybersecurity from the University of Arkansas, USA, in 2021. His research interests include Cybersecurity, Cloud and Edge Computing Security, Machine Learning and the Internet of Things.

**Naif Al Mudawi** is assistant Professor, Department of Computer Science and Information system, Najran University. He holds a PhD from the Collage of Engineering and Informatics at University of Sussex in Brighton, UK in 2018. He graduated from the Australian La Trobe University with a master's degree in computer science in 2011 during his academic journey to obtain a master's degree, he was a member of the Australian Computer Science committee. Dr. Naif has has many published research and scientific papers in many prestigious journals in various disciplines of computer science.

**Touseef Sadiq** is a PhD researcher at University of Agder, Norway. His current research focuses on deep multimodal learning for descriptive object identification in urban environments. He obtained his B.E degree in computer engineering from Bahria University Islamabad, Pakistan and completed his MS degree in communications and computer networks engineering from Polytechnic university of Turin, Italy on fully funded scholarship. His primary research interests include machine learning, computer vision, deep multimodal learning, and their applications.

**Bayan Alabduallah** received the Ph.D. degree in informatics from the University of Sussex, Brighton, U.K., in May 2022. She is currently an Assistant Professor with the Department of Information System, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University. She teaches several courses with the Information System Department, such as data governance, system security, and database system. Her research interests include machine learning, data science, privacy, and security.

**Hameed ur Rahman** chair of the Department of Computer Games Development at Air University, Islamabad, boasts a robust research profile. With a Ph.D. in Computer Vision and expertise in augmented reality, virtual reality, image processing, and more, he demonstrates a commitment to cutting-edge technology. As a pivotal member since 2018, Dr. Rahman contributes to AI/Data Science, Cybersecurity, Computer Science, and Gaming Departments, Mentoring Students and Fostering Interdisciplinary Research. Notably, his leadership in the Ignite (Pakistan) Project showcases practical applications of his research, emphasizing his dedication to knowledge dissemination and skill development in emerging fields.

**Asaad Algarni** is working as Assistant Professor at the Department of Computer Sciences in the College of Computing and Information Technology, Northern Borders University, Kingdom of Saudi Arabia. He holds a PhD in Software Engineering from North Dakota State University, USA. His research interests revolve around Software Engineering, Computer Vision applications and Machine Learning.

**Ahmad Jalal** is currently an Associate Professor from Department of Computer Science and Engineering, Air University, Pakistan. He received his Ph.D. degree in the Department of Biomedical Engineering at Kyung Hee University, Republic of Korea. Now, he was working as Postdoctoral Research fellowship at POSTECH. His research interest includes Multimedia Contents, Artificial Intelligence and Machine Learning.