International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 1, February 2026, pp. 216~229
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i1.pp216-229

AI SWLM: artificial intelligence-based system for wildlife monitoring

Arun Govindan Krishnan[1], Jayaraman Bhuvana[2], Mirnalinee Thanga Nadar Thanga Thai[3], Bharathkumar Azhagiya Manavala Ramanujam[4]
[1]Department of Computer Science, SIVET College, Chennai, India
[2]Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Article Info

Article history:

Received May 14, 2025
Revised Sep 22, 2025
Accepted Nov 23, 2025

Keywords:

Animal intrusion
Camera trap images
CSPDenseNet
Deep learning
PANet

ABSTRACT

Detection and recognition of wild animals are essential for animal surveillance, behavior monitoring and species counting. Intrusion of animals and the disaster to be caused can be averted by the timely recognition of intruding animals. An artificial intelligence-based system for wildlife monitoring (AI SWLM) is designed and implemented on the camera trap images. The challenges such as detecting and recognizing animals of different sizes, shape, angles and scale, recognizing the animals of same and different species, detecting them under various illumination conditions, with pose variants and occlusion are addressed by identifying the optimal weights of the deep learning architecture, AI SWLM. Models were trained using Gold Standard Snapshot Serengeti dataset with random weights and the best weights of model were used as initial weights for training the augmented data. This has doubled the performance in terms of mean average precision, which can be interpreted.

Corresponding Author:

Jayaraman Bhuvana
Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai, India
Email: bhuvanaj@ssn.edu.in

1. INTRODUCTION

Recognizing animals irrespective of wild or domestic is essential in a variety of applications namely species counting, surveillance of tress-passing of the animals, and monitoring their behaviors for managing them effectively. By detecting the presence of animals, the disaster caused by their intrusion could be reduced. Also, the welfare of the animals is most essential in balancing the ecosystem. Counting the animals, with their species manually in applications like census will be time consuming and expensive operation. Involving humans to monitor the intrusion of animals will be tedious and risky.

Humans perceive, visualize what they see and act upon accordingly. Human visual recognition system possesses object constancy, ability to recognize object across different viewpoint conditions such as orientation, lighting, and object size variability. We can interpret the entities in each scene, irrespective of their size, scale, angles, rotated or translated. Semantic meaning of images and videos are useful information for any scene interpretation with several applications involving self-driving cars, navigation in mobile robotics, street traffic observations, soccer game analysis, smart room cameras, monitoring of elderly. The detection of animal intrusion can be modelled as object recognition problem.

Animal classification and recognition play a major role in surveillance, automatic car driving to prevent accidents, animal population survey for endangered species, animal surveillance. To balance the wildlife ecology, monitoring and surveillance will be the inherent part of the system. There are few

successful works to maintain the biodiversity of the birds, where for wild animal monitoring such sophisticated system techniques that have been deployed have few inherent drawbacks such as lack in robustness, coverage area, reliability of the equipment and the delay in informing the decisions to the authorities. Early decision-making system should be in place wherever we have human wild conflict. The solution to handle this issue is to imitate the cognitive functionality of brain in recognizing objects. This motivated us to investigate different theories to design novel computational frameworks to solve significant visual perception tasks. There is a growing need for AI-based systems that can automatically detect and classify wildlife species in real-world environments for proactive conservation and to address human-animal conflict.

We aim to design an automatic computational framework to provide efficient solutions for animal recognition, performed effortlessly by a human being. Several challenges in animal detection and recognition are, Animals of different sizes (small and large), Occlusion, multiple species in same frame, animal looking similar to background, partially visible animals without occlusion, counting the number of animals in given frame, various illumination conditions, with pose variant, detecting multiple instances of same species animals in a single frame, locating the detected animal in a cluttered background. This paper presents an end-to-end deep learning based artificial intelligence-based system for wildlife monitoring (AI SWLM) for animal detection from camera trap images. The novelty of the work is in the use of transfer learning on you only look once version5 (YOLOv5) variants with a class-balanced augmentation strategy that significantly improves performance in terms of mean average precision (mAP), precision, and recall when compared with the existing approaches on the Serengeti dataset.

An extensive set of experiments were conducted to identify best suitable model for animal detection; Class imbalance issue is handled by applying augmentation and the best weights are used to initialize the training of the enhanced dataset; detailed qualitative and quantitative analysis were done on the performance of the proposed system.

The article is organized as follows: section 2 discusses the existing systems for object detection and classification for animals. Section 3 proposes AI-SWLM architecture and the design of its functional components. Section 4 discusses the implementation related concepts of AI-SWLM. Section 5 provides the plan of different experiments; section 6 presents the detailed qualitative and quantitative analysis of the results and comparison followed by the conclusion in section 7.

2. SURVEY OF EXISTING WORKS

Humans perceive visual information through the retina, which is transmitted via the optical nerve to the brain, where it is interpreted into objects and scenes. Researchers have found that neuronal firing patterns in the inferior temporal cortex strongly correlate with successful object recognition tasks. The human visual recognition system includes neuronal representations capable of pattern discrimination. Artificial intelligence (AI), a domain of computer science, has developed mechanisms to incorporate such intelligence through algorithms that automate human-like perception and object recognition.

Incorporating the neuron representation patterns of the human brain into computational algorithms can lead to efficient object recognition. Object detection remains one of the most challenging tasks in computer vision, requiring identification of object instances varying in color, shape, location, pose, illumination, and background. It serves as the foundation for applications such as segmentation, captioning, object tracking, and scene understanding. Real-world applications include autonomous vehicles and surveillance systems [1].

Earlier, machine learning algorithms were widely used for object detection. The work in [2] focuses on efficient multiscale features for image retrieval. However, shape features often struggle under varying shadows and illumination. Extracting edges in wildlife imagery remains difficult. Multi-resolution features [3] are well suited for detecting objects of varying shapes. Domain generalization challenges are addressed in object detection, especially in wildlife datasets where environmental variation affects the performance [4].

The evolution of deep learning algorithms and supporting high-end systems has significantly advanced computer vision. Various deep learning techniques [5]–[11] now allow automatic extraction of features from images and videos. Prior work on camera trap images can be broadly classified into two categories: application of pre-trained models and use of object detection and recognition models.

A notable example is multi-task generative adversarial network (MTGAN) [12], an end-to-end framework developed to detect small-scale objects, in which a generator upscales image resolution and a discriminator simultaneously evaluate authenticity and the presence of the object. This is evaluated on common objects in context (COCO) and WIDER FACE datasets, their model used ResNet50 as its backbone and incorporated a regression module to refine details. This multi-task structure helps maintain object-level clarity in low-resolution regions, making it well-suitable for wildlife monitoring applications.

Mask Region-based convolutional neural network (R-CNN) [13], derived from faster R-CNN, has been used for cattle detection and counting, successfully handling occlusion and overlap by leveraging binary mask classification. Simpler CNN-based models have been used to classify images into mammals and reptiles [14], or more granularly into Snakes, Lizards, and Toads/Frogs [15]. Camera traps are widely used to capture wildlife images for population surveys. However, these traps also record humans and false triggers due to wind or vegetation [16]. To classify such images into wildlife, human, or empty, a deep learning approach used AlexNet-96 to segment foreground objects and address class imbalance by color augmentation, achieving 73.13% recall.

Two-level classification on the Snapshot Serengeti dataset was performed in [17]. The first stage was a binary classifier for animal presence, followed by multi-class classification into 26 species using pre-trained models such as AlexNet, visual geometry group (VGG), GoogLeNet, and various ResNet versions, achieving 93.6% with ensemble methods. This work used the same dataset as ours but focused on classification, not object detection. Other efforts used pre-trained models [18] like DenseNet201, Inception-ResNet-V3, and NASNetMobile to classify 35 animal species in the Parks Canada dataset. Augmentation techniques helped mitigate class imbalance, improving performance to 71.2% after ensemble. Similarly, ResNet-18 was employed in [19] to classify animals across 58 classes from camera trap images taken in ten U.S. states. Pre-trained models like InceptionV3, MobileNet, and VGG-16 were used for classifying six animal categories [20]. A robust, location-invariant classifier trained on datasets like FlickR and iNaturalist was proposed in [21]. Using Keras-RetinaNet, their models achieved a mAP of 82.33%–88.59% when tested on Snapshot Serengeti. Facial detection using Faster-RCNN was explored in [22] using the animal face database (AFD), achieving 87.03% accuracy. YOLOv2 was used in [23] for species recognition.

Recent surveys and model innovations emphasize the growing role of deep learning in ecological monitoring. For instance, Zhao *et al.* [24] provides a detailed review of CNN-based wildlife classification from camera trap images, highlighting challenges such as class imbalance and feature extraction in uncontrolled environments. Bhattacharjee *et al.* [25] proposes YOLO-based architectures customized for animal detection under varying environmental conditions, showing improved detection precision and robustness across real-world datasets.

3. METHOD

The proposed artificial intelligence-based system for wildlife monitoring (AI SWLM) will recognize the category of the species in the given camera trap image. The object of interest is the animal, which is detected by the popular and efficient object detection algorithm YOLOv5. The images of different animals captured in trap cameras under different lighting conditions are fed to train the proposed object detection model for localization of animal species, recognition of species and counting of species. This will enable us to monitor animal movements, locations and further notify the respective forest departments regarding their movement near agricultural fields and residential areas. Statistics of animals can be used by the forest department to maintain the ecosystem.

The proposed AI SWLM accepts the inputs in the form of images captured and applies the YOLOV5 architecture that has a backbone system, neck and detection head to localize and classify the animal. The input images in batches will be processed through the backbone, neck and the head outputs the localized as the wild animals along with their names and count. The proposed AI SWLM system combines standard deep learning components such as the YOLOv5 detection architecture with novel enhancements including a two-stage training procedure using pretrained weights, class-balanced data augmentation, and evaluation across different model configurations. The novelty is in the structured augmentation pipeline and reusing best-trained weights to improve generalization of the wildlife detection model challenges, including poor illumination, cluttered backgrounds, and different species. The functional components of AI SWLM and the identifying the best suitable model for detection is shown in Algorithm 1 and Figure 1. The working of functional components is elaborated in the following 3 subsections.

Algorithm 1. Artificial intelligence-based system for wildlife monitoring (AI SWLM)

```
Input:
- Training: Wildlife images, labels, bounding box coordinates
- Testing: Images
Output:
- Recognized objects, labels, bounding box coordinates, counted species
Step 1: Let X ← Original imbalanced training dataset with labels and bounding boxes
Function Main ()
1. WLM_O ← WLM(X)
2. WLM_A_RW ← WLM (X_Enhanced with random weights)
3. WLM_A_BW ← WLM (X_Enhanced with best weights of WLM_O)
```

```
4. WLM_O_Count ← Counting_Species (WLM_O)
5. WLM_A_RW_Count ← Counting_Species (WLM_A_RW)
6. WLM_A_BW_Count ← Counting_Species (WLM_A_BW)
7. AI_SWLM ← Performance_comparison (WLM_O, WLM_A_RW, WLM_A_BW)
Return: AI_WLM model
Function Augment(X)
1. X_Enhanced ← manual augment
2. X_imglevel ← Image_level_augment(X_Enhanced)
3. X_pixellevel ← Pixel_level_augment(X_Enhanced)
4. X_augmented ← X_imglevel + X_pixellevel
Return: X_augmented
Function WLM(X)
1. X.remove_duplicates ()
2. X.remove_corrupted ()
3. X_preprocess ← X.reshape (640, 640)
4. X_augmented ← Augment(X_preprocess)
5. X_featuremap ← CSP_Network (X_augmented)
6. X_featuremap ← Spatial_Pyramid_Pooling (X_featuremap)
7. X_feature_Pyramid ← PANet(X_featuremap)
8. X_PANet ← X_feature_Pyramid
9. (Class_prob, Obj_scores, b_boxes) ← Detection_Head(X_PANet)
Note:
- Class_prob: class probabilities
- Obj_scores: objectness scores
- b_boxes: bounding boxes
Return: WLM model
```

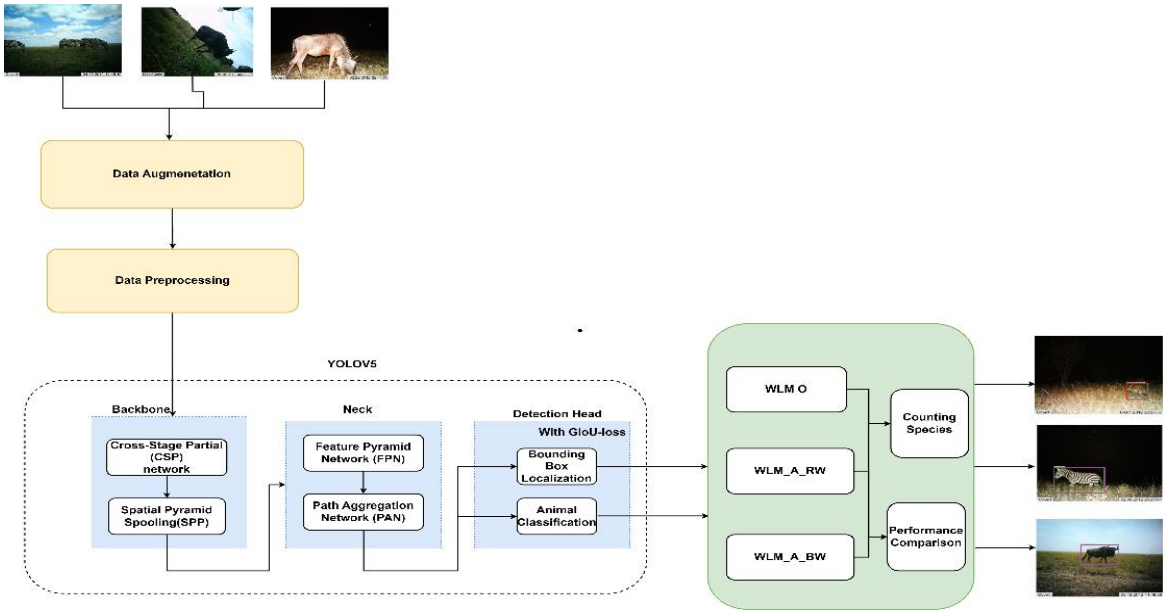


Figure 1. Overview of proposed AI SWLM

3.1. Feature extraction network

The feature extraction part of the architecture will serve as the backbone and help to extract features from the input images. The Backbone of the AI SWLM has the cross stage partial network (CSPNet) and spatial pyramid spooling as the major functional units that extract the features from the input images. The rich features of the wildlife species in each frame will be extracted using a lightweight network called CSPNet, where the feature map is divided into halves and are combined after passing them through different layers. Similarly, the gradient information is also made to flow through different paths and are concatenated and transitioned while passing during the back propagation. The basic building block of this backbone structure is dense block, which will have several dense layers in it. In a dense block the input of one dense layer will be the concatenation of previous dense layer's output and its input. This arrangement will help in accumulating knowledge from all of the previous layers. Multiple dense blocks will be separated by transitional layers. The transitional layer has set of convolutional layers and an average pooling layer of 1×1 and 2×2 respectively.

CSPNet is made up of partial DenseNet block and Partial transitional layers. Partial DenseNet block will divide the feature map into two say f_i and f_j from the base layer, where one half f_i will pass through the dense block and the other half will be concatenated with the input of transitional layer. In the partial transitional layers, the first layer accepts the output of its previous partial dense block as input. The output of the transitional layer is now concatenated with the other half of the feature map f_j and served to the next transitional layer. The CSPNet with its hierarchical feature fusion approach will strengthen the learning ability by giving the innermost layers with the features extracted from the early dense layers. Due to its partial connections, CSPNet extracts very diversified set of features that will help to discriminate against the wildlife of same and different species.

The spatial pyramid pooling (SPP) in the backbone stage of YOLOV5, is a variant of Bag of Words (BoW) model removes the limitation of Convolutional layers working with fixed sized inputs. This characteristic of the SPP makes the model scale invariant and avoids overfitting. The output from CSPNet is passed to SPP before the features are sent to the Neck phase of the network. SPP makes multiple copies of the features and applies maxpooling of different sized kernels and concatenated them and can generate output of fixed length irrespective of the input size using the multi-level spatial bins.

3.2. Feature pyramid path aggregation network (PANet)

The next step in object detection of animals is the construction of feature pyramids by path aggregation network (PANet) in the neck stage of YOLOV5. PANet performs the instance segmentation that serves as the neck part of the single stage object detection model. The purpose of the feature pyramid is to generalize the model on object scaling and to segment animal instances in the camera trap images by maintaining their spatial information. The model needs to detect the same wildlife species in different sizes and scales. This feature pyramid is designed to extract multi-scale feature maps and performs well on unseen or hidden data. The reason why PANet's chosen is because it helps in proper localization of pixels for mask formation. PANet helps in bottom-up path augmentation, adaptive feature pooling, fully connected fusion.

Features will flow via both bottom-up and top-down pathways that work around the spatial resolution before sending them for prediction stage of the network. The Bottom-up network uses ResNet architecture, through which the features flow, that helps in semantic detection and reduces the spatial dimension into half. The top-down flow, up samples and augments the previous layer's output and propagates the features that are semantically significant.

3.3. Object localization and prediction using detection head

The third stage of Proposed AI SWLM is the head of YOLOV5, which predicts the bounding box coordinates, objectless score along with the label of the predicted animal. It applies anchor boxes on features maps from PANet and generates final output vectors with class probabilities, objectless scores, and bounding boxes. From the detected animals, the count of the species belonging to the same or different wildlife species in the scene is processed which can be communicated to the authorities concerned. The detection head will have 3 layers that accept the feature maps of sizes namely, 80×80, 40×40 and 20×20 respectively to detect the animals of different sizes. These detection layers generate an output vector with predicted bounding box coordinates, class probability and category of the animal predicted.

4. IMPLEMENTATION

4.1. Dataset description

Snapshot Serengeti is one of the world's largest camera trap projects with 7.1 million images across 12 seasons. In those 7.1 million images, over 76% of images were empty. Serengeti National Park in Tanzania is best known for the massive animal migrations of Wildebeest, Zebra that drive the cycle of its dynamic ecosystem. The most common wildlife species in the dataset are Wildebeest, Zebra and Gazelle Thompsons. Totally 225 cameras were deployed to capture the wildlife images in Serengeti National Park, Tanzania, East Africa. Citizen volunteers have been involved in this project to annotate the images that have 48 classes of wildlife species in it. The labelled dataset named Gold Standard Snapshot Serengeti with 46 classes is used in this work for species detection and recognition. Bounding box coordinates around each animal in a camera trap image are provided with the data set.

4.2. Data pre-processing and augmentation

The Gold Standard Snapshot Serengeti has the width, height, x_{min}, y_{min}, x_{max} and y_{max} of the bounding boxes for each of the instance of animals present in an image. As the first step of pre-processing, these measures of the bounding box coordinates are converted into x_{center}, y_{center}, width and height. 70% of

images in the dataset are used for training the species detection model, where 20% of images are used for validation and 10% of images are used for testing.

Dataset images were in different sizes, so they were converted into a standard image size of 640, 640. Data cleaning operations were performed to remove the duplicate, corrupt and unused images. Along with that, the images for which the label and annotations are not given were removed from the dataset. Few images with mismatched filenames have also been removed from the dataset as a part of data cleansing. Additional images with labels namely scout guard, jeep, trucks, rocks, sky and images without animals are removed from the dataset.

Augmentation is the process of adding new learning samples either from the existing data or by generating new synthetic data that increases the size of the dataset to enhance the learning of the model and hence enhance the performance with better accuracy. As one of the regularization techniques, data augmentation avoids model overfitting by introducing a diverse set of learning instances.

Several data augmentations techniques have been developed to achieve better performance in computer vision problems, that can be categorized into image level and at pixel level augmentation, to improve the learning ability and to increase the variance in the training data. At the pixel level the proposed AI SWLM system performs basic linear and affine transformations namely the flipping, rotation, clipping, adaptive scaling and modifying the brightness and contrast data augmentations. At the image level, the proposed system employs copy-paste, letterbox to maintain aspect ratio and mix-up, mosaic methods for data augmentation. Cut mix and mosaic methods combine 2 and 4 images respectively together to generate a new learning sample.

In the Gold Standard Snapshot Serengeti dataset, it has been observed that the number of images in a few of the classes is not sufficient to train a object detection and recognition model, since the dataset is imbalanced. The greater number of samples are found in the class Wild beast, and very a smaller number of samples are found in classes namely Hare, Water buck, Vervet Monkey, and Leopard Rhinoceros. To enhance the variance of the training data, image augmentation is the best approach before applying any deep learning framework. Apart from the above techniques, the proposed AI SWLM also augments the dataset with new images taken from the web to form enhanced dataset. This is done to overcome the data imbalance problem across all the classes. These manually augmented images of the enhanced dataset are annotated and the bounding box coordinates are then used by the proposed system during training. The enhanced dataset was split into 3 parts in the same proportion as that of the original dataset and used for training, validation and testing the animal object detection and recognition model, AI SWLM.

AI SWLM model is developed and trained in a NVIDIA GeForce RTX 2080 GPU 11GB system with CUDA version 9.1 using PyTorch version 1.7.1 and Python version 3.6.10. The models were trained using stochastic gradient descent (SGD) with an initial learning rate of 0.01 and momentum of 0.937. Loss was computed using GIoU for bounding box regression and binary cross-entropy for classification and objectness. We have chosen YOLOv5 for its speed and accuracy in real-time object detection tasks and hence suitable for deployment in wildlife surveillance. CSPNet enhances the learning capability by enabling feature reuse and gradient flow. PANet helps to retain spatial features and improve localization in cluttered wildlife scenes. To address severe class imbalance in the Serengeti dataset, a two-tier augmentation strategy was used.

5. EXPERIMENTS

This section discusses the various experiments conducted to identify the most suitable AI SWLM model for detecting the animals.

a. To detect and recognize the animals in the Gold Standard Snapshot Serengeti dataset with the original set of images, that are imbalanced across the classes.
   - Experimentation with a smaller model, YOLO V5m referred to as wildlife monitoring with original dataset (WLM-O1).
   - Experimentation with larger models with original dataset, YOLO V5l referred as WLM-O2.

b. Detecting and recognizing animals with augmented dataset and different weight initialization methods.
   - Experimentation with smaller model, YOLO V5m with randomly initialized weights referred as wildlife monitoring with augmented dataset and random weights (WLM-A-RW1).
   - Experimentation with larger model on augmented dataset, YOLO V5l with randomly initialized weights referred (WLM-A-RW2).
   - Experimentation with smaller model on augmented dataset, YOLO V5m with using best weight from experiment WLM-O1 referred as wildlife monitoring with augmented dataset and best trained weights (WLM-A-BW1).
   - Experimentation with larger model on augmented dataset, YOLO V5l with best weight from experiment WLM-O2 referred as WLM-A-BW2.

5.1. Building WLM-O1 model with original dataset

At first the original dataset was used to train the YOLO V5m model which has 369 layers. This model used the pre trained weights and hyper parameters values of YOLO V5 model. Then the YOLO V5m model was made to run for 500 epochs with 16 as batch size. The model training stopped at 311 epochs since it has shown no improvement in learning after that epoch. Their best, last weights are saved for detection and future use. The trained WLM-O1 model is used for testing and found that the model detected for most test images and failed to detect wildlife species in the few of the test images. The outcome of the WLM-O1 model during testing will have the bounding box around the detected animals and counts the number of species present in the test image. The layers of YOLO V5m were not sufficient to detect the animal instances with poor light conditions and due to class imbalance in the original dataset.

5.2. Building WLM-O2 model with original dataset

The original dataset was then used to train the YOLO V5l model which has more layers when compared to the YOLO V5m model. YOLO V5l has 468 layers. This model also used the pre trained weights and hyper parameters values provided with the YOLO V5 model. Then the YOLO V5l model was made to run for 500 epochs with 16 batch size. The model training stopped at 292 epochs and no improvement was observed in learning after that epoch. Their best, last weights are noted for detection and future use. When WLM-O2 model is used for testing, analysis on detections, found that there were wrong detections of wildlife species and some remained undetected because the training was not sufficient because of the unavailability of enough data across classes. The WLM-O2 model was able to detect species under poor lighting conditions but still class imbalance has played in dragging the performance down.

5.3. Building WLM-A-RW1 model from scratch with augmented dataset

The augmented dataset was used to train the YOLO V5m model labeled as WLM-A-RW1 from scratch without using any special weight initialization. This experiment also used hyper parameters provided with YOLO V5. Then the model was made to run for 500 epochs with 16 as batch size. Since it is training from scratch the early stopping was not used and the model was run for complete 500 epochs. The training results of WLM-A-RW1 have shown good performance in terms of learning and variance between the animals. The test results of the experiment show that it has detected the species and labeled them correctly without any problem. Though the training and detection were good, the drawbacks were that it was not detecting a few multiple species in the same image, it just detected one or two species and ignored the remaining. And observed that YOLO V5m layers were not enough to detect species in few images with poor quality and lighting conditions.

5.4. Building WLM-A-RW2 model from scratch with augmented dataset

The same procedure as in experiment in WLM-A-RW1 were used in YOLO V5l from labelled as WLM-A-RW2. The only positive in the WLM-A-RW2 model is that it detected species even in images with poor quality and lighting conditions. Like WLM-A-RW1, this model lacks performance by not detecting a few multiple species in the same image. The additional observation made is that the WLM-A-RW2 model detects a few species wrongly. Other than the few drawbacks the WLM-A-RW2 showed good performance when compared to the previously built ones.

5.5. Building WLM-A-BW1 model using best weight from WLM-O1 that used original dataset

The augmented dataset was once again used to train the YOLO V5m model referred as WLM-A-BW1. This time the model was given the best weights of WLM-O1 which was trained on the original dataset with same hyper parameters. Since it uses weights from the previous model, we used early stopping to stop the model when there is no improvement in learning. The model stopped training at 388 epochs. The training results were encouraging in terms of learning. The testing result shows the best performance, when compared with the previous augmented models. Multiple species detection was also found to be improved but still performed poorly on images with poor quality, lighting conditions and anomalies.

5.6. Building WLM-A-BW2 model using best weight from WLM-O2 that used original dataset

The final experiment was building WLM-A-BW2 model using augmented dataset using the best weights from WLM-O2 that was trained on the original dataset. The model was then made to run for 500 epochs with 16 as batch size. The model stopped its training at 416 epochs with no improvements in learning after that. The test results showed that WLM-A-BW2 has given better results for images with poor quality, lighting conditions and anomalies. Multiple species detections were also improved, and the misclassification was drastically reduced in WLM-A-BW2 model when compared to all the previous models.

6. RESULTS

6.1. Quantitative analysis

Wildlife species detection and identification model detects the animals and recognizes them by bounding boxes and generates objectiveness score along with class names. Quantitative performance analysis is performed to evaluate the measurable factors of the results generated by the detection and identification model on the test set. Precision, Recall, mAP quantitative measures are used for the evaluation.

The three losses calculated are bounding box loss, objectness loss and classification loss for both training and validation. Bounding box loss is the loss computed for the localization phase of animal detection where it calculates the mean squared error between the ground truth and the predicted box. The probability of the bounding box having an animal is calculated by the objectness score. Binary Cross-Entropy was used to compute the classification loss during animal species prediction. The training loss among these models shows a decline from 0.03 to less than 0.025, similarly the objectiveness loss has also got reduced from 0.015 to less than 0.01. The classification loss has also decreased to 0.005 for WLM-A-BW2. From the results, we observed that the precision values were deteriorating with the sample of the original dataset. When the augmented dataset is used the precision values are consistent and increasing for most of the time and reaching above 0.8 for WLM-A-BW2. Similar behavior was noticed for recall values among the models. From the analysis it is understood that increased precision and recall values lead to better object detection results of WLM-A-BW2.

Considering the mAP values obtained for the same three YOLO V5l models for two different threshold values,0.5 and 0.5:0.95, mAP values obtained from the original dataset were not continuously increasing. From this it can be understood that the detection obtained for the experiment with the original dataset was not better. But when the mAP values of the other two models namely WLM-A-RW1 and WLM-A-BW2 are observed, they are continuously increasing and become constant after some time. Though both model's mAP values increase and become constant, the mAP values of the experiment with augmented dataset using best weights, WLM-A-BW2 were slightly better when compared to the other. With this it is found that experiments with augmented datasets using already trained weights give better detection when compared to all other experimental models. As can be seen from the values reported in Table 1, mAP values are very low for WLM-O1 and WLM-O2. This was also observed from the detection of these models, where many animals were left unidentified and many were falsely detected, and these models could not detect many challenging images as well.

Table 1. Performance metrics of all AI SWLM models

Model	Precision	Recall	mAP [0.5:0.95]
WLM-O1	66.41	50.54	32.85
WLM-O2	76.17	61.91	35.94
WLM-A-RW1	80.51	74.50	62.36
WLM-A-RW2	80.44	76.55	62.69
WLM-A-BW1	77.43	77.47	63.97
WLM-A-BW2	81.28	77.88	64.27
WLM-O1	66.41	50.54	32.85

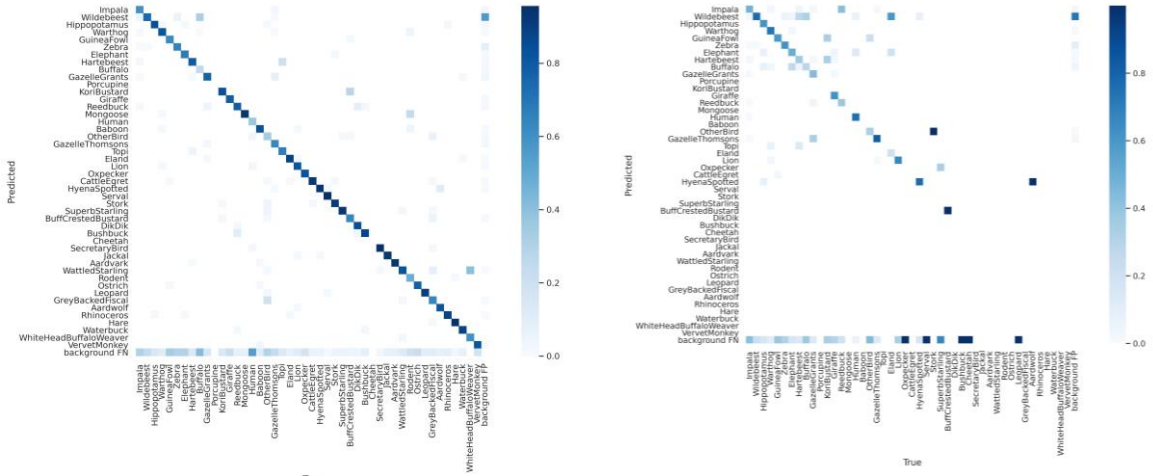
From the detections made by WLM-O1 and WLM-O2 models, it was observed that the detection has several false positives where buffaloes were detected as wild beasts with poor objectiveness score and many animals were not detected due to low mAP values. From the Table 1, the mAP values for the augmented models such as WLM-A-RW1 and WLM-A-RW2 are twice higher than non-augmented WLM-O1 and WLM-O2 models where their False Positives were comparatively reduced with WLM-O1 and WLM-O2. And we found some animals are not detected in images with multiple species. Though the mAP values are relatively high but not sufficient to improve the detection for multiple species in a single camera trap image. Figure 2(a) to (c) shows the performance of WLM-A-BW2 under varied background conditions: Figure 2(a) clear-sky illumination, Figure 2(b) dense forest, and Figure 2(c) shadow dominated scenes.

From the detections of WLM-A-RW1 and WLM-A-RW2 models, it was clearly seen that the detection of false positives was reduced with increase in mAP values but still several species are not detected when there are multiple species in a single image. For the models, WLM-A-BW1 and WLM-A-BW2 that used augmented dataset and the fetched best weights from WLM-O1 and WLM-O2 models, it is observed that the mAP values are relatively higher than the other models. The detection from these models were better than the previous models. And we found that these models overcome challenges such as misclassification, less objectness score and multiple species detection that occurred in other models.

From these images one can clearly see that the detection of multiple species in a single image was improved, which in turn contributed to the increase in mAP values. The confusion matrices in Figure 3 show that WLM-A-BW2 as shown in Figure 3(a) performs significantly better species recognition than WLM-O1 as shown in Figure 3(b). The augmentation and the use of the best weights for training WLM-A-BW2 lead to better performance as shown in the diagonal of the confusion matrix. Too many species are left undetected by the WLM-O1 model due to the unavailability of the sufficient learning samples across the classes.

(a) (b) (c)

Figure 2. WLM-A-BW2 in different backgrounds (a) a clear sky (b) forest, and (c) shadow



(a) (b)

Figure 3. Confusion matrix of (a) WLM-O1 model and (b) WLM-A-BW2 model

6.2. Qualitative analysis

The detection results obtained from all the 6 models were analyzed in this section with respect to challenging situations namely different illumination conditions, background, clutter, same species single instance, different species different instances.

6.2.1. Different illumination conditions and similar background

Wildlife species were shot under different lighting conditions (*i.e.*, different illuminations) in that many species were pictured under poor illumination conditions. To detect images under the poor illumination conditions was one of the major challenges faced by the object detection models. With the presence of CSPDarknet with 468 layers in its backbone made the SWLM models possible to overcome the poor illumination challenge.

The animals detected in Figure 4 are the results of the best performing WLM-A-BW2 model that has been trained on the augmented data and used the best weights from WLM-OM1 model. It can be observed

that the test images are taken at night with different illuminations and images that are difficult to differentiate from the background. WLM-A-BW2 can detect animals with challenging backgrounds such as cloudy, sunset and sunrise. These detections are attributed to the working of CSPDarknet that clearly differentiates background and foreground information during detection. This capability of the SWLM model will allow the AI SWLM to detect the animal even outside the forest or countryside irrespective of the background.

Figure 4. WLM-A-BW2 performance in poor illumination conditions

6.2.2. Clutter

Cluttered images have the focus on the different objects than on the desired objects. So, the wild species captured in clustered images are either blurred or not seen brightly. Detecting species in the cluttered images is the next challenge of AI-SWLM. The presence of PANet as its neck in the architecture of AI SWLM plays a major role in detecting species in cluttered images and makes it possible to recognize each of the animals in the clutter. The bi-directional feature fusion technique helps the network train on different input features. Detection of multiple wild animal species in cluttered images by WLM-A-BW2 model is shown in Figure 5.

Figure 5. Detection of cluttered images by WLM-A-BW2

6.2.3. Single species single instance, single species multiple instances and different species multiple instances

The other challenge of AI-SWLM system is to detect the single species in a single instance. CSP and YOLO detection head play the main role in the same species in a single instance and can be seen in Figure 6. Basic recognition of animals is performed well by WLM-A-BW2 model with good objectiveness score greater than 95% and better mAP when compared with other models under consideration through their quantitative measures such as precision and recall. The learning ability of the best performing model is achieved due to data augmentation and having multiples layers of CSPDarknet as its backbone. Partial transition layer in CSPDarknet, with its feature fusion strategy in a hierarchical fashion contributed to classification of multiple animal species in the camera trap images. The performance of WLM-A-BW2 model in detecting the single animal species with multiple instances is shown in Figure 7 and different animal species with multiple instances in Figure 8.

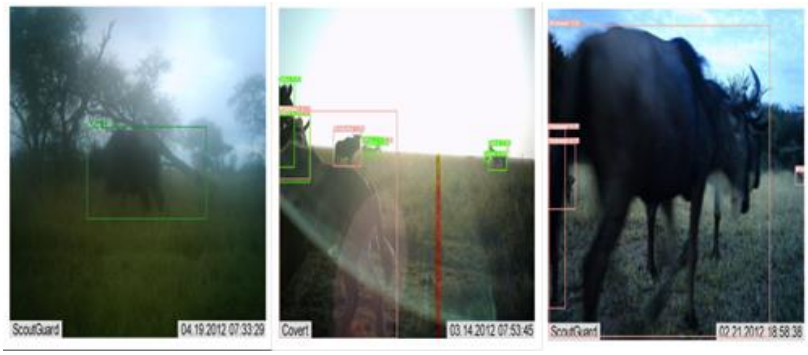


Figure 6. Single instance recognition of WLM-A-BW2 model

Figure 7. Multiple instances of same wild animal species recognized by WLM-A-BW2 model

Figure 8. Different species with multiple instances recognized by WLM-A-BW2 model

6.2. Comparison with existing works

The performance of the AI SWLM models is compared with existing state-of-the-art methods applied on the Gold Standard Snapshot Serengeti to evaluate the effectiveness of the proposed work. Very few works have used the Serengeti dataset for the purpose of animal detection and recognition. Two works that have used the same dataset as AI SWLM are [16] and [19]. These prior works have applied binary classification task of presence of the animal.

Our model focuses on object-level detection and species recognition. And hence recall is used as a primary metric to evaluate the detection of objects in wildlife images. ALexNet-96 was used for classification of Serengeti images into three classes in [16] and has achieved a recall of 73.13%. ResNet-18 architecture was employed for binary classification into empty or animal classes in [19] and has achieved an accuracy of 94.1%.

Since the AI SWLM is an object detection model, for which accuracy cannot be good measure and not a potential metric for imbalanced data. From the values reported in Table 2, it can be observed that the

recall measure of models of the proposed system perform better than the existing works in literature. The data augmentation has helped the variant of proposed model WLM-A-BW2 to perform well in detecting the wild animals with a recall of 77.88% which is 4.75% greater than the result reported in [16]. These results indicate that the AI SWLM is suitable for real-time wildlife monitoring and can provide proactive measures to address human-animal conflict problems.

AI SWLM offers fine-grained species-level object detection using YOLOv5. The use of augmented training data and pretrained weights have greatly improved generalization ability of the model to challenges such as clutter, low illumination, and multiple animal instances per frame. This shows that combining class-balancing augmentation with deep architectural variants leads to effective animal monitoring solutions.

Table 2. Comparison of AI SWLM models with existing state-of-the-art approaches on Snapshot Serengeti

Model	Classes	Recall	Precision	mAP	Remark
Animal Scanner [16]	3	73.13	-	-	Binary classification
ResNet-18 [19]	2	-	-	-	Binary classification
WLM-A-RW1 (Ours)	46	74.50	80.51	62.36	Augmented dataset, YOLOv5m, random weights
WLM-A-RW2 (Ours)	46	76.55	80.44	62.69	Augmented dataset, YOLOv5l, random weights
WLM-A-BW1 (Ours)	46	77.43	77.43	63.97	Augmented dataset, YOLOv5m, best pretrained weights
WLM-A-BW2 (Ours)	46	77.88	81.28	64.27	Augmented dataset, YOLOv5l, best pretrained weights; best overall performance

7. CONCLUSION

An automatic system to detect and recognize wild animals was designed and implemented. The AI SWLM can detect animals of different species, multiple instances of the same species and have also detected the animals with occlusion, poor illumination and from cluttered backgrounds. Extensive sets of experiments were conducted, with and without augmentation, with random and pre-trained weights to identify the most suitable model for detecting the animals. The class imbalance problem has dragged the performance down and was handled by data augmentation. From the results it was observed that the model trained on augmented data which has also used the weights from the already trained model exhibited a better performance that has almost doubled the mAP score.

The performance of AI SWLM can still be improved by fine tuning the parameters of backbone network to reduce the misclassification rate and to improve the mAP score further. With the performance achieved, the AI SWLM can be used to annotate the unseen images that can help to create a new corpus thereby reducing the manpower required for labelling. This system is suitable for integration into real-time edge AI-based surveillance systems by applying model compression techniques for deployment on embedded resource constrained devices.

FUNDING INFORMATION

No funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

All authors have made significant contributions to the research work and preparation of this manuscript. Mr. Arun G K is responsible for Conceptualization, Methodology and Writing Original Draft. Dr. J. Bhuvana led the idea formulation and model design, structured the methodology, and helped in the preparation of initial draft of the manuscript. Dr. T. T. Mirnalinee has helped in Investigation, Data Curation, Formal Analysis Visualization and providing domain knowledge with necessary computational resources. She supervised the overall progress of the project. Mr. Bharath Kumar A.M.R was responsible for dataset preprocessing, model training, and performance evaluation.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Arun Govindan Krishnan			✓	✓		✓		✓	✓		✓			
Jayaraman Bhuvana	✓	✓			✓				✓	✓		✓		
Mirnalinee Thanga Nadar Thanga Thai	✓	✓			✓							✓		
Bharathkumar Azhagiya Manavala Ramanujam			✓	✓		✓		✓			✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

No conflict of interest.

REFERENCES

[1] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, Feb. 2020, doi: 10.1007/s11263-019-01247-4.

[2] P. Arjun and T. T. Mirnalinee, "An efficient image retrieval system based on multi-scale shape features," *Journal of Circuits, Systems and Computers*, vol. 27, no. 11, p. 1850174, 2018.

[3] M. Dhinesh, S. Das, and K. Varghese, "Automatic curvilinear structure detection from satellite images using multiresolution GMM," *International Journal of Imaging Science and Engeneering*, vol. 2, no. 1, pp. 154–157, 2008.

[4] S. Beery, G. Van Horn, and P. Perona, "Recognition in Terra incognita," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11220 LNCS, pp. 472–489, 2018, doi: 10.1007/978-3-030-01270-0_28.

[5] M. A. Tabak *et al.*, "Machine learning to classify animal species in camera trap images: Applications in ecology," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 585–590, 2019, doi: 10.1111/2041-210X.13120.

[6] T. Bruce *et al.*, "Large-scale and long-term wildlife research and monitoring using camera traps: a continental synthesis," *Biological Reviews*, vol. 100, no. 2, pp. 530–555, 2025, doi: 10.1111/brv.13152.

[7] G. Jocher, A. Chaurasia, J. Qiu, and L. Stoken, "YOLOv5: Open-source object detection architecture and training method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 2240–2249. doi: 10.1109/CVPRW59100.2023.00235.

[8] A. Gomez Villa, A. Salazar, and F. Vargas, "Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks," *Ecological Informatics*, vol. 41, pp. 24–32, 2017, doi: 10.1016/j.ecoinf.2017.07.004.

[9] S. Beery, D. Morris, and S. Yang, "Efficient pipeline for camera trap image review," *arXiv preprint arXiv:1907.06772*. 2019. [Online]. Available: http://arxiv.org/abs/1907.06772

[10] S. Binta Islam, D. Valles, T. J. Hibbitts, W. A. Ryberg, D. K. Walkup, and M. R. J. Forstner, "Animal species recognition with deep convolutional neural networks from ecological camera trap images," *Animals*, vol. 13, no. 9, p. 1526, May 2023, doi: 10.3390/ani13091526.

[11] J. Bhuvana, T. T. Mirnalinee, B. Bharathi, and I. Sneha, "Efficient generative transfer learning framework for the detection of COVID-19," *Computer Science and Information Systems*, vol. 19, no. 3, pp. 1241–1259, 2022, doi: 10.2298/CSIS220207033B.

[12] Y. Zhang, Y. Bai, M. Ding, and B. Ghanem, "Multi-task generative adversarial network for detecting small objects in the wild," *International Journal of Computer Vision*, vol. 128, no. 6, pp. 1810–1828, 2020, doi: 10.1007/s11263-020-01301-6.

[13] B. Xu *et al.*, "Automated cattle counting using Mask R-CNN in quadcopter vision system," *Computers and Electronics in Agriculture*, vol. 171, p. 105300, 2020, doi: 10.1016/j.compag.2020.105300.

[14] E. M. T. A. Alsaadi and N. K. El Abbadi, "An automated classification of mammals and reptiles animal classes using deep learning," *Iraqi Journal of Science*, vol. 61, no. 9, pp. 2361–2370, 2020, doi: 10.24996/ijs.2020.61.9.23.

[15] S. B. Islam and D. Valles, "Identification of wild species in Texas from camera-trap images using deep neural network for conservation monitoring," in *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020*, 2020, pp. 537–542. doi: 10.1109/CCWC47524.2020.9031190.

[16] H. Yousif, J. Yuan, R. Kays, and Z. He, "Animal Scanner: Software for classifying humans, animals, and empty frames in camera trap images," *Ecology and Evolution*, vol. 9, no. 4, pp. 1578–1589, 2019, doi: 10.1002/ece3.4747.

[17] M. S. Norouzzadeh *et al.*, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 25, pp. E5716--E5725, 2018, doi: 10.1073/pnas.1719367115.

[18] S. Schneider, S. Greenberg, G. W. Taylor, and S. C. Kremer, "Three critical factors affecting automated image species recognition performance for camera traps," *Ecology and Evolution*, vol. 10, no. 7, pp. 3503–3517, 2020, doi: 10.1002/ece3.6147.

[19] M. A. Tabak *et al.*, "Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: MLWIC2," *Ecology and Evolution*, vol. 10, no. 19, pp. 10374–10383, 2020, doi: 10.1002/ece3.6692.

[20] N. Sheikh, "Identification and classification of wildlife from camera-trap images using machine learning and computer vision," National College of Ireland, Dublin, 2020.

[21] A. Shepley, G. Falzon, P. Meek, and P. Kwan, "Automated location invariant animal detection in camera trap images using publicly available data sources," *Ecology and Evolution*, vol. 11, no. 9, pp. 4494–4506, 2021, doi: 10.1002/ece3.7344.

[22] S. Guo *et al.*, "Automatic identification of individual primates with deep learning techniques," *iScience*, vol. 23, no. 8, p. 101412, 2020, doi: 10.1016/j.isci.2020.101412.

[23] G. Falzon *et al.*, "Classify me: A field-scouting software for the identification of wildlife in camera trap images," *Animals*, vol. 10, no. 1, p. 58, 2019.

[24] J. Zhao, W. Zhang, G. Hu, and C. Xu, "Deep learning-based wildlife species classification using camera trap images: A survey," *Ecological Informatics*, vol. 73, p. 102116, 2023.

[25] D. Bhattacharjee, A. Mukherjee, and N. Dey, "Hybrid deep learning for animal detection in natural environments using optimized YOLO variants," *Applied Soft Computing*, vol. 144, p. 110053, 2024.

BIOGRAPHIES OF AUTHORS

Arun Govindan Krishnan is an assistant professor in the Department of Computer Science in SIVET College Chennai, India. He is pursuing his research under Anna University under the guidance of Dr. J. Bhuvana. He can be contacted at email: arungk@ssn.edu.in.

Jayaraman Bhuvana is an associate professor in the Department of Computer Science and Engineering with 25 years of experience in teaching. Before joining SSN College of Engineering in 2006, she worked as an Assistant professor in AVC College of Engineering for 8 years. She received her Ph.D. from Anna University, Chennai in 2015, with master's degree, M.E. in CSE from Annamalai University, Chidambaram in 2004, with First class and Distinction. She completed B.E. in CSE from the University of Madras in 1998. Her research interests include deep learning, multiobjective optimization, memetic algorithms, evolutionary algorithms, machine learning. She can be contacted at email: bhuvanaj@ssn.edu.in.

Mirnalinee Thanga Nadar Thanga Thai currently the head of the department of Computer Science and Engineering. She received her B.E. degree from Bharathidasan University, Trichy, M.E. degree from the College of Engineering, Guindy, Anna University, Chennai, and Ph.D. from Indian Institute of Technology Madras (IITM), Chennai, India. Her research interests include Computer vision, Machine learning, Green Networks and Software Defined Networks. Seven research scholars have completed PhD under her supervision, and she is currently guiding seven more scholars. Mirnalinee has completed three research projects and has published about 80 papers in international journals and conferences. She has reviewed several papers in international journals and chaired several sessions in conferences. She can be contacted at email: mirnalineett@ssn.edu.in.

Bharathkumar Azhagiya Manavala Ramanujam is pursuing his under graduation at the Department of Computer Science and engineering at Sri Sivasubramaniya Nadar College of Engineering Chennai, India. He can be contacted at email: bharathkumar18034@cse.ssn.edu.in.