**A Paper on Object Detection using Ensemble Learning**

Allaparthi Sree Roja Rani1, Dr. G Jeyakumar2

Department of Computer Science Engineering,

Amrita School of Engineering, Coimbatore

Amrita Vishwa Vidyapeetham, Ettimadai

[cb.en.p2aid20011@cb.students.amrita.edu1](mailto:cb.en.p2aid20011@cb.students.amrita.edu1), [g\_jeyakumar@cb.amrita.edu2](mailto:g_jeyakumar@cb.amrita.edu2)

Abstract - This paper mainly focuses on Object Detection using ensemble learning. Object Detection is a challenging task in deep learning and computer vision. Also, this can be used in many real-world applications like self-driving cars, image classification, and face recognition. Basically, Object Detection is to identify all the objects in an image or a video. From the survey, the YOLO algorithm (You Only Look Once) performed better than CNN, RCNN, and the genetic algorithm. The objective of this research work is to detect objects by performing ensemble learning using multiple YOLO-V3 variants in images. As ensemble learning is a combination of multiple individual algorithms, this gives more performance to the predicted model. Here, YOLO-320, YOLO SPP, and YOLO Tiny variants are ensembled. The initial step is to predict objects using the above YOLO variants individually. Finally, to combine all the individual predicted models and detect objects in an image. This ensemble framework had overcome a challenge by reducing execution time. This model will be tested using evaluation metrics like IoU (Intersection over Union) and mAP (mean Average Precision). The output of this work will be an object with a bounding box and predicted class label and an IoU score. The ensemble model predicted objects correctly and scored an IoU of more than 0.7 out of 1 and mAP of more than 80%.

**Keywords – Object Detection, Ensemble Learning, Deep learning, Computer Vision, Self-driving cars, Image Classification, Face Recognition, YOLOV3(You Only Look Once), CNN, RCNN, Genetic Algorithm, SPP, Tiny, Iou and mAP.**

1. **INTRODUCTION**

**Problem Statement:** “Design and implementation of an algorithmic framework to solve Object Detection problems using Ensemble Learning.”

This paper mainly focuses on Ensemble Learning. Ensemble Learning (EL) is one of the types in machine learning. Machine learning is the type of learning which makes computers learn without being explicitly programmed. And the types are supervised machine learning, unsupervised machine learning, reinforcement learning, and ensemble learning. Supervised learning trains the input data and gives desired output in such a way that a function is mapped. Unsupervised learning only trains the input data and has no desired output. Reinforcement learning has a sequence of states and actions to achieve maximum rewards. This work focuses on ensemble learning, where predictions are based on combined results of individual models.

**1.1.a ENSEMBLE LEARNING**

Ensemble learning is a type of learning where predictions are based on combined results of individual models. This gives better performance and reduces variance and bias. The result is based on maximum or average voting. Types are Stacking, Blending, Bagging, and Bootstrap. Stacking is the method that creates a new model from multiple individual predicted models (Random Forest, Decision Tree, KNN). Blending is the special case of Stacking that used a validation set in place of test set to make predictions. Examples for Stacking are Stacked Models, Super Ensemble. Bagging follows bootstrapping. Bootstrapping is a technique in which the dataset is divided into several subsets with replacement. The size of the subset will be the same size as the dataset. Bagging uses these subsets to get combined results of the voting. All subset models will run at a time and so these models are independent of each other. Bagging reduces variance. Examples are Random Forest, Extra Tree, Bagged Decision Trees.

Boosting is an ensemble model where each successive or next model learns from the errors of the previous model. Each model increases the performance of the ensemble model. Boosting reduces both bias and variance. Also, uses a learning rate, a hyper-parameter. Types are: Adaboost (Adaptive Boosting) adds weights to the points which are incorrectly predicted and the successive model will predict these values correctly, Gradient Boosting (GBM) uses regression trees, a base learner, and each successive tree is built on previous tree errors, Extreme Gradient Boosting (XGBoost) also known as Regularized Boosting and works faster than other algorithms. Also uses regularization techniques to reduce overfitting and improve overall performance, Light GBM performs better than other algorithms when the data is large and follows a leaf-wise approach, Cat GBM mainly handles categorical variables(classification) in data and does not require any data pre-processing.

1. LITERATURE SURVEY

In modern times, forest fire detection [1] is a challenging task because of its shapes, textures, and colors. This work is implemented using the ensemble learning method. The models Yolov5 and EfficientDet are combined to detect the fire in this problem. And, EfficientNet is used to avoid false positives. Final predictions are based on the decision of all these models. Images were collected from multiple public fire datasets: BowFire, FD-dataset, ForestryImages, VisiFire and integrated all above into one which contains 10581 images with both fire and non-fire images. Non-maximum suppression is applied after integration of Yolov5 and EfficientDet and the model EfficientNet gives classified results. These results are evaluated based on metrics like frame accuracy (FA) and false positive rate (FPR). Finally, the ensemble model performed better in dark climates.

This paper work [2] focuses on detecting moving objects from traffic video surveillance. Continuous video tracking leads to some issues like a greater number of moving items suddenly, alter in light conditions, size variety, and darkness conditions. Firstly, extract necessary video frames and apply a genetic vehicle detection algorithm. Then, generate population by randomization method, Calculate the value of fitness and choose chromosomes randomly and then apply crossover and do validation of chromosomes. This proposed system is quick, easy, efficient. This system detects images, background in a video and gives bound boxes with a detected class label.

Federated Learning [3] is basically collaborative machine learning without centralized training data. Object detection is a sub-task of Face recognition. This system proposes a model trained with ensemble learning and evolutionary methods. First, the population is initialized and performs genetic operators like mutation, crossover, selection. From this, population gives inference to ensemble learning (to combine individual models). This work uses Random Forest of Ensemble learning with multiple CNNs. Datasets namely GFR-R and GFR-V are taken from the NDWD platform, which contains all human faces over the globe. This federated ensemble learning model performs better than the federated averaging and federated file system.

The objective of this project work [4] is Real-time object detection of genetic algorithms. This system detects distance and key elements on a football. This depends on the correlation between captured images and information gained in key elements. In this process, the population will be initialized randomly and calculate the value of fitness. Information from previous populations gives the correlation between consecutively captured images. This technique yield results in short execution time by reducing the number of iterations and individuals. This system gives good performance and efficiency.

This paper [5] is about video segmentation and moving object detection to categorize defined classes. Video segmentation is performed using GDSM and foreground detection. These techniques are used for identifying the moving object and the max distance covered by an object in the group of boxes. Steps for this proposed technique: Video Frames, Pre-processing, Background Modelling, Foreground Detection, Data Validation, and Foreground Masks. Finally, this system detects the object by removing background subtraction.

The paperwork on COVID-19 detection [6] in chest radiographs are classified based on ensemble learning. Here, multiple CNNs are used to perform ensemble and get good performance in Computer Vision tasks. This study has many limitations like random noise during training, artificial images, deep learning behaviour to take decisions, variability affecting learning and evaluation. The steps in overcoming these limitations are pre-training using ensemble learning and then statistical analysis at learning stages.

The state-of-the-art research works in the area chosen for this project work are discussed here. The work presented in [7] mainly focuses on implementing object detection using stacked YOLO v3(You Only Look Once) by finding bounding boxes. The model is evaluated using COCO dataset and Intersection over Union (IoU) is the suitable metric for evaluating results. The class of object is classified correctly if the IoU of the box is greater than 0.5. NMS is used to find the best bounding box. Yolov3 is the fastest algorithm for real-time object detection which uses the Darknet-53 network and has 53 CNN layers to recognize 80 classes. Two different types of threshold values were used to evaluate the model and both performed with a good accuracy score and detected all objects in an image at threshold 0.5.

This paper [8] works on deep learning-based image analysis for road damage detection. To solve this object detection problem, three ensemble learning approaches were implemented:

1. Ensemble Model Approach (EM) – which uses multiple trained models for prediction
2. Ensemble Prediction Approach (EP) – which applies an ensemble of the predictions obtained from images generated by the test time augmentation procedure.

Test Time Augmentation (TTA): This applies several transformations like horizontal flipping, increasing image resolution to test an image.

1. Hybrid Approach (EM+EP) – which uses an ensemble model from EM for generating predictions for the images generated by the TTA procedure.

The image dataset is taken from IEEE Big Data 2020 Global Road Damage Detection Challenge was collected from three countries: the Czech Republic, India, and Japan. Here, the training set consists of 21,041 images with some classes. Models are implemented using Faster-RCNN, YOLOv3, u-YOLO approaches. In these, u-YOLO achieved the highest F1 scores and this model is used for further evaluation by considering hyperparameters on augmented data.

This paperwork [9] discusses object detection using the SSD MobileNet algorithm. While working with large datasets, the model should be trained with high-performance machines like GPUs. Here, Single-shot detector-MobileNet (SSD) is used to predict multiple class objects in images. This research work is to train the model using CNN. Also, single-class object detection classifiers are also performed and this is given to android devices. SSD yields good advantages of speed and performance. The proposed system uses SSD and MobileNet algorithms for larger data and faster performance.

This paper [10] mainly focused on Object Detection by combining CNN and Adaptive Color Prior Features. One of the challenges in this problem is insensitive to scale, light, and dark conditions. To improve the accuracy of predicted models, this paper has found an approach to model color priors. The initial step is to get the color of prior features of target samples by a cognitive-driven model. Then, these features are weighted accordingly and obtain a prior-based saliency image. These images are called as features maps and merged with a convolutional neural network at the extraction stage. Here, this proposed system has experimented on Cascade R-CNN, SSD300, Libra R-CNN, Retina Net.

Object Detection and tracking [11] is one of the challenging tasks in computer vision and have been used in many sectors like health care, self-driving cars. By the growth of deep learning and GPU computation, the performance of object detection models has been improved. This paper is the study of all models to perform object detection. Using GPU may increase computation power. This paper covered two-stage detectors like RCNN, Fast RCNN, R-FCN, FPN, Mask RCNN, and one-stage detectors like YOLO, YOLOv2, YOLOv3, SSD, RetinaNet, DSSD, and RefineDet. And also discussed some applications like face detection, pedestrian detection, and object tracking.

This paper [12] deals with the object detection process in image processing applications. For machines to detect the object, some algorithms like feature descriptors are used. Generally, there are two approaches like global and local descriptors. But some difficulties arise in light conditions, positioning, mirror images, rotation, or flipping. Because of these difficulties, a lot of descriptors need to be used. Now the problem is to select the optimal number of descriptors to give a good accuracy score. This can be achieved by using an optimization algorithm named differential evolution (DE). This approach is implemented and tested on some images.

This paperwork [13] is done on atomic clocks using an ensemble algorithm. The principle of atomic clocks is to work on frequent times generated by the atoms of the elements. This timescale is generated from atomic clock frequencies. Here, ensemble learning performs the combination of atomic clocks to yield an optimal clock. This clock is stable and optimal in frequency. The proposed system performed an artificial neural network (ANN) ensemble approach to observe the changes in clock behavior. For implementing this, overlapping deviations are given as input to the model.

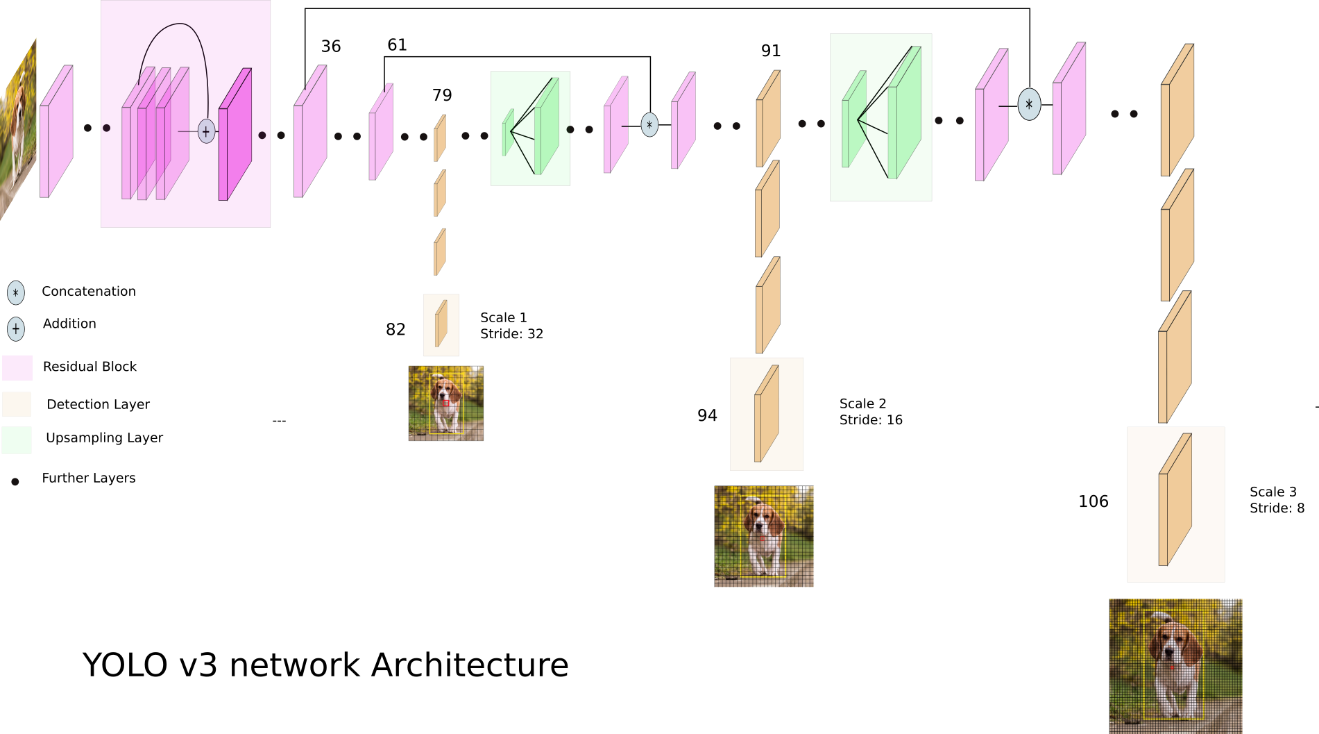
Object detection [14] is an important task in deep learning. This work is done on a smart surveillance system architecture. Object detection can be used in many applications like voice recognition, face recognition, image classification. This also identifies the latest trends in the dataset and analysed them. These algorithms are capable of identifying an object in the region of interest and in the pooling layer. Smart technology is used to continuously monitor the conditions. In this comparative study of all deep learning algorithms for smart surveillance systems, YOLO architecture and its versions got efficient results for object detection.

1. PROPOSED SYSTEM

The block diagram of the proposed system is as follows:



And the architecture of each YOLOV3 model will be as follows:



YOLO is an object detection algorithm that identifies objects in images and videos. YOLOV3 uses a variant of Darknet which has 53 CNN layers. Another 53 layers are stacked to detect the object. Totally, 106 CNN layers architecture is used for yolov3.

Here, three YOLOV3 models are stacked. They are YOLOV3-320 weights, YOLOV3-Tiny, YOLOV3- SPP.

YoloV3 uses down-sampling(stride=2) in convolutional layers.

YoloV3 Tiny uses down-sampling(stride=2) in Max-pooling layers.

YoloV3 SPP uses down-sampling(stride=2) in Convolutional layers and Max-pooling layers.

These three models are combined using the ensemble boxes library to classify the object in an image.

**3.2 DATASET**

**COCO Dataset:**

COCO dataset is a large dataset that consists of 330K images and 80 object-defined classes used for object detection, object recognition, and image segmentation problems. Many datasets can be created for required classes by using this coco dataset. This dataset is taken from the link: <https://cocodataset.org/>

YOLO models are already trained on COCO dataset and gave a good performance. By using this pre-trained model, there is a chance to reduce execution time. This can solve a challenge. These pre-trained models are taken from: <https://pjreddie.com/darknet/yolo/>

These trained models consist of weights and configuration files. All the layers in a network are optimized and mapped in weights and configuration files.

**3.3 RESULTS:**

The object detection problem has some specific evaluation metrics. They are IoU, mAP.

IoU: It is the ratio of area of intersection to area of union in bounding boxes.

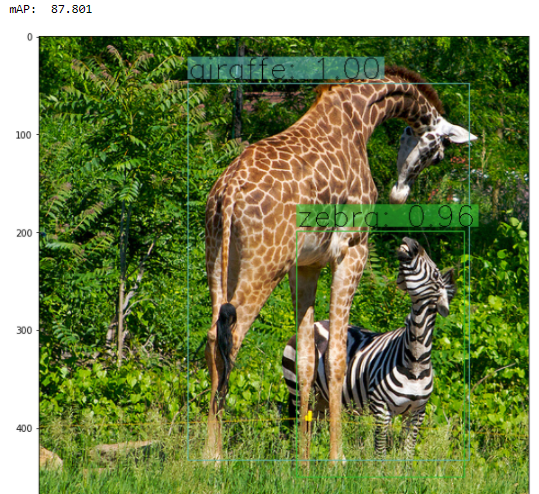
If IoU is greater than 0.5, the class is predicted correctly.

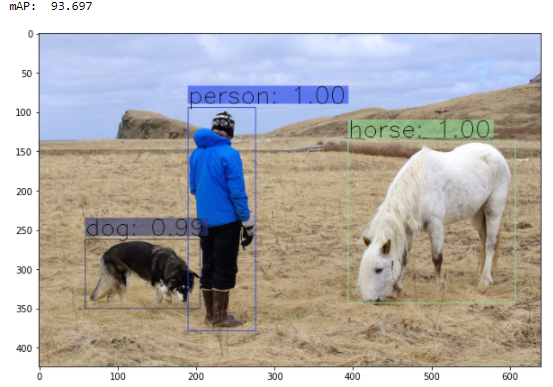
If IoU is less than 0.5, the class may not be predicted correctly.

Mean Average Precision(mAP): It is calculated by considering average AP over all classes or overall IoU thresholds.

**3.3.1 YOLOV3-320 Weights:**

Here, the objects in the images are classified correctly with IoU more than 0.9.

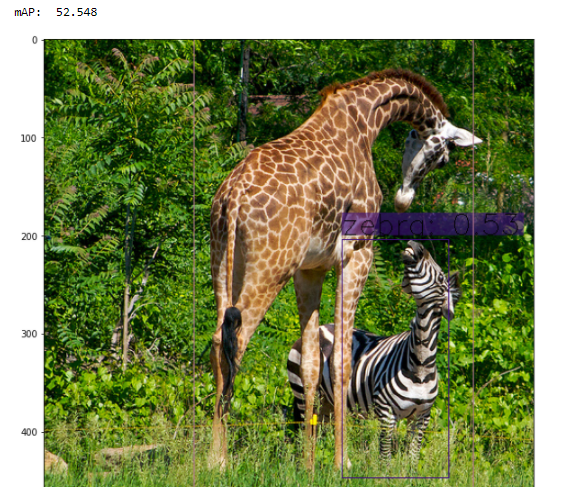
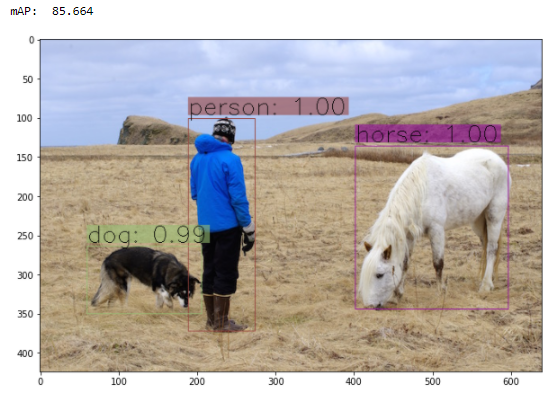




Each image is classified with bounded boxes and predicted class, IoU score is displayed above the bounded box. All the objects in the images are classified correctly.

This model performed well in all three models with good IoU and mAP Scores.

* + 1. YOLOV3 SPP:



Here, the model had predicted objects with good IoU and mAP. But it couldn’t predict some objects like giraffe and horse in the above images.

**3.3.3 YOLOV3 Tiny:**



A giraffe and a zebra in a zoo exhibit

Description automatically generatedA picture containing text, grass, outdoor, mammal

Description automatically generated

In this model, some objects are predicted differently from the above two models like sheep.

**3.3.4 Ensemble Model:**

As Ensemble is a combination of individual models, this model is predicted based on three predicted models.

The predictions of ensemble yolo model are as follows:

A picture containing text, mammal

Description automatically generated



This model has predicted objects by merging all the three different yolov3 predicted models.

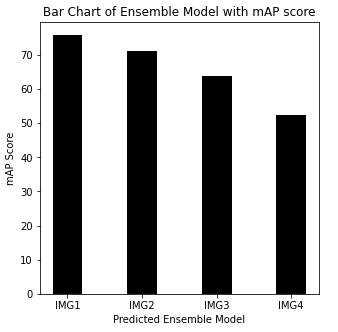
The results of all individual models and predicted ensemble model are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **METRICS** | **YOLO-320** | **YOLO SPP** | **YOLO TINY** | **ENSEMBLE MODEL** |
| **mAP** | 93.6,90.8,85.9,90.8 | 85.6,86.7,64,52.5 | 75.9,71.2,68,87.8 | 75.9,71.2,64,52.5 |
| **IMG-1**  **IMG-2**  **IoU IMG-3**  **IMG-4** | P-1, H-1, D-0.99  B-0.99, D-1, T-0.94  H-0.88,0.9,0.96,0.9  G-1, Z-0.96 | P-1, H-1, D-0.99  B-0.9, D-0.98, T-0.87  H- 0.56, 0.7, 0.8  Z-0.53 | P-0.98, S-0.7, D-0.89  D-0.83, C-0.73  H-0.8, 0.98,0.59, 0.98  G-0.95, Z-1 | P-0.98, S-0.7, D-0.89  D-0.83, C-0.73  H-0.56, H-0.70  Z-0.53 |

Here, P- Person, H-Horse, D-Dog, S-Sheep, B-Bicycle, C-Car, T-Truck, G-Giraffe, Z-Zebra.

From the table, we can see the results of all individual YOLO v3 variant models and an ensemble model. In these individual models, the YOLO-320 model predicted the models correctly and got a good IoU and mAP score. Our proposed system Ensemble model combined the results of all individual predictions and gives a new model. This model predicted objects based on all individual models and got a good score. Therefore, one individual model may not give correct predictions. So, our proposed system ensemble model is required to predict objects correctly by combining individual models.

The bar chart of the proposed system with mAP score for four different images:



**IV. CONCLUSIONS AND FUTURE ENHANCEMENTS**

This paper presented work on object detection using ensemble learning. This ensemble model gave good performance by combining the results. The evaluation metrics like Intersection over Union and mean Average precisions are calculated, got the good score performance in each model. All the objects are predicted correctly. While in all three models, yolov3 320 weights model had performed better. Also, this research work has solved one of the challenges by using trained yolo models.

REFERENCES

1. 1, R. X. (2021). A Forest Fire Detection System Based On Ensemble Learning. Multidisciplinary Digital Publishing Institute.
2. Dey, J. (2016). Moving Object Detection Using Genetic Algorithm For Traffic Surveillance. International Conference On Electrical, Electronics, And Optimization Techniques (Iceeot) - 2016.
3. Li, L. (2021). Evolutionary-Based Federated Ensemble Learning On Face Recognition. Imcec.
4. Mart´Inez-G´Omez, J. (2010). Using Genetic Algorithms For Real-Time Object Detection. Springer.
5. R. Aruna Jyothi, K. R. (2019). Moving Object Detection Using The Genetic Algorithm For Real Times Transportation. International Journal Of Engineering And Advanced Technology (Ijeat).
6. Rajaraman, S. (2020). Analyzing Inter-Reader Variability Affecting Deep Ensemble Learning For Covid-19 Detection In Chest Radiographs. Plos One.
7. Sai Shilpa Padmanabula, R. C. (2020). Object Detection Using Stacked Yolov3. International Information And Engineering Technology Association.
8. Hegde, V. (2020). Yet Another Deep Learning Approach for Road Damage Detection using Ensemble Learning. *IEEE International Conference on Big Data (Big Data)*.
9. Dr. Senthil Kumar T., D. L., & Subramani, G. (2021). A Mobile-Based Framework for Detecting Objects Using SSD-MobileNet in Indoor Environment. *Intelligence in Big Data Technologies–-Beyond the Hype, Springer Singapore, Volume 1167, Singapore, p.65-76.*
10. Peng Gu, X. L. (2021). Object Detection Combining CNN and Adaptive Color Prior Features. *MDPI*.
11. Sankar K Pal, A. P. (2021). Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence*.
12. Dr. Jeyakumar G., S. K. (2020). An Evolutionary Computing Approach to Solve Object Identification Problem in Image Processing Applications. *Journal of Computational and Theoretical Nanoscience,vol. 17, pp. 439-444.*
13. Dr. Rajathilagam B., S. N., Maharana, S., Subramanya Ganesh, T., & Krishnamoorthy, S. (2020). An Artificial Neural Network Model for Timescale Atomic Clock Ensemble Algorithm. *MAPAN- Journal of Metrology Society of India, vol. 35, no. 4, pp. 547 - 554*.
14. Dr. Senthil Kumar T., U. S., Kumar, D. K., & Parameswaran, L. (2020). An Extensive Study and Comparison of the Various Approaches to Object Detection using Deep Learning. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*.