Received: April 19, 2023
Revised: September 19, 2023
Accepted: September 20, 2023

NILE UNIVERSITY OF NIGERIA
HONORIS UNITED UNIVERSITIES

Comparative Study of Some Deep Learning Object Detection Algorithms: R-CNN, FAST R-CNN, FASTER R-CNN, SSD, and YOLO

Oluwaseyi Ezekiel Olorunshola[1*] | Adeniran Kolade Ademuwagun[2]| Charles Dyaji[3]

[1]Department of Computer Science, Air Force Institute of Technology, Kaduna, Nigeria
[2,3]Electrical Electronic Engineering Department, Air Force Institute of Technology, Kaduna, Nigeria

Corresponding Author's Name:
Oluwaseyi Ezekiel Olorunshola
Corresponding Author's E-mail:
seyisola25@yahoo.com

ABSTRACT

Due to its numerous applications and new technological advancements, object detection has gained more attention in the last few years. This study examined various uses of some deep learning object detection algorithms. These algorithms are divided into two-stage detectors like Region Based Convolutional Neural Network (R-CNN), Fast Region Based Convolutional Neural Network (Faster R-CNN), and Faster Region Based Convolutional Neural Network (Faster R-CNN), and one-stage detectors like Single Shot MultiBox Detector (SSD) and You Only Look Once (YOLO) algorithms that are used in text and face detection, image retrieval, security, surveillance, traffic control, traffic sign/light detection, pedestrian detection and in medical areas among others. This research primarily focuses on three applications: drone surveillance, applications relating to traffic, and medical fields. Findings from the performed analysis indicate that YOLO stands out as the predominant algorithm for drone surveillance among different deep learning models used in various application fields and being a one-stage detector. In terms of usage in traffic-related applications, SSD proved to be a prominent one-stage detector alongside Faster R-CNN which gained popularity as a two-stage detector preferred for applications in the medical field.

KEYWORDS: R-CNN, Fast R-CNN, Faster R-CNN, SSD, YOLO

DOI: *https://doi.org/10.5455/NJEAS.150264*

NJEAS
NILE JOURNAL OF ENGINEERING & APPLIED SCIENCE

1 | INTRODUCTION

One of the most important computer vision (CV) tasks is object detection, which is concerned with identifying instances of specific classes of visual objects (like people, animals, or vehicles) in digital images. The goal of object detection is to create computational models and methods that produce one of the key bits of data needed by CV applications [1]. Both the spatial information of the picture and semantic cues should be well understood by a good detection algorithm. In reality, the first stage in many CV applications, including face recognition, pedestrian detection, video analysis and logo detection, is object detection.

The foundation of many other CV tasks, including instance segmentation, picture captioning, object tracking, etc., is object detection which is one of the fundamental CV snags. When it comes to applications, object detection can be divided into two categories: general object detection and detection applications. The former focuses on exploring ways to find various types of objects within a single framework to mimic human vision and cognition, while the latter discusses detection in relation to particular application scenarios, such as pedestrian detection, face detection, text detection, etc. According to [2], deep learning-based object detection methodologies have undergone intensive research in the past few years as a result of the enormous success of deep learning-based picture classification. From traditional methods of detection to two-stage detectors and most recently, one-stage detectors. Object recognition algorithms have made significant strides, in terms of speed of detection, model weight, model architecture and model generalization. Enormous advancements in object identification are being achieved.

The goal of object detection is to determine whether any instances of specific objects like people, cars, bicycles, dogs, or cats can be found in an image and, if so, to return the spatial position and scale of each instance. Robot vision, consumer electronics, security, autonomous driving, Human Computer Interaction (HCI), content-based picture retrieval, intelligent video surveillance, and augmented reality are just a few of the uses that can benefit from object detection [3]. The goal of this research is to compare various object detection techniques that have been employed over time and identify the object detection algorithm that is most appropriate for various CV applications.

The rest of this paper is presented as follows; Section 2 briefly reviews different object detection algorithms. Section 3 contains different application areas of the detection algorithms reviewed. Section 4 analyses the result of the comparison done and finally Section 5 details the conclusion of this study

2 | LITERATURE REVIEW

In the deep learning era, object recognition algorithms were divided into two genres: two-stage detection and one-stage detection. While two-stage detection performs its detection process as a "coarse-to-fine" process, one-stage detection frames the detection process to finish in a single phase. While the one-stage detectors include SSD and YOLO, the two-stage detectors include R-CNN, Fast-RCNN, and Faster-RCNN. The following is a further explanation of these recognition algorithms:

2.1 Region-Based Convolutional Neural Networks (R-CNN)

A detection algorithm called Region Based Convolutional Neural Network (R-CNN) employs selective search to generate 2000 regions from input images, which are referred to as region proposals [4]. The first of the two-stage detectors, which conceptualizes detection as a "coarse-to-fine" procedure, it was the first to be used. There are three components that make up the object detection system. The first creates category-independent region proposals. These suggestions outline the range of potential detecting candidate detections. A large convolutional neural network (CNN) in the second module extracts a fixed-length feature vector from each area. A collection of class-specific linear Support Vector Machines (SVMs) makes up the third module. Figure 1 shows the R-CNN algorithm in action.

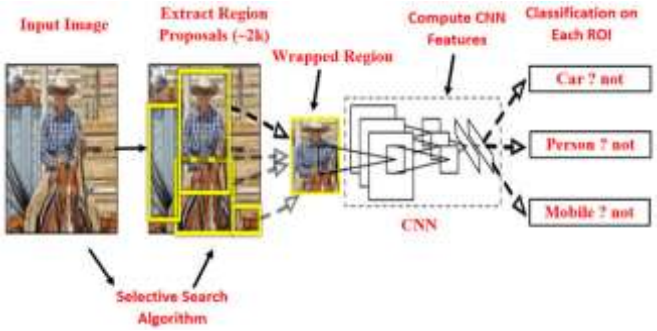

Figure. 1. R-CNN Algorithm [5].

The CNN receives the 2000 suggested regions and outputs a 4096-dimensional feature vector. With a mean Average Precision (mAP) of 31.4%, R-CNN considerably outperforms OverFeat's second-best performance of 24.3%.

The cutting-edge R-CNN object detector's subspace alignment-based domain adaptation was suggested [5]. This paper's primary concern was obtaining localized domain adaptation for object detector adaptation. By employing a method based on subspace alignment that skillfully transfers the original subspace to the target subspace, the adaptation was successfully completed. The suggested technique enhanced object recognition.

Based on the obvious finding that contextual cues are important in determining what action a person is

undertaking, R-CNN was modified to use more than one region to model a prediction [6]. The outcome demonstrated that R-CNN can be used effectively for projects like attribute categorization.

Even with the increase brought about by R-CNN, the model's redundant large feature computation made it very slow at detection. This flaw in R-CNN created space for additional advancements in object recognition

2.2 Fast Region-Based Convolutional Neural Network (Fast R-CNN)

An improvement over R-CNN is Fast Region Based Convolutional Neural Network (Fast R-CNN). The image is inputted in order to create a convolutional feature map rather than the 2000 region plan [7]. Fast R-CNN uses a number of innovations to increase detection accuracy while also increasing training and testing speeds. Fast R-CNN achieves an elevated mean average precision (mAP) on the PASCAL Visual Object Classes (VOC) 2012 benchmark by training the very deep Visual Geometry Group (VGG16) network 9 times faster than R-CNN. It is 213 times faster during testing, and it is faster overall. Fast R-CNN trains VGG16 3x quicker, tests 10x faster, and is more accurate than spatial pyramid pooling (SPPnet).

The fast R-CNN technique has the following benefits: Higher detection quality (mAP) than R-CNN, updating all network levels during training, and feature caching without the need for disk storage are just a few of the benefits. The following three major findings lend credence to this: Modern mAP on VOC 07, 2010, and 2012, as well as fast training and testing in comparison to R-CNN and SPPnet and fine-tuning convolutional layers in VGG16 enhances mAP. The fast R-CNN's architecture is shown in Figure 2.

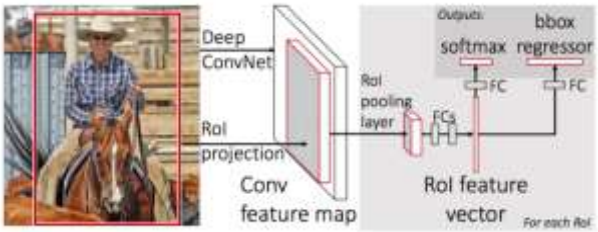

Figure. 2. Fast R-CNN Architecture [7]

A CNN that is built on the fast R-CNN architecture was used to spool reliable pedestrian features for effective and practical pedestrian detection in complex environments [8]. Investigations revealed that the suggested technique performs satisfactorily on the INRIA and ETH datasets.

Using fast R-CNN features, Li et al sought to recognize and classify fish species from underwater images [9]. The results of the experiment showed that the fast R-CNN performed admirably, increasing mAP by 11.2% in comparison to the Deformable Parts Model (DPM) baseline, reaching a mAP of 81.4%, and detecting 80 times fast than previous R-CNN on a single fish image.

2.3 Faster Region-Based Convolutional Neural Network (Faster R-CNN)

For effective and accurate region proposal generation, Region Proposal Networks (RPNs) was introduced [10]. The region proposal phase is essentially cost-free due to the down-stream detection network's sharing of convolutional properties. A unified, deep-learning-based object detection system can now operate at frame speeds that are nearly real-time. The total object detection accuracy is increased by the learned RPN, which also improves the quality of region proposals. R-CNN and fast R-CNN use selective search, which is slow and time-consuming to find region proposals but faster R-CNN employs RPN which is faster. The Faster R-CNN algorithm is shown in Figure 3.

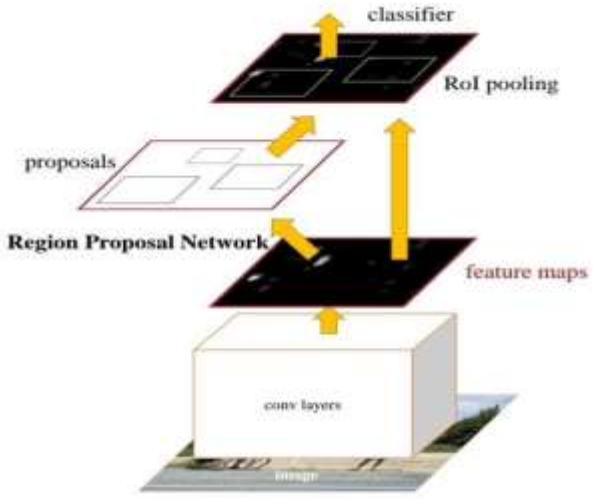

Figure. 3. Faster R-CNN Algorithm [10]

For the purpose of recognizing facial expressions, [11] suggested faster R-CNN (Faster Regions with Convolutional Neural Network Features). The performance and generalizability of the faster R-CNN for facial expression recognition were demonstrated by experimental findings. The mAP number was around 82%.

The faster R-CNN framework was modified by [12] for the purpose of detecting objects buried beneath the earth. The proposed method is quite optimistic to deal with this kind of specific ground penetrating radar (GPR) data even with a few training samples, according to preliminary detection findings, which demonstrate that it can lay out substantial developments compared to classical CV methods.

2.4 You Only Look Once (YOLO)

By dividing a picture into S x S grids, YOLO creates individual grids with unique bounding boxes. The network produces a class probability and an offset number for each bounding box. To find the object in the image, bounding boxes with class probabilities above a threshold number are used. YOLO, in contrast to the R-CNN series, a fresh strategy to object detection [13].

Classifiers are used in earlier work on object detection to accomplish detection [13]. Instead, bounding boxes and related class probabilities were spatially separated using frame object detection as a regression problem. Bounding boxes and class probabilities are immediately predicted by a single neural network from complete images in a single evaluation. Since the entire detection pipeline consists of a single network, detection performance can be optimized from beginning to finish. This network divides the image into regions and concurrently forecasts the bounding boxes and probabilities for each region. Later series built on the foundation of YOLO presented its enhancements in the v2 and v3 editions, further improving the detection accuracy while maintaining a very high detection speed [14], [15].

The YOLOv4, YOLOv5, YOLOv6, YOLOv7 and most recently the YOLOv8 are versions of YOLO. The earlier versions aim to build on the earlier ones. Figure 4 depicts the standard YOLO algorithm.

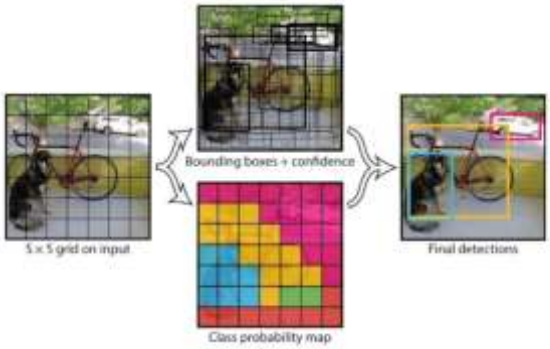

Figure. 4. YOLO Algorithm [13]

In the Liris Human Activities dataset, [16] showed that YOLO is an efficient technique and relatively quick for recognition and localization.

Early versions of YOLO experience a drop in localization accuracy compared to two-stage detectors, particularly for some small objects, despite their significant increase in detection speed. Later versions of YOLO attempt to resolve the issue, and Single Shot Multibox Detection (SSD), another one-stage detector, was also proposed.

2.5 Single Shot Multibox Detection (SSD)

The second one-stage detector to be proposed during the deep learning era was the Single Shot Multibox Detector (SSD). The introduction of multi-reference and multi-resolution detection techniques is the main contribution of SSD, which greatly raises the detection accuracy of a one-stage detector, particularly for some small objects. According to the results of its detection using the following datasets with the corresponding results: VOC07 dataset with a mAP of 76.8%, VOC12 dataset with a mAP of 74.9%, and COCO dataset with mAP@.5 of 46.5% and mAP@0.5:0.95 of 26.8%, a fast version runs at 59fps. SSD has advantages in terms of both detection speed and accuracy. The primary distinction between SSD and any prior detectors is that the former runs detection of objects of 5 different scales on different layers of the network while the latter run detection on their top layers [3]. SSD architecture is shown in Figure 5.

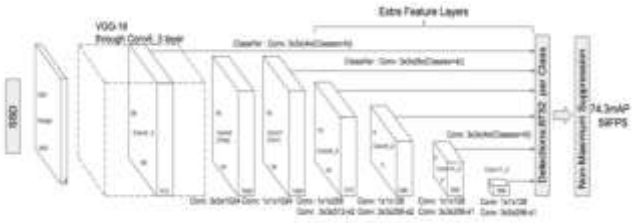

Figure. 5. SSD Architecture [3].

Over the years, numerous writers have made significant contributions to the advancement of object detection. Figure 6 shows the authors' contributions as well as the years in which they put forward various object detection models. Over the years, authors like Girshick and Redmon have continued to work to improve object detection by continuously introducing new ideas and features for a better detection algorithm, and more works are being implemented using the authors' suggested model and improvements on them.

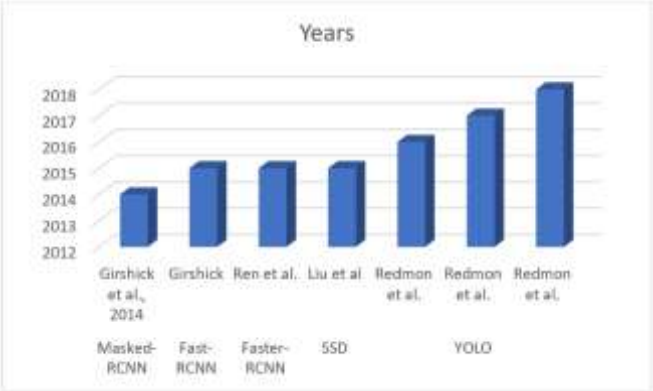

Figure. 6. Authors and Years for the Object Detection Algorithm.

The Spatial Pyramid Pooling Networks (SPPNet) [17] and the Feature Pyramid Networks (FPN) [18] are additional two-stage detectors worth mentioning from the contributions of the previous authors and additional advancements on object detection, while RetinaNet is a one-stage detector [18].

3 | APPLICATION AREAS OF THE DETECTION ALGORITHMS

Three application areas were examined in this study because they significantly influenced the use of object detections and offered significant advancements over the conventional approaches to managing them. These include medical disciplines, traffic-related detection, and drone surveillance.

3.1 Drone Surveillance

In order to make object detection in Unmanned Aerial Vehicle (UAV) imagery fast and extremely accurate, Vandersteegen et al. suggested a methodology [19]. The basic architecture used was standard YOLOv3, which can identify 80 different object classes and support a variety of square input resolutions (320 x 320, 416 x 416 or 608 x 608). The Microsoft Common Objects in Context (MS COCO) and Visdrone 2018 databases were used to train the detector. They were able to implement this methodology by introducing a multi dataset learning strategy that, on Visdrone 2018 and MS COCO, significantly outperformed transfer-learning by 3.5% mAP and 50% mAP, respectively. This led to the creation of a general model that could be applied to all UAV operations.

In order to track UAV targets, [20] introduced a novel technique known as tracking by detection. Real-time, high-precision monitoring was accomplished using this technique. In order to obtain a detection speed of 54 FPS and a multiple object tracking accuracy (MOTA) of 94.4%, this process used YOLOv3 as the detector and deep simple online and real time tracking (DeepSORT) as the tracking mode. Based on these findings, the YOLOv3 network was altered by altering the loss function of the model to account for the drone target's traits and by including the spatial pyramid pooling (SPP) module to gather drone data and produce the initial anchor. With a fast detection speed of only 5 FPS, these operations improved the MOTA of the modified YOLOv3-SPP network by 2%. The anchor-free option was used by the authors to follow UAV targets as a comparison. With CenterNet acting as the target detector and DeepSORT acting as the tracker, this method is frequently used for target tracking. This model's final recognition speed was 25 FPS, and its MOTA was 66.4%, which only marginally improved real-time tracking. The experimental findings supported the effectiveness of the suggested strategy and demonstrated that YOLOv3-SPP + DeepSORT is a highly effective method for tracking multiple targets in real time for UAVs.

An online multiple object tracking (MOT) framework based on a UAV system was suggested by Huang et al [21], and it featured a cutting-edge, high-performance target detector called Hierarchical Deep High-Resolution Network (HDHNet). The HDHNet blends the benefits of high-resolution representation networks and hierarchical aggregation networks. It is easy to extract multi-scale and high-resolution characteristics, which were then applied to various prediction networks. To train their model, an adjustable loss function was also suggested, which can also address the issues of class imbalance and difficult data. In order to achieve a lower computational complexity than high resolution network (HRNet), the suggested model had significantly fewer parameters and floating-point operations per second (FLOPs). On the VisDrone2019 multiple object tracking (MOT) benchmark, the experiments and results demonstrated that the technique was capable of the highest MOTA and precision when compared to state-of-the-art methods.

[22] proposed an approach for real-time implementation of CNN-based object detector and tracking system for Augmented Reality (AR) Drone 2 using SSD. Implementing SSD with one class detection in real-time systems, where computational time is a major factor and the system's effectiveness is a difficult job. A method that could be used in real-time that would maximize SSD architecture's peak accuracy while minimizing computational time was created. The suggested system had two components: target object tracking and object detection. Based on several experiments, the target object tracking accuracy was 96.5%, the speed was 58 frames per second, and the efficiency for object detection was 98%. The obtained results demonstrated that more accurate results can be obtained within an acceptable computation time by using an SSD object detector. Notably, implementing such an approach in a sophisticated system like a drone is very doable.

CNN algorithms' basic architecture was examined by Jain et al. [23], along with a synopsis of YOLO's real-time object detection algorithm. CNN architecture models are able to identify objects in each picture and remove highlights. When used correctly, CNN models can address deformity identification and the development of educational and instructional applications. In contrast to other CNN algorithms, YOLO has a lot of advantages in practice. YOLO can train the complete model in parallel because it is a unified object detection model that is easy to build and train in accordance with its simple loss-function. The state-of-the-art, optimal tradeoff between speed and accuracy for object detection is offered by YOLOv2, the second main edition of the software. The most advanced object detection method, YOLO, may be suggested for real-time object detection because it is more effective than other object detection models at generalizing object representation.

According to [24], CNNs have improved over time, offering better detection rate, precision, and performance. SSD is one of the CNNs that has attained performance that is almost at the cutting edge while maintaining a low processing cost. Modern F-RCNN's hidden layer has been optimized and minimized, enabling SSD to perform multiple times as many detections per second. For a hybrid static-mobile drone surveillance system, a pipeline for human action tracking and recognition was suggested. The suggested pipeline was created using the KERAS library and its backend was TensorFlow from Google. The experimental outcomes demonstrated that the suggested pipeline can recognize gaits with 99.51% accuracy using the KTH dataset.

In [25], drone recordings were used to train and assess a YOLOv4 model. At three various altitudes: 20 feet, 40

feet, and 60 feet. The trained YOLOv4's performance was evaluated in real time. The performance was then determined using the mAP and FPS evaluation metrics. The YOLOv4 obtained a mAP of 74.36% at an IoU of 50 and FPS of 19.0 for the DJI MAVIC Pro and FPS of 20.5 for the DJI Phantom III using a Tesla T4 graphics processing unit (GPU) and OpenCV (3.2.0).

3.2 Traffic: Recognition of Sign, Vehicle, Pedestrian, and Light

With the introduction of autonomous driving, tracking traffic offenders, among other things, object detection in traffic has received a lot of attention recently. It is used to identify traffic signs, vehicles, pedestrians, and traffic lights. The use of CV in traffic is not a straightforward issue, but rather another difficult one. According to Zou, et al. [1], some of the difficulties encountered when applying CV to traffic include poor weather, change in lighting conditions, motion blur, and real-time detection. Many studies and techniques have been put forth to use detection algorithms in this situation.

Pedestrian recognition is one of the first CV tasks to use deep learning, according to Hosang et al. [26]. To test how well the algorithm works for detecting pedestrians using feature fusion, Zhang et al. [27] built Faster R-CNN.

For intelligent transportation systems that classify and detect vehicles, Vijayaraghavan and Laavanya [28] suggested a technique combining the RPN, Faster-RCNN, and Fast-RCNN by training and retaining the model with all three algorithms. Bus, car, and motorbike detection classes were used in the study, and they produced better results than Fast R-CNN when compared to those classes. Bus detection accuracy ranged from 0.9 to 0.85, car detection accuracy ranged from 0.87 to 0.8, and motorbike detection accuracy ranged from 0.85 to 0.76.

One of the major difficulties encountered with traffic sign detections is the false detection of traffic signs. You et al in a bid to solve this challenge proposed a lightweight SSD network algorithm by downscaling the baseline SSD network's 3 x 3 convolution kernel to a 1 x 1 convolution kernel, removing some of the convolution layers to help reduce the calculation load, and incorporating an enhancement strategy to enhance the dataset's overall balance [29]. Compared to the baseline SSD network with a 1.2 times faster detection speed, the experiment results revealed that the suggested method is 3% more accurate in detection.

A model for automatic traffic density estimation by vehicle counting using SSD and MobileNet-SSD was put forth by Biswas et al. [30]. The algorithm was applied to the use of cameras for real-time traffic density estimation. The results showed a significant increase in terms of MobileNet-SSD being faster but SSD achieved a higher accuracy of 92.97% compared to 79.30% of MobileNet-SSD.

A technique for enhancing the SSD for vision-based surrounding car detection was put forth by Chen et al. [31]. The goal of the model was to make autonomous driving systems superior. In order to make use of the SSD, the MobileNet v2 was chosen as the model's backbone as one of three improvements (i.e., a lightweight backbone network), followed by the use of an attention mechanism and, lastly, a deconvolution module. The proposed model gave improvements in autonomous driving systems.

Chen et al. [32] added an inception block to the SSD model for better vehicle detection. The algorithm achieved a higher mAP and also kept a fast speed of detection when compared with SSD using the KIITI and UVD datasets for verification. According to the experimental findings, the suggested method on the BDD100K and KITTI datasets achieved average precision of 82.59% and 84.83%, respectively, with faster inference times. When applied to various complicated traffic scenarios, the proposed model could identify vehicles automatically.

Using the YOLOv5 object detection algorithm as a foundation, Chen et al. [33] suggested a vehicle detection method based on high-resolution images taken by UAVs. This method addressed the issue of conventional object detection algorithms' limitations on image quality and object size. According to the experimental findings, their algorithm's precision improved from 79.5% to 91.9%, recall from 44.2% to 82.5%, and mAP@0.5 increased from 47.9% to 89.6% when compared to the original YOLO object detection algorithm.

Bilinear interpolation was used in the RoI pooling operation to address the positioning deviation brought about by the conventional method. Based on Faster R-CNN, an improved LIIOU loss function for positioning was proposed, which addresses the problem that the L1 norm and L2 norm loss functions cannot accurately reflect the overlap between the prediction region proposal and the ground truth. The outcomes demonstrated that the suggested algorithm performed well on traffic signals with resolutions between 0 and 32; its recall rate was 90% and its accuracy rate was 87%. [34]

The two current models for object detection and object counting of people were compared in Gupta et al. [35]. The SSD architecture and YOLOv3 were examined. Videos were used for object counting, while image datasets were used for object identification. Precision, recall, and F1 were the three parameters used in the studies. Precision, recall, and F1 measure average scores for SSD were 95, 84, and 90 respectively.

A modified YOLOv1-based neural network for object detection was suggested by Ahmad et al. [36]. The new network was trained on an end-to-end method. The extensive experiment on a difficult Pascal VOC dataset, 2007/2012, demonstrated the effectiveness of the improved new network, with the detection findings being 65.6% and 58.7% respectively.

[37] suggested using a feature expression enhanced SSD (ESSD) which extracts semantic information in a lightweight manner, in addition to fusion using detailed information and forming a new feature map to enhance the feature expression, to increase the accuracy of traffic sign detection [37]. The suggested ESSD puts more emphasis on the traffic indicator. The experimental results demonstrated an improvement in mAP of 81.26% and 90.52% with robustness on verification with PASCAL VOC dataset, which exhibited better detection for smaller objects, when compared with SSD by training both the proposed model and SSD with TT100K and CCTSDB datasets.

An SSD model with Receptive Field Module (RFM) and Path Aggregation Network (PAN) was suggested by Wu et al in an experiment using the GTSDB and CCTSDB dataset to help intelligent transportation systems in terms of detection accuracy and efficiency in traffic sign detection [38]. In comparison to other detection algorithms like Faster R-CNN, RetinaNet, and YOLOv3, the suggested model achieved a better balance between detection time and precision.

3.3 Medical

Johnson showed how to perform highly effective and efficient automatic segmentations of a wide range of microscopy images of cell nuclei, for a variety of cells acquired under a variety of conditions [39]. Masked-RCNN is a state-of-the-art model for instance segmentation developed on top of faster R-CNN. A backbone of the feature pyramid network was used with the masked-RCNN model. On a single NVIDIA Titan Xp GPU, a group size of six was also applied. A mAP as defined for the MS COCO challenge dataset of 59.40% was obtained in the mode with ResNet-101 as the backbone and the same parameters and training methods.

Lemay reviewed the effectiveness of the YOLOv3 object detection algorithm on kidney localization from CT scans in both 2D and 3D. Real-time object detection models like YOLOv3 or SSD demonstrated encouraging kidney detection findings, indicating that they could recognize organs well even with a small number of CT scans as their training set [40]. The outcomes showed that the model could apply to a variety of kidney morphologies.

The use of R-CNN for detecting ROI only using a limited number of labelled WSIs for training was examined by Nugaliyadde et al. [41]. The R-CNN was trained and tested using 60 WSIs and 12 WSIs, respectively. The R-CNN was able to acquire the features of the ROI well due to the use of the center patch. The candidate area proposal was created using selective search and greedy algorithm. Features were then extracted using VGG 16 that had been trained on ImageNet, and the classification was created using a softmax dense layer. Results demonstrated that the well-established R-CNN can be used to locate ROI on WSI, which could help pathologists find potential GC-containing areas within healthy and benign lymph nodes.

Brain CT and Magnetic Resonance Imaging (MRI) pictures were combined and classified as normal or abnormal using YOLOv2 [42]. Segmentation, feature extraction for texture and shape, and an SVM classifier are used in conventional brain tumor classification; these methods require a lot of computation time and intricacy.

Faster R-CNN was used by Rahmat et al for the categorization of chest x-ray images. CNN implemented a RPN to increase object recognition accuracy [43]. Once multiple layers had been combined into a single network (fast R-CNN), it was improved in terms of time execution by removing the region proposal extraction at an earlier level (i.e. faster R-CNN). The writers divided a subset of the chest x-ray14 dataset into training and testing sets, using 80% of the data for each. According to the findings, the accuracy was 62% higher than that of the medical representatives.

Faster R-CNN was used by Albinali et al to identify items in micrograph images that had been manually annotated [44]. Because of its speed and high level of detection accuracy, the faster R-CNN model was taken into consideration. The study's goals were to identify items in medical images and assess how well faster R-CNN performed in this regard. When no overlapping objects in the images were taken into account, faster R-CNN algorithm produced highly accurate findings in image detection. 65 micrograph images with about 1000 objects total were used in the experiment. The results showed that faster R-CNN produced highly accurate results with mAP 76% for non-overlapping objects.

In order to help radiologists identify pulmonary nodules more precisely, Cai et al suggested the methods of detection and segmentation for pulmonary nodule 3D visualization diagnosis using Masked R-CNN and ray-casting volume rendering algorithm [45]. To test and assess the suggested technique in their paper, experiments were conducted using the independently generated datasets from the Ali TianChi challenge and the publicly accessible LUNA16 dataset. At 1 and 4 false positives per scan, the suggested Masked R-CNN of weighted loss achieved sensitivities of 88.1% and 88.7% respectively. On the labelme LUNA16 dataset, a Masked R-CNN of weighted loss produced an AP@50 score of 0.882.

4 | RESULT AND DISCUSSION

From the research, it can be inferred that remarkable advances in object detection have been made over time. The reviewed object detectors demonstrate not only how the object detection process has been improved, but also their various applications and advantages when used in the particular area covered in this paper.

During this survey, more than 30 papers were consulted, and it was found from the analysis shown in Table 1 that, among the various applications and

improvements on object detection models, YOLO models are the most popular algorithm for drone surveillance due to its superior performance in that area of use, SSD takes the lead for object detection and application in traffic, and faster R-CNN is the most popular in the medical industry.

Table 1. Results on Application of Object Detection Algorithm

Detectors	Algorithm	Drone Surveillance	Traffic	Medical
Two-stage Detectors	Masked R-CNN	1		3
	Fast R-CNN		1	
	Faster R-CNN	1	3	4
One-stage Detectors	SSD	2	7	
	YOLO	4	3	2

The YOLO models due to its performance in speed is mostly preferred than other models but in terms of accuracy, the two-stage detectors such as the faster R-CNN are widely used.

In real-time detection scenarios, where speed and efficiency are crucial, one-stage detectors have an advantage. These detectors directly predict the bounding boxes and class labels in a single pass through the network. By eliminating the need for a separate region proposal step, they can achieve faster inference times.

One-stage detectors, such as YOLO or SSD are known for their real-time capabilities. They trade off some accuracy for speed by making coarse predictions at multiple scales. While they may not achieve the highest possible accuracy compared to two-stage detectors, they still provide satisfactory results in many applications.

If achieving the highest possible accuracy on the other hand, is of utmost importance, two-stage detectors offer better performance. These detectors typically consist of RPN followed by a classification network. The RPN generates potential object proposals which are then refined and classified by the subsequent stages. A comparison of the one-stage and two-stage detectors application is shown in Figure 7 below.

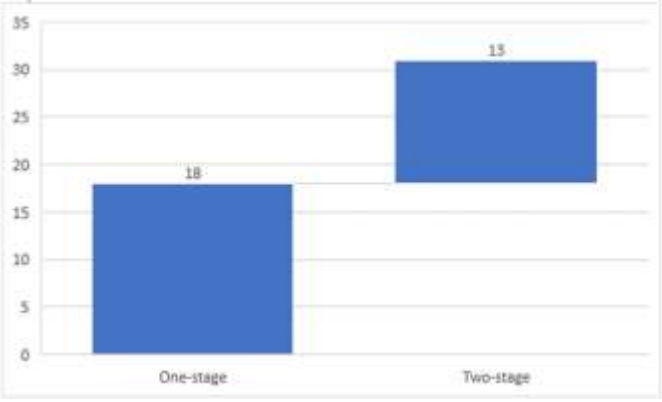

Figure. 7. One-stage Detectors and Two-stage Detectors Applications.

Two-stage detectors like Faster R-CNN have gained popularity in the field of object detection due to their ability to achieve high accuracy. By using a region RPN followed by a classification network, these detectors can generate potential object proposals and then refine and classify them in subsequent stages. This multi-stage approach allows for more precise localization and classification of objects, resulting in improved performance compared to single-stage detectors. The additional computational complexity of two-stage detectors is often justified by the significant boost in accuracy they provide, making them a preferred choice in applications where accuracy is paramount.

The analysis conducted on Figure 7, based on the data presented in Table 1, indicates a higher prevalence of one-stage detectors compared to two-stage detectors in multiple domains of object detection.

5 | CONCLUSION

This analysis of object detection algorithms based on the consulted papers revealed interesting trends in their application areas. YOLO models emerged as the most popular algorithm for drone surveillance. This is likely due to YOLO's real-time processing capabilities, which are crucial for tracking objects in dynamic aerial environments.

On the other hand, SSD models gained popularity in applications involving traffic. The ability of SSD to efficiently detect objects at different scales makes it well-suited for analyzing traffic scenes with varying distances and sizes of vehicles. Its speed and accuracy make it a preferred choice for monitoring road conditions and managing traffic flow.

In the medical field, Faster R-CNN stood out as the most popular algorithm for object detection. This is not surprising considering that medical imaging often requires precise localization and identification of abnormalities or anatomical structures. Faster R-CNN's two-stage architecture, with its region proposal network, enables accurate detection and classification of objects within medical images.

Interestingly, one-stage detectors were found to be more frequently used across various applications However, the preference for one-stage detectors in various applications does not diminish the significance of Faster R-CNN in the medical field. The precise localization and identification of abnormalities or anatomical structures in medical imaging demand a higher level of accuracy, which Faster R-CNN provides. Its two-stage architecture, although slightly more complex, offers superior performance in terms of object detection and classification within medical images.

Furthermore, the use of Faster R-CNN in the medical field extends beyond just object detection and

classification. Its two-stage architecture allows for the incorporation of additional modules, such as instance segmentation and key point detection, which are crucial for more advanced medical imaging tasks. These additional capabilities enable the precise delineation of boundaries and the identification of specific points of interest within medical images, aiding in the diagnosis and treatment planning processes. Therefore, while one-stage detectors may be more commonly used in various applications, the unique advantages of Faster R-CNN make it an indispensable tool in the field of medical imaging.

The performance metrics considered in this paper are speed, accuracy, efficiency and memory usage to the author's choice of models for various applications. The choice between the models depends on the specific requirements of the task at hand. If speed and efficiency are a priority, one-stage detectors might be more suitable. However, if achieving the highest possible accuracy is more important, two-stage detectors would be a better choice.

In resource-limited settings or when working with extensive datasets, the authors might have taken memory usage into account as a measure. Assessing memory usage allows the authors to gauge the model's capacity to handle and process data effectively without overburdening the system.

Speed is an essential metric as it directly impacts real-time applications. For tasks that require quick responses, such as autonomous driving or video surveillance, models with high speed capabilities are preferred. These models can rapidly process incoming data and make prompt decisions, ensuring timely actions.

Accuracy plays a vital role in applications where precision is of utmost importance. Tasks like object detection in medical imaging or facial recognition demand high accuracy to avoid misdiagnosis or false identifications. Two-stage detectors often excel in accuracy as they employ complex architectures and multi-step processes to refine predictions.

Efficiency is another crucial aspect that influences the overall performance of a model. It encompasses both speed and accuracy while considering resource utilization. Efficient models strike a balance between these factors, delivering satisfactory results without sacrificing too much computational power or sacrificing too much accuracy.

More papers and detection algorithms will be used to further enhance general detection processes and performance as more research and development are done in various application areas.

Acknowledgement

Funding Statement
The authors received no specific funding for this study.

Authors Contributions

The authors confirm contribution to the paper as follows: study conception and design: O.E. Olorunshola; analysis and interpretation of result: A.K. Ademuwagun; draft manuscript preparation: C. Dyaji. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials

Data available on request from the authors.
The data that support the findings of this study are available from the corresponding author, O. E. Olorunshola, upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055v2*.
2. Wu, X., Sahoo, D., & Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing*, *396*, 39-64.
3. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, *128*(2), 261-318.
4. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587).
5. Raj, A., Namboodiri, V. P., & Tuytelaars, T. (2015). Subspace alignment based domain adaptation for Rcnn detector. *arXiv preprint arXiv:1507.05578*.
6. Gkioxari, G., Girshick, R., & Malik, J. (2015). Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1080-1088).
7. Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1440-1448).
8. Zhao, Z. Q., Bian, H., Hu, D., Cheng, W., & Glotin, H. (2017, August). Pedestrian detection based on fast R-CNN and batch normalization. In *International Conference on Intelligent Computing* (pp. 735-746). Springer, Cham.
9. Li, X., Shang, M., Qin, H., & Chen, L. (2015, October). Fast accurate fish detection and recognition of underwater images with fast r-cnn. In *OCEANS 2015-MTS/IEEE Washington* (pp. 1-5). IEEE.
10. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

11. Li, J., Zhang, D., Zhang, J., Zhang, J., Li, T., Xia, Y., ... & Xun, L. (2017). Facial expression recognition with faster R-CNN. *Procedia Computer Science*, *107*, 135-140.
12. Pham, M. T., & Lefèvre, S. (2018, July). Buried object detection from B-scan ground penetrating radar data using Faster-RCNN. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 6804-6807). IEEE.
13. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).
14. Redmon, J., & Farhadi, A. (2017). Yolo9000: better, faster, stronger. *arXiv preprint*.
15. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
16. Shinde, S., Kothari, A., & Gupta, V. (2018). YOLO based human action recognition and localization. *Procedia computer science*, *133*, 831-838.
17. He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European conference on computer vision.* (pp. 346-361) Springer Publication.
18. Lin, T.-Y., Dollar, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. In *CVPR*, 1(2), p. 4.
19. Vandersteegen, M., Van Beeck, K., & Goedemé, T. (2019, May). Super accurate low latency object detection on a surveillance UAV. In *2019 16th International Conference on Machine Vision Applications (MVA)* (pp. 1-6). IEEE.
20. Hong, T., Yang, Q., Wang, P., Zhang, J., Sun, W., Tao, L., ... & Cao, J. (2021). Multitarget Real-Time Tracking Algorithm for UAV IoT. *Wireless Communications and Mobile Computing*, *2021*.
21. Huang, W., Zhou, X., Dong, M., & Xu, H. (2021). Multiple objects tracking in the UAV system based on hierarchical deep high-resolution network. *Multimedia Tools and Applications*, *80*(9), 13911-13929.
22. Rohan, A., Rabah, M., & Kim, S. H. (2019). Convolutional neural network-based real-time object detection and tracking for parrot AR drone 2. *IEEE access*, *7*, 69575-69584.
23. Jain, S., Singh, A., Shah, S. N., Lalam, R., & Saxena, D. (2021, October). Machine Learning-Based Real-Time Traffic Control System. In *2021 IEEE Mysore Sub Section International Conference (MysuruCon)* (pp. 92-97). IEEE.
24. Sien, J. P. T., Lim, K. H., & Au, P. I. (2019, April). Deep learning in gait recognition for drone surveillance system. In *IOP Conference Series: Materials Science and Engineering* (Vol. 495, No. 1, p. 012031). IOP Publishing.
25. Singha, S., & Aydin, B. (2021). Automated Drone Detection Using YOLOv4. *Drones*, *5*(3), 95.
26. Hosang, J., Omran, M., Benenson, R., & Schiele, B. (2015). Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4073–4082).
27. Zhang, L., Lin, L., Liang, X., & He, K. (2016). Is faster rcnn doing well for pedestrian detection? In European Conference on Computer Vision. Springer, 2016, (pp. 443–457).
28. Vijayaraghavan, V. & Laavanya, M. (2019). Vehicle Classification and Detection using Deep Learning. *International Journal of Engineering and Advanced Technology (IJEAT), 9(*155*)*.
29. You, S., Bi, Q., Ji, Y., Liu, S., Feng, Y. & Wu, F. (2020). Traffic Sign Detection Method Based on Improved SSD. *Information, 11*(10), (p. 475) DOI: 10.3390/info1100475
30. Biswas, D., Su, H., Wang, C., Stevanovic, A. & Wang, W. (2019). An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD. *Physics and Chemistry of the Earth, Parts A/B/C, 110*, (pp. 176-184).
31. Chen, Z., Guo, H., Yang, J., Jian, H., Feng, Z., Chen, L. & Gao, T. (2022a). Fast vehicle detection algorithm in traffic scene based on improved SSD. *Measurement, 201*(7), 111655.
32. Chen, W., Qiao, Y., & Li, Y. (2022b). Inception-SSD: An Improved Single Shot Detector for Vehicle Detection. *Journal of Ambient Intelligence and Humanized Computing*, 13 (pp. 5047-5053).
33. Chen, Z., Cao, L., & Wang, Q. (2022c). YOLOv5-Based Vehicle Detection Method for High-Resolution UAV Images. *Mobile Information Systems*.
34. Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., ... & Wu, Z. (2019). An improved faster R-CNN for small object detection. *IEEE Access*, *7*, 106838-106846.
35. Gupta, P., Sharma, V., & Varma, S. (2021). People detection and counting using YOLOv3 and SSD models. *Materials Today: Proceedings*.
36. Ahmad, T., Ma, Y., Yahya, M., Ahmad, B., & Nazir, S. (2020). Object detection through modified YOLO neural network. *Scientific Programming*s
37. Wu, J. & Liao, S. (2022). Traffic Sign Detection Based on SSD Combined with Receptive Field Module and Path Aggregation Network. *Interpretable Methods of Artificial Intelligence Algorithms, 2022*.
38. Johnson, J. W. (2018). Adapting Mask-RCNN for Automatic Nucleus Segmentation. *arXiv preprint arXiv:1805.00500*.
39. Lemay, A. (2019). Kidney Recognition in CT using YOLOv3. *arXiv preprint arXiv:1910.01268*.
40. Nugaliyadde, A., Wong, K. W., Parry, J., Sohel, F., Laga, H., Somaratne, U. V.,... & Foster, O. (2020, November). RCNN for Region of Interest Detection

in Whole Slide Images. In *International Conference on Neural Information Processing* (pp. 625-632). Springer, Cham.
41. Sabbava, S. R., Boddapati, S. P., Dasari, T., & Bollam, N. (2022). CNN based multi modal medical image fusion with classification using YOLO-V2
42. Rahmat, T., Ismail, A., & Aliman, S. (2019). Chest X-ray image classification using faster R-CNN. *Malaysian Journal of Computing (MJoC)*, *4*(1), 225-236.
43. Albinali, H., & Alzahrani, F. A. (2021, March). Faster R-CNN for detecting regions in human-annotated micrograph images. In *2021 International Conference of Women in Data Science at TAIF University (WiDSTaif)* (pp. 1-6). IEEE.
44. Cai, L., Long, T., Dai, Y., & Huang, Y. (2020). Mask R-CNN-based detection and segmentation for pulmonary nodule 3D visualization diagnosis. *IEEE Access*, *8*, 44400-44409.