BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI

STUDY-ORIENTED PROJECT

A REPORT

On

OBJECT DETECTION USING ML MODELS

Submitted To

Dr. L Rajya Lakshmi

Submitted By:

Archi Jain

2020B1A71380P

TABLE OF CONTENTS

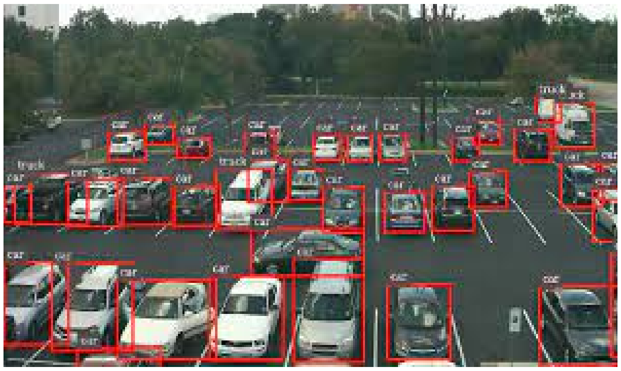
INTRODUCTION

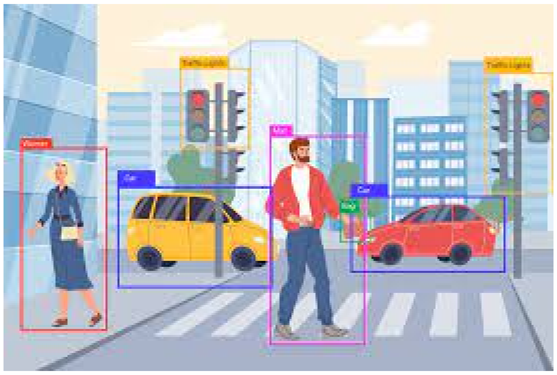
A crucial problem in computer vision is object detection, which involves locating and identifying things inside an image or a video frame. It is a fundamental part of many applications, including augmented reality, medical imaging, surveillance systems, and autonomous vehicles. Object detection has seen a revolution with the introduction of machine learning, especially deep learning, which has allowed for impressive improvements in efficiency and accuracy.

Classical methods in object detection are as follows: Hough transform method, frame-difference method, background subtraction method, optical flow method, sliding window model method, and deformable part model method. Conventional methods for object detection frequently needed help to handle changes in object appearance, scale, and occlusion because they mostly depended on manually created characteristics and intricate algorithms. However, object detection has seen a paradigm shift with the advent of deep learning techniques, especially Convolutional Neural Networks (CNNs). CNN-based models are skilled at handling challenging visual tasks because they can automatically learn discriminative features from raw pixel input.

As object detection techniques have developed, a variety of architectures, such as region-based and single-shot detectors, have emerged. Each has advantages and disadvantages of its own. YOLO (You Only Look Once), SSD (Single Shot Multibox Detector), and Faster R-CNN are a few notable architectures that have pushed the limits of object recognition performance, attaining amazing accuracy and real-time processing capabilities.

We explore the theories, practices, and developments in object detection with machine learning models in this study in the healthcare sector. We examine the underlying methods, assess the effectiveness of well-known models, and talk about the difficulties that the area is currently facing as well as potential future developments.





CLASSICAL METHODS

- **Hough Transform:** It can convert parameter space into image space. Each pixel in the picture space corresponds to a curve in the parameter space, and the parameters of a curve in the image space are the coordinates of the majority of curves' intersections determined by voting in the parameter space. Only object detection tasks where the object shape can be described using an analytical function—such as roundness, straightness, etc.—may utilize the common rough transform.

- **Frame difference method:** The idea is to subtract the two neighboring frame images to create the difference image, which is then denoised via binarization processing and morphological filtering to obtain the object motion area.

- **Background subtraction method:** The background subtraction method works similarly to the frame-difference approach; the only distinction is that the former requires the definition of a background frame and frequent updates. In order to define the background frame, background modeling technologies such as Mixture of Gaussians Models (GMM) and Local Binary Patterns (LBP) are typically combined with picture data, which are typically brightness, texture, and spatial information.

- **Optical flow method: T**he method assumes that the gray change is merely related to the object's motion and delineates the motion of the image pixel by establishing an optical flow equation, thereby delineating the motion of an object.

- **Sliding window model:** Fixed-size sliding windows are used to collect characteristics from images and slide them on the image in accordance with predetermined strategies before being classified by a classifier. Features in the model have the option of selecting a color histogram, gradient histogram, or shift. Additionally, SVM and the Adaboost classifier are options for the classifier.

- **Deformable Part Model (DPM) method:** The primary model and the sub-model are both included. A sub-model is used for local feature extraction, whereas the main model is used for global feature extraction. The sub-model separates the object into multiple segments and extracts features from each segment. In the meantime, the confidence in the deformation is described by a cost function of position offset between the main model and sub-model. The sliding window model is also used by the object detection approach based on the deformable component model to extract features and categorize them.

MODELS BASED ON REGION PROPOSAL

- **R-CNN:**
  - The principle of R-CNN is that it utilizes the region segmentation method of selective search to extract the region proposals in the image.
  - The feature vectors will be classified using the classifier SVM to determine the classification outcomes for each region proposal.
  - The model outputs exact object classifications and object bounding boxes.

- **SPP-net:**
  - It is a deep neural network based on spatial pyramid pooling.
  - The crop/warp procedure on the input image in the previous method can be eliminated by using the spatial pyramid pooling layer.
  - Additionally, it allows input images of various sizes to flow through the convolution layer and link to the full connection layer using a feature vector of the same dimension.
  - Solves the problems of object image incompleteness and deformation.

- **Fast R-CNN:**
  - Unlike R-CNN, Fast R-CNN extracts the region proposal from the input image using a selective search technique and then uses ROI pooling to execute pooling on the mapped region proposal of the feature layer.
  - The role of ROI pooling is just like the spatial pyramid pooling of SPP-net.

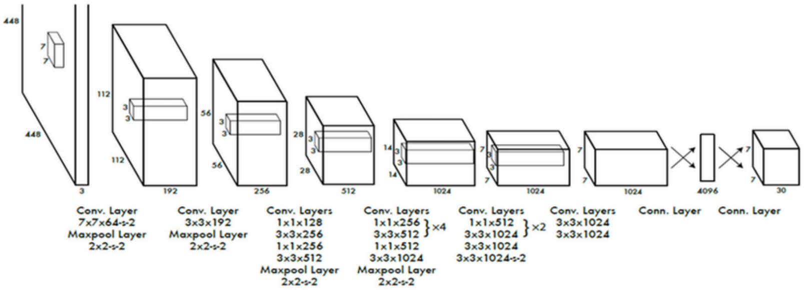
- **Faster R-CNN:**
  - The region proposal network (RPN) is utilized by Faster R-CNN to address the problems of excessive processing and subpar real-time resulting from the selective search technique employed by R-CNN and Fast R-CNN.
  - The technique known as the anchoring mechanism allows RPN to split the feature layer into n×n areas and produce feature regions that are centered on the region at different scales and aspect ratios.

- **R-FCN:**
  - Is a full convolution neural network based on regions, having solved the problem that RoI can’t share the computation.
  - With the position-sensitive score maps (k*k*(C+1) dimensional Convolutional Network).
  - R-FCN could do the recognition and location simultaneously to achieve object detection.

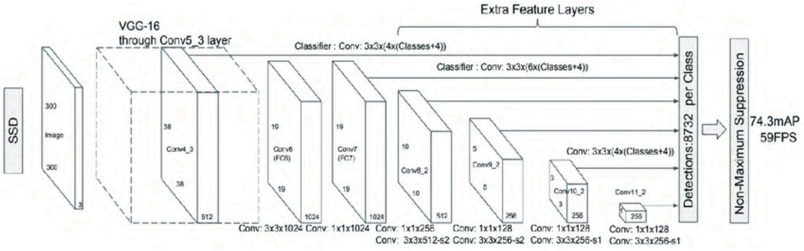
MODELS BASED ON REGRESSION

- **YOLO:**
  - Is a convolution neural network for real-time object detection and can accomplish end-to-end training.
  - Because of the cancellation of the RoI module, YOLO won't extract the object region proposal anymore.
  - YOLO divides the input image scale into 7*7 grids, each of which will produce two bounding boxes.
  - To obtain the detection results, YOLO sets a threshold, filters the object proposals with low confidence, and removes the redundant object proposals.

- **SSD:**
  - The design of SSD has integrated YOLO's regression idea and Faster R-CNN's anchors mechanism.
  - With the anchors mechanism, SSD can extract the features of different scales and aspect ratios to guarantee detection accuracy.
  - The local feature extraction method of SSD is more reasonable and effective compared with the general feature extraction method of YOLO.
  - One of the disadvantages is its weak detection capacity for small objects.



YOLO



SSD

Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism

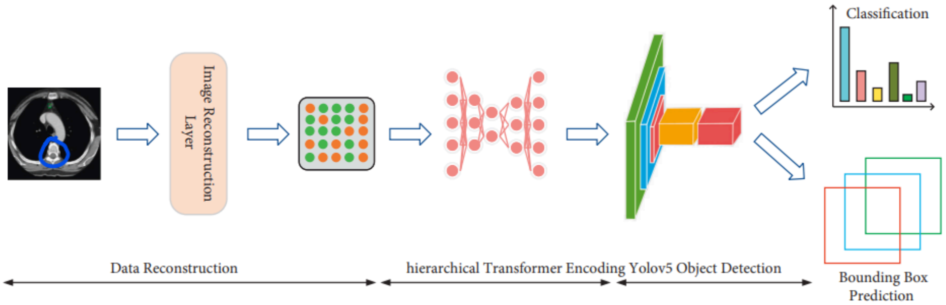
INTRODUCTION: One major issue in the field of object detection is still making the machine filter out most of the background information and reliably identify the small lesions in the images since the objects to be recognized in medical images are small. This paper proposes the MS Transformer framework model to address the above problems. To make the model's detection efficiency higher, we use the YOLOv5 single-stage object detection head to complete the bounding box regression task and the object recognition classification task.

DATASET USED:

- DeepLesion, the world's largest dataset of CT medical images thus far, was developed by the NIHCC team.
- On 32,120 axial slices, the dataset contains 10,594 CT studies from 4,427 different patients, and it contains 32,735 lesion annotations.
- The BCDD dataset is a benchmark on blood cells containing 4,888 blood cell images.
- The BCDD dataset has three categories: WBC, RBC, and Platelets.

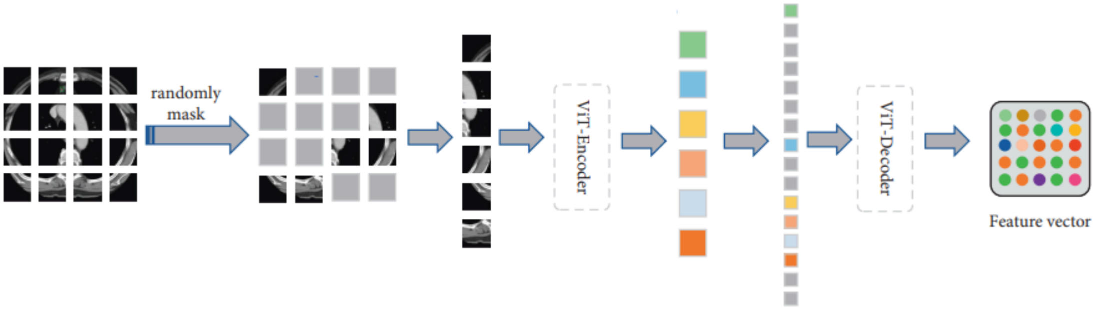
METHODOLOGY:

- Framework consists of a mask self-supervised pretraining model, a hierarchical Transformer model, and a single-target detection head YOLOv5.
- First, this paper divides the input image into multiple regular patches and performs mask operations on some of the patches.
- Secondly, we encode the unmasked patches on this basis to obtain the potential distribution of image features.
- Then we input the latent feature vector obtained after encoding together with the feature vector without mask operation into the decoder for self-supervised learning to reconstruct the missing pixels.
- Finally, to improve the detection efficiency of the model for medical images, we input the feature vector with attention weight into the YOLOv5 single-object detection head, and the regression and classification tasks are performed for the bounding box to be predicted.



COMPONENTS:

1. **Image Reconstruction Layer -** A self-supervised learning technique based on the masking mechanism is used to rebuild images. An autoencoder is used to recover the original signal.
2. **Masking -** each medical image is segmented into regular patches, and then, we sample them randomly and mask them.



3. **Encoder -** The encoder used in this work has the same architecture as ViT, and we feed the encoder the unmasked patches to be encoded. We embed the location vectors corresponding to the feature vectors into the associated patches and process them through a series of transformer blocks to reflect the position differences between the various feature vectors.
4. **Decoder -** A feature vector that can be trained to predict and reconstruct the missing pixels is represented by each masked token in the image. To reflect their position information in the image, we embed the position vector of the entire ensemble sequence—just like the encoder does—into the entire token ensemble sequence. The feature vectors from the rebuilt image are further processed by an additional set of transformer blocks.
5. **Self-Attention Mechanism -** We present the mechanism of self-attention. Because of the low resolution of the image, the model finds it challenging to learn relevant feature information. In medical photos, the items that need to be detected are typically contained in a tiny region and contain a lot of noise information in the image. Consequently, we apply greater weights to the most valuable semantic information in the input feature vector by using the attention method.
6. **Self-Attention in Local Windows -** To make the model focus its attention on the small objects to be detected, we design the windows in a way that the images are uniformly segmented and introduce a self-attention mechanism in the nonoverlapping sliding windows as a way to give a higher attention score to the small objects needing to be detected.

7. **YOLOv5 Architecture -** The feature vector that was extracted following processing by the mask self-supervised learning mechanism is fed into the backbone. A convolutional neural network with 32 convolution kernels, a 3 x 3 filter size, and a stride of 2 makes up the convolutional layer. Additionally, YOLOv5 provides an FPN + PAN framework. The PAN uploads semantic data for localization from bottom to top, whereas the FPN layer transmits multiscale semantic data from top to bottom. The bounding box coordinates and the classification result of the medical picture disease are output by the model in the prediction.

RESULTS:

- The MS Transformer performs significantly better than previous models on the benchmark dataset in terms of recognition accuracy for several categories.

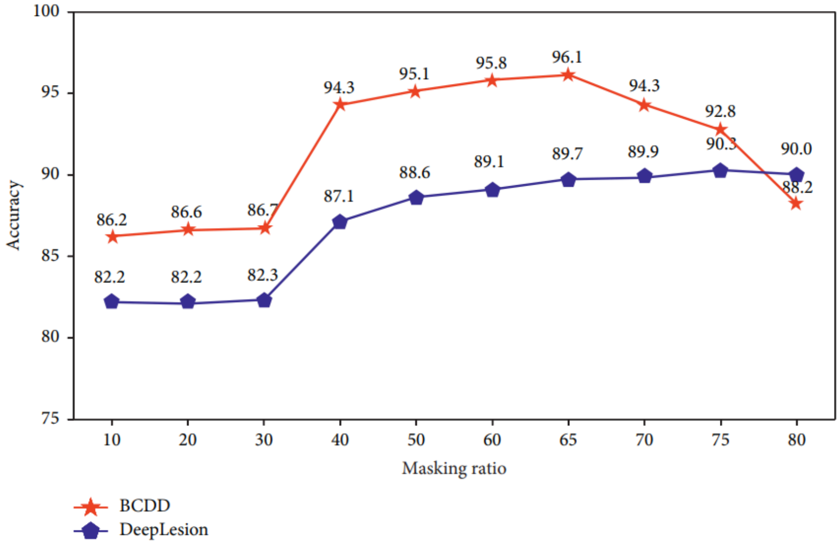


FIGURE 3: Relationship between mask rate and accuracy on BCDD and DeepLesion benchmark datasets. A high mask rate (65% 75%) can achieve better results.

CONCLUSION:

- In contrast to previous research, the suggested model gives a richer feature vector for the model and accounts for the poor resolution, excessive noise, and small items that must be detected in the medical area.
- To increase the model's capacity for generalization, we will take into consideration further research projects employing RPN networks to extract multiscale information from images.

Detecting multiple lesions of lung cancer-caused metastasis with bone scans using a self-defined object detection model based on the SSD framework

INTRODUCTION: One of the most commonly used clinical techniques for detecting bone metastases from a range of solid tumors, including lung cancer, is bone scintigraphy, also known as bone scans. The low specificity and poor resolution of 99mTc-MDP SPECT imaging (Nathan et al., 2013) greatly impede human manual examination of bone scan pictures for the detection of bone metastases. In the realm of automated medical image analysis, the automatic identification and localization of lesions by object detection algorithms is essential. Not only can a lesion's disease kind be identified automatically, but it may also be used to pinpoint the exact position of a lesion.

METHOD:

- An inputted 256 × 256 image is first convoluted using a 7 × 7 filter(7 × 7 Conv,c_out = 64) to produce feature maps. [c_out = channel number]
- Followed by a down-sampling using a 3 × 3 pooling layer(3 × 3 MaxPool, S = 1). [S = stride length]
- Lesion locations in photos (feature maps) are located using the lesion localization and boxing stage, which also involves boxing each area with a rectangle.

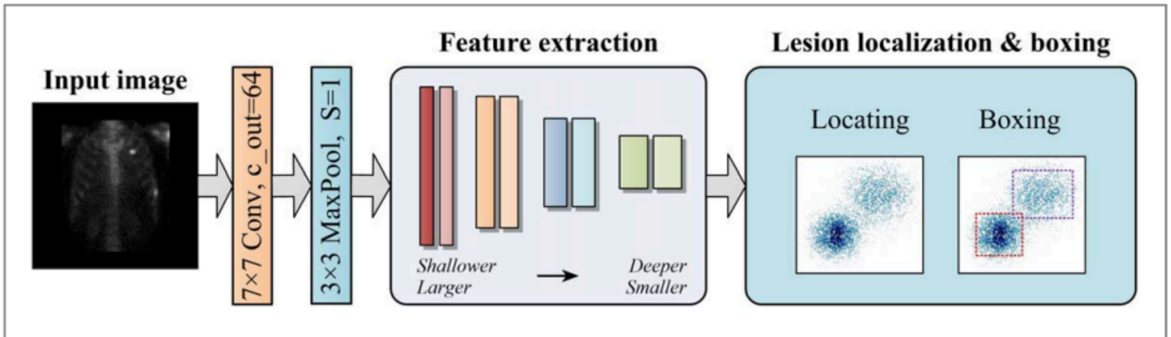


Figure 1. An overview of the proposed bone metastasis lesion detection method. (Conv = Convolution, MaxPool = Max pooling).

1. **Feature extraction -**
   - Four groups of convolution blocks are included in the defined feature extraction sub-network, with each block consisting of a 3 × 3 convolution layer(3 × 3 Conv,c_out) and a 1 × 1 convolution layer(1 × 1 Conv).
   - The number of blocks in these groups is indicated by {3, 3, 5, 3}.

- Two convolutional layers within a block have a residual connection (also known as an intra-res connection), while two convolutional layers of different blocks have a residual connection (also known as an inter-res connection).
- An outline of the suggested technique for detecting bone metastase lesions is shown in Figure. (MaxPool = Max pooling, Conv = Convolution).

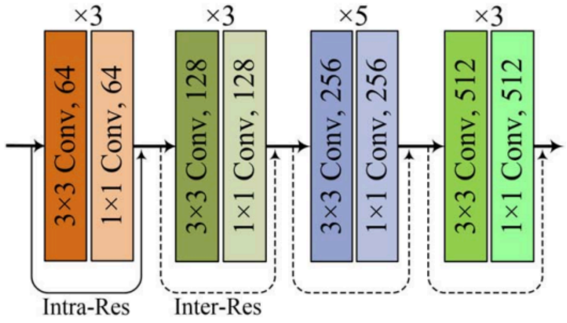


Figure 2. The structure of the feature extraction sub-network consisting of blocks.

2. **Lesion localization and boxing -**
    - Output is a group of varied-size feature maps of {32 × 32, 16 × 16, 13 × 13, 11 × 11, 9 × 9, 7 × 7, 5 × 5, 3 × 3, 1 × 1}.
    - With these feature maps, a two-stage operation consisting of candidate box (CB) generation and valid candidate box (VCB)selection is conducted to locate and box each lesion area in an image.
    - A CB is called a positive sample if it partially or fully covers a real lesion; it is a negative sample otherwise.
    - A positive CB will be selected as a VCB if it has IoU > θ (strong positive sample). [IoU measures the overlap between this CB and its ground truth]

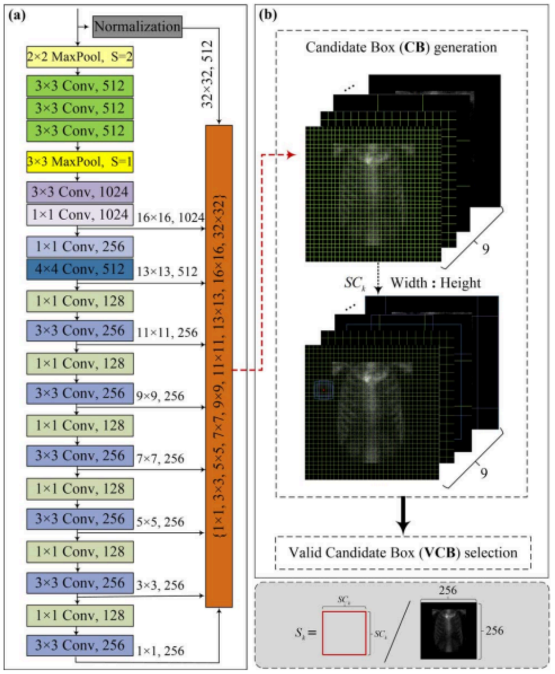


Figure 4. The overview flowchart of locating and boxing lesion areas.

EVALUATION METRICS AND RESULTS:

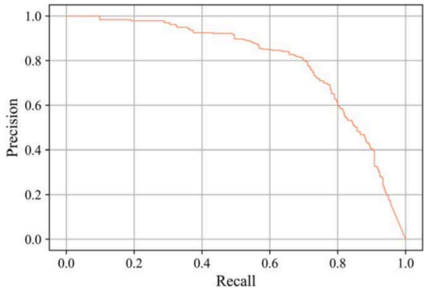


Figure 5. The *P–R* curve of the proposed model on the test set.

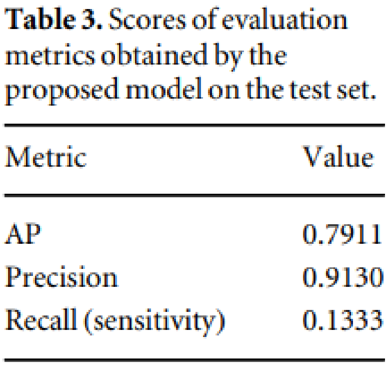
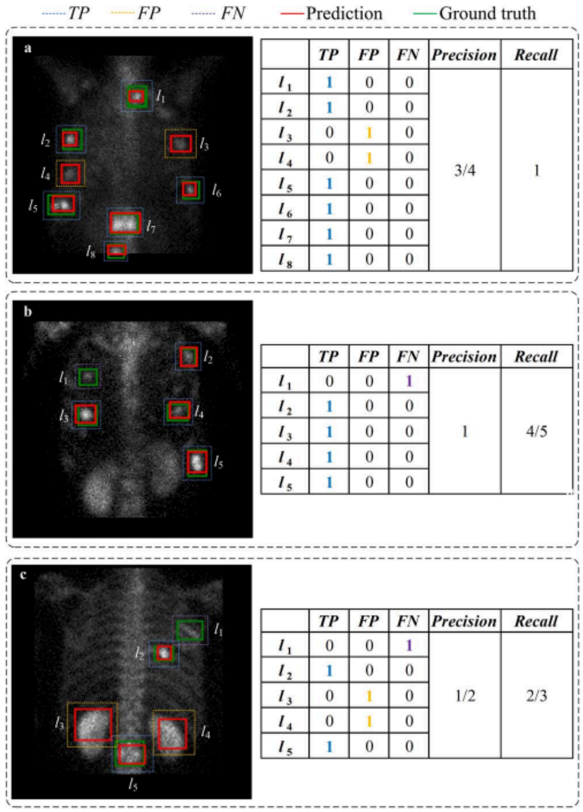


Table 3. Scores of evaluation metrics obtained by the proposed model on the test set.

Metric	Value
AP	0.7911
Precision	0.9130
Recall (sensitivity)	0.1333



	TP	*FP*	*FN*	*Precision*	*Recall*
l_1	1	0	0	3/4	1
l_2	1	0	0		
l_3	0	1	0		
l_4	0	1	0		
l_5	1	0	0		
l_6	1	0	0		
l_7	1	0	0		
l_8	1	0	0		

	TP	*FP*	*FN*	*Precision*	*Recall*
l_1	0	0	1	1	4/5
l_2	1	0	0		
l_3	1	0	0		
l_4	1	0	0		
l_5	1	0	0		

	TP	*FP*	*FN*	*Precision*	*Recall*
l_1	0	0	1	1/2	2/3
l_2	1	0	0		
l_3	0	1	0		
l_4	0	1	0		
l_5	1	0	0		

Figure 9. Illustration of the detected multiple lesions in three images by the proposed model with the prediction and ground truth being labeled with red box and green box, respectively. (a) The case contains false-positive predictions in a posterior-view image; (b) The case contains false-negative prediction in an anterior-view image; and (c) The case contains false-positive predictions in the organs (i.e. kidneys) in a posterior-view image.

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = Sensitivity = \frac{TP}{TP + FN},$$

- Fortunately, the composite metric AP that measures the area under the Precision–Recall (P–R) curve obtains a relatively high score, which is depicted in the figure.
- A high score of Precision reveals that the proposed detection model can successfully identify true positives while suppressing false positives.
- The model obtains a low score for Recall, which is contributed by the high false negatives.

Cervical Cancer Diagnostics Healthcare System Using Hybrid Object Detection Adversarial Networks

INTRODUCTION: One of the most frequent cancers in women is cervical cancer, which has a high death rate in many underdeveloped nations. Acetic acid staining or visual inspection are the two methods used to diagnose cervical lesions. However, the majority of approaches only take into account the segmentation and labeling of cervical spots. To address cervical screening, cervical cancer detection, and cancer type utilizing digital colposcopy pictures, this research seeks to introduce the Faster Small-Object Detection Neural Networks (FSOD-GAN). To identify cervical cancer and predict the different stages of the disease using colposcopy images, this research presents FSOD-GAN. The proposed FSOD-GAN will automatically perform localization of cervical spots without manual intervention and multiclass classification of cervical malignant conditions based on fine-tuned deep features.

ARCHITECTURE:

1. **Data Collection and Denoising -**
   - The images were gathered into two categories comprising normal and abnormal cervical images.
   - 3105 colposcopy images were collected from heterogeneous sources.
   - Of the 3105 cervical images, 1993 fall into the normal categories, which are composed of three types, and 1112 fall into the pathological classes, which are composed of three stages.
   - The performance of standard stacked autoencoders (SAE) was enhanced by the introduction of stacked denoising autoencoders (SDAE) [26]. Stacking layers of denoising autoencoders are used by the SDAE to transform corrupted inputs into uncorrupted ones.

2. **Pre-Processing and Data Augmentation -**
   - Enhancing the gathered photos to prevent overfitting brought on by the subsequent training stage's smaller sample size.
   - Data augmentation processes, including rotation, flip, shift, and zoom, have been done.
   - Input images were augmented and pre-processed by resizing and cropping to the width and height of 100 × 100 and recoloring the grayscale color channel.

3. FSOD - GAN -

- After pre-processed and denoised from FSDAE, the input images are fed into the GAN generator and discriminator.
- The discriminator will receive the fake images produced by the generator with random noises in addition to the actual input images from FSDAE.
- The classifier CNN will use the bounding boxes to identify which image classes are normal and abnormal.
- The GAN generator is added on top of the FR-CNN convolution layer to produce false images.
- To distinguish between generated and original images, a discriminator is added to the FR-CNN's dense layer.
- Additionally, the class labels and cervical location are extracted from the FR-CNN's fully connected layer as outputs.
- The generator will generate the bias, non-zero values for normal images, and zero values for fake images
- The capacity of the generator to produce fictitious images that resemble actual ones via parameter learning is used to gauge its performance.

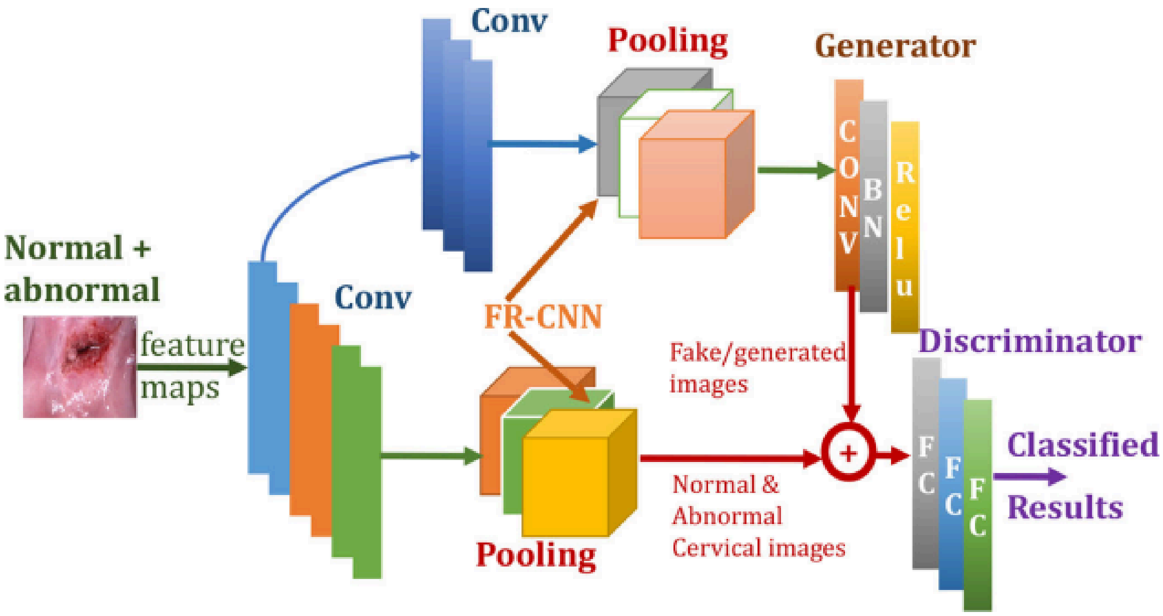
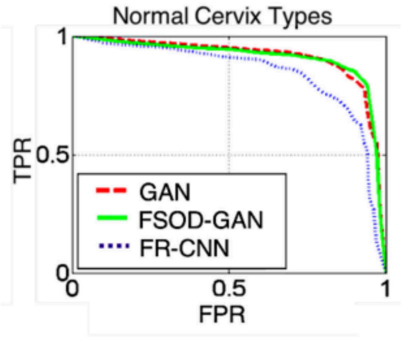


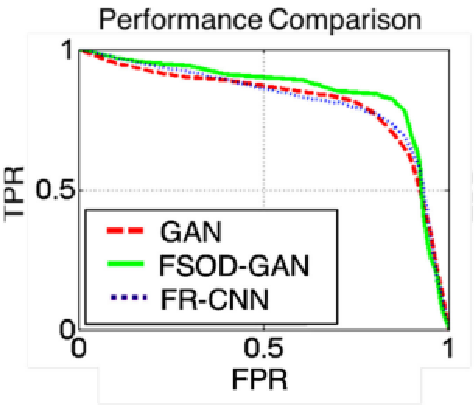
Fig. 5. Faster R-CNN integrated GAN Architecture.

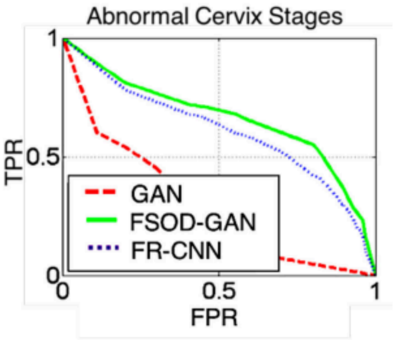
EVALUATION AND RESULTS:

- Evaluation criteria such as accuracy, specificity, sensitivity, precision, recall, F1 score, true positive (TP) rate, true negative (TN), false positive (FP), and false-negative (FN) rate are computed for the proposed FSOD-GAN.



(b)





(c)

CONCLUSION: The suggested FSOD-GAN architecture is the first to categorize cervical images as normal or abnormal, as well as the find and stage of infection. It does this by hierarchical multiclass classification. Additionally, experimental results demonstrated that the suggested FSOD-GAN performs better in cervical cancer screening and diagnosis than other cutting-edge methods. It is now established that the suggested FSOD-GAN can be used in real-time scenarios for the purpose of employing colposcopy pictures for cervical cancer screening, diagnosis, and prognosis.

REFERENCES

1. Elakkiya, R., Subramaniyaswamy, V., Vijayakumar, V., & Mahanti, A. (2021). Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, *26*(4), 1464-1471.

2. Lin, Q., Chen, X., Liu, L., Cao, Y., Man, Z., Zeng, X., & Huang, X. (2022). Detecting multiple lesions of lung cancer-caused metastasis with bone scans using a self-defined object detection model based on SSD framework. *Physics in Medicine & Biology*, *67*(22), 225009.

3. Yuntao Shou, Tao Meng, Wei Ai, Canhao Xie, Haiyan Liu, Yina Wang, "Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 5863782, 12 pages, 2022. https://doi.org/10.1155/2022/5863782

4. Tang, C., Feng, Y., Yang, X., Zheng, C., & Zhou, Y. (2017, July). The object detection based on deep learning. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)* (pp. 723-728). IEEE.