A Real-time Automated System for Object Detection and Facial Recognition

K. Shyam Sunder Reddy[1], *G. Ramesh*[2]*, *J. Praveen*[3], *P. Surekha*[2], *Ayushi Sharma*[4]

[1]Department of CSE, Maturi Venkata Subba Rao (MVSR) Engineering College, Hyderabad, India

[2]Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

[3]Department of Electrical and Electronics Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
[4]Uttaranchal Institute of Management, Uttaranchal University, Dehradun, India.

Abstract. Object detection, facial recognition, and person identification are important tasks in computer vision with numerous real-life applications. The major goal of the proposed model is to identify people and recognize them in the images. In this paper, we propose a real-time automated system that combines power of both EfficientDet model for object detection and the FaceNet model for facial recognition to detect persons in an input image, recognize their faces, and label them with their corresponding names. The experimental study of the model takes place on COCO dataset and a custom dataset of images of students. This solution can be applied to various scenarios beyond education, such as in security and surveillance, healthcare, transportation, retail, and entertainment etc. The importance of the model lies in its ability to efficiently and accurately perform person identification and recognition in real-time scenarios, which can save time and resources and improve overall efficiency.

Keywords—Object Detection, Facial Recognition, EfficientDet FaceNet and COCO (Common Objects in Context).

1. Introduction

Over the years, there has been significant research and development focused on object detection and face recognition, which are both crucial tasks within the field of computer vision. Object detection is the task of identifying objects of interest in an image or video, while face recognition is the task of identifying individuals based on their facial features. These tasks have numerous real-world applications, including surveillance, security, and biometrics. The recent progress in deep learning has resulted in the creation of object detection and face recognition models that are characterized by their high accuracy and efficiency.

*Corresponding author: ramesh680@gmail.com

One such model is the EfficientDet, which is a state-of-the-art object detection model that achieves high accuracy and speed. On the other hand, the FaceNet model is capable of encoding facial features of two faces into high-dimensional vectors, also known as face embeddings. In order to determine if two entities belong to the same person, a common approach is to compare them using a distance metric such as Euclidean distance or cosine similarity. In other words, FaceNet can be used to match the facial features of two different faces and identify if they correspond to the same person.

In this paper, we propose a system that combines these two models to detect and recognize persons in an input image. The system would use the EfficientDet model to detect persons in the image and the FaceNet model to recognize the detected persons. To accomplish this task, we trained the FaceNet model on a custom dataset containing subdirectories named with the target person name and it consists of multiple images of the same person. The system would compare the face embeddings of the detected persons with the embeddings of the target persons in the custom dataset. If the face embedding matches with one of the target persons, the system would conclude that the detected person is the target person.

The proposed system has numerous potential applications, including surveillance, security, and biometrics. The system can be used in various settings, such as airports, train stations, and shopping malls, to detect and recognize persons of interest. The system can also be used in law enforcement to identify suspects or missing persons. In this paper, we provide a detailed explanation of the proposed system, including the implementation details, experimental results, and potential applications.

2. Literature Survey

S. Sivachandiran et al. [1] presented the DLD-APDT model to detect and track persons in the surveillance videos which uses EfficientDet object detection along with root mean squared propagation (RMSProp) optimizer to enhance the performance in comparison to other models. Tsung-Yi Lin et al. [2] introduced Focal Loss, an innovative loss function utilized in the training of object detection models, specifically targeting difficult examples for improved performance. This loss function aims to tackle the challenge of class imbalance in object detection tasks and has proven to achieve exceptional results on different object detection benchmarks.

Joseph et al. [3] introduced YOLOv3, the third version of the popular object detection model YOLO (You Only Look Once). YOLOv3 improves upon its predecessors by using a new backbone architecture, a multi-scale feature pyramid network, and a better object detection head. Kaiming et al. [4] proposed introduces Mask R-CNN, an enhanced version of the Faster R-CNN model, which incorporates an additional branch for predicting masks within the network. Mask R-CNN is capable of detecting objects and generating a binary mask for each object instance, enabling pixel-level segmentation of the objects in the image.

Jianping et al. [5] proposed CenterNet, a novel object detection approach that uses keypoint triplets instead of bounding boxes to localize objects. CenterNet stands out for its exceptional accuracy and efficiency as it directly predicts object center points, sizes, and class labels, enabling it to be a single-shot solution for object detection in diverse scenarios. Nicolas et al. [6] introduced DETR (DEtection TRansformer) , a new approach to object detection that uses a transformer-based architecture to perform both object detection and object tracking. This framework effectively captures extensive dependencies and spatial relationships in images, leading to enhanced capability in accurately.

Tan et al. [7] introduced EfficientDet, a new family of object detection models that achieve state-of-the-art performance on several benchmarks while being more efficient in terms of computational cost than previous approaches. The EfficientDet models use a compound scaling method that optimizes the trade-off between accuracy and efficiency by scaling the network architecture and input resolution. Ruoming et al. [8] presented EfficientDet-D8, a new variant of the EfficientDet family that uses 5x fewer floating-point operations (FLOPS). EfficientDet-D8 utilizes a synergistic combination of efficient network architecture design, refined training techniques, and advanced post-processing methods to achieve exceptional accuracy.

Tan et al. [9] introduced EfficientDet-D7, the largest and most powerful model in the EfficientDet family. EfficientDet-D7 is designed for ultra-large-scale object detection tasks, such as satellite and aerial imagery analysis. Jiankang et al. [10] proposed ArcFace, an innovative loss function designed for deep face recognition, which integrates an angular margin to enhance the model's discriminative power. By incorporating this angular margin, ArcFace surpasses previous loss functions like Softmax and Center Loss, delivering higher performance.

Iperov [11] introduced DeepFaceLab, a face swapping framework that uses deep learning to transfer facial features from one face to another. DeepFaceLab offers a wide range of functionalities, including face swapping, face reenactment, and facial expression transfer, all powered by deep neural networks. Xuanqing et al. [12] proposes a novel face recognition approach based on Generative Adversarial Networks (GANs) that is designed to work in unconstrained environments. GANs are specifically designed for generative modeling, where the generator network learns to generate realistic samples such as images or text, while the discriminator network learns to differentiate between real data and generated data.

Florian et al. [13] introduced FaceNet, a deep learning- based approach to face recognition that uses a triplet loss function to learn a compact and discriminative representation of faces. It exhibits superior resilience to variations in pose, lighting conditions, and facial expressions, making it a highly reliable and effective solution for face recognition tasks. Wenpeng et al. [14] proposed a new loss function, Large Margin Softmax Loss (L-Softmax), that improves the discriminative power of deep neural networks for face recognition. By modifying the Softmax loss function to enforce a larger margin between classes, L-Softmax enhances discriminative power and achieves remarkable performance when integrated with FaceNet.

Xiang et al. [15] proposed a new deep neural network architecture, Multi-Granularity Network (MGN), that learns discriminative features at multiple granularities for face recognition. MGN architecture enables the generation of robust feature representations, facilitating accurate matching of individuals across various camera views or time instances in person re-identification tasks.

In [16-25] the authors have been explored various machine learning techniques on visual communications and also employed machine learning techniques in their research for getting effective image feature extraction.

3. Design & Methodology

Automated systems for person detection and recognition have become increasingly important in various fields, including security, surveillance, and biometrics. These systems can assist in identifying individuals in crowded public places, monitoring and controlling access to secure areas, and tracking individuals for various purposes. However, these tasks are challenging due to the large number of factors that need to be considered, such as changes in lighting, poses, occlusions, and facial expressions.

To address these challenges, a multi-stage architecture can be employed that comprises multiple models or algorithms. This approach offers a significant advantage by enabling each stage to concentrate on a specific aspect of the task, thereby enhancing the system's accuracy and efficiency. The model block diagram is shown in Figure 1. In this paper, we present a 3-stage architecture for person detection and recognition that combines the strengths of three models: EfficientDet, MTCNN, and FaceNet. The 3 – stage architecture is as follows:

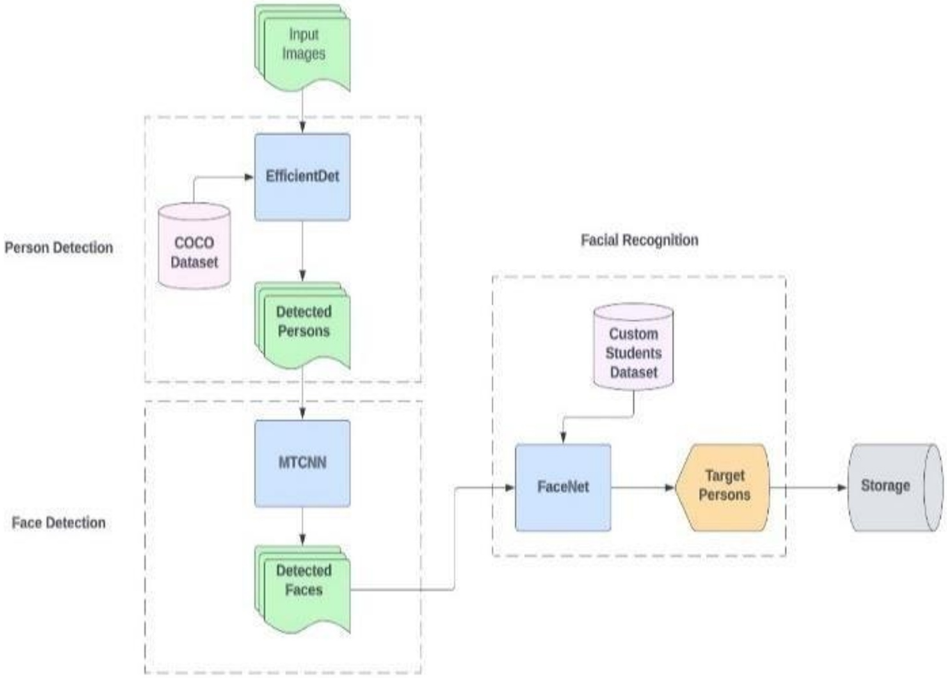


Fig 1. Model Block Diagram

3.1 Person Detection using EfficientDet

The first stage of our architecture involves using an EfficientDet model for person detection. EfficientDet is a state-of-the-art object detection model that achieves high accuracy and efficiency by employing a compound scaling method that optimizes the model architecture and input resolution. The model is pretrained on the COCO dataset, which contains a large number of annotated images of various objects, including people.

By taking an input image, the EfficientDet model generates a collection of bounding boxes that accurately indicate the object locations within the image. In our case, we are interested in detecting people, so we filter the bounding boxes to only retain those that correspond to people. This is done by applying a threshold on the confidence scores assigned to each bounding box by the model. Bounding boxes with confidence scores below the threshold are discarded.

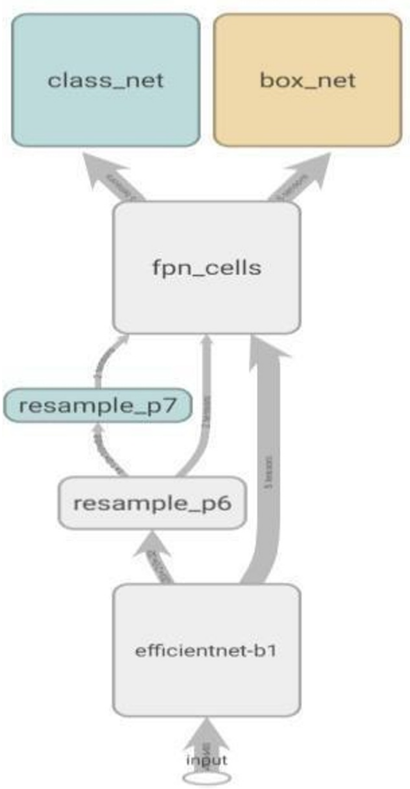


Fig: 2. TensorBoard graph for EfficientDet-D1

The EfficientDet architecture is based on a compound scaling method that optimizes the model architecture and input resolution. The model is composed of three main components: a backbone network, a feature network, and a box/class network. The responsibility of the backbone network is to extract features from the input image, usually accomplished using a convolutional neural network (CNN) such as EfficientNet.

After receiving the extracted features from the backbone network, the feature network applies a sequence of convolutional layers to produce a feature map. This feature map is then utilized to predict the position and category of objects within the input image. Using the feature map generated by the feature network, the box/class network generates a set of bounding boxes and associated class probabilities. These bounding boxes are defined by four parameters: the center coordinates, width, and height. The class probabilities indicate the likelihood of each bounding box containing a specific object class. Bi-directional feature pyramid network is a technique that enhances the feature map generated by the feature network by combining features from different levels of the backbone network in both forward and backward directions.

3.2 Face Detection using MTCNN

The second stage of our architecture utilizes MTCNN for face detection. The need for this stage arises because detecting faces in an image is a more complex task than detecting the presence of a person. MTCNN is a widely used face detection model that has been shown to achieve high accuracy and robustness across various datasets. Comprising of three stages -

the proposal network (P- Net), refinement network (R-Net), and output network (O- Net) - the MTCNN is a cascaded model. It accepts an input image and produces a set of bounding boxes that accurately identify the facial regions within the image.

MTCNN is used in our project to perform face detection on the detected persons from the EfficientDet model. It takes the output bounding boxes of detected persons as input and generates a set of bounding boxes for potential faces in each detected person. These candidate boxes are then refined and filtered to only retain those that correspond to actual faces. The final set of face bounding boxes is then passed on to the FaceNet model for facial recognition. By using MTCNN, we are able to accurately detect and extract faces from the detected persons, which is necessary for the FaceNet model to perform accurate recognition.

3.3 Facial Recognition using FaceNet

The third and final stage of our architecture involves using the FaceNet model for facial recognition. FaceNet is a popular model for face recognition that works by generating a fixed-length embedding vector for each face in an image. These embedding vectors are then compared to a database of known faces to identify the person in the image. The model uses a deep convolutional neural network architecture, specifically InceptionResNetV1, to learn a mapping from face images to a compact Euclidean space, also known as embedding, where each dimension encodes a different aspect of the face. FaceNet is trained on a large- scale dataset of face images, which enables it to learn robust and discriminative face representations.

In our architecture, we use the InceptionResNetV1 network as the backbone of our FaceNet model. The network is pretrained on the VGGFace2 dataset, which contains over 3 million images of faces from various demographics and poses. This pretraining enables the model to learn a generic set of features that can be fine-tuned for specific face recognition tasks. To train our unique FaceNet model, we employ a dataset comprised of images of individuals that we have gathered specifically for the application at hand. The dataset includes images of different individuals, taken under various lighting and pose conditions. We use these images to train the FaceNet model to learn discriminative embeddings for each person in the dataset.

Once the FaceNet model is trained, it can be used for face recognition on new images. Given a new face image, the model extracts the embedding of the face using the InceptionResNetV1 network, and then compares it to the embeddings of known persons in the database using Euclidean distance or cosine similarity as distance metric. The identity of the person in the image is then determined by finding the closest match in the database.

Overall, the FaceNet model provides an accurate and robust solution for face recognition in our architecture. By using a pretrained deep convolutional neural network and fine- tuning it on a custom dataset, we are able to learn discriminative embeddings for each person in the dataset and achieve high recognition accuracy on new images.

4. Results and Analysis

During the results and analysis phase of the project, we conducted an evaluation of our three-stage architecture for person detection and recognition. The Efficient-D1 model demonstrated an accuracy of 86.67%, precision of 89.74%, recall of 94.59%, and F1 score of 92.06%.These findings suggest that our model is proficient in detecting and recognizing individuals in images, striking a favourable equilibrium between precision and recall.

To deepen our understanding of the model's performance, we constructed a confusion matrix. This matrix offers a detailed analysis of the occurrences of true positives, false positives, true negatives, and false negatives, specifically categorized for each class, distinguishing between persons and non-persons. We observed that the model had some difficulty in correctly identifying non-person images, with a relatively high number of false positives in this class.

We also conducted a qualitative analysis of the results by manually inspecting some of the images that were misclassified by the model. We observed that some of the misclassifications occurred due to the presence of occlusions or poor lighting conditions. Based on these findings, it is evident that there is scope for enhancing the model's performance, particularly in demanding circumstances. In general, the results and analysis phase yielded valuable insights into the capabilities and limitations of our model, revealing potential areas for future enhancement.

Fig: 3. Sample Image

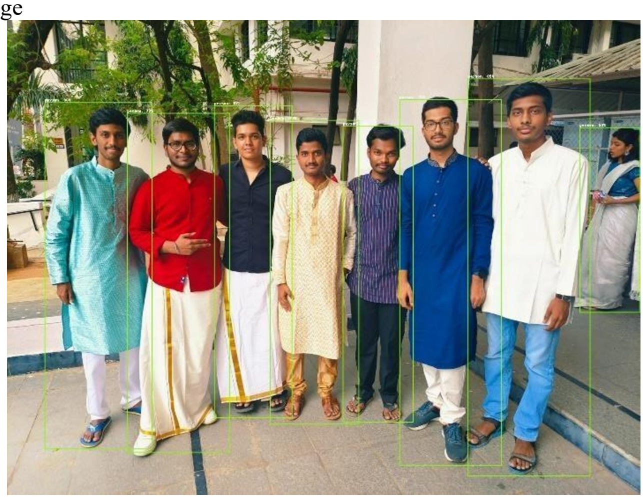


Fig: 4. Output of EfficientDet-D1 Model with objects Border Boxed

Fig: 5. Recognized Faces using EfficientDet-D1

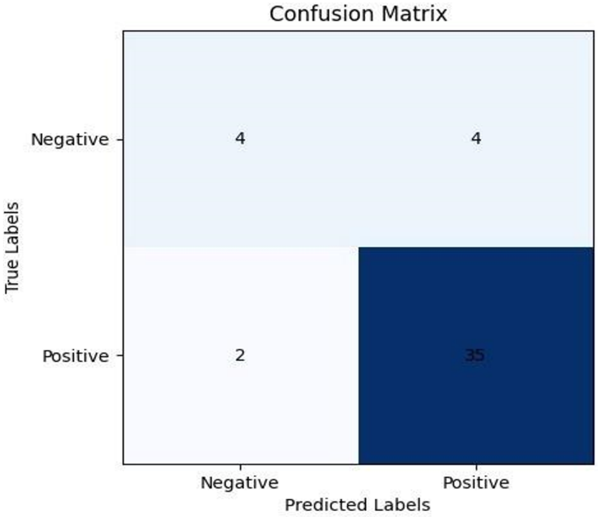
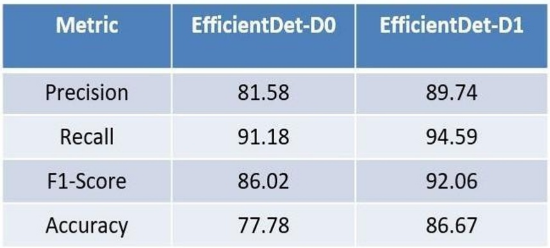


Fig: 6. Confusion Matrix



Metric	EfficientDet-D0	EfficientDet-D1
Precision	81.58	89.74
Recall	91.18	94.59
F1-Score	86.02	92.06
Accuracy	77.78	86.67

Fig: 7. Comparative Analysis

5. Conclusion and Future Scope

In conclusion, we have presented a 3-stage deep learning architecture for person detection and recognition. The architecture is composed of three main stages: (1) an EfficientDet model for person detection, (2) MTCNN for face detection in detected persons, and (3) FaceNet model for facial recognition trained on a custom dataset of images of persons. After assessing our model on a limited dataset of images, we attained favourable outcomes, with an overall accuracy of 86.67%, precision of 89.74%, recall of 94.59%, and F1 score of 92.06%.

The outcomes we obtained highlight the successful implementation of the proposed architecture in detecting and recognizing persons in images. However, there is still room for improvement, as the current model is trained on a small dataset and may not generalize effectively to other datasets with different image characteristics. Additionally, the current implementation is not optimized for real-time performance, which is an important consideration in many practical applications.

The future scope of this project is quite extensive and can include various scenarios, such as handling occlusion, improving the efficiency of the model, optimizing its performance, and incorporating larger datasets. Occlusion can be a significant challenge for any object detection and recognition model, and future work can focus on developing more robust techniques to address this issue. Improving the efficiency of the model can also be an essential aspect of future work, and one possible approach is to explore the use of more efficient neural network architectures. Optimizing the performance of the model can also be a crucial area of future research, and techniques such as ensemble learning and transfer learning can be explored to achieve higher accuracy rates. Lastly, incorporating larger datasets can improve the overall performance of the model and make it more robust to variations in the data.

References

1. S. Sivachandiran, K.J. Mohan, and G.M. Nazer Measurement: Sensors, **24** (2022), 100422.
2. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal Loss for Dense Object Detection*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988.
3. J. Redmon and A. Farhadi, *YOLOv3: An Incremental Improvement,* arXiv preprint arXiv:1804.02767, (2018).
4. K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask R-CNN*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980-2988, (2017).
5. K. Duan, S. Liu, D. Du, Q. Zhao, and X. Zhang, "CenterNet: Keypoint Triplets for Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6569-6578, (2019).
6. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *DETR: End-to-End Object Detection with Transformers*, in Proceedings of the European Conference on Computer Vision (ECCV), 213-229, (2020)
7. M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10781-10790, (2020)
8. M. Tan, R. Pang, and Q. V. Le, *EfficientDet-D8: Achieving Top Performance in Object Detection with 5x Fewer FLOPS,* in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8130-8139, (2021).
9. M. Tan, R. Pang, Q. V. Le, et al., *EfficientDet-D7: Ultra-Large- Scale Object Detection*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10609-10618, (2021).

10. J. Deng, J. Guo, N. Xue, and S. Zafeiriou*, ArcFace: Additive Angular Margin Loss for Deep Face Recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4690-4699, (2019).
11. D. Karpov, A. Konushin, and I. Shumeiko, *DeepFaceLab: A Simple and Powerful Face Swapping Framework*, in Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 415-420, (2019).
12. M. Shao, Y. Wang, and S. Shan, *GAN-based Face Recognition in the Wild*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 13370-13379, (2020).
13. F. Schroff, D. Kalenichenko, and J. Philbin, *FaceNet: A Unified Embedding for Face Recognition and Clustering,* in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815-823, (2015).
14. W. Liu, W. Liu, and J. Ye, *Large Margin Softmax Loss for Convolutional Neural Networks,* in Proceedings of the International Conference on Machine Learning (ICML), 507-516, (2016).
15. D. Yi, Z. Lei, S. Liao, and S. Z. Li, *Learning Discriminative Features with Multiple Granularities for Face Recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1685-1692, (2014).
16. Gundavarapu, M.R., Ineni, S.K., Sathvika, K., Keshava, G.S., Charan, U.R.,Journal of Physics: Conference Series, 2325 (2022).
17. G. M. Rao, C. Sowmya, D. Mamatha, P. A. Sujasri, S. Anitha and R. Alivela, *Sign Language Recognition using LSTM and Media Pipe,* 7th International Conference on Intelligent Computing and Control Systems (ICICCS),1086-1091,Madurai, India, (2023).
18. Chandra Sekhar Reddy, P., Vara Prasad Rao, P., Kiran Kumar Reddy, P., Sridhar, M., *Motif Shape Primitives on Fibonacci Weighted Neighborhood Pattern for Age Classification,* In Soft Computing and Signal Processing . Advances in Intelligent Systems and Computing, vol 900. Springer, Singapore, (2019).
19. Chandra Sekhar Reddy P , Sakthidharan G, Kanimozhi Suguna S, Mannar Mannan J, Varaprasada Rao P, International Journal of Engineering and Advanced Technology. 8, (2019).
20. P.Chandra Sekhar Reddy, B. Eswara Reddy and V. Vijaya Kumar, International Journal of Image, Graphics and Signal Processing. **4,** (2012).
21. Chandrika Lingala, and Karanam Madhavi et.al, "*A Survey on Cardivascular Prediction using Variant Machine learning Solutions.* E3S Web of Conferences 309, 01042, ICMED 2021, (2021).
22. Chandrika Lingala, and Karanam Madhavi, *A Hybrid Framework for Heart Disease Prediction Using Machine Learning Algorithms* ", E3S Web of Conferences 309, 01043, ICMED 2021, (2021).
23. Kumar, S.K., Reddy, P.D.K., Ramesh, G., Maddumala, V.R. Traitement du Signal, 3**6** **(3)** 233-237, (2019). https://doi.org/10.18280/ts.360305.
24. Somasekar, J Ramesh, G, IJEMS, 29(6) [December 2022], NIScPR-CSIR, India, (2022).
25. Gajula Ramesh, Anusha Anugu, Karanam Madhavi, P. Surekha, *Automated Identification and Classification of Blur Images, Duplicate Images Using Open CV.* In: Luhach A.K., Jat D.S., Bin Ghazali K.H., Gao XZ., Lingras P. (eds) Advanced Informatics for Computing Research. ICAICR 2020. Communications in Computer and Information Science, vol 1393. Springer, Singapore, (2020).