Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Applications Of Machine Learning In Computer Vision: A Review

M. Senthil Kumar[1] , S Venkata Achuta Rao[2]

[1]Professor, *Electronics and Communication Engineering, Sree Dattha Group of Institutions, Hyderabad, India. Email: professor.msk@gmail.com*
[2]*Professor, Computer Science and Engineering, Sree Dattha Institute of Engineering & Science, Hyderabad, India. Email: sreedatthaachyuth@gmail.com*

The field of computer vision that deals with the effort of fostering capabilities of the computers in analyzing the visual information and making decisions on the same is enhanced by the achievements in machine learning (ML) techniques. This paper presents a detailed survey on how machine learning is used with different fields of computer vision including image classification, detection, facial recognition, and medical applications. Deep learning, neural network, generative models, performance, and limitations, and further outlook are also discussed. This paper also provides an overview of datasets, algorithms as well as measures adopted in computer vision applications and trends for the future research are also presented.

Keywords: Machine Learning, Computer Vision, image classification, object detection, facial recognition, and medical imaging.

1.Introduction

CV is a subset of both AI and computer science that focuses on teaching computers how to understand and analyze visual data in comparable to a human being [1]. This was however made possible by the introduction of machine learning especially the deep learning methods that do not require feature engineering from designers. Thanks to this capability of learning from data, CV has greatly advanced and is no longer a theoretical but a set of tools applicable in various fields including but not limited to healthcare, automobile, farming, security, entertainment and many others.

One fact that facilitates this success of machine learning on CV is that the field has rapidly grown in terms of both algorithmic innovations and the computational power accessible. Major advancements in algorithms like CNNs, GANs and most recently transformers models have helped increasing solution accuracy within visual tasks ranging from image classification, object detection and segmentation [2]. Similarly, advancement in the large dataset such as ImageNet, COCO and PASCAL VOC, GPU and cloud computing has by far boosted this growth.

In this paper is to offer a literature review of modern use cases of machine learning in computer vision. It mainly covers basic types of tasks, such as image recognition, object recognition and detection, face recognition, image segmentation and analysis, and medical image analysis. Based on these applications, we explain the most common approaches to using machine learning methods, the types of data and annotations needed for each application, and the performance metrics for real-world applications. In this frequency-based review, it fills the gaps and shortcomings of the existing ML approaches: collect massive, labelled data; contain bias; and privacy-related ethics [3]. The purpose of this review is to help researchers and practitioners have an appreciation of the current scholarship in the field and potential future developments in the field. In this way, we hope to expose potential improvement avenues by fleshing out the essential attributes and consequences of different approaches in the CV domain, especially in improving the performance, explainability, and fairness of novel ML methodologies desired for extending CV app ranges.

2.Background

Recent years witnessed proliferation in the field of machine learning in computer vision where computer's ability to analyze images and videos was progressed through various stages of rule-based systems to complex neural networks. CV, in its principle, is concerned with the process of training machines to recognize and analyze input images or videos to inform decisions and subsequent actions [4]. The elements of this change have been brought by the machine learning algorithms which allow models to learn on the data and not rules of patterns.

2.1 Development process of Machine Learning in the field of Computer Vision

Originally computer vision was using hand engineered features and statistical techniques for activities like edge detection, shape recognition and simple matching [5]. These approaches involved significant amounts of manual labour to engineer features and required specialised expert knowledge to do so and as such were rigid and inflexible with low generalization capabilities. This is true owing to the discovery of Machine learning, especially Deep learning (DL), which the CV has been transformed. CV machine learning approaches let models train directly from inputs, bypassing feature definition while yielding a far higher level of accuracy and consistency across a staggering array of processes.

The first breakthrough came in the 1980s with what's called the convolutional neural network (CNN). CNN architectures are mostly pyramidal in that while the initial layers in the network learn fine grained features such as edges and intensity variations, the higher-level layers elaborately extract complex features as shapes, identification and location of objects. The advancement of CNNs could therefore be said to have been sparked by AlexNet in 2012, which went on to win the ILSVRC to record breaking accuracy thus marking the start of Deep Learning for CV [6].

2.2 The Two Techniques of Machine Learning in Computer Vision

Modern CV applications predominantly utilize deep learning, which includes various types of neural networks optimized for handling visual data [7]:

- **Convolutional Neural Networks (CNNs):** CNNs are intentionally created to deal with data that is laid out on a grid, for example, images. They are used in almost all CV tasks with primary applications in image classification, object detection as well as segmentation. There is also the CNN architectures like pre-trained ResNet, Inception, and EfficientNet enable better accuracy, and efficient computation hence essential for CV.

- **Recurrent Neural Networks (RNNs):** Although originally intended for use in NLP, they can be effectively applied to video analysis in CV, since temporal dependencies are critical in this context. They are especially beneficial for those problems where sequence learning is necessary, for instance, recognition of actions in videos.

- **Generative Adversarial Networks (GANs):** Proposed by Ian Goodfellow in 2014, GANs are comprised of two models an unsupervised discriminator model and a generator model. They can be applied to images generation, style transfer and data augmentation especially for situations where there is limited data.

- **Transformers:** Originally proposed for NLP tasks, the transformers are relatively newly applied to the CV domain, specifically as Vision Transformers (ViTs) demonstrated best performance in image classification and in other areas. Transformers use self-attention, which for that reason, it has been applied to tasks where long-range contextual information in images is critical.

- **Deep Reinforcement Learning (DRL):** DRL is well suited to situations where there is a need to choose between sequences in decision making including robotics and autonomous driving. However, by using reinforcement learning with the deep learning, the DRL enables systems to learn from practice in the simulated or real environment.

2.3 Datasets and Benchmarks

The availability of large volume high quality datasets has been instrumental in the progress of the ML in the application domain of CV. Datasets make available the large amount of labelled data that is needed to train and evaluate machine learning models. Key datasets include [8]:

- **ImageNet:** A collection of images with over 14M labeled images across thousands of categories. ImageNet has been ar important resource for image classification and has been used as a standard dataset to evaluate new models of ML.

- **COCO** (Common Objects in Context): This dataset include over 330K images with object segmentation and captioning annotation which makes it suitable for object detection, segmentation and captioning.

- **PASCAL VOC:** This dataset contains slabelled images for object detection, segmentation, and action classification which makes it a starting point for many CV methods.

In addition to serving as training data, these benchmarks have fostered competition and advances in the field of CV by establishing goals and pushing researchers to addresses large scale performance issues.

2.4 Evaluation Metrics

Machine learning models in CV are evaluated using metrics tailored to specific tasks [9]:

- Accuracy, Precision, and Recall: Mainly applied in classification problems where standard measures its capacity to categorize items properly; they include precision that is the proportion of the number of actual positive in the given community with the number of positive stated by the program and recall that is the ratio of actual positive cells with the positive stated by the program.

- Intersection over Union (IoU): When applied in object detection as well as segmentation tasks, IoU compares the amounts of overlapping of the proposed and the ground truth bounding boxes, providing a level of localization precision.

- Mean Average Precision (mAP): Most often used for object detection, mAP takes the average of precision across the classes ensuring balanced measures of both precision and recall.

- Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC): These metrics used in medical image analysis and more often used when the classes are imbalanced, measure the cost of predicting one class over the other.

2.5 difficult technical problems in machine learning for computer vision

Though nice progresses have been made in developing ML to enhance CV, few issues remain. Most learning models rely on large datasets that have human labels, which learning models that are expensive and time-consuming to acquire [10]. Furthermore, most of the current ML models, specifically deep learning structures, are high processing and high memory requirement models. Privacy is a core problem in implementing CV applications, especially in security systems such as surveillance, health, and police work; another is the presence of bias within models.

3.Key Applications of Machine Learning in Computer Vision

The combination of machine learning (ML), specifically deep learning (DL), have significantly enlarged the possibilities of computer vision (CV) to interpreted sophisticated visual information by computing devices [11]. Some of the most important CV applications with the help of ML include image classification, object detection, facial recognition, semantic segmentation and medical imaging all of which have significant uses in fields such as healthcare, autonomous vehicles, security and personal entertainment respectively.

3.1 Image Classification

Image classification is one of basic types of image analysis tasks and encompasses the , recognition of the images belonging to a predefined set of classes. Due to achievement of ML, especially CNNs, has created very high accurate models for classification of thousands different objects. Image classification applications include [12]:

- **Healthcare:** Ines and Ernstson mentioned that it is employed to find pathological signatures within the X-Ray, MRI scans, and other images for diseases such as pneumonia, cancer, and cardiovascular diseases.
- **Autonomous Vehicles:** It is used to identify road signs, vehicles and pedestrians to enable autonomous systems to make right navigational decisions.
- **Retail and E-commerce:** According to the classification, products are sorted and tagged automatically, which helps to improve the task of cataloguing items and search in the sphere of online retailing.

Challenges

Generally, image classification models depend on the availability of large, labelled datasets and collecting these in specialized domains such as healthcare can be very challenging. This can make a neural network deal with intra-class variability where the objects belonging the same class differ significantly in their appearances, and inter-class similarity where contrasting objects in the image bear resemblance to one another.

3.2 Object Detection

Object detection goes beyond the methods used in image classification where the algorithm not only determines what object is in the image, but where they are; this is often shown with boxes drawn around such objects. By now, we have Faster R-CNN, YOLO and SSD which facilitate real-time object detection, which is vital for applications that need dynamism in action. Major applications include [13]:

- **Surveillance and Security:** Object detection is applied in surveillance systems with the purpose of identifying suspicious objects, everyone, and detect any out of the ordinary activities in public areas.
- **Autonomous Driving:** Self-driving cars apply object recognition to detect and therefore follow other automobiles, people, signs, and road barriers, etc.
- **Industrial Automation:** The object detection brings quality assurance since it is able to detect faults in manufactured items, conformity, and productivity.

Challenges

High-speed object detection is critical in real-time applications, but it comes with the trade-off of accuracy – especially in self-driving cars. Partial occlusion, where an object appears masked and the identification of small objects within large image continues to be a hurdle to many detection techniques.

3.3 Facial Recognition

The utilization of Facial recognition in ML has stood out that relabelling or verifying faces in CV based on features is quite widespread. Facial recognition ML models are typically trained to identify distinguishing patterns, even in different lighting and pose, and different expressions. Common applications include [14]:

- **Security and Access Control:** Major applications of facial recognition include use in airports, identification at workplace, and smartphones for user authentication and security clearance.
- **Consumer Electronics:** The unlock features of face in smartphones and the other personal devices have been efficient and effective as methods of user authentication.
- **Marketing and Retail:** Identifying susceptible for specific ideas or perceptions in the store, for purposes of target marketing and client satisfaction.

Challenges

Sub issues or risks are then that FRS technologies are ethical problematic due to privacy issues, misapplication, and racist/disparaging effects through biases in ML Algorithms. Some of the technical issues include how to deal with pose and lighting changes as well as occlusions that hampers recognition.

3.4 Semantic Segmentation

Semantic segmentation is even more advanced than object detection since it labels each pixel in a picture or scene. Some methods used to obtain good results in pixel-weighted predictions include fully convolutional networks (FCNS) as well as the U-Net. Applications include [15]:

- **Autonomous Driving:** Semantic segmentation is essential to the identification of drivable areas, sidewalks, pedestrians and obstacles; for a self-driving system it is a way to keep the environment safe.
- **Medical Imaging:** In healthcare, segmentation models are applied to localize areas of interest, like a tumor or organ from images with the purpose of accurate diagnosis and therapy planning.
- **Augmented Reality (AR):** Correct segmentation will enable AR systems to place virtual information onto a scene by identifying how the scene is laid out with practical applications in gaming, learning, and construction.

Challenges

Semantic segmentation is very computationally intensive especially when real time applications like self-driving cars are in consideration. Keeping precision in complex or chaotic scenes is also difficult since object separation is not very clear compared to simple geometry.

3.5 Medical Imaging

Artificial intelligence has predefined considerable changes within medical imagery as it helps radiologists as well as clinicians by offering them tools that can detect all sorts of pathologies and point out specific anomalies. Applications of ML in medical imaging include [16]:

- **Cancer Detection:** ML models can also recognize tumours in mammography, CT scans and MRI scans, helping to diagnose cancers including breast, lung and brain cancer at the initial stage.
- **Cardiology:** In cardiac imaging, the applying of ML helps in the examination of the heart, diagnosis of diseases such as; arrhythmia and the identification of plaque in arteries.
- **Neurology:** With regards to neurology, ML helps in the analysis of imaging scans of the brain to identify abnormalities associated with diseases such as Alzheimer's, Parkinson's and Multiple sclerosis.

Challenges

There was a high level of accuracy required by the medical imaging applications since their results are vital. They also call for large, well-annotated datasets and again in healthcare domain, it is challenging to address these two constraints mainly because of data privacy and paucity of manually annotated data. Another strategy is also needed to overcome the problem of interpretability of the model and its predictions, which is crucial for getting the approval of clinicians and, consequently, regulatory authorities.

3.6 Anomaly Detection

Abnormality identification in CV means discovering those patterns which contain events and objects that are significantly different from the rest. Machine learning techniques, especially unsupervised learning and deep anomaly detection, are applied to various fields where anomaly identification is critical [17]:

- **Industrial Inspection:** In an instance, the ML models help in defining the quality of the products in a manufacturing plant through distinguishing the factory-produced shapes, sizes or patterns that do not fit a particular standard.
- **Surveillance and Public Safety:** Security detects anomaly or suspected behavior in the public realm and can aid law enforcement in observation in real-time.
- **Healthcare:** In diagnostics, it is applied to look for such features in the images that would suggest disease or abnormality and help the early diagnosis.

Challenges

The major difficulty of the anomaly detection is the absence of the amount of data with labels of the events that should be considered as deviations. It is theoretically instructive to note that high false positive means non-anomalous data is flagged, and this causes inefficiency and mistakes in important activities.

3.7 Scene Understanding

Scene understanding is the integration of several CV tasks which include detection, segmentation and depth learning to understand a given image or video scene expit. This holistic approach has applications in [18]:

- **Robotics:** With CV, self-driving robots can decode their environment to allow them to move, interact with objects, and the people around them.
- **Virtual Reality (VR) and Augmented Reality (AR):** Scene understanding is very significant for VR/AR system since it enables the systems to place VR/AR objects at the pertinent scenes in the actual environment, supporting the formation of immersive scenes.
- **Smart Homes:** Interactive space has scene understanding that helps home automation systems categorize and analyze scenes within the house, as well as human activity, in order to provide prompt help and automation.

Challenges

In addition to it, Scene understanding models should sustain complex and crowded scenes, and different forms of illumination, occlusion, and additional objects. When essentially multiple CV tasks are to be incorporated, additional computational load is imposed; hence, time-efficient models are desirable.

4.Machine Learning Methods in Computer Vision

The progress that has characterised the increased potency of computer vision (CV) has been fuelled by the efficient algorithms in machine learning (ML) particularly deep learning (DL). These methods allow reducing the human intervention during the processing, analysis and understanding of visual information and acting on this information with the help of machines. Some of these techniques adopted by CV include CNNs, RNNs, transformers, GANs, and deep reinforcement learning. These methods are quite different from each other and are most appropriate for various tasks of CV such as image classification, scene understanding.

4.1 Convolutional Neural Networks CNNs

Convolutional neural networks (CNNs) or also known as the base model of CV, used for the structure of hierarchical relationship into the noticeable features of the specific set of visual information by applying node layers of the convolution. CNNs work through filter sliding convolution over input images feature maps and output results of the pattern detection where each layer extracts high complexity and abstract features ranging from edges to complete features of the actual object [19].

- **Applications:** CNNs have been utilized for image classification, object detection and image segmentation for a long time. For example, ResNet, VGGNet has been vital to ensure better accuracy in large scale dataset such as ImageNet.
- **Key Variants:**
  - **AlexNet:** In 2012, another work called AlexNet presented the world how CNNs can perform image classification, bringing a big step forward in CV accuracy.

- **ResNet:** This deep CNN introduced residual learning, in other words, allowed training very deep networks with the help of skip connections to overcome the problem of vanishing gradients.
- **EfficientNet:** This family of models improves both precision and figures of merit, effectively extendible CNNs with limited resources in terms of depth, width, and resolution.

Strengths and Limitations

CNNs are general structure able to process data organized in a grid such as images. But to do so they need large datasets for training and substantial amount of computing power. They also do not work well when dealing with temporal data, and this is where the recurrent architectures are most useful.

4.2 Recurrent Neural Networks (RNNs:

Recurrent neural networks, as the name suggests is better for datasets having sequential relations between the elements. While developed primarily for data such as language, RNNs are valuable in CV and especially for resolving issues such as video analysis and sequence prediction. Two primary types of RNNs are LSTM, and GRU specifically developed to combat the vanishing gradient problem while capturing long dependencies between data [20].

- **Applications:** In video analysis, RNNs are applied to activities like activity recognition and object tracking, and video description generation.
- **Key Variants:**
  - **LSTM:** The memory cell and gating mechanism are responsible for storing information over a sequence where LSTM yields its best performance in such sequences.
  - **GRU:** Another version of LSTMs are GRUs they also memorize over longer sequences but have fewer parameters due to which it less time consuming.

Strengths and Limitations

RNNs are concretely useful with sequential information and dedicated to temporal patterns identification, which makes them suitable for processing videos. However, they are time-consuming and sometimes fail to handle sequences of very large size, in which the transformer architecture has exhibited more efficiency.

4.3 Transformers

Transformers that were designed for NLP tasks have been applied to the CV tasks with networks such as the Vision Transformers (ViT). Unlike CNNs where convolutions are used to capture local attributes, transformers utilize self-attention for depiction of cross image relation. This helps to get a broader understanding of the visual data, as transformers are very efficient for almost any CV tasks [21].

- **Applications:** Vision transformers are primarily employed in image classification, segmentation and any task which demands spatial attention. They also perform well when dealing with big data problems.
- **Key Variants:**

- Vision Transformer (ViT): The original transformer used in CV, ViT splits an image into patches and feeds them as sequence and methods this to compete with CNNs especially on large datasets.

- Swin Transformer: This one is better to implement than LM and has both advantages of transformers and500 improved local attention to make it more suitable for high-resolution images.

Strengths and Limitations

The advantages of transformers are they are effective in perceiving the global context in images, and do not strictly need inductive bias as the CNNs do. However, they mostly demand massive data to work and at times can be very greedy in terms of computational resources especially while in their basic form.

4.4 GANs:

GANs are a kind of ML that is made of two unites; a generator models and a discriminator model that plays an adversarial game to generate realistic data. In the generator, the idea is to generate data points that are as close as possible to a set of real samples, and in the discriminator case the aim is to distinguish that data point as being either real or simulated. This adversarial process yields virtually photo-realistic synthetic images [22].

- Applications: GANs find their applications in image synthesis, augmentation, styling and super-resolving. It can be used in synthesising training data for obscure cases, increasing the resolution rate in medical images, and developing quality textures in anima-tion.

- Key Variants:

- DCGAN (Deep Convolutional GAN): As one of the first successful approaches to image generation with the use of GANs, DCGAN incorporates CNNs into the structure of a GAN.

- CycleGAN: This model is famous for style transfer tasks, applying certain style to images of any other style without having corresponding pairs for training (e.g., applying winter images to summer images).

Strengths and Limitations

GANs can create realistic data that can be applied to more creative tasks or to the problems of data expansion. But they are hard to teach, may depend on the big sets of data, and might exhibit the problem of mode collapse, where generator offers minimum variation in the result.

4.5 Deep Reinforcement Learning (DRL)

The proposed architecture termed as deep reinforcement learning (DRL) integrates reinforcement learning with the help of agents that develop through interaction with environments and deep learning. In CV, DRL is applied in scenarios where the decisions should be made successively for instance, self-driving, robotics, and gaming [23].

- Applications: By interacts with ant robot or application, DRL optimizes them in object manipulation and path planning in robotics, decision making and navigation in autonomous cars and automated diagnosis in medical imaging.

- Key Variants:

- DQN (Deep Q-Network): A reinforcement learning method used together with CNNs for image-based game playing and applicable in cases where the action space is discrete.

- A3C (Asynchronous Advantage Actor-Critic): An efficient DRL method for a continuous action space, which can be useful in complex environments, such as robotic navigation systems.

Strengths and Limitations

Dynamic environments involving real-time decision making are well supported by DRL. However, it needs a huge amount of computation and is very sensitive to the choice of hyperparameters, and thus difficult to be used efficiently in complicated applications.

4.6 Hybrid Approaches

Recent developments in CV suggest that the field has been opting for more and more approach that combine one or more ML algorithms. The mixed models enhance performance in task execution by interfacing CNNs, transformers or GANs for moderation of spatial feature learning and context extensiveness. For instance [24]:

- CNN-Transformer Hybrids: These models take advantage of the locally connected feature extraction of the CNN with transformers' global attention for Imaging cycle efficiency.

- GAN-Augmented CNNs: Moreover, CNN training is augmented using synthetic data, which originates from GANs; it helps to overcome performance reduction due to diverse obstacles that appear amidst limited data.

- Reinforcement Learning with CNNs: While for example images are used in the autonomous driving case, CNNs are used to process images while DRL algorithms take decisions with processed information.

4.7 Unsupervised and Self Supervised Learning

Since labelling of large datasets is still a problem, we have seen the recent development of unsupervised and self-supervised learning techniques to take advantage of the concept of using unlabelled data. These strategies enable models to be trained with no reliance on explicit supervisory information but rather rely on the internal structure or on standard tasks to create the representations.

- Applications: Self-supervised learning is applied for downstream tasks in a low label scenario as in medical imaging. It also supports anomalous data learning representations that highlight numerical data anomalies.

- Key Techniques:

- Autoencoders: These networks learn efficient representaion by passing through the input data through a compression and reconstruction process such as for feature extraction and for detecting the abnormality.

- Contrastive Learning: It thus elaborates on contrastive methods, like SimCLR and MoCo, that learn the representation by comparing similar and dissimilar data pairs and enhance representation learning for the downstream tasks.

Strengths and Limitations

Notably, self and unsupervised methods eliminate to a larger extent the requirement of labeled data and provide more flexibility to models. However, these methods need comparatively more computational resources for deriving sufficiently effective representation vectors; their utility subsumed to the quality and variability of data.

5.Datasets and Evaluation Metrics

Evaluation of datasets, and measures are essential in creating, comparing, and validating Supervised and Unsupervised learning-based machine learning (ML) models in the computer vision (CV) [25]. We have training data and testing data in datasets which is useful for the ML model learning phase and generalization stage respectively, and we have evaluation parameters for an ML model performance that's useful for improvement of the same model.

5.1 Datasets

Lots of datasets have been introduced for various CV tasks with some special features of annotations for the particular application areas like object detection, image segmentation and face recognition. The following are some of the most widely used datasets in CV research and applications:

5.1.1 Image classification datasets

- **ImageNet:** A massive image collection that with more than 14 million images categorized into over 20000 classes. There are thousands of projects every year to improve the CV, especially in the context of the ILSVRC which is annually held on the ImageNet.
- **CIFAR-10 and CIFAR-100:** These are relatively small, general purpose image classification datasets comprising of 60 000 images curated into 10 classes (CIFAR-10) and 100 classes (CIFAR-100). They enable assessment of new architectures and procedures especially when computational resources are limited.
- **MNIST:** a large database consisting of 70,000 numeric digits consisting of 60,000 for training and 10,000 for testing in greyscale. Despite this, MNIST is a ground norm for simple image classification and pattern recognition models at their base.

5.1.2 Object Detection Datasets

- **COCO (Common Objects in Context):** COCO has more than 330k images with pixel-level annotations of segmentation of objects falling under 80 categories. This capability includes object detection, segmentation, keypoint detection, and captioning making it very general and commonly used in evaluating object detection and segmentation models.

- **PASCAL VOC:** This dataset provides images that it is possible to make annotations for object detection, segmental and action, while sharing 20 overall tags of object. It has become instrumental in evaluating object detection algorithms and is always used together with the COCO dataset.

- **Open Images:** A publicly available dataset consisting of over nine millions of images together with object annotations as well as visual relationships. Open Images involves large scale research on object detection, object relationship and scene understanding.

5.1.3 Semantic segmentation datasets

- **Cityscapes:** A dataset on urban street scene with 5,000 highly annotated images and containing 19 semantic classes. Due to its pixel-level annotations on various objects enable separation of images into meaningful segments Carmero is widely used for creating models for autonomous driving applications.

- **ADE20K:** This dataset contains more then 20K images with semantic segmentation and instance segmentation annotations for 150 classes of objects. Due to its designed objective of supporting scene parsing and segmentation, ADE20K can be useful for robotics and scenarios concerning urban scene understanding.

- **CamVid:** A different dataset of 701 labeled images for the road scene segmentation problem. CamVid is employed in evaluating and training semantic segmentation models for use in self-driving and navigation.

5.1.4 Facial recognition datasets

- **LFW (Labeled Faces in the Wild):** LFW is a large-scale dataset containing more than 13 Thousands images extracted from web, each image contains faces of individual(s) labeled with their identity. It is mostly applied in assessing face verification and recognition solutions in realistic scenarios.

- **VGGFace2:** As a large database of more than 3,322,000 images of over 9,000 individuals, VGGFace2 offers various pose, age, and lighting variability to facilitate research on facial recognition and other CV tasks wherein face information is vital.

- **MS-Celeb-1M:** This dataset contains face images of the order of multiple millions for about a hundred thousand identities and as such it serves as a rich source for training and testing face recognition systems.

5.1.5 Medical Imaging Datasets

- **ChestX-ray14:** A benchmark dataset of more than 100,000 chest X-ray images with 14 disease labels for the creation of models for use in radiology.

- **ISIC** (International Skin Imaging Collaboration): Arising from the image analysis competition ISBI2017, this dataset provides dermoscopic images of skin lesions for melanoma detection and other dermatological purposes.

- **BraTS** (Brain Tumor Segmentation): The BraTS dataset is specifically concerned with brain MRI scans and includes multimodal imaging data labeled to guide brain tumor segmentation, which is critically integral to studies in neuro-oncology.

5.2 Evaluation Metrics

Measures are very significant in judging the performance of an ML model in CV. Heterogeneous types of tasks demand different measurement attributes to reflect the trends in model accuracy, precision, recall and robustness.

5.2.1 Image Classification Metrics

- **Accuracy:** Measures the percentage of correctly generated items out of the total items primarily for use with balanced sets of data. It is not useful when the data set being analyzed is unbalanced or has more positive cases than negative ones, precision, and recall may be more beneficial.

- **Precision and Recall:** Accuracy is the percentage of right predictions as positive across all such predictions while recall is the percentage of actual positives found among all those predicted as such. These metrics are very important in the fields that are strictly divided by positive and negative results; for instance, patients' diagnosis from images.

- **F1 Score:** The H average of precision and recall, giving a value that is the average of the two values. F1 score is handier when working on problems based on the multiply imbalanced classes.

5.2.2 Object Detection Metrics

- **Intersection over Union (IoU):** The mean of the per-image intersection over union between predicted boxes and ground truth boxes. IoU plays vital role in the object detection for measuring the accuracy of the object localization.

- **Mean Average Precision (mAP):** A metric that averages the per-image precision for different IoU thresholds and object classes, a single scalar value mAP, is used for evaluating the object detection models on COCO and other similar datasets.

- **Average Precision (AP):** As with mAP, AP considers the area under the precision-recall curve for a single IoU threshold and provides an accuracy of each class in an object detection problem.

5.2.3 Semantic Segmentation Metrics

- **Pixel Accuracy:** The ratio of correctly segmented pixels in an image or a dataset to assess a model's basic performance of semantic segmentation.

- Mean Intersection over Union (mIoU): The global intersection over union, mean intersection over union mIoU is one of the most popular evaluation measures in studies on semantic segmentation evaluating the quantity of overlap between the predicted and the ground truth segmentations over all classes.

- Dice Coefficient (F1 Score): Dice coefficient, as IoU, is a popular metric in segmentation task in medical image and is calculated between the predicted and ground truth masks.

5.2.4 Facial Recognition Metrics

Because the accuracy of live face images is an important factor in biometric face recognition, there is a need to consider the following facial recognition metrics:

- Verification Accuracy: Evaluate how accurately the face verification models determine whether or not two face images can be from the same person.

- True Positive Rate (TPR) and False Positive Rate (FPR): TPR means the percentage of proper face matching while FPR signifies the corresponding wrong face matching. Such rates are normally depicted in a form of a receiver operating characteristic (ROC) curve to show model performance for various thresholds.

- Area Under Curve (AUC): Area under the Receive Operating Characteristic curve, AUC offers a single index of accuracy irrespective of the choice of cut-point. The last row displays the extent to which the models perform in terms of AUC where a higher AUC represents a better classifier of positive and negative classes than the other.

5.2.5 General Evaluation Metrics

- Confusion Matrix: The table which shows true positive, false positive, true negative, false negative. The confusion matrix is specifically valuable for determining bias or weak points in the areas of interest; it comments on model performance by class.

- Logarithmic Loss (Log Loss): Log loss works well for the case of models where outputs are probabilities as is the case with classification models. It has a way of punishing wrong classifications especially if the model was very sure about the classification it did.

- ROC-AUC Score: This score defines the balance between TPR and FPR, appropriate for applications with a high level of imbalance between classes or where the proper distinction between them is essential.

Table 1-7 contains the applications of Machine Learning in Computer Vision, Machine Learning Algorithms in Computer Vision, number of images in training datasets, model Accuracy Improvement over years, accuracy of Models Across Datasets and Tasks, classification task in Computer Vision and model training and validation accuracy over epochs.

Table 1 Applications of Machine Learning in Computer Vision

Application	**Usage Percentage (%)**
Object Detection	40
Image Classification	25
Semantic Segmentation	15
Face Recognition	10
Medical Image Analysis	10

Table 2 Machine Learning Algorithms in Computer Vision

Algorithm	**Usage Percentage (%)**
CNN (Convolutional NN)	50
RNN (Recurrent NN)	10
GAN (Generative Adversarial NN)	20
SVM (Support Vector Machine)	10
Other	10

Table 3 Number of Images in Training Datasets

Image Count Range	**Frequency (Number of Datasets)**
0–5,000	5
5,000–10,000	10
10,000–50,000	15
50,000–100,000	8
100,000+	12

Table 4 Model Accuracy Improvement Over Years

Year	**Model Accuracy (%)**
2015	75
2016	78
2017	82
2018	85
2019	88
2020	90

2021	92
2022	94
2023	95

Table 5 Accuracy of Models Across Datasets and Tasks

Dataset/Task	**Object Detection**	**Image Classification**	**Semantic Segmentation**
COCO	85	90	75
ImageNet	88	95	-
Cityscapes	-	-	80
PASCAL VOC	82	88	72
Medical Imaging Dataset	-	92	65

Table 6 Classification Task in Computer Vision

	Cat	**Dog**	**Bird**	**Car**	**Plane**
Cat	50	2	1	3	0
Dog	3	45	2	5	1
Bird	1	2	55	1	0
Car	4	3	2	47	4
Plane	0	1	2	3	54

Table 7 Model Training and Validation Accuracy Over Epochs

Epoch	**Training Accuracy (%)**	**Validation Accuracy (%)**
1	60	58
2	65	62
3	70	68
4	75	72
5	78	74
6	80	77
7	82	79
8	84	81
9	85	83
10	86	84

6.Discussions

Although, machine learning (ML) has brought a revolution in computer vision (CV), several constraints that exist do not offer full-fledged fruitful result for the best use in the real world. It is safe to conclude that challenges regarding component emphasize data demands, computational complexity, interpretability, and ethical concerns [26]. However, there are research directions, which have great potential to overcome the mentioned limitations and can contribute to the development of a sophisticated and widely applicable ML approach for CV.

6.1 Data Dependency and Scalability

A primary drawback in using ML in CV is the need for big, labelled data sets. CNN's and other deep learning models often depend on large dataset which are tough to prepare and at times may lead to financial burdens. Moreover, the access to data differs between domains; for instance, autonomous driving and facial recognition domains are data exhaustive, but domains such as medical imaging might not have sufficient datasets for labelled data [27]. Furthermore, datasets can be tranquil unrepresentative, which causes bad generalization of models in various settings, demographics, and environmental situations.

6.2 Computational and Energy Concerns

Current day Machine Learning algorithms for CV are usually complex and need a lot of resources even in terms of processing power for training and use. Transformer and GAN models are computationally expensive and memory-seeking, so it appears that they are not affordable by most organizations. Moreover, the process of training and deployment of such models is time-consuming and often expensive to power, and thus contributes to the emission of copious amounts of carbon. This is an important limitation especially with environmental and energy efficiency concerns are on the rise.

6.3 Interpretability and Transparency

The interpretability of such models in the ML context in CV remains as an open problem. First, given the high order of some structures like deep neural networks (DNNs), we cannot recognize how exactly these models make decisions, so they are called "black boxes". Such opacity is undesirable for AI applications that require interpretability, which includes, medical diagnosis, self-driving cars, and Law. We trust these models as they make decisions that often cannot be justified due to lack of interpretability.

6.4 Adversarial attack susceptibility

Despite the development of Convolutional – Neural – Network based CV models, recent discoveries show that such models contain vulnerability to adversarial attacks, wherein the perturbation of pixels within an image can cause the model to output a prediction that is entirely different from the actual one. For instance, an automated car whose line of sight is only its sensors can misread some tiny alterations on the road signs meaning high safety consequences [28]. This weakness poses great threats to CV applications in the security, surveillance, and autonomous applications areas and hence requires better models to be designed.

6.5 The ethical and privacy.

Some of the issues attributable to the deployment of CV systems include ethical and privacy issues include the use Facial recognition, Surveillance, and data sharing. CV applications often use and process personal information and that creates a potential for misuse, bias and privacy violation. Moreover, the learning dataset biases may provide unsuitable treatment of the discriminated demography causing unfair societal treatment. Such ethical considerations and risks explain the need for responsible AI and encourage the need to improve CV models to be fairer and, free from bias and respect people's right to privacy.

Nevertheless, current research and the development of newer technologies suggest new courses to overcome these constraints and make ML in CV more effective, fast, and fair.

6.6 Data-Efficient Learning

According to our analysis, we can expect that future CV research of interest will examine the aspects of self-supervised, unsupervised, and few-shot learning. Such approaches enable a model to rely on data patterns encompassed and inherent structures to perform feature extraction when trained on a minimal set of samples labelled data [29]. Recent techniques such as self-supervised learning, for instance, use big datasets of unlabelled data, and the method produces its own training signals by coming up with fake labels and thus minimizing the need for labelled data. This approach is most beneficial in situations where in specific domains, such as clinical or medical, procuring labelled training data is expensive or scarce.

6.7 Models of Lightweight and Energy-Efficient Cars

When the computational requirements increase, one hears demands for lightweight models designed to fit energy requirements. Decreasing model size through methods including pruning, quantization, and knowledge distillation are taken to minimize the dependence on hardware resources while maintaining model performance. For instance, pruning eliminates small weights in a neural network, and quantization changes a neural network's parameters to a fewer number of bits. Moreover, since on-device processing and edge computing are prominent, mirroring does not necessarily require much cloud-based resource hence increasing efficiency when it comes to real time applications.

6.8 Explainable and Transparent AI

There are rising attempts in improving the explainability in AI for CV applications. Saliency maps, Grad-CAM, and SHAP (Shapley Additive Explanations) describe model decision-making by extracting and displaying important features or regions from inputs. It may be demonstrated that creating models, which in turn may be interpreted by analyzed images, can increase faith in CV applications and their potential application in areas such as healthcare, finance, and security. Thus, the future studies will be mainly dedicated to the idea of designing the complex model architectures and making them genuinely interpretable.

6.9 Adversarial robustness

To mitigate threats posed by adversarial attacks, investigations are increasingly dedicated to improving the models' resilience to manipulation. These include the training of models on the adversarial perturbed data, adversarial training for instance. Moreover, employing defences such as requisite feature extraction, and specialized GAN for adversarial detection could fortify the CV systems. With such applications as autonomous driving and surveillance, being characteristic features of modern technology, such strong defences will be instrumental in safeguarding these applications.

6.10 FAIR Associated initiatives, Fairness, Accountability and Privacy (FAccT)

Mitigating ethical factors will entail CV research to incorporate Fairness, Accountability and P Privacy (FAccT) into modelling. Positive representation of demographics in the training data set and deposition are major ways through which CV systems can be proven to be fair. Other methods such as federated learning, which allows different models to learn from separate datasets, without having to merge the data together, are also able to maintain the privacy of data by keeping it locally on devices instead. Legal requirements also exist through regulations and ethical standards will also come into force that will push developers and researchers to embrace ethical AI across various industries.

6.11 Multimodal Learning

Multimodal learning where data is collected in different form including text, audio, and visuals is still in its early-stage development in CV. As we mentioned earlier, multimodal models can consequently give more detailed and varied descriptions of visual data derived from different sources. For example, in the application of medical diagnosis, an incorporation of an image processing model into patient record analysis model may yield more accurate results. It also finds application in AV or self-driving capability as collecting information from LiDAR, radar, camera, etc, make it safer and gives a better capability in decision making.

6.12 Real Time and Continual Learning

The performance of applications that require an ability to learn and adjust on the fly, on the example of surveillance, self-driving cars or robots. Subsequent generations of CV systems will continue with an enhanced approach of continual learning in which data is updated recurrently without having to be trained all over again. These models will be able to make quick decisions if they are implemented in an environment that which is dynamic because of real-time processing. These capabilities are made feasible with the recent developments in reinforcement learning, online learning, and transfer learning. **Fig 1-6** depicts the usage percentage of various machine learning algorithms, displays the distribution of image counts in different dataset size ranges, illustrates the accuracy improvement of image classification models over time, represents model accuracy across various datasets and tasks, shows classification performance for different object categories and compares training and validation accuracy over epochs respectively.

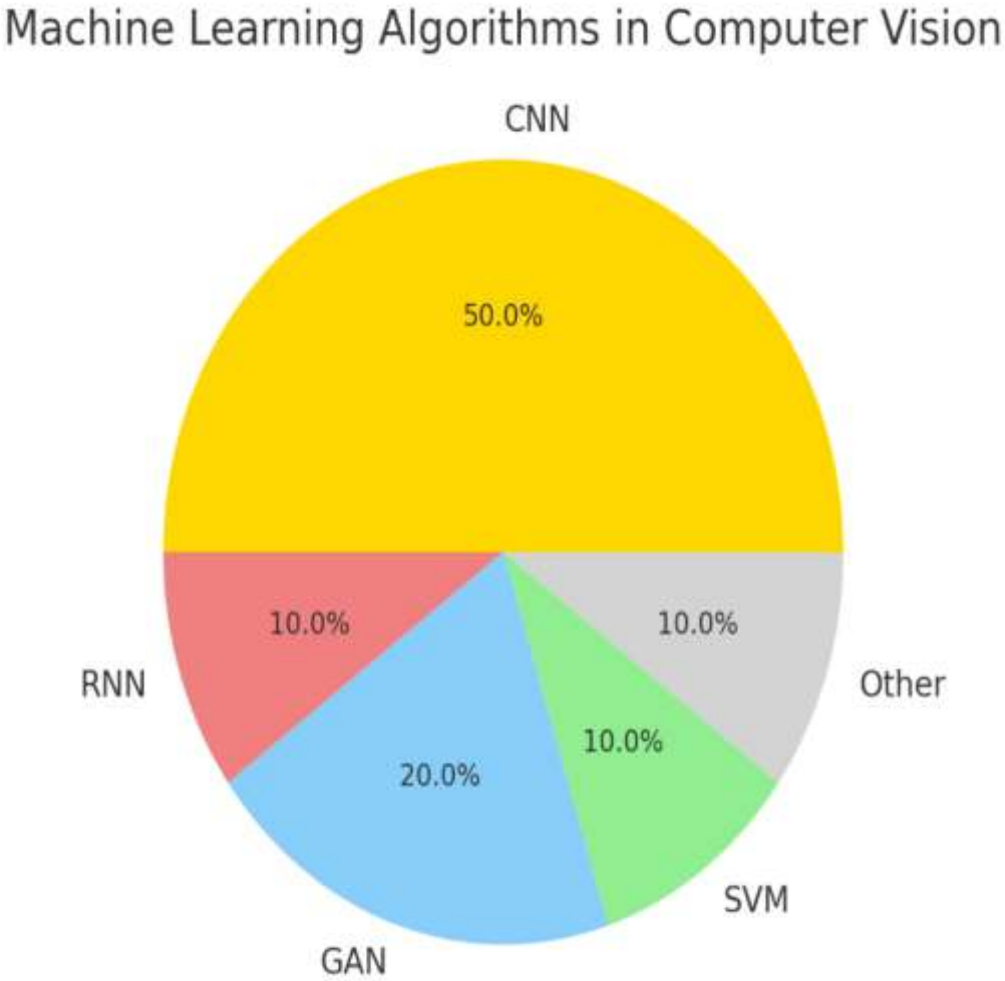

Fig 1 Depicts the usage percentage of various machine learning algorithms.

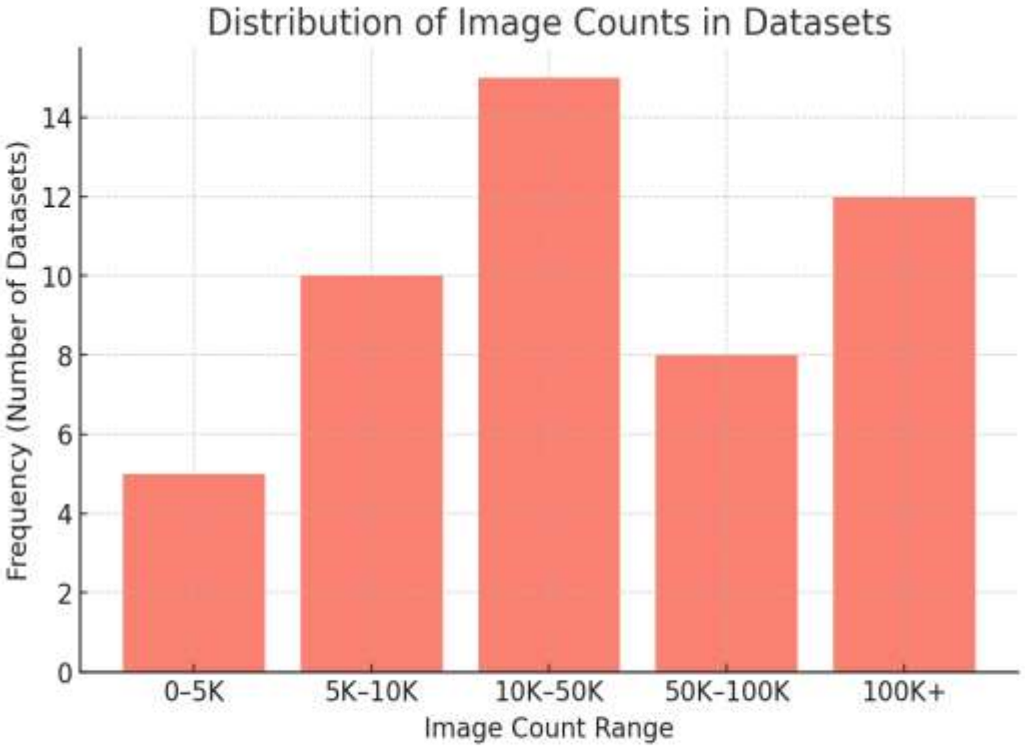

Fig 2 Displays the distribution of image counts in different dataset size ranges.

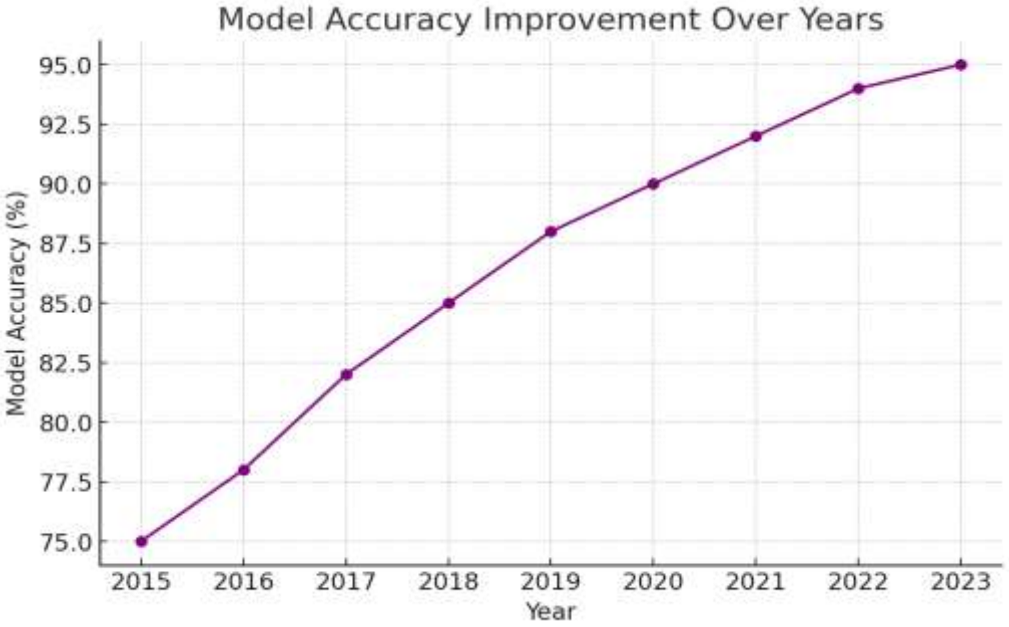

Fig 3 Illustrates the accuracy improvement of image classification models over time.

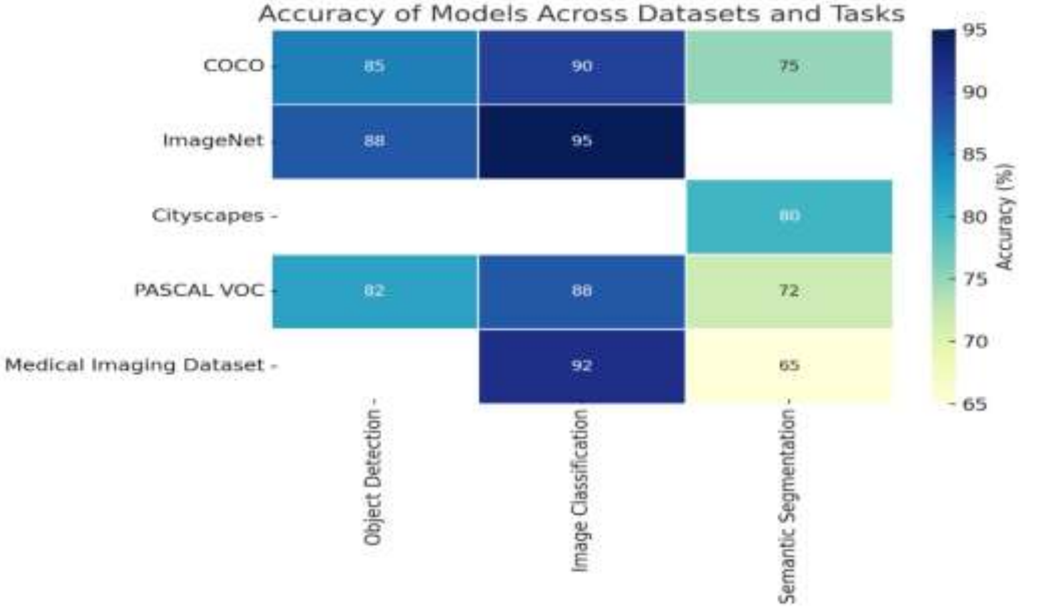

Fig 4 Represents model accuracy across various datasets and tasks.

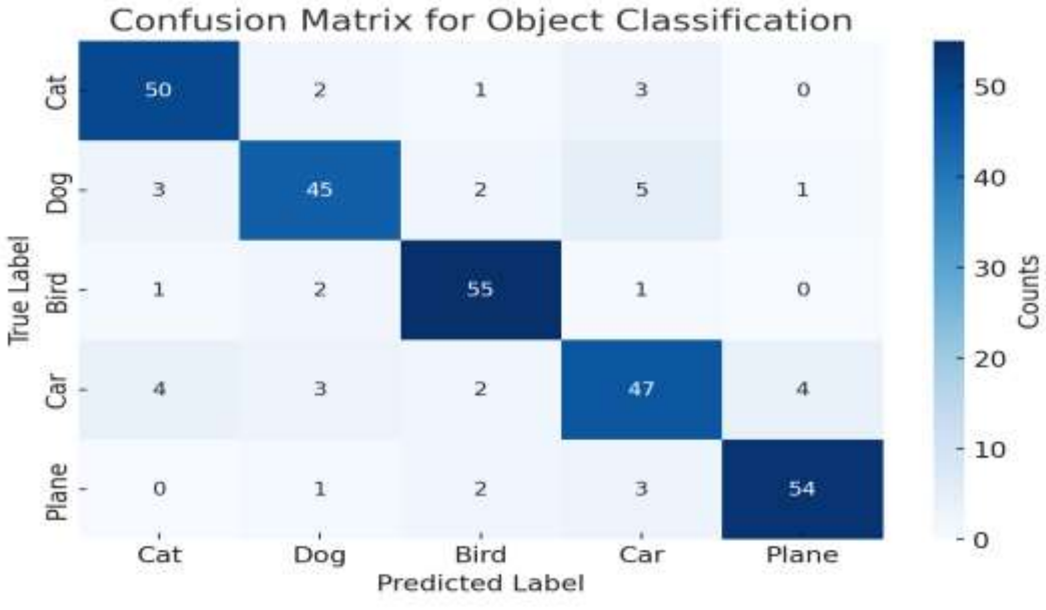

Fig 5 Shows classification performance for different object categories

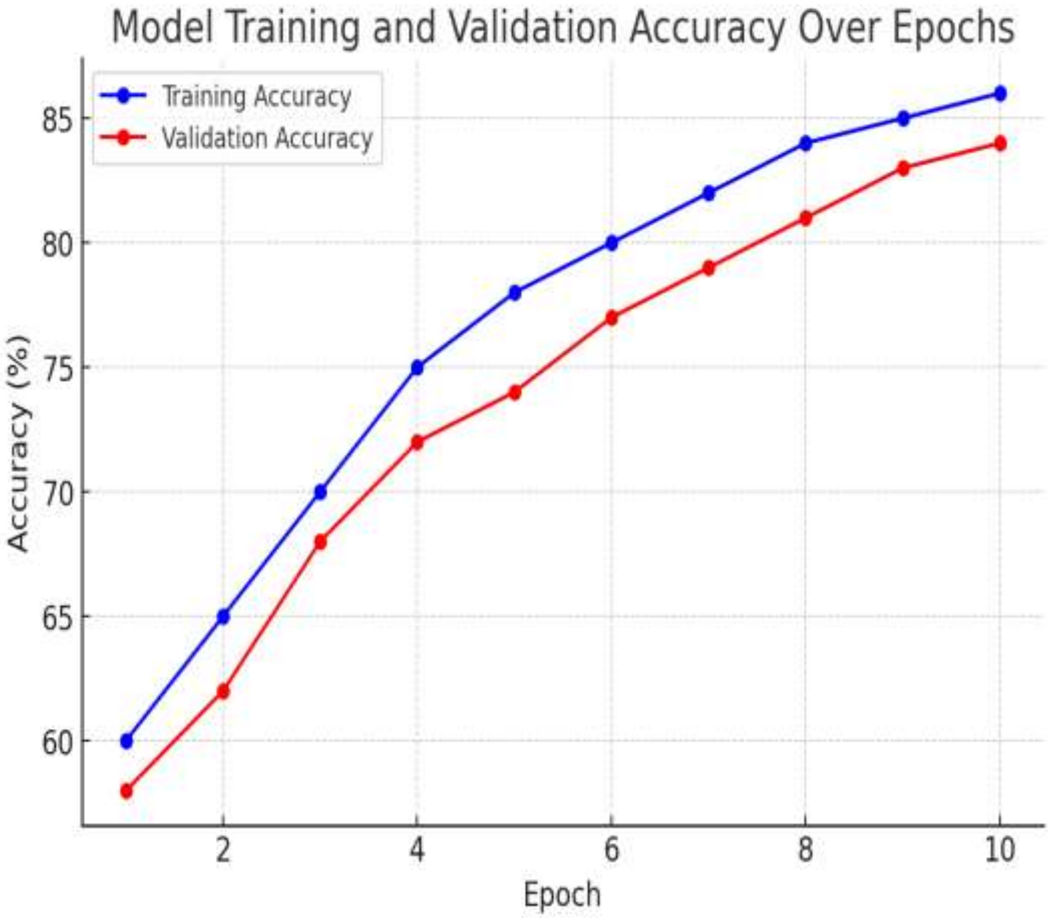

Fig 6 Compares training and validation accuracy over epochs.

7. Conclusion

Computer vision has been enhanced by machine learning to accomplish a plethora of challenging vision-based problems in various sectors. However, there is a vast area of work that requires a lot of computation, large amounts of data, and more crucially, ethics. The work presented here also shows and supports the idea that future developments in computer vision will be composed by more collaborative and secure models of AI that can be explained and extended to a higher number of fields and areas, namely in autonomous systems as well as in the corresponding adoption in personalized healthcare systems.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, (2012), doi:10.1145/3065386.
2. He, K., Zhang, X., Ren, S., Sun, J., "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, (2016), doi:10.1109/CVPR.2016.90.
3. Redmon, J., Divvala, S., Girshick, R., Farhadi, A., "You Only Look Once: Unified, Real-Time Object Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, (2016), doi:10.1109/CVPR.2016.91.
4. Long, J., Shelhamer, E., Darrell, T., "Fully Convolutional Networks for Semantic Segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440, (2015), doi:10.1109/CVPR.2015.7298965.
5. Szegedy, C., Liu, W., Jia, Y., et al., "Going Deeper with Convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, (2015), doi:10.1109/CVPR.2015.7298594.
6. Ren, S., He, K., Girshick, R., Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, issue 6, pp. 1137–1149, (2017), doi:10.1109/TPAMI.2016.2577031.
7. Liu, W., Anguelov, D., Erhan, D., et al., "SSD: Single Shot MultiBox Detector," Proceedings of the European Conference on Computer Vision (ECCV), pp. 21–37, (2016), doi:10.1007/978-3-319-46448-0_2.
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), (2021), arXiv:2010.11929.
9. Parkhi, O. M., Vedaldi, A., Zisserman, A., "Deep Face Recognition," British Machine Vision Conference (BMVC), vol. 1, pp. 6, (2015), doi:10.5244/C.29.41.
10. Deng, J., Dong, W., Socher, R., et al., "ImageNet: A Large-Scale Hierarchical Image Database," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255, (2009), doi:10.1109/CVPR.2009.5206848.
11. Ronneberger, O., Fischer, P., Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241, (2015), doi:10.1007/978-3-319-24574-4_28.
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, (2014), doi:10.5555/2969033.2969125.
13. Radford, A., Metz, L., Chintala, S., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," International Conference on Learning Representations (ICLR), (2016), arXiv:1511.06434.
14. Carion, N., Massa, F., Synnaeve, G., et al., "End-to-End Object Detection with Transformers," European Conference on Computer Vision (ECCV), pp. 213–229, (2020), doi:10.1007/978-3-030-58452-8_13.
15. Zhu, X., Su, W., Lu, L., et al., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," International Conference on Learning Representations (ICLR), (2021), arXiv:2010.04159.
16. Lin, T.-Y., Dollár, P., Girshick, R., et al., "Feature Pyramid Networks for Object Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125, (2017), doi:10.1109/CVPR.2017.106.

17. Lin, T.-Y., Goyal, P., Girshick, R., et al., "Focal Loss for Dense Object Detection," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988, (2017), doi:10.1109/ICCV.2017.324.
18. Howard, A. G., Zhu, M., Chen, B., et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint, (2017), arXiv:1704.04861.
19. Tan, M., Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning (ICML), pp. 6105–6114, (2019), arXiv:1905.11946.
20. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D., "mixup: Beyond Empirical Risk Minimization," International Conference on Learning Representations (ICLR), (2018), arXiv:1710.09412.
21. Girshick, R., "Fast R-CNN," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, (2015), doi:10.1109/ICCV.2015.169.
22. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., "Hypercolumns for Object Segmentation and Fine-Grained Localization," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 447–456, (2015), doi:10.1109/CVPR.2015.7298652.
23. Oord, A. V. D., Kalchbrenner, N., Kavukcuoglu, K., "Pixel Recurrent Neural Networks," International Conference on Machine Learning (ICML), pp. 1747–1756, (2016), arXiv:1601.06759.
24. Xie, S., Tu, Z., "Holistically-Nested Edge Detection," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1395–1403, (2015), doi:10.1109/ICCV.2015.164.
25. Liu, C., Yuen, J., Torralba, A., et al., "Nonparametric Scene Parsing: Label Transfer and Query Expansion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, issue 12, pp. 2368–2382, (2011), doi:10.1109/TPAMI.2011.163.
26. M.Senthil Kumar et al, "Facial emotion recognition using geometrical features based deep learning techniques", International Journal of Computers Communications & Control, Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 18, Issue: 4, Month: August, Year: 2023 DOI: https://doi.org/10.15837/ijccc.2023.4.4644
27. Tiwari, R., Senthil Kumar, M., Diwan, T. D., Pinjarkar, L., Mehta, K., Nayak, H., … Shrivastava, R. (2023). Enhanced Power Quality and Forecasting for PV-Wind Microgrid Using Proactive Shunt Power Filter and Neural Network-Based Time Series Forecasting. Electric Power Components and Systems, 1–15. https://doi.org/10.1080/15325008.2023.2249894
28. S.V. Achuta Rao “Severity of Defect: an Optimized prediction”, International Journal of Advanced Intelligence Paradigms, Vol.13, No.3-4, August 28,2029, PP 334-345; ISSN (print): 1755-0386•ISSN (online),e): https://doi.org/10.1504/IJAIP.2019.101983
29. S.V. Achuta Rao “Software Defect Prediction in Class Level Metric Aggregation Using Data Mining Techniques”, Research Journal of Applied Sciences, Engineering and Technology 13(7): 544-554, 2016 DOI:10.19026/rjaset.13.3014 ISSN: 2040-7459; e-ISSN: 2040-7467 © 2016 Maxwell Scientific Publication Corp. Submitted: March 14, 2016 Accepted: June 25, 2016 Published: October 05, 2016 http://dx.doi.org/10.19026/rjaset.13.3014