Extract Rich Information from Images and Video using Custom Vision Cognitive Services

Amr Elmaghraby
October University of Modern Sciences and Arts
6th of October -Cairo-Egypt

ABSTRACT
Computer vision is a branch of AI that allows computers and systems to extract useful information from digital photos, movies, and other visual inputs in order to address real-world visual problems. Artificial intelligence has an area called machine learning. Machine learning has a subfield called deep learning. Cognitive Services are a set of data-mining-based machine learning methods. Cognitive Machine Learning is a type of artificial intelligence that was created to address issues (AI).Deep Learning (DL)-powered computer vision technology adds real-world benefit to a variety of businesses. Deep learning is the use of neural networks containing more than one hidden layer of neurons to solve problems in domains such as computer vision which are more accurate quality inspectors than humans, make fewer mistakes, and don't mind doing tedious, repetitive duties all day. [1].Cognitive services algorithms are used in a variety of industries to help businesses and improve our daily lives. One of these domains is image classification, which uses convolutional neural networks to help humans discover key components of a picture. The purpose of this paper is to introduce the Microsoft Azure framework's Custom Vision Service. The Azure Custom Vision Service allows you to create, deploy, and develop high image identification modelsand how to make your Custom Vision Service model better. The amount, quality, and variety of labelled data you offer, as well as the entire dataset's balance, determine the quality of the classifier or object detector. A good model will have a well-balanced training dataset that is indicative of the data it will be given. The process of creating such a model is iterative, and it's normal to go through several rounds of training before getting the desired results..Convolutional neural networks, a cutting-edge technology with massive learning capacity, are used in the Custom Vision Service. Because constructing a convolutional neural network is a time-consuming activity that most engineers lack, a Custom Vision Service supplies this component for constructing a classifier. The Custom Vision service analyses photographs using a machine learning algorithm. You can use Custom Vision to create your own labels and train custom models to detect them. Each label denotes a different set of classes or objects. By submit groups of images that have and don't have the characteristics in question. The images have been labeled.at the time of submission. Then the algorithm trains this data and calculates its own accuracy by testing itself on those same images. Train the model by iterating over the entire dataset several times. On the basis of the test results, the model was evaluated. The model can be downloaded and utilized without having to be connected to the internet. Azure Cognitive Services provides a wide range of Artificial Intelligence (AI) solutions. Because the Custom Vision service is optimized for fast detecting significant differences between photographs, we can begin constructing our model with a small amount of data. We'll use 15 images in Custom Vision (the minimum required). Microsoft recommends using at least 50 different images to improve prediction accuracy (with different types of images).The suggested system can handle JPEG images, MPEG-1 bitstreams, and live video inputs. It is also possible to operate the procedures on an individual and autonomous basis. Once the training is complete, the model can be published, and you should be able to access it using the Custom Vision API..Azure Custom Vision's primary goal is to aid in the picture prediction process. The second suggested experiment, will use Java to build an integration with the Video Indexer service in order to improve it even further.

General Terms
Cognitive Services, Azure cognitive Services, Computer Vision, Pattern Recognition, Custom Vision Service.

Keywords
Artificial Intelligence, Big Data, Cloud Computing, Deep Learning, Machine Learning, image classification, Custom Vision Service, convolutional neural network, precision, and recall.

1. INTRODUCTION
The term "intelligence" refers to one's ability to comprehend. It comes from the Latin word intelligere, which means "to understand," which makes sense given that it relates to someone's ability to comprehend information. One of AI's goals was to develop artificial systems that could outperform, if not outperform, human talents. The scientific study of intelligence is divided into several fields.Cognitive Services are a set of machine learning techniques created to solve challenges in the field of AI (AI). The majority of early research focused on mathematical models of intelligence, ignoring the human factor. Cognitive modelling is the process of creating artificial systems that can imitate mental functions and processes using experimental data. Computer vision is an artificial intelligence field in which software systems are developed to visually perceive the world utilizing cameras, pictures, and video [2].The new problem is to get the computer to 'understand' what it's seeing by breaking images down into atomic components and evaluating them using prior indexing. This is accomplished through the machine learning method, which involves training neural networks using millions of images. The problem is that when people and computers look at the same object, they view things differently. A machine perceives an array of pixel values where a human sees an apple (object) (image color data). We employ pixel values as numeric characteristics to train a machine learning model to offer machines a higher-level grasp of what the picture data represents.

2. OBJECTIVES AND SCOPE
The goal of this paper is to give an overview of Custom

Vision Service, a new framework available on Microsoft Azure that allows developers to create something unique and easy to use in their own projects. In an interdisciplinary discipline, it aids developers in the development of image classifiers. Because they work on a data and intelligence model that we give, such as bespoke vision, some cognitive services require machine learning algorithms to be trained beforehand. A good model will have a well-balanced training dataset that is indicative of the data it will be given. Custom Vision in Azure is an image recognition service that allows you to create, deploy, and upgrade your own image identifiers. Others can be contacted simply by sending them requests. Developers can use the Computer Vision API to comprehend the contents of any image [3]. It develops tags to recognize items, people such as celebrities, or actions in a picture, and then constructs meaningful words to explain them. In addition to landmarks and handwriting, the Computer Vision API can recognize them in photographs. The paper focuses on how to increase the quality of your Microsoft Azure Custom Vision Service model. Microsoft Azure is a cloud computing service developed by Microsoft for developing, testing, deploying, and managing applications and services through Microsoft-managed data centers. The study focuses on using a variety of strategies to improve the accuracy of a custom picture categorization or object detector model.

3. PREVIOUS EXPERIENCE

Convolutional neural networks, a form of neural network with a large learning capacity, can be used to create such a computer vision model. They are a significant family of algorithms that have been demonstrated to be state-of-the-art in large object recognition, picture classification, and a variety of other image and natural language processing problems. They were not well received in previous years due to the high cost of applying them to some photos. It takes a long time to train a deep neural network. To train, you'll require dedicated hardware (for example, high-powered GPUs). By uploading an image or supplying an image URL, the Computer Vision API offers developers with access to complex algorithms for processing images and producing information. It analyses visual content in numerous ways based on inputs and user choices. Has the ability to recognize and classify items using a complicated algorithm that executes a whole mathematical operation [4].There are many distinct types of cloud computing service models, each with its own set of business requirements, such as SaaS, PaaS, and IaaS, and deciding which one is best for a company can be difficult. The most crucial decisions are how much you can and want to manage yourself versus how much you want to delegate to your service provider. In terms of who manages what, here's how infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) compare.

Infrastructure as a Service (IaaS) is a concept that refers to infrastructure as a service and pay-as-you-need services that provide basic processing, storage, and networking capabilities on demand. Microsoft Azure, Digital Ocean, Linode, Rackspace, Amazon Web Services (AWS), Cisco Metapod, and Google Compute Engine are just a few examples (GCE)

PaaS is a complete cloud development and deployment environment with resources that enable you to produce everything from simple cloud-based apps to sophisticated, cloud-enabled business systems. Users pay for the resources they require from a cloud serviceprovideover a secure Internet connection,developer manage the applications, dataand the cloud service provider typically manage everything else. Windows Azure, AWS Elastic Beanstalk, Heroku, Force.com, Google App Engine, Apache Stratos, and OpenShift are some examples.

PaaS, like IaaS, contains infrastructureservers, storage, and networkingas well as middleware, development tools, Business Intelligence (BI) services, database management systems, and other services. PaaS is intended to support the entire web application lifecycle, including development, testing, deployment, management, and update.

SaaS is a comprehensive software solution that users can acquire from a cloud service provider on a pay-as-you-go basis. Employees rent the usage of an app for their enterprise, and they connect to it via the Internet, usually through a web browser. The service provider's data center houses all of the underlying infrastructure, middleware, app software, and app data. The service provider is in charge of the hardware and software, and with the right service agreement, they will ensure the app's availability and security, as well as the protection of your data. With SaaS, your company can quickly get an app up and operating for a low upfront cost.Common Examples: Common examples are email, calendaring, and office tools (Microsoft Office 365, Google Applications (G Suite)), Zoom , Salesforce, Dropbox, ZenDesk, Slack, Hubspot, Shopify,Netflix. ,Cisco WebEx, Concur, GoToMeeting. Developers can construct highly available systems on Linux or Windows hosts using Microsoft Azure or Google Cloud. While both systems have similar capabilities, the resources that support those capabilities are frequently arranged differently since Microsoft Azure and Google Cloud evolved their capabilities independently over time, resulting in significant implementation and design variations. Innovative machine learning tools and services on a secure platform are now available in Azure, a Google Cloud service that allows you to train, deploy, automate, and manage machine learning models in the cloud. Now it is become Easy-to-deploy and automatically configured third-party applications, including single virtual machine or multiple virtual machine solutions

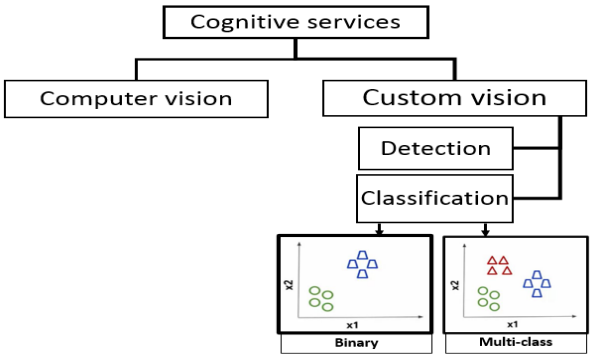


Figure 1: Hierarchical model of cognitive services

Custom Vision gives us the ability to create our own labels and train custom models to recognize them. There are two types of Custom Vision functionality.

1- An image classification model is a form of Supervised Machine Learning technique that uses some training to categories data. The Classification model assigns one or more labels to an image and categorizes it into one of several groups. You create image features for the entire image (using classic or deep learning methods). These characteristics are image aggregates that give an image one or more labels (names).

2- The image classification model is comparable to the object detection model. It gives coordinates of the images where the specific labelled image is found by doing this on a more fine-grained, granular, regional level of the image. For example, we can use it to count the number of instances of an object by doing this on a more fine-grained, granular, regional level of the image.

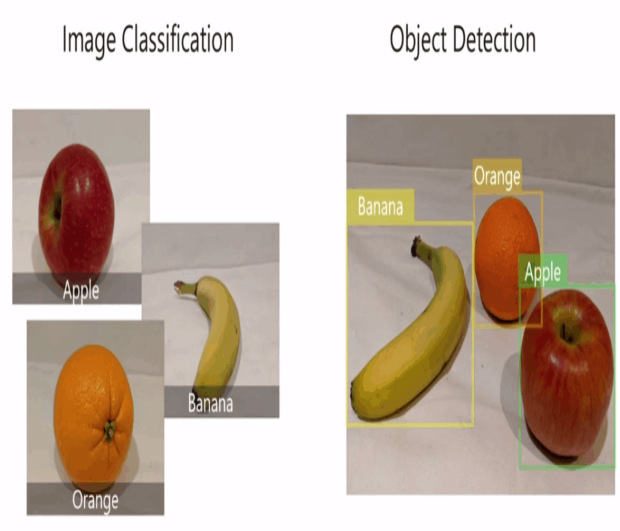


Figure 2: Image Classification vs. Object Detection

Labeled data is required to train a custom model. Images with appropriate bounding box coordinates and labels are labelled data in the context of object detection. That is, the (x,y) coordinates at the bottom left and top right, plus the class. Figure(2) the Banana bounding box coordinates, for example, are [0.1, 0.44, 0.34, 0.56] and Applebounding box [0.72, 0.57, 0.87, 0.77]. This model works like a pattern-detection algorithm, probabilistically converting computer-friendly characteristics (pixel values) to human-friendly labels (objects, attributes). When we feed this model an image, it can now predict an appropriate label along with a confidence value.

Object detection works similarly to tagging, except that the API gives the bounding box coordinates (in pixels) for each object discovered. If an image contains a banana, apple, orange, for example, the Detect operation will list those objects in the image along with their positions. This feature can be used to process the relationships between objects in an image. It also allows you to see if a picture contains several instances of the same tag. The tags are applied by the Detect API based on the objects or live things found in the image. The tagging taxonomy and the object detection taxonomy do not yet have a defined link.

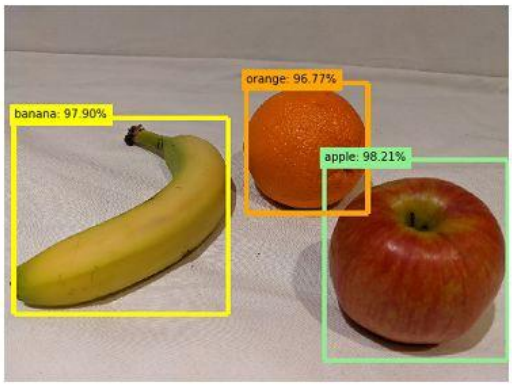


Figure 3:Object detection bounding boxes

We need to correlate photographs with tags in Custom Vision Service, such as the type of clouds in an image. After you've done this for enough photographs, you'll use a machine learning method to teach the Cognitive Service to recognize the different types of items in each image [5]. It will recognize the types of items in photographs that you feed it once it has been taught, for example (fruit types). In this example, we utilize Smart Labeler to generate suggested tags for images. When training a Custom Vision model, this allows us to classify a huge number of images more quickly[19].

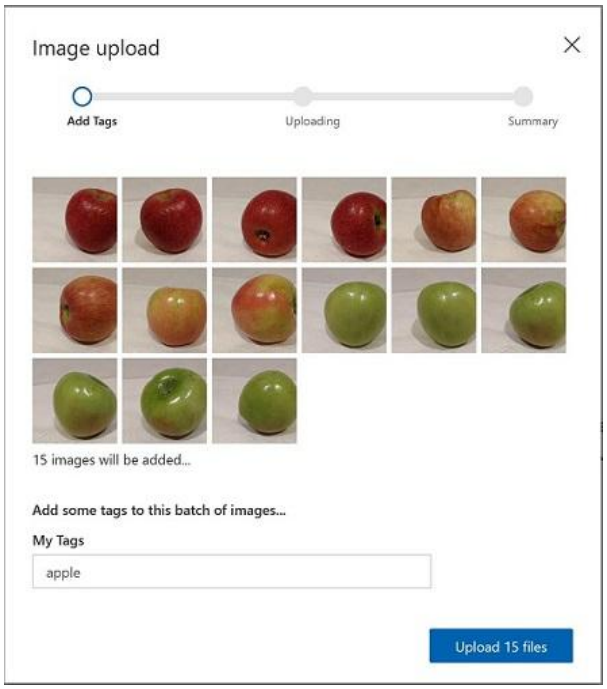


Figure 4: smarter tag

Custom Vision's main purpose is to establish unique image identifiers. customize the model to match your needs. Assume you want a service that can tell the difference between a banana and an apple. You name various photographs of Apple with the word 'apple,' and the same goes for the word 'banana' After training, you can upload a fresh image that the model hasn't seen before, and you'll get feedback on whether the image accurately depicts a banana or an apple.

Object detection works similarly to tagging, except that the API gives the bounding box coordinates (in pixels) for each object discovered. If an image contains a banana, an apple, and a person, the Detect operation will list those things in the image along with their locations. This feature can be used to

process the relationships between objects in an image. It also allows you to see if a picture contains several instances of the same tag.The Detect API only discovers objects and living things on a conceptual level, whereas the Tag API can contain contextual phrases like "indoor," which cannot be localized with bounding boxes.

The tags are applied by the Detect API based on the objects or live things found in the image. The tagging taxonomy and the object detection taxonomy do not yet have a defined link.

Figure(4) show how to use Smart Labeler by select all of the files in the training-images/apple folder by clicking Add images. Then, specifying the tag apple, upload the image files.

To upload the images in the banana folder with the tag banana, and the images in the orange folder with the tag orange, repeat the previous step.

Examine the images you've submitted to the Custom Vision project - each class should include 15 images.You may also create an area by dragging around the object. Add a new tag with the appropriate item type (apple, banana, or orange) once the region has encircled the object, as seen here:

You may also construct an area by just dragging around the object. Add a new tag with the proper item type (apple, banana, or orange) when the region surrounds the object, as illustrated here:

Select and tag each other object in the image, resizing sections as needed and adding new tags.

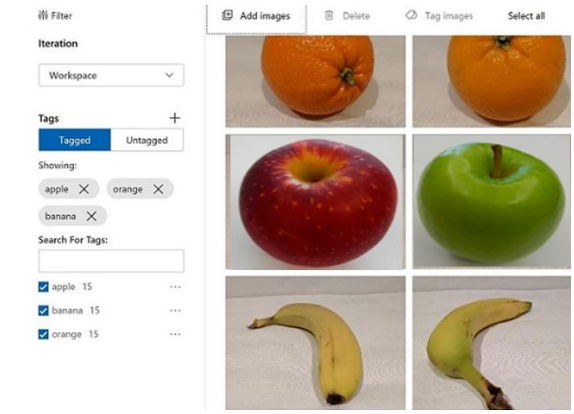


Figure 5:Smart Labeler workflow

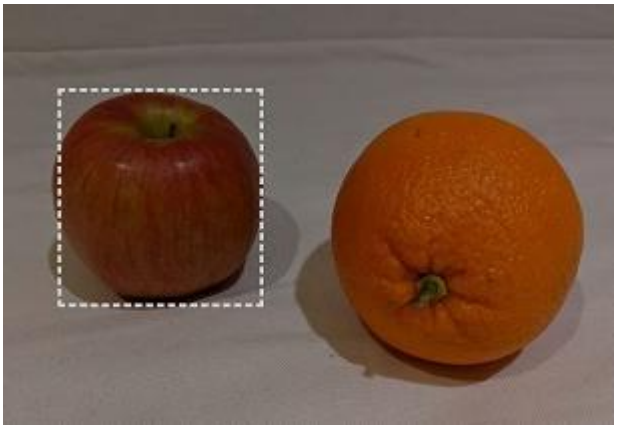


Figure 6: indicate bounding boxes for each object

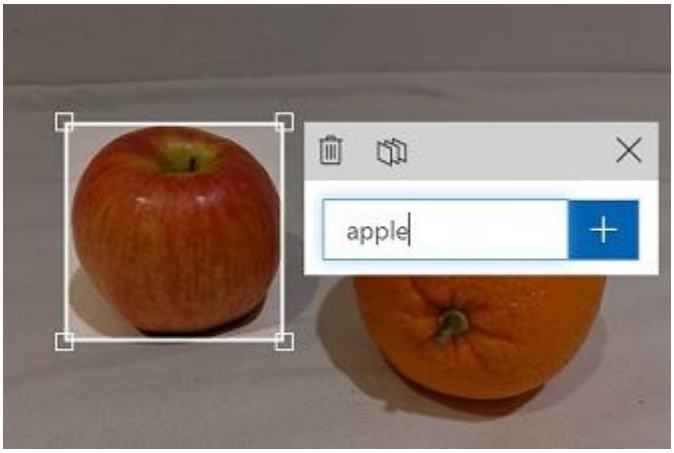


Figure 7: Add a new tag with the appropriate Apple object

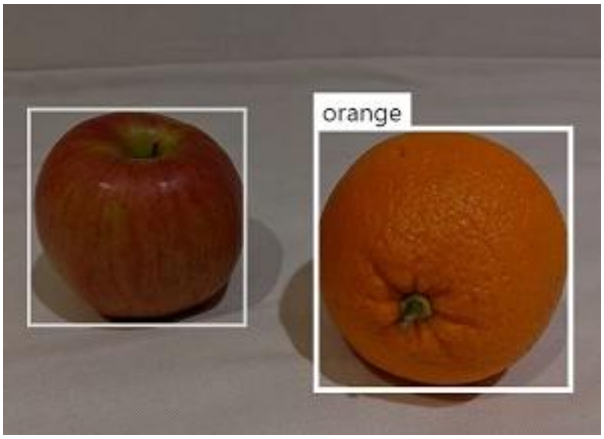


Figure 8:Add a new tag with the appropriate Orange object

4. RELATED LITERATURES

To improve the contrast of grayscale photographs, one publication proposes using adaptive region-based histogram equalization [6]. The "Yale B" database, the "Extended Yale B" database, and the "CMU-PIE" dataset, all of which contain a high number of single-sided illumination and low intensity face images, are used in the study. To classify images, the paper employs the "Euclidean distance nearest-neighbor" classifier. Experiments have shown that adopting an adaptive region-based histogram equalization technique can enhance image categorization accuracy to 100%.Convolutional Neural Networks (CNN) have shown to be particularly effective in natural image classification systems. Further application of such systems to medical picture categorization offers a lot of promise, especially given the inherent nature of medical images, which makes them excellent for deep-learning..The "Euclidean distance nearest-neighbor" classifier, on the other hand, performs less well in picture classification than the CNN model. This thesis also attempted to experiment with the datasets, but because the CNN model performed exceptionally well on the original dataset and can obtain 100% accuracy, it cannot be concluded that image enhancement can help CNN perform better on these datasets. When CNN is used for natural scene character recognition, studies have shown that employing bimodal image augmentation can considerably improve recognition accuracy [7]. The accuracy of image recognition will be affected by various bimodal image improvements. The usage of edge systems combined with machine learning approaches offers a way to improve current procedures by providing speedier tools to aid inspection

diagnostics[8].In addition, employing image enhancement functions with the Laplace operator has been found to increase R-CNN and fast R-CNN performance in pedestrian identification tasks [9]. The Laplace operator can increase the detection rate of 2% and 1% in the two R-CNN models respectively. The paper's experiments used transfer learning, fine-tuning a pre-trained R-CNN model. However, the experiment did not use cross validation and did not use hypothesis testing, which indicates that there is a high risk in the conclusion of the experiment. there is also a difference between target detection and image recognition.

4.1 Performance Metric

Modelling, evaluation methods are critical. Standard metrics for evaluating classification predictive models, such as classification accuracy and classification error, are frequently employed.We all desire a detection system that is 100 percent accurate. but developing a more realistic model would be prohibitively expensive. The heading of subsections should be when utilizing categorization to construct a model to predict events. As demonstrated in the accompanying display, we create a confusion matrix for the model based on test data.A confusion matrix is a summary of classification problem prediction outcomes. The number of correct and bad predictions is totaled and broken down by class using count values. The confusion matrix's key is this. The confusion matrix depicts the various ways in which your classification model becomes perplexed when making predictions. It reveals not just the number of errors made by your classifier, but also the types of errors that are being made. This breakdown addresses the drawback of relying solely on classification accuracy [24].

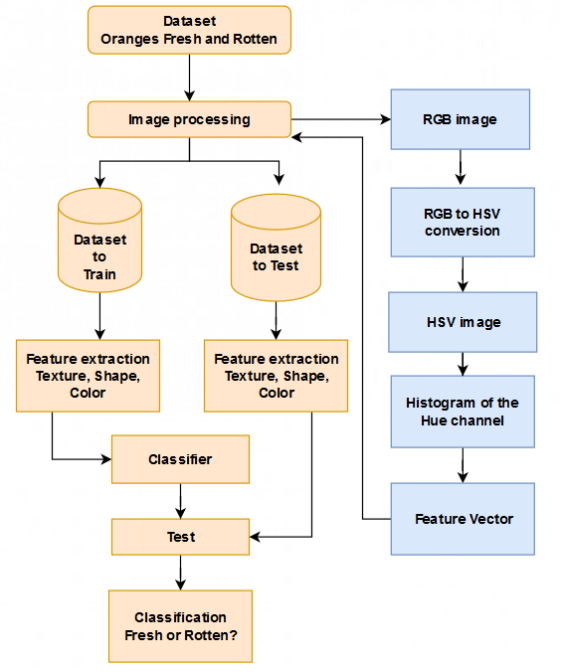


Figure 9: Computer-Vision-based machine learning algorithm creation process.

4.2 What is a confusion Matrix and how calculate it?

The steps for calculating a confusion matrix are outlined below.

1. A test or validation dataset with expected outcome values is required.

2. For each row in your test dataset, make a prediction.

3. Count the following from the expected outcomes and predictions count:

a) The number of correct predictions for each class.
b) The number of incorrect predictions for each class, organized by the class that was predicted.

These numbers are then organized into a table or a matrix as follows:

a) Expected down the side: Each row of the matrix corresponds to a predicted class.
b) Predicted across the top: Each column of the matrix corresponds to an actual class.

The counts of correct and incorrect classification are then filled into the table. The total number of correct predictions for a class is recorded in the expected row and predicted column for that class value [10]. The total number of inaccurate predictions for a class is entered into the expected row and predicted column for that class value.

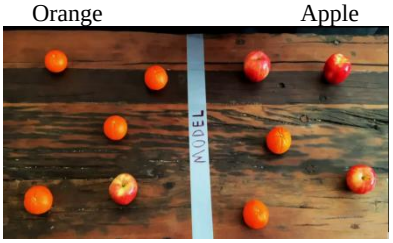


Figure 10: summary of prediction results on a classification problem.

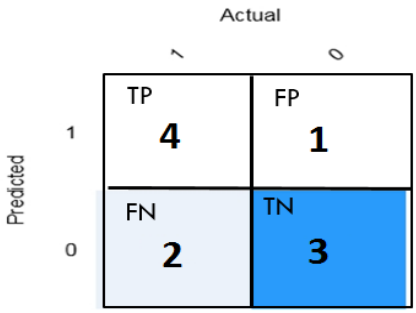


Figure 11: Confusion Matrix result

- TP:"true positive" for correctly predicted event values.
- FP:"false positive" for incorrectly predicted event values.
- TN:"true negative" for correctly predicted no-event values.
- FN:"false negative" for incorrectly predicted no-event values.

Accuracy and error rate are the most prevalent and relevant metrics in all classification issues, whether they are two-category or multi-category tasks. The accuracy ratio is the number of correctly classified samples divided by the total number of samples, whereas the error rate is the proportion of samples identified incorrectly. The number of correctly identified samples divided by the total number of samples is known as accuracy[14].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

The error rate and accuracy, on the other hand, do not meet all of the task criteria, and alternative metrics are needed for different activities. For example, in a system for capturing

fruit, it is important to recognize photos of the orange in order to capture an orange. Because the system does not need to catch all of the oranges, it is critical that as many as feasible are caught among the fruit caught. The image recognition system is now mainly concerned with "the proportion of the actual orange to all the fruit that are called orange," or the "Precision" rating.The formula for Precision's calculating method is as follows: True Positive is the number of correctly identified oranges, whereas False Positive is the number of fruits that are not orange but are mistaken for oranges[15].

Accuracy for orange = total correct/total observable

Accuracy for orange =(4+3)/10 =7/10 = 70%

Form Figure (12)the Accuracy for Apple = total correct/total observable Accuracy for Apple =(1+2)/10 = 3/10 = 30%

Accuracy with imbalanced classes cause (problem) .The main problem with classification accuracy is that it hides the detail you need to better understand the performance of your classification model. The examples describes where used to encounter this problem. Imbalanceclass is occurred when some datasets, considerably more instances in some classes than others. If the imbalance in the training set is not represented in the real data stream, machine learning classification can suffer from poor average precision.Imbalance class leads to believe that model is better than it really is.

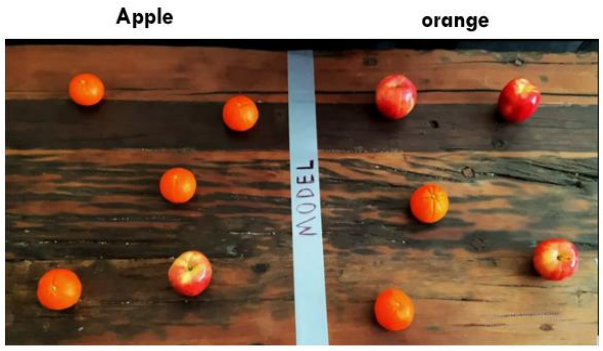


Figure 12: Machine Learning Model for Apple and orange prediction

		Actual	
		Apple	Orange
predicted	Apple	1	4
	orange	3	2

Figure 13:Confusion Matrix result

1. When there are more than two classes in your data. With three or more classes, you might receive a classification accuracy of 90%, but you don't know if that's because all classes are predicted equally well or if the model is overlooking one or two classes[16].

2. When the number of classes in your data is not even. You may reach a score of 99 percent or more, however this isn't a good result if 990 records out of 1000 belong to one class, and you get this number by always guessing the most common class value.Classification accuracy can hide the detail you need to diagnose the performance of your model. But thankfully we can tease apart this detail by using a confusion matrix. for Example: consider we have 990 orange and 10 Apple The predicted System give high accuracy and say everything is orange

		Actual	
		Orange	Apple
predicted	Orange	990	10
	Apple	0	0

Figure 14: imbalanced classes

Accuracy= total correct/total observable

Accuracy=990/1000 = 99 %

The main problem is that Apple has been completely misclassified.Accuracy is no longer a valid measure in the context of imbalanced data sets since it does not distinguish between the numbers of correctly identified samples of various classes. As a result, it may lead to incorrect conclusions.Properties including dataset size, label noise, and data dispersion add to the difficulty of imbalanced classification.

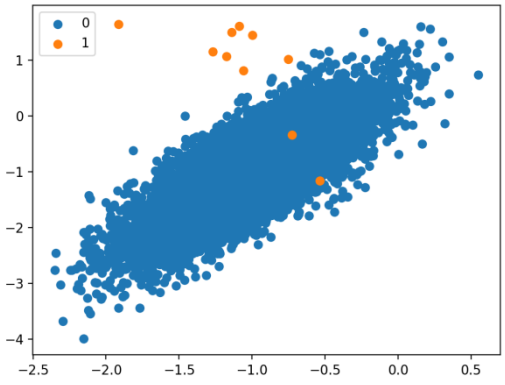


Figure 15: Imbalanced Classification

4.3 Precision , Recall ,F1 score

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.

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

"The proportion of correctly identified objects in all object cases", that is, the recall rate. Recall's calculation method is shown in the formula, where “True Positive” means the number of samples that are correctly detected, and “False Negative” means the number of samples of diseases that have not been detected.

Recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score :The harmonic mean of precision and recall, also known as the classic F-measure or balanced F-score, is a measure that combines precision and recall:

When the two numbers are near, this measure is about the average of the two, and it is more generally the harmonic mean, which coincides with the square of the geometric mean divided by the arithmetic mean in the case of two integers. Due to its bias as an evaluation tool, the F-score might be criticized for a variety of reasons in specific situations.

Precision and Recall are equivalent in the unbiased two-class scheme. In this situation, the F1 score can be utilized as a benchmark for measuring performance. The formula demonstrates the F1 score calculating process.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

According to confusion matrix

predicted \ Actual	Orange	Apple
Orange	4	1
Apple	2	3

Figure 16: Confusion matrix result

$$\text{Precision} = \frac{TP}{TP+FP} = 4/(4+1) = 4/5 = 80\%$$

$$\text{RECALL} = \frac{TP}{TP+FN} = 4/(4+2) = 4/6 = 66\%$$

$$\text{F1} = \frac{2 \times precision \times recall}{precision + recall} = 2*.8*.66/(.8+.66) = 72\%$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} = (4+3)/10 = 70\ \%$$

$$\text{Efficiency} = \frac{TN}{TN+FF} = 3/(3+1) = 75\ \%$$

predicted \ Actual	Apple	Orange
Apple	TP 1	FP 4
orange	FN 3	TN 2

Figure 17: Confusion matrix result

$$\text{Precision} = \frac{TP}{TP+FP} = 1/(1+4) = 1/5 = 20\%$$

$$\text{RECALL} = \frac{TP}{TP+FN} = 1/(1+3) = 1/4 = 25\%$$

$$\text{F1} = \frac{2 \times precision \times recall}{precision + recall} = 2*.2*.25/(.2+.025) = 44\%$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} = (1+2)/10 = 30\ \%$$

$$\text{Efficiency} = \frac{TN}{TN+FF} = 2/(2+4) = 33\ \%$$

Probability threshold

This is the level of confidence required for a prediction to be judged correct (for the purposes of calculating precision and recall).When using a high probability threshold to interpret prediction calls, they tend to produce findings with great precision at the expense of recall—the detected categories are correct, but many are missed. A low probability threshold has the reverse effect: the majority of true classifications are recognized, but there are more false positives in the set. The thresholds are plotted in increasing order on a line plot with recall on the x-axis and accuracy on the y-axis.

5. PROPOSED TRAINING MODEL

How to Improve Custom Vision Model is the topic of the proposed research and the ways to increase the Azure Custom Vision Service model's quality to aid in the development of a more accurate model. The amount, quality, and variety of labelled data offered, as well as the entire dataset's balance, determine the quality of classifier or object detector. A good model will have a well-balanced training dataset that is indicative of the data it will be given. The process of creating such a model is iterative, and it's normal to go through several rounds of training before getting the desired results.

5.1 Data quality

The quantity of data has an impact on AI machine vision performance and can be improved.Machine vision based on AI can only operate if you have a large, high-quality dataset from which to detect predictable patterns. Although the number as well as the quality of your data are critical.Because of the product's inherent variability in shape and the large array of things that can be acquired in the field, a food classification application may require more data to determine if it is viewing a good or bad Apple. However, because there is less diversity in the inspected object and the types of faults, checking a metal bar for defects may not necessitate such a large dataset.The quality of the data is just as crucial.Badly annotated datasets, inconsistent data, defective recordings, and other factors all contribute to poor machine vision performance. Because a neural network requires images to be of a defined sizeA common image size to feed into an object detector is 512 x 512 pixels or smaller., data cleansing for image classification is critical. There are a variety of methods for standardizing photographs, and it's crucial to note that none of them is inherently better or worse. Each has its own set of disadvantages and applications. Improving the quality of your data can help you reduce system latency by reducing complexity. The amount of training photos in your dataset is the most essential component. It's difficult to say how well deep learning models will perform if you don't have a lot of data. As a starting point, we recommend utilizing at least 50 photos per label. We use this data to predict classification accuracy at a given training sample size using the learning curve approach..There's a bigger risk of overfitting with fewer photos, and while your performance figures may indicate

good quality, your model may struggle with real-world data.

5.2 Prohibit Overfitting

Overfitting happens when a model's variance is high.The model fails to generalize the model's learning because it catches the noise in the training datai.e., the model performs well on training data but not properly on the evaluation set.The model will learn to generate predictions based on arbitrary features shared by your photographs and memorize data patterns in the training dataset, but it will fail to generalize to new samples.Overfitting happens when:

1- the data used for training is not cleansed and contains junk values, overfitting occurs.
2- The model has a lot of variation.
3- The training data is insufficient, and the model is forced to train on it for numerous epochs.
4- Deep neural networks are complex and several neural layers are placed together in the model's design.

Making a classifier for apples vs. citrus and utilize photos of apples in hands and citrus on white plates, the classifier can place too much emphasis on hands vs. plates instead of apples vs. citrus. Unexpected classification image Provide photographs with varied angles, backdrops, object sizes, groupings, and other modifications to solve this challenge.

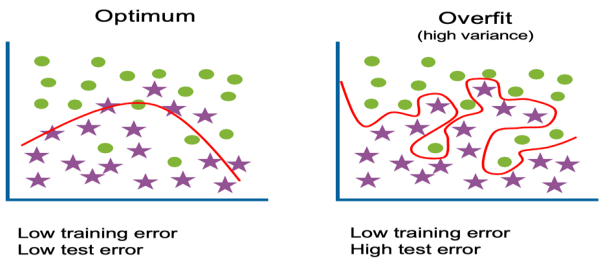
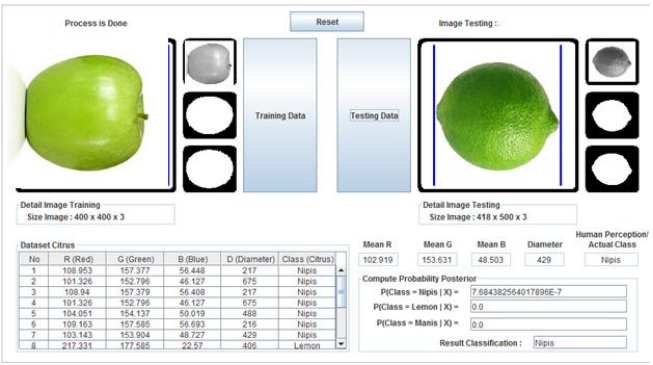


Figure 18: Optimum and Overfitting in Machine Learning model



5.3 Balance training dataset

It's also important to think about the relative sizes of your training data. A training dataset with 990 images for one label Apple and 10 images for another label orange, for example, is unbalanced. As a result, the model will be better at predicting one label than another. Maintaining a 1:2 ratio between the label with the fewest images and the label with the most images is likely to yield better outcomes. If the label with the most images has 1000, the label with the fewest images should have at least 500 for training. During testing, it was discovered that the model accurately identified all of the apples in the images, but not the oranges. You notice that there are 10,000 photographs of apples in your training dataset but only 100 pictures of oranges. This is an example of data imbalance, where datasets for each object class do not have a proportionate amount of instances. To limit the possibility of biased outputs, it's critical that we pay special attention to training data. It's also crucial to think about the relative sizes of your training data.

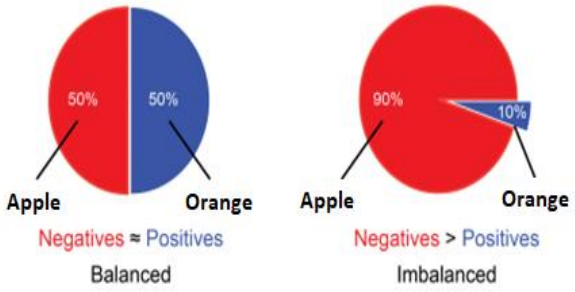


Figure 19:Example of balanced and imbalanced data

Visual inspection can reveal trends that can be corrected by adding new training data or altering existing data. A classifier for apples vs. limes, for example, might mistakenly categorize all green apples as limes. Then, by adding and giving training data that includes labelled images of green apples, you may correct the problem[11].

5.4 Best practice

A digital camera, a light diffusion chamber, a distance adjustment pedestal, and a personal computer make up the created system. For further investigation, the digital camera's images were saved as (RGB). The photos were taken in three directions for a total of 50 samples of each cultivar[17]. To make our training more diverse, we did the following setup:

5.4.1 Background: Include photographs of your thing against a variety of backgrounds. Photos taken in natural settings are preferable than photos taken in front of neutral backgrounds because they give the classifier more information.[18]

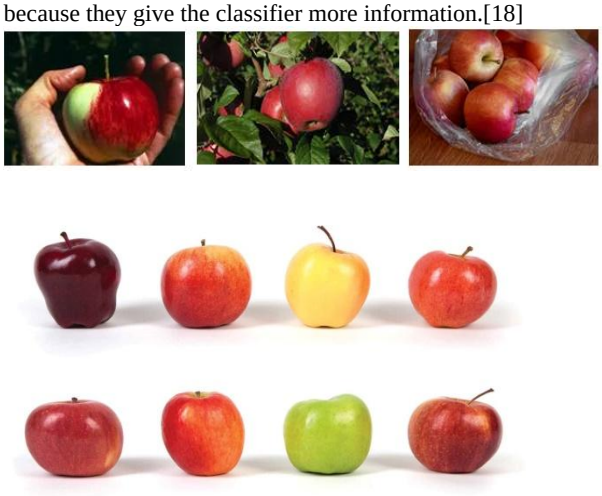


Figure 20:image of background Samples

5.4.2 Lighting: Lighting is crucial for the development of vision systems for industrial inspection applications such as measuring, grading, sorting, monitoring, and control. It is critical to acquire knowledge about the real world that portrays things in various lighting conditions for computer vision. The main lighting aims are to create a uniform light area across the field of vision and to use as many grey levels as feasible to differentiate the feature or portion under scrutiny from the surrounding background. Provide

photographs with varying lighting (e.g., flash, high exposure, etc.), especially if the images used for prediction contain varying lighting. An orange may appear red or yellow through a filter, but it is still orange. Under a red light, a red object becomes practically invisible [16].

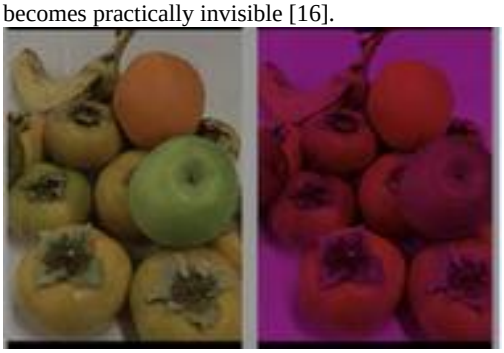


Figure 21:red light effect on red object

Figure 22: lighting object effect

5.4.3 Camera Perspectives: Include photos captured from various camera angles. If all of your photographs must be taken using fixed cameras (such as surveillance cameras), make sure to give each regularly-occurring object a different label to avoid overfitting—interpretation of irrelevant objects (such as lampposts) as the main feature.

Figure 23:Camera Perspectives effect

5.4.4 Object Size: Include photos with a variety of object sizes and numbers (for example, a photo of bunches of bananas and a close-up of a single banana). Different sizing aids the classifier's ability to generalize

Figure 24: Object size effect

5.4.5 Differentstyles: Provide images of various styles within the same class (for example, different varieties of the same fruit). If you have objects with dramatically different styles (for example, the Apple logo vs. a real-life Apple), we recommend labelling them as separate classes to properly convey their respective properties. Recommend to label them as separate classes to better represent their distinct features.

Figure 25: different styles

5.4.6 Resolution is an important factor to consider when selecting a machine vision camera for your application. The level of detail with which an image can be reproduced or recorded is known as resolution. It is critical to have a vision camera with enough resolutions. Resolution determines how detailed an image is. The following are some of the factors that influence resolution: The number of pixels in the image sensor, as well as the quality of the optics that map the image to the sensor. The pixel size shrinks as the overall number of pixels on an image sensor grows, necessitating a higher-quality lens to obtain the best focus [12].

5.5 Model design and evaluation

The proposed Custom Vision service analyses photographs using a machine learning algorithm. We submit groupings of photographs with and without the desired features. At the time of submission, we label the photographs. The algorithm then uses this information to train itself and compute its own accuracy by testing itself on the same photos. We can test, retrain, and eventually use the algorithm in your image recognition software to identify images once we've taught it. We can also download the model to use offline.varying sizes can be found in the photographs in the dataset [13]. The photographs do not have a uniform background. Different poses of the same sorts of fruits can be found in the dataset. Fruits are shown in a variety of postures and angles, including top view, side view, various backgrounds, half cut, sliced on the dish, chopped into pieces, half-eaten, showing the seed, and partially occluded. Fruits might be fresh, rotting, or packaged in bundles.Some photographs have low lighting, various light effects, are covered with snow or net, are decorated, painted, and have leaves on trees.

6. FIGURES/CAPTIONS

6.1 Dataset description and Create a vision AI project

The datasets Fruits 360 from Kaggle are used in this suggested model. Dataset attributes for a dataset of 90380 photos of 131 fruits and vegetables. There are 90483 photos in all.67692 photos in the training set (one fruit or vegetable per image).22688 photos in the test set (one fruit or vegetable per image).There are 131 classes in all (fruits and vegetables).100x100 pixels is the image size.The following are the steps to model and predict this problem:

1. Create a Custom Vision AI project
2. Add Images to the project

3. Labeling and tagging
4. Train the classifier and create the model
5. Evaluate the classifier and measure performance
6. Publish the model and expose the endpoint for use by other clients
7. Use the exposed endpoint and predict using new images

6.2 Create a Custom Vision AI project

In this practical experiment, I'll create a vision AI model that recognizes cars in video for a workplace safety issue.based on Image classification: Does not draw bounding boxes and analyses the entire frame as a picture. Per frame, only one item can be identified. It's simple to learn.

Object detection: Detects several items in a single video frame and draws bounding boxes around them. It takes a little longer to train because the object must be identified each uploaded image.

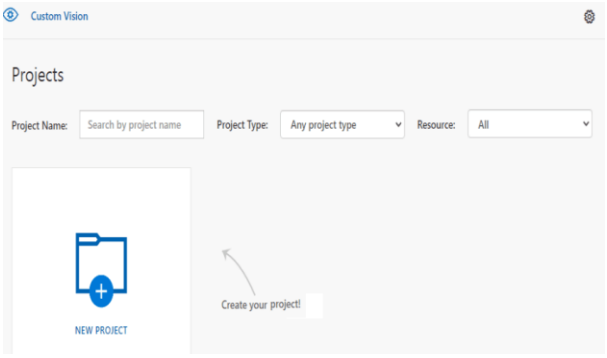


Figure 26: Create a Custom Vision AI project

6.3 Add images into the project

Select Add images and then browse local files to add images. To begin tagging, select Open. Because your tag selection will be applied to the full group of photographs you've chosen to upload, it's better to submit images in separate groups based on the tags they've been assigned. Individual photographs can also have their tags changed after they've been submitted. photos uploaded to the model

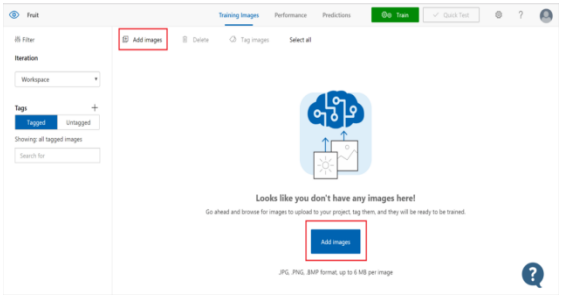


Figure 27:Image Tagging Automation

6.4 Tagging and labelling

We usually categorize objects and notice distinct patterns when we think of images. Convolutional neural networks will decompose images into numbers and process them. It compares the components of these numbers, known as features, and tries to discover a match in the images being compared at about the same positions. Enter text in the My Tags field and press Enter to create a tag. A dropdown selection will show if the tag already exists. You can add many tags to your photographs in a multilabel project, but only one in a multiclass project. Use the Upload [number] files button to finish uploading the photographs. set the appropriate tag (s) Apple, banana, and orange.

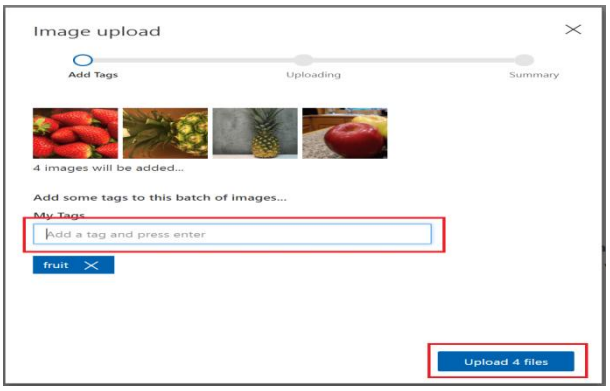


Figure 28:Data Labeling for Computer Vision

6.5 Create train the classifier

It takes time to train on the dataset. After that, we can examine the categorization metrics such as Precision, Recall, and Accuracy. Select the Train button to begin training the classifier. The classifier creates a model that detects the visual features of each tag by using all of the current photos. You establish a fresh iteration with updated performance metrics each time you train your classifier. To get perfect precisions, a great deal of training is required. Any photos that are uniquely associated with an iteration are deleted when you remove it. Select the image, then the tag, and then Save and Close to add it to the training data. The image has been moved from Predictions to Training.

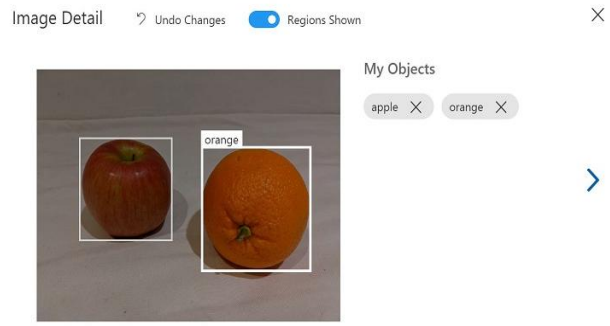


Figure 29:Trainthe Image Classifier

7. CUSTOM IMAGE CLASSIFIER

(Practical Experiment 1)

The custom vision service is a simple tool for prototyping, improving, and deploying a custom image classifier to the cloud. We'll create our model, and then go on to custom vision AI and signing. Create a new project and name it example (cats or dogs classifier) .Simply upload and name a few photographs to begin training your computer vision model. As you upload photographs, the model tests itself on them and improves precision through a feedback. For project types, we'll use classification; for domains, we'll use general compact and then create project. We'll need to add some tags to distinguish which image belongs to which name; the cat images belong to the cat tag, and the dog images belong to the dog tag, so we'll create a tag that says "cats or dogs." then make a new tag with the word "cats" in it. We'll select cats and add photographs by browsing local folders, selecting images, and adding them to the cats tag. Notice that I have 30 images; it's advised that you have at least 30, but the more images you have, the more accurate this classifier will be. I'll

unselect cats and then select dogs so that we can add these photographs, then browse local files, pick your images, and then open and tag this with dogs. 31 puppy pictures to upload Then we train our models and use clickers to see if they can distinguish between cats and dogs. As a result, when you put it to the test, it will correctly recognize it. The difference between the two is that I upload one more image of dogs as opposed to cats. I recommend that your precision rate be above 85 percent, and if it isn't, the problem is that you need to add more images. The more images you add, the more accurate your image classifier will be, and downloading images doesn't take much time. So, to test image classifier, we'll go to quick test and choose one of two options: submit an image by entering a URL or browse local files and select the image to open. It instantly recognized that it's 100% a cat and 0% a dog, making it very simple to create your machine learning model.

Figure 30: sample of cats &dog dataset

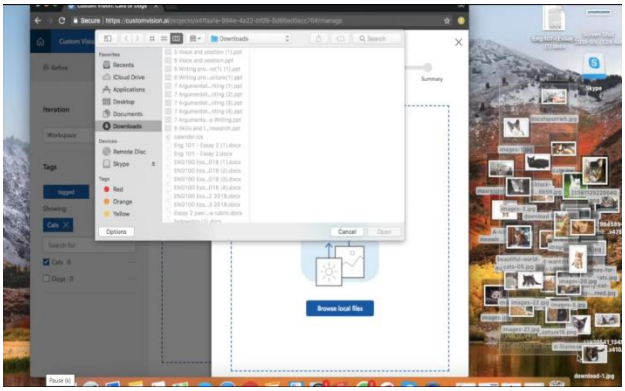


Figure 31: Deploy Custom Vision and load data set for cats

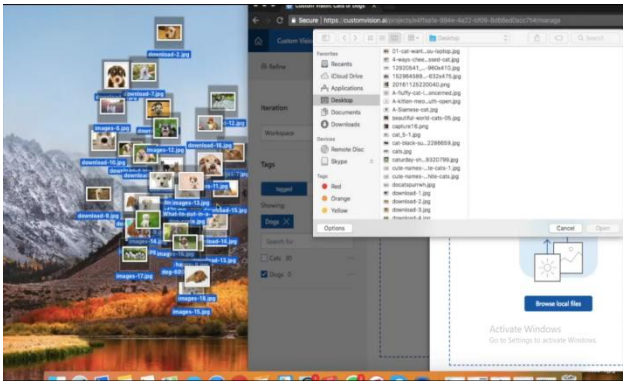


Figure 32:Deploy Custom Vision and load data set for cats

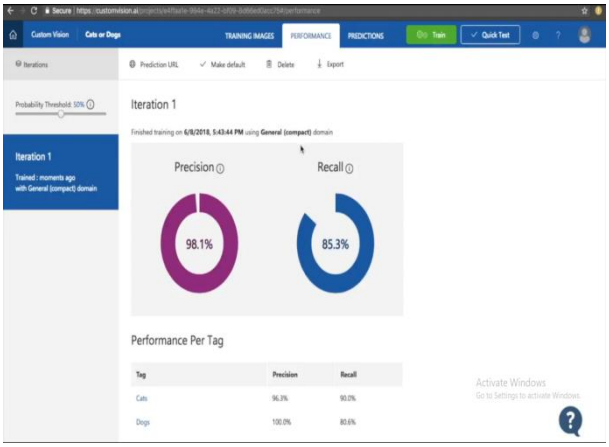


Figure 33:Azure Cognitive Services Precision and Recall Measures

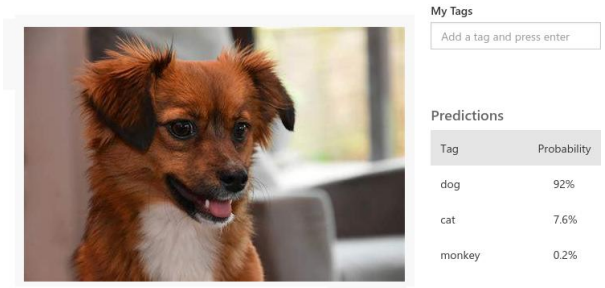


Figure 34:Image Classification test Results

7.1 Evaluate the classifier's performance

The model's performance is estimated and shown after training is done. A model evaluation can be done in a variety of ways, but in this case, a typical methodology was utilized. This is a k-fold cross validation method for estimating how well a classifier will perform in real-world circumstances. There is a display for precision and recall rate for each iteration and model training.

Precision and recall are two different measures of a classifier's effectiveness:

-Precision refers to the percentage of correctly detected classes the model identified.

-Recall indicates the fraction of actual classifications that were correctly identified.

- The average value of the average precision is the mean average precision (AP). The area beneath the precision/recall curve is referred to as AP (precision plotted against recall for each prediction made).

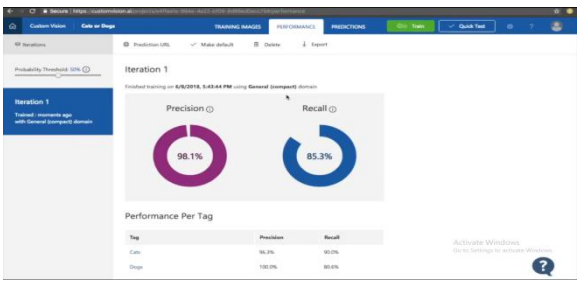


Figure 35: performance per Tag

8. IDENTIFY CARS IN REAL-TIME VIDEO

(Experiment 2)

More and more, real-time video is being used as the media feed to be processed for object detection. So, for this procedure, the proposed experiment useVideo Indexer as an Azure Media service integrated with CustomVision.ai. It allows us to retrieve some valuable information and insights.By integrating the two services, Custom Vision and Video Indexer are merged. "Animated characters (preview)" is the name of the feature.

To get the most out of proposed model, follow these steps:Create a schema: 1- Video Indexer is used in the proposed experiment to cut the video into shots and key frames. All of the key frames are listed and available (with IDs)

Data labelling quality is an important component in determining model performance. Avoid class ambiguity by ensuring that classes are easily distinguishable from one another (Cars, Trucks, Busses, etc..), particularly in single label classification.

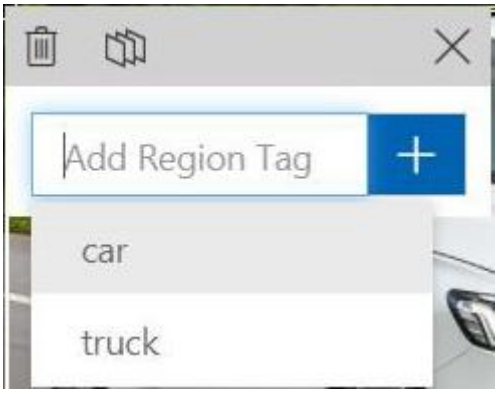


Figure 36:Adding tags

2- Model training: model begins to learn from your labelled data.

3- Details on the model evaluation can be found here: Examine your model's evaluation details to see how well it performs when presented with new data.

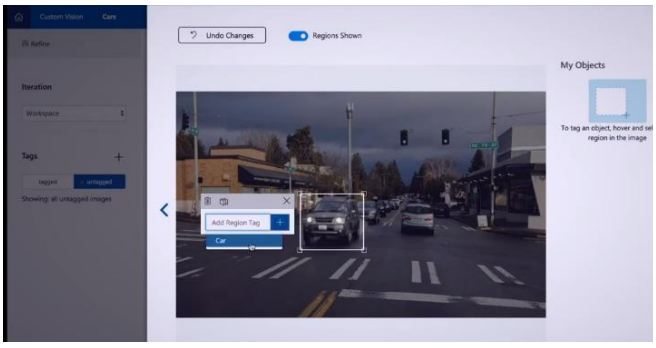


Figure 37: Custom Classification Model Training

4- Improve your model's performance by looking at the inaccurate model predictions and looking at the data distribution.

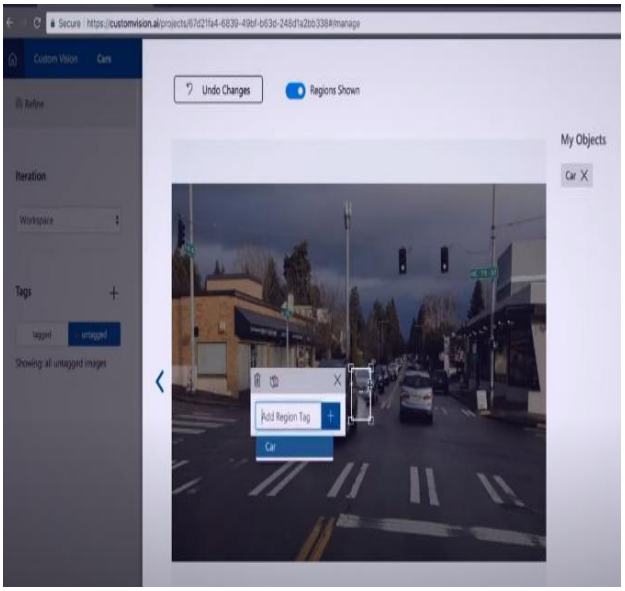


Figure 38: Custom Classification Model Training

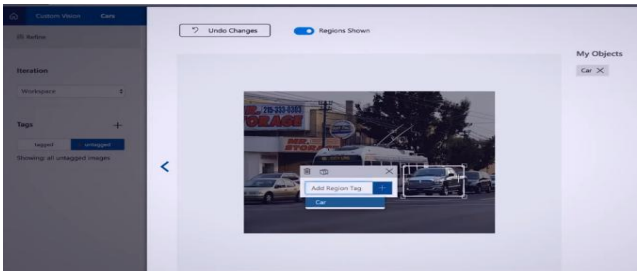


Figure 39: Custom Classification Model Training

5- Evaluate the classifier.

The model's performance is estimated and shown after training is done. Using a procedure known as k-fold cross validation, the Custom Vision Service calculates precision and recall using the photos you submitted for training. Precision and recall .

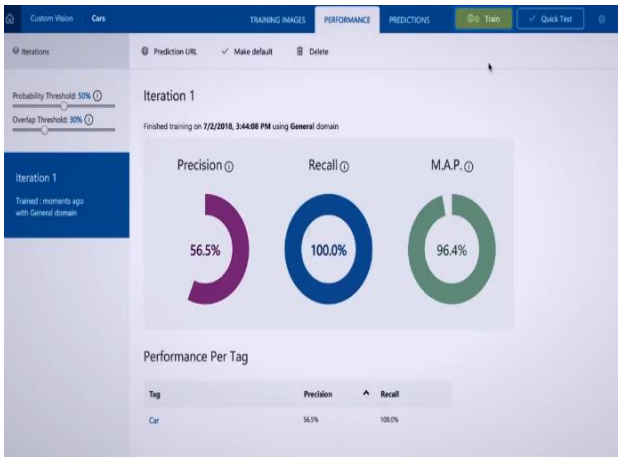


Figure 40:Evaluating Models in Azure Machine

With this in mind, you should set the probability threshold according to the specific needs of your project. Later, when you're receiving prediction results on the client side, you should use the same probability threshold value as you used here [20].

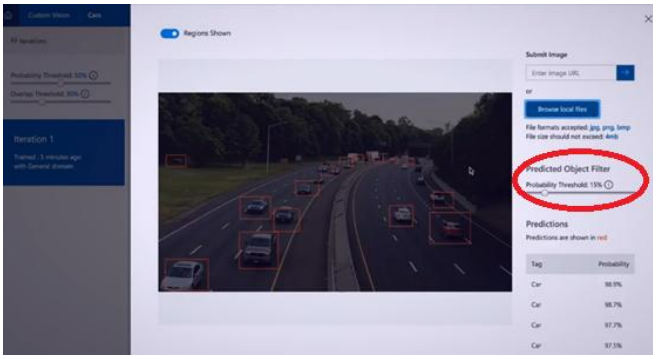


Figure 41:Azure Dynamic Thresholds

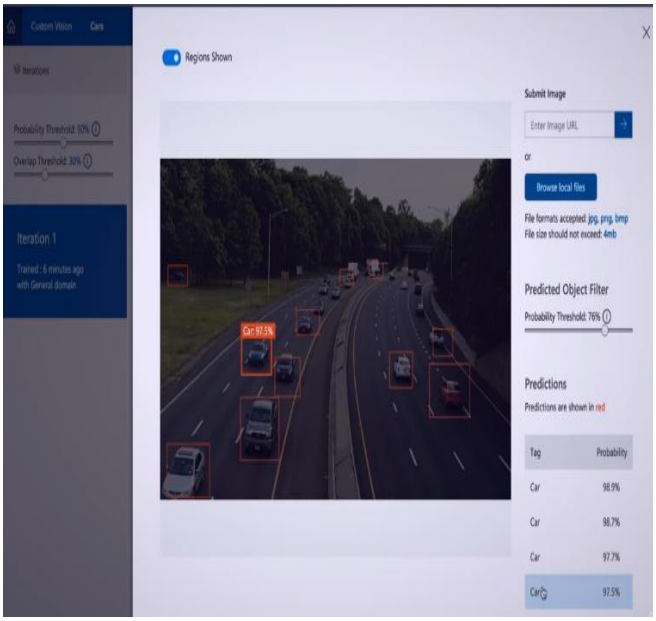


Figure 42: Image Object Detection using Azure Custom Vision

7-Deploy a model: When you deploy a model, it becomes available for use.

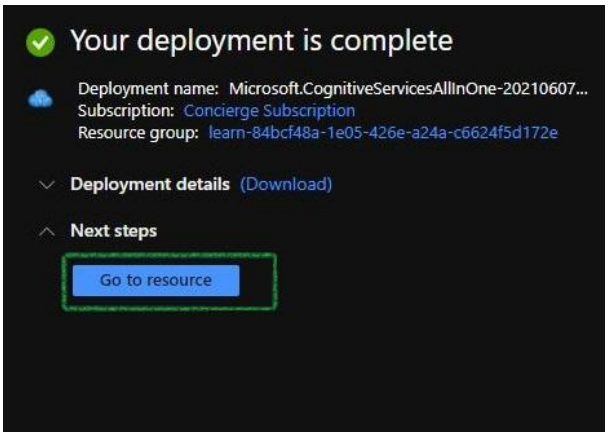


Figure 43:deploy a model

9. CONCLUSION

Different image processing-based categorization algorithms for computer vision are discussed in this work. Custom vision based on Azure AI can be used to create Microsoft Azure framework apps. Convolutional neural networks-based Custom Vision Service assists developers in the trans disciplinary sector in creating models that match their demands. Different picture segmentation, feature extraction, training, and classification methods are utilized, with improved results. The train/test split method is used in this paper to measure the model's capacity to generalize to new data. The computer vision results will improve if the recommendations in this study are implemented.We learned about deep learning and Azure cognitive services software, which may be applied to a variety of areas.. We hope that the results and methods presented in this paper can be further expanded in a bigger project

10. REFERENCES

[1] Frida Femling, Adam Olsson and Fernando Alonso-Fernandez 2018 Fruit and Vegetable Identification Using Machine Learning for Retail Applications 14th Int. Conf. on Signal-Image Technology & Internet-Based Systems (SITIS) , p. 9-15

[2] Y LeCun, L Bottou, Y Bengio and P Haffner 1998 Gradient-based learning applied to document recognition Proc. of the IEEE vol86 , p. 2278-2324

[3] X Yuan et al 2018 Deep Learning-Based Feature Representation and Its Application for Soft Sensor Modeling with Variable-Wise Weighted SAE IEEE Transactions on Industrial Informatics 14 , p. 3235-43

[4] Z Gao, L Wang, L Zhou and J Zhang 2017 HEp-2 Cell Image Classification with Deep Convolutional Neural Networks IEEE J. Biomed. Heal. Informatics vol21 , p. 416-28

[5] T Kooi et al 2017 Large scale deep learning for computer-aided detection of mammographic lesions Med. Image Anal 35 , p. 303–12

[6] S. Du and RK Ward, "Adaptive Region-Based Image Enhancement Method for Robust Face Recognition Under Variable Illumination Conditions," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, (9), pp. 1165-1175, 2010.

[7] Y. Zhu, J. Sun, and S. Naoi, “Recognizing Natural Scene Characters by Convolutional Neural Network and Bimodal Image Enhancement,” in Camera[1]Based Document Analysis and Recognition, vol. 7139, M. Iwamura and F. Shafait, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 69–82

[8] Q. Tian, G. Xie, Y. Wang, and Y. Zhang, “Pedestrian Detection Based on Laplace Operator Image Enhancement Algorithm and Faster R-CNN,” in 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 2018, pp. 1–5.

[9] Lucas, SM, Panaretos, A., Sosa , L., Tang, A., Wong, S., Young, R.: ICDAR 2003 robust reading competitions. In: 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, vol. 2, pp. 682–687 (2003)

[10] Alex Krizhevsky, Geoffrey Hinton, Technical Report, Learning Multiple Layers of Features from Tiny Images, vol. 1, University of Toronto, 2009. No. 4.

[11] Yann LeCun, et al., Gradient-based learning applied to document recognition, Proc. IEEE (1998) 2278–2324.

[12] Feng, X., Jiang, Y., Yang, X., Du, M. and Li, X., 2019.

Computer vision algorithms and hardware implementations: A survey. *Integration*, 69, pp.309-320.

[13] Ying, X., 2019. An Overview of Overfitting and its Solutions.*Journal of Physics: Conference Series*, 1168, p.022022.

[14] Bossert, L. and Hagendorff, T., 2021. Animals and AI.The role of animals in AI research and application – An overview and ethical evaluation.*Technology in Society*, 67, p.101678.

[15] Oosthuizen, K., Botha, E., Robertson, J. and Montecchi, M., 2020. Artificial intelligence in retail: The AI-enabled value chain. *Australasian Marketing Journal*, 29(3), pp.264-273.

[16] Plinere, D. and Borisov, A., 2015. Case Study on Inventory Management Improvement.*Information Technology and Management Science*, 18(1).

[17] B., 2014. A STUDY ON THE IMPORTANCE OF IMAGE PROCESSING AND ITS APPLICATIONS.*International Journal of Research in Engineering and Technology*, 03(15), pp.155-160.

[18] L'Heureux, A., Grolinger, K., Elyamany, H. and Capretz, M., 2017. Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, 5, pp.7776-7797.

[19] Razali, M. and Manshor, N., 2018. Object Detection Framework for Multiclass Food Object Localization and Classification. *Advanced Science Letters*, 24(2), pp.1357-1361.

[20] Kene, Y., Khot, U. and Rizvi, I., 2018. A Survey of Image Classification and Techniques for Improving Classification Performance.*SSRN Electronic Journal*,.