REPUBLIC OF TURKEY

AKDENIZ UNIVERSITY

MACHINE LEARNING BASED CIGARETTE BUTT DETECTION USING YOLO FRAMEWORK

Hasan Ender YAZICI

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

JANUARY 2022

ANTALYA

REPUBLIC OF TURKEY

AKDENIZ UNIVERSITY

MACHINE LEARNING BASED CIGARETTE BUTT DETECTION USING YOLO FRAMEWORK

Hasan Ender YAZICI

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

JANUARY 2022

ANTALYA

REPUBLIC OF TURKEY

AKDENIZ UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

MACHINE LEARNING BASED CIGARETTE BUTT DETECTION USING YOLO FRAMEWORK

Hasan Ender YAZICI

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

This thesis unanimously accepted by the jury on 24/01/2022.

Asst. Prof. Dr. Taner DANIŞMAN (Supervisor)

Asst. Prof. Dr. Hüseyin Gökhan AKÇAY

Asst. Prof. Dr. Shahram TAHERİ

ÖZET

YOLO FRAMEWORK KULLANARAK MAKİNE ÖĞRENME TEMELLİ SİGARA İZMARİTİ TESPİTİ

Hasan Ender YAZICI

Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Taner DANIŞMAN

Ocak 2022; 45 sayfa

Bu tez çalışmasında sokak veya tarımsal arazilerde sigara izmariti tespiti yapılabilmesi için gereken teorik ve pratik çalışmalar yer almaktadır. Tez çalışmaları You Look Only Once(YOLO)v5 framework kullanılarak uygulanmaktadır. Bu tez çalışması için videoların parçalanması ile framelerden oluşturulan verisetleri dışında internet ortamında yapılan araştırmalar sonucunda bir veri seti bulunamamıştır. İnternet ortamında dijital şekilde sentetik hazırlanan veri seti mevcuttur. Bu nedenle tez çalışmasının tamamlanabilmesi amacıyla 2000 fotoğraftan oluşan bir veri seti manuel olarak dijital herhangi bir işlem gerçekleştirmeden hazırlanmıştır.

Kamu kuruluşları, temizlik hizmetlerindeki insan sayısını artırarak sigara izmaritlerinden kaynaklanan kirliliği önlemeye çalışmaktadır. Ancak bu durum maliyeti önemli ölçüde artırmaktadır. Bu durumda görüntü işleme teknikleri kullanılarak sigara izmaritleri tespit edilerek, sokaklarımızın ve temizlememiz gereken alanların çevre kirliliğini önlememiz daha kolay olacaktır.

Bu çalışmada sigara izmaritleri gerçek görüntüler kullanılarak tespit edilmiştir. Önerilen yöntem, sigara izmaritini tespit etmek için kullanılabilecektir. Tezin sonucunda farklı alanlarda izmarit tespiti yapılacak ve GPS üzerinden en yoğun sigara izmaritinden kaynakları çevresel atığa sahip bölge gösterilebilecektir. Bu sayede sadece gerekli bölgelere otonom sistemler veya ilgili görevliler yönlendirilerek hem zaman tasarrufu hem de kaynak yönetimi verimli şekilde kullanılacaktır. Bu sayede çevresel atık sorununa yardımcı bir çözüm sunulabilecektir.

Amacımıza ulaşmak için gerçek hayatta farklı açılar, eğimler ve mesafeler için görüntüler alınarak manuel olarak bir görüntü veri seti oluşturulmuştur. Sigara atıkları kamusal,

tarım alanlarında herhangi bir yerde olabileceğinden, bu görüntülerin farklı zamanlarda farklı yerlerde olması gerekliliği çalışmada deneyimlenmiştir. Görüntü veri seti, öğrenme, doğrulama ve test için toplam 1000×800 çözünürlükte 2100 görüntüden oluşmaktadır. Her görüntü çalışma için manuel olarak etiketlenmiştir. Deneylerimize göre, veri setimizde umut verici sonuçlar (0.889 ortalama hassasiyet) elde edilmiştir.

ANAHTAR KELİMELER: Çevresel atık, nesne tespiti, sigara izmariti tespiti, YOLO Framework

JÜRİ: Dr. Öğr. Üyesi Taner DANIŞMAN

Dr. Öğr. Üyesi Hüseyin Gökhan AKÇAY

Dr. Öğr. Üyesi Shahram TAHERİ

ABSTRACT

MACHINE LEARNING BASED CIGARETTE BUTT DETECTION USING YOLO FRAMEWORK

Hasan Ender YAZICI

MSc Thesis in Computer Engineering
Supervisor: Asst. Prof. Dr. Taner DANIŞMAN
January 2022; 45 pages

In this thesis, there are theoretical and practical studies required to detect cigarette butts on streets or agricultural lands. Thesis studies are implemented using You Look Only Once(YOLO) v5 framework. For this thesis, apart from the datasets created from the frames by the fragmentation of the videos, a dataset could not be found as a result of the researches made in the internet environment. There is a digitally prepared dataset on the internet. For this reason, in order to complete the thesis, a data set consisting of 2100 photographs was prepared manually without any digital processing.

Public organizations are trying to prevent the pollution that exists because of cigarette butts, by increasing number of people in cleaning service. However, this situation significantly increases the cost. In this case by detecting cigarette butts with using image processing techniques, it would be easier to prevent environmental pollution of our streets and the area we need to clean.

In this study, cigarette butts were detected using real images. The proposed method can be used to detect cigarette butts. As a result of the thesis, cigarette butts will be detected in different areas and the region with the most intense cigarette butt sources will be shown via GPS. In this way, both time saving and resource management will be used efficiently by directing autonomous systems or relevant officials to only the necessary regions. In this way, a helpful solution to the environmental waste problem will be offered.

To achieve our goal, an image dataset was created manually by taking images in real life for different angles, slopes, and distances. The necessity of these images to be in different places at different times is experienced in the study since cigarette waste could be anywhere in public, agricultural areas. The image dataset consists of 2100 images at

1000×800 resolution in total for train, validation and test. Every image has been labeled manually for the study. According to our experiments, we obtained promising results (0.889 mean average precision) on our dataset.

KEYWORDS: Cigarette butt detection, environmental waste, object detection, YOLO framework

COMMITTEE: Asst. Prof. Dr. Taner DANIŞMAN
Asst. Prof. Dr. Hüseyin Gökhan AKÇAY
Asst. Prof. Dr. Shahram TAHERİ

ACKNOWLEDGEMENTS

I wish to thank my committee members for their time and dedication. A special thanks to Asst. Prof. Dr. Taner DANIŞMAN my advisor, for believing in me and my family who showed every kind of understanding and support during my graduate education and thesis work.

LIST OF CONTENTS

TEXT OF OATH

I declare that this study "Machine Learning Based Cigarette Butt Detection Using YOLO Framework", which I present as master thesis, is in accordance with the academic rules and ethical conduct. I also declare that I cited and referenced all material and results that are not original to this work.

24/01/2022

Hasan Ender YAZICI

LIST OF ABBREVIATIONS

AI	: Artificial Intelligence
ANN	: Artificial Neural Network
CNN	: Convolutional Neural Network
CPU	: Central Processing Unit
DNN	: Deep Neural Network
DL	: Deep Learning
FNN	: Feed-Forward Neural Network
HOG	: Histogram of Gradients
IoU	: Intersection Over Union
LBP	: Local Binary Patterns
mAP	: Mean Average Precision
ML	: Machine Learning
MS COCO	: Microsoft Common Objects in Context
NN	: Neural Network
RBF	: Radial Basis Function
SDG	: Stochastic Gradient Descent
YOLO	: You Look Only Once

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Environment is an essential element of life. If the air, water, and soil are contaminated, it is impossible to imagine a healthy and strong person. The physical, mental, cognitive, and social structures of man are all influenced by environmental influences. Currently, cigarette filters (butts) are the most common anthropogenic trash on the globe.

Cigarette butts are the most common waste that we can see in our streets everyday. This waste is hard to collect by human beings that is why we may use machines to handle this problem in our living spaces. Since cigarette's butt waste is small object, this pollution is ignored by the cleaning service employees. Due to its small size, it can fit in various places and workers can easily ignore it because it requires a high labor force and serious level of concentration. Filters also contain a variety of compounds derived by smoking tobacco when discarded, and some still include unsmoked remains. Cigarette butts have the potential to increase soil pollution due to the possibility of remaining in the soil. Therefore, they are one of the most dangerous reasons of environmental pollution and it must be cleaned before it affect the society's health.

Instead of using workforce of human, we may able to clean this kind of waste with machines. To do this, we need machines or algorithms that detect cigarette's butt. The only way to teach something to the machine is to have datasets that related with the problem. Systems or machines, that have ability to learn itself, would be solution for pollution and act as its programmed. Which brings us to the topic of machine learning.

ML is a branch of artificial intelligence(AI) and computer science which focuses on the use of raw data and algorithms to mimic the way that humans learn, gradually improving its accuracy. It's main focus is that creating machines which are able to learn from raw data and make predictions for given task related with the raw data. It can find patterns on labeled or unlabeled data. ML is technology that paves the way for computers to gain learning capability from data without programming explicitly.

The way data is collected and represented has influence on the outcomes of machine learning systems. As a result, extracting features with high representation from the available data is given specific attention in terms of improving the efficiency and classification performance of machine learning(ML) algorithms(Bengio et al. 2013). In the 1970s, the process of selecting optimal features to create feature vectors and solve the problem man-

ually became a complicated process. Since there is human design in rule-based machine learning approaches, making decisions on which information is related to solve the problem is harder. Therefore, selected features as optimal features were not even important features or information to classify an object/entity by using machine learning approaches. In order to avoid such situations and speed up the process, Deep Learning(DL) is suitable.

DL replaced the old method by bringing a different perspective. DL is form of ML that allows systems to discover for a fact and grasp the world in relation to the hierarchical order of ideas(Goodfellow et al. 2016). DL, a type of ML, is used to learn complex datasets which is not pre-processed. Deep learning is being utilized to develop AI technologies that can imitate the brain functions of humans(LeCun, Bengio, et al. 2015). It is a representation/unsupervised learning methods, made it possible for the machine to gain self-learning capability with raw data.The detection studies which use deep learning mostly depend on automatic detection. Because of using less effort for the learning stage, it is getting more popular for studies which can work on raw data.

In this thesis, we work on studies required for detecting cigarette butt with machine learning techniques under the framework of YOLO. To do this objective, an image dataset was created manually by taking images in real life. For creating our model and teaching it to the machine, every image is labeled manually. To make our study successful, we standardized images into 1000×800 resolution. At the end of our study, we obtained promising results such as 89% mean average precision(mAP), precision of 95% and recall of 93%. F1 score of 98%, precision of 96% and accuracy of 97% were calculated on the testing dataset.

These result will lead us to possibility of autonomous system that can clean cigarette's butt waste in the streets and needed areas. It makes streets clean so the society live in healthy conditions. In addition, since it uses this human resource efficiently, it needs less workforce. In this way, the workforce can be shifted to the required areas. Especially, addition of GPS location information in to the systems that will build to deal with this type of pollution will solve this problem with less effort.

2. LITERATURE REVIEW

With the growing technologies in computer science there is a field occur named as computer vision. Computer vision is a scientific field which tries to improve understanding of computers on digital data such as images, videos, speech and text. With the increasing technology and hardware of computers become more powerful and capable. As a result of the studies and researches done over time, the tendency to use computer vision ML has improved since ML offers powerful hardware such as GPU,CPU. It become faster than traditional ways. With the improvement in computer power and the diversification of data, data analysis and detection models can be established in images.

In the field of ML, images are processed also the specific features of the images has been extracted in order to create a model. By using the features an object in the image is able to detect. In the challenges of detecting an object from images, classification studies on whether there is the object in the picture were done using machine learning.

The capacity to describe objects through images has now transferred to computers as a result of technological advancements, however not to the same extent as the ability of the human brain to classify objects. With their tremendous performance in image classification challenges, deep learning algorithms have shown themselves many times over. Image processing projects or computer vision in artificial intelligence have captivated the public's interest in recent years and have been the center of attention.

Computers can identify objects using powerful artificial intelligence algorithms. The volume of data and processing power has gotten increasingly difficult for computers as image quality and pixel count have increased. The usage of GPU in computers has grown prevalent in order for technology to tackle this problem. In image data, CNN is commonly utilized for several applications such as segmentation, object identification, and picture recognition.

Deep learning algorithms conduct actions on points called pixels in image processing procedures using a matrix structure. Computers require powerful processors to process any image or photo. One of the main reasons is that it converts the image into a matrix format and attempts to capture the relationships. As a result, deep learning algorithms employ a significant number of multidimensional matrix operations.

2.1. Image Processing

Image processing is the process of transferring an image to digital form and processing it to extract valuable information from it. Computers, software, and a variety of computational approaches are used to process images electronically. There are three steps to image processing. The first step is to digitize the image using various methods. The transmitted image is cleaned, analyzed, and processed in the second stage in order to reduce noise. The image can be used for the output stage.

In the 1920s, image processing concept was born with the digitization of a newspaper photos received through undersea cable.(Perihanoglu 2015) In 1961, the notion of digital light-sensing was introduced at a space conference , and a panel, that is done in 1968, with the intended logic was created. Steven Sasson created the world's first digital camera in 1975.(Galal 2016)

Image processing is a technique for digitizing images and then performing operations on them. These operations is done either in order to produce a specific image or extract relevant information from them. There must be a pre-proccessing stage to use an image in the image processing operations. The pre-processing stage affects the results of the image processing operation directly. To increase the effectiveness of the operation, noise and noise ratio on the image should be improved. There are many techniques for reducing noise. Filtering process is one the most common techniques.

Because of the many effects applied to the picture during the filtering process, it is possible to more easily interpret the image by allowing the distinction between physical attributes to be clarified or erased(Ma and et al. 2019 , Michalak and Okarma 2019). To decrease color disparities in the image and highlight sparse structures, image smoothing filters are utilized.(Fan and et al. 2017). For many computer vision applications, smoothing is a necessary step.(Liu and et al. 2020). Gradient size is commonly utilized as hint in image smoothing techniques.(Fan and et al. 2018).

2.1.1. Image captioning

Image captioning is a way of creating a textual description of an image using both natural language processing and computer vision. There has been an increase in the number of research aimed at analyzing the pattern and developing image captioning due to the

massive rise in the number of photos shared on the Internet. Due to its potential applications, defining the visual content of an image to a machine has gotten a lot of interest in recent years. The automatic creation of a caption on an image is the problem of creating a textual description of a rendered image. Therefore, the image captioning problem is seen as part of both science of computer vision and natural language processing.

Studies based on creating image captions are roughly divided into two groups which are template-based and retrieval-based. Template-based methods such as generative-based approach detect a specified set of visual concepts and generate captions based on sentence template or grammar rule(Yang et al. 2011,Mitchell et al. 2012) while access-based methods generate captions from a set of predetermined sentences. Since image captioning can be used for indexing purposes, these captions can be used during the retrieval process.

2.1.2. Image classification

Image classification is the another important topic in the field of image processing research. Since an image is big matrix of numbers for computers, there is no way to understand what is in the image for computers. To understand the content inside an image one must use image classification techniques which are used algorithms to extract a feature from an image.

Image Classification is basically assigning a label to given image from defined category list.Using their advanced intelligence, computers can classify objects as if they were human decisions. With the quality of the photo and the number of pixels, the volume and processing of data has become difficult. In order to eliminate this problem of the technology, the use of GPU Graphics "Processor Units" has been expanded.(Cınar 2020)

To analyze the contents of an image, feature extraction is the way of doing this traditionally. Image classification algorithms need a dataset to extract features of given predefined classes. After taking an input image, obtaining a feature vector that describes the image is called as feature extraction. In classification research, when there are hundreds of alternative choices, choosing the optimal variable and feature becomes the concentrate. Feature selection is a strategy for reducing feature size while increasing classifier output by picking a subset of relevant features from the original data(Danışman et al. 2013). Applying features like Histogram of Gradients(HOG), Local Binary Patterns(LBPs) used to finalize the process traditionally. Other than traditional methods, ML algorithms and DL

algorithms are used for image classification as well.

2.2. Artificial Intelligence

In variety of industrial, intellectual, and social software applications, Artificial Intelligence (AI) has the same disruptive potential for augmenting and perhaps replacing human functions and activities. With recent discoveries in algorithmic ML and autonomous decision making, new avenues for development are opening up(Dwivedi et al. 2021). AI software imitates the learning process by imitating people's thinking and evaluating the results obtained from past experiences. Since people act through the learning process, more successful results can be obtained in solving problems compared to rule-based software (Erler et al. 2003).

With the increasing capacity of computers, there are four important areas where AI is used such as voice recognition and understanding, image processing, natural language processing and understanding are reasoning. The use of artificial intelligence in daily life has expanded. Even applications in mobile phones has AI for humans to handle with their daily life issues. Applications can understand what people ask and how they feel, more human-like AI and robots have been developed. AI usually studied in the form of various Neural Network architectures depending on the problem.

2.2.1. Machine learning

The widespread use of technology has resulted in an increase in the amount of data produced. At the same time, storing and accessing data has become both easier and less costly compared to the past. Depending on these, data has started to be considered as valuable as finance in our world.

ML is a sub-branch of artificial intelligence that includes of modeling and algorithms that use mathematical and statistical approaches to draw conclusions from current data and make predictions about the unknown. The main purpose of ML is to produce predictions that is accurate. Estimation functions, on the other hand, might be difficult to understand and link to a certain probability model(Caglayan 2018).

The point that needs to be emphasized in the context of machine learning is that the learning ability of hardware such as computers, telephones, air conditioners, cars and

televisions is gained by computer software. In the software, instead of explicitly programming the hardware to do a task, the approach of developing programs that can learn to do the task is used. In other words, in traditional programming, every step necessary for the fulfillment of a task is coded by the programmers, while in machine learning, algorithms learn to solve the same task themselves by training on data (Samuel 1988). Machine learning can be defined as any change that will enable a system to perform the same task for the second time or to perform better in a new task created on the same data. Briefly, ML is predicting future by using past experiences.

Summarizes the important elements of machine learning,

- Today, when machine learning is mentioned, computers can be brought to mind as the machine that performs learning.
- It is noteworthy that in the definitions of machine learning, especially the emphasis on experience and performance.
- What is considered as past experience for the machine is actually the data required for learning to take place and the size of this data.
- Certain criteria have been defined to measure the success of learning. Thus, model performance is evaluated at the end of the process.
- If there are parameters of the installed model so that the machine can learn, these parameters are changed. (Kartal 2015)

Many various issues are solved with ML, including optical character recognition, facial identification, spam e-mail filtering, spoken language understanding, medical diagnosis, consumer segmentation, fraud detection, and weather forecasting. Next section will briefly explain the methods of ML.

There are four known methods of machine learning as presented in Table 2.1.

Table 2.1. Describing different types of ML (Seetharam et al. 2020)

Types of machine learning	Function	Examples
Supervised learning	The dataset has labels and outcomes, infers from data for prediction purposes	Encompasses logistic regression, ridge regression, elastic net regression, Bayesian network, artificial neural network
Unsupervised learning	The dataset contains no labels, detects pivotal relationships	This contains hierarchical clustering, k-means clustering, principal component analysis
Semi-supervised learning	A combination of supervised and unsupervised learning	Frequently used in image and speech recognition
Re-enforcement learning	Utilizes reward function to execute tasks	Commonly seen in medical imaging, analytic and prescription selection

2.2.2. Neural networks

The neural network(NN) means a collection of nodes that are bounded together in network and the attributes of the network are determined by topology and the attributes of every node in the network.

In 1943, the term of NN is first mentioned by Warren McCulloch and Walter Pitts on a paper on how they could work. They simulate first NN using electrical/logical circuits. After releasing paper about NN, in 1949 Donal Hebb pointed in his work that neural are getting stronger each time they are used. By developing technology, in 1950's first hypothetical NN was simulated and first attempt was failed.

In 1959, Marcian Hoff and Bernard Widrow developed some models and both model were able to recognize the path so they also were able to predict the next bit in phone line. In 1962, they improve their previous model so it will become more stable.

During the same time period, a research was published claiming that a single layered neural network could not be extended to a multiple layered neural network. Furthermore, many people in the area were relying on a learning function that was essentially incorrect because it did not differ across the entire line. As a result, research and funding have plummeted. Later that in 1975, the first multilayered and unsupervised network was developed.

Trying to find the best weights and biases for minimizing loss function is the main

purpose of NN systems. To minimize the loss function, training data are used for detecting the difference between output and the real output. If there is a resulting error, then updating the weights of neurons process will be done. The updating process done by using backpropagation algorithm(LeCun, Boser, et al. 1990)(LeCun, Boser, et al. 1998).

As a result of developing technology and studies day by day, there are many types of NN developed and become available for ML and DL methods. However, there are 4 common types of NN which are shown below for these purposes.

- Feed-forward Neural Network(FNN)
- Radial basis function Neural Network(RBF)
- Recurrent Neural Network(RNN)
- Convolutional Neural Network(CNN)

2.2.2.1. Feed-forward neural network

Feed-forward neural networks(FNN) are the basis of several modern notable neural networks, such as convolutional neural networks(CNN) and recurrent neural networks (RNN). There are no feedback connections or loops in feed-forward neural networks, which makes them a type of directed acyclic graph.

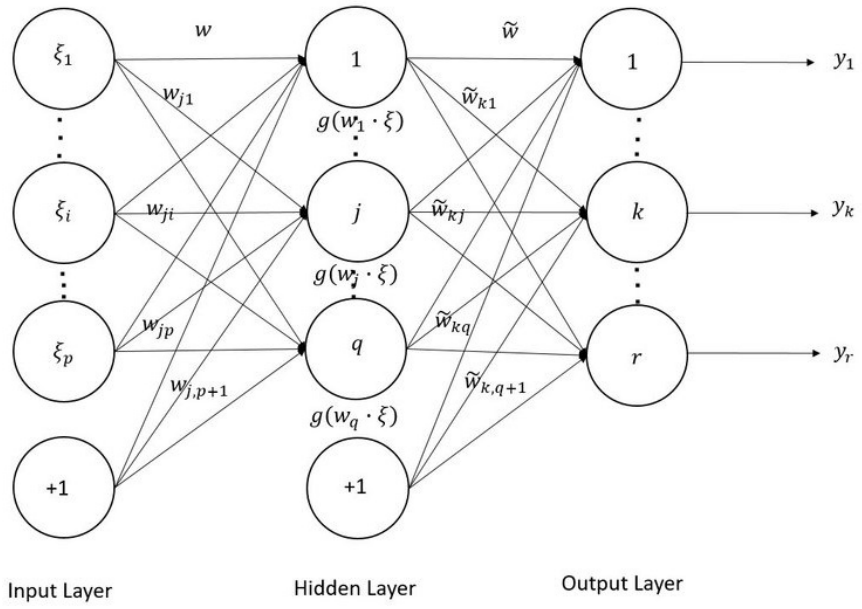

Figure 2.1. FFN with one hidden layer(Alemu et al. 2018)

FNN is not capable to solve problem which its data is not separable linearly. Feed forward neural networks are used to learn the link between independent variables (inputs) and dependent variables (outputs). FNN is generally use supervised learning method. It makes predictions depending on the labeled data. Its prediction can be either label or numerical. However, by adding hidden layers to FNN, it is able to solve this type of problem. Generally, there are multiple hidden layers for handling with complex situations. FNN can be used in problems such as car purchase price or house rent price prediction.

2.2.2.2. Radial basis function neural network

Radial basis function(RBF) network is a derivation of feed-forward neural network. It uses supervised algorithm for training. Generally, they are created with a single hidden layer of units that is selected from basis functions.

RBF has more advantages than back propagation. RBF's training process is way more faster than networks that uses back propagation algorithm. The radial basis function hidden units are less susceptible to problems with non-stationary inputs because of their nature. RBF networks used in various problems. Some major applications are below.

- Image processing
- Speech recognition
- Time-series analysis
- Adaptive equalization
- Radar point source location
- Medical diagnosis
- Process faults detection
- Pattern recognition(Sahin 1997)

RBF networks are primarily used to solve classification challenges. The pattern recognition issue is the most common classification problem requiring RBF networks. The second most prevalent application area for RBF networks is time-series analysis(Sahin 1997).

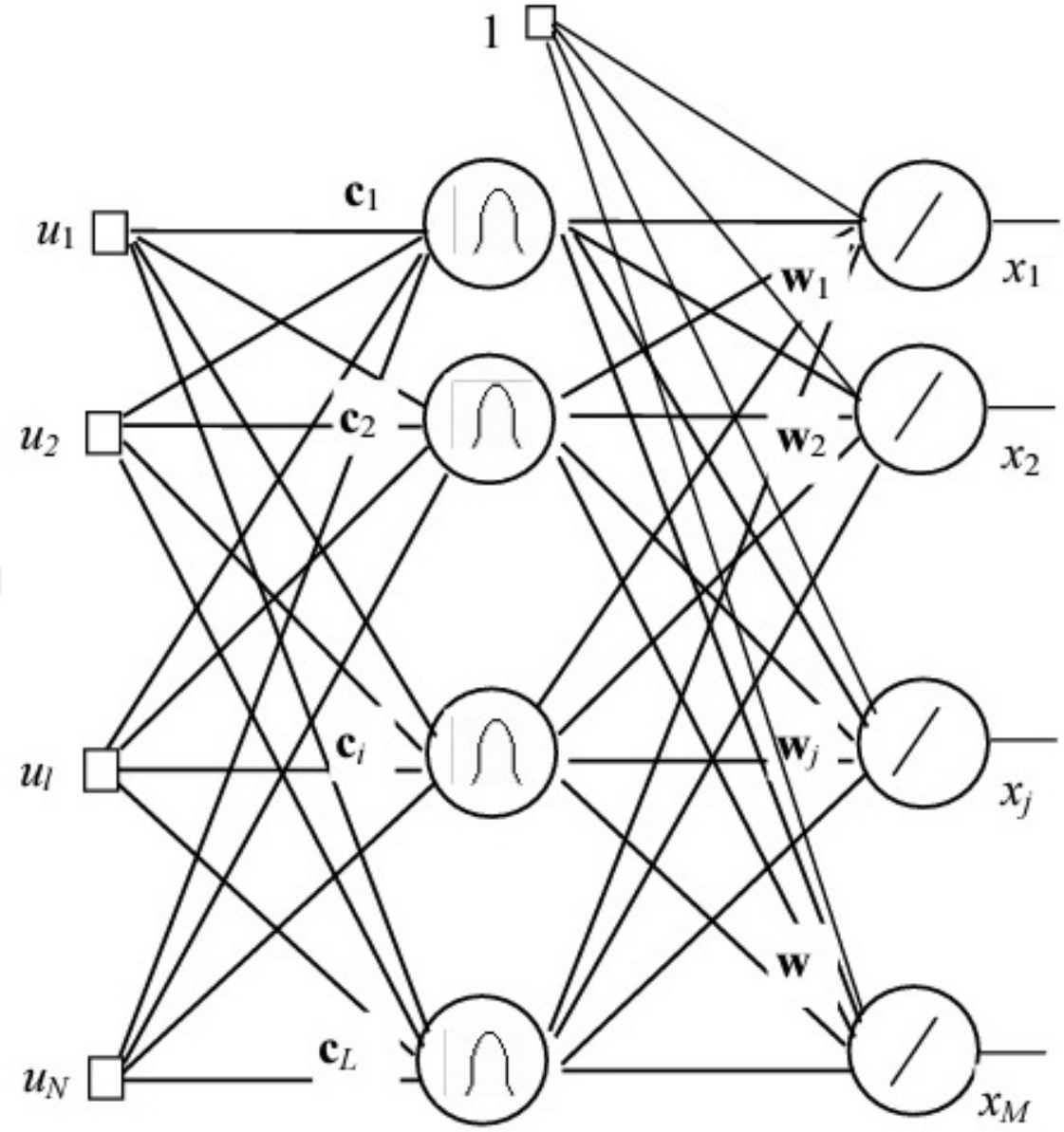

Figure 2.2. Structure of the standart RBF network(Halıcı 2004)

2.2.2.3. Recurrent neural network

As mentioned before, Recurrent Neural Network(RNN) is a type of FNN. It works like FNN only with minor difference. On hidden state, RNN has a recurrent connection. The looping constraint make sure that sequential information is captured in the input data. Hidden cell gets its output as input parameter with fixed delay to ensure the result. Context is crucial for RNN, since decisions that made in the previous iterations can affect the newer ones. RNN is commonly used for machine translation, speed recognition, video tagging, generating text. These problems are usually can be grouped into data problems

as below.

- Time Series
- Text
- Audio

RNN's architecture is shown in the Figure 2.3.

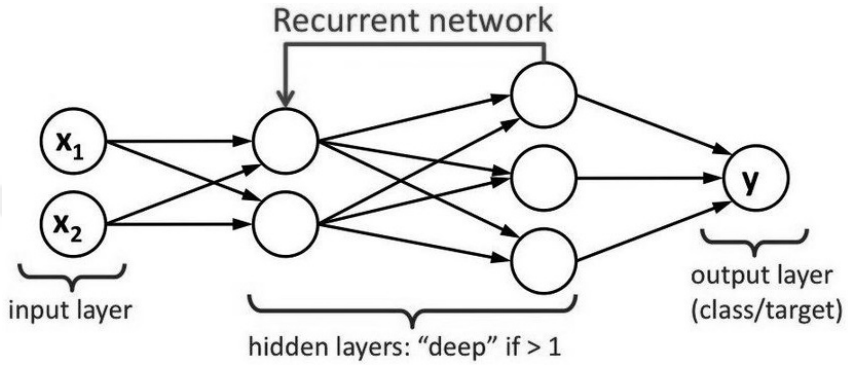

Figure 2.3. Recurrent neural network(Mishra et al. 2018)

2.2.2.4. Convolutional neural network

Convolutional Neural Network(CNN) is a DL technique based on artificial neural processing, classification and object detection (Albawi et al. 2017). Local receptive fields, shared weight networks (ANN) that is extensively used in the fields of computer vision, object recognition, images (or weight replication), and spatial or temporal subsampling are three architectural structure that Convolutional Networks use to assure some degree of shift, scale, and distortion invariance. Each unit in layer takes input from group of units that is being in the prior layers that are positioned in a small neighborhood(LeCun, Boser, et al. 1998).

Actually, FNN and CNN are quite similar. They are made up of neurons that have learnt weights and biases. Every neuron takes some inputs, makes a dot product, and then, if desired, does a non-linearity. The network as a whole still has a single differential score function. And they still have a loss function, as well as all of the training and optimization approaches for learning feed-forward neural networks.

CNN captures a photograph's spatial characteristics. Spatial features refer to how pixels are arranged and how they interact in a picture. They help us recognize an item, as well as its position and relationship to other objects in a picture.

CNN does not need feature extraction. The machine detects the feature between layers by itself. The most important achievement is that computer hardware is now able to work under more load. Moreover, graphical process units(GPU) have made a lot of progress in terms of increasing their capacity and workload. Because of these developments, CNN is used in image and video classification by using its multiple layers.

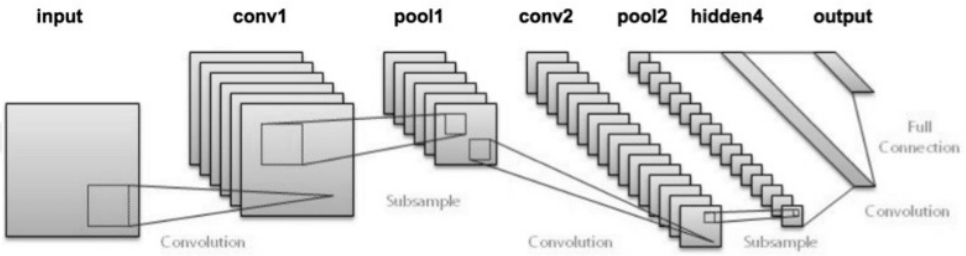

Figure 2.4. Convolutional neural network

Convolutional Neural Networks generally consist of many layers:

- Input layer
- Convolutional layer with nonlinear activation function
- Pooling layer
- Fully connected layer
- Softmax layer
- Output layer (Hariri 2021)

Artificial neural networks presently are dominated by CNN. Convolution cells (or pooling layers) and kernels are used, with each having a distinct role. Convolution kernels analyze the incoming data, whereas pooling layers simplify it and remove extraneous features. They're most commonly employed for picture recognition, and they only work on a tiny subset of images. Pixel by pixel, the input window moves along the picture. The information is sent into convolutional layers, which form a funnel. The first layer

recognizes gradients, second lines, third forms, and so on to the scale of specific objects in terms of picture identification.

2.2.2.5. Deep learning

As mentioned in ML section, a human needed to extract data from big dataset. The problem is that anyone is capable to do it. Sometimes people can make mistakes so the model created to handle with the problem may fail. That is the point Deep Learning(DL) started to become popular. DL replaces the traditional ML approach, in which human specialists define the collection of features to be extracted from images. A DL network takes images, or regions in images, as input and transforms them into a conclusion through multiple layers of processing stages. Feature extraction occurs in these intermediary layers, and these features are not intentionally generated by the system's designers, but rather learned from the data throughout the training process. This is a total paradigm shift that has been dubbed "the end of code" by some.

DL is a strong type of ML which allows computers to handle sensory problems like image and speech recognition is increasingly being used in research.

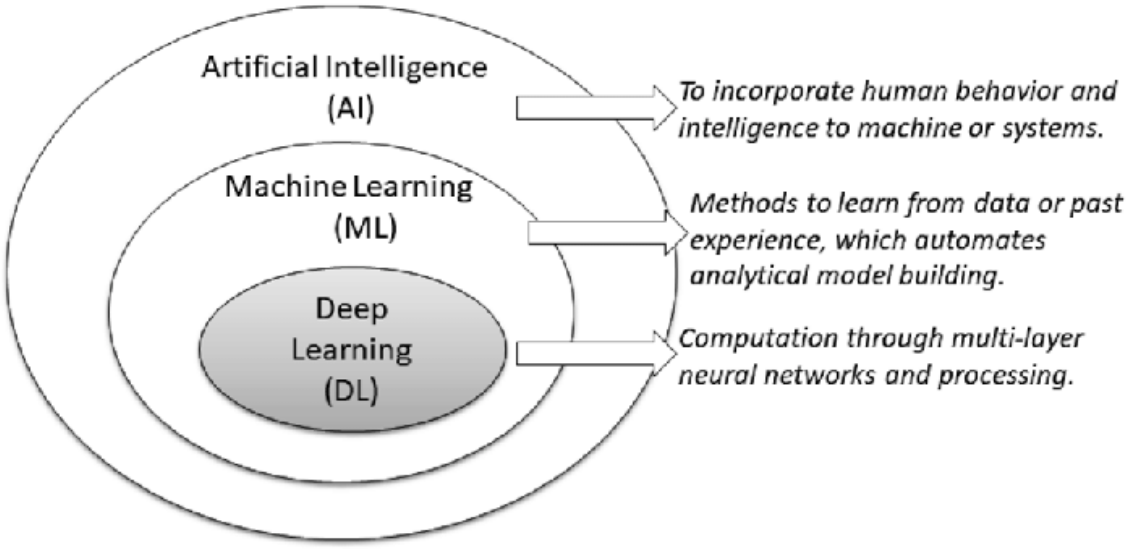

Figure 2.5. Relation of DL with ML and AI(Sarker and Iqbal 2021)

DL techniques, such as deep artificial neural networks(DNN), employ numerous processing layers to find structure and pattern in huge datasets. Each layer learns a notion from the input, which is then built upon by successive layers; the higher the level, the

more abstract the concepts learnt. Deep learning does not require any pre-processing data stage and extracts features automatically.

In the initial layer, a deep neural network tasked with interpreting forms would learn to detect basic edges, and then in following layers, it would learn to recognize more complicated shapes made of those edges. Even though there is no strict rule regarding how many layers deep learning requires, most experts agree that more than two are required (Rush 2016). The Figure 2.6 shows the Deep Neural Network's structure.

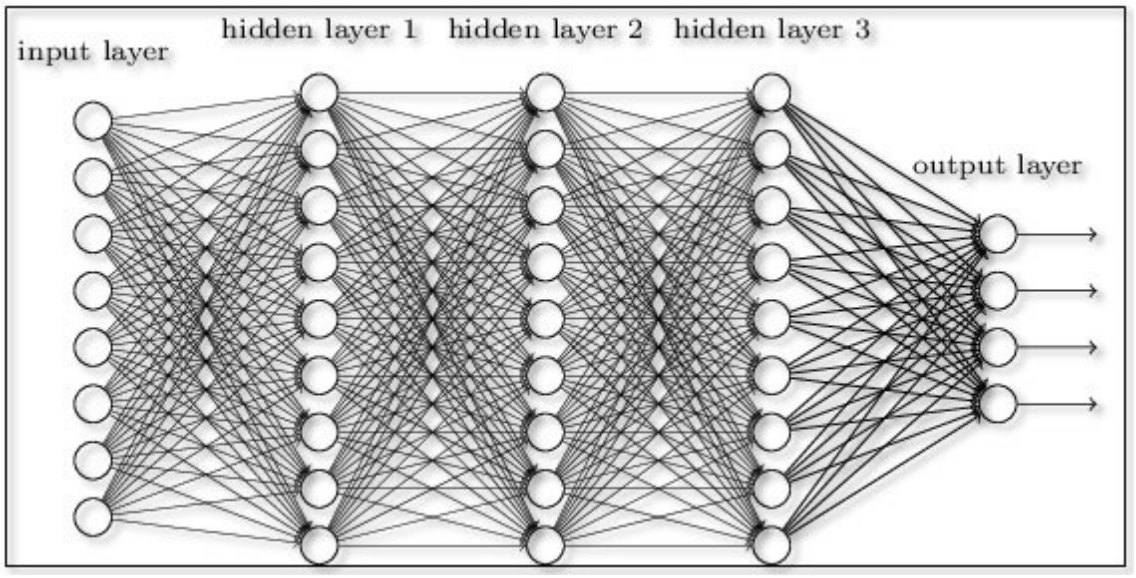

Figure 2.6. Deep neural network structure

Because of working on training dataset first, detection of features takes a long time from given dataset. Also one more crucial point is data. Data must be in proper and complete. If there is a missing columns in dataset, data engineers should work on the given dataset to make it complete. Since there is no need to extract features manually from big datasets, deep learning become popular. In deep learning machine can understand which feature is important to make a classification or make a prediction.

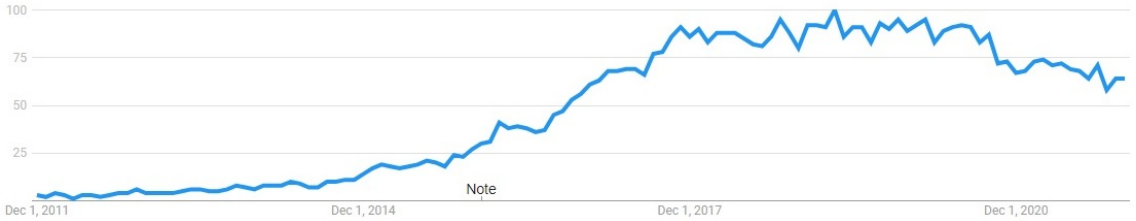

Figure 2.7. Deep learning search statistics (2011-2021)

2.3. YOLO Framework

You Only Look Once (YOLO) is a clever CNN that detects objects in real time. Using a single neural network processed to the full image, this approach divides the image into regions and forecasts bounding boxes and probabilities for each area. These bounding boxes are weighted based on the estimated probability. One convolutional neural network predicts various bounding boxes and class probabilities for those boxes at the same time. YOLO increases detection performance by training on entire images. Compared to traditional way, this integrated model has a lot of advantages. The YOLO network is made up of two completely connected layers and twenty four convolutional layers. They just use 1 x 1 reduction layers followed by 3 x 3 convolutional layers instead of the inception modules used by GoogLeNet(Abrol and Mahajan 2015). The Figure 2.8 shows the YOLO's architecture.

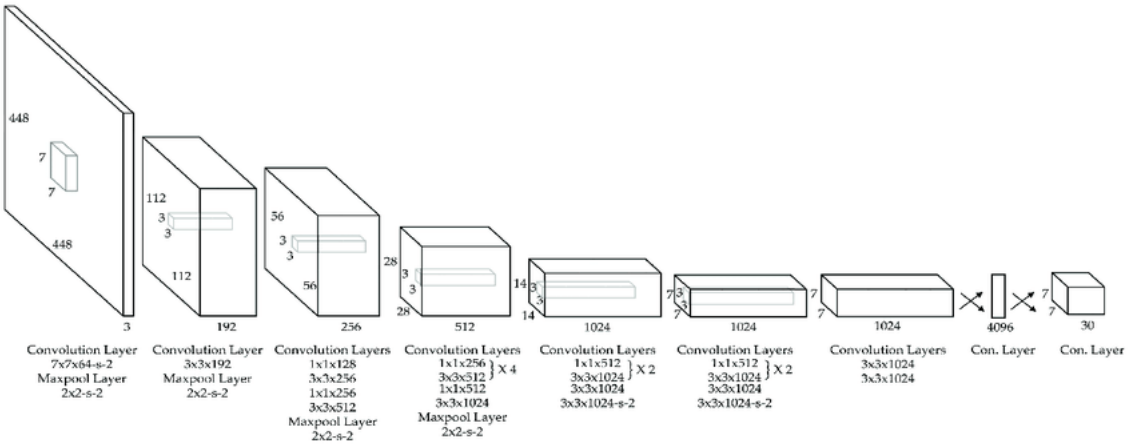

Figure 2.8. Architecture of YOLO

Yolov5's network architecture is separated into three sections. The backbone (CSPDarknet), the neck (PANet), and the head (Yolo Layer). The data is initially sent through the backbone, which extracts features, before being fed into the neck, which fuses them together. Finally, the head displays your detection findings. In this study, we are going to use YOLOv5 which is the newest version of YOLO Framework(Xu et al. 2021).

The Figure 2.9 shows the architecture of YOLOv5.

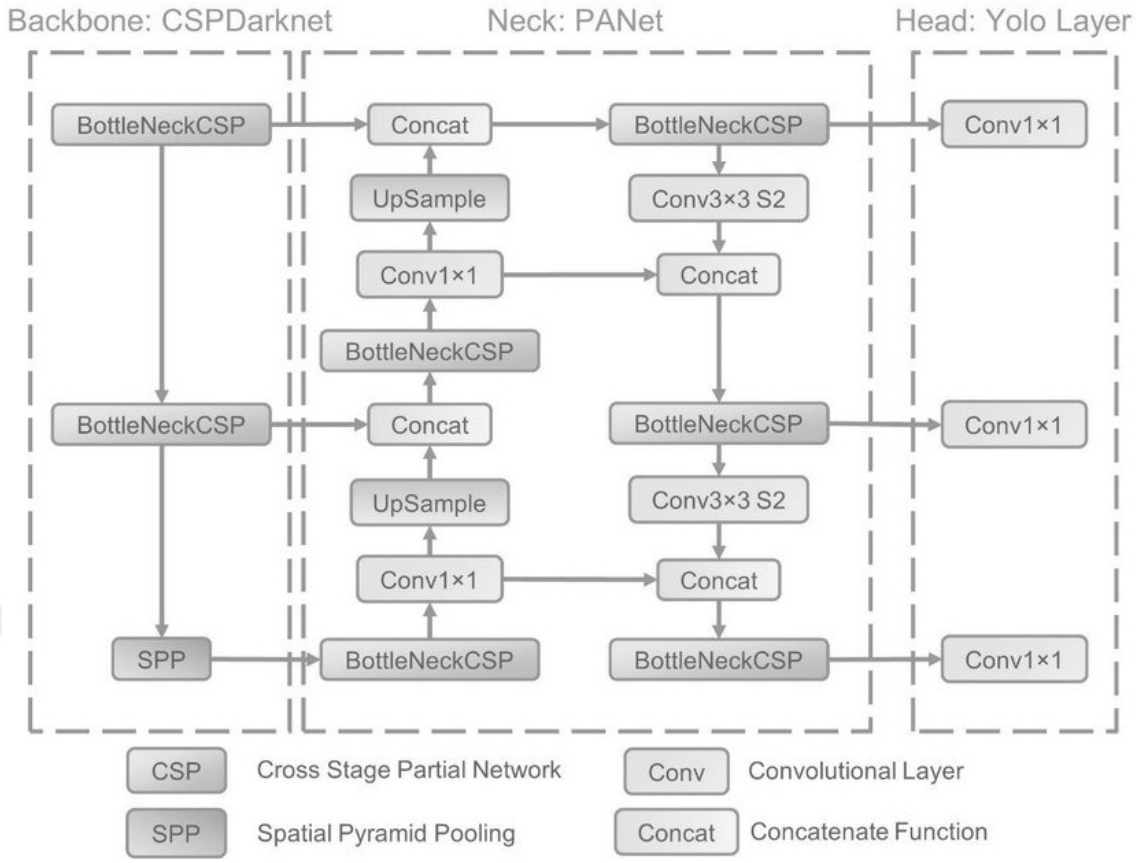

Figure 2.9. Architecture of YOLOv5(Xu et al. 2021)

Joseph Redmon et al. created YOLO, a deep learning-based object recognition method. It estimates bounding boxes and class probabilities with a single neural network and reframes object detection as a single regression problem (Redmon et al. 1990). It provides numerous constraint boxes and class choices for these boxes using a single neural network. YOLO divides the input image into cells using the S $\times$ S grid, evenly dividing the x, y, and image into the grid. If an object has a central cell, that cell is in charge of identifying it. Each grid cell calculates the probability of finding things.

YOLOv5 is different from earlier versions in that it is based on PyTorch rather than the traditional Darknet. Like the YOLO v4, the YOLO v5 has a CSP backbone and a PA-NET neck. Two of the most notable improvements are mosaic data augmentation and auto learning bounding box anchors.

YOLO is a one-stage object detection technique that has been around for a long time. It transforms the detection issue into a regression issue. Instead of extracting RoI, it uses the regression approach to obtain the bounding box coordinates and probability of

each class. It considerably enhances detection speed when compared to quicker R-CNN. utralytics created the fifth generation of YOLO in 2020. It outperforms all prior versions in terms of speed and accuracy. The YOLOv5 algorithm adjusts the width and depth of the backbone network using the depth multiple and width multiple parameters, yielding four different models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The simplest version is YOLOv5s.

2.3.1. Object and event detection studies using yolo framework

Variety of objects such as small arms, ceramics, glasses, smoke, fire etc. is detected by using algorithms like YOLO that use own unique dataset. Studies in object detection have been carried out deep learning in recent years.

Li M., Zhang Z., Lei L., Wang X., Guo X. (2020) in their work, they presented an Agricultural Greenhouses Detection in High-Resolution Satellite Images Based on Convolutional Neural Networks: Comparison of Faster R-CNN, YOLO v3 and SSD in China. For detecting Agricultural Greenhouses from satellite photos, they used Convolutional Neural Network(CNN) based models such as Single Shot Multi-Box Detector(SSD), You Look Only Once(YOLOv3) and Faster R-CNN. Their performance, which primarily included accuracy and efficiency, was analyzed and assessed using test sets in a variety of settings. According to mean avarage precision and frame per seconds metrics, the YOLO v3 network produced the greatest results. They concluded that while SSD and Faster R-CNN both indicate certain capabilities in terms of detecting efficiency and accuracy, they both have struggle to meet the demands of quick and accurate object detection at the same time. In operational monitoring work, YOLOv3 can generally accomplish accurate and fast Agricultural Greenhouses detection using satellite images, which can be used as ground for planning and management by government(Li et al. 2020).

Basaran, E., Comert, Z., Celik, Y.,Velappan, S., Togacar, M. (2020) in their work presented an application of Determination of Tympanic Membrane Region in the Middle Ear Otoscope Images with Convolutional Neural Network Based YOLO Method. They discovered that LBP, HOG, IFT, and histogram-based algorithms like k-Means are the most common methods for detecting objects. Object detection with computer vision and deep learning methods such as YOLO, Fast R-CNN, SSD, Faster R-CNN and R-CNN are practiced. Traditional algorithm and deep learning algorithm, that use object-region detection,

are both efficient when these experiments are reviewed. They applied object detection algorithm which has been demonstrated to be an effective method in the literature in their research(Basaran et al. 2020).

Dikbayir and Bülbül (2020) realized that when they add a Faster R-CNN neural network before YOLO, their results are getting higher by 4.3% and their input values reach 60 fps in their work which is Real-Time Vehicle Detection by Using Deep Learning Methods. However, they observed that the performance and accuracy of the developed algorithm decreased as the input size increased(Dikbayir et al. 2020).

Cagıl and Yıldırım (2020) presented Detection of an Assembly Part with Deep Learning and Image Processing. In their experiment, they used the YOLOv3 algorithm with DarkNet Neural Network and they saw that the study works correctly at a rate of 84% (Cagil and Yıldırım 2020).

Shahriar S. S., Junzo W., Anurava R., Dra R. (2020) in their work, discussed which method is better between Mask R-CNN and YOLO. In their experiments they did on their custom image dataset, YOLO had magnificent results on detecting human figures. They also discovered that Mask R-CNN takes much longer to detect human figures, whereas YOLO can detect any form of human figure. YOLO can be used to detect any form of item in real time and is believed to be the better model of the two (Shahriar et al. 2020).

Wei J. et al.(2021) in their work, they focus on motorcylce driver's safety in accidents. Since helmets are crucial for motorcycle drivers, helmets may save driver's life. Therefore, they tried to find a solution by detecting non-helmet users on the road with using real time traffic cameras. They made an automated system for this purpose. Also they improved their model with using YOLOv5 Detector. The upgraded YOLOv5 detector is used in the initial phase of their research to detect motorbikes from video surveillance. The second stage continues to employ the upgraded YOLOv5 detector to detect whether the motorcyclists are wearing helmets, using the motorbikes detected in the previous step as input. At the end of their work the proposed method is verified by experiments and compared with other methods. Their method achieved mAP of 97.7%, F1-score of 92.7% and frames per second (FPS) of 63 , which outperforms other detection methods. They also indicated that in the future, they may add a tracking algorithm to YOLOv5 so that the same item is only detected once, which is critical for real-time traffic video monitoring.(Jia et al. 2021).

Kasper-Eulaers M. et al.(2021) in their work, they pointed out that for improving road safety, it is important to organize rest areas for truck drivers. It is easier with using YOLO Framework with real time cameras. They utilized heavy vehicle detection at rest areas throughout the winter to enable for real-time forecast of parking place availability to make rest facilities better for truck drivers. With that their rest time become more useful for heavy truck drivers. They created The Rest Area Dataset in order to accomplish their task. They installed a real time camera in a fixed position. They first demonstrate an advancement in order to detect heavy goods vehicles using their front side and rear side instead of the entire vehicle when winter wheather conditions result in challenging images with a large number of overlaps and cut-offs, and secondly, they demonstrate thermal network imaging to be hopeful in vehicle detection as a result of their work (Kasper-Eulaers et al. 2021).

Yao et al.(2021) in their work, they mentioned that detecting defection is very important process for kiwifruit. Moreover, detecting small defection is even harder to do. These problem can be solved with real time detection by using YOLOv5 Framework. To address these issues, they created a defect detection model based on YOLOv5, which can identify faults reliably and quickly. The findings of their study suggest that the network's overall performance has improved. When compared to the original algorithm, YOLOv5's mAP@0.5 attained 94.7% , which was a nearly 9% improvement. Our model just takes 0.1 second to recognize a single picture, demonstrating its efficacy. As a result, YOLOv5 meets the real-time detection criteria and provides a solid approach for the Kiwi defect detection system(Yao et al. 2021).

Wenbo Lan et al.(2018) in their work, they worked on pedestarian detection system. However, they figured out that when they train their network there are some loss of information about pedestrian information and it cause the disappearance of gradients and make the network predict inaccurately. Therefore, they were working on improving their network structure and they named it as YOLO-R. They applied Three Passthrough layers to the existing YOLO network. The Passthrough layer is made up of two layers which are the Route layer and the Reorg layer. They created for to integrate high and low resolution pedestrian features, as well as shallow and deep layer pedestrian features. Their Route layer's purpose is to provide the current layer's pedestrian characteristic information, then use the Reorg layer to reorganize the feature map such that the currently

introduced Route layer feature may be matched with the following layer's feature map. Their three Passthrough layers used in this approach may successfully convey the shallow pedestrian great properties of the network to the deep network, allowing the deep network to better learn shallow pedestrian feature information. To strengthen the network's power to extract information from shallow pedestrian features, the layer number of the Passthrough layer link in the original YOLO approach is also altered from Layer 16 to Layer 12 in their study. They evaluated their improvements on the network on the INRIA pedestrian dataset. Their experiments show that their approach improves pedestrian detection accuracy while lowers false detection and missed detection rates, and detection speed can reach 25 frames per second(Lan et al. 2018).

Dewi et al.(2021) in their work, they tried to create identification system of traffic lights with using Conveolutional Neural Network. They used their own dataset since different countries use different signs. Their study blends synthetic and original images to improve datasets and test synthetic datasets' efficiency. For training, they employed various numbers and sizes of images. They used YOLO V3 and YOLO V4 for their work. The recognition performance was increased by blending the genuine image with the synthetic image, obtaining an accuracy of 84.9% on YOLO V3 and 89.33% on YOLO V4.

As mentioned before, there are many object detection researches with different version of YOLO framework. As can be seen, the higher the version, the higher the success rate for object detection. Even in comparison researches between different frameworks or algorithms, YOLO versions reaches the best mAP rate generally. Although there are detection studies for many different objects, there is no study on cigarette butts. Even if YOLO frameworks earlier version has some problems with detecting small objects such as cigarette butts, the latest version of YOLO framework will reach promising success rate.

3. MATERIAL AND METHOD

As discussed in the literature and background section, it is known that feature extraction is crucial in ML and DL techniques. Moreover, both techniques needs data as key point of their processes. Data must be in a form that model can be created from. Also ML techniques needs human for feature extraction and data preparation. Unlike ML, DL does not need human for detecting the most important features to create model from data that is given or collected for the purpose of that study. It makes feature extraction process by itself. It detects the key points of the image or given data by itself. Therefore, automated systems or real time camera detection systems generally prefer to work with DL algorithms.

DL algorithms needs big datasets to detect important notion about the given data such as image or audio. Since size of the datasets are too large, it takes more time for training stage. In our case, there is a need of dataset which include at least 2000 images. Since there are a lot images in the dataset, powerful computers with strong GPU needed. There is also one more requirement to accomplish the task of our study. That is a platform to run framework on it without any interrupt. Working environments such as Google Colaboratory, Paperspace Gradient, Kaggle, Amazon SageMaker, Floydhub can provide services for ML researches without any interruption. All of them contain and host machines with high GPU bar. In this thesis, Google Colaboratory will be preferred because it supports Jupyter Notebook and provides consistent and high service quality in computers which has high-capacity GPU.

3.1. Gathering Data

There are different types of filters for cigarettes such as different shapes, colour etc. Therefore, the dataset that we create must have different types of filters to make the classifier better generalization for detecting cigarette's butt. Also, areas that human can be in at any time should have been covered in dataset. That means the dataset have to have images in different areas such as streets, agricultural areas, ground and even on grass. Since there is an important element of training stage and creating our image classifier, there should be different images in different areas. Moreover, to obtain images from different devices, we collected the dataset using 4 different devices. Part of the dataset

created by optical cameras, the rest is collected by mobile devices.

3.2. Data Preparation

In this section, the preparation process of collecting images to create a dataset are explained. In order to execute any ML or DL algorithm, data should have been prepared. There should be no blank information in any row of the data set. However, for the purpose of this case, the dataset should have contain images that does not have cigarette's butt inside of it. Also there are some blurry images. These types of images should have cleaned from noise. Such images should have been de-blurred.

These steps include preparing the dataset, labeling, and preprocessing of the dataset. To ensure our learner can handle different background and cigarette's butt(orange, white), we researched the Internet for a dataset. Even though there are some datasets such as Immersive Limit Cigarette Butt dataset, which are either synthetically composed images of cigarettes on the ground or almost the same pictures taken from video frames so we decided not to use them as dataset. Then, every image, that in the dataset, is manually taken for labeling process.

Instead of using frames from videos, real life images taken and used as dataset for detecting cigarette's butt. Although the idea of creating a dataset may seem like a simple case, there is a wide variety of cigarette butts in different colors and sizes. Today, cigarette butts are produced in 2 different colors, white and orange. In addition, some cigars or different tobacco group products have a wide variety of colors. The Figure 3.10 shows some examples.

Figure 3.10. Cigarette butts in different colors

The size of a cigarette butt waste can vary depending on the situation or the person who throws it. In addition, the status of cigarette waste may differ from each other. They may still be lit or extinguished. In this study, all different cases were tried to be added to the manually created dataset. The Figure 3.11 , 3.12, 3.13 shows different kind of cigarette butt wastes.

Figure 3.11. Cigarette butts in different shapes

Figure 3.12. Lit cigarette butt

Figure 3.13. Extinguished cigarette butt

2100 image used in total for this work. A non-synthetic dataset was prepared, with all photos taken in real life. The images were distributed in different environments and in different situations. All images are 1000×800 pixels. According to the Pareto Principle, we separated our dataset into validation and training datasets in an $80 \ / \ 20$ rate rule. The Pareto Principle says that around 80% of the consequences are caused by 20% of the causes. Therefore, there are 1600 images for the train process and 400 for validation.

Also, after getting weights, there are 100 more images for the test process. The Table 3.2 shows image distribution respect to Pareto Principle.

Table 3.2. Image distribution respect to Pareto Principle

Number of Images	Training Set	Validation Set	Test Set
2100	1600	400	100

3.2.1. Labeling process

makesense.ai is used for labeling process. It is working on Web Browser without uploading any images. It saves a lot of time for labeling purposes such as image recognition and object detection which is our case. It is possible to label more than one cigarettes in the same image. There are many samples which include more than 10 cigarettes in the dataset. There is also possible to label more than one class so detection more than one class is possible with it. It also outputs the format YOLO Framework needs and .csv file format. Every image is labeled for training process of our classifier and YOLOv5 Framework. The figure below shows the labeling process of an image which is in the dataset.

At the end of this process we get the txt file that contains the height, width, X, Y for the cigarette butt in every image. It also outputs the height, width, X, Y for the cigarette butt in .csv format.

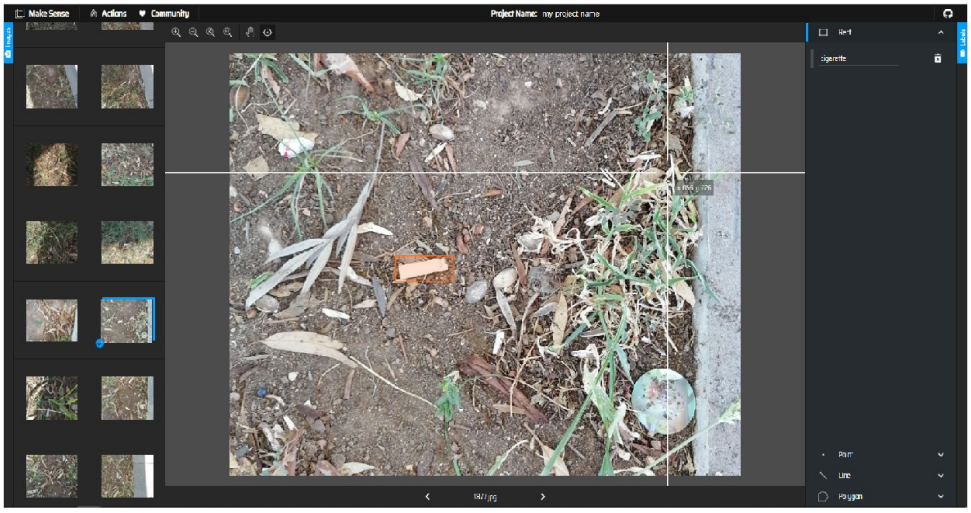

Figure 3.14. Labeling procedure with makesense.ai

3.2.2. Setting up working environment

There is a need for strong computers with GPUs since there is big dataset for training process. To speed up the training process we need a faster GPU and more RAM. Therefore we used Google Colaboratory Environment. Therefore we are able to work on 2100 pictures in 18 batches which use 16GB RAM for GPU.

3.3. Training Process

In the experiments that made by researchers, Adam optimizer can work well on smaller custom datasets, and can provide good initial results on larger datasets, whereas Stochastic Gradient Descent(SGD) tends to outperform in the long run, especially on larger datasets, and seems to generalize better to real world results. Therefore, SGD is selected as an optimization algorithm.

$$\theta_j = \theta_j - \alpha(\hat{y}^i - y^i)x_j^i \tag{3.1}$$

The batch size is chosen 18 for the Tesla P100-PCIE-16GB machine on Google Colaboratory Environment. The batch size means the number of samples that propagates through the network before updating the model parameters. Each batch goes through one full forward and backward propagation. That means 18 will be taken at a time to train the network. For going through every 1600 images that we divided as training set, it takes 88 iterations which is 1 epoch.The network model has trained 60 epochs. That means training process continues 60 times(epochs).

Batch size directly effects the time of training process. However, it may be limited by RAM and GPU size. Even if the network for our case is working on Google Colaboratory platform, it is still limited since there is a big need for resources to do object detection on the images. Although, larger batch sizes have faster progress in training, they do not always converge as fast. Even though small batch sizes train slower, they can converge faster.

The Table 3.3 shows the parameters of training process.

Table 3.3. Parameters for training process

Number of Images	Batch Size	Iterations	Epochs
1600	18	66	60

There are many types of pretrained checkpoints which are YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5n6, YOLOv5s6, YOLOv5m6, YOLOv5l6, YOLOv5x6+TTA.

YOLOv5x, which is XLarge model, pre-trained model were picked for this case to fit our study. After that with the dataset, we created manually, the machine will learn the features of our model from the labels that we made by itself. In this case, 30-35 epochs is enough to learn our model. However we tried with 60 epochs to get better results.

Intersection Over Union(IoU) metric used and set its rate as 0.72. It is commonly accepted and total pixel based accuracy metrics to measure the quality of the segmentation process(Danışman 2020).

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{3.2}$$

In the training process, validation batch prediction results are between 0.3 to 0.9 in confidence. Most of them are more than 0.7. However, we set the confidence threshold to 0.33 for not capturing the predictions even if there is less confidence rate. We also set the Intersection over Union(IoU) rate to 0.72 to get better results with 60 epochs. It takes 3.52 hours to get the best weights from our dataset.

The model checkpoint class is used to select optimal model weights with respect to the validation loss value. The Figure 3.15 shows train loss and validation loss values for 60 epochs.

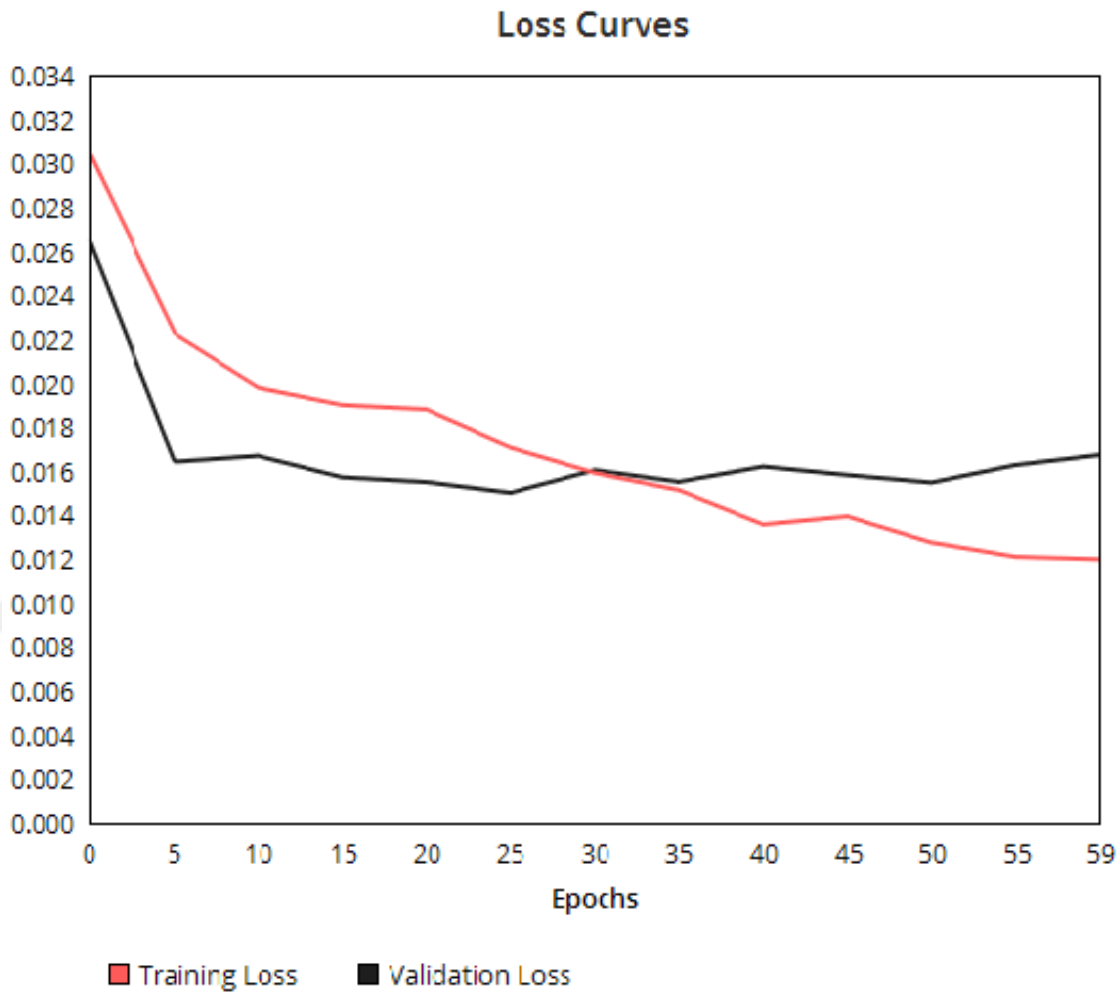

Figure 3.15. Loss curves after 60 epochs

After training the network, there is a file called as weights. There are last weight and the best weight of the training process. By using the best weights it is possible to validate the model by using n fold cross validation technique and test the model and network with new images.

4. RESULTS AND DISCUSSION

In this thesis, I experimented detecting cigarette butts by using YOLO Framework. Even with some limitations for the YOLO Framework, especially in detecting small objects, I obtained good results.

4.1. Testing the Network

After the CNN has been trained, the trained model's generalization capacity should be assessed. However there are some problems with splitting data into training and test data. To eliminate the problem that is created by seperating dataset into training and validation dataset as one group, I tried to use n fold cross validation technique to validate my model. One separates whole dataset into five or ten groups generally when using N fold cross validation technique. I separated my dataset into five subset and one of these subgroups will be the validation set and the rest will be train set. For each time, validation set will change with another subset so that every image in the dataset will be used and it will be possible to validate the model that we trained.

Also I want to test my network with another test set which includes 100 images that contains cigarette's butt. However, some of them are not contain any cigarette waste. It is obvious that this is done for testing purposes.

By using weight that we get from training, we are able to detect our model in a given image, video or live stream video as it is shown in the Figure 4.16.

Figure 4.16. Detecting cigarettes butt in different areas

4.2. Results

There are some measurement metrics and some mathematical equations to find accuracy, sensitivity, specificity, precision about the process that the network done which is object detection. Moreover, confusion matrices is used to measure these parameters. At the end of this process, there is accuracy rate is found. Some definitions of these parameters is explained in the Table 4.4.

Table 4.4. Definition of parameters

Parameter	Definition
True Positive (TP)	Number of images that classify cigarette's butt correctly
True Negative (TN)	Number of another waste images correctly classified
False Positive (FP)	Number of another waste images incorrectly classified
False Negative (FN)	Number of cigarette's butt images incorrectly classified

These parameters will lead us to find accuracy, sensitivity, specificity, precision. For finding these, we need some calculations with the parameters that we mentioned.

$$Accuracy = \frac{TP + TN}{(TP + TN) + (FP + FN)} \tag{4.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$Recall = \frac{TN}{TN + FP} \tag{4.5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.6}$$

$$F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.7}$$

As I mentioned earlier, I used n fold cross validation technique for validating my model. I used five fold cross validation. As a result, the dataset was divided into five subsets. There were 420 images in these subsets. Then, each time, one of these subsets will be used as a test set for validating the trained model. After training five times with different validation sets, I get confusion matrix that can be seen in 4.5.

Table 4.5. Five fold cross validation confusion matrix

n=420		Prediction	
		Cigarette	Not Cigarette
Truth	Cigarette	TP: 400	FN: 4
	Not Cigarette	FP: 11	TN: 5

Table 4.6. Results of the five fold cross validation

Accuracy	Precision	Recall	F1 Score
0.964	0.973	0.99	0.981

After getting promising validation results, there are also other metrics which are important the measure performance of the model. The mean average precision (mAP) is used to evaluate models in the field of object detection. By comparing the ground-truth bounding box to the detected box, the mAP derives a score. This metric measures how well the model detected a ground-truth bounding box. The model is more accurate if the final score is higher. In our study, we have reached 0.8897 in metric/mAP_0.5 which is the most important metric. It is seen the Figure 4.17.

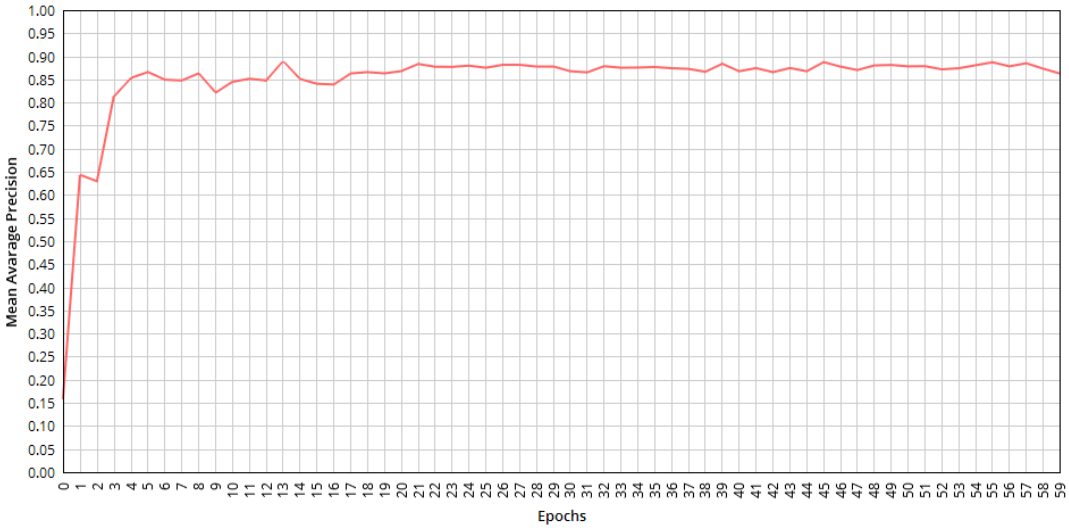

Figure 4.17. Mean average precision(mAP)

The better mAP results depend on loss function results. In the training process our box_loss value decreased to 0.014897 and obj_loss value decreased to 0.0072. Also, in

the validation process box_loss is 0.026 and obj_loss is 0.0108. If an object is detected in the grid cell, the loss function penalizes classification error (Redmon et al. 1990). Number of images in the process matters. We also have these results shown in the Figure 4.18 for metric/precision and in the Figure 4.19 metric/recall.

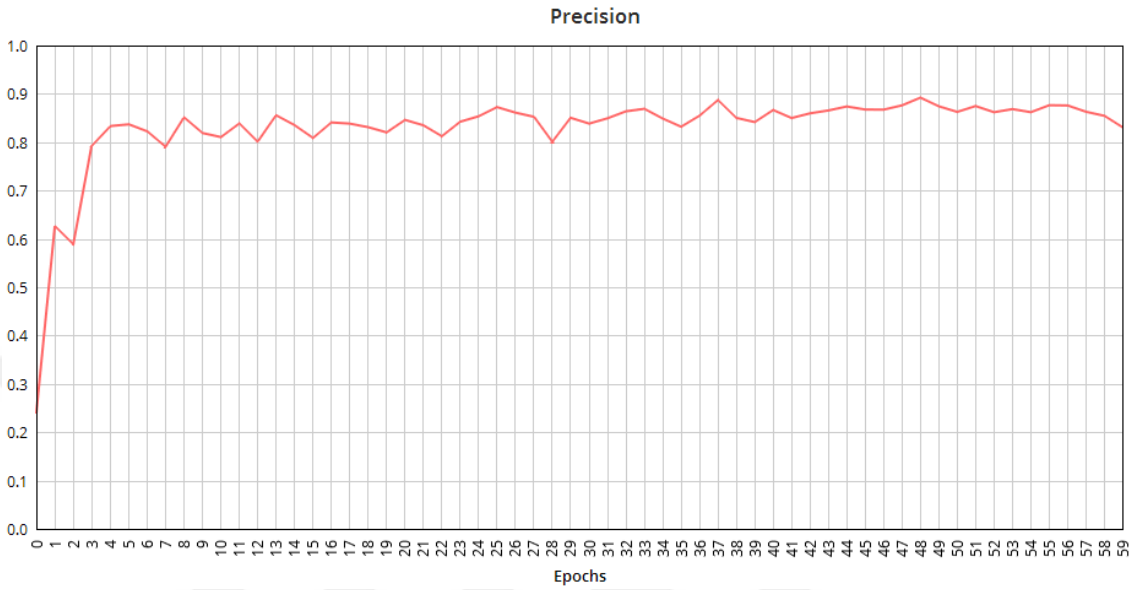

Figure 4.18. Metric/precision graph

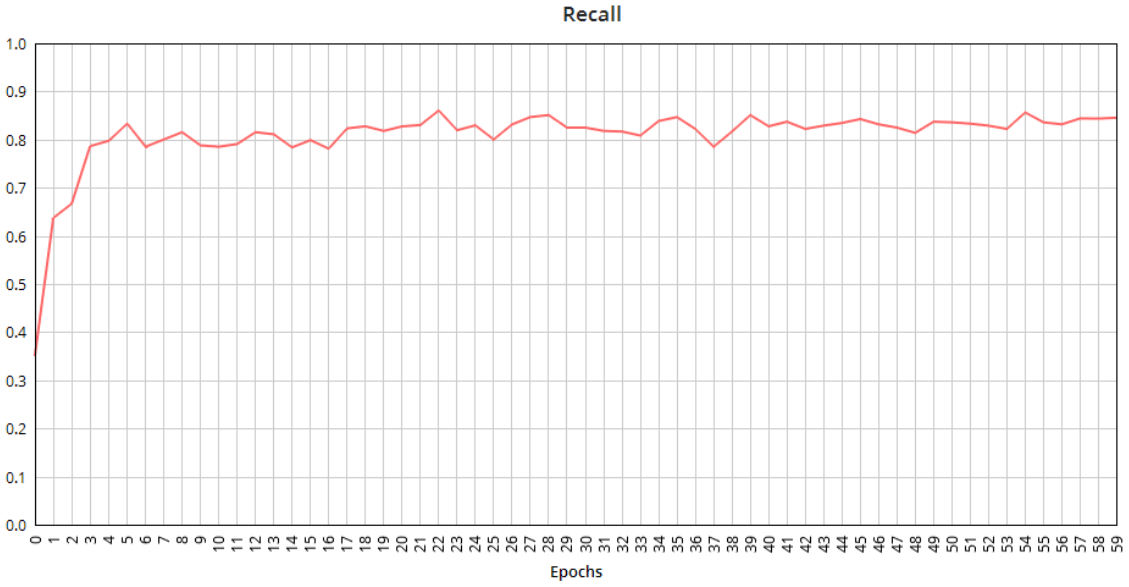

Figure 4.19. Metric/recall graph

Another important metric is the Precision-Recall curve which should be like that the recall increases, the precision decreases since the number of positive samples increases,

accuracy decreases. This is an expected situation. When a looking Precision Recall curve graph, it is easy to determine which point is the highest precision and recall value. The Precision-Recall curve graph, Precision-Confidence, the Recall-Confidence and F1 curve graph can be seen in the Figure 4.20.

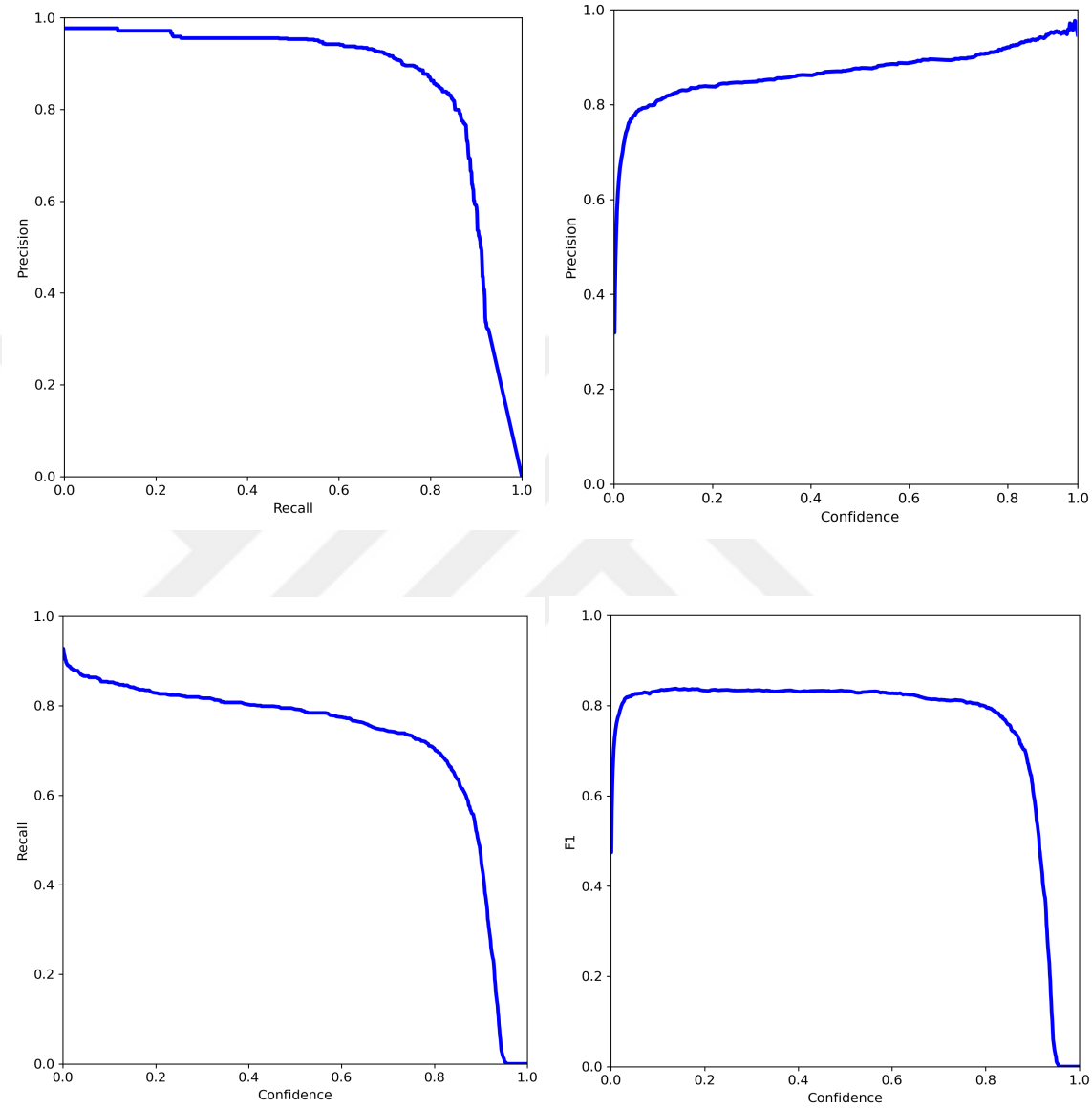

Figure 4.20. Metric graphs

After getting the weights and results that I showed above, I wanted to test 100 more pictures to try the model. I took another 100 images which 90 of them include cigarettes, 10 of them do not include cigarettes inside of images. Our model detects 90 images as True Positive from our test images. It also detects 3 of them as False Positive(different

object as cigarette) and 7 of them as True Negative. None of the images detected as False Negative. The Table 4.7 shows the confusion matrix of our test images after the detection process.

Table 4.7. Confusion matrix

n=100		Prediction	
		Cigarette	Not Cigarette
Truth	Cigarette	TP: 90	FN: 0
	Not Cigarette	FP: 3	TN: 7

Finally, the performance metrics of our network is shown in the Table 4.8.

Table 4.8. Results of the network

Accuracy	Precision	Specificity	F1 Score
0.97	0.967	0.7	0.983

There are detection and confidence value of test images shown in the Figure 4.21.

Figure 4.21. Results of test images

Also, false positive sample shown in the Figure 4.22.

Figure 4.22. False positive sample

5. CONCLUSION

Living in a non-polluted area is crucial for human beings and any living being. Cigarette butts are one of the most dangerous reasons for environmental pollution. Public organizations are trying to prevent the pollution that exists. Also cigarette's butts create pollution on the streets and many other places that we live in our daily life. Because of the waste that created by cigarette butts, public organization and cleaning services needs more human resources to handle with the pollution. Increasing the number of people in cleaning service creates a problem such as budget and cost of maintenance of additional workers. Therefore this situation significantly increases the cost. To solve this problem and decrease the cost of the additional ones, we may use the technology. We may use machines such as computers, cameras, robots, drones to speed up the process of making environment clean.

In our study, a network is designed for cigarette butt detection. It has been shown that Machine Learning based on Cigarette's Butt Detection using YOLO Framework enables us to detect cigarette's waste at a good rate. With 1000×800 resolution images, we obtained promising results such as 89% mean average precision(mAP), precision of 95% and recall of 93%. F1 score of 98%, precision of 96% and accuracy of 97% were calculated on the testing dataset that manually created for this work.

With the network created in this thesis, it is possible to develop integrated system that can detect cigarette's butt in real time and also clean the waste immediately when it is detected. Morever, by increasing the number of images in the dataset, detecting becomes even more accurate. By using this network, it can lead to an autonomous system that detects cigarette butts in the popular streets. So, it makes our environment in city life healthier.

As future work, autonomous system which is integrated with the model can easily detect and clean the areas which is polluted. In other words this work can be extendable to create autonomous systems to clean our streets and needed areas. It can be used by many public organizations. It is also can be used for detecting the waste on beaches by drones and clean the waste by autonomous systems or human workforce from cigarette's waste. Also, by adding GPS Location stickers in video taken by autonomous systems, it will be easier to detect which location is more polluted. It makes also create heat map

of cigarette's butt waste so it can be evaluated which place needed the clean more easier. Moreover, labor can be saved by shifting the workforce to other areas and by reducing the high budget of cleaning service, public organizations will shift their resources to other important things.

6. REFERENCES

Abrol, S., and Mahajan, R. 2015. Artificial neural network implementation on fpga chip. *Int J Comput Sci Inf Technol Res*, *3*, 11–18.

Albawi, S., Mohammed, T. A., and Al-Zawi, S. 2017. Understanding of a convolutional neural network.

Alemu, H., Wu, W., and Zhao, J. 2018. Feedforward neural networks with a hidden layer regularization method. *Symmetry*, *10*, 525. https://doi.org/10.3390/sym10100525

Aydin, I., and Degirmenci, C. 2018. Yapay zeka, 106–107.

Basaran, E., Comert, Z., Y., C., Velappan, S., and Togacar, M. 2020. Determination of tympanic membrane region in the middle ear otoscope images with convolutional neural network based yolo method. *DEUFMD*, *66*, 919–928.

Bengio, Y., A.Courville, and P.Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *35*, 8., 1798–1828.

Cagil, G., and Yıldırım, B. 2020. Detection of an assembly part with deep learning and image processing. *Zeki Sistemler Teori ve Uygulamaları Dergisi*, *3(2)*, 31–37.

Caglayan, A. 2018. A new horizon in econometrics: Big data and machine learning. *Social Sciences Research Journal*, *7*, 41–53.

Chao, W. 2011. Machine learning tutorial. https://tcxsproject.com.br/dev/Biblioteca%5C%20Livros%5C%20Hacker%5C%20Gorpo%5C%20Orko/Machine%5C%20Learning%5C%20Tutorial.pdf

Cınar, U. K. 2020. Derin öğrenme algoritmalarını kullanarak görüntü sınıflandırma.

Danışman, T. 2020. Segmentation of portrait images using a deep residual network architecture. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *22(65)*, 569–580.

Danışman, T., Bilasco, I. M., Martinet, J., and Djeraba, C. 2013. Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron. *Special issue on Machine Learning in Intelligent Image Processing*, *93(6)*, 1547–1556.

Dewi, C., Chen, R.-C., Liu, Y.-T., Jiang, X., and Hartomo, K. D. 2021. Yolo v4 for advanced traffic sign recognition with synthetic training data generated by various

gan. *IEEE Access*, *9*, 97228–97242. https://doi.org/10.1109/ACCESS.2021.3094201

Dikbayir, H., Bülbül, S., and H.İ. 2020. Derin öğrenme yöntemleri kullanarak gerçek zamanlı araç tespiti. *Türk Bilim Araştırma Vakfı*, *13:3*, 1–14.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. 2021. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Erler, M., Sagiroglu, S., and Besdok, E. 2003. Mühendislikte yapay zeka uygulamaları 1 / yapay sinir ağları.

Fan, Q., and et al. 2017. A generic deep architecture for single image reflection removal and image smoothing. *Proc. EEE Int. Conf. on Comput. Vision*, 3238–3247.

Fan, Q., and et al. 2018. Image smoothing via unsupervised learning. *ACM Trans. on Graph. (TOG)*, *37*, 1–14.

Galal, A. 2016. An analytical study on the modern history of digital photography. *Int. J. Des.*, *26*, 1–13.

Goodfellow, I., Y.Bengio, and A.Courville. 2016. Deep learning. *MIT Press*, *1*.

Halıcı, U. 2004. Artificial neural networks,chapter 9 radial basis function networks. *EE543 LECTURE NOTES. METU EEE*, 139–147.

Hariri, M. 2021. Detection of calcium deficiency and physiological status in strawberry leaves using deep learning.

Ibadullayeva, J., Jumaniyazova, K., Azimzadeh, S., Canıgür, S., and Esen, F. 2019. Çevre kirliliğinin İnsan sağlığı üzerindeki etkileri. *Türk Tıp Öğrencileri Araştırma Dergisi*, *1*, 52–58.

Jia, W., Xu, S., Liang, Z., Zhao, Y., Min, H., Li, S., and Yu, Y. 2021. Real-time automatic helmet detection of motorcyclists in urban traffic using improved yolov5 detector. *The Institution of Engineering and Technology*, *15*, 3623–3637. https://doi.org/10.1049/ipr2.12295

Kartal, E. 2015. *Sınıflandırmaya dayalı makine öğrenmesi teknikleri ve kardiyolojik risk değerlendirmesine İlişkin bir uygulama*, Master's thesis. İstanbul Üniversitesi. İstanbul.

Kasper-Eulaers, M., Hahn, N., Berger, S., Sebulonsen, T., Myrland, Ø., and Kummervold, P. E. 2021. Short communication: Detecting heavy goods vehicles in rest areas in winter conditions using yolov5. *Algorithms*, *14*, 4. https://doi.org/10.3390/a14040114

Kotsiantis, S. 2007. Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268.

Lan, W., Dang, J., Wang, Y., and Wang, S. 2018. Pedestrian detection based on yolo network model. *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, 1547–1551. https://doi.org/10.1109/ICMA.2018.8484698

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*, *521*, 436–444.

LeCun, Y., Boser, B., Denker, J., D.Henderson, R.E.Howard, W.E.Hubbard, and Jackel, L. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

LeCun, Y., Boser, B., Denker, J., D.Henderson, R.E.Howard, W.E.Hubbard, and Jackel, L. 1990. Handwritten digit recognition with a backpropagation network. *In Advances in neural information processing systems*, 396–404.

Li, M., Z.Zhang, L.Lei, X.Wang, and X.Guo. 2020. Agricultural greenhouses detection in high-resolution satellite images based on convolutional neural networks: Comparison of faster r-cnn, yolo v3 and ssd. *Sensors*, *20*.

Liu, W., and et al. 2020. Improvement of image binarization methods using image preprocessing with local entropy filtering for alphanumerical character recognition purposes, 562.

Ma, W., and et al. 2019. Adaptive median filtering algorithm based on divide and conquer and its application in captcha recognition. *Comput., Mater. and Continua*, *58*, 665–677.

Michalak, H., and Okarma, K. 2019. Improvement of image binarization methods using image preprocessing with local entropy filtering for alphanumerical character recognition purposes. *Entropy*, *21*, 562.

Mishra, V., AGARWAL, S., and PURI, N. 2018. Comprehensive and comparative analysis of neural network. *INTERNATIONAL JOURNAL OF COMPUTER APPLICATION*, *2*. https://doi.org/10.26808/rs.ca.i8v2.15

Mitchell, M., A., D. J. G., K., Y. K. S., X., H., A., M., and Daume, B. H. 2012. Generating image descriptions from computer vision detections. *13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756.

Perihanoglu, G. 2015. *Feature extraction from images by using digital image processing techniques*, Master's thesis. Istanbul Technical Univ.

Redmon, S., Divvala, R., Girshick, and Farhadi, A. 1990. Handwritten digit recognition with a backpropagation network. *In Advances in neural information processing systems*, 396–404.

Rush, N. 2016. Deep learning. *Nature Methods*, *13*, 35.

Sahin, F. 1997. *A radial basis function approach to a color image classification problem in a real time industrial application*, Master's thesis. Virginia Polytechnic Institute and State University. Virginia.

Samuel, A. 1988. Some studies in machine learning using the game of checkers, 210–229.

Sarker, and Iqbal. 2021. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*. https://doi.org/10.1007/s42979-021-00815-1

Seetharam, K., D, B., PD, F., and PP, S. 2020. Role of artificial intelligence in cardiovascular imaging: State of the art review. front. *Front. Cardiovasc. Med.* https://doi.org/https://doi.org/10.3389/fcvm.2020.618849

Shahriar, S., W., J., Anuvara, R., and R., D. 2020. *Journal of Physics: Conference Series*.

Sucu, I., and Ataman, E. 2020. Dijital evrenin yeni dünyası olarak yapay zeka ve her filmi üzerine bir çalışma. *e-Journal of New Media*, *4*, 40–52.

Sutton, R. S., and Barto, A. G. 2014. Reinforcement learning: An introduction.

Wang, H., Chen, X., Wu, Q., Yu, Q., Hu, X., Zheng, Z., and Bouguettaya, A. 2017. Integrating reinforcement learning with multi-agent techniques for adaptive service composition. *ACM Transactions on Autonomous and Adaptive Systems*, *12*, 1–42. https://doi.org/10.1145/3058592

Xu, R., Lin, H., Lu, K., Cao, L., and Liu, Y. 2021. A forest fire detection system based on ensemble learning. *Forests*, *12*, 217.

Yang, Teo, C.L., H., D., and Y., A. 2011. Corpus guided sentence generation of natural images. *Conference on Empirical Methods in Natural Language Processing*.

Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., and Li, X. 2021. A real-time detection algorithm for kiwifruit defects based on yolov5. *Electronics*, *10*, 14. https://doi.org/10.3390/electronics10141711

CURRICULUM VITAE

Hasan Ender YAZICI

EDUCATION DETAILS

Graduate 2018-2022	Akdeniz University Institute of Natural and Applied Sciences, Department of Computer Engineering, Antalya
Undergraduate 2010-2013	Technical University of Sofia Computer Science Computer Engineering, Sofia

WORK EXPERIENCES

Bilgisayar Mühendisi 2017-Present	Antalya Gıda Kontrol Laboratuvar Müdürlüğü
Programcı 2013-2017	T.C. Tarım ve Orman Bakanlığı Bilgi İşlem Dairesi Başkanlığı