LAKEHEAD UNIVERSITY

MASTER THESIS

Improving Cataract Surgery Procedure using Machine Learning and Thick Data Analysis

Author:
Chandrashekhar Singh

Supervisor:
Dr. Jinan Fiaidhi

A thesis submitted in fulfillment of the requirements for the degree of Msc in computer science

in the

Department of Computer Science

January 6, 2023

LAKEHEAD UNIVERSITY

Abstract

Department of Computer Science

Msc in computer science

Improving Cataract Surgery Procedure using Machine Learning and Thick Data Analysis

by Chandrashekhar Singh

Cataract surgery is one of the most frequent and safe Surgical operations are done globally, with approximately 16 million surgeries conducted each year. The entire operation is carried out under microscopical supervision. Even though ophthalmic surgeries are similar in some ways to endoscopic surgeries, the way they are set up is very different. Endoscopic surgery operations were shown on a big screen so that a trainee surgeon could see them. Cataract surgery, on the other hand, was done under a microscope so that only the operating surgeon and one more trainee could see them through additional oculars. Since surgery video is recorded for future reference, the trainee surgeon watches the full video again for learning purposes. My proposed framework could be helpful for trainee surgeons to better understand the cataract surgery workflow. The framework is made up of three assistive parts: figuring out how serious cataract surgery is; if surgery is needed, what phases are needed to be done to perform surgery; and what are the problems that could happen during the surgery. In this framework, three training models has been used with different datasets to answer all these questions. The training models include models that help to learn technical skills as well as thick data heuristics to provide non-technical training skills. For video analysis, big data and deep learning are used in many studies of cataract surgery. Deep learning requires lots of data to train a model, while thick data requires a small amount of data to find a result. We have used thick data and expert heuristics to develop our proposed framework.Thick data analysis reduced the use of lots of data and also allowed us to understand the qualitative nature of data in order to shape a proposed cataract surgery workflow framework.

Acknowledgements

I would like to show my gratitude to my supervisors Dr. Jinan Fiaidhi and Dr. Sabah Mohammed. Thank you for spending all those hrs to guide and mentor me through out the process. You are knowledgeable, resourceful and very supportive. Thank you for your patience while sending you numerous emails to clear my doubts.I would like to be grateful to all of my professors for transferring their knowledge to us and enriching our understanding of the course materials. I would like to thank the examination committee for reviewing the thesis report and assisting me with their valuable feedback ...

Contents

Abstract **ii**

Acknowledgements **iii**

1 Introduction **1**
1.1 Background 1
1.1.1 Why Doing Analytics on Cataract? 2
1.1.2 Identify Actions In a Video To Analyze It 3
1.1.3 Why Segmenting Cataract Surgeries Videos is Advantageous? 3
1.1.4 Importance of Analyzing the Phases of Cataract Surgeries 4
1.1.5 The importance of ML and Thick Data Analytics in developing framework 6
1.1.6 Problems in training method for Cataract surgery . . . 6
1.1.7 Research Questions 7
1.2 Objective 8

2 Related Research Work **10**
2.1 Deep Learning application in Surgery 11
2.1.1 Computer vision applications in Surgery 11
2.2 Overview On ML/Deep Learning Research for Analyzing Cataract Videos: 11
2.3 Multi-label Classification of Surgical Tools with Convolutional Neural Networks 11
2.4 Code-free machine learning model for cataract surgery phases detection 12
2.5 Surgical phases extraction using Inception V3 in Real Time . . 14
2.6 Dataset used in this work 15
2.7 Methodology used in this work 16
2.8 Surgery Tools identification using Yolcat cnn 17
2.9 Laparoscopic Video Recognition Tasks using deep learning model EndoNet 18
2.10 Surgical Tool Detection Using Attention-Guided Convolutional Neural Network 19
2.11 Convolutional neural with multi-image fusion for surgical tool identification during cataract surgery 20
2.12 Thick Data Analytic for Small Training Samples Using Siamese Neural Network and Image Augmentation 21
2.13 The promise of big data technologies and challenges for image and video analytic in healthcare 22

2.14 Deep Bayesian networks with LSTM for surgical workflow analysis . . . 24
2.15 CNN methodology for Automatic Pupil and Iris Detection . . 24

3 Theoretical Background for Cataract Video Analytic 26
3.1 Thick Data Overview . . . 26
3.2 Convolution Neural Network . . . 26
3.3 Convolution Neural Network Architecture . . . 28
3.4 Pooling layers . . . 29
3.5 Activation Function used in CNN layers . . . 30
3.6 Dropout . . . 31
3.7 Batch Normalization . . . 31
3.8 Recurrent neural network . . . 31
3.9 Long Short Term Memory (LSTM) . . . 32
3.10 Transformer Model . . . 33
3.11 Transformer Attention Mechanism . . . 33
3.12 Mask R-CNN . . . 34
3.12.1 Semantic Segmentation . . . 35
3.12.2 Instance Segmentation . . . 36
3.12.3 Mean Average Precision (mAP) . . . 36

4 Developing a Methodology for Identifying Patterns from Cataract Videos 37
4.1 Overview . . . 37
4.2 Step 1. Cataract detection . . . 38
4.2.1 Dataset . . . 38
4.2.2 Training and model compilation . . . 39
4.3 Step 2: Video analysis of cataract surgery . . . 39
4.3.1 Dataset for Video Analysis . . . 39
4.3.2 Annotated CSV . . . 39
4.3.3 Phase CSV . . . 41
4.3.4 Video CSV . . . 42
4.4 Data Preprocessing for phases detection . . . 42
4.5 CNN-LSTM implementation for Phases Detection . . . 45
4.5.1 Training and Model Compilation. . . . 45
4.6 Transformer model implementation for Phase Detections . . . 45
4.6.1 Training and Model Compilation . . . 46
4.7 Step 3: Iris Pupil Detection . . . 46
4.7.1 Dataset For Iris and Pupil detections . . . 47
4.7.2 COCO Dataset . . . 48
4.8 Format of COCO Dataset . . . 48
4.8.1 Image List . . . 48
4.8.2 Annotations List . . . 49
4.8.3 Categories list . . . 49
4.8.4 Transfer Learning Using Mask R-CNN for iris pupil dataset . . . 50
4.8.5 Data Preprocessing for Iris and pupil dataset . . . 51
4.8.6 Training . . . 51

4.9 Thick data: Surgeon experience as a expert heuristic to improve model performance. 51

5 Results and Discussions **53**
5.1 Step 1. Cataract detection 53
5.2 Step 2 : Iris Pupil Results 54
5.2.1 Mean Average Precision Result for Mask R-CNN . . . 56
5.3 CNN-LSTM Result Analysis 57
5.4 Transformer Result Analysis 58
5.5 Limitation of predicted results of CNN-LSTM and Transformer 58
5.5.1 Github repositores of our tasks 59

6 Conclusion and Future Work **60**
6.0.1 Conclusion 60
6.1 Future Work 60
6.2 Outline 61
6.2.1 Chapter 1 61
6.2.2 Chapter 2 61
6.2.3 Chapter 3 61
6.2.4 Chapter 4 61
6.2.5 Chapter 5 61
6.2.6 Chapter 6 61

List of References **62**

List of Figures

1.1 Normal and Cataract eye 2
1.2 Frame Image of Hydrodissection phase 5
1.3 Frame Image of Irrigation/Aspiration phase 5
1.4 Frame Image of Phacoemulsification 5
1.5 Frame Image of Lens Implant phase 6
1.6 Cataract Surgery Workflow Framework 9

2.1 phase-cataract 13
2.2 Performance table for Auto ML 14
2.3 surgical instruments, lighting methods, and nuclear extraction techniques 15
2.4 Dataset breakdown of cataract surgery phases 16
2.5 images of three surgical phases from the dataset which was used in this work 16
2.6 segmentation masks' intersection-over-union and per-class average bounding box accuracy on the test set 18
2.7 Images shows predicted tip positions for cataract surgical tools,segmentation masks, and anatomical landmarks 19

3.1 A simple feed forward neural network image 27
3.2 Overall architecture of CNN 28
3.3 Convolution Operation 29
3.4 Max pooling between filter and feature map at stride 2. 30
3.5 Recurrent Neural Network 32
3.6 Architecture of Transformer 34
3.7 Attention Mechanism 35
3.8 Mask R-CNN 35
3.9 Segmentation 36

4.1 Stages Cataract detection 38
4.2 Stages Cataract detection 38
4.3 Stages Cataract detection 38
4.4 Video Data from Dataset 101 40
4.5 Data preparation 44
4.6 ResNet and DenseNet architecture 46
4.7 Dataset for iris and pupil segmentation 47
4.8 image information of coco like iris pupil dataset 49
4.9 Annotation list of Image ID 105 49
4.10 Category list of iris pupil coco like datset 50

5.1 Cataract Detection Confusion Matrix 53
5.2 Predicted image 53
5.3 Iris pupil detection result 1 55
5.4 Mean Average Precision of Mask R-CNN 56
5.5 Predicted result 57
5.6 Predicted result metrics Of CNN-LSTM 57
5.7 Predicted result metrics of Transformer 58
5.8 Accuracy and Loss graph of LSTM 59
5.9 Accuracy and Loss graph of Transformer 59

List of Tables

4.1 Phase CSV File Data Sample 40
4.2 Phase CSV File 41
4.3 Video CSV File Data Sample 42
4.4 Accuracy rate of CNN-LSTM and Transformer with two different dataset 52

5.1 Inference Configuration set up parameters 54

Chapter 1

Introduction

1.1 Background

One in seven Canadians is predicted to experience the onset of at least one of glaucoma, retinal disorders or cataracts, at some point in their lives.Given that the majority of vision loss and eye problems in Canada are preventable, these are frightening statistics. In reality, 75% of the nation's prevalent eye issues can be successfully addressed or avoided. A change in lifestyle, early detection, and treatment are frequently the keys to good eye health. Regular eye exams from an ophthalmologist are essential because so many eye disorders and major visual issues do not have any symptoms when they first appear [9].

An intensive cataract surgery training is important for new surgeon or students to perform cataract surgery. For a long time, ophthalmology training was based on Halsted's method. In this type of training, the trainee should: spend a lot of time caring for patients under the direct supervision of a qualified surgeon; learn the science behind the disease that needs surgical treatment; and develop the skills to do operations that get more complicated as they go on. Also, only people who had completed a predetermined number of procedures were considered able to accomplish surgery successful. But this method is time taking and trainee spends lots of time and energy to learn procedure because learning based only on the number of procedures done and direct practice with the patient which has some limits and risks. One of these problems is that the level of skill gained isn't always the same because there are different ways to learn, and one of the risks is that a patient might be treated by a surgeon who doesn't know what they're doing. There are different ways to teach cataract surgery For example, many new trainees use the animal eye model, in which pig or rabit'eyes use used, to understand the workflow of cataract surgery. In many countries , this type of practice is not admissible. Killing animals is not permitted and is frowned upon in many Islamic countries. Using our cataract framework based on machine learning , training could be easy for a new surgeon to understand cataract surgery workflow. Fig 1.1 shows the difference between normal and cataract eye. We can see in Normal eye light is focused sharply while in cataract eye, light is scattered or blocked by cloudy lens [7].

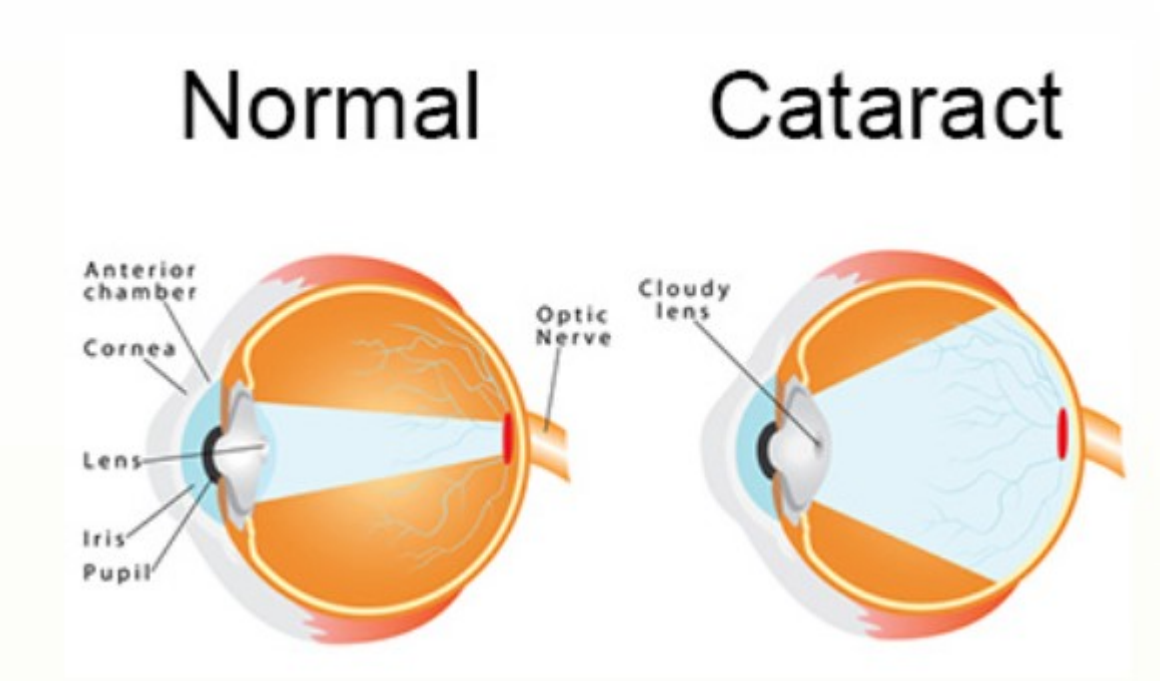


FIGURE 1.1: Normal and Cataract eye [34]

1.1.1 Why Doing Analytics on Cataract?

The surgery is one of the most common microscopic procedures in the area of ophthalmology. The purpose of this kind of surgery is to replace the human eye lens with an artificial one. The whole operation is conducted under microscope[4]. Currently, the recorded video is utilized for documentation and training purpose after surgery. It also involves analyzing surgical recordings to evaluate surgical proficiency. Another advantage of recording cataract is that they allow video analytic to be used to improve cataract surgical workflow. This video can be deleted after some time but using this video, we can make useful for a new trainee or surgeon. As part of surgical training in ophthalmology, doctors learn how to do cataract surgery very well. Surgical educators provide a training plan for ophthalmologists who are just going to start training [2]. The processing is very lengthy, and some surgeons do not show interest because of the length of training.

AI can be used to improve training for cataract surgery by identifying the different parts of the surgery on videos taken during surgery[16]. Videos of cataract surgery are often available to teachers and students, but they are not very useful for training right now. AI can be used to make tools that can easily break up videos of cataract surgery into its different parts so that automated skill testing and feedback can be done afterward[13]. Expertise in cataract surgery is important for public health. The cataract surgery videos are available at large scale. Surgical videos vary greatly in terms of image quality, objects in the field, and movement artifacts[43]. Surgeon experience is also important. Experienced surgeon perform surgery well and

complete surgery less than an average time. Other hand less experience performing surgery under supervision could take longer time than average time of surgery.

1.1.2 Identify Actions In a Video To Analyze It

Action recognition is the process of recognizing different actions from a series of two-dimensional frames of video. Action recognition is mostly a change from classifying a single image to classifying a series of images in a video. [35]. Frames are also images. Sequences of frames turns into small clip of videos. When we work with video data set, we create frames(images) to perform any deep learning task.

1.1.3 Why Segmenting Cataract Surgeries Videos is Advantageous?

Video analytics take video in real time and turn it into data that can be used to make decisions. They automatically make information that describes what is going on in the video and are used to find and follow objects in the video stream, which could be people, vehicles, or other items. Based on this information, actions are taken, Simple IP video is turned into business intelligence with the help of video analytic. Video analytic is a much better way to find incidents that are relevant to what we are looking for than watching video. The video classification method is now became very effective method to recognize various real world activities such as video summarize, traffic problems and facial recognition.Video analytic is also widely used in health care making surgical workflow easy and smooth. Using video analytic procedure will make our framework more effective for example recommend next suitable cataract surgery phases based on current predicted phase [40].

It is now common practice in the field of ophthalmology to record every surgery and keep the videos for documentation or educational purpose. On average, the surgery takes 5 to 10 minutes to finish. Some surgeons have a extensive experience and can do their tasks in even less time than the average. Because this surgery is done under a microscope, only the surgeon and one other person with extra glasses can see it. This makes it hard for trainee surgeons to learn, which is especially frustrating because ophthalmic surgery is one of the hardest kinds of surgery and requires special operation techniques and psycho motor skills that need to be trained intensively. The average surgical video duration is 3 to 5 min. Such a small duration video is not beneficial for understating entire surgical workflow. Segmenting the whole cataract video is very important to make it more understandable and more focused. The surgeon-in-training needs to carefully understand each phase of a cataract such as sequence of cataract ,next recommended phases , what complexity could be possible.

1.1.4 Importance of Analyzing the Phases of Cataract Surgeries

There are ten phases of cataract surgery. To complete the surgery, the surgeon must complete these ten phases. In each phase, different tiny tools are used. Understating each phase would allow students to become experts in cataract surgical workflow. The sequence of these phases is as follows:

- **Incision**: Surgeon uses a sharp blade to cut through the cornea, which gives instruments access to the inside of the eye. After the paracentesis, a "clear cornea incision" less than three mm wide is created. This incision is big enough to fit the phaco handpiece [30].

- **Viscous agent injection**: Ophthalmic viscoelastic agent gel is injected into the anterior chamber while performing surgery to preserve the depth of the chamber, shield the corneal endothelium, and maintain vitreous stability [30].

  Item textbfRhexis: The front of the lens capsule has been opened. The central radial cut is the first thing the surgeon does. After the cut, a tear is made, which lets the anterior capsule fold over itself. This tear is grabbed, and a flap is moved in a circle around it.[30].

- **Hydrodissection**:Fig 1.2 shows Hydrodissections. The surgeon injects electrolyte solution and epinephrin under the rhexis to separate the lens's periph- eral cortex from the capsule. This makes it easier for the nucleus to turn and hydrates the outer cortext[30].

- **Irrigation and aspiration**:Fig 1.3 shows irrigation/ aspiration phase. In this process doctor Uses irrigation fluid to keep the anterior chamber clean,and remove excess anterior chamber material[30].

- **Phacoemulsification**:Fig 1.4 shows phacoemulsification phase. The anterior central cortex is broken up by the ultrasonic power of the phaco tip. A deep, straight groove is cut through the center of the nucleus, and the lens is split in two. The lens is turned and cut into small pieces that can be mixed together. It is important to keep the posterior capsule whole during this procedure [30].

- **Capsule polishing**: To prevent capsule opacification, the posterior capsule is polished. Lens Implant Setting-Up:In this process surgeon Inserts the foldable artificial lens. and the lens is slowly unfolding and being pushed into the capsular bag is the lens [30].

- **Lens implant setting-up**:FIGURE 1.5 shows Lens implantation process. It is the process to set up IOL carefully.A clear artificial lens implant is placed into the capsule after the cataract has been removed. For the rest of the patient's life, the lens should be positioned such that it is just behind the pupil [28].

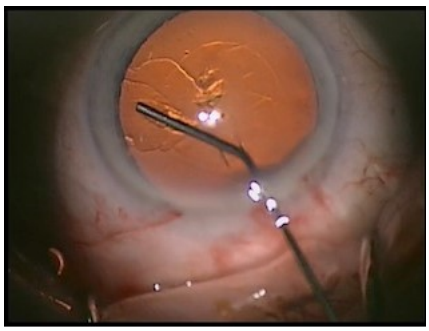


FIGURE 1.2: Frame Image of Hydrodissection phase [31]

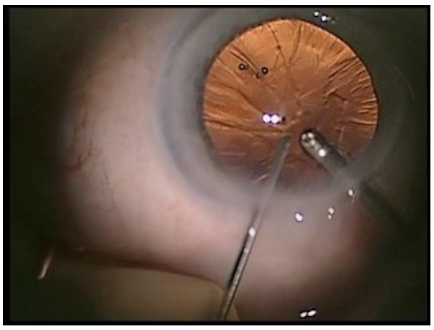


FIGURE 1.3: Frame Image of Irrigation/Aspiration phase [31]

- **Viscous agent removal**:In order to preserve the corneal endothelium during cataract surgery and to make intraocular lens (IOL) implantation easier, viscoelastic materials are often used [39].The viscous elastic agent is taken out of the front chamber and the capsule bag [30].
- **Tonifying and antibiotics**: This is the last process in which doctor put antibiotic which prevent from from endophthalmitis. Electrolyte solution and antibiotics are injected into the wound in the cornea. This causes the stroma to swell temporarily and the cut to close. A suture is only needed if it leaks. [30].

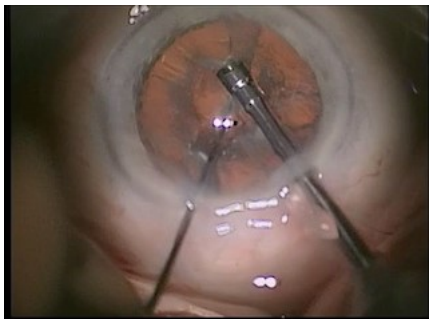


FIGURE 1.4: Frame Image of Phacoemulsification phase [31]

FIGURE 1.5: Frame Image of Lens Implant phase [31]

1.1.5 The importance of ML and Thick Data Analytics in developing framework

Thick data is made up of qualitative information, like observations, feelings, and reactions, that shows how people feel in their everyday lives. Thick data tries to find out about people's feelings, stories, and models of the world they live in. Thick data allows us to gain a deeper understanding of a dataset, such as emotions, experience. While big data is a lot of complicated, unstructured information, it is large in size, and to extract meaning and support information, further preprocessing is needed for unstructured and semi structured data sources, including text, audio, and video. To perform any deep learning task, we need a lot of data, which is either very expensive or not readily available. Thick data uses smaller samples of data to find patterns, whereas big data uses a lot of data to find patterns at a large scale in deep learning. We designed our surgical workflow framework using thick data analytics. The use of transfer learning enables one to avoid the requirement for a large amount of new data [11] because In transfer learning, a model that has already been trained on a task with plenty of labeled data. Transfer learning is one way to reduce the size of the datasets needed for training a deep learning model. Understanding the depth of insight in our dataset, such as surgeon experience, and the emotions of trainees such as difficulties, helped us design our framework perfectly.The surgeon experience heuristic assisted in the creation of a good dataset, and the model result was better for recommending next sequence of phase based on current predicted phase.

1.1.6 Problems in training method for Cataract surgery

Cataract surgery training is different around the world. There are different ways training is provided some of them such as, Human eye, Animal Eye Models or Synthetic Eye Models. Human cadaver eyes are the best way to teach eye surgery, but they are hard to find in eye banks and cost significantly high. Because of this, many trainee programs look for other ways for surgery training. which is cheaper and easier to get. In this situation, the eyes of animals like pigs, goats, sheep, and rabbits can be used for surgical procedures. In many Western countries, pig eyes are easier to get than rabbit

eyes. But pig eyes can't be used in places like Sudan, which is mostly Muslim, because of religious beliefs. In addition to that there is a lot of motivation and hard procedure to study how they can be used in labs for training. Also, To keep the eco system from getting out of balance, many countries don't let people kill animals. On the other hand, synthetic eye model was developed , which have some advantages over biological materials. The problem with the synthetic eye model is that it is expensive to set up.

To learn a practical procedure of cataract surgery, it's important to understand what's going on at each step. To do this, new trainees should first watch and ask the trainer what's going on. When a skilled surgeon works, he or she uses many small tricks or tasks that may not be obvious to someone who hasn't seen it before [4]. Writing down the steps of a procedure in a notebook can be helpful but not always easy to take notes. It might be possible that they miss notes while doing training.The old training method of "see one, do one, teach one" doesn't work when it comes to surgery.Another reason is that changes in the surgical training curriculum have forced residency educators to keep looking for new ways to teach surgical skills. These changes were made because of tight working schedule for resident , worried about patient safety, and limited availability of resources in the operating room[2] [35].

1.1.7 Research Questions

Keeping many problems in cataract surgical training in mind, we created a machine learning-based cataract surgical framework. We developed frameworks for surgical workflow analysis in ophthalmic surgery. The frameworks will accelerate the learning process for students or surgeons, and they do not need to depend on Animal eye Modal or synthetic Eye Modal.The framework will describe surgical workflow from beginning.

The framework have three steps.Using this framework we are going to answer these research questions.

1. First, we need to detect cataracts. If a person has cataracts, what is the level of the cataracts? e.g., normal, cataract, or severe cataract. Do they need surgery?

2. How many phases of surgery are there if patients require surgery? Based on the current predicted surgery phases, how does the model recommend the next sequence of surgery phases? Does surgeon experience helps to improve model?

3. There could be many complications during surgery, but detecting the iris and pupil is difficult because the tools used in surgery make the iris and pupil unstable.

1.2 Objective

This study aims to expedite the cataract surgery training process. We have decided to create a framework for the surgical workflow that will assist trainees in their understanding of the cataract surgical workflow. There are three steps in the framework, and for each stage, we have employed a separate data collection and model to represent the surgical process from cataract diagnosis through surgery.

To accomplish these task , we have used machine learning and deep learning procedure. In this framework,there are three steps, and in each step, different datasets and models have been used. Mainly, we have used thick data analytic using transfer learning to perform each tasks such as severe cataract detection, phase detection, and surgery complexity. In addition to that, Surgeon experience as expert heuristics used to enhance the model accuracy.

1. Step 1: In this step, a classification model was used to detect how bad the cataract was. We used a set of images with more than 30 images that show different levels of cataract. There are four labels on the dataset, such as "Normal," "Cataract," and "Severe Cataract." We used a transfer CNN learning model, which correctly identified normal, cataract. In the stage, trainee can understand when patients need surgery, How bad cataract is?

2. Step 2: This step was completed using the video classification model. We used the Cataract 101 video dataset for this task. The cataract dataset contains over 30 videos of cataract surgery performed by both inexperienced and experienced surgeons. The average video length is 8 to 10 minutes, and some cataract surgery videos are shorter than average because they were completed by highly experienced surgeons who could take less time to complete the surgery. In this video, the surgeon follows all phases of surgery, which is very beneficial for our task. The dataset of surgery performed by a highly experienced surgeon provided a better accuracy rate for the model because the extracted frames of each phase were more accurate. In this stage, trainee will understand the sequence of 10 phases of cataract and what type of tools being used.

3. Step 3: In this step, we have used an object detection model to detect the iris and pupil from cataract images. We have used the ITEC Cat21 dataset for this work. Using this dataset, we solved one complexity of cataract, which is iris pupil detection. The surgeon must understand the iris and pupil reaction, as well as their sizes, during surgery. In this stage,Trainee will focus on complexity during surgery.

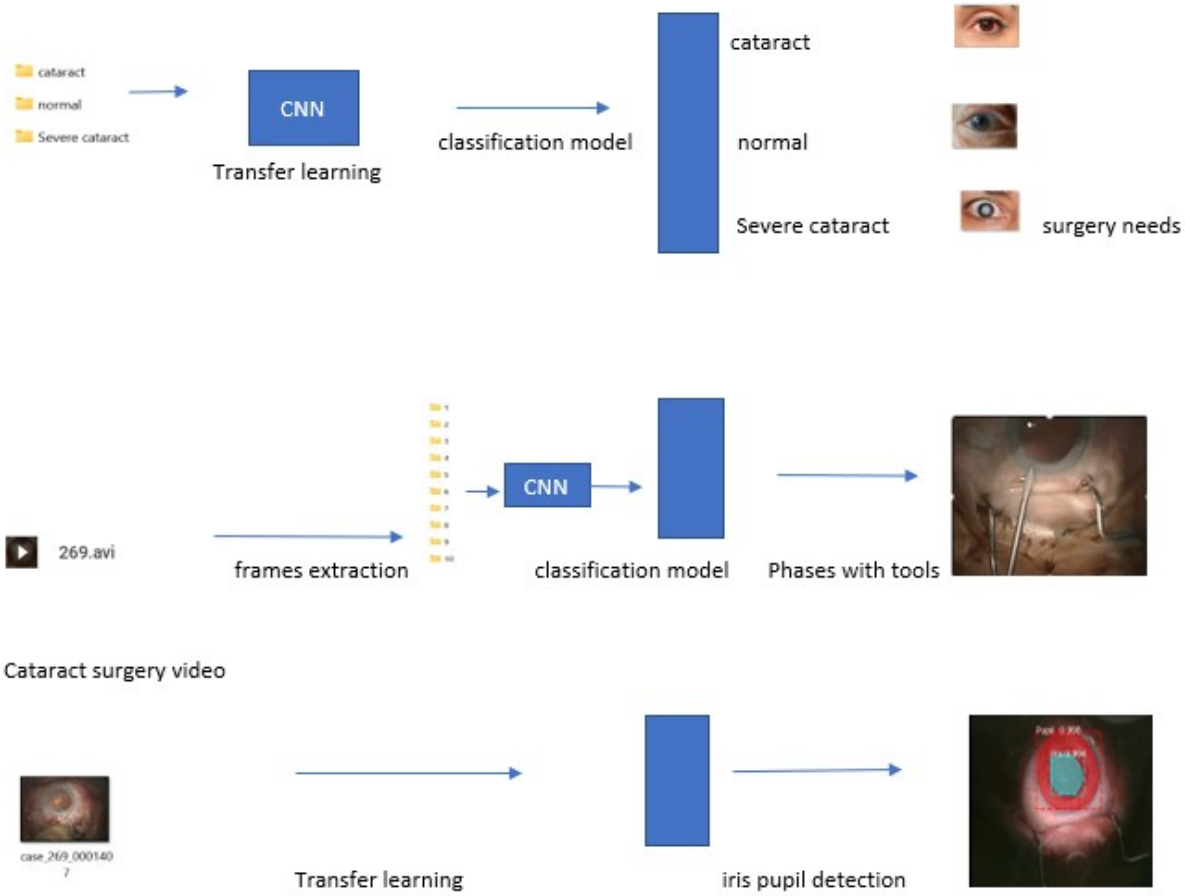


FIGURE 1.6: Cataract Surgery Workflow Framework

.

Chapter 2

Related Research Work

Deep learning techniques are widely used to get meaningful information from image data. For example, pattern or action recognition from video data for autonomous driving, and intelligent monitoring are some of the common uses of video analysis There are different ways for deep learning to learn visual representations from the Data. Convolution Neural Network is being exploited a lot for action recognition from video or image data. In addition to that there are several methods have been developed in the field of CNN for video or images analysis[35].

2.1 Deep Learning application in Surgery

Deep learning models have millions of neurons that can perform like the human brain, which means they can visualize like humans or more accurately than a human brain can do Healthcare systems, for example, can be improved using deep learning techniques, allowing surgeons to provide the best input possible. This AI technique helps operating rooms make surgeries safer and more likely to be successful. Recent improvements in deep learning algorithms for classifying objects and recognizing actions can help medical imaging in a big way. Deep learning has been used successfully in healthcare research to find tumors and cancer, monitor patients remotely, help doctors diagnose patients, and train doctors. [35].

2.1.1 Computer vision applications in Surgery

Computer vision is a field of research that covers ideas related to how computers extract data from digital photos and videos. In comparison to the human eye, it may help us see more clearly and with a wider field of vision and also help in examine objects that are challenging for the human eye to see. Medical image processing is being accelerated while also boosting accuracy thanks to computer vision-based solutions.There are several uses for computer vision in the medical field already. It has an effect on medical specialties including dermatology, oncology, radiology, and physiology.Many different types of diseases can be detected from medical dataset such as Tumor Detection, Cancer Detection, Health monitoring.The computer vision has revolutionized the medical industry and is used in several medical research projects [6].

2.2 Overview On ML/Deep Learning Research for Analyzing Cataract Videos:

This chapter provides an overview of the existing literature view on cataract surgeries Analysis using deep learning and Big data. Most of research has been done using deep learning, Convolution Neural Network and computer vision algorithms. To produce meaningful results, deep learning requires a significant amount of data. However, thick data was crucial in avoiding massive data. By combining expert heuristics with little amounts of data, we may create a model that will accomplish our goal..

2.3 Multi-label Classification of Surgical Tools with Convolutional Neural Networks

In this research work, authors looks at different ways to find surgical tools in videos using convolutional neural networks. According to Authors ,past few

years, CNNs have made great improvement in solving vision-based problems, making them a good selection for this work. Authors used ResNet model for their research works because ResNet is one of the best models in solving number of vision problems. Authors used CNNs which classify each frames. Using a 50-layer deep residual network, Authors developed a model to identify various surgical instruments in videos of cataract surgery.They have used ResNet50 model architecture .It contains 48 CNN layers with one global pool layer and max pool [35]. They made ResNet fined tuned model for their work by randomly freezing the first few layers which is smartest way to decrease memory necessity and saved lots of time by training whole network. Other than Authors also looked for max pooling and global average pooling by pooling out image features by using fixed weights and First random number.

Similar to this, to identify surgical instruments, feature vectors derived from the feature extractor are then passed to either the average pooling or max pooling layer. ResNet is trained using every sixth picture frame from each operation video.ResNet is trained using every sixth picture frame video, which mitigate both the number of redundant frames in the data and the model's training time. They also use data augmentation technique on images taken from video. The authors came to the conclusion that fine-tuned ResNet performed better than ResNet50 when employed as a fixed feature extractor

2.4 Code-free machine learning model for cataract surgery phases detection

Initially, the rise of surgical video analysis began with the identification of surgical tools and phases. Automated tool and phase identification in surgical images or videos can significantly improve the process of surgical workflow analysis . For evaluating surgical workflow, researchers investigated several feature extraction methods and machine learning models. Some of these earlier efforts were focused mainly with tool identification and phases detection in cataracts surgery.Tool identification is a critical step in acquiring information about tool trajectories and tools use in a surgical video. Phases detection using tools are done with previous works. The surgical tool study and phase recognition are then used to teach new trainers or surgeon . I am going to discuss some research works which has already been done in cataract surgeries using deep learning and CNN methodology.

This researchers [21] focused at how well automated machine learning (Auto ML) did at sorting surgical videos into phases of cataract surgery. Two ophthalmology residents who didn't know how to code made a deep learning model in Google Cloud Auto ML Video Classification to classify the 10 different phases of cataract surgery. For model training, validation, and testing, we used two public datasets with a total of 122 surgeries. Ten surgeries that came from another dataset were checked from the outside. The Auto ML model did a great job of differentiating, even doing better than custom deep learning models made by experts.Precision area curve and recall was 0.855.

At the 0.5 confidence threshold cut-off, the overall performance metrics were: sensitivity (81.0%), recall (77.1%), accuracy (96.0%), and F1 score (0.79). During the different phases of surgery, the per-segment metrics ranged from 66.7 to 100% for precision, 46.2 to 100% for recall, and 94.1 to 100% for specificity. The phases that were most accurately predicted were hydro dissection and phacoemulsification, which were right 100% of the time and 92.31% of the time, respectively. Throughout external validation, the precision was 54.2% (0.00–90.0%), the average recall was 61.1% (0.00–100%), and the average specificity was 96.2% (91.0–99.0%). In the end, a code-free Auto ML model can classify the phases of cataract surgery from videos as well as or better than models made by experts.

The algorithm was trained and tested with videos of cataract surgery. Figure 2.1 shows phases of the surgery.The model was trained and tested with videos from two datasets, Cataract 101 and 21. It is also mentioned from researchers prediction that Viscous agent injection (Phase 2) was the one that occurred the most, since surgery can be done more than once. All of the

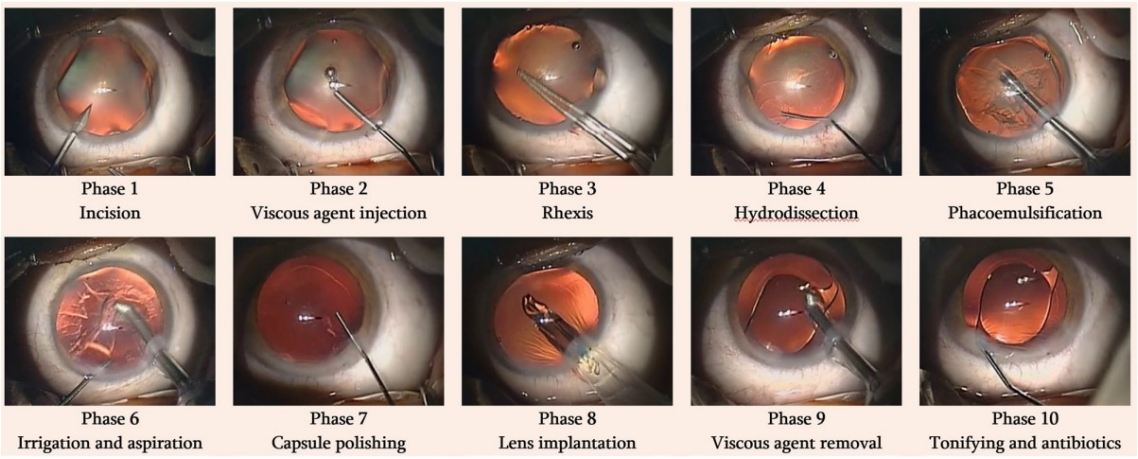


FIGURE 2.1: Ten phases of cataract surgeries from the 101 dataset [21]

videos were encoded as MP4 files with a frame rate of 25 frames per second and a resolution of 720 x 540 pixels. The Cat-21 dataset had 21 videos and the Cat-101 dataset had 101 videos, for a total of 122 videos. The videos ran for a total of 16 hours, 26 minutes, and 11 seconds. The datasets have comma-separated value (CSV) files that show when each phase starts and ends.

The algorithm was tested on a different dataset outside of the model to make sure it worked properly. Authors used 10 videos from the CATARACTS dataset, which can be found at https://ieee-dataport.org/open-access/cataracts and is open to the public. The original labels looked at tool identification, so these had to be re-labeled and re-marked. Since no capsule polishing was done in those videos, authors could only look at the first 9 phases. To make it look like the other videos, the size was changed to 720 540 pixels.

Authors developed performance metrics of Automated ML which used in the AI community. These are precision (positive predictive value, or PPV),

recall (sensitivity), and the area under the precision/recall curve (AUPRC). AUPRC(The area under the precision-recall curve) is a metric which has ranges from 0.5 to 1.0.Higher values describes better discriminate performance13. Model also generated F1 score, which represents harmonic mean and recall.

Overall, the AutoML model did a great job, with an AUPRC of 0.855. The confidence level was 0.5, the F1 score was 0.79, the recall was 77.1% ,the precision was 81.0%,and the confidence threshold was 0.5. For all phases, the average calculated accuracy was 96.0

Figure 2.2 shows whole model performance of each phases. From table it can be seen that Precision varied from 66.7% for capsule polishing to 100% for both lens implantation and hydrodissection. The range for specificity was between 94.1 and 100% whereas the range for recall was between 46.2% and 100%. The rhexis covers lowest sensitivity and F1 score which is only 0.63 and F-1 score hydrodissection phase was 0.89 which highest for any phase.

	Number	TP	FP	TN	FN	AUPRC	PPV	Sensitivity	Specificity	F1 score
Overall	144	NR	NR	NR	NR	0.855	81.0%	77.1%	98.0%	0.79
Incision	12	10	3	129	2	NR	76.9%	83.3%	97.7%	0.80
Viscous agent injection	26	21	7	111	5	NR	75.0%	80.8%	94.1%	0.79
Rhexis	13	6	0	131	7	NR	100.0%	46.2%	100.0%	0.63
Hydrodissection	12	12	3	129	0	NR	80.0%	100.0%	97.7%	0.89
Phacoemulsification	13	12	3	128	1	NR	80.0%	92.3%	97.7%	0.86
Irrigation & aspiration	16	11	3	125	5	NR	78.6%	68.8%	97.7%	0.73
Capsule polishing	14	10	5	125	4	NR	66.7%	71.4%	96.2%	0.68
Lens implantation	12	9	0	132	3	NR	100.0%	75.0%	100.0%	0.86
Viscous agent removal	13	9	1	130	4	NR	90.0%	69.2%	99.2%	0.78
Tonifying & antibiotics	13	11	0	131	2	NR	91.7%	84.6%	100.0%	0.88

FIGURE 2.2: Performance table for Auto ML [21]

2.5 Surgical phases extraction using Inception V3 in Real Time

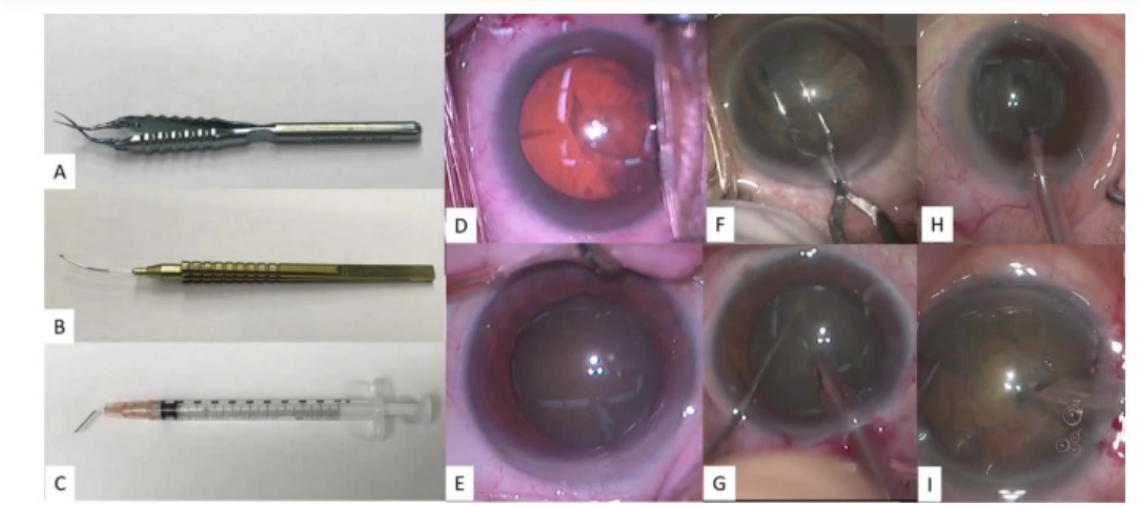
The current work [20] used AI to do a real-time automated analysis of continuous curvilinear capsulorrhexis (CCC), nuclear extraction, and three additional cataract surgical stages. 303 cataract surgery cases from Tsukazaki Hospital's clinical database were utilized. Each frame was taken from surgical recordings at 299 168 and 1 FPS.The collected images were labeled accurately depending on the start and finish timings of each surgical step recorded by an ophthalmologist.InceptionV3 was used to identify each image's surgical phase. The images were then analyzed in chronological order using a Inceptionv3 to get the moving average of five consecutive outputs. Surgical phase was the highest-output class. For each surgical step, the start time was when it was initially recognized, and the end time was when it was last identified. The performance was assessed by calculating the mean absolute error between the start and finish times of each significant phase recorded

by the ophthalmologist and the start and end times provided by the model,. Correct cataract surgery phase categorization was 90.7% for CCC, 94.5% for nuclear extraction, and 97.9% for other phases, with a mean of 96.5%. The ophthalmologist's start and finish timings differed from the neural network model's. CCC's start and finish timings were 3.34 and 4.43 seconds, whereas nuclear extraction's were 7.21 and 6.04 seconds, for a mean of 5.25 seconds. This study's neural network model classified the surgical phase using just the final 5 seconds of video.

2.6 Dataset used in this work

In this study, video recordings of cataract surgery at Saneikai Tsukazaki Hospital which is a social medical care corporation, were used to recognise surgical phases. The videos had a resolution of 1920 X 1080, a frame rate of 30 FPS, a mean duration of about 534 seconds, and a standard deviation (SD) of about 237 seconds. The average (SD) lengths of each phase were: about 42 (44) seconds for CCC, 133 (85) seconds for nuclear extractions, and 359 (163) for other phases, Out of the 303 surgical videos, 245 were used for training, 10 were used for verification, and 48 were used as tests.

There were 303 surgical videos with 17 different surgeons. For CCC, three types of forceps were used, and for nuclear extraction, four types of surgical techniques were used. For both CCC and nuclear extraction, two types of lighting were used. [20].



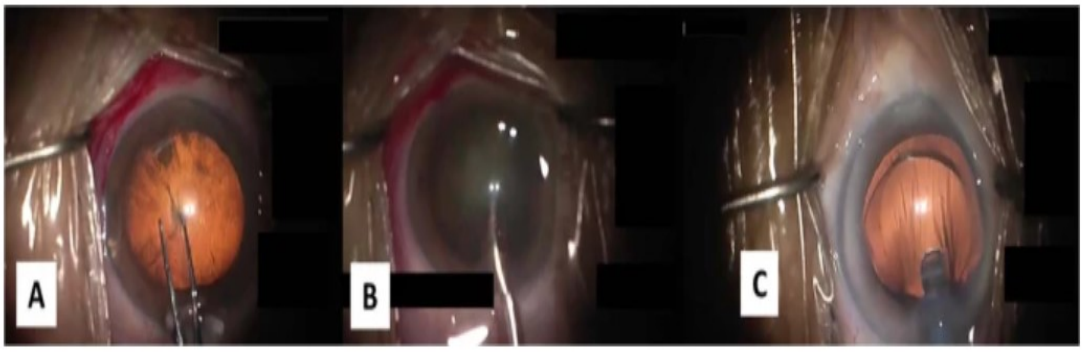
Examples of surgical instruments, lighting methods, and nuclear extraction techniques. (**A**) Inamura forceps, (**B**) Ikeda forceps, (**C**) cystotome, **(D**) retro illumination, (**E**) direct illumination, **(F**) phaco-prechopper method, (**G**) phaco-chopper method, (**H**) divide and conquer method, (**I**) central-divide method.

FIGURE 2.3: Surgical instruments, lighting methods, and nuclear extraction techniques [20]

In addition to that Videos were scaled down to 256 168 at 1 FPS so that surgical phase recognition could be done on each image. In the end, 161,140 images were taken from 303 videos, for a total of 161,140. The surgical phases were correctly called CCC, nuclear extraction, and others. The labels were based on the start and end times of each phase of surgery, which was recorded by an ophthalmologist.Figure 2.4 shows how many image datasets were collected for each phase of surgery, and Figure 2.5 shows three phases of a real surgery.

Recognition class	Training data (images)	Validation data (images)	Test data (images)	Total (images)
CCC	10719	211	1725	12655
Nuclear extraction	33020	976	5995	39991
Others	90023	2376	16095	108494
Total	133762	3563	23815	161140

FIGURE 2.4: Dataset breakdown of cataract surgery phases [20]



Sample images of three surgical phases. (**A**) CCC (Inamura forceps, retro illumination method), (**B**) nuclear extraction, (**C**) others (intraocular lens insertion).

FIGURE 2.5: images of three surgical phases from the dataset which was used in this work [20]

2.7 Methodology used in this work

Authors used the Inception V3 model17, which is a convolution neural network model, used to recognise the 3 phases of surgery. By replacing n X n convolution with 1 X n convolution and n X 1 convolution, the Inception module reduced the good amount of computation and also gradient elimination. The input was a 299X168X3 color image, and the number of neurons in

the output layer was 3, which is the number of surgical phases to recognize. The model was trained by starting with trained values for each parameter from the ILSVRC 2012 data[18].To standardize pixel values between 0 and 1, images were preprocessed, and the preprocessing methods were done randomly to avoid over learning. In training The batch size was used 32, the loss function used was Multi-class log loss, the the maximum number of epochs was 300 and optimization function was Momentum SGD with learning rate 0.0001, momentum was 0. This model's classification results were as follows: 94.5% for nuclear extraction,90.7% for CCC,a mean response rate of 96.5% and 97.9% for others.The recognition error rates were as follows:nuclear extraction was misidentified as "others"5.5% of the time,CCC was misidentified as "others" 9.3% of the time,and "others" were misidentified as CCC 0.9% of the time and nuclear extraction 1.2% of the time. The rate that the model could not identify the distinction between CCC and nuclear extraction was less than 0.01%.

2.8 Surgery Tools identification using Yolcat cnn

In this works author [42] took 1156 frames from the 9 main steps of 268 cataract surgery videos and marked 8 different surgical tools, the pupil border, and the limbus.Researchers pretrained YOLACT which is a real-time object detection and segmentation model, on the CaDIS dataset. This is public dataset for semantic segmentation of videos of cataract surgery. After that pretrained model was fined tuned on dataset which going to use in this work .Researchers trained a model to find the accuracy of surgical instruments and tool tips in real time. Using videos of cataract surgery, the model could be used to provide an automated feedback system for assessing surgical rate performance.Object detection was evaluated by the average precision score (AP), which was calculated by averaging the precision of the bounding boxes along the precision-recall curve. Segmentation was assessed by intersection-over-union (IoU), which was calculated by taking the intersection of the predicted mask and the true mask over their union. The position of the tool tip was assumed by finding the edge point of the predicted mask that was closest to the center of the screen. The center of the pupil was estimated by fitting an ellipse to the edges of the pupil mask and finding where the center of the ellipse was. Across various object classes, the average AP and IoU were 0.78 and 0.82, respectively. Blade, weck sponge, and phaco tools had the best segmentation performance, while the performance of the needle of instruments was the poorest Fig 2.6. The predicted phaco tip locations' average error from the ground truth positions was 6.13 pixels. In Figure 2.7 , examples are shown. The typical sensitivities and precisions were 81% and 100%, respectively, when true positives were defined as predictions within 10 pixels of the actual location.

class name	average precision	intersection-over-union
limbus	1.00	0.82
pupil	0.98	0.90
blade	0.80	0.92
forceps	0.64	0.68
second instrument	0.48	0.70
weck sponge	0.66	0.92
needle or cannula	0.65	0.67
phaco	0.92	0.92
irrigation/aspiration	0.90	0.81
lens injector	0.79	0.88

FIGURE 2.6: segmentation masks' intersection-over-union and per-class average bounding box accuracy on the test set [42]

2.9 Laparoscopic Video Recognition Tasks using deep learning model EndoNet

Authors [37]proposed EndoNet to perform various tasks on the pictures of laparoscopic ivdeo from the m2cai2016 dataset, which is created by fine-tuning with the AlexNet network and utilizing weights from the ImageNet dataset [29]. EndoNet was developed to simultaneously identify phases and find tools in the images. This was accomplished by having the third fully connected layer output with scores probability for each of seven nodes, each of which relates to the likelihood that an image includes a certain tool. To provide features for phase recognition, the output from the first two fully connected layers is concatenated with another dense layer called fc phase. A Support Vector Machine (SVM) is used to generate a probability score for the observed phase using feature vectors learnt from the previous dense layer. The probability score was passed to 2-level hierarchical HMM from SVM.The confidence score derived by SVM is used by the HHMM as the input observations. The forward algorithm of the HHMM is used to make the final forecast for the phase. By experimenting with various designs based on the AlexNet basic network configuration, they have expanded their work utilizing the Cholec80 dataset.EndoNet is a model created by fine-tuning the AlexNet network and utilizing weights from the ImageNet dataset by researchers to execute various tasks on the images taken from laparoscopic video from the m2cai2016 dataset. EndoNet was developed to simultaneously identify phases and find tools in the images of laparoscopic video. EndoNet was developed to simultaneously identify phases and find tools in the

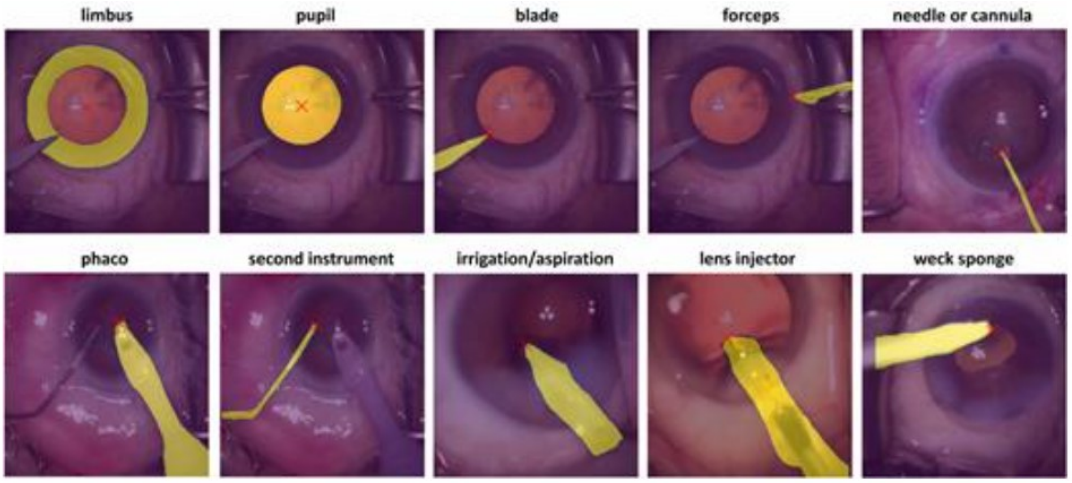


FIGURE 2.7: Images shows predicted tip positions for cataract surgical tools,segmentation masks, and anatomical landmarks [42]

photos. This was accomplished by having the third fully connected layer provide probability scores for each of seven nodes, where each node denotes the likelihood that an image includes a certain tool. To provide features for phase recognition, the output from the first two fully connected layers is concatenated with the output from a third dense layer called fc phase. A Support Vector Machine (SVM) is used to generate a probability score for the observed phase using the feature vectors learnt from the last dense layer. A 2-level hierarchical HMM is again given the SVM probability score. The confidence score derived by SVM is used by the HHMM as the input observations. The forward algorithm of the HHMM is used to make the final forecast for the phase. By experimenting with various typologies based on the AlexNet basic network configuration, they have expanded their work using the Cholec80 dataset [35].

2.10 Surgical Tool Detection Using Attention-Guided Convolutional Neural Network

In this research work Authors[33] Used Faster RCNN to built a modulated anchoring network for laparoscopic video surgical equipment detection. The three components of a modulated anchoring network are the modulated feature module, the anchor box location prediction, and its shape prediction. The purpose of the anchor position branch is to produce a probability map, and the map indicates the potential location of any item or instrument's center. The anchor shape prediction branch attempts to determine the shape of the instrument at the position of the discovered anchor box. To further compare the features with the shape information of accessible tools in the operation, a modified feature module combines the shape information of the

instrument or item provided by the shape prediction branch into a feature map. A relation module that tries to calculate the relative relationship of various instruments in every situation of the videos is incorporated in the network. Authors used ResNet-101 to combined with a Feature Pyramid Network is used to create the fundamental feature detection network[14] [15]. Each surgical instrument in a video frame receives bonding box labeling from the modulated anchoring network. By creating heat maps for each surgical instrument, the authors investigated the movement of the devices that had been observed further. By analyzing tool use patterns using heat maps and the chronology of how long each instrument is used throughout a video, it is possible to assess the operational efficiency in surgical videos. The author demonstrated that Faster RCNN-based networks outperformed other current methods in terms of tool identification accuracy.

2.11 Convolutional neural with multi-image fusion for surgical tool identification during cataract surgery

Another study recognizing tools from video surgery,Authors[1] using a multi-image fusion technique to enhance the ability to recognize tools in cataract videos, which serves as the first stage in completing surgical workflow analysis. In this work, rather of using a single video frame, each video is represented by a series of related frames. Authors used convolutional neural network (CNN) which contains few CNN layer. A pooling layer comes before each convolutional layer. The dropout layer is once again placed after the last two fully connected layers, which are designed to forecast tool presence. A group of 16 consecutive video frames are sent into the CNN. The optical flow between each successive image in a series is calculated, and the results of this calculation are processed by the first few layers of the CNN. The activation map for one image is fused to the activation maps of the subsequent successive images in order to combine information from one image to the next in a sequence. The activation map of the last image in the series is then fused once again with the activation map that resulted from the previous two images. The CNN is first trained on each individual video frame before being retrained on image sequences to further examine how the model's performance varies for various input data. The authors demonstrated that using an image sequence is superior to the traditional method of having a CNN analyze each individual video frame for the purpose of enhancing the tool detection performance of the CNN.

2.12 Thick Data Analytic for Small Training Samples Using Siamese Neural Network and Image Augmentation

Although deep learning and machine learning have offered solutions to a range of complicated problems, they need to be trained with a huge quantity of labeled data in order to execute with high accuracy. Collecting large amounts of data is sometimes not practical in many applications, such as in healthcare, large data is not available for train model.These problems could be solved by thick data procedure.It solves this problem by adding more qualitative steps, like using experts' heuristics to annotate and add to the training data.

In this research work,Authors[11] looked into how the heuristics of a human radiologist can be used to find COVID-19 in a few cases of CT-Scan images by using clusters of image annotation and augmentation techniques. To find new COVID-19, a unique structure Siamese network is used to rank the commonality between new COVID-19 CT Scan images and images that a radiologist has already identified as COVID. Using a similarity ratio, the Siamese network takes features from the augmented images and compares them to the new CT-Scan image to figure out if the new image is COVID-19 positive or not. Author showed in result that proposed model, which uses augmentation heuristics and is trained on a small dataset, does better than the advanced models, which are trained on dataset with many samples. The Author investigated some important questions, like why we need CT scans to diagnose COVID-19, what "Thick Data" is, and how "image augmentation" and "Siamese Neural Network" help them learn from small samples.

In this area of research, Authors[11] used thick data heuristics into the CT image dataset using Siamese neural network to reduce trainee time and provide better result rather than using traditional deep learning model. When trying to classify complicated data sets and images like CT scans, there are a number of problems that can come up. Author mentioned that Traditional networks take a long time in image classification, and they also need a lot of data to be able to make accurate conclusions. Even when given a lot of data, popular neural networks like VGG and ResNet can't classify images accurately and consistently for sensitive tasks like finding COVID-19 in a CT lung scan. To solve this problems they have used Siamese neural network architecture on a sequential network, which has been said to reduce training times and the amount of training data needed. To make this network even stronger, they added thick data heuristics into the CT image dataset.Authors mark important parts of the images that a radiologist will indeed look for to diagnose, such as ground glass opacity.Their proposed network is about 3% better than the top five neural networks for image classification. By using thick data heuristics, researchers have shown that accuracy is improved, and also they think that accuracy will continue to rise. Basically , they annotate

2.13 The promise of big data technologies and challenges for image and video analytic in healthcare

In this research work, Authors described different types of advantages and challenges of big data. Data is produced from sources including next-generation sequencing (NGS) technologies, imaging ,continuous bio metric data monitoring with the help of embedded tiny sensors, lab data and clinical data from electronics health records. Authors addressed that Heterogeneity-related problems must be addressed in data processing. A system of systems may be used to handle heterogeneity by separating and analyzing different big data analysis systems for each data domain. The strategy is to use metadata to leverage the newly discovered structures in a semantically linked way for collaborative study and the development of new knowledge.

Big healthcare data research is anticipated to have a significant impact on the delivery of accurate and timely healthcare services, which will become more preventative and individualized. Predisposition to illness, early and more precise diagnosis, post-hospitalization tailored rehabilitation, and healthcare interventions are just a few of the many possible paradigms that might potentially revolutionize today's clinical care in the future. The quality of life of both citizens/patients and their family will be considerably improved by improving the quality of care, and this will provide the groundwork for cutting healthcare costs and opening up new market opportunities.

Authors describes different types of challenges associated with medical video analytic, issue with storage, explains the processing paradigms for large-scale video data sets. Medical video as a Big data is not easy to analyze so easily. The authors described challenges through Volume, velocity, variety, veracity ,visualization and value.

Volume:- Digital video clearly meets the volume needs of big data systems more than anything else. To see this, consider that a standard video with a resolution of 720X576 (PAL) that needs to be sent at 30 frames per second using ITU-R BT 601 requires about 166 mbps. This is the same as 74.7 Gigabytes an hour. Because of this, both storing and sending video requires compression. In terms of volume we can see that a single medical video can produce data sizes that are much bigger compared to any of the modern clinical imaging methods. In fact, medical video data is being measured in Terabytes or even Petabytes. Such volume of data is very expensive to process and store.

- Velocity:- Using different kinds of medical video modalities adds variety such as video of ultrasound, and endoscopy, and emergency videos. Within each modality, there are also important subcategories and each subcategories all produce ultrasound videos. Trauma videos and videos of the scene of an emergency are also examples of emergency videos. Clinical video acquisition devices are also very different. The employment of various resolutions, frame rates, and video

compression methods adds variety. Then, there are even more differences when different teams use different clinical protocol which produce more variety.

- Veracity:- In health care systems, clinical diagnostic quality can be linked to veracity. Video compression can make it much harder to tell if a video is real or not because it can introduce large artifacts in diagnostically important areas. The use of various clinical procedures that have a significant impact on the length of the videos and how organs are imaged makes maintaining veracity even more challenging.

- Variability:- Variation is also a problem. As an example, a lot of the available video data comes from follow-up video examinations such as mass screening of a population that results in an only one video capture, endoscopy clips of a single operation and exam, trauma video files of an emergency video is spread out all across hospitals and healthcare data storage.

- Value:- Supporting personalized interventions and making better decisions in emergencies and new situations are two examples of value. Ultimately, medical video analysis can be very valuable because it can improve the quality of care while lowering costs. This can be done by redesigning clinical care paths to be more effective based on what we've learned from using medical video analysis methods. For such a complicated job to be done well, a number of problems must be solved, like data privacy and ownership, legal, data storage and data transfer, normalization and data processing and so on.

- Visualization:- Visualization should make it easier to understand for both patients and the public, as well as for healthcare professionals. The challenge is to give each stakeholder information that is specific to them and has meaning. Also, the generated metadata needs to be linked to other healthcare data in a way that makes sense, so that it can be used for further display and analysis.

  The author gives an overview of the most important problems with big data in medical video analytic. The main issue with big data is its size, which can reach petabytes or terabytes in the future. The process of handling such large amounts of data is very expensive. Big data technologies, on the other hand, can be used to make good use of a lot of computing power and extract useful clinical information from unstructured video footage, but compressing video without sacrificing diagnostic quality and processing large amounts of unstructured video data is a difficult task.

2.14 Deep Bayesian networks with LSTM for surgical workflow analysis

In this work Authors [5] examined the performance of an active deep learning system based on Deep Bayesian Network (DBN) for annotating surgical videos and images captured in real-time surgery. The suggested DBN for active learning is designed using the standard CNN AlexNet.DBN had already been trained on ImageNet before modified by author. AlexNet is made up of three fully connected layers and four 2-dimensional convolutional blocks, each of which is followed by a pooling layer. The base AlexNet is extended to a DBN for instrument identification in laparoscopic images by adding a dropout layer after each convolution layer. Each dropout layer of the DBN is followed by another dropout layer. To recognize surgical tools using images, a basic AlexNet network with simply a dropout layer is used.A complicated kind of recurrent neural network, such as Long Short Term Memory, is added at the end of the first completely connected layer in the basic DBN [5] to perform video-based phase recognition. With the help of the new Recurrent DBN, the system is now able to view any surgical video from the viewpoint of successive frames that show the connections between each frame throughout the whole clip. Video-based surgical phase segmentation is enabled because to the LSTM layer's ability to aggregate previously learnt information from one frame to the next successive frame. Processing a full video might take some time since surgical videos are sometimes lengthy. In order to solve this problem, each video is divided into shorter video snippets with a short length . The Cholec80 dataset's video frames were processed at a resolution of 384 x 216 pixels. Adam [10] was used as an optimizer during DBN training. On the Cholec80 dataset from the m2cai2016 competition, they tested their system and demonstrated that the DBN may be effective for image frame or video frame annotation.

2.15 CNN methodology for Automatic Pupil and Iris Detection

Techniques for monitoring the eyes that are underpinned by deep learning are quickly gaining traction in a diverse array of application domains.In this research Authors[3] investigated the feasibility of using eye tracking methods in ocular proton therapy applications. They implemented a totally automated method that is based on two-stage convolutional neural networks (CNNs): the first stage conducts a rough identification of the eye location, and the second stage performs a fine detection of the iris and pupil. They selected 707 images from the video that were recorded during clinical procedures and OPT treatments that were carried out at institution. 650 images were utilized for training, and 57 were used for testing in a blind condition. Estimates of the iris and pupil were compared to manually labeled contours drawn by a clinical operator. These comparisons were made using the manual data. Quantification was done on different parameters such as: on the

Szymkiewicz–Simpson coefficient the median was 0.97 and 0.98, on the Dice coefficient the median was 0.94 and 0.97, on the Hausdorff distance the median = 11.6 and 5.0 (pixels) and last on the Intersection over Union coefficient the median was 0.88 and 0.94. This was done in order to make predictions about the iris and the pupil. It was found that the iris and pupillary areas were equivalent to the manually labelled ground truths. Their system may give an automated method for quantitatively analyzing pupil and iris misalignment, and it could be utilized as an extra support tool for clinical activity. The suggested technique used two cascaded U-Nets from coarse (ROI U-Net) to find the pupil and iris using a dual stage convolutional neural network pipeline (Pupil U-Net and Iris U-Net). The pipeline's initial step was preliminary eye localization inside the video frame utilizing the U-Net architecture that was made available for ROI detection (ROI U-Net). The extracted ROI represented the pupil and iris, and it was passed into the U-Nets in the second pipeline step (the pupil U-Net and the iris U-Net), which were developed to provide accurate pupil and iris identification.The objective of the ROI identification step was to isolate a region of the picture where the important ocular characteristics—the pupil and iris—should most likely be present while excluding potential background components like the eyelid, eyelashes, and retractors. Additionally, by cropping the image, it was possible to increase pupil and iris resolution while minimizing image noise and future illuminating inhomogeneities. In fact, the final two U-Nets (iris and pupil) worked on a higher iris and pupil resolution, resulting in a higher degree of feature extraction accuracy. Between the first and second stages of the suggested technique, the detection of the ocular center was enhanced by 7%. Several CNN approaches to classifying objects in an image worked better when the recognition problem was limited to a certain part of the images. The ROI extraction in Author 's suggested technique,automatically generates the input picture for the next U-Nets. Due to two primary factors, researchers decided against building a multi class U-Net and instead chose single class networks: first one was that they though that this dataset was sufficient for training single class networks but not multi class networks and The Heavier and deeper networks are needed for feature recognition and discrimination. They decided to use two quicker and lighter structures rather than one heavier one because they needed a framework that could be able to do real-time analysis. Additionally, the suggested method's architecture is adaptable enough to allow for the prospective incorporation of other U-Net cascades intended for various supplemental ocular feature extractions.

Chapter 3

Theoretical Background for Cataract Video Analytic

We have used different types of datasets in every step of the framework. We used an image dataset in the first step for cataract detection, a cataract video dataset in the second step for phase detection, and a coco like image dataset in the final step to detect the iris and pupil. There are different types of deep learning tools we have used to implement our frameworks, e.g., deep learning, convolutions neural networks, and thick data analytic. We are not using big data instead we taken help from thick data analytic.

3.1 Thick Data Overview

Thick Data is information that has been obtained using anthropological approaches and reveals the real feelings that consumers have for the products. In other words, thick data refers to the qualitative information discovered from examining global mindsets and human experiences.Thick Data uses human learning to reveal the social context concealed in the data, while Big Data relies on machine learning to deploy, process, and analyze the provided collection of data. Thick Data just needs a little quantity of quantitative information to depict human patterns, but Big Data needs a big amount of data to understand information from the information's hidden patterns [41].

3.2 Convolution Neural Network

Before jumping into CNN, lets discuss about ANN. Artificial Neural Networks (ANNs) are computer processing systems that are based on how biological nervous systems, like the brain, work. ANNs are mostly made up of a large number of connected computational nodes called neurons. These neurons work together in a distributed way to learn from the input and improve the output as a whole. Figure 3.1 shows how to model the basic structure of an ANN. We'd send the multidimensional vector as input, to the input layer and then send it to the hidden layers. The hidden layers is responsible for making decisions based on the previous layers and think about movement

change of stochastic inside itself improves the ultimate output, which is referred to as the learning process of model. Deep learning is a term used to describe systems with many hidden layers built on top of one another[24].

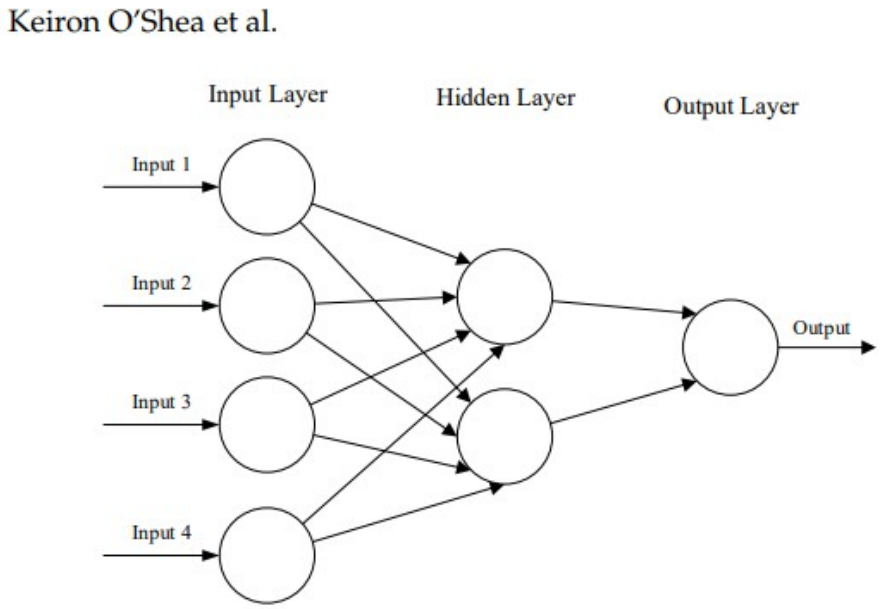


FIGURE 3.1: A simple feed forward neural network diagram [24]

The two primary forms of learning employed in image processing takes are supervised learning and unsupervised learning. Supervised learning involves learning using labeled inputs that are utilized as objectives. Each training example will include a collection of input values (vectors) and one or more predefined output values. On the other hand, the purpose of this method of training is to lower the overall classification error of the model by ensuring that the output value of each training sample is accurately computed. Unsupervised learning differs from supervised learning in that there are no labels in the training set. Typically, the effectiveness of a network is determined by its ability to minimize or maximize an associated cost function. It is crucial to remember, however, that the majority of image-based pattern-recognition tasks depend on categorization through supervised learning [24].

(CNNs) are similar to conventional ANNs in that they are composed of neurons that may develop. As is the foundation of all ANNs, each neuron will continue to receive input and perform an operation (such as a scalar product followed by a non-linear function). The whole network will continue to express a single perceptual scoring function from the raw image vectors input to the final output of the class score (the weight). All the techniques used for conventional ANNs still apply to the final layer, which has loss functions for each class.

3.3 Convolution Neural Network Architecture

As was previously said, CNNs are mostly based on the idea that the input values will be images (in the form of vector). This makes it clear that the architecture needs to be set up in a way that works best for a certain type of data.The layers of the CNN are made up of neurons that are arranged in three dimensions, including the height and breadth of the input as well as the depth. For example input will have dimensions of 8 X 8 X 3 (height, width, and depth), and The final output layer will be 2 X 2 X n in size (n is the number of potential classes), since we need to condense the whole input dimension into a smaller amount of class scores recorded along the depth dimension.There are three basic types of layers in CNNs. These layers are fully connected, convolutional, and pooling. A CNN architecture is created by stacking these layers.[24].

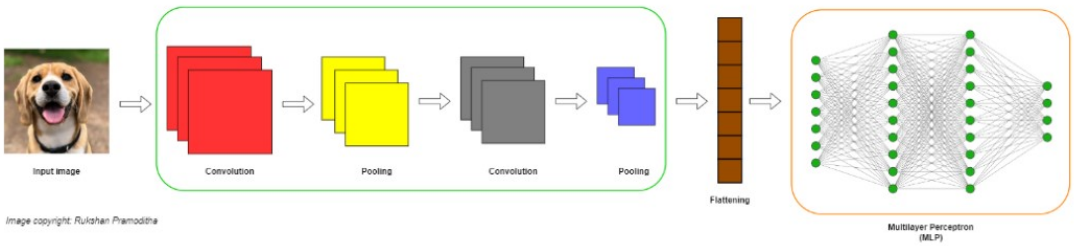


FIGURE 3.2: Overall architecture of CNN [27]

Convolutional layers and its operations.A CNN's first layer is a "convolutional" layer. A CNN can have more than one convolutional layer. The first layer of CNN takes the images as input and begins to process them. This layer is responsible for extracting features from images while keeping the connections between the pixels. It has three elements input image, Filters and feature map.on the other hand, CNN operation is nothing but element-by-element multiply-sum operation.

Filter: It is one of the most important concept here. Filter is also known as a Kernel Detector or a Feature Detector. This matrix is small. In a single convolutional layer, there can be more than one filter. In a convolutional layer, filters of the same size are used. Each filter does certain function. Different features of the image are identified with the help of different filters. We can decide number of filters and sizes as hyper- parameters.The size should be less than the size of the input image. The filter's configuration is set by the elements inside it. and these elements are type of CNN parameter that is learned during training [27].

Image section: The size of the image section should match the size of the filter(s) we choose. We can move the filter(s) up and down or side to side on the input image to make different parts of the image. Depending on the Stride, the number of image sections varies .

Feature map:The feature map stores the results of different convolution operations between different sections of the image and the filter (s) and this

will go into the next layer of pooling. The number of "elements" in the feature map is the same as the number of different sections of the image we received by moving the filter(s) around [27].

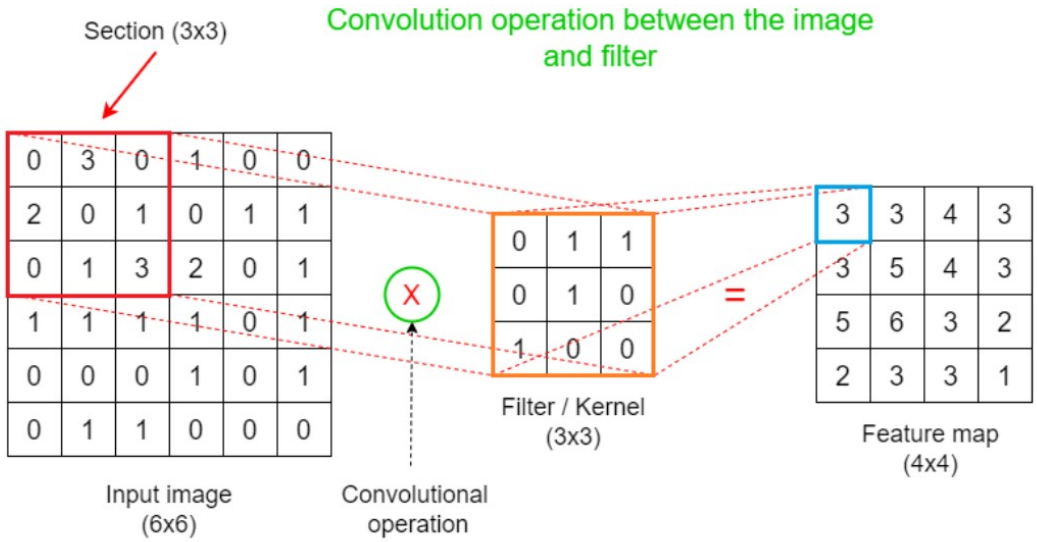


FIGURE 3.3: Convolution Operation [27]

3.4 Pooling layers

The second type of layer used in a CNN is called a "pooling layer." A CNN can have more than one pooling layers. After each layer of convolutions is a layer of pooling. So, pooling layers and convolution are used in pairs. The main objective of pooling layer is to find most important relevant features by getting the highest number or averaging the numbers. It helps to minimize the dimensionality of previous layers'output values. Other than that it also minimize any kind of noise exist in the features and also help in boosting the accuracy. Max pooling and average pooling are two types of pooling operations. There are two distinct kinds of pooling operations: max and average pooling. While average pooling identifies the average of values in the filter's application region, max pooling assists in locating the largest number in that area.[27].

In the pooling layer, there are three parts: the Feature map, the Filter, and the Pooled feature map. In each pooling layer, the pooling operation takes place. The

Filter: Filter is just a window and does not have elements inside it. Therefore, there are no parameters in the pooling layer to learn. The filter is only used to specify a certain area of the feature map.The size need to be less than that of the feature map. A filter with the same number of channels should be used if the feature map includes multiple channels. On each channel separately, the pooling operations will be carried out[27].

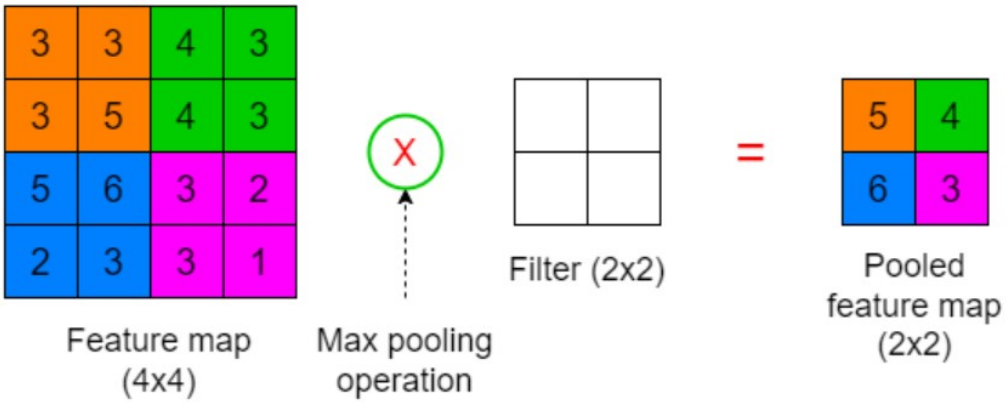


FIGURE 3.4: Max pooling between filter and feature map at stride 2[27]

Pooled feature map:The pooled feature map holds the results of various pooling operations between the filter and other feature map sections.

Feature map section: The size of the filter should be matched the size of the feature map area. To create different sections, filters could be moved on the feature map both vertically and horizontally. Depending on the Stride value, the number of sections will vary. Stride should be equal size of filer and padding is added to the feature map to maintain the size of the pooled feature map. These are important parameters (stride and padding) and should be applied in pooling layer. The number of channels remains unchanged when pooling is applied to the feature map. This indicates that both the feature map and the pooled feature map include the same number of channels. feature map having multiple channels should use filter with having same number of channels . Each channel will individually perform the pooling operations. Once its is done then Flatten operation starts.

Fully connected layer: The classification layer is another name for the fully connected layer. It is an useful method for identifying high-level with non-linear combination features from the convolution layer's output feature map. In order to be sent to the Fully Connected layer, feature vectors acquired from the convolutional layers are transformed into a 1-dimensional column vector. A fully connected neural network with one or more layers dedicated to classification makes up the FC layer. Every time the model is trained, backpropagation is applied to the fattened column vector before being passed to the feed-forward neural network..

3.5 Activation Function used in CNN layers

The activation methods used by CNN include Sigmoid, ReLU, and Leaky-ReLU. These are a non-linear alteration that assists the neuron to decide whether or not to pass information to the next layer as input.

3.6 Dropout

During a neural network's training phase, the neurons become more dependent on each other, which leads to over fitting. The network requires to be made more regular, which can stop it from being overloaded. Dropout is a common method for making a neural network more stable. Regularization makes the neural network's loss function worse so that it learns a set of weights that don't depend on each other. At each training iteration in this procedure, a neuron is deactivated with some probability value. All of the inputs and outputs that the neuron is connected to are cut off. At each training step, a neural network goes back to the neurons with dropped out at a rate of some probability values and a dropped-out neuro can get activated during the subsequent training phase. Dropout regularizes a neural network by minimizing the interdependence of the neurons' learning. The dropout method helps a neural network learn more stable features.

3.7 Batch Normalization

A network layer called batch normalization allows every layer to adapt more independently. The output of the earlier layers is normalized using it. It is possible to prevent the model from over fitting and increase learning efficiency by using batch normalization. The sequential model is integrated with this layer to normalize the inputs and outputs. Between the model's layers, it may be applied in a number of places. It typically follows the sequential model, the convolution layer, and the pooling layer.

3.8 Recurrent neural network

A feed forward neural network extension with internal memory is called a recurrent neural network. Given that it performs a same task for each data input and that the outcome of the current input depends on the computation of the previous input, RNNs are recurrent in nature. When the output is finished, a copy of it is forwarded back into the recurrent network. Before making a decision, it considers both the current input and the learning gained from the previous input's outcome. It features an encoder/decoder design.

RNNs use their internal state (memory) for processing inputs sequence, which is different from feed forward neural networks. This means they can be used for tasks like recognizing handwriting that is not broken up into separate lines or recognizing speech. In other neural networks, each input is separate from the others. But in RNN, all of the inputs are connected.

It first takes X(0) from the input sequence and then produces h(0), which, combined with X(1), is the input for the following step. As a result, the following phase's inputs are h(0) and X(1). Similarly, h(1) from the previous step is the input for the following phase, and so on. This way, it remembers the context while training. The advantage of Recurrent Neural Network is,

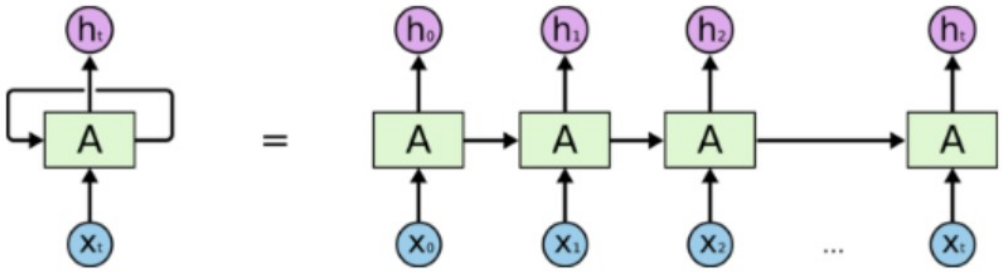


FIGURE 3.5: Recurrent Neural Network [19]

it simulates a data series such that prior samples are considered to be dependent on the current samples. RNN can be used with CNN layers to increase the effective pixel neighborhood. But there is also disadvantages to using RNN, such as gradient vanishing and explosion problems. It is not easy to train an RNN easily, It is a very difficult task, and last, it is not comfortable processing long sequences if relu and tanh is being used as an activation function [19].

3.9 Long Short Term Memory (LSTM)

LSTM is advanced version of RNN which helps to remember past data in memory. It is based on encoder-decoder method and capabilities to retain long memory. One issue, as discussed earlier in RNN is vanishsing gradient which is solved by LSTM. LSTM is supposed to be best and well suited for processing and classifying for predicting time series having lags of unknown duration. LSTM trains model use back-propagation methodology. As we can see in this Figure , LSTM has three gates present:

Input Gate: It helps to find out which input value must be used to modify the memory.The Sigmoid activation function figures out which values 0 or 1 can pass through. And the tanh function provides the weightage on the basis of passed values and decide their levels from -1 to 1 [19].

Output gate:The output is determined by the input and the block's memory. The sigmoid function determines which values to pass through 0,1, and the tanh function assigns weight to the values that are passed by determining their significance level, which ranges from -1 to 1, and multiplying with the output of the sigmoid [19].

Forget gate:it determines which information should be removed from the block. The sigmoid function makes that decision. It examines the input content (Xt) and the previous state (ht-1), and produces a number between 0 and 1 for each number in the cell state Ct1 [19].

LSTM has ability of learning long term dependencies. It was designed to overcome short-term dependency problems.It is capable of retaining informations for long time. But there are downside of LSTM. LSTM takes long

time to train as result it requires more memory to train. It has tendency to overfit and dropout is much difficult to implement. LSTM is also considered slow as it does too much computation.In addition to that , It is recursive nature and cannot be trained parallel .

3.10 Transformer Model

LSTMs have trouble with more difficult modern problems, such as machine translation across multiple languages or text generation that is indistinguishable from human-written text. A new architecture is introduced to solve such a complex task called Transformer.The first time the transformer was talked about was in the paper "Attention is All You Need," which was about translating languages. The architecture of the transformer is very complicated. But the most important part is the idea it is attention based models [38].A transformer is type of neural network that discovers context and subsequently meaning by tracing relationships in sequential data, such as the words in this phrase.Transformers are one of the newest and most powerful types of models that have been made so far. It is causing a wave of progress in machine learning that some people have called "transformer AI" [18]. Text and speech are being translated by "transformers" in almost real-time, making meetings and classrooms accessible to people with hearing loss.It helps scientists figure out amino acids in proteins and how the chains of genes in DNA works which helping speeding up the process of making new drugs.Transformers can find patterns and outliers to stop fraud, speed up production, make online suggestions, or improve health care. It is more easier to train than LSTM and less complexity and computational cost.

Transformer is replacing modern deep learning networks such as CNNs and RNNs.Transformers, unlike neural networks, make it easier to train on large sets of labeled data in less time. The model has a self-attention mechanism, which makes it easier to train on larger dataset. Prior to the Transformer model, training on large datasets with neural networks was a costly and time-consuming process. Furthermore, the computation performed by the transformer is simpler than that performed by other neural networks [18].

3.11 Transformer Attention Mechanism

Transformer model is also other type of neural network which contains encoder/decoder architecture Like LSTM or Other RNN. Transformers are incredibly strong due to small and specific innovations in the blocks. Positional encoders are used by transformers to identify data items entering and leaving the network. These tags are followed by attention units, which compute a kind of algebraic map showing how one element connects to the others. Self-Attention compares all the members of the input sequence to each other and changes the output sequence positions that match. In other words, the self-attention layer searches the input sequence for each key-value pair in a different way and adds the results to the output sequence. On the other

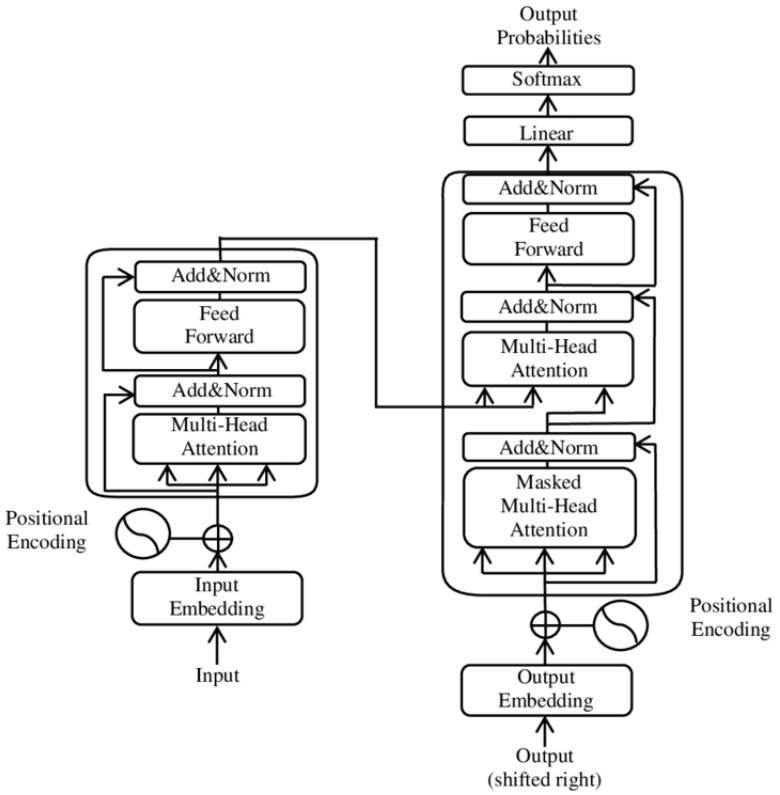


FIGURE 3.6: Architecture of Transformer [36]

hand Transformer is easy to train and has less computational complexity than LSTM .

Let's discuss it in action: As we previously discussed, Transformer is an attention-based model. What if the decoder had access to all the encoder's previous states rather than depending just on the context vector? That is precisely what attention does. The decoder may examine any specific state of the encoder at any stage of the decoding process. This is what distinguishes Transformer from LSTM [23].

The input sequences are supplied both forward and backward in this instance, and they are concatenated to create a content vector (Ci) before being handed on to the decoder. The attention model creates a context vector that is uniquely filtered for each output time step rather than placing the input sequence value into a single fixed context vector.Attention weights are used to ensure that the content vector of each output time step varies in accordance with how dependent it is on the input time steps[23]. .

3.12 Mask R-CNN

Mask R-CNN is a convolutional neural network (CNN) and the most advanced method for image segmentation or object detection. It is a region-based convolutional neural network known as Faster R-CNN, which served

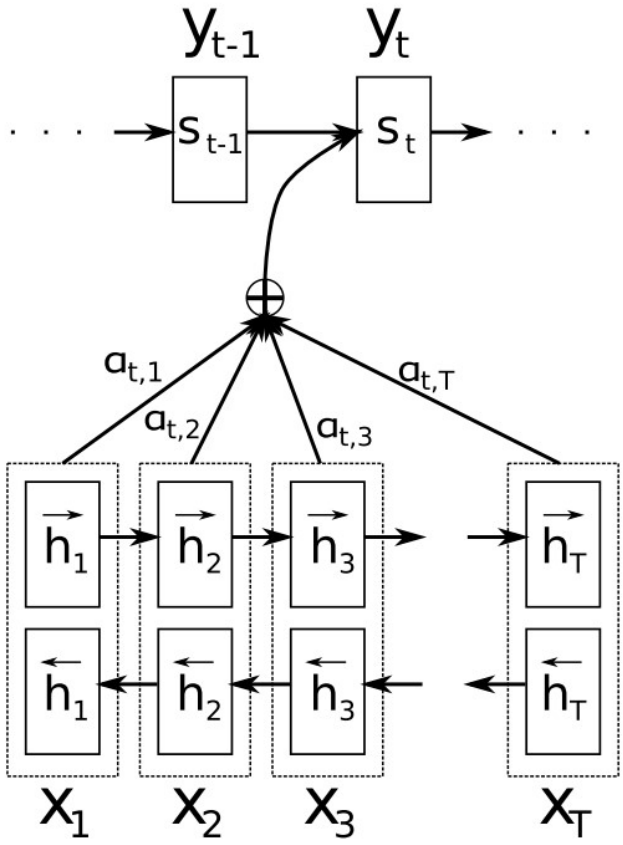


FIGURE 3.7: Attention Mechanism [23]

as the foundation for Mask R-CNN. Before we can understand how Mask R-CNN works, we need to know what image segmentation is ?. Image segmentation involves dividing a digital image into multiple parts . Under Mask R-CNN, there are two main kinds of image segmentation:

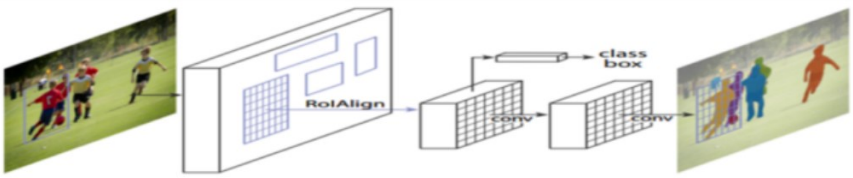


FIGURE 3.8: Mask R-CNN [25]

3.12.1 Semantic Segmentation

Semantic segmentation is the process of finding and grouping objects that are similar into a single class at the pixel level. As shown in the picture above, all of the objects were put into one group which is person. Since semantic

segmentation separates the objects in an image from the background, it is also known as background segmentation.

3.12.2 Instance Segmentation

Instance segmentation, also called instance recognition, is the process of finding all of the objects in an image and separating them into their own parts. So, it is the combination of finding objects, figuring out where they are, and classifying them. Alternatively said, this kind of segmentation goes further to demonstrate a clear distinction among each object that has been classified with other objects that are comparable. [25]. As shown in Fig. 3.9, instance segmentation was capable of detecting each object in the image (for example, the iris and pupil).

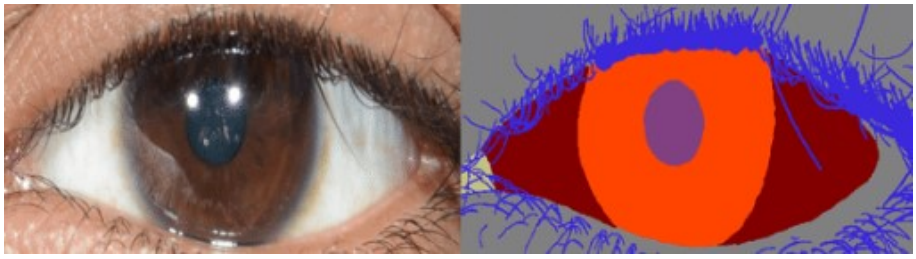


FIGURE 3.9: Example of segmentation [17]

3.12.3 Mean Average Precision (mAP)

A statistic called Mean Average Precision is used to assess object detection models, including Fast YOLO, and Mask R-CNN. Recall values between 0 and 1 are utilized to get the median of average precision values.It includes Confusion Matrix, Intersection over Union (IoU), Precision, and Recall through which the mAP performs object identification.

Chapter 4

Developing a Methodology for Identifying Patterns from Cataract Videos

4.1 Overview

Cataract surgery is a procedure in which surgeons remove the clouded lens and implant an artificial lens. This surgery is supposed to be one of the safest and most effective in the world. There are 10 phases of cataract surgery, such as incision, viscous agent injection, rhexis, hydrodissection,Irrigation and aspiration, phacoemulsification, capsule polishing, lens implant setting-up, viscous agent removal and tonifying, and antibiotics These are the main steps performed by surgeons in the cataract procedure. This surgery is performed under microscopical conditions, and it is recorded. These videos can be very helpful for research or training purposes. Cataract training is a complex procedure. Before performing actual surgery, a trainee must practice surgery several times. For this purpose, many trainees use animal eyes or synthetic eyes. Animal eyes are not easily available, and the setup for synthetic eyes is expensive. There are also different types of curricula, which include structured exams of surgical skills and help trainee surgeons become proficient in their learning and keep patients safe while doing surgery. Even though residents in residency training have to follow complex procedures to become successful doctors, some of their learning strategies are still outdated. While keeping these issues in mind, we developed a framework for cataract training based on deep learning, surgical data sets, and thick data analytic to bring some innovation to the learning process without using any animal eye modal. The framework will have three steps, and in each step different types of data and models will be used.

4.2 Step 1. Cataract detection

In this step, we'll talk about the dataset and classification model that were used to figure out the stages of cataract development. This stage will explain why cataracts are a serious medical condition. The learner was able to comprehend the appearance of cataracts and recognize when individuals need surgery.

4.2.1 Dataset

The three kinds of labels in this dataset are "normal eye, cataract, and "severe cataract eye. Each category in the dataset has more than 20 images. Surgery is necessary for the Server cataract since the patient is completely blind. The cloudy lens has entirely obstructed or dispersed light.

FIGURE 4.1: Normal Eye

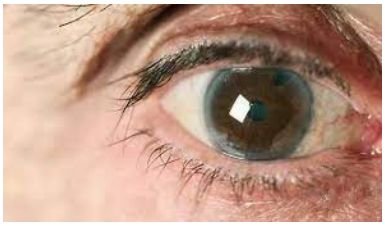


FIGURE 4.2: Cataract Eye

FIGURE 4.3: Sever Cataract Eye

We can see a normal eye, a cataract eye, and a severe cataract in the images above. When a cataract is severe, the lens becomes so densely clouded that it totally blocks light, rendering the patient entirely blind.

4.2.2 Training and model compilation

The dataset is divided into train and test dataset.Used data augmentation process to make our training more robust. Since we have small dataset, we have used Nasnetlarge pretrained CNN model for feature extraction. Then prepare Sequential model with 8 layers. Compiled model with 100 epochs, ,Adam optimizer, and a loss function categorical cross-entropy. The result comes out with a 78% accuracy rate.

4.3 Step 2: Video analysis of cataract surgery

This step will help the trainee understand the 10 phases of cataract surgery. Each phase is performed in sequence, using different tiny tools. The whole video of cataract surgery might be confusing for trainees, but breaking down each phase will help to provide a better understanding of the phases and their sequence. The proposed model will recommend the right sequence of cataract phases based on the current phase.

4.3.1 Dataset for Video Analysis

The dataset used in this work is taken from ITEC (Institute of Information Technology), which is part of ALPEN-ADRIA University KLAGENFURT (Austria). This institute conducts research in a variety of areas, including multimedia communication and information systems, as well as distributed and parallel systems. The dataset Catarct-101 [32] made up of videos of 101 cataract surgeries done by four different surgeons at the Department for Ophthalmology and Optometry at Klinikum Klagenfurt Austria's largest public hospital. These videos were collected and annotated by a senior ophthalmic surgeon with different phases of cataract surgery. Surgery was of recorded video with different experience level of doctors. Some of them had level one experience and others have more than level 2 experience. All of the videos have a PAL resolution of 720x540 pixels and are encoded as MP4 files using H.264/AVC with profile High as the video codec (25 frames per second, about 1.25 MBit/s bitrate) [30]. The dataset was divided into two groups. We extracted frames from a video of a cataract procedure performed by a less experienced surgeon as well as frames from a video of a procedure performed by a highly experienced surgeon. The duration of the "high experience" cataract surgeon video was less than low experience.

With 101 video data Figure 4.1 ,There are three CSV (comma-separated values) files that describe the 101 videos cataract dataset: annotations.csv, phases.csv, and videos.csv.

4.3.2 Annotated CSV

The annotation csv file has 1266 values. In that file, there are three columns named Video ID, Frame No, and Phase Table 4.1. For example, we can see Video ID 269, and the starting frame number of phase 1 is 68. For same video

FIGURE 4.4: Video Data from Dataset 101

the phase 2 starts from Frame No 1043 and phase 3 starts of video ID 269 from 1228.

TABLE 4.1: Phase CSV File Data Sample

Video ID	Frame No	Phase
269	68	1
269	1043	2
269	1228	3
269	2118	4
269	3478	5
269	6864	6
269	7727	7

This is just an example of an annotated CSV file for video ID 269. There many video which has specific ID in this format **case_**_video ID_**.mp4**

4.3.3 Phase CSV

Like annotation csv file , the phase cvs file describes cataract phases and their descriptions as shown in Table 4.2. We can see that phase.csv file has phase number in the first column and phase name in the second column.

TABLE 4.2: Phase CSV File

Phase	**Meaning**
1	Incision
2	Viscous agent injection
3	Rhexis
4	Hydrodissection
5	Phacoemulsificiation
6	Irrigation and aspiration
7	Capsule polishing
8	Lens implant setting-up
9	Viscous agent removal
10	Tonifying and antibiotics

4.3.4 Video CSV

The video csv file has five columns, eg Video ID, Frames, FPS, Surgeon and Experience. For this work, we are using just the Video ID and Frames columns. From Table below, we can see that video ID 269 has 14734 frames at rate 25 FPS.

TABLE 4.3: Video CSV File Data Sample

Video ID	Frames	FPS
269	14734	25
270	20500	25
271	16633	25
278	14185	25

4.4 Data Preprocessing for phases detection

We used the annotation.csv file for data preprocessing. We created a python algorithm for phase extractions. Let's see step by step how we created an algorithm for frame extractions. We used only one video to create data for further processing. We used only video number 269, Let's say we have video number 269; the start frame is 68 for phase 1, and the end of frame number for phase 1 is 1043. So here we are extracting 500 frames between 68 and 1043 for phase 1 and storing them in a folder "named phase 1". I followed the same steps for the other phases. We are extracting 500 frames per second and storing them in a folder and creating a name for the corresponding phase . We are creating 20 frames per second for each phase till 500. So, in other words, there will be 500 frames of each phase at a rate of 20 fbs. For further processing, all extracted datasets will be treated as a single dataset. Each phase folder has frames at a rate of 20 frames per second, which are converted into small video clips. We have given an input size of 512 X 512 X 3 for a convolutional neural network because it needs input data to be the same size and shape . We have also added padding to shorter video datasets to make it easier for CNN to pull out features. On the other hand, for coco-like datasets, we used the transfer learning module to build custom datasets for training using the Dataset class. The Dataset class makes it easy to work with any dataset.

```
df ← pd.readcsv(r'annotations.csv', sep ←';' )
df ← df[df['VideoID'] ← 269]                                  ▷ using video 269
FrameLists ← df['FrameNo'].to_list()
cap ← cv2.VideoCapture(r'case_269.mp4')
phase ← df['Phase'].to_list()
width ← int(cap.get(cv2.CAP_PROP_FRAME_WIDTH))
eight ← int(cap.get(cv2.CAP_PROP_FRAME_HEIGHT))
size ← (width, height)
fourcc ← cv2.VideoWriter_fourcc(*'XVID')
framecount ← 0
start ← 0
end ← 0
idx ← 0
while cap.isOpened() do
    ret, frame ← cap.read()
    if framecount in FrameLists then
        idx ← FrameLists.index(framecount)
        start ← framecount
        createDir(os.path.join(r'cataract101/', str(phase[id])))
        out ← cv2.VideoWriter(os.path.join((r'cataract101/'
, str(phase[idx])),
"video.avi"), fourcc, 20.0, size)
        if idx ≤ len(FrameLists) − 1 then
            end ← FrameLists[idx + 1]
        else
            end ← start + 500
        end if
        if start ≠0 end ≠0 framecount ≤start+500 then
            out.write(frame)
        else
            framecount+ = 1
        end if
        if idx ←len(FrameLists)-1 framecount ←end  then
            break
        end if
        cap.release()
        cap.release()
        cv2.destroyAllWindows()
```

end if

end while

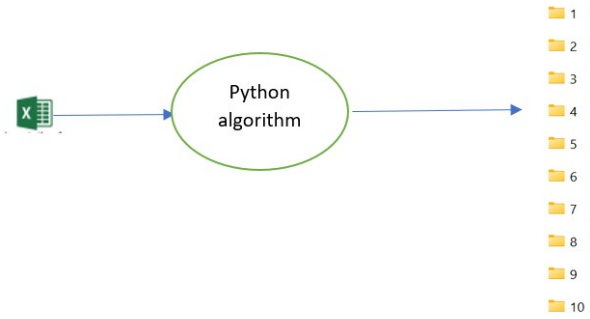


FIGURE 4.5: Data preparation

Figure 4.5 shows the algorithms in action. Algorithms extracted frames and put in folder corresponds to their phase name.

4.5 CNN-LSTM implementation for Phases Detection

The CNN-LSTM model can be created by first adding CNN layers, then LSTM layers with Dense dense layer on the output. In the first experiment, LSTM was used for phase predictions. We used Inception ResNet V2 for feature extractions with "image net" weights. The I Inception-ResNet-v2 is trained on more than a million images from the Image Net database. The network has 164 layers and can divide images into 1000 different types of objects. The InceptionResNetV2 has learned features in detail to represent a wide range of images. Other than that, the labels on the videos are strings. Neural networks don't understand string values, so they have to be changed into numeric values before they can be fed to the model. Here, we have used the StringLookup layer to turn the class labels into numbers.

4.5.1 Training and Model Compilation.

The whole dataset was divided into test data, test labels, train data, and train labels. The size of the feature frames in the train set is (10,500 ,1536) and the frame mask feature in train set is (10,500). First, we initialized frame feature input with parameters (MAX_SEQ_LENGTH,NUM_FEATURES) and mask_as_input with parameter (MAX_SEQ_LENGTH, type bool). The input of the first layer of LSTM was frame features and masks (labels). In all the layers, the activation function was used and also added dropout in the layers with a value of 0.1 . The final layer has soft max as an activation function. The model was compiled with Adam optimizer. This optimizer was a replacement for SGD (stochastic gradient descent), and the adam optimizer is very straightforward and efficient in terms of computation. Another benefit of using the Adam optimizer is that it is very suitable for noisy gradients. The cost function we have used is sparse categorical cross-entropy. This loss function is used for multi-class classification models where the output label is given an integer value, like our dataset.

4.6 Transformer model implementation for Phase Detections

The Transformer model is also another type of neural network that contains an encoder/decoder architecture like LSTM or other RNN. The main difference is that it is an attention-based model. In this experiment, we used a different feature extractor, which is DenseNet121. The input size and weights of the CNN model were the same as we used for InceptionResNetV2. Inception-ResNetV2 having very small difference with DenseNet. InceptionResNetV2 uses an additive method, which means that the output from one layer is used as the input for the next layer. Dense Net, on the other hand, uses the output from every layer as the input for the next layer.

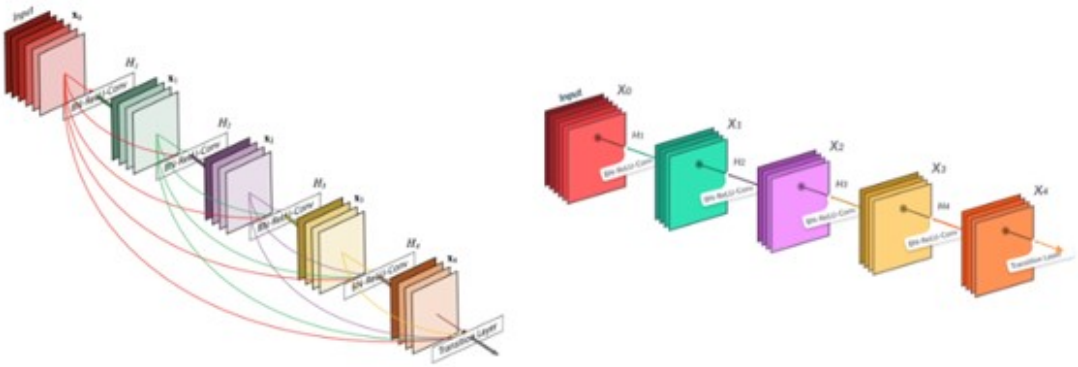


FIGURE 4.6: ResNet and DenseNet architecture

4.6.1 Training and Model Compilation

The dataset was already divided into training and test sets. Again, we have used the StringLookup layer to turn the class labels into numbers. First, the self-attention layers that make up a Transformer don't care about sequence. Since videos are made up of a series of frames OR frames are always in order taken from videos, our Transformer model needs to take this into account. Positional encoding is the way we do this. We used an embedding layer to store the positions of the frames in a video. Then, we added these positional embed dings to the CNN feature maps that have already been made. AS i said earlier the attention layer does not care about the sequence, so we need to use positional embedding. In simple words, we can say that it is used to tell a transformer about the position of an input vector, which is very important if we have data in a sequence. Other than Positional Embedding layers on top , we added GlobalMaxpooling1D layer which takes the max vector over the step dimension and is followed by a dropout layer at a rate of 0.1 and a final layer added with softmax as an activation layer. We used Adam as the optimizer and the same loss function, which is sparse categorical cross entropy.

4.7 Step 3: Iris Pupil Detection

In this section, trainees will learn about iris and pupil detection, which is one of the more complex aspects of surgery. There are also many complications that exist in these surgeries. One of the main complications is iris pupil detection. During cataract surgery, pupil and iris sizes are variable. Pupil responses, or the dilatation or constriction of the pupil during the surgical operation, are one such occurrence that might result in difficulties.During surgery, the eyes move due to instruments used in the surgical procedure. Also, it's not always possible to find circles in the frames, like when the visible part of the pupil or iris doesn't look round because of the tools used.

4.7.1 Dataset For Iris and Pupil detections

We used another dataset from ITEC [30] which keeps 82 frames from 35 videos of cataract surgery, and this dataset is also annotated with areas of iris and pupil. The dataset was recorded at Klinikum Klagenfurt(Austria). The resolution is 540X720pixels with frame rate of 25 fps. Since this dataset was taken during cataract surgeries, images have different surgical tools and depict the eyes in different states, such as with artificial lenses, without lenses, and tools used. This dataset is also annotated and it is in COCO Format. The dataset is already annotated and divided into test, train and validation. We just prepared using python methods which is described in transfer learning section.

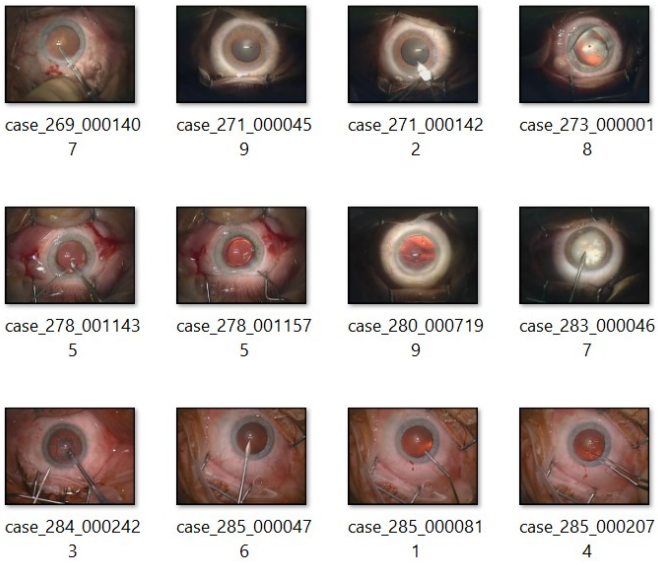


FIGURE 4.7: Dataset for iris and pupil segmentation [30]

4.7.2 COCO Dataset

The MS COCO dataset, which is a large-scale object detection, segmentation, and captioning dataset, was made available by Microsoft. Engineers who work with machine learning and computer vision often use the COCO dataset for different computer vision projects. Understanding visual scenes is one of the main goals of computer vision. This means figuring out what objects are there, where they are in 2D and 3D, what their properties are, and how they relate to each other. This dataset is very helpful in the field of computer vision and helps algorithms easily identify images or objects in images. This dataset can be used to train algorithms for finding objects and putting them into groups. COCO, or Common Objects in Context, was created to improve the goal of image recognition. The dataset has a high-quality visual dataset for work in the field of computer vision. It is also used to test algorithms and compare how well they can detect objects in real time. In this dataset, there are a total of 330 000 images, and more than 200 000 of them are labeled[22].

4.8 Format of COCO Dataset

The COCO format specifies how exactly bounding boxes, object classes, and image metadata such as height, width, and image sources are stored on disk. The COCO dataset is in JASON format. The format contains five sections of information for the entire dataset [22].

- Info- This section contains general information about dataset.
- Licenses- List of licenses with unique IDs that can be specified by images.
- Categories-Each classification category has its own ID. Possibly linked to a super category that can cover more than one class.
- Images- This section keeps list images and their relevant informations.
- Annotations- This section contains list of annotations information such as bounding boxes.

In all these fields,we need only Images, annotations and categories . We do not need Info and Licenses part [8].

4.8.1 Image List

The Image List has all kinds of information about images in a dataset, such as Unique ID (required field), width(required field) , height(required field) of images, file name(image file name which is required), data_captured (required field, date and time when the image was captured) and other information which is really not required, such as license, coco_url, and flickr_url. From table 4.3, we see that the image object of ID 105 contains all the relevant information related to this ID.

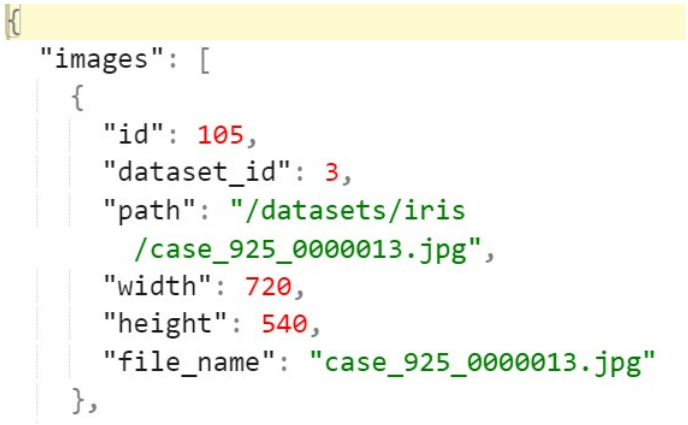


FIGURE 4.8: image information of coco like iris pupil dataset [30]

4.8.2 Annotations List

In the annotations list, information about the bounding boxes of all objects on all images is kept. One annotation object has information about the object's bounding box and its label on an image. For every object in an image, there is an annotated object. The list's required fields are image_id (which tells which image belongs to this annotation object), category_id(This identifier of the label that identifies the object inside a bounding box and also points to the id field of the categories array.), and bbox(which contains the coordinates of the bounding box of the image object. (Coordinates are in pixels.) There are no required fields in the list for segmentation, area, or is crowd. From table 4.4 , we see that bbox information and segmentation information are added for Image ID 105 of iris and pupil coco dataset.

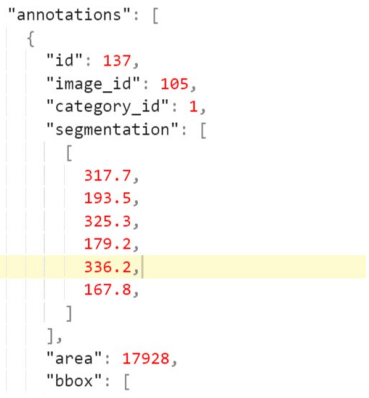


FIGURE 4.9: Annotation list of Image ID 105 [30]

4.8.3 Categories list

This list contains label information about objects. The required fields are ID(identifier of label; in an annotation object, the id field is linked to the category

id field) and name (the label name of the image). From figure4.5, we can see that in the iris pupil coco dataset, the category list has two IDs. ID no.1 belongs to the pupil and also a color code is added so that we can identify it during our training model. ID no.2 belongs to the iris label and its color is given.


```
"categories": [
  {
    "id": 1,
    "name": "pupil",
    "supercategory": "",
    "color": "#3917ee",
    "metadata": {}
  },
  {
    "id": 2,
    "name": "iris",
    "supercategory": "",
    "color": "#d0226a",
    "metadata": {}
  }
],
```


FIGURE 4.10: Category list of iris pupil coco like datset [30]

4.8.4 Transfer Learning Using Mask R-CNN for iris pupil dataset

We used Mask R-CNN for this work. Mask R-CNN is a deep learning neural network that is used for segmentation.This model has a branch of bounding box and classification regression, and it has a mask classifier which will generate a mask for every class. In deep learning, we need good data amounts to build our model. As we do not have enough iris pupil data, we have used pretrained model, which is also called transfer learning [26]. We used the Mask-R-CNN repository, which is MatterPort Mask R-CNN, to work on this dataset [12] and also downloaded the pre-trained weights for the Coco model from the same source. We imported all of the necessary libraries and overrides various classes such as CustomConfig classes, which are derived from config classes that helped to set up for training such as set up classes (we have three classes in the images, iris, pupil, and background), image dimension (512X512X3), and GPU capacity.

4.8.5 Data Preprocessing for Iris and pupil dataset

Once we are done with configuration , we have built an iris pupil custom dataset. For this work, we created a class(Iris_pupil_coco_like_dataset) which is derived from the Dataset class. This class helped to process the iris pupil JASON dataset. Dataset class allows for the simultaneous loading of many object from datasets. When we want to detect various objects but they are not all present in one data set, this class is really useful and provide different helpful methods to work on. There are different methods in this class that have been used such as load mask, load_datset,add_class and add_images.load_dataset (load_data_iris_pupil) method helped to iterate through all object of dataset such as image object,annotation object and categories in Jason file and using add_class(load_data_iris_pupil) and add_image, a custom dataset was created. Another important class that we used is load_mask class (iris_pupil_load_mask). This method generates masks for every object in the image, such as the mask over the iris or pupil. It will return class ids, one mask per instance, and a one-dimensional array of class ids for the mask instance.

4.8.6 Training

The dataset was divided into validation, test, and training datasets. We used the instance of the dataset derived class(Iris_pupil_coco_like_dataset) to prepare for training. We loaded pre-trained weights for Mask R-CNN from COCO data for better results. We run through 250 epochs at a learning rate of 0.0001. Once model training was completed, we saved this model in a folder and later used this trained model for prediction.

4.9 Thick data: Surgeon experience as a expert heuristic to improve model performance.

Cataract 101 has numerous videos of cataract surgery performed by both experienced and inexperienced surgeons. For example, take video ID 269, which was done by a less experienced surgeon, and video ID 350, which was done by a highly experienced surgeon. We extracted frames from video ID 350 to create a new dataset. We used this dataset to train our models. We found that the model's performance was better than the dataset (frames taken from video))269, the reason may be that frames extracted from video ID 350 of each phase are more accurate because a highly experienced surgeon performed each phase very well. There is a possibility that the framed work of a less experienced surgeon will be less accurate. Because a less experienced surgeon may be repeating some phases or surgery performing timing is more than high experience surgeon, the frames of each phases are not collecting accurately from video.

TABLE 4.4: Accuracy rate of CNN-LSTM and Transformer with two different dataset

	CNN-LSTM	Transformer
Type of dataset	Accuracy(%)	Accuracy(%)
Less Experience dataset	10	20
High Experience dataset	20	40

From Table 4.4. We can see that both model performance and accuracy were enhanced by using dataset taken from highly experienced surgeons.

Big data has been employed in several earlier studies for all deep learning tasks. Using big data or video data sets, a lot of study has been done in the topic of cataract. Big data is not always simple to access. Processing large amounts of data is always costly. We have used thick data analytic with deep learning to accomplish our task.Using thick data expert heuristic , our model was improved . To avoid huge data, we used transfer learning as support.

Chapter 5

Results and Discussions

5.1 Step 1. Cataract detection

The main goal of this step was to detect the stage of cataract eg. Normal, Cataract, and Sever Cataract. The accuracy is 78%. The used call back function to save best model because model was overfiting. We used early model check point function to save best model and used this model for inference. F-1 score is an error matrix which range between 0 for worst and 1 for best. We can from confusion matrix that our model is preforming good in all three dataset. The f-1 score is between 0 and 1.

	precision	recall	f1-score	support
cataract	0.57	0.72	0.63	18
normal	0.74	0.88	0.80	16
severe_cataract	0.67	0.35	0.46	17
accuracy			0.65	51
macro avg	0.66	0.65	0.63	51
weighted avg	0.65	0.65	0.63	51

FIGURE 5.1: Cataract detection confusion matrix

We used sever cataract image for testing and we got correct result.

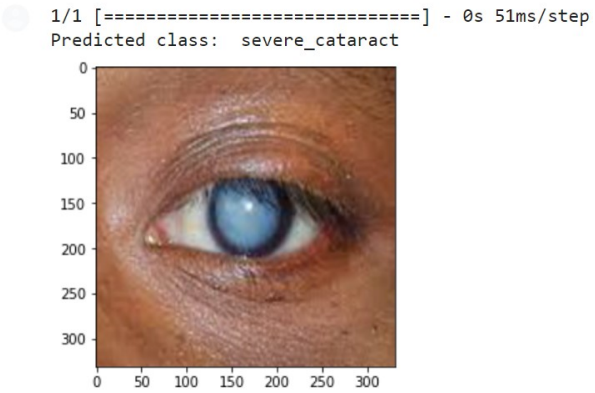


FIGURE 5.2: Cataract detection model inference result image

5.2 Step 2 : Iris Pupil Results

The main goal is to detect iris pupil from images with bounding box. We created new inferenceConfig class for predictions. The model was trained for 150 epochs and saved in a folder for inference. In the InferenceConfig class, we added the following parameters so that we can get better results.

TABLE 5.1: Inference Configuration set up parameters

Parameters	**Value**
GPU_COUNT	1
IMAGES_PER_GPU	1
IMAGE_MIN_DIM	512
IMAGE_MAX_DIM	512
DETECTION_MIN_CONFIDENCE	0.9

The DETECTION_MIN_CONFIDENCE is very important parameters as it helps to detect bounding boxes, classes and confidence percentage without any restrictions. The confidence score displays the likelihood and degree of certainty with which the classifier believes that the box contains an item of interest. The confidence score should be zero if there is nothing in that box.In our case, we have given 90 percent, which means that a model could detect objects with bounding box very less restrictively. We preprocessed test images using skimage library and then looped through all test images for prediction. Predicted results are

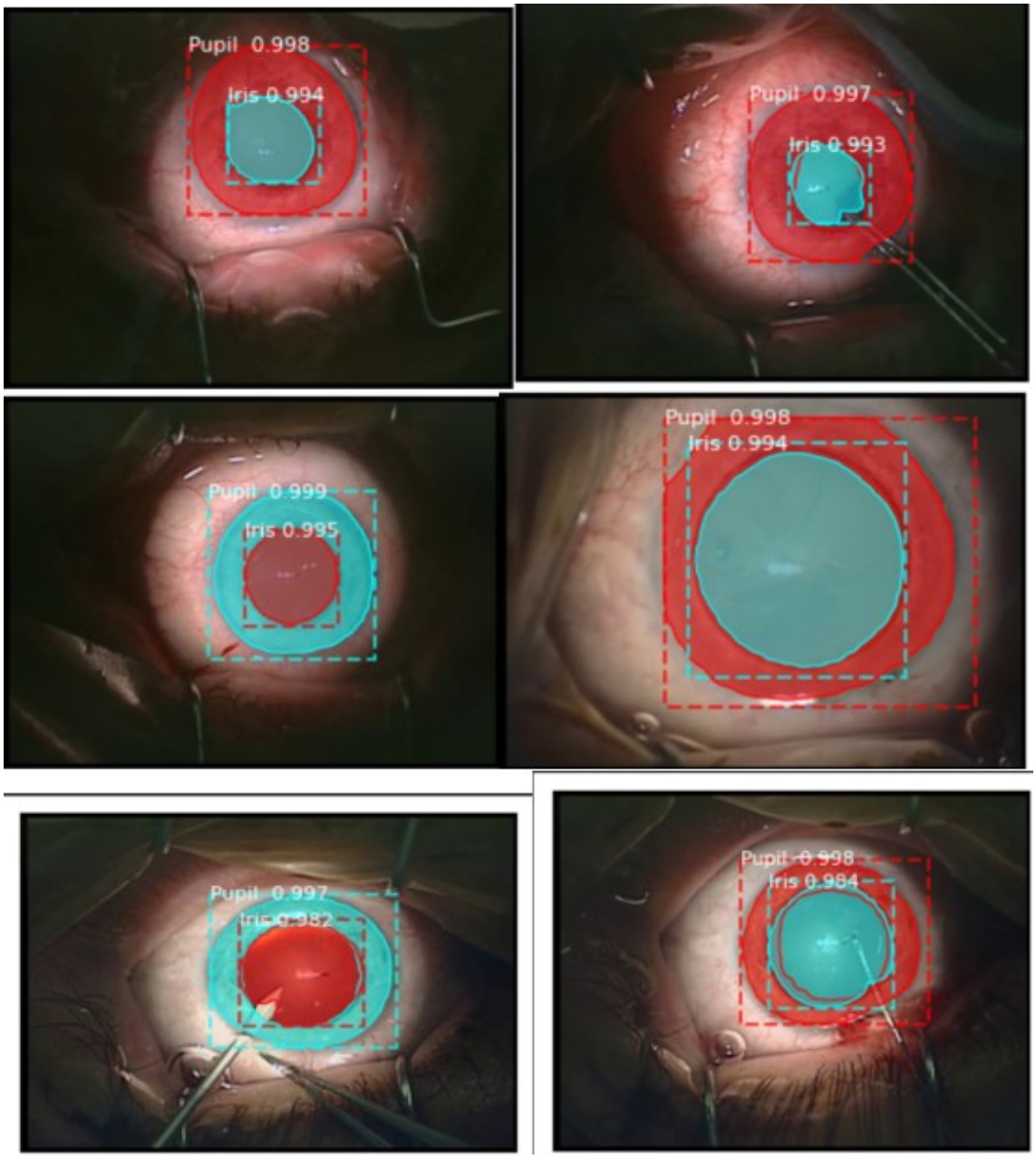


FIGURE 5.3: Iris pupil detection result

5.2.1 Mean Average Precision Result for Mask R-CNN

mAP metric is used to evaluate object detection model results. Throuhgh mAp, Precision has been calculated by whole image dataset.We used mAP to understand how our model performed. Mean Average Precision metrics are used for object detection models to get an overview of the model's performance. The mAP metrics were calculated for Mask RNN to understand how this object detection model performed over the coco dataset. Based on the mAP metrics, the Mask RNN model has performed well. For example, in Figure 5.2, the model processed the first images taken from the validation folder with only a ground truth vect is 2, indicating that the images have two objects, iris and pupil, and a predicted vect value is 2, indicating that the model detected two objects (iris and pupil) correctly and created a bounding box over them .The average precision of particular image is 1 which also states that model were able to identify correctly two objects. The actual mean average precision for whole image is also 1. The precision is called positive predictive value. It is the ratio of correct predictions to the total predicted positives [12].

```
Processing 1 images
image                    shape: (512, 512, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 512, 512, 3)      min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 15)               min:    0.00000  max:  512.00000  int64
anchors                  shape: (1, 65472, 4)         min:   -0.70849  max:    1.58325  float32
the actual length of the ground truth vect is :  2
the actual length of the predicted vect is :  2
Average precision of this image :  1.0
The actual mean average precision for the whole images 1.0
Processing 1 images
image                    shape: (512, 512, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 512, 512, 3)      min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 15)               min:    0.00000  max:  512.00000  int64
anchors                  shape: (1, 65472, 4)         min:   -0.70849  max:    1.58325  float32
the actual length of the ground truth vect is :  4
the actual length of the predicted vect is :  4
Average precision of this image :  1.0
The actual mean average precision for the whole images 1.0
Processing 1 images
image                    shape: (512, 512, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 512, 512, 3)      min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 15)               min:    0.00000  max:  512.00000  int64
anchors                  shape: (1, 65472, 4)         min:   -0.70849  max:    1.58325  float32
the actual length of the ground truth vect is :  6
the actual length of the predicted vect is :  6
Average precision of this image :  0.3333333432674408
The actual mean average precision for the whole images 0.777777781089147
```

FIGURE 5.4: Mean Average Precision of Mask R-CNN

5.3 CNN-LSTM Result Analysis

The result of LSTM was not really good it has only. The accuracy rate for the model was not satisfactory. Initially, we used one video to extract frames and prepare the dataset to run the model. The accuracy of the model was bad on this dataset. Secondly, we increased the number of frames by extracting more frames from two videos, but again we did not get a good result. We used one video clip for phase 1 from the test dataset. Figure 5.2 shows the predicted result.

FIGURE 5.5: Predicted result

```
10: 20.71%
1: 18.37%
9: 13.93%
2:  8.21%
6:  7.90%
8:  7.61%
5:  6.66%
3:  6.33%
4:  5.97%
7:  4.31%
```

FIGURE 5.6: Predicted result metrics CNN-LSTM

From Figure 5.3, we can see that predicted result metrics are provided by the trained model. In simple words, it describes the possibilities of a predicted video clip that belongs to phase 1. It described that there was an 18.37% chance that it belonged to phase 1, while a 20.71% chance that the

test video clip belonged to phase 10. The model provided a false prediction here. We have used a video clip of phase 1 from the test dataset. Here, the model is totally confused about this and is not able to understand sequence of frames. Other than that, there might be issues with the feature extraction process that is why the model is not able to perform very well.

5.4 Transformer Result Analysis

The prediction result was slightly better than CNN-LSTM. We used the same video clips that we used in the prediction for the LSTM model. We got better results than LSTM.

```
1: 25.67%
10: 18.33%
8: 17.74%
6: 12.25%
2: 10.35%
7:  7.61%
3:  4.30%
5:  2.45%
9:  0.88%
4:  0.41%
```

FIGURE 5.7: Predicted result metrics of Transformer

We can see that the model predicted a better result than CNN-LSTM and that it described the correct result. It described that 25.6%7 chances, which was higher than other phases values, that the video clip was from phase 1. Even though, model was not perfect, it provided a good overview. But again, the model was confused about Phase 1 and Phase 10. This model has said that this video clip could be phase 10 because the model provided a higher chance value, which is 18.13% after Phase 1. According to me, Phase 1 and Phase 10 are very different steps which also can not be repeated.

5.5 Limitation of predicted results of CNN-LSTM and Transformer

The accuracy rate of CNN-LSTM is only 10% and Transformer model is only 20%. Both models were not able to perform very well on the data. We added more frames, epochs and layers in both model but did not get satisfactory result. Fig show the loss graphs of CNN-LSTM and Transformer.

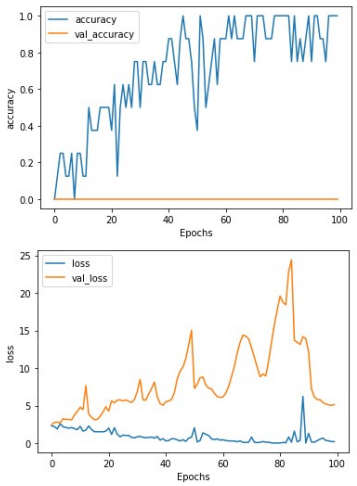


FIGURE 5.8: Accuracy and Loss graph of LSTM

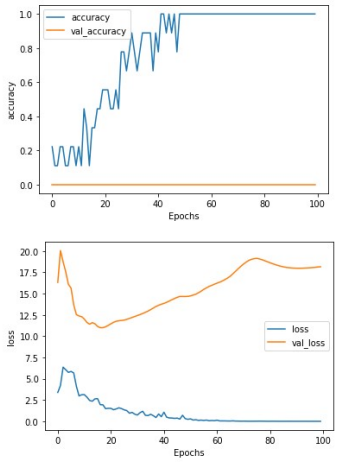


FIGURE 5.9: Accuracy and Loss of Transformer

5.5.1 Github repositores of our tasks

- Step 1:`https:https://github.com/Csingh1s/thesis_work_work_cataract_detection`
- Step 2: `https://github.com/Csingh1s/thesis_work_work_cataract_detection`
- Step 3: `https://github.com/Csingh1s/thesiswork-Iris_pupildetection`

Chapter 6

Conclusion and Future Work

6.0.1 Conclusion

We offer a multitasking learning process for cataract surgery from detection to surgery. Utilizing three different dataset types, such as image datasets, video datasets, and coco-like datasets of Cataract, we were able to complete three linked tasks in the same domain. The framework will assist students and trainees expedite their training for cataract surgery. The major goal of the research was to improve the coherence of cataract instruction. A skilled surgeon can complete the treatment in 10 to 15 minutes. Surgeon residents do not get enough time to observe the process. Only the surgeon performing the surgery and a second person with an additional ocular can see the entire process because it is being done under a micro- scope. It is not similar to endoscopic surgery or other surgeries in any way. To improve training and make it simpler to understand, The major goal of this study is to mitigate issue found by new training and make their learning smooth. On the other side, this framework is helpful for both catract patients and trainee surgeons. A person with cataracts must take a number of steps, including consult an eye surgeon, getting an ultrasound, and waiting for the surgeon's turn. All of these procedures are quite expensive and lengthy, and in many developing nations, thousand of people do get such a facility where patients can go check their eye and find cataract severity. This framework (Step 1) makes it simple to demonstrate whether or not they have a contract.

6.1 Future Work

The accuracy rate of the LSTM and transformer must be improved. Datasets play a crucial role in any machine- or deep learning task. In this case, a well-organized dataset for cataract surgery would be helpful for phase extractions and tool identification to improve the accuracy rate. Other than during cataract surgery, some phases are done twice. This could happen because the surgeon does not have enough experience. There could be risk and complexity during cataract surgery, such as bleeding, IOL instability, retinal detachment, and so on. A trainee surgeon should be aware of these types of problems. A new model or frame- work should be designed to address these complexities

6.2 Outline

6.2.1 Chapter 1

This chapter briefly introduced the thesis, including the background and problem motivation, overall aim, detailed problem statement, scope, and proposed solutions. For example, the background of cataract surgery, the research question, "What are problems in cataract training methods and how could they be improved?"

6.2.2 Chapter 2

This chapter summarizes existing research on video analysis of cataract surgery using deep learning and big data, as well as how thick data could be used in research. It describes how researchers exploited deep learning and big data methodologies for their findings using a surgical dataset.

6.2.3 Chapter 3

This chapter provides a brief overview of the technology used to complete the thesis, such as deep learning, convolutions neural networks, the transformer model, and the LSTM model, as well as thick data analytics, and also explains why the transformer model is superior to the LSTM model and what metrics are used to evaluate the models. also describes an object detection model and the metrics that will be used to evaluate the model.

6.2.4 Chapter 4

This chapter talks about the implementation of the entire framework of the cataract surgery workflow. It interprets the overall system's design and implementation. It states what steps are taken to design and build each step of the framework and how thick data helped to improve of the model. How transfer learning helped us avoid using big data and also provided background on different types of datasets, such as image datasets and video datasets, used in each step of the framework.

6.2.5 Chapter 5

The performances of the framework was assessed in this Chapter.Model results and their evaluation matrices are explained, and what is unsolved and problems that could be further investigated are mentioned. Predicted images, result tables, and graphs are clearly explained in this chapter, along with limitations and task Github link.

6.2.6 Chapter 6

This chapter talks about Conclusion and Future work.

List of References

[1] Hassan Al Hajj et al. "Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network". In: *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2017, pp. 2002–2005.

[2] Abdullah M. Alfawaz. "Ophthalmology resident surgical training: Can we do better?" In: *Saudi Journal of Ophthalmology* 33.2 (2019), pp. 159–162. ISSN: 1319-4534. DOI: https://doi.org/10.1016/j.sjopt.2018.11.009. URL: https://www.sciencedirect.com/science/article/pii/S1319453418302145.

[3] Luca Antonioli et al. "Convolutional Neural Networks Cascade for Automatic Pupil and Iris Detection in Ocular Proton Therapy". In: *Sensors* 21.13 (2021). ISSN: 1424-8220. DOI: 10.3390/s21134400. URL: https://www.mdpi.com/1424-8220/21/13/4400.

[4] Larry Benjamin. "Training in surgical skills". en. In: *Community Eye Health* 15.42 (2002), pp. 19–20.

[5] Sebastian Bodenstedt et al. "Active learning using deep Bayesian networks for surgical workflow analysis". In: *International journal of computer assisted radiology and surgery* 14.6 (2019), pp. 1079–1087.

[6] Gaudenz Boesch. *Top 10 Applications Of Deep Learning and Computer Vision In Healthcare - viso.ai*. [Online; accessed 2022-08-22]. June 2022.

[7] *Cataract surgery training around the world*. https://eyewiki.aao.org/Cataract. [Online; accessed 2022-11-10].

[8] *COCO format - Rekognition*. https://docs.aws.amazon.com/customlabels-dg/md-coco-overview.html. [Online; accessed 2022-08-29].

[9] *Common Eye Problems - Glaucoma, Cataracts, AMD More | Seema Eye Care Center*. URL: https://seemaeye.com/common-eye-conditions/. (Accessed: Aug. 22, 2022).

[10] K Diederik and J Adam Ba. "A method for stochastic optimization. arXiv 2014". In: *arXiv preprint arXiv:1412.6980* (2015).

[11] Jinan Fiaidhi, Darien Sawyer, and Sabah Mohammed. "Thick Data Analytics for Small Training Samples Using Siamese Neural Network and Image Augmentation". In: *LISS 2021*. Springer, 2022, pp. 57–66.

[12] . *GitHub - matterport/Mask_RCNN: Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*. [Online; accessed 2022-08-30]. Mar. 2019.

[13] Michal Kawka et al. "Intraoperative video analysis and machine learning models will change the future of surgical training". In: *Intelligent Surgery* 1 (2022), pp. 13–15. ISSN: 2666-6766. DOI: https://doi.org/10.1016/j.isurg.2021.03.001. URL: https://www.sciencedirect.com/science/article/pii/S266667662100003X.

[14] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[15] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[16] Daniel Josef Lindegger, James Wawrzynski, and George Michael Saleh. "Evolution and Applications of Artificial Intelligence to Cataract Surgery". In: *Ophthalmology Science* 2.3 (2022).

[17] Natalia Mathá et al. "Pixel-Based Iris and Pupil Segmentation in Cataract Surgery Videos Using Mask R-CNN". In: Apr. 2020, pp. 1–4. DOI: 10.1109/ISBIWorkshops50223.2020.9153367.

[18] Rick Merritt. *What Is a Transformer Model? | NVIDIA Blogs*. 2022. URL: https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/. (Accessed: Aug. 28, 2022).

[19] Aditi Mittal. *Understanding RNN and LSTM*. Aug. 2021. URL: https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e.

[20] Shoji Morita et al. "Real-Time Extraction of Important Surgical Phases in Cataract Surgery Videos". In: *Scientific Reports* 9.1 (Nov. 2019), p. 16590. ISSN: 2045-2322. DOI: 10.1038/s41598-019-53091-8. URL: https://doi.org/10.1038/s41598-019-53091-8.

[21] Shoji Morita et al. "Real-time extraction of important surgical phases in cataract surgery videos". In: *Scientific reports* 9.1 (2019), pp. 1–8.

[22] *MS COCO Dataset: Using it in Your*. https://datagen.tech/guides/image-datasets/ms-coco-dataset-using-it-in-your-computer-vision-projects. [Online; accessed 2022-08-29]. June 2022.

[23] Harshith Nadendla. *Why are LSTMs struggling to matchup with Transformers? | by Harshith Nadendla*. URL: https://medium.com/analytics-vidhya/why-are-lstms-struggling-to-matchup-with-transformers-a1cc5b2557e3. (Accessed: Aug. 28, 2022).

[24] Keiron O'Shea and Ryan Nash. "An introduction to convolutional neural networks". In: *arXiv preprint arXiv:1511.08458* (2015).

[25] Elisha Odemakinde. *Everything about Mask R-CNN: A Beginner's Guide - viso.ai*. [Online; accessed 2022-08-30]. Mar. 2022.

[26] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

[27] Rukshan Pramoditha. *Convolutional Neural Network (CNN) Architecture Explained in Plain English Using Simple Diagrams*. URL: https://towardsdatascience.com/convolutional-neural-network-cnn-architecture-explained-in-plain-english-using-simple-diagrams-e5de17eacc8f/. (Accessed: Aug. 27, 2022).

[28] P.A R. C. Consultants. *Dislocated Lens Implant in Sarasota Manatee County, FL*. 2022. URL: https://oaklandeye.com/iol-options/. (Accessed: Aug. 22, 2022).

[29] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[30] Klaus Schoeffmann et al. "Cataract-101: Video Dataset of 101 Cataract Surgeries". In: *Proceedings of the 9th ACM Multimedia Systems Conference*. MMSys '18. Amsterdam, Netherlands: Association for Computing Machinery, 2018, pp. 421–425. ISBN: 9781450351928. DOI: 10.1145/3204949.3208137. URL: https://doi.org/10.1145/3204949.3208137.

[31] Klaus Schoeffmann et al. "Cataract-101: video dataset of 101 cataract surgeries". In: *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*. Ed. by Pablo César, Michael Zink, and Niall Murray. ACM, 2018, pp. 421–425. DOI: 10.1145/3204949.3208137. URL: https://doi.org/10.1145/3204949.3208137.

[32] Klaus Schoeffmann et al. "Cataract-101: video dataset of 101 cataract surgeries". In: *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*. Ed. by Pablo César, Michael Zink, and Niall Murray. ACM, 2018, pp. 421–425. DOI: 10.1145/3204949.3208137. URL: https://doi.org/10.1145/3204949.3208137.

[33] Pan Shi et al. "Real-Time Surgical Tool Detection in Minimally Invasive Surgery Based on Attention-Guided Convolutional Neural Network". In: *IEEE Access* PP (Dec. 2020), pp. 1–1. DOI: 10.1109/ACCESS.2020.3046258.

[34] *StackPath*. URL: ttps://www.naturaleyecare.com/blog/lens-cataracts-stem-cells/. (Accessed: Aug. 22, 2022).

[35] Ummey Hani Tanin. "Deep Video Analysis Methods for Surgical Skills Assessment in Cataract Surgery". PhD thesis. Carleton University, 2022.

[36] *Transformer (machine learning model) - Wikipedia*. https://en.wikipedia.org/wiki/Transfo [Online; accessed 2022-12-16]. Aug. 2019.

[37] Andru P Twinanda et al. "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos". en. In: *IEEE Trans Med Imaging* 36.1 (July 2016), pp. 86–97.

[38] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[39] *Viscoelastic Substance - an overview | ScienceDirect Topics," Oakland Ophthalmic Surgery, P.C.* URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/viscoelastic-substance/. (Accessed: Aug. 22, 2022).

[40] *What are video analytics?* https://www.axis.com/learning/web-articles/video-analytics. [Online; accessed 2022-12-09].

[41] *Why Big Data Needs Thick Data?* https://www.datascienceacademy.io/blog/why-big-data-needs-thick-data/. [Online; accessed 2022-12-10]. July 2020.

[42] Hsu-Hang Yeh et al. "PhacoTrainer: A Multicenter Study of Deep Learning for Activity Recognition in Cataract Surgical Videos". In: *Translational vision science & technology* 10.13 (2021), pp. 23–23.

[43] Felix Yu et al. "Assessment of Automated Identification of Phases in Videos of Cataract Surgery Using Machine Learning and Deep Learning Techniques". en. In: *JAMA Netw Open* 2.4 (Apr. 2019), e191860.