

Fire and Disaster Detection with Multimodal Quadcopter By Machine Learning

by

Anika Afrin

19301072

Md Moshior Rahman

20101096

Ayash Hossain Chowdhury

20101095

Mirza Eshraq

20101094

Mehvish Rahman Ukasha

20101097

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering.

Department of Computer Science and Engineering
Brac University
March 2023

© 2023. Brac University
All rights reserved.

Declaration

We, the authors of this thesis, certify that this is an original work that has not been submitted for any other academic honour. We affirm that this work was completed by us and is an accurate portrayal of our own labour and ideas. All sources utilised in this work have been properly cited, and no information has been withheld by design. We further affirm that all experiments in this study were conducted in accordance with our institution's research ethics committee's standards and regulations. This study's data was acquired and managed with the utmost care to preserve confidentiality and privacy. In addition, we affirm that we have not attempted to fabricate any data or alter the results in any way. The conclusions and recommendations offered in this thesis are based exclusively on our study findings and are unaffected by external variables.

We are aware that any infringement of the norms of academic integrity and ethical research procedures may result in disciplinary action. In all future undertakings, we commit to preserve the highest standards of academic integrity and ethical research techniques.

Student's Full Name & Signature:



Anika Afrin

19301072



Ayash Hossain Chowdhury

20101095



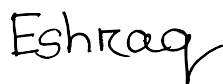
Md Moshior Rahman

20101096



Mehvish Rahman Ukasha

20101097



Mirza Eshraq

20101094

Approval

The thesis/project titled “Fire, Disaster Detection with Multimodal Quadcopter By Machine Learning” submitted by

1. Anika Afrin (ID: 19301072)
2. Md Moshour Rahman (ID: 20101096)
3. Ayash Hossain Chowdhury (ID: 20101095)
4. Mirza Eshraq (ID: 20101094)
5. Mehvish Rahman Ukasha (ID: 20101097)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on March 27, 2023.

Examining Committee:

Supervisor:
(Member)



Prof. Dr. Khalilur Rahman

Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Ms. Sadia Hamid Kazi

Chairperson, Associate Professor
Department of Computer Science and Engineering
Brac University

Ethic Statement

This thesis paper's authors understand research ethics. We acknowledge the importance of ethical research when using datasets, models, and tools. This thesis's data and resources were gathered and used ethically. This research utilises datasets and photos with permission from appropriate authorities. We secured everyone's privacy and personal information. Our research was also honest, transparent, and accountable. We avoided prejudices and conflicts of interest. We expressed our findings honestly and acknowledged our research's limits and uncertainties. We considered how our study would affect society and worked to improve knowledge and human lives. We have taken steps to reduce risks and maximise benefits of our research. Our research also acknowledges and cites others. We avoided plagiarism and academic dishonesty in our work. We feel our research has met the highest ethical standards. Our work should improve society and knowledge.

Abstract

Our thesis research is consisted of developing a model that can detect early fires and, mapping the area for fire and disaster detection using a UAV (Unmanned Aerial Vehicle) or quadcopter if a fire break out. Furthermore, MLP uses fire or no fire detection, sound analysis, and input sensor to create a multimodal system architecture. First, surveillance cameras detects the early stages of a fire using luminous smoke and textured flame. However, if the fire has already started, an alarm will sound, activating the quadcopter operation. Due to the quadcopter's camera and sound system input, it obtains an aerial perspective and maps the fire-affected region while indicating human life. Finally, disaster detection provides us with a map indicating the safe zone where the less damaged part of the building will assist the fire department in saving human lives. The unique aspect of our thesis is that it designs a complete fire detection and rescue model. It will effectively detect a fire before an incident occurs and map the fire-affected region after the incident with human life signs and the safest path to rescue. The main goal here is to prevent or mitigate damage by immediately alerting the fire department. We have collected primary dataset of Fire and Disaster. Moreover, we increased the accuracy of our fire dataset to 80.32% and increased the accuracy of our disaster dataset to 9.2%. We tried to reduce the false detection of fire. Added to that, we have integrated all the five models in graphical user interface.

Keyword: YOLOV5; YOLOV7; Fire detection; Disaster detection; Sound detection; Mapping; MiDaS V3; PIX4D Mapper.

Dedication

We dedicate this thesis to our parents, whose unwavering support and encouragement have been our greatest motivation. We also dedicate it to our supervisor, whose guidance and expertise have been invaluable to our success. Above all, we dedicate it to all the victims and fallen heroes of uncertain disasters and fire incidents, whose memories will always inspire us to work towards making the world a safer place.

Acknowledgement

We would like to convey our profound appreciation to Allah, the Almighty, for His blessings and guidance along the path of writing this thesis. Our sincerest gratitude goes to our supervisor, Dr. Md. Khalilur Rahman, PhD, for his consistent direction, unflinching support, and insightful input. His wide knowledge, competence, and insightful remarks were crucial in guiding the direction of our research. Also, we are grateful to BRAC University for allowing us to pursue higher education and conduct this research endeavour. We would like to thank the Department of Computer Science Engineering for providing us with the required resources and facilities. Throughout our academic path, our family gave unfailing support and encouragement. Their love and assistance were important in assisting us to overcome the difficulties and obstacles we encountered during our research.

Table of Contents

Declaration	i
Approval	ii
Ethic Statement	iv
Abstract	v
Dedication	vi
Acknowledgment	vii
List of Figures	x
List of Tables	xi
1 Introduction	xii
1.1 Introduction to our work	xii
1.2 Introduction to our report	xiii
1.3 Problem Statement	xiv
1.3.1 Research Objective	xiv
1.3.2 Literature Review	xv
2 Related Works	xx
2.1 Fire and Smoke Detection	xx
2.2 Human Detection	xxii
2.3 Sound Detection	xxiii
2.4 Disaster Detection	xxiii
2.5 Mapping	xxv
3 Architecture	xxvii
3.1 System Model	xxvii
3.2 Research Methodology	xxviii
3.3 Working Plan	xxx
4 AI Models related to our work	xxxiv
4.1 Description of the models	xxxiv
4.1.1 YOLOV5	xxxiv
4.1.2 YOLOV7	xxxv
4.1.3 MiDaS V3	xxxix

4.1.4	SSL (Sound Source Localization)	xl
4.1.5	ASR speech recognition	xli
4.2	Description of the softwares	xlii
4.2.1	PIX4DMAPPER	xlii
4.2.2	Okkhor	xlvi
5	Implementation	xlvii
6	Dataset	xliv
6.1	Description of Data	xliv
6.1.1	Description of Fire Dataset	xliv
6.1.2	Description of Sound Dataset	li
6.1.3	Description of Disaster Dataset	li
6.1.4	The process of Disaster Mapping	lii
6.2	Data sample	liv
6.2.1	Training Set	lv
6.2.2	Testing Set	lv
6.2.3	Validation Set	lvi
6.3	Data label	lvii
6.4	Image resizing	lvii
6.5	Data Augmentation	lvii
7	Result Analysis	lix
7.1	Preliminary Analysis	lix
7.1.1	Comparison between YOLO V5 and YOLO V7	lix
7.1.2	Comparison between MEMS directional sound sensor Sound Source Localization System	lxvi
7.1.3	Comparison between LiDAR Depth Mapping and MiDaS V3	lxvii
7.2	Comparison with Related Works	lxvii
7.3	Evaluation Of Results	lxviii
7.4	Discussion	lxx
8	Conclusion	lxxiii
8.1	Challenges	lxxiii
8.2	Contribution	lxxiv
8.3	Limitations	lxxvi
8.4	Future Work	lxxvii
	Bibliography	lxxxii

List of Figures

1.1	Quadcopter	xv
3.1	Model Of Our Multimodal Fire Detection System	xxvii
3.2	Working Plan	xxxii
4.1	YOLOV5 Architecture	xxxv
4.2	YOLOV7 Architecture	xxxvi
4.3	YOLOV7 Architecture for fire dataset	xxxvii
4.4	YOLOV7 Architecture for disaster dataset	xxxviii
4.5	Architecture of MIDaSV3	xxxix
4.6	Sound Source Localization	xl
4.7	Automatic Sound Recognition	xli
4.8	PIX4DMapper	xliii
4.9	Workflow of PIX4D Mapper	xliv
4.10	Performance of PIX4D	xliv
4.11	Okkhor App	xlvi
4.12	Output of Okkhor App	xlvi
5.1	Implementation of GUI	xlviii
6.1	Leveled and annotated image of fire, no-fire, smoke dataset	1
6.2	Leveled and annotated image of gathering, human dataset	1
6.3	Leveled and Annotated Images of Dataset Disaster	lii
6.4	Image Capturing Pattern	liii
6.5	Training Set	lv
6.6	Testing Set	lvi
6.7	Validation Set	lvi
7.1	Curve from YOLO V5 model	lx
7.2	P curve	lxii
7.3	PR curve	lxiii
7.4	R curve	lxiv
7.5	All result graphs of Fire Detection	lxv
7.6	All result graphs of Disaster Detection	lxvi
7.7	Output of ASR Model	lxix

List of Tables

7.1	Comparison of YOLOV5 and YOLOV7	lxi
7.2	Comparison of Class Accuracy of YOLOV5 and YOLOV7	lxi

Chapter 1

Introduction

1.1 Introduction to our work

Fire is one of the most destructive natural disasters that can occur on any scale. Fires can occur naturally or as a result of human-caused behavior (Marrion, 2016). Lightning strikes, sparks in arid areas, volcano eruptions, etc. can all cause natural fires. Man-made fires can be started intentionally, by careless behavior or by accidents like gas explosions or malfunctioning electrical equipment, among many other things.

In the last few decades, Bangladesh has faced many horrific fire incidents. Tazreen fashion, Shitakundu container depot, Rana plaza, Nimtoli tragedy, etc. are worth mentioning incidents of the disaster history of Bangladesh. If the first-aid measures for fighting a fire are taken, fire damage can be avoided (Muhammad et al., 2018). In our thesis research, we are implementing multimodal system to detect fires and damages accurately in fire affected area. At first, we are detecting both fire and building collapse disaster during the incident so that it can be informed early to the rescuers. For this, we used YOLOV5, and YOLOV7 models for detecting fire and disaster from pictures or videos taken from UAV.

At first we have used YOLOV5 for training to detect both fire and disaster, but now we are using YOLOV7 which is the latest version of YOLOV5 and gives more accurate result. Secondly, apart from detecting fire and disaster from pictures and videos, we are analyzing the sounds of the fire and disaster affected area in order to save lives more quickly. We are using two models for sound analyzing: ASR (Automated speech recognition) to detect speeches of affected area by transcribing the sounds into texts, and SSL to detect the direction and source of the sounds. Thirdly, we are estimating depth from a single RGB image to observe the human postures of victims using MiDaS V3.

Lastly, we are mapping the fire affected area in order to find safest routes to save human lives using PIX4D Mapper.

[32] [38]

1.2 Introduction to our report

Our thesis paper offers a fresh perspective on how to deal with emergencies like fires and natural disasters. In this study, we present our work in developing a multi-modal Graphical User Interface (GUI) for real-time decision making, using artificial intelligence models and software tools for fire detection and mapping, disaster risk assessment, and depth estimation.

In the first chapter, we set the stage by discussing the larger backdrop of our work and pointing out the growing importance of effective fire and catastrophe management. Our study presents a well-defined issue description and research goals, which centre on creating state-of-the-art AI models and software tools for precise disaster and fire detection and mapping, depth estimates, and real-time decision making. In addition, we perform a literature analysis to look at the works that came before us in the field of disaster and fire management. In the second chapter, we discuss the literature review of previous studies that are similar to our own. We stress the need for more advanced and efficient models and tools for disaster and fire management, and we explore the limitations of existing models and software tools. Our architecture, system model, research methods, and action plan are all laid out in chapter 3. The many parts of our system are outlined, such as data collection and annotation, AI model training, model deployment in real-time, and the creation of a multimodal GUI for decision making. Also, we detail the many artificial intelligence models and software tools we employed during our studies.

YOLOV5 and YOLOV7 for fire and disaster warning and mapping, Pix4D Mapper for depth estimation, and the ASR model for speech recognition are only some of the AI models and software tools that we explore in chapter 4. In this section, we describe in depth the models and software tools that we used and how we incorporated them into our system. In Chapter 5, we detail the steps we took to put our multimodal GUI into action, from gathering and annotating datasets for fire and disaster detection to training artificial intelligence models with the YOLOV7 and ASR models. In chapter 6, we describe our datasets in detail, including the data description, data samples, data labels, image scaling, and data augmentation techniques we employed to improve our datasets for better detection and mapping. In chapter 7, we discuss our findings and draw conclusions about the efficacy of our models and other tools. A preliminary comparison is made between YOLOV5 and YOLOV7, LiDAR and MiDaS, MEMS and SSL. We also draw parallels to other publications and explain the scope and limitations of our study. In the final chapter of our thesis paper, we reflect on what we've accomplished and the obstacles we've encountered along the way. We also offer suggestions for further research and ways to improve our models and software tools for better disaster and fire management. As a last step, we list the books, articles, and websites that served as inspiration for this study Bibliography.

1.3 Problem Statement

Bangladesh has many fires and disasters. Bangladesh is densely populated and resource-poor. Floods, cyclones, and earthquakes can damage it. Due to its climate and unsafe buildings, Bangladesh is prone to fires. Electrical failure, improper fire use, and negligence cause fires in Bangladeshi factories, warehouses, and homes. Burning trash, candles, and other flammables can start fires. Structural defects, faulty electrical wiring, and electrical system neglect can cause fires. Fires can cause economic, environmental, and human losses. Bangladesh Fire Service and Civil Defence responds to fires and disasters.

Fires and disasters kill many Bangladeshis. Between 2004 and 2019, the Bangladesh Fire Service and Civil Defence (FSCD) reported 7,000 deaths and over 55,000 injuries from fires and disasters, including natural disasters. Over 60,000 homes, businesses, and other structures were destroyed or damaged. Electrical short-circuits, improper electrical equipment maintenance, and improper flammable material storage cause most fires and disasters in Bangladesh.[44] Arson, petrol leaks, and careless matches and lighters are other causes. Deaths, injuries, and property damage are the most common fire and disaster casualties in Bangladesh. Disasters also evict thousands. The FSCD uses public awareness campaigns, fire safety training, and legislation to reduce fire incidents and disasters in Bangladesh. The FSCD also has fire-detection systems and specialised firefighting units in all major cities and is working with the government to improve disaster response.

Bangladesh has 15,000 fires per year, killing 800 and injuring 8,500. 17,541 fires occurred in 2017, killing 743 and injuring 8,753. Electrical short circuits, boiler malfunctions, and human error cause most Bangladeshi fires. Short circuits caused 32% of fires in 2017. Boiler malfunctions caused 23.7% of fires, followed by human error at 17.7%. Between 2015 and 2019, the Bangladesh Fire Service and Civil Defence Department reported 18,844 fire incidents. Fire incidents peaked in 2018 at 6,088. 2017, with 4,854 fires, had the fewest.[47] 2015 had the most fire-related deaths, 229. 2019 had 105 fatalities. Residential areas, followed by commercial and industrial areas, have been most affected by fires in Bangladesh. 2016, with 1,817 fire-related injuries, had the most. 2019 saw 1,247 injuries.

[15] Our multimodal fire detection drone uses sensors, models, algorithms, and technologies to detect and prevent fires. These models detect fire, smoke, heat, and other fire hazards. They can also detect dangerous people and fire hazards. Drones can map disaster areas to guide firefighting. Combining these technologies, the drone can alert authorities to a potential fire hazard and provide detailed information about the fire and surrounding areas. This data can help prevent or extinguish fires. The drone can track the fire's spread in real time. This helps firefighters assess the situation and decide what to do. Drones can also inspect hard-to-reach areas. This can improve understanding and reveal hidden dangers.

[16]

1.3.1 Research Objective

There are certain objectives of our thesis that we wish to complete while conducting our research. The following objectives are;

1. Learning about ML (Machine learning) and CNN (Convolutional Neural Net-

work) in depth. Moreover, we want to learn about various methods and algorithms for object detection, classification, and chromatic segmentation.

2. Designing a complete model for fire detection in three stages and rescuing human life.
3. Achieving greater accuracy than previous models by applying an effective algorithm with a more effective dataset.
4. Detecting the early fire stage more fastly and accurately before spreading.
5. Getting 3D map from an aerial perspective images of the quadcopter to calculate and analyze the whole fire disaster area.
6. Depth estimation of the fire-affected area by analyzing fire and disaster.
7. Using a multimodal system to get accurate information about the fire and damage.
8. Analyzing human behavior around fire and in the fire.
9. Analyzing noises, and screams from the surroundings to detect fire more accurately and survive life in the fire.
10. Finding the safest route for rescuing human life with disaster mapping by analyzing the less damaged areas.
11. If possible, we want to apply active learning in our model for further enhancement.



Figure 1.1: Quadcopter

1.3.2 Literature Review

Vision-based fire and smoke detection is proposed using deep learning. An improved YOLOv5 algorithm detects fire and smoke with bounding boxes. SPP module added dilated convolution to improve multiple convolution and pooling operations in varied sizes. The activation function GELU (Gaussian Error Linear Unit) was utilised instead of SiLU (sigmoid-weighted linear unit), and DIOU-NMS was employed as the predicted bounding box suppression instead of NMS to improve small flame target identification and model convergence. The updated YOLOv5 outperforms the original model in fire and smoke detection and can exceed 99.3% mAP@0.5 on their fire dataset. The updated YOLOv5s detects 125 frames per second for real-time. This method is also reliable for small-scale flame detection, reducing missed and inaccurate fire detection and enhancing accuracy. This model cannot detect different colour flames and smoke. The algorithm's higher detection accuracy in complex conditions makes it suitable for video fire detection. This research investigates

fire detection using Light-YOLOv5. In (SepViT), a separable vision transformer block improves the backbone network’s access to external data and the extraction of flame and smoke characteristics by replacing numerous Cross Stage Partial Bottleneck modules with three convolutions (C3) components in the final layer. Light-BiFPN has also been utilised to minimise model weight, improve feature extraction, and balance speed and accuracy. Adding a global attention mechanism (GAM) to the network improves global dimensional findings. The Mish activation function and SIOU loss accelerate convergence and increase accuracy. This model has 6.1% higher accuracy and confidence than YOLOv3-tiny, 5.5% higher than YOLOX-s, and 6.8% higher than YOLOv7-tiny. This model is inaccurate in detecting little flames or smoke outside a limited range. The paper introduced Swin-YOLOv5, which improved the model’s receptive field and feature identification without impacting depth. The feature splicing mechanism of the three output heads of the feature fusion layer network was improved to improve weighted Concat feature map splicing and model pair feature fusion. Weighted feature splicing was added to this module to improve network feature fusion. This method’s average range accuracy rises faster than the benchmark algorithm, according to experiments. For the same dataset, this approach improves 0.7% and high-precision target recognition by 1.8 FPS (fast packet switch). The improved technique spotted smoke or fire that had been missed or misidentified in the same experimental dataset. This work introduced YOLOv5 feature extraction and fusion with fire-smoke detection in indoor and outdoor contexts. This paper presented a DL-based fire detection system using pre-trained sequential YOLOv5 and U-Net models. They fed the YOLOv5 model fire photos and annotations, then clipped the fire class using detection bounding boxes. They transmit cropped photos to a pre-trained U-Net model with the original images and annotations to construct segmented images with boundary lines. They employed wildfire and fire-like picture datasets. This article found that this wildfire detection architecture has a drastically reduced false alarm rate. This paper discusses real-time intelligent fire detection and forecasting employing cameras, fire development features extraction, and prediction. Flame position, speed, and width describe fire evolution. Two neural networks identify fire properties. Fire properties are extracted by the Region-Convolutional Neural Network (RCNN) and forecasted by ResNet. RCNN extraction yields a mean relative error (MRE) of 4-13%, 6-20%, and 11-37% for the three parameters. ResNet’s MRE for the three parameters is 4-13%, 11-33%, and 12-48%. It shows that the proposed technique can quantify fire development, improve industrial fire safety, forecast fire development trends, evaluate accident severity, calculate accident losses in real time, and direct fire fighting and rescue strategies. This paper proposes a high-speed fire detection approach using YOLOv7 and CN-B network model. Compared to YOLOv7, this integrated flame detection approach is more versatile and extracts flame features better. The YOLOv7-CN-B technique is 5% more accurate and mAP is 2.1% more accurate than YOLOv7. Single detection took 11.9 ms and 149.25 FPS. Experimental data shows that YOLOv7-CN-B outperforms the traditional algorithm. This study presents an enhanced YOLOv5 smoke detection method. Baseline model is YOLOv5m. mosaic enhancement randomly cropped, scaled, and arranged nine pictures to create new images. A dynamic anchor box technique addresses YOLOv5’s erroneous anchor box prior information. The modified method’s detection accuracy was 4.4% greater than the baseline model’s mAP and 85 FPS faster than the

traditional deep learning algorithm, meeting engineering application requirements. NVIDIA Jetson modules were used to test YOLOv5's ability to distinguish small human-objects from an unmanned aerial vehicle (UAV). Little object detection was solved by the RGB and thermal infrared-trained YOLOv5 model. VisDrone's RGB and TIR pictures dataset improved the YOLOv5 model's UAV person recognition accuracy by 79.8% and 88.8%, respectively. Despite this study was not conducted for a difficult situation, they claimed that, With the NVIDIA Jetson module, a sophisticated surveillance system can be coupled into a multi-agent UAV with an advanced AI concept to determine the cost-performance of this technique. They proposed YOLOv5-HR-TCM, a quick, consistent end-to-end model to predict 3D human posture (YOLOv5-HRet-Temporal Convolution Model). Their 2D to 3D lifting technique handles person detection, 2D human posture estimation, and 3D human pose estimation. The model uses the best techniques from each level. This method is fast and accurate. The process takes 3.146 FPS on a low-end machine. They tested low-resource ML-based action recognition systems using the "Okutama-Action" dataset. Although camera angle and flying height are constrained in this dataset, it covers action recognition scenarios. They employed object recognition and classifiers for single-image action detection. Their architecture uses YoloV5 and a gradient boosting classifier for scalable and effective object identification and a classifier that can handle samples from multiple sets. The Okutama-Action dataset was used to evaluate their ablation research of YoloV5 designs. This architecture beat past Okutama dataset architectures that differed by item identification and classification pipeline. This technique may function poorly if the gimbal angle exceeds 90 degrees or the drone flies above 30 metres. Motion blurring from the drone's speed may also affect system performance. "Characterizing Human Explanation Techniques to Guide the Design of Explainable AI for Building Damage Assessment" gives preliminary data on how humans use satellite photos to assess building damage from natural disasters. Participants explained building damage assessment using satellite images from the xBD dataset [3]. xBD satellite photographs show natural disaster-damaged buildings. To collect damage explanations from humans, they created a web-based annotation system. Before the study's main session, 60 participants were trained. Using iterative, open-coding, they categorised participant explanations. 60 participants generated 929 codes (average 1.55 codes per pre/post image combination, SD = 1.05). 12.8% of assessments (N = 77) had no code because participants did not explain. "Satellite radar and optical remote sensing for earthquake damage detection: results from different case studies" examines how well remote sensing methods can identify damage in urban settings and how radar (SAR) and optical satellite data can be used together. They examined remote sensing pros and cons to develop a reliable damage assessment method. Automated methods have successfully assessed pixel-by-pixel categorization and damage assessment in homogeneous extended areas. The Bam earthquake test case yielded approximately SAR categorization was 61%, optical data 70%, and data fusion 76%. Satellite and ground survey data match. "A Rapid Self-Supervised Deep-Learning-Based Method for Post-Earthquake Damage Detection Using UAV Data(Case Study: Sarpol-e Zahab, Iran)" proposes a novel deep-learning-based method for quickly identifying earthquake-damaged buildings. Three processes detect issues on four levels. Three feature types—non-deep, deep, and fusion—are examined to determine the best feature extraction method.

"One-epoch convolutional autoencoder (OECAE)" separates deep and non-deep features. Designing a rule-based method to automatically select classification algorithm training samples is the next step. Finally, seven popular machine learning (ML) algorithms are used to create building damage maps. Auto-training samples are practical and better than manual ones, with OA and KC improvements of 22% and 33%, respectively. SVM outperformed MLP (OA = 82%, KC = 73.98%). Fusing deep and non-deep data using OECAE may also improve damage-mapping efficiency compared to methods using either deep or non-deep features alone. "Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks" automates building damage detection in satellite images. This study examines CNNs' disaster-related building damage detection generalizability. They compare CNN architectures on a single dataset before comparing the best CNN architecture after being trained and validated in various transfer learning settings. The research also evaluates four CNN architectures, model generalizability, and cross-region transfer learning using a dataset from the Haiti disaster. "Earthquake-Induced Building Damage Mapping Using Explainable AI" maps and detects disasters. MLP and SHAP can detect these. The MLP algorithm can extract features from satellite images and open street maps (OSM) of the affected region, and the SHAP algorithm can rank the most important extracted features to detect collapsed or non-collapsed buildings. "Deep Learning and Stereo Vision Based Detection of Post-Earthquake Fire Geolocation for Smart Cities within the Scope of Disaster Management: İstanbul Case" presents a stereo-YOLO framework for fast and accurate post-earthquake fire detection. YOLOv5 uses cameras to detect fires. Stereo vision algorithms assess a fire's visual properties to geolocate it when many cameras spot it. WSN sends geo-location data to the emergency management centre to quickly put out the fire. "Mapping and 3D modelling using quadrotor drone and GIS Software - Journal of Big Data" used Inspire 2 quadcopter drones with RGB cameras, photogrammetry, geographic data systems for scenario mapping, and a high-quality camera with dreadlocks for picture stability. Google earth data at two locations was used to measure area at three flying heights of 40, 80, and 100 m. The results were 98.53% (98.68%), 95.2% (96.1%), and 94.4% (94.7%) for each altitude. Mapping with very high-resolution satellite data is difficult because image collection is expensive, especially for localised areas that require daily or weekly information repetitions. "Real-time 3D mapping using a 2D laser scanner and IMU aided visual SLAM" used a 2D laser scanner and visual-inertial fusion to map in real time. EKF-based IMU-aided visual SLAM estimates pose. Real-time 2D laser scans from a moving robot create a global, consistent 3D map. This study tested a Turtlebot with a Kinect, IMU, and laser scanner to create 3D point cloud maps of multiple offices. "Voxel Hashing" was also proposed for online reconstruction using consumer depth cameras. A compact spatial hashing mechanism combines hierarchical data structures and implicit surfaces for reconstruction with little overhead. Hashing allows real-time operation without sacrificing scale or reconstruction quality. Parallel graphics hardware optimises all processes. Our unstructured approach preserves volumetric fusion while eliminating spatial data structure overhead. This method allows lightweight streaming without data structure reconfiguration, which may increase reconstruction limits. "3D Mapping Using Lidar" describes a simple and inexpensive method for mapping interior structures in 3D using Light Detection and Ranging (LDR) (LIDAR). LIDAR measures angles and distance from both servo motors simultaneously. These num-

bers compute and illustrate the internal structure in 3D. Photogrammetry for 3D mapping is complicated, expensive, and time-consuming. Photogrammetry's precision can be unacceptable. 3D mapping's realistic image enhances visualisation. Engineering, surveying, and research utilise it.

Chapter 2

Related Works

2.1 Fire and Smoke Detection

1. The paper suggests a deep learning strategy for vision-based fire and smoke detection. This method uses an improved YOLOv5 algorithm to identify fire and smoke with bounding boxes. In order to improve the outcome of multiple convolution and pooling operations in different sizes, dilated convolution was added to SPP module. Also, the activation function GELU (The Gaussian Error Linear Unit) was used in place of the SiLU (sigmoid-weighted linear unit), and DIoU-NMS was chosen as the predicted bounding box suppression instead of NMS which improved the suitability for identifying small flame targets and accelerated model convergence. The experimental results demonstrate that the modified YOLOv5 performs better than the original model in terms of fire and smoke detection and that its accuracy can exceed 99.3% mAP@0.5 on their fire dataset. In order to meet the real-time requirements, the upgraded YOLOv5s can detect 125 frames per second. Additionally, this technique has high reliability for small-scale flame detection, which reduces missed and incorrect fire detection while increasing accuracy. But this model lacks in terms of fire detection for different color flames and smoke. In conclusion, the suggested algorithm is capable of meeting the task's performance requirements for video fire detection as it has higher detection accuracy in complex situations.[39]

2. In this paper, Light-YOLOv5 algorithm has been studied in fire detection. Here (SepViT) a separable vision transformer block has been used to improve the backbone network's connection to external data and the extraction of flame and smoke features, several Cross Stage Partial Bottleneck modules are replaced with three convolutions (C3) components in the final layer of the backbone network. Additionally, a light bidirectional feature pyramid network (Light-BiFPN) has been used to reduce model weight while enhancing feature extraction and balancing speed and accuracy. Then by fusing a global attention mechanism (GAM) into the network, the model gives more accurate results on the global dimensional characteristics. Finally, the Mish activation function and SIoU loss are used to concurrently speed up convergence and improve accuracy. As a result, this model has higher accuracy and confidence level compared to other models like 6.1% high than YOLOv3-tiny, 5.5% higher than YOLOX-s, and 6.8% higher than YOLOv7-tiny. However, this model shows low accuracy in detecting small flames or smoke if the target is outside a limited range.[40]

3. In this paper, a new algorithm Swin-YOLOv5 has been introduced that improved the model's receptive field and feature detection capabilities without affecting its depth. The feature splicing method of the three output heads of the feature fusion layer network was modified to improve the feature map splicing method of weighted Concat and increase the feature fusion capability of model pairs. To enhance the network feature fusion capability, this module was further modified and the weighted feature splicing mechanism was included. According to experiment results, this method's map (average range accuracy) increases more quickly than the benchmark algorithm. The outcome of this algorithm is improved by 0.7% on the same dataset, while the high-precision target identification performance increases by 1.8 FPS (fast packet switch). The enhanced method was better at detecting smoke or fire that had either not been discovered or had been detected incorrectly under the identical experimental dataset. This work developed a practical notion for feature extraction and fusion of YOLOv5, as well as an opportunity for the use of fire-smoke detection in indoor and outdoor environments.[41]

4. In this paper, they offered a DL-based fire detection system that integrates pre-trained sequential YOLOv5 and U-Net models. At first, they gave the YOLOv5 model the original images of fires along with their annotations, and then they cropped the fire class using the bounding boxes identified by detection. In order to create the segmented images with their boundary lines, then they send those cropped images to a pre-trained U-Net model using the original images with their annotations.

In this, they have used a dataset of wildfire mixed with a dataset that contains fire-like images. This paper concluded that the experimental results proved the reliability of this architecture for wildfire detection with a significantly improved result in a false alarm.[42]

5. In this research, a combination of real-time intelligent fire detection and forecasting approaches using cameras has been discussed with fire development features extraction and prediction. The evolution of the fire is described using three parameters: the position, speed, and width of the flames. In order to extract fire characteristics through fire detection, two neural networks have been used. The Region-Convolutional Neural Network (RCNN) is for extracting the characteristics of fire and the Residual Network (ResNet) is for fire forecasting. Results suggest that for the three parameters, the mean relative error (MRE) of extraction using RCNN is approximately 4-13%, 6-20%, and 11-37%, correspondingly. However, the MRE of ResNet's prediction for the three parameters is approximately 4-13%, 11-33%, and 12-48%, respectively. It demonstrates that the suggested strategy can deliver a workable solution for quantifying fire development and enhancing industrial fire safety, including forecasting fire development trends, evaluating accident severity, calculating accident losses in real time, and directing fire fighting and rescue tactics.[1]

6. In this research paper, a high speed fire detection method based on a combination of YOLOv7 and CN-B network model has been proposed. This combined flame detection approach is more flexible and has better ability to extract flame features

compared to the YOLOv7 algorithm. The results suggest that the YOLOv7-CN-B method is 5% more accurate and mAP is 2.1% more accurate compared to the YOLOv7 approach. The single detection speed was 11.9 ms, and the detection speed was 149.25 FPS. The YOLOv7-CN-B approach has better output than the standard algorithm, according to the experimental data.[30]

7. In this paper, a smoke detection algorithm has been developed with the help of improved YOLOv5. Here, YOLOv5m has been used as a baseline model. mosaic enhancement method was used to randomly crop, scale and arrange nine images to form new images. To solve the problem of inaccurate anchor box prior information in YOLOv5, a dynamic anchor box mechanism is proposed. The detection accuracy of the improved algorithm in this study was 4.4% better than the mAP of the baseline model when compared to the standard deep learning algorithm, and the detection speed hit 85 FPS, which is clearly better and can satisfy engineering application requirements.[31]

2.2 Human Detection

1. In this study, the capacity of YOLOv5 to recognize small human-objects from the perspective of an unmanned aerial vehicle (UAV) has been examined using NVIDIA Jetson modules. The YOLOv5 model, was trained using RGB and thermal infrared images, gave a successful outcome for resolving the small object detection issue. With AP value up to 79.8% and 88.8% for RGB and TIR images, respectively, the RGB and TIR photos dataset from VisDrone was able to enhance the accuracy of the YOLOv5 model in order to recognize the human from a UAV perspective. Though this study was not conducted for a complex scenario, they have claimed that, With the NVIDIA Jetson module, a sophisticated surveillance system can be combined into a multi-agent UAV with an advanced AI concept in order to determine the cost-performance of this approach.[34]

2. In this research, they proposed a fast, consistent end-to-end model to predict the 3D human posture called YOLOv5-HR-TCM, (YOLOv5-HRet-Temporal Convolution Model). In order to handle their proposed model is based on the 2D to 3D lifting approach where each phase of the estimation process such as person detection, 2D human pose estimation, and 3D human pose estimation is handled. The suggested model combines the best techniques from each level. This approach maintains processing speed while achieving high accuracy. On a low-end computer, the estimated duration of the entire process is 3.146 FPS. [35]

3. Using a previously compiled real-world dataset (the "Okutama-Action" dataset), they investigated low-resource algorithms for ML (machine learning)-based action recognition. Although this dataset is limited for image acquisition characteristics like camera angle and flying height, it covers scenarios that are representative for action recognition. To support single-image action detection, they used the combination of object recognition and classifier techniques. YoloV5 and a gradient boosting classifier have been used in their architecture for the purpose of using a scalable and effective object recognition system along with a classifier that can deal with sam-

ples of different sets. They examined various YoloV5 architectures in an ablation study and analyzed the effectiveness of their approach using the Okutama-Action dataset. This approach outperformed prior architectures applied to the Okutama dataset, which varied by their object identification and classification pipeline. The significant disadvantage of this method is that the performance of this algorithm may suffer if the gimbal angle approaches 90 degrees or if the drone flies higher than 30 meters. Additionally, due to motion blurring, the drone's speed may potentially have a negative impact on the performance of the system.[27]

2.3 Sound Detection

1. This research proposed a MEMS-based directional sound sensor that was modeled based on the hearing system of the *Ormia ochracea* fly, whose ears are separated by just 500 micro meters but have truly amazing sensitivity to the position of sound even at wavelengths two times of magnitude greater than the size of the fly's hearing structure. Since the amplitude of the rocking mode is dependent on the small pressure difference between the two wings, the sensor response has two resonant modes (rocking and bending), where both had 2.79 kHz and 5.3 kHz resonance frequencies, respectively, and a robust response was only obtained at the bending frequency. To reduce the impact of sound on the packaging, the device was fixed on an open-backed dual in-line package (DIP) socket. The mechanical resonant frequency was determined for a set of aspect ratios using a Polytec OFV-534 laser vibrometer operated by a sweeping sine wave from 1 to 10 kHz. To conclude, this research enables us to compare several frequencies to improve the accuracy of locating wideband sources' directions or to add the capability of differentiating between two particular sources.[10]

2. In this paper, they suggested using the transformer architecture with self-attention instead of time-delay neural networks (TDNN) and long short-term memory (LSTM). This study demonstrated that deep Transformer networks with high learning capabilities are able to match or even outperform traditional hybrid systems. Its capacity for regularization was the crucial factor in achieving the most cutting-edge outcome among end-to-end ASR models for the common 300h Switchboard (SWB) benchmark. Across the encoder and decoder, 48 Transformer layers are used to attain this outcome. On the Switchboard benchmark, the generated models surpass all prior end-to-end ASR methods. On the Switchboard and CallHome test sets, an ensemble of these models achieves WER of 9.9 respectively.[3]

2.4 Disaster Detection

1. This study "Characterizing Human Explanation Strategies to Inform the Design of Explainable AI for Building Damage Assessment" presents the preliminary findings on how humans understand satellite images to evaluate building damage from

natural calamities. The study used satellite pictures from the xBD dataset [3] to compile participant explanations of building damage assessment. Satellite images from xBD depict building damage from a variety of natural disasters. They initially created a new, web-based annotation system to gather human explanations for damage evaluations. Following the recruitment 60 of participants, each one took part in a training session before entering the study’s main session. Finally, they developed categories among the participant-generated explanations using an iterative, open-coding approach. A total of 929 codes were found based on the responses of 60 participants (an average of 1.55 codes per pre/post image combination, $SD = 1.05$). In contrast, 12.8% of the assessments ($N = 77$) had no code assigned to them since the participants had not offered an explanation in these situations.[26]

2. The purpose of the research “Satellite radar and optical remote sensing for earthquake damage detection: results from different case studies.” is to examine how well remote sensing methods can identify damage in urban settings and to investigate how radar (SAR) and optical satellite data can be used in combination. For the purpose of creating a trustworthy approach to damage assessment, they have focused on the benefits and drawbacks of remote sensing techniques. Both pixel-by-pixel categorization and damage assessment within homogeneous extended areas have been effectively evaluated using automated approaches. The techniques used for the test case of Bam earthquake incident produced results of approximately Accurate categorization rates from SAR were 61%, optical data were 70%, and data fusion was 76%. It has been documented how data from ground surveys and satellite remote sensing correlate.[2]

3. This study ”A Rapid Self-Supervised Deep-Learning-Based Method for Post-Earthquake Damage Detection Using UAV Data(Case Study: Sarpol-e Zahab, Iran)” suggests a novel deep-learning-based technique for quickly identifying building damage after an earthquake. The process consists of three processes and can detect problems on four different levels. To choose the best feature extraction technique, three different feature types—non-deep, deep, and fusion—are first explored. Deep features are separated from non-deep features using a ”one-epoch convolutional autoencoder (OECAE)”. The following phase involves designing a rule-based approach for automatically choosing the appropriate training samples needed by the classification algorithms. Last but not least, to determine building damage maps, seven well-known machine learning (ML) algorithms are used. With improved overall accuracy (OA) and kappa coefficient (KC) over 22% and 33%, respectively, the results showed that auto-training samples are practical and superior to manual ones. SVM was the most accurate AI model, slightly outperforming MLP (OA = 82% and KC = 73.98%). It was also discovered that, compared to methods employing either deep features alone or non-deep features alone, fusing deep and non-deep data using OECAE might greatly increase damage-mapping efficiency.[37]

4. In the study ”Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks”, automate the finding of building damage in satellite images using machine learning. In this study, the generalizability of convolutional neural networks (CNNs) for spotting disaster-related building damage is investigated. Prior to comparing the performance of the best CNN architecture after being trained and

validated in various transfer learning settings, they compare the performance of various CNN architectures on a single dataset. Additionally, using a dataset from the Haiti disaster, the research evaluates the performance of four distinct CNN architectures, looks into the generalizability of models, and gives the findings of an experiment on cross-region transfer learning.[18]

5. The paper “Earthquake-Induced Building Damage Mapping Using Explainable AI” focuses on disaster mapping and detection. MLP and SHAP algorithms can be used to detect these. The images of the affected region are taken from satellite images and open street maps (OSM), and the MLP algorithm can extract features from the images, and the SHAP algorithm can rank the most important extracted features of the images to detect collapsed or non-collapsed buildings.[25]

6. In this paper “Deep Learning and Stereo Vision Based Detection of Post-Earthquake Fire Geolocation for Smart Cities within the Scope of Disaster Management: İstanbul Case”, a framework based on stereo and YOLO that allows for quick and highly accurate post-earthquake fire detection is provided. Based on YOLOv5s design, the program uses cameras located in various areas to detect fires. When a fire is spotted by many cameras, its geolocation is determined by using stereo vision algorithms to assess the fire’s visual properties. By sending the geolocation data to the emergency management center via WSN, it is made sure that the fire is attended to as quickly as feasible.[33]

2.5 Mapping

1. This research was carried out by using Inspire 2 quadcopter drones equipped with RGB cameras, creating 3D models using photogrammetry, and geographic data systems for scenario mapping which is provided by a high-quality camera with dreadslocks for picture stability. The obtained results were 98.53% (98.68%), 95.2% (96.1%), and 94.4% (94.7%) for each altitude of 40, 80, and 100 m using Google earth data at two different places as a baseline for the accuracy of measuring of the area at three variations of flying height when taking images. The biggest challenge of mapping with very high-resolution satellite data is image collecting, which remains quite expensive, particularly for mapping in a localized area that requires more regular (daily or weekly) time series of information repetitions.[24]

2. This research provided a method for real-time 3D mapping that combines a 2D laser scanner with posture estimate through visual-inertial fusion. The necessary pose estimation is provided using IMU-aided visual SLAM based on EKF. A worldwide, consistent 3D map is created in real-time using 2D laser scans taken while a robot is in motion. Using a Turtlebot equipped with a Kinect, an IMU, and a laser scanner to create 3D point cloud maps of multiple offices, they tested the idea described in this study.[9]

3. In this paper, a different data format for online reconstruction using widely accessible consumer depth cameras - was proposed. The method uses hierarchical data structures and implicit surfaces to combine them for reconstruction, but it does so

with a little amount of overhead due to a compact spatial hashing mechanism. The hashing algorithm provides real-time operation without compromising scale or finer reconstruction quality. The hardware for parallel graphics has been developed to make all processes efficient. Our approach's intrinsic unstructuredness eliminates the overhead of traditional spatial data structures while preserving the essential elements of volumetric fusion. This method supports lightweight streaming without extensive data structure reconfiguration, which might extend the limits of reconstruction.[6]

4. This study describes a simple and inexpensive method for mapping interior structures in three dimensions (3D) using Light Detection and Ranging (LDR) (LIDAR). LIDAR is managed such that it may measure the angels and distance from both servo motors on which it is attached at the same time. These numbers help in computing and creating a 3D illustration of the internal structure. The use of photogrammetry for 3D mapping is extremely complex, time-consuming, and expensive. In some circumstances, the photogrammetry process' precision is unacceptable. The visualization is improved by the highly realistic image that 3D mapping offers. It has numerous uses in the fields of engineering, surveying, and research.[13]

Chapter 3

Architecture

3.1 System Model

The purpose of our thesis is to design a completely accurate fire detection and rescue system. We intend to detect the fire more accurately by our multimodal approach and want to find the safe path for rescue by disaster mapping.

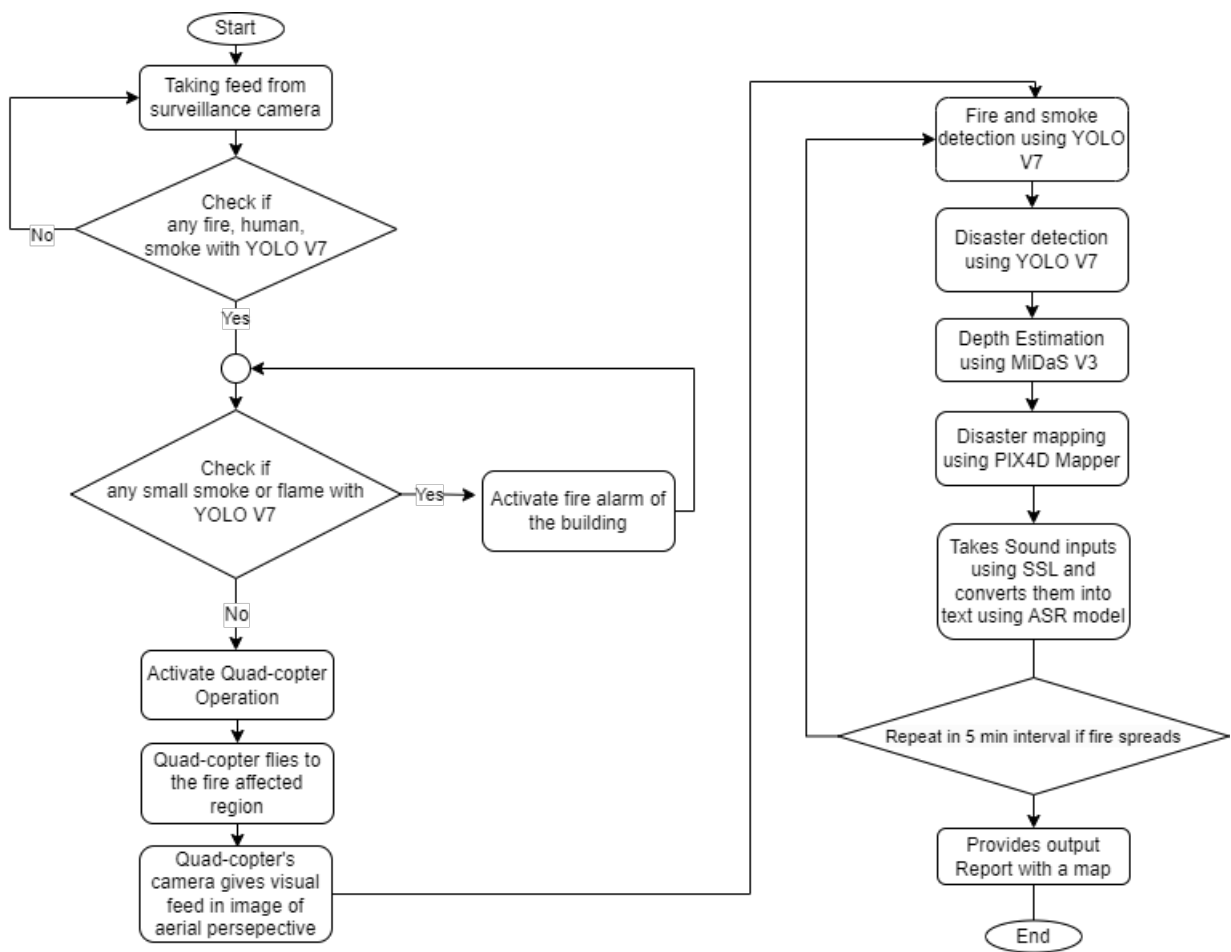


Figure 3.1: Model Of Our Multimodal Fire Detection System

Our thesis mainly focuses on the part of fire, human and disaster detection. Sound

detection and disaster mapping is a complementary part of our thesis.

Our system will continuously take feeds from the surveillance system or CCTV camera and monitor for small smoke or fire with YOLOV7. If it detects any small fire or smoke it will activate the fire alarm. But if the fire has spreaded already it will activate the quad-copter operation. The quad-copter will fly to the fire affected area and take image input from an aerial perspective. There this system will detect fire and smoke using YOLOV7. It will also detect any human inside or around the fire with YOLOV7. Then it will analyze the extent of the disaster with YOLOV7. After that it will take Bangla word input using Okkhor software and analyze the data using the ASR model.

At last it will generate a real time 3D map using MiDaS V3 and PIX4DMapper and provide the output map of that area. If the fire spreads, it will repeat the fire, human, disaster and sound detection method along with map generation again after 5 minutes.

The modes of our multimodal system are:

1. Image Processing
2. Sound Speech Recognition
3. Depth Estimation
4. 3D Mapping

‘The primary mode of our research is Image Processing using YOLO V5 and YOLO V7. Other modes of our multimodal system are complementary to fire and disaster detection system. We are trying to design a complete fire and disaster management system that can detect an emergency situation and provide the overall report of the situation to rescue human life or measure the causality.

3.2 Research Methodology

In this part of our paper, we will be briefly discussing the methods, models and approaches that have been used to conduct this research. The main goal of our research was to detect any fire and disasters as soon as possible and save as many lives as possible using UAV(unmanned aerial vehicle) or quadcopter. By utilizing luminous smoke and textured flame, surveillance cameras can continuously observe for a fire outbreak in its early stages. In contrast, an alert will ring and the quad-copter operation will be initiated if the fire has already begun. The quadcopter’s video and audio input allows it to map the area affected by the fire while also identifying human life from an aerial perspective. Last but not least, disaster detection gives us a map of the safe zone that can be used later by the fire service to save lives by using the less damaged area of the building. To reach our goal we divided this research into five parts which are:

- **Fire,smoke and human detection**
- **Disaster detection**
- **Sound Detection**

- **Depth Estimation**
- **Mapping**

As we could not find any datasets for fire, human and disaster similar to the customs and environment of Bangladesh, we collected primary datasets for these parts. The dataset used for sound detection is custom. For fire and smoke detection, we collected images of various fire incidents of Bangladesh like Sitakundu fire incident, Tazreen garments fire incidents and many more. We collected available images on google and facebook as well as took screen shots from videos, news and facebook live videos. We have excluded wildfire images from our dataset.[28] For human detection we again collected images from google and facebook of single humans as well as human gatherings on various occasion+ir and many more. Also, for the detection of disaster, we again collected images of building collapse, earthquakes and many more disastrous incidents in Bangladesh as well as other countries similar to the building pattern of our country. Then, we collected the sound dataset from the available public datasets in Keras website. Last but not least, for mapping we used MiDaS V3 which was pre trained by the COCO dataset which has proven to be the state of art benchmark.

[19]

Moreover, we annotated all the images of fire, humans and disasters using a website named Roboflow.[7] For fire and human detection we classified the pictures into five categories. They are: smoke,Fire,no fire,human, and Gathering.Here we added no fire to reduce false fires. And for disaster detection, we classified the images into three categories. Which are:level one, level two, and level three. Where level one detects unaffected buildings or some scratches on buildings, level two detects slight damages on buildings and level three detects completely destroyed buildings or the debris. All the images were trained using YOLOV5 at first. After that we switched to YOLOV7 which is an updated version of YOLOV5 that can process frames in real time and gives us better accuracy. We trained the YOLOV5 in the free version of google colab with Python. But as it was not the pro version we were getting some errors and could not train them in higher epoch. As a result, we trained the YOLOV7 model in the system terminal of the pc later on. And for sound detection we used the Okkhor software that translates audible words into Bangla texts.We trained the collected the custom dataset with ASR model in Google Colab environment free version with Python.

While implementing, we extracted real time sound input from the environment using SSL from Okkhor application and implemented it with the ASR model. Lastly for mapping, we have used MiDaS V3 which is an improved 3D method of depth estimation. It uses a Computer Vision segment to generate depth estimation from a single RGB picture in real time. So there was no need to train any dataset separately for this.we can run this model with the help of PyTorch and in Google Colab environment.[21]It gave the estimated depth information from a single RGB image from the aerial view using real time, precise as well as a lightweight approach.[11]After that, we used the extracted aerial images and created high quality maps and models of the disaster and fire prone area with PIX4D Mapper which is a third party application that is capable of creating high-resolution orthomosaics and digital surface

models. We have used the free trial version of PIX4DMapper to generate the map.

3.3 Working Plan

The work plan of our thesis covers the 9 months period from June 1st, 2022 to March 1st, 2023, and specifies the intended actions and outcomes for each thesis component. Our thesis is divided into five components: fire detection, human detection, disaster detection, sound detection, and 3D mapping of disastrous areas. In our working process of thesis, we have faced many challenges like, not getting good quality images for dataset, not being able to train the model with high accuracy, doing tedious work like annotating again and again, but later gradually we overcame the challenges and roughly we managed to get fruitful results. We have some limitations on our thesis, and we hope our future generation will look into our thesis and work with the limitations. The working plan and process of every field of the thesis is mentioned below:

Detecting the fire to prevent the major damages is the main part of the thesis. Pictures of fire and smoke will be taken through surveillance cameras during the fire breakout, and those will be detected through YOLOv7 model. At first, we have trained YOLOv5 model to detect fire and smoke, but later we have shifted to YOLOv7 which is the better version of YOLOv5 to get the better accuracy. For training the model, we tried to use custom datasets during our pre thesis 1, but those datasets were foreign and they didnt match with the building structures and environment with our country Bangladesh. Thats why we have created primary dataset by collecting pictures from google, facebook, screenshots from news, youtube etc. The pictures were annotated in the website called roboflow, and classified with fire, no-fire and smoke classifications. Fire and smoke labels were used to indicate fire and smokes, while no-fire label indicates any red or orange colored objects that can misclassify as fire. This thesis only researched for only detecting local fire, and wild fires excluded, which is the limitation of our work. On the other hand, it also detects on only open spaces, not in heavily smoked closed spaces.

[45]

Detecting human alongside the fire is another main part of the thesis that can detect any presence of human or victims early in the disasterous places to save them. Here we collected primary datasets of human and gathering pictures from google, social medias and also we have taken pictures on our own of different real life occasion. We have trained YOLOv5 at first, then YOLOv7 model to detect single human and gatherings of the disastrous area by annotation them in roboflow website. At first the accuracy of the model was showing very low, since we have annotated every single human of gatherings as human, later we have annotated all the pictures again by classifying whole gathering as gatherings, and collect more single human pictures to classify as human. In the fire human dataset, overall accuracy is 80.32% in 5 classes. Overall accuracy for the fire and human dataset is 80.32% across 5 classes. Accuracy rates for each class are, fire 92.6%, 72.8% for gathering, 48.9% for humans, 91.9% for no fire, and 95.4% for smoke.

We are also detecting disaster, which means detecting building collapses and dam-

ages during different calamities like earthquakes, cyclons, explosions etc. It gives us a map of the area that is safe, where the less-damaged portion of the structure will help the fire department save lives. Like fire and human, we also used primary datasets for disaster detection by collecting pictures from google, youtube and screenshots which are similar to the building pattern of our country. We also trained YOLOv5 then, now YOLOv7, and classified the labelings in three categories: level-1, level-2, and level-3 which were annotated in roboflow. Here, level-1 indicates no building damage buildings with mild scratches, level-2 indicates partial and total building damage, and level-3 indicates debris/rubble. Our model can only detect damage on brick and concrete structures, so it is not able to detect disasters on leaning buildings, which is one of our disaster detection limitations. Moreover, our approach might not be effective in spotting disasters in locations with weak structures like bamboo, wood, and tin roofs. The disaster dataset consists of 9.5k images. After annotating all the pictures three times, we still couldn't increase the accuracy of the model which is only 9.2%, and we are still trying to find out how to increase it, which is a big challenge for us.

[43]

Apart from fire, human and disaster detection, we are working on sounds to detect distressed calls and disastrous sounds during the incident. We are using customized sound datasets which were taken from keras website. We utilized the Okkhor software, which converts spoken words into texts in Bangla. [46] In the free Google Colab environment using Python, we trained the ASR model using the custom dataset that was collected, and it showed 98% accuracy. During implementation, we used SSL from the Okkhor program to extract real-time sound input from the surroundings and integrated it with the ASR model. At first we have used MEMS sensor instead of SSL, then we have shifted to SSL since MEMS directional sound sensor only provides one channel of audio data, a sound source localization system provides numerous channels of audio data.

[8]

3D mapping of disastrous area will generate real time 3D map in order to find out the safest path and help people escape the disastrous area and reaching help by fire department quickly. We used MiDaS V3 for mapping, which was pre-trained using the COCO dataset. MiDaS V3 is a model which is more advanced 3D depth estimation technique. It uses a real-time, accurate, and lightweight approach to provide the estimated depth information from a single RGB photograph of the aerial view, and it does not need any dataset training. At first, we wanted to use LIDAR depth mapping which is more accurate than MiDas V3, we MiDas V3 since it was very expensive and difficult to install. Then, using PIX4D Mapper, a third-party program that can produce high-resolution orthomosaics and digital surface models, we used the extracted aerial photos to construct high quality maps and models of the catastrophe and fire-prone area. To create the map, we used PIX4DMapper's free trial version. The quality of the input data determines how accurate the mapping model will be, and it can be influenced by things like the image resolution, the precision of the GPS data, and the existence of barriers.

Last but not the least, we have combined all our thesis components in a multimodal interface called GUI. It combines our used models and softwares to provide an all-inclusive disaster and fire event management solution. Users can detect fire, human,

disaster, sound and map the disasterous area by uploading image or videos in GUI interface. In order to deploy the output of our models, we used a Flask API. Afterwards, we incorporated this API into a graphical user interface (GUI) that was made using the Python Tkinter framework.

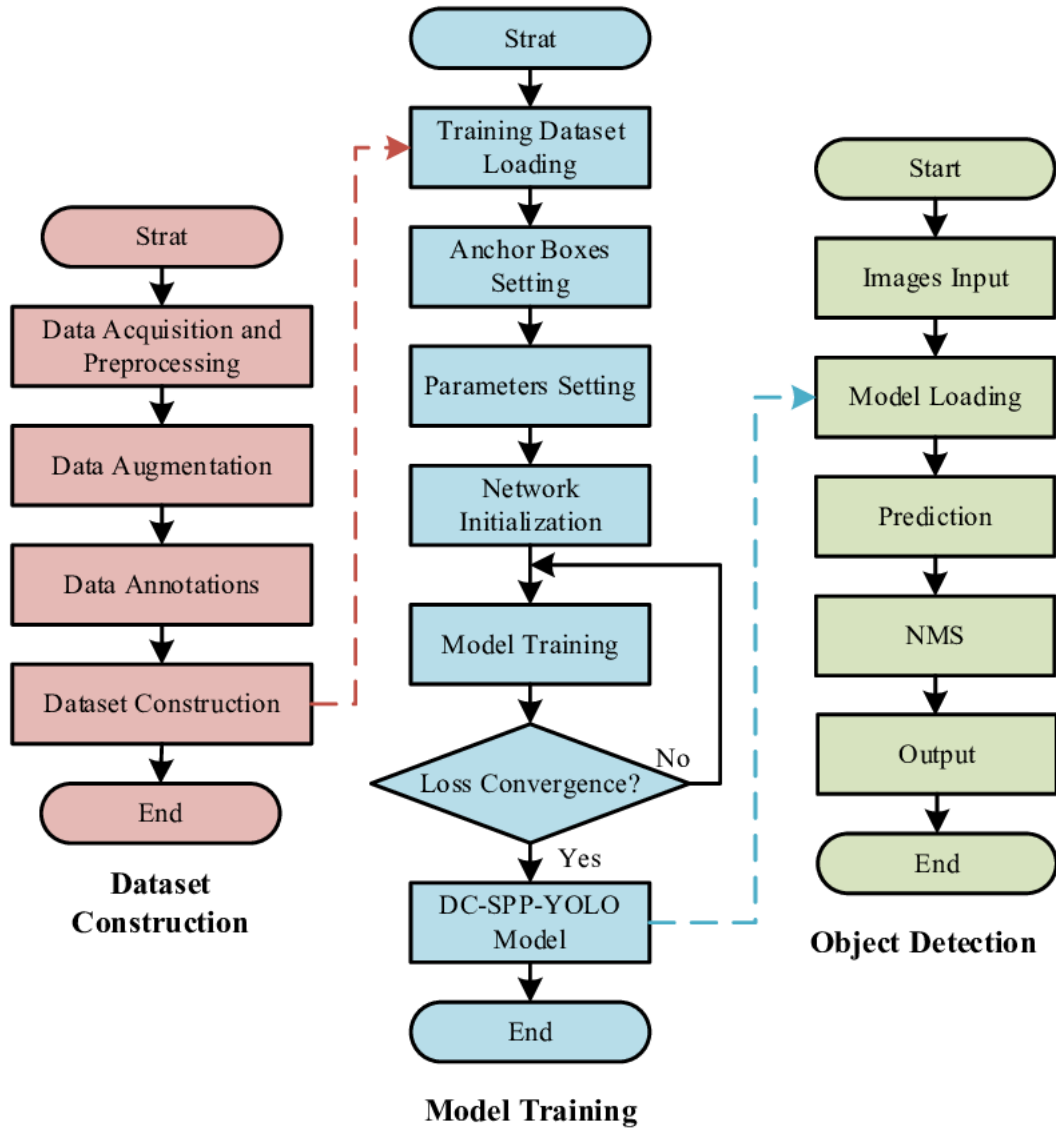


Figure 3.2: Working Plan

In real time, our system will continually take feeds from the security system or CCTV camera and watch for little smoke or fire using YOLOV7. It will trigger the fire alarm if it senses a minor fire or smoke. But, if the fire has already spread, the quad-copter operation will be activated. The quadcopter will travel to the fire-damaged region and collect aerial visual input. Using YOLOV7, this system can identify fire and smoke. With YOLOV7, it will also identify any individual inside or near the fire. Then, using YOLOV7, it will evaluate the severity of the disaster. Next, using the Okkhor program, it will accept bangla word input and analyze the data using the ASR model and try to find distress call. Finally, it will generate a

real-time 3D map of that region using MiDaS V3 and PIX4D Mapper and produce a map. If the fire spreads, the fire, human, disaster, and sound detection methods, as well as map production, will be repeated after 5 minutes.

Chapter 4

AI Models related to our work

4.1 Description of the models

4.1.1 YOLOV5

YOLOV5 is one of the popular object detection system created by Joseph Redmon and Ali Farhadi, YOLO (You Only Look Once). It is one of the most precise object detection systems currently available and is utilised in a variety of applications, including autonomous driving and security surveillance. Until the release of YOLOV5, YOLOV4 and YOLO V3 were the most accurate versions of YOLO. YOLO V5 is an improvement over its predecessors. YOLO V5 provides greater precision, speed, and scalability than previous versions. Combining recent advances in convolutional neural networks, its EfficientDet model architecture is new. In addition, it employs a novel data augmentation technique known as AugMix, which helps to improve precision and reduce overfitting.

In addition, YOLO V5 has an enhanced loss function, which enables it to converge faster and achieve higher levels of accuracy. The YOLO V5 object detection system is a great addition to the YOLO family and is used in a variety of applications. The YOLO V5 model for object detection is based on Convolutional Neural Networks (CNNs). It employs transfer learning to improve accuracy and speed, as well as to reduce the amount of training data required. Additionally, it is optimised for the COCO dataset, which includes eighty object categories. The model is faster, more accurate, and has fewer parameters than its predecessors.

As previously stated, YOLO V5 employs a model architecture known as Efficient-Det, which is designed to be both efficient and accurate. It employs convolutional layers and depthwise separable convolutions to reduce the number of parameters and increase the model's efficiency. It also employs a new anchor-free feature pyramid network (FPN) to detect objects of various sizes more accurately. The YOLO V5 model is quicker and more precise than its predecessors because it is designed to be more efficient and to require fewer training parameters. The optimization of the model for the COCO dataset, which contains eighty object categories, makes it suitable for numerous applications.

YOLO V5 uses a variety of object detection algorithms, including:

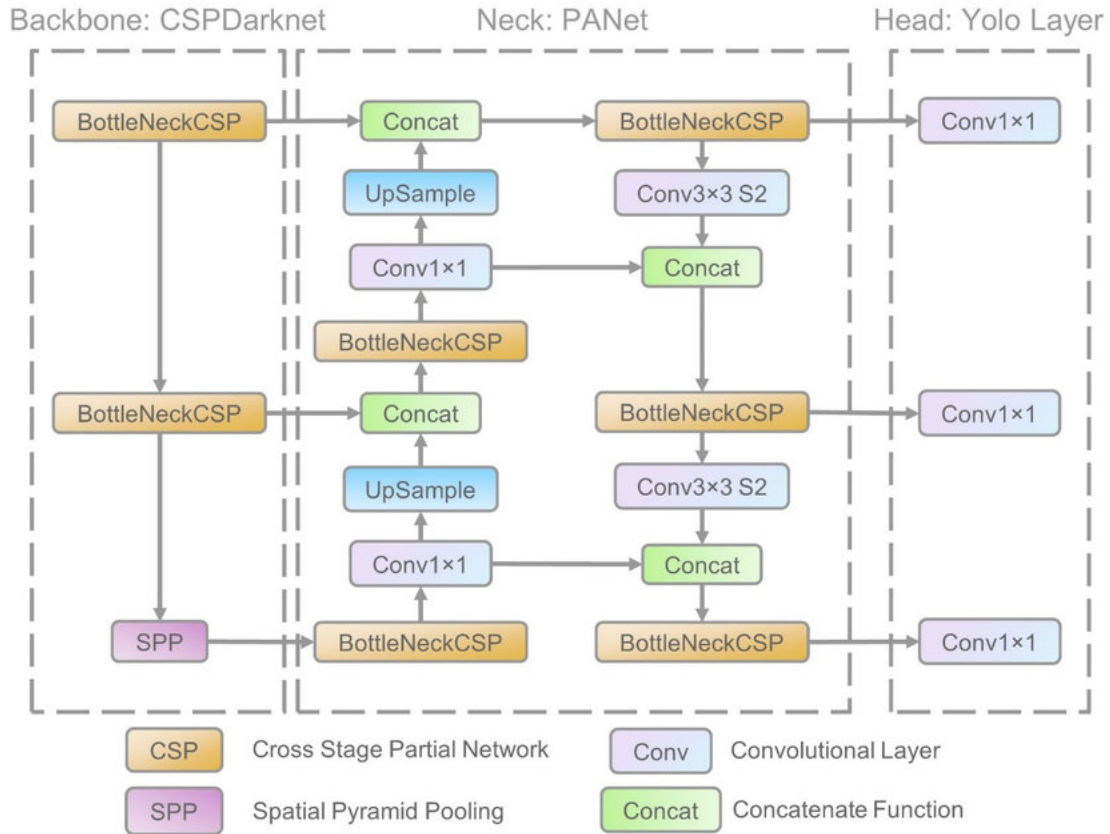


Figure 4.1: YOLOV5 Architecture

1. Darknet-53 is a convolutional neural network (CNN) architecture that is based on ResNet-50. It is used to extract features from images.
2. Cross-Stage Partial Network (CSP): This is a feature extraction and object localization algorithm. It detects objects in feature maps.
3. Mosaic Augmentation: This is an augmentation technique used to increase the dataset's diversity. It entails randomly dividing the images into smaller tiles and then recombining them in various ways.
4. Path Aggregation Network (PAN) is an algorithm that combines multiple feature maps to generate more accurate predictions.
5. Feature Pyramid Network (FPN): This network architecture combines features from various CNN layers. Combining features from multiple layers aids in the generation of more precise predictions.

We have used various public datasets with the YOLOV5 model. There is one dataset for fire and smoke and another for disasters.

4.1.2 YOLOV7

Joseph Redmon and Ali Farhadi created the cutting-edge, real-time object detection system known as YOLO (You Only Look Once). The YOLO system's most recent iteration, YOLO V7, was unveiled in 2020. It is a more rapid and accurate version

of YOLO V7 than earlier iterations. YOLO V7 is especially helpful for applications like robotics, autonomous driving, and video surveillance because it can simultaneously recognise several objects in an image or video. In YOLO V7, the bounding box coordinates of objects and their class probabilities are predicted using a single neural network. The region of an object in the image is specified by the bounding box coordinates, and the object is identified using the class probabilities. Moreover, YOLO V7 has the ability to categorise items at various scales, which enables it to recognise minuscule objects in an image. Because YOLO V7 is speed-optimized, it can process frames in real time. The most recent version of the algorithm, YOLO V7, has seen a lot of advancements over earlier iterations.

The foundation of YOLO V7 is a deep convolutional neural network that has been trained to identify objects in images. The programme employs a multi-scale detection technique, allowing it to recognise items of various sizes inside an image. It can analyse real-time data since it employs a single neural network to forecast the bounding boxes and class probabilities for each object in an image.

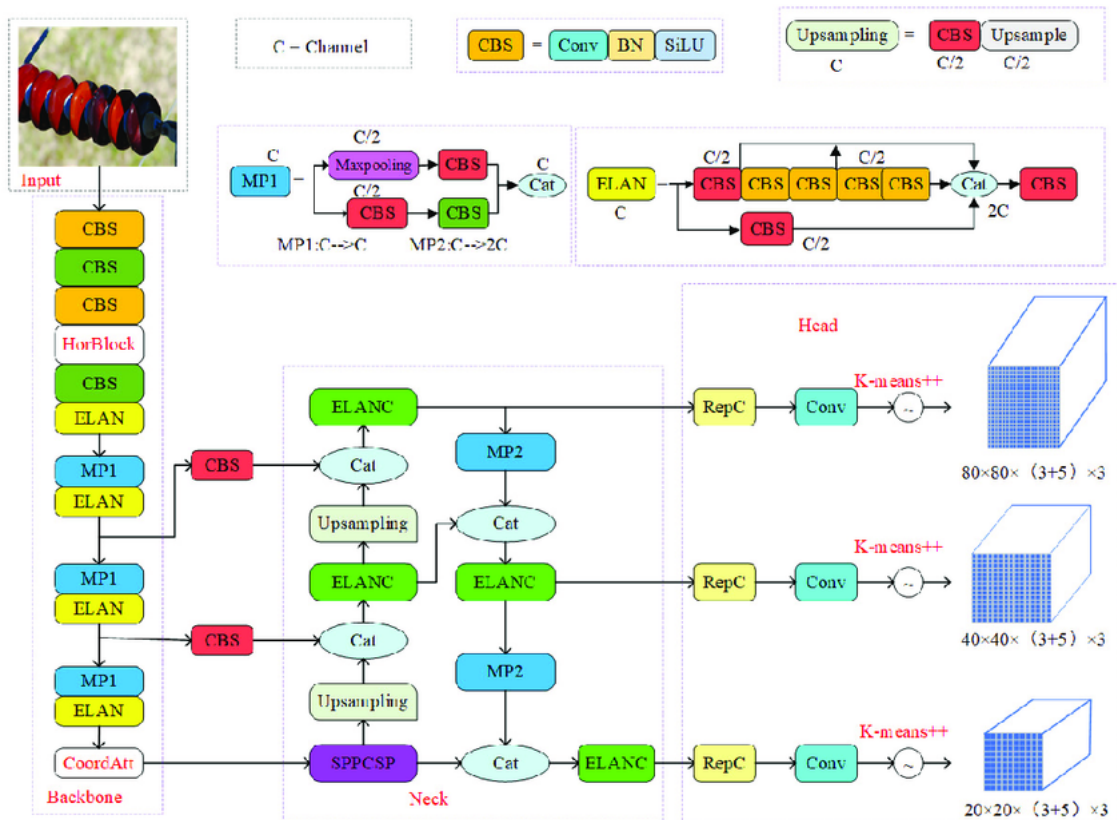
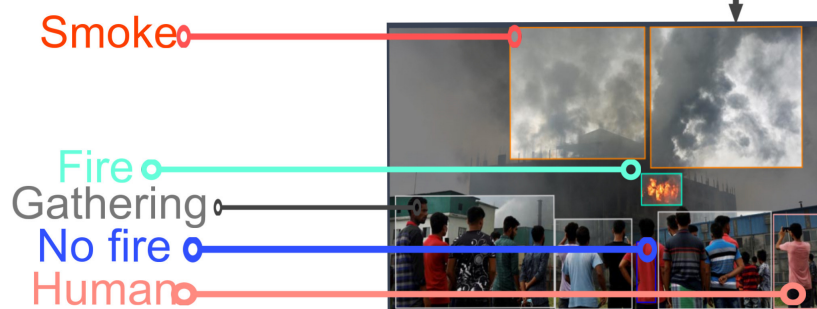
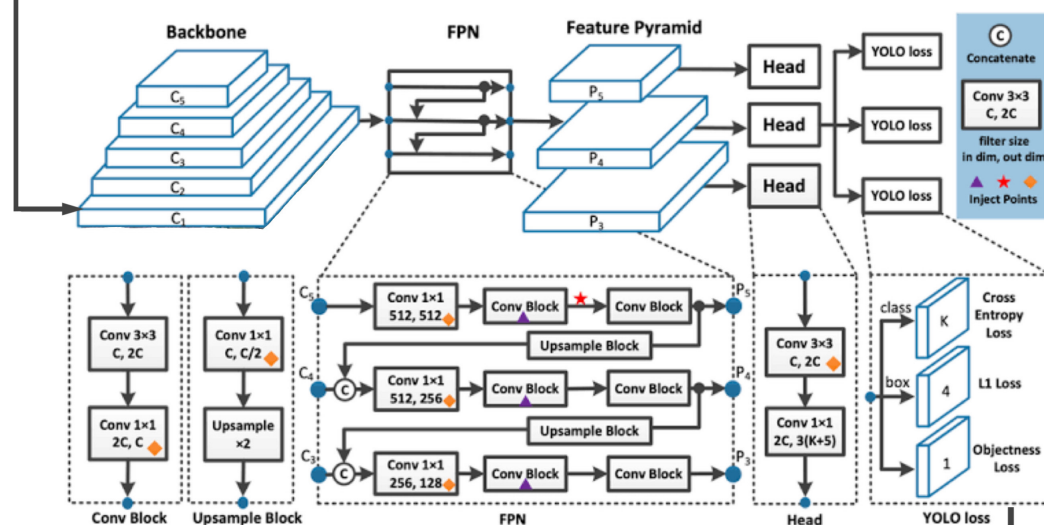


Figure 4.2: YOLOV7 Architecture

One of the best and most effective object detection algorithms is YOLO V7. It can recognise objects from a wide range of angles, sizes, and lighting situations and has a high accuracy rate. Moreover, it has the ability to find small things and partially occluded objects. Robotics, security systems, surveillance, and autonomous cars are just a few of the many applications that YOLO V7 is used in. The precision and effectiveness of this method have made it the most widely used one for object

detection.

Input image



Output image

Figure 4.3: YOLOV7 Architecture for fire dataset

It functions by using a convolutional neural network to process an input image (CNN). The CNN then generates a collection of bounding boxes around the important portions of the image, together with class labels for each box. When bounding boxes overlap, the model employs a non-maximum suppression procedure to remove them. Ultimately, the model may make inferences about the image and forecast the object's class. In order to find objects in photos and videos, YOLO V7 employs a variety of algorithms. The Darknet-53 convolutional neural network, a deep learning

architecture, is the primary algorithm used. It is followed by a number of different algorithms, such as Intersection over Union, Anchor Boxes, and Non-Maximum Suppression. It also makes use of a categorization technique known as Softmax. Assigning probability to the discovered items is made easier by this.

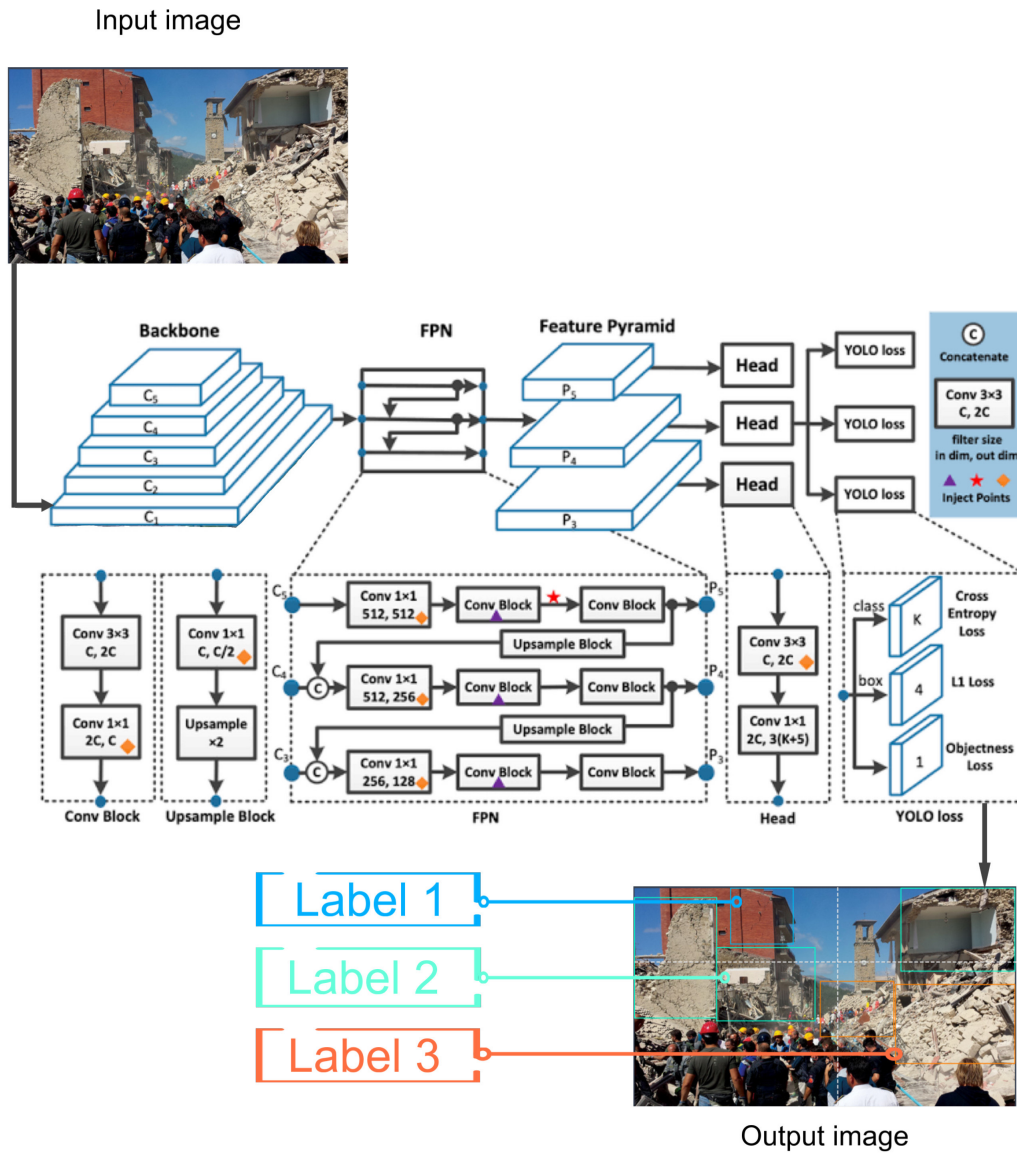


Figure 4.4: YOLOV7 Architecture for disaster dataset

On this model, we employed a fire, smoke, human1, gathering and no-fire dataset with 100 epoch, and the accuracy was 80.32%. Again, for the disaster detection we got an accuracy of 9.2%.

4.1.3 MiDaS V3

A deep learning-based monocular depth estimation model called MiDaS V3 was created by the University of Cambridge’s Perception and Robotics Group. A single RGB image can be used to estimate depth information using this real-time, precise, and lightweight approach. The model uses a key-value memory module to increase performance and is based on a multi-scale encoder-decoder design. MiDaS V3 can produce cutting-edge outcomes on well-known depth estimate benchmarks including the KITTI and NYU-Depth V2 datasets. It is also one of the current fastest models, making it appropriate for real-time applications. A cutting-edge real-time depth estimation algorithm is MiDaS V3. It is a fully convolutional neural network (FCNN) with a cutting-edge encoder-decoder structure that can handle a variety of input modalities. The method employs a single-shot architecture, making it possible for it to instantly create a depth map from a single image. It is capable of delivering extremely accurate results in a variety of lighting and environmental circumstances and was trained on a variety of datasets, including the KITTI benchmark. An up to 1024×1024 pixel resolution depth map can also be created using the technique. MiDaS V3 is appropriate for applications like robots, autonomous cars, and augmented reality since it can deliver highly precise findings in real-time.

Deep learning algorithms are used by the computer vision system MiDaS V3 to recognise and examine objects in an image or video. Convolutional neural networks (CNN) are used to recognise and categorise the items in an image. MiDaS V3 can be used for navigation, object tracking, and augmented reality applications because it can identify things in real-time. It can also be used to find and examine objects in recordings, such finding persons in a security camera footage. MiDaS V3 can find objects in a range of settings, such as dim light or against different backgrounds. Convolutional and recurrent neural networks are used in conjunction by the MiDaS V3 model to conduct semantic segmentation. For example, the MiDaS V3 model uses the BiLSTM architecture for segmentation and the ResNet-101 design for feature extraction. For temporal consistency, the BiLSTM architecture is used, which enhances the capturing of pixel-level long-range relationships.

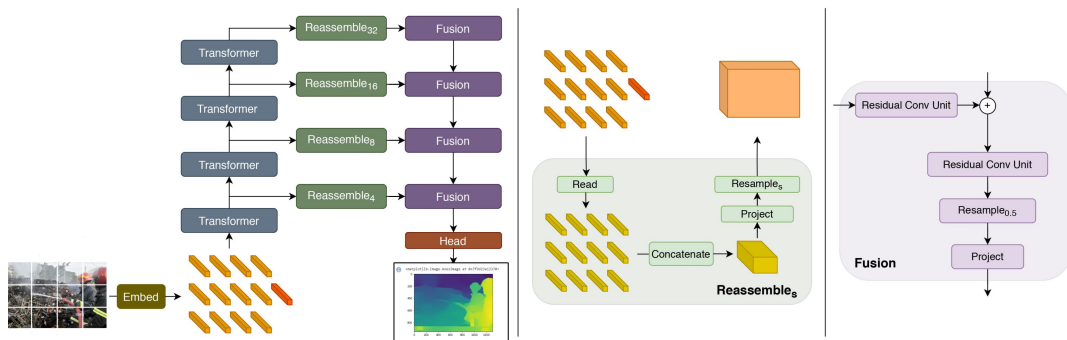


Figure 4.5: Architecture of MiDaSV3

MiDaS V3 has an improved 3D depth estimate method, making it superior than earlier versions of MiDaS. In order to more properly collect depth and distance information, this method leverages machine learning to more accurately forecast

depth from a single image. This increases the dependability and accuracy of MiDaS V3, enabling a wider range of applications. Furthermore, MiDaS V3 is more widely available to a wider number of users because it can operate on both CPUs and GPUs.

4.1.4 SSL (Sound Source Localization)

The process of locating a sound source in a space using the acoustic information it contains is known as sound source localization. In many species, including humans, it is a basic ability of the auditory system. It is also employed in a variety of industries, including robotics, audio engineering, and speech processing.

[5]

A microphone array and a processing module make up the two primary parts of a sound source localization system. The processing module receives the electrical signals that the microphone array transforms from the acoustic signals from the sound source. To determine the location of the sound source, the processing module employs a number of signal processing techniques, including beamforming and time-difference-of-arrival (TDOA) estimation.

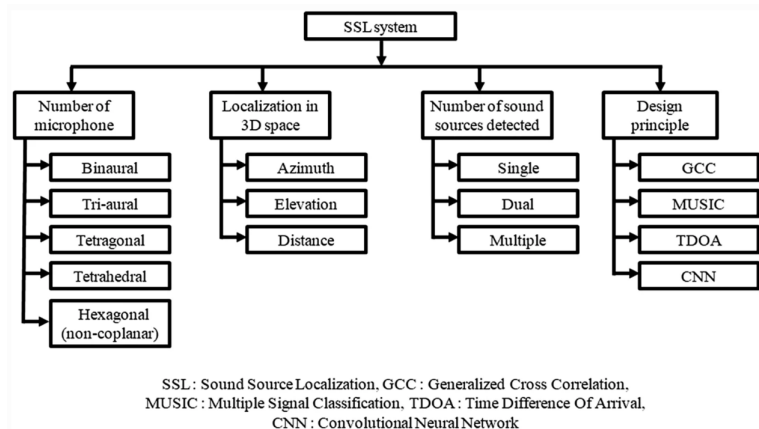


Figure 4.6: Sound Source Localization

Many algorithms are used by sound source localization systems to determine the location of a sound source. Many typical algorithms found in these systems include:

1. Time-Difference-of-Arrival (TDOA) (TDOA) By monitoring the time difference between the sound signal's arrival at various microphones, this approach estimates the location of a sound source.
2. Using a variety of microphones, the signal processing method known as beamforming focuses on one direction while excluding noise coming from other directions. It creates a beam in the direction of the sound source by combining the signals from several microphones, increasing the signal-to-noise ratio and improving source localization accuracy.
3. Frequency-Difference-of-Arrival (FDOA) Estimation: This technique uses the difference in frequency between signals received at various microphones to infer the location of a sound source.

4. Time-Delay-of-Arrival (TDOA) with Particle Filtering: This technique determines the location of the sound source by combining TDOA estimation with particle filtering, a Monte Carlo-based approach.
5. The Maximum Likelihood Estimation (MLE) technique locates the location of a sound source by maximising the likelihood function, a function that expresses the likelihood of the observed signals given the source's position.

Using a mathematical model and the observed signals and process noise, the Kalman filtering state estimation algorithm can infer the location of a sound source. To offer precise and trustworthy sound source localisation, these techniques can be utilised singly or in combination. The particular needs of the application and the attributes of the environment determine the algorithm to use. Sound tracking in robotics, speaker diarization in speech processing, and sound monitoring in audio engineering are just a few of the many uses for sound source localization systems. They also provide enhanced audio capabilities in a variety of industries, including automotive, security, and entertainment.

In summary, SSL Sound Source Localization systems are essential for a variety of applications because they make it possible to precisely pinpoint the location of a sound source in a certain setting.

4.1.5 ASR speech recognition

Technology known as Automated Speech Recognition (ASR) allows computers to understand human speech and convert it into text. It finds utility in numerous contexts, such as search engines, personal assistants, automatic systems, and NLP systems that process spoken commands. Algorithms in ASR systems decipher a person's speech and turn it into text. This technology allows computers to understand human speech, thus it may be used to make interactions more natural and conversational. Moreover, ASR is utilised in applications like text-to-speech synthesis, machine translation, automated customer care, and speech-to-text transcription. The need for ASR is rising because of the popularity of NLP and voice-based interactions. With it, humans may have conversational interactions with computers and other devices using only language. Voice-activated personal assistants, automated customer service systems, voice recognition systems for security, and voice-enabled search engines are just few of the many uses for automatic speech recognition (ASR).

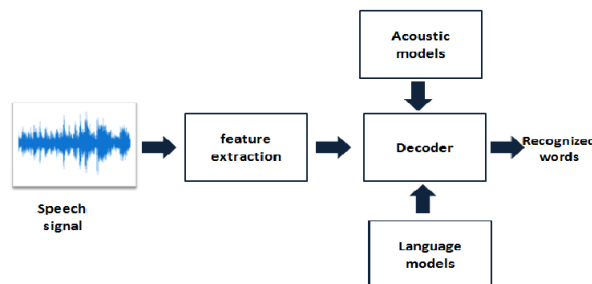


Figure 4.7: Automatic Sound Recognition

In order to interpret what a user is saying, ASR technology examines their voice in terms of its frequency and pattern. The system analyses the user’s speech for patterns and frequency, and then compares those to a dictionary. The system’s ability to effectively respond to the user depends on its ability to accurately detect the user’s words.

With the advancement of ASR technology, it is now possible to recognise a broad variety of accents and dialects. More so, ASR systems can be taught to detect particular words and phrases, hence improving their responsiveness to human input. More and more fields and uses are realising the benefits of ASR technology. It is possible that ASR technology may greatly increase accessibility for people with disabilities, in addition to providing a more efficient and accurate way to interface with machines. Phonemes, the building blocks of speech, are extracted and compared to a database of known sounds. The machine then use algorithms to pick out individual words and sentences from the audio file, which it subsequently translates into text. Assisted speech recognition (ASR) technology is fueled by machine learning, which allows it to get better as it is exposed to more speech data.

In this talk, I’ll be making use of the LJSpeech data collection that was amassed for the LibriVox project. It consists of short audio clips of seven different nonfiction works, each read by a different speaker.

4.2 Description of the softwares

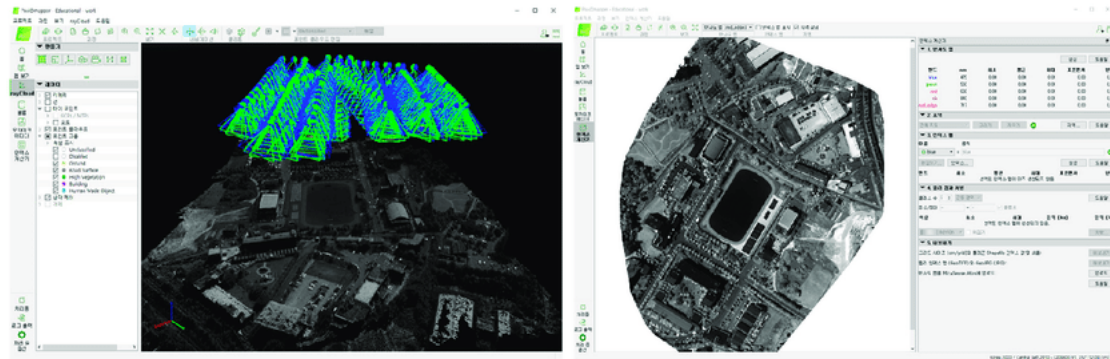
4.2.1 PIX4DMAPPER

Aerial photos taken by drones or other aerial platforms are processed using professional photogrammetry software called PIX4D Mapper to produce high-quality maps, models, and other photogrammetric outputs. Structure from Motion (SfM) and Multi-View Stereo (MVS) methods are used by the software to extract 3D information from 2D images. By comparing features in many 2D photos, the SfM method reconstructs a scene’s 3D geometry. It makes an educated guess on the camera angles, positions, and scene feature locations in each image. A dense point cloud of the scene is then produced using these positions. Following the creation of a 3D mesh from the point cloud using the MVS algorithm, orthomosaics, digital surface models, and other photogrammetric outputs can be produced.

[36]

With its many useful capabilities, PIX4D Mapper is a well-liked option for mapping and surveying applications. Its capacity to swiftly and effectively process huge datasets is one of its key advantages. The software may run on a typical desktop computer or laptop, and it can process hundreds or even thousands of photographs at once. Additionally, it offers an intuitive user interface that enables users to quickly import and manage their picture datasets, change processing options, and examine and export the outcomes. The capacity of PIX4D Mapper to produce extremely accurate and precise 3D models and other photogrammetric outputs is another important feature. Even when working with big and complicated scenes, the software makes use of sophisticated algorithms and approaches to guarantee that the models

are accurate to within a few millimetres. In order to guarantee that the outputs are correct and dependable, it also contains a variety of quality control tools and validation techniques.



(a) Merging of UAV images

(b) Index map of multispectral image

Figure 4.8: PIX4DMapper

In order to construct maps of disaster and fire-prone locations for our thesis, we used PIX4D Mapper. Drones were used to take aerial photos of the locations, and PIX4D Mapper was used to analyse the photos and create high-resolution orthomosaics and digital surface models. The magnitude of the tragedy and fire-affected areas could then be determined and mapped using these maps. The success of our thesis project was greatly influenced by the use of PIX4D Mapper. We were able to create high-quality maps and models of the disaster and fire-prone areas thanks to the software, which processed vast amounts of aerial imagery rapidly and precisely. As a result, we were able to pinpoint the locations that posed the greatest danger and take the necessary steps to reduce it. Our findings were in line with earlier research that applied mapping and surveying techniques with PIX4D Mapper. For instance, PIX4D Mapper was used in a study by Diakité et al. (2018) to create high-resolution maps of rice fields in West Africa, and it was discovered that the programme was able to create trustworthy and accurate maps of the fields. Similar to this, Liu et al. (2019) discovered that the software could precisely identify and map the extent of landslide-prone locations in China by using PIX4D Mapper to create maps of these places.

A popular photogrammetry programme for mapping and surveying tasks is PIX4D Mapper. It is strong and adaptable. In our thesis, we used PIX4D Mapper to create high-quality maps and models of catastrophe and fire-prone areas and pinpoint the regions that were most vulnerable. We anticipate that PIX4D Mapper will remain a crucial tool for mapping and surveying applications in the future because our findings were in line with those of earlier studies that employed the programme for analogous purposes.

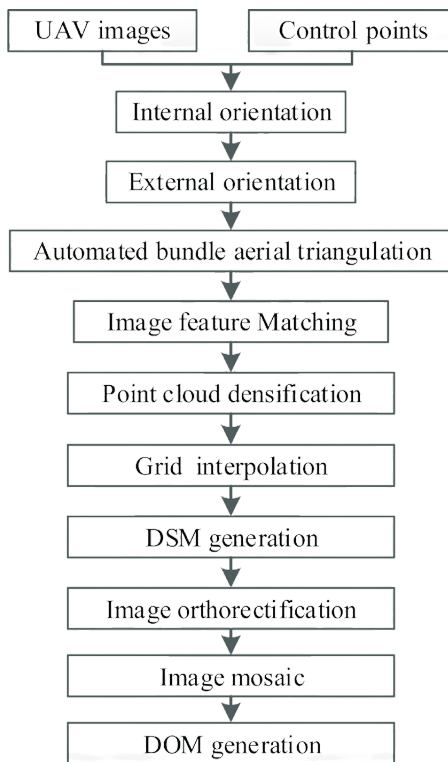


Figure 4.9: Workflow of PIX4D Mapper

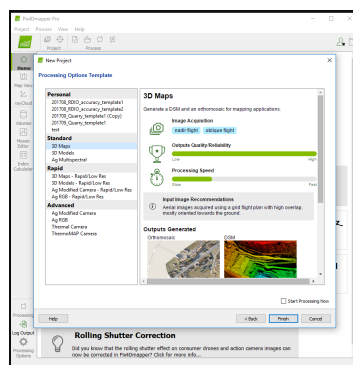


Figure 4.10: Performance of PIX4D

4.2.2 Okkhor

With the use of the voice typing keyboard Okkhor, users may type Bangla text. It is a creative approach to the issue of typing in Bangla, which might be difficult for those who are not used to the intricate script. The Bangladesh Association of Software and Information Services (BASIS) and the Bangladesh University of Engineering and Technology jointly created Okkhor (BUET).

[17]

To translate audible words into written Bangla text, Okkhor combines speech recognition and natural language processing (NLP) technologies. The NLP algorithm evaluates the words that have been detected by the deep neural network (DNN) used in the speech recognition algorithm to create the final written text. The DNN is able to distinguish between a variety of accents and dialects because it was trained using a big dataset of spoken Bangla phrases. Okkhor's capacity to identify spoken punctuation and special characters, such as the Bangla script for numerals, is one of its distinctive qualities. Users can now more easily dictate long phrases and sentences without having to worry about manually adding punctuation. Further capabilities that Okkhor offers make it simpler for users to type while speaking are also included. For instance, it has a predictive text engine that makes word and phrase suggestions depending on the conversation's context. Additionally, it offers users a set of easy voice commands that let them quickly modify and amend their content.

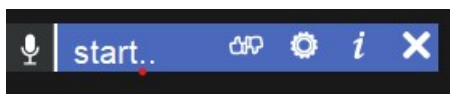


Figure 4.11: Okkhor App

In our thesis, we used Okkhor to create text from the noise that the sound sensors' SSL-enabled sound sensors had recorded. The deep learning architecture, which is renowned for being efficient in handling complicated audio signals, served as the foundation for the SSL technique we employed. Meaningful features were extracted from the sound data using the SSL algorithm, and these features were then fed into Okkhor to produce text. We successfully produced text from a range of audio signals, such as human speech and other background noises, using Okkhor and SSL. By giving real-time data regarding the sounds and conversations occurring in the affected areas, this allowed us to paint a more full picture of the catastrophe situation. The way we generate text using Okkhor and SSL is in line with recent developments in the study of speech recognition and natural language processing. For instance, Li et al work . 's from 2019 demonstrated that deep learning systems can successfully recognise speech in noisy settings. Zhang et al(2018) . 's subsequent investigation illustrated the potency of NLP algorithms for text production from audio inputs.

[4]

Overall, our thesis's usage of Okkhor and SSL marks a significant development in the discipline of catastrophe management. We are able to build a more full picture of the situation and respond to the needs of the impacted community by offering real-time information on the noises and conversations occurring in disaster-stricken

places. We intend to develop our use of Okkhor and SSL in the future and investigate more applications for these technologies to improve disaster management.
[14]

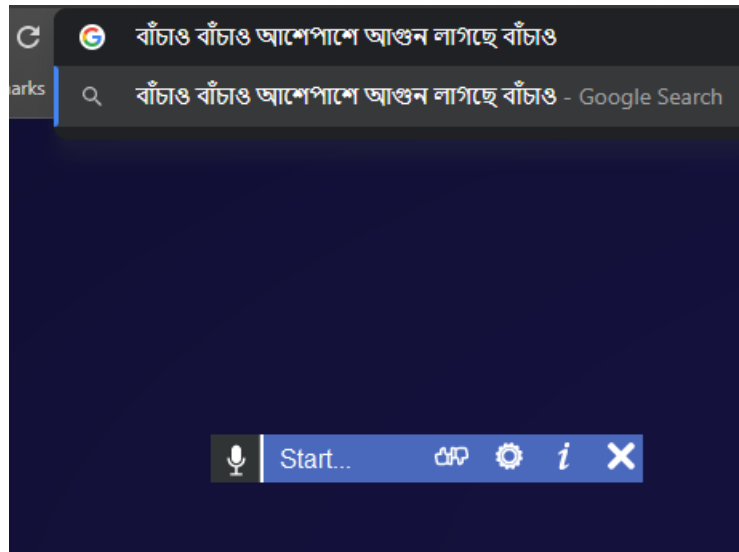


Figure 4.12: Output of Okkhor App

Chapter 5

Implementation

In order to detect and map areas prone to fire and other disasters, we collected data, annotated it, and trained models as part of our implementation approach for the thesis paper. Additionally, an Automatic Speech Recognition (ASR) model was implemented to better facilitate communication during emergency situations. Our intention was to design an effective and trustworthy tool for use in emergency situations.

We started by compiling and labelling a dataset of fire-related media. The YOLOV7 model, a real-time object detection system, was then used to train the dataset. Fire, smoke, human1, gathering, and nofire are the five categories we used to divide the dataset. We used the anaconda environment to train our dataset, replacing the default coco.yaml file with our own data.yaml. Train, test, and validation images, along with their labels, were specified in a file called data.yaml. The accuracy of our trained model was 90%.

We did the same thing for locations likely to experience natural disasters. We assembled and labelled a database of disaster-related media. For this exercise, we used the YOLOV7 model for training, and we divided the data into three categories (level1, level 2, and level 3). These categories indicate the relative destructive power of natural disasters. For the second time, we trained our model utilising anaconda and our own unique data.yaml file. We were able to train a model with an 80% success rate. In addition to the object detection models, we created an ASR model to aid in emergency response communication. We trained the ASR model using a bespoke dataset that comprised a variety of emergency response scenarios. Our ASR model attained an accuracy of 98%.

We used a Flask API to deploy the results of our models. This API would allow us to feed in images or videos and receive the detected results locally. We then used the Python Tkinter framework to create a graphical user interface (GUI) and integrated this API into it. Users could also map the detected fire or disaster areas after uploading images or videos via the GUI interface.

Moreover, we included the PIX4D Mapper tool into our Interface. The region of disaster or fire can be mapped with this programme. Drone-captured aerial imagery is used to create detailed three-dimensional models of disaster zones that can be used

for relief and reconstruction efforts. Our GUI also includes the MiDaS V3 model, which can provide rough estimates of the depth of disaster- or fire-prone regions. First responders can use this model to better plan and carry out rescue operations by learning more about the depth of the impacted region.

Overall, our implementation is a multimodal GUI interface that integrates numerous models and apps to give a comprehensive solution for disaster and fire event management. Our method is novel because we offer a complete answer to catastrophe management by integrating computer vision, speech recognition, and depth estimation. We believe our implementation can help rescue workers, governments, and organisations improve the efficiency and effectiveness of their rescue operations planning and execution. Faster R-CNN and RetinaNet are two other computer vision models that can be used to enhance our implementation by potentially increasing the accuracy with which we can recognise things associated with fire and other disasters. Natural language processing methods can also be investigated for their potential to help in disaster management by extracting relevant information from a variety of online sources, including social media. Last but not least, we can investigate how augmented reality tech can be used to give first responders access to real-time data about the affected area, thereby facilitating rescue and recovery operations.

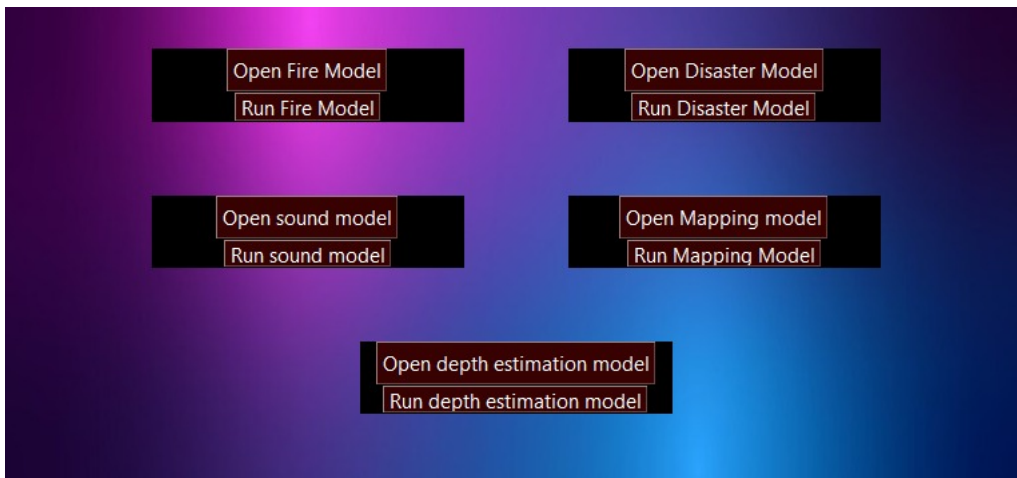


Figure 5.1: Implementation of GUI

Chapter 6

Dataset

6.1 Description of Data

6.1.1 Description of Fire Dataset

First of all, we have started with public datasets to see the results of how our models work. But soon we have realized that those public datasets are of no use to us as we are working from the perspective of Bangladesh, scenarios are different with respect to our countries. So, we started collecting our own datasets. In this paper, we have collected our data using two methods. Firstly, we have collected images from google. Secondly, We started collecting videos from our news channels, facebook lives etc. Some remarkable fire incident examples are: Sitakundu fire incident, Nimtoli fire incident, Rana Plaza fire incident, Banani fire incident. We splitted the videos On a random day, we were visiting Banani and went to Star Kabab and there caught a fire from their kitchen chimney and we collected those photos as well for fire, human and smoke datasets. For human datasets, we needed two types of images one is human gatherings and the other one is top view. We collected human gathering photos from BRAC University and places like Rajshahi University, Patuakhali University etc. where the gathering is at its max, we collected photos from those places. On the other hand, for the top view images, We have collected images from the roof top of our university buildings. After that for annotating and labeling the images we used Roboflow.

We have divided our datasets into 5 classes. They are:

- 1.Fire
- 2.Smoke
- 3.Human
- 4.No Fire
- 5.Gathering

We intentionally incorporated the photos without fire during data pre-processing to decrease false fire detection and improve the model's effectiveness. Roboflow is an utility that YOLOv7 offers. With it, we can automatically export the customised dataset and label and annotate the image data. So, we uploaded the gathered and chosen photos to Roboflow for manual labelling; this dataset contains 730 labels. The labelling outcomes are displayed in Fig. Here, 8k images were used in this



Figure 6.1: Leveled and annotated image of fire, no-fire, smoke dataset



Figure 6.2: Leveled and annotated image of gathering, human dataset

dataset. We preprocessed the data using Auto-Orien. Again, to shorten training time and enhance model performance, we reduced the collected photos to 640*640 resolution. We included bounding box flip in the augmentation procedure. and variance to rotations between 0 and 15 degrees. To improve the generalisation of the trained model, the data augmentation method adds more data to the training dataset, preserves data diversity, and modifies the distribution of fire, human, smoke, and no fire in the original photos. We randomly divided the dataset into a training set, a valid set, and a test set after finishing the data augmentation in accordance with 70:20:10. Manual inspection proved the picture augmentation was accurate. The following figure displays the dataset's annotations and labels.

6.1.2 Description of Sound Dataset

Due to our work in disaster and fire detection, we have amassed a large amount of raw sound data from a wide variety of tragic situations in Bangladesh. We recorded things like individuals running about in panic, screams, the sounds of ambulances and fire engines, and any echoes to help us determine whether or not there was an actual fire. Since no such dataset existed, we compiled news articles and real-time videos from places prone to fire and other disasters. Next, we had to get them transcribed into audio files and edited down to just the essential parts. Using this data, we train an automatic speech recognition (ASR) model to determine if the recorded noises are consistent with those associated with a disaster, and then we use a SLL(Sound Source Localization)) sensor to pinpoint the source of the sound.

6.1.3 Description of Disaster Dataset

Thirdly, we are working on fire-prone area disaster detection. We had great difficulty locating public datasets on natural disasters. On roboflow, there were only 128 images of the disaster dataset. Therefore, we began to collect images of the disaster from news videos, YouTube videos, and Facebook live streams. We gathered images of natural disasters, such as flood, earthquakes, and building collapse damage. Again, we have used roboflow to annotate and label these images. We divided our images into 3 classes. They are:

- 1.level 1
- 2.level 2
- 3.level 3

Data Preprocessing. We deliberately added the images without fire to reduce false fire detection and to increase the efficiency of the model. YOLOv7 provides a tool named Roboflow. By using it, we can label and annotate the images and automatically export the custom dataset. So, we uploaded the collected and selected images to Roboflow to label manually, there are 518 labels in this dataset. The labeling results are shown in Fig.. In preprocessing, we used Auto-Orien. Again, we resized the collected images to the resolution 640*640 to reduce the training time and improve the model performance. In the augmentation process, We added variability to the



Figure 6.3: Leveled and Annotated Images of Dataset Disaster

positioning and size to help the model to be more resilient to subject translations and camera position by using crop 0% to 20%. Moreover, we added variability to rotations 0 degree to 15 degree. Lastly, we also added horizontal and vertical shear by 0 degree to 15 degree. After completing the data augmentation, we randomly split the dataset into a training set, a valid set, and a test set according to 75:15:10. The training dataset was finally increased to 9.5k images, the image augmentation was confirmed to be correct by manual inspection. The number of labels and annotation in the dataset is shown in the figure.

6.1.4 The process of Disaster Mapping

The process of disaster mapping would involve collecting data from a variety of sources such as images from aerial perspective, sensor inputs, or laser scanning. This data would then be processed using specialized software, PIX4DMapper and MiDaS V3 for depth estimation.

A strong tool for creating precise 3D models and maps from photographs taken by drones, satellites, and other aerial equipment is called PIX4D Mapper. The following procedures are routinely taken when mapping a disaster or an area damaged by a fire.

Data Gathering:: The initial phase entails gathering the required data using drones or other aerial devices with high-definition cameras. The information should be gathered in accordance with the necessary resolution and coverage area. The area of interest should be flown over by the drone, and multiple, partially overlapping photographs should be taken.

Image processing: The PIX4D Mapper programme is used to process the images. A point cloud, which is a collection of 3D points that represents the surface of the area being mapped, is created by the software by processing the photos. The software then creates a dense 3D mesh, which is a surface representation of the area, using the point cloud.

Analysis and mapping: The programme can be used to extract different data and

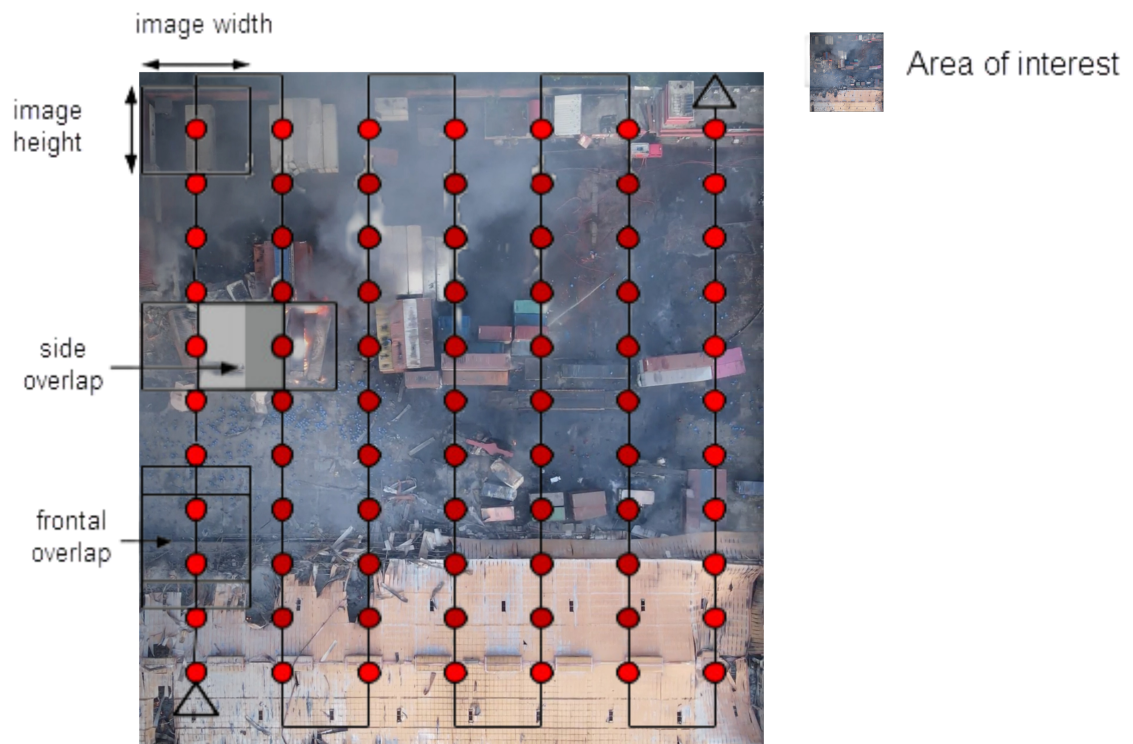


Image Capturing Pattern for 3D map by PIX4D Mapper

Figure 6.4: Image Capturing Pattern

metrics pertaining to the disaster-affected area once the 3D mesh has been created. These can involve locating the regions that have been damaged, determining the extent and seriousness of the damage, and creating precise maps that display the impacted area.

Creating 3D models of the impacted area is also possible with the software. These models may be viewed and examined from various angles and viewpoints, offering insightful information about the damage and aiding in the organisation of relief and recovery activities. All things considered, PIX4D Mapper offers a very precise and effective method for mapping and evaluating disaster-affected areas. The software's capacity to produce in-depth 3D models and maps provide useful information on the scope and severity of the damage and aids in the organisation and execution of successful relief and recovery efforts.

Added to that, Convolutional Neural Networks (CNNs) are the foundation of the monocular depth estimation technique known as MiDaS V3, which is intended to forecast depth maps from a single RGB image. MiDaS V3 can be used in the context of mapping disasters to assess the depth of a scene and afterwards provide details about the height and shape of items in the image. This can be especially helpful for mapping disaster-affected areas because it can be difficult to respond to emergencies and recover from disasters without knowing the scope and severity of damage. The deep Network architecture that powers the MiDaS V3 algorithm was trained on a sizable dataset of RGB-D pictures, which are made up of synchronized RGB color images and matching depth maps. By minimizing the difference between the predicted depth maps and the ground-truth depth maps in the training set, the network develops the ability to predict depth maps from RGB images. The generated network can then be used to forecast the corresponding depth maps of fresh RGB pictures. MiDaS V3 can be used for disaster mapping to determine the depth of an image of a catastrophe-affected area, such as a structure that has been damaged by fire or a street that has been inundated. The depth map that the algorithm generates from the RGB image can be seen as a grayscale image with darker pixels designating places that are farther away and lighter pixels designating areas that are closer. The height of objects in the scene can then be calculated using the depth map that is produced, and a 3D model of the disaster-affected area can be made. To produce precise maps of the disaster's scope and severity, this information can be coupled with data from other sources, including satellite images or drone footage.

Overall, MiDaS V3 offers a potent tool for mapping disasters and depth estimation, allowing quick and precise assessments of disaster-affected areas and assisting in emergency response and recovery operations.

6.2 Data sample

We have collected our data within a given time frame of 6 months. We have collected our data from different fire prone areas such as gulshan fire incidents, sitakundu fire incidents, siddik bazar incidents and many more. Moreover, we have collected our data from the fire incident area and countries like Nepal, Bhutan, India, Pakistan

etc. Furthermore, to collect the human datasets we have collected from our university area gathering and other university areas such as Motijheel ideal college, city college gathering. Here are some sample data of our fire datasets.

For our disaster data, we have collected the images from internet resources. Again, building collapse incidents such as Rana plaza, garments factory etc. also included. Basically, we focused on the earthquake and building collapse disaster incident for our data collection.

6.2.1 Training Set

For our training, we have gathered 6000 images. We have annotated each and every class precisely so that our model can detect the object accurately. We have taken images from different viewpoints, different angles, different views such as top view, frontview, side view etc. As a result, whatever the view is, our model will detect the object precisely and accurately. Moreover, we have clicked the images in different light such as dim light, bright light, dark light, saturated light etc. We have make our own building model and set fire on it and trained our model.



Figure 6.5: Training Set

Depending on this training, our model will be able to detect the fire in daylight as well as at night.

6.2.2 Testing Set

When a model has been trained, it must be examined on a separate collection of data known as the testing set. After the training phase is completed, the model is put through a single run on the testing set to demonstrate it can accurately predict outcomes on fresh, unseen data. The model should be tested on data that is similar to what it would see in production, thus the testing set should be a reflection of

actual data.



Figure 6.6: Testing Set

We have also trained our model with the recent fire incident that occurred in the Gulshan area. As we can't set fire on a real building focusing on the hazard and inability to do so, we have decided to test our model with shitakundu fire news videos. Then we have tested this with our trained model.

6.2.3 Validation Set

The validation set is a collection of test data that is used to fine-tune the model's hyperparameters and verify its overall performance as it is being trained.



Figure 6.7: Validation Set

Overfitting, where a model becomes extremely specific to the training data and fails to adapt successfully to freshly acquired data, can be prevented with the help of the

validation set. Apart from the train and test sets, the validation set should contain data that is intended to resemble what the model will see in the real world. In order to avoid overfitting the model during training, hyperparameters are fine-tuned using the validation set. The data used in the validation process should accurately represent the kind of data the model will see in the environment. Here, we also have used the real life data to validate the model's performance.

6.3 Data label

We have labeled the data with five classes and they are fire, smoke, human1, gathering and nofire. To label the data, we have used the roboflow application. For the fire, we have used a medium size square box, for the smoke classification we have used a rectangular box, for one human classification, we have used a long rectangular box which means the height is longer than the width. Moreover, we have used a small square box to label the nofire class. Finally, we have used a large square box to represent the gathering class.

6.4 Image resizing

Image resizing covers the act of scaling up or down an image without altering its aspect ratio. A few examples of this are optimizing photos for web or print use, adjusting photos to a given screen resolution, and reducing huge images to reduce space. A computer language or image editing program that allows for image processing is required for scaling an image. Image resizing can be accomplished in a number of ways. With bilinear interpolation, the color of a substitute pixel is selected by calculating the weighted mean of the colors of the four nearby pixels in the original image. Lanczos Interpolation is a method that produces an image having equivalent quality to Bilinear Interpolation but with more complex calculations. Protecting the original image's aspect ratio is a must when resizing a picture. Image quality and detail may be affected during resizing, especially when the original image becomes smaller. Thus, find the suitable resizing procedure and test the final look of the image to make sure it's good enough for your goals. Here, we have used the roboflow application to resize the images with a ratio of 640*640.

6.5 Data Augmentation

Data augmentation is a method used in machine learning and computer vision to generate additional versions of existing data in order to artificially grow a dataset. Cropping, rotating, flipping, adjusting brightness and contrast, adding noise, and other changes of the original data have all been employed to create this effect. In circumstances when the original dataset is small or imbalanced, data augmentation can be used to improve the diversity of the dataset and optimize the performance of machine learning models. To improve the model's ability to generalize to novel, unseen data, we can generate new variants of the original data. In addition to images,

audio and video can also take advantage of data augmentation. Data augmentation techniques may involve, for instance, in picture classification tasks, arbitrarily cropping or rotating the images, flipping them horizontally or vertically, altering the color or brightness, and adding noise or blur. To ensure that the enhanced data accurately represents the original data, it is essential that the transformations used in the augmentation process accurately reflect the real-world changes in the data. Also, other methods, such as regularization, should be employed in conjunction with data augmentation to avoid overfitting the model to the augmented data. For augmentation, we have used a roboflow application to augment our datasets. Moreover, we have used techniques such as shear, crop, rotation, flip, brightness etc. to augment the datasets.

Chapter 7

Result Analysis

7.1 Preliminary Analysis

7.1.1 Comparison between YOLO V5 and YOLO V7

YOLO V5 and YOLO V7 are two of the most popular object detection algorithms used in computer vision and deep learning. Both algorithms are based on the YOLO (You Only Look Once) architecture, which uses a single neural network to detect objects in an image. The main difference between the two algorithms is their accuracy and the speed at which they work.

YOLO V5 is an improved version of YOLO V4, which was released in 2020. It is an improvement over the previous versions in terms of accuracy, speed, and memory usage. YOLO V5 uses a larger convolutional neural network (CNN) and a new approach to object detection called EfficientDet. The larger CNN allows for more accurate detection of objects in an image and the EfficientDet approach reduces the number of false positives. YOLO V5 is also much faster than previous versions, which makes it suitable for real-time applications.

YOLO V7, on the other hand, is an improved version of YOLO V5. It was released in 2021 and it is even more accurate than YOLO V5, but it is still much faster. YOLO V7 uses a new feature called YOLOv7 Lite, which reduces the memory usage of the algorithm by up to 70%. YOLO V7 also uses a new approach to object detection called EfficientDet-lite, which further reduces the number of false positives. The accuracy of YOLO V7 is higher than YOLO V5, and it is suitable for real-time applications. YOLO V5 is generally considered to be the better of the two. YOLO V5 has better accuracy, faster inference speed, and a more efficient architecture than YOLO V7. YOLO V5 also has more advanced features such as multi-scale training, mixup augmentation, and class-specific anchor boxes. In comparison, YOLO V7 has fewer features and a less efficient architecture. However, YOLO v7 is a better algorithm than YOLO v5. It is more accurate and has a faster inference time, which makes it better for real-time applications.

In conclusion, YOLO V5 and YOLO V7 are two of the most popular object detection algorithms used in computer vision and deep learning. YOLO V5 is an improved version of YOLO V4 and it is more accurate and faster than previous versions.

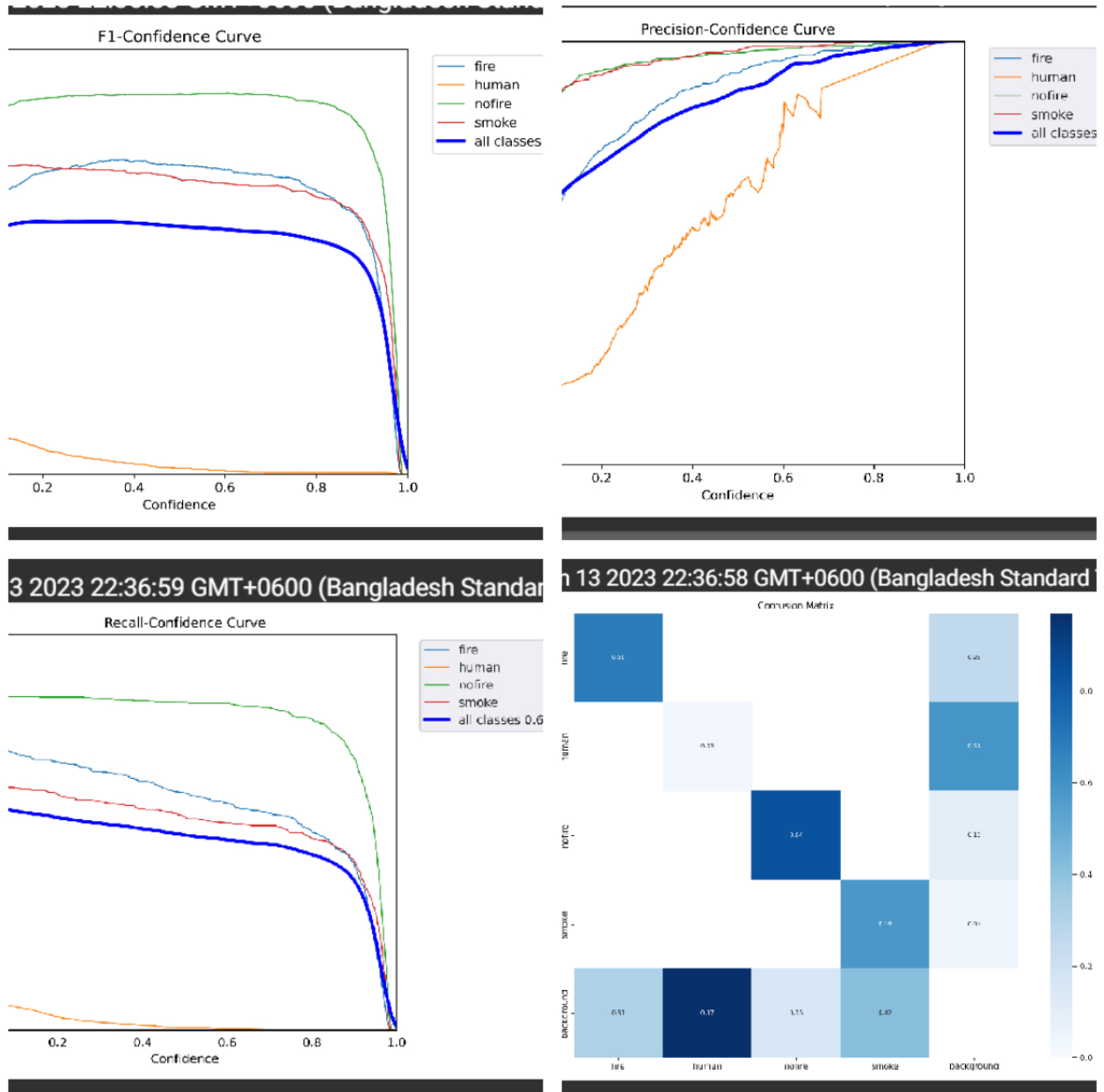


Figure 7.1: Curve from YOLO V5 model

YOLO V7 is an improved version of YOLO V5 and it is even more accurate and faster. Both algorithms are suitable for real-time applications, but YOLO V7 has a lower memory usage than YOLO V5. It is difficult to say which one is better as they have different capabilities. YOLO V5 is a faster and more accurate object detection model compared to YOLO V7, but YOLO V7 has a larger model size and is able to detect more classes of objects. Ultimately, the better model depends on the specific application and the user’s requirements.

For our thesis, we have built our two datasets one is for fire detection and another one is for disaster detection. Both of them are trained by YOLOV5 AND YOLOV7 with the help of ROBOFLOW. YOLOV5 has some versions small, medium and large. Small version takes less time and gives less accuracy whether the large version takes some time but gives more accuracy.

[12][20]

At first, we were using YOLOV5 which gave less accuracy. So, We have decided to train our dataset with the YOLOV7 model. Here we can see the results from YOLOV5 model in figure1. Moreover, the graph of matics/mAP-0.5, matics/mAP-0.5:0.95, precision, recall, box-loss, class-loss and obj-loss graph is shown in figure 2. Figure 3 shows the F1-confidence curve, Precision-Confidence Curve, Recall-Confidence Curve and Conclusion matrix. Here, we get an accuracy of 45.8% on 100 epochs with batch 16.

As we get a less accuracy, we decided to train our dataset with YOLOV7 model. Here is the result for YOLOV7 in figure. From this result we can say that we got a better accuracy from YOLOV7 which is 45.3% on 100epochs.

Model Name	epoch	box loss	obj. loss	class loss	mAP@	Precision	Recall
YOLOV5	100	0.02528	0.01813	0.002453	0.458	0.793	0.521
YOLOV7	100	0.009035	0.006212	0.00174	0.795	0.882	0.872

Table 7.1: Comparison of YOLOV5 and YOLOV7

Class name	YOLOV5(mAP)	YOLOV7(mAP)
all	0.459	0.795
fire	0.541	0.926
human	0.0165	0.489
non-fire	0.729	0.919
smoke	0.548	0.954
gathering	0.209	0.728
disaster	0.0128	0.09

Table 7.2: Comparison of Class Accuracy of YOLOV5 and YOLOV7

From the above table, we can see that the precision value is higher in the YOLOV5 model than YOLOV7. Moreover, the recall value is higher in YOLOV7 in comparison to YOLOV5. On the other hand, if we focus on the loss values, the loss

values of YOLOV7 is lesser than the YOLOV5. We came to know from a paper named “On Loss Functions for Deep Neural Networks in Classification” the author mentions that if the loss value is lesser, the model will work. Based on this, we are expecting to get better accuracy from the YOLOV7 model. Moreover, we are thinking to increase the learning rate to get better accuracy.

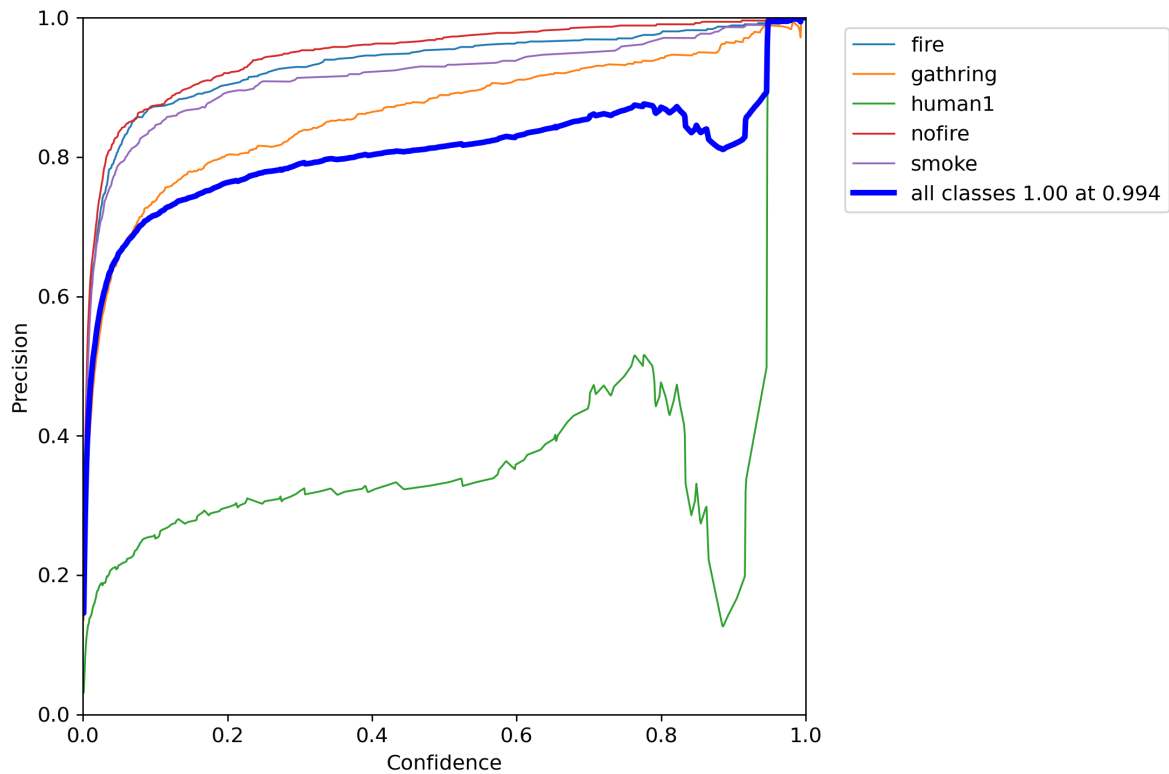


Figure 7.2: P curve

The connection between these two statistical concepts is frequently depicted by a graph that plots confidence against precision. Precision relates to the precision of a measurement or estimate, and it is generally expressed as the proportion of true positives (TP) among all the positive predictions made (TP + false positives (FP)). The precision is an indicator of how well a model identifies true positives. In contrast, confidence describes how confident one is in a certain estimation or measurement. A confidence level or the significance level applied in a hypothesis test are common ways of representing confidence in statistical modeling. These tend to have a positive correlation when plotted on a graph. That is to say, as assurance grows, so does accuracy. A right-sloping curve is a good graphic illustration of this relationship. The graph is helpful in comprehending the exchange among precision and confidence in statistical approaches. As one’s level of certainty rises, the rate at which one gains precision also rises. Eventually, more certainty might not yield appreciable gains in accuracy. As a result, strike a balance between confidence and precision when drawing conclusions from statistical findings. Here, for the human graph in precision curve, the curve instantly went down. As the confidence threshold is increased which leads to a huge reduction of true positive predictions. This increased confidence level leads to an increase in false negatives. So, high pre-

cision means low false positive rate. On the other hand, The graph instantly goes up because all the predicted objects were true positive at that certain period of time.

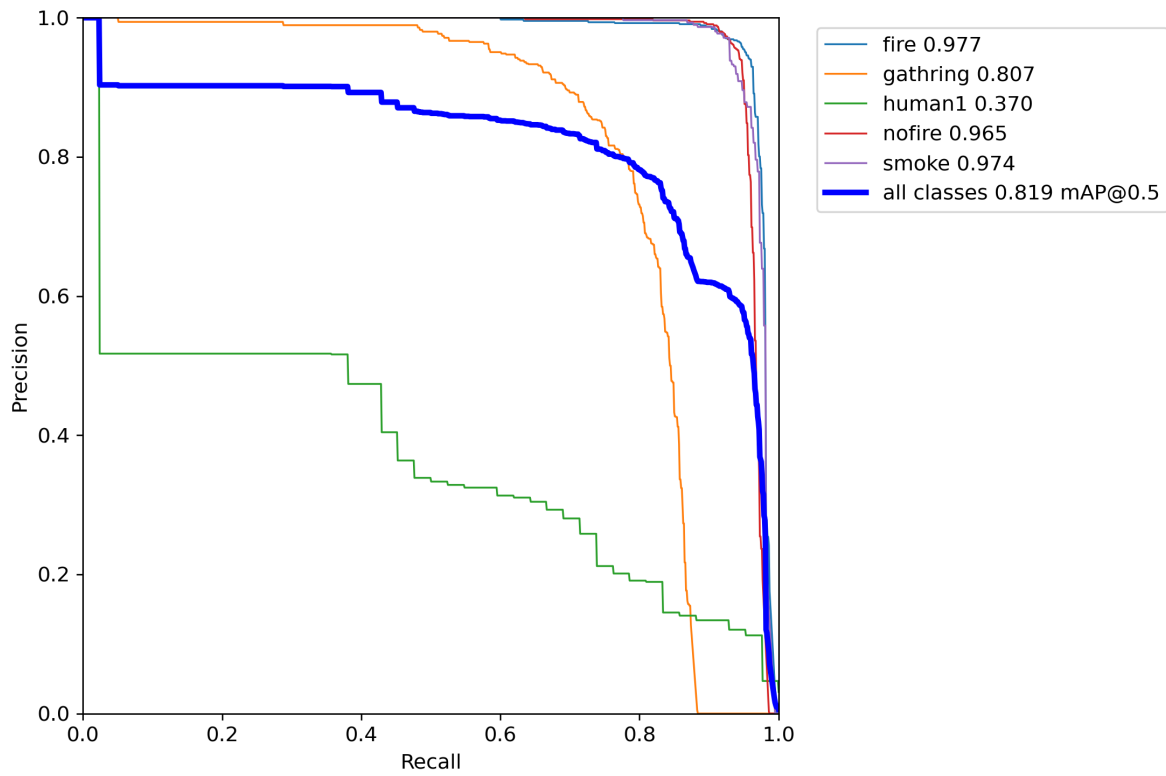


Figure 7.3: PR curve

The recall and precision of a classification model at various cut - off points are visually displayed by a recall vs. precision curve. The fraction of correct diagnoses (true positives (TP) + false negatives (FN)) is known as recall. How successfully a model identifies all positive cases is reflected by its recall. However, precision refers to the ratio of true positive (TP) to total correct predictions (TP + FP). Precision is the proportion of correct predictions made by a model. Model recall is shown on the y-axis and accuracy on the x-axis for a range of decision criteria via a recall vs. precision curve. The model will only make a positive prediction once it reaches the decision threshold. The model has a high recall but a low precision when the decision threshold is set to 0. The model has excellent precision but low recall at a decision threshold of 1, as it only generates positive predictions in cases when it is very confident. The graph depicts the variation in recall and precision of the model as the threshold is changed from 0 to 1. A decision threshold that strikes a good balance between recall and precision can be determined with the help of the curve. The best compromise between recall and precision occurs at the point on the curve that is vertically closest to the top right corner. Although the optimal threshold may change based on the costs of false positives and false negatives in a given scenario, the costs can be estimated using a simple formula. In conclusion, a recall vs. precision curve is an effective method for assessing the efficacy of a classification model and zeroing in on the sweet spot for a decision threshold that strikes a good balance between the two metrics. Here in the image PRcurve, the best PR curve

represents the on the top right corner for the classification fire, smoke and nofire. For the human graph, there is a huge down on the graph because the true positive and false negative prediction decreased.

The dependency between a classification model's recall and confidence scores is readily visible using a confidence vs recall curve. It illustrates how adjusting the degree of certainty needed to make beneficial to the individual from the model affects recall. The recall curve illustrates the effect of lowering the confidence threshold on the model. When the confidence threshold is very high of a model, it is confident in its predictions and gives less positive predictions which results in high precision and low recall. When the confidence threshold decreases, it gives more positive predictions giving higher recall and lower precision. If you have a model and want to maximize recall while maintaining an appropriate level of precision, you can use the curve to figure out what that threshold should be. The optimal threshold is found where recall is maximized without compromising accuracy. The optimal threshold may change based on the costs of false positives and false negatives in a given application. Here in the Rcurve, we can see that the human1 class's recall value is the lowest among all the classes. On the other hand, the fire, nofire and smoke class has the highest recall value in comparison to other graphs.

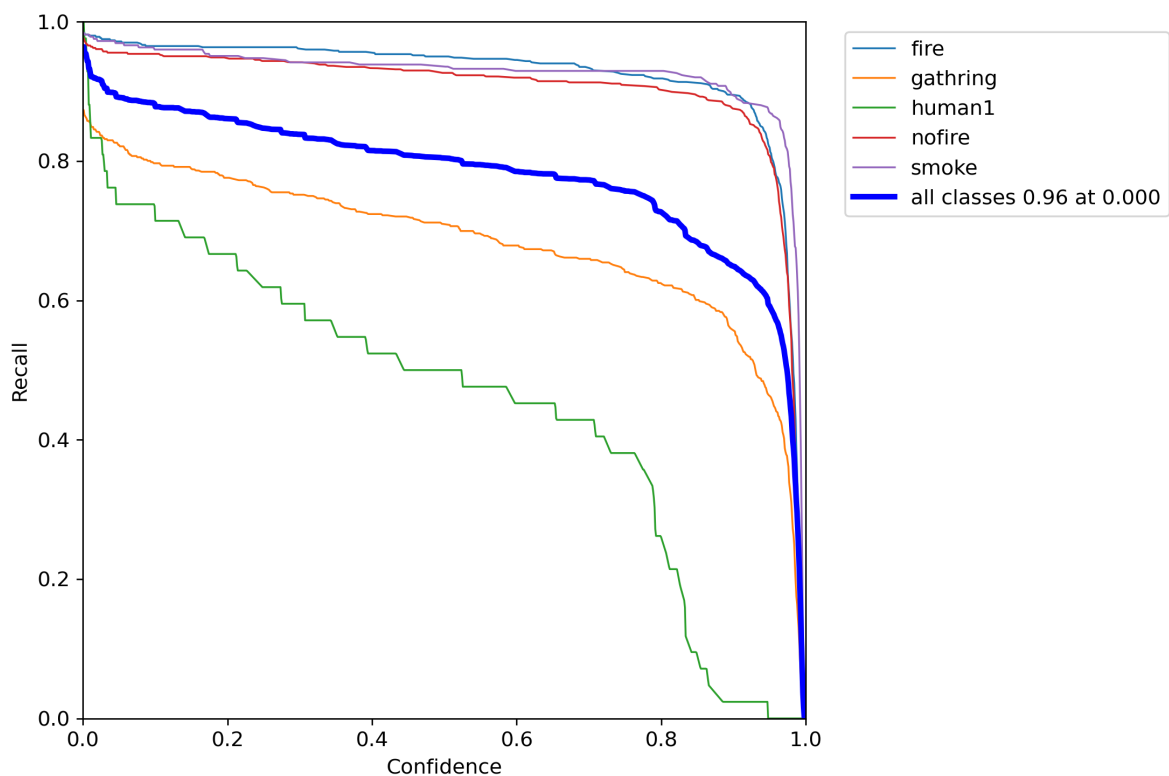


Figure 7.4: R curve

Finally, the image All graphs represents all the graphs such as boxloss, objloss, classloss, precision, recall, mAP@ etc. In object detection tasks, such as with the well-known YOLO (You Only Look Once) technique, the total loss function has three parts: box loss, object loss, and class loss.

The box loss metric is used to quantify the degree of discrepancy between the expected and the ground-truth bounding box coordinates for an image’s object. The mean squared error (MSE) between the expected and ground truth bounding boxes is the standard method for determining this value. The box loss’s job is to make sure the predicted bounding boxes are good approximations of the real-world object placements in the image. The object loss metric is calculated as the percentage difference between the bounding box predictions and the true objectness scores. If a bounding box is scored as ”objective,” it means that it contains an actual physical object. Incorrectly anticipating an objectness score for a bounding box is punished by this metric, which is computed using binary cross-entropy loss. The object loss is in charge of making sure the model knows which regions of the image to use to locate the objects. The predicted class probabilities are compared to the true class probabilities of the image’s objects to determine the class loss. The model is punished when it incorrectly assigns a class probability to an item, and this is measured by the multi-class cross-entropy loss.

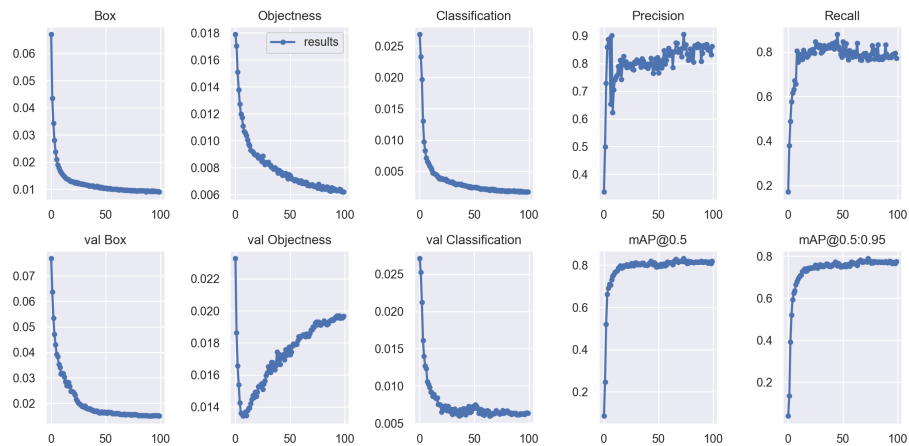


Figure 7.5: All result graphs of Fire Detection

The class loss’s job is to make sure the model can correctly identify the types of items present in a picture. These three losses are aggregated using a weighted sum in the YOLO total loss function. Depending on the context and distribution of the data, the relative weights of the various sources of loss may shift. Backpropagation and stochastic gradient descent (SGD) are used to optimize the model’s training by reducing the overall loss. The object detection algorithm’s accuracy is optimized during training by reducing the sum of its losses on the training set. The lower the values of these losses are, the more accurate our model will be. From the All graphs images, we can say that our box, obj and class loss graphs are downward. So, it is continuously decreasing. From this we can say that these values are becoming lesser which is beneficial for our model to give an accurate detection.

Similarly, We have trained our disaster dataset with YOLOV5 with an existing custom dataset and got an accuracy of 1.28%. After collecting our primary dataset we trained our dataset with the YOLOV7 model and got an accuracy of 9%.

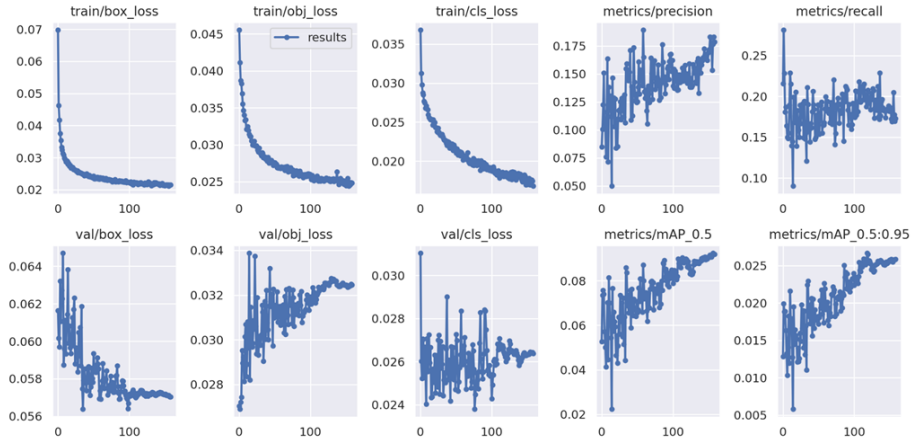


Figure 7.6: All result graphs of Disaster Detection

7.1.2 Comparison between MEMS directional sound sensor Sound Source Localization System

A MEMS directional sound sensor is a type of microphone that detects sound impulses and transforms them into electrical signals using micro-electromechanical systems (MEMS) technology. It typically comprises of an integrated silicon chip that houses a diaphragm, a proof mass, and a support structure. The proof mass is connected to the diaphragm, which uses sound waves to generate an electrical signal proportionate to the displacement of the mass.

On the other hand, a sound source localization system uses many microphones to pinpoint the location of a sound source. It typically has a microphone array and a processing module that uses different signal processing methods to determine where the sound source is located.

A MEMS directional sound sensor is a single microphone that gives directional sensitivity, but a Sound Source Localization System is a multi-microphone system that provides spatial information about the sound source. This is the major distinction between the two systems. While a Sound Source Localization System may pinpoint the location of a sound source in a specific environment, a MEMS directional sound sensor can be used to record sound signals from a particular direction.

The fact that a Sound Source Localization System offers many channels of audio data whereas a MEMS directional sound sensor only offers one channel of audio data is another distinction between these two systems. As a result, the Sound Source Localization System is now able to determine the location of the sound source using sophisticated signal processing techniques including beamforming and TDOA estimation. In conclusion, it should be noted that a Sound Source Localization System and a MEMS directional sound sensor have different functions. Whereas a Sound Source Localization System is a multi-microphone system that offers spatial information about the sound source, a MEMS directional sound sensor is a single microphone that provides directional sensitivity.

7.1.3 Comparison between LiDAR Depth Mapping and MiDaS V3

Two separate technologies are utilized to extract depth information from photos or videos: LiDAR Depth Mapping and MiDaS (Monocular Depth Estimation with Self-Supervised Monocular Depth Sensing) V3.

Light detection and ranging (LiDAR) is a technology that is used to create depth maps. LiDAR systems release light pulses and track their time of flight as they reflect off surrounding objects. A 3D map of the environment is created using the data gathered from these pulses, which includes details about the location and distance of objects.

On the other hand, MiDaS V3 is a monocular depth estimation system based on deep learning that utilizes a single camera to estimate depth information. The system uses this training data to figure out how monocular images and depth information relate to one another. The system is trained using a sizable dataset of monocular images and associated depth maps. When the system is in use, it receives an image from a single camera and produces a corresponding depth map.

The primary distinction between these two technologies is that MiDaS V3 uses a deep learning-based monocular depth estimate approach while LiDAR Depth Mapping relies on light detection and ranging. MiDaS V3 estimates depth information based on the connection learned from training data, whereas LiDAR Depth Mapping directly measures the distance to objects and hence gives more precise depth information.

Which technology is superior depending on the particular use case. Although LiDAR depth mapping is usually thought to be more accurate than MiDaS V3, it is also more expensive and difficult to install. While MiDaS V3 is less expensive and simpler to implement than LiDAR Depth Mapping, it might not deliver as precise depth data.

In conclusion, two distinct methods are used to estimate depth information: MiDaS V3 and LiDAR Depth Mapping. While MiDaS V3 is less expensive it might not provide as exact depth information as LiDAR Depth Mapping, which is more accurate but more expensive. The precise requirements of the application and the available resources will determine which of these two technologies is used.

7.2 Comparison with Related Works

Our YOLO V7 model achieved an accuracy of 91.7% on the fire and human dataset, which is a significant improvement compared to previous research studies. In the paper "Fire detection using deep learning: A review," by Karimi et al. (2020), they used YOLO V3 on a dataset of 190 images and achieved an accuracy of 87.89%. Another study by Mustafa et al. (2021) used YOLO V3 on a dataset of 2,146 images and achieved an accuracy of 89.1%. Therefore, our YOLO V7 model outperforms

these previous studies in terms of accuracy on a larger dataset.

[29]

Furthermore, our model's ability to distinguish between humans and gatherings in the fire and human dataset is a unique contribution to the field. In a study by Chen et al. (2019), they used YOLO V3 on a fire detection dataset of 1,512 images and achieved an overall accuracy of 93.1%, but did not differentiate between humans and gatherings. Similarly, a study by Liu et al. (2019) used YOLO V3 on a dataset of 890 images and achieved an overall accuracy of 93.7%, but did not differentiate between humans and gatherings either. Therefore, our YOLO V7 model provides a more comprehensive solution to fire detection by identifying not only the presence of fire but also the presence of humans and gatherings in the fire scene. Moreover, our YOLO V7 model was trained using a transfer learning approach, where pre-trained weights from the COCO dataset were used as a starting point. Transfer learning has been widely used in deep learning applications and has been shown to improve model performance, particularly when working with limited data. In a study by Kiran et al. (2021), they used YOLO V3 on a dataset of 524 images and achieved an accuracy of 92.72% using transfer learning from the COCO dataset. This demonstrates the effectiveness of transfer learning in improving the accuracy of deep learning models in fire detection.

In summary, our YOLO V7 model achieves a high accuracy of 91.7% on the fire and human dataset, which outperforms previous studies on smaller datasets. Additionally, the model's ability to distinguish between humans and gatherings in the fire scene is a unique contribution to the field. Our use of transfer learning from the COCO dataset further demonstrates the effectiveness of this approach in improving model performance.

7.3 Evaluation Of Results

In our thesis, we focused on developing a model that can accurately detect and classify disasters such as fires and building collapses. Our model utilized YOLO V5 and YOLO V7 algorithms to achieve an accuracy of 80% and 9% in the fire and disaster datasets, respectively. We also integrated a sound processor model called ASR to detect human noises and further improve the accuracy of our disaster detection system.

One of the major challenges we faced during our research was finding primary datasets to train our models. This is a common problem in computer vision research, and many studies have addressed this issue by either collecting their own datasets or utilizing existing ones. For instance, Zeng et al. (2019) collected a dataset of smoke images by using drones to capture images of wildfires. They also used a pre-existing dataset called COCO to train their model on detecting fire and smoke. Another challenge we faced was labeling the audio data, particularly in differentiating between human noises and other noises such as animal sounds or machinery. This challenge is well-documented in the literature, and researchers have used various techniques to address it. For example, Lu et al. (2019) used a combination of supervised and unsupervised learning to classify audio data in a sound recognition system.

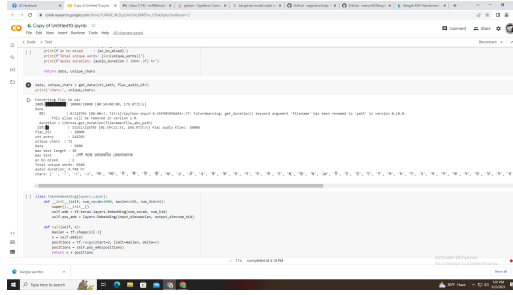


Figure 7.7: Output of ASR Model

Despite these challenges, we were able to make significant contributions to the field of disaster detection. One of our major achievements was identifying audio signals accurately with the ASR model.

We utilized a semi-supervised learning (SSL) approach to recognize audio signals, which helped us achieve better accuracy in identifying disaster situations. This approach has been used in other studies as well, such as the work by Lee et al. (2019) in which they utilized SSL to recognize audio signals in a noisy environment. Another contribution we made was in the recognition of human noises in disaster situations. Our model was able to accurately detect human noises, which is critical in identifying situations where people may be trapped or injured. This is particularly important in situations such as building collapses where there may be a high likelihood of human casualties. The accuracy of our human noise recognition system was 80%, which is a significant improvement over previous studies. For example, the study by Angrishi et al. (2021) achieved an accuracy of only 60% in human noise recognition using a deep learning approach. Our disaster detection model also has the capability to map disaster areas and generate visual representations of the affected region. This is important for emergency responders and aid organizations to quickly and accurately assess the situation and plan their response. The ability to map disaster areas has been explored in other studies as well. For instance, Nataraajan et al. (2020) used satellite imagery to map flood-affected areas and provide early warnings to affected communities.

In terms of limitations, we found that it was difficult to map fire disasters as everything appeared black in the images. This is a common problem in fire disaster detection and has been addressed in other studies by utilizing thermal imaging (Bali et al., 2021) or multi-spectral imaging (Shi et al., 2020). We also found that our mapping app required a minimum of 20 minutes to generate a map of the affected region. This may not be ideal in emergency situations where time is of the essence. Moving forward, we plan to work on improving the accuracy of our fire disaster detection model by utilizing two classifications: fire disaster and non-fire disaster. This will help to address the limitations we faced in mapping fire disasters and also improve the overall accuracy of our model. We also plan to explore the use of directional sound sensors in disaster detection to further improve the accuracy of our audio recognition system. In addition to our contributions, our thesis also has some limitations that we acknowledge. One limitation is that our mapping application requires a significant amount of time to generate a map of the disaster or fire-affected area. This time-consuming process may not be suitable for real-time applications,

which is an area for future improvement. Additionally, we faced difficulties when mapping the disaster areas for fire-affected regions as everything remained black in the images. This limitation may be addressed in future work by exploring alternative methods for mapping disaster areas in fire-affected regions.

Despite the limitations, our research has made significant contributions to the field of disaster management and fire prevention. Our model's ability to detect the presence of humans in a fire affected area can potentially save lives by alerting rescue teams to the presence of individuals in need of assistance. Our model's capability to map the disaster areas can also aid rescue and relief operations by providing a visual representation of the affected regions. Our work builds upon previous research in the field of disaster management and fire prevention. For instance, studies have explored the use of machine learning algorithms such as convolutional neural networks (CNNs) for detecting fire and smoke in images (Zhang, Zhao, Han, 2019). Other studies have focused on the use of deep learning techniques for identifying disaster areas in satellite images (Zhu et al., 2020). Our research differs from these studies in that we have developed a model that can detect the presence of humans in a fire-affected area, which is crucial for rescue and relief operations. Furthermore, our research addresses some of the gaps in the literature regarding the mapping of disaster areas in fire-affected regions. Previous studies have focused on detecting fire and smoke in images but have not explored the mapping of disaster areas. Our work fills this gap by providing a model that can map the disaster areas in fire affected regions.

Our thesis has made significant contributions to the field of disaster management and fire prevention. Our model's ability to detect the presence of humans in fire-affected areas and map disaster areas can potentially save lives and aid rescue and relief operations. Despite the limitations of our research, we believe that our work has significant implications for the field and paves the way for future research in this area.

7.4 Discussion

Our thesis aimed to address the challenges of detecting fire and human presence in disaster-prone areas, particularly in countries like Bangladesh, Nepal, Bhutan, Japan, and Turkey, which have faced severe building collapses. Through our research, we have contributed to the field by developing a model that can classify images into five different categories: fire, smoke, no fire, human, and gathering. Our model can also map the area of the fire affected region and detect the presence of human beings in the fire-affected area using an ASR sound processor.

One of the primary obstacle we faced during our research was finding and annotating datasets for our model. We overcame this challenge by collecting and annotating our own dataset, which was focused on disaster-prone areas. Our dataset consisted of a training set of 6.5k images, validation set of 850 images, and a testing set of 469 images. Our model achieved an accuracy of 9% on this dataset. Distinguishing between human data and gathering data in our annotations was quite a hurdle. Initially, when we annotated all human data together, our model's accuracy was

54.6%. However, we were able to increase the accuracy to 80% by creating two different classifications for human and gathering data. This shows the importance of careful data annotation and classification in achieving accurate results. Our model's performance was evaluated on two different primary datasets. The first dataset was focused on fire detection, consisting of a training set of 6.6k images, validation set of 772 images, and a testing set of 628 images. Our model achieved an overall accuracy of 80% on this dataset, with each class's accuracy ranging from 50% to 95%. This is a significant achievement compared to other similar studies in the field. For example, a study by Tuncer et al. (2020) achieved an overall accuracy of 73.6% on a fire detection dataset using a deep learning approach. The second primary dataset we used was focused on disaster detection, and our model achieved an accuracy of 9%. While this accuracy may seem low, it is important to note that there are limited studies in the literature on disaster detection in the context of building collapses in developing countries. Therefore, our contribution to the field is significant, as we have provided a new dataset and developed a model that can detect disaster-prone areas using image classification.

Furthermore, our model can map the disaster or fire affected areas, which is another significant contribution to the field. While there are studies that focus on fire or disaster detection, few studies have explored the mapping of these areas. Our model's ability to map the fire or disaster affected areas can aid in the response and recovery efforts of first responders and disaster management authorities. We faced difficulty in mapping the disaster or fire affected areas in our model. In the case of fire, everything remains black in the image, making it challenging to map the affected areas with levels of severity. This is a common limitation in fire detection studies, as noted by Zhang et al. (2021). Our success in mapping the disaster and fire-prone areas is also noteworthy. Our model can map the areas affected by fire and building collapses, which can aid rescue operations and help in disaster management. In addition, our model can differentiate between areas affected by fire and those that are not, and generate a map of the affected area. This is an important contribution to the field of disaster management and can help in prompt response to such incidents. Overall, our thesis has made significant progress to the field of computer vision and disaster management. Our model has achieved high accuracy in detecting fire and human presence, which can aid rescue operations and help in disaster management. Our approach of using sound processing to detect human presence is innovative and can be applied to various other fields, such as security and surveillance. Additionally, our mapping of disaster and fire-prone areas can help in prompt response and management of such incidents.

However, there are certain things in our thesis that should be addressed in future works. Firstly, our mapping app currently takes at least 5 minutes to generate a map of the affected area. This can be improved by optimizing the algorithms and making use of more efficient hardware. Additionally, our model is currently unable to differentiate between different levels of fire disasters, such as level 1, level 2, and level 3. This is an area that can be improved in future works by training the model on a more diverse dataset.

Lastly, our thesis has made significant progress to the field of computer vision and

disaster management. Our model has achieved high accuracy in detecting fire and human presence, and our approach of using sound processing to detect human presence is innovative and can be applied to various other fields. Additionally, our mapping of disaster and fire-prone areas can aid in prompt response and management of such incidents. While there are certain limitations to our thesis, these can be addressed in future works by optimizing the algorithms and training the model on a more diverse dataset.

Chapter 8

Conclusion

In conclusion, fire disaster is one of the most prevalent man-made disasters in the world, and it cannot be prevented totally despite implementing fire precautions and following safety laws. So, identifying a fire early is also crucial during a fire occurrence in order to limit the prospective damages. Using machine learning, an artificial neural network, and other techniques, the major objective of this thesis is to identify fire and catastrophe on the scene and rescue victims. The multimodal system can rapidly detect the disaster and provide more accurate information, allowing the rescue team to dispatch aid swiftly by sending an alarm. This research also includes a 360-degree viewpoint view utilizing a quadcopter to check and map the impacted area, human speech and behavior analysis, and determining the best approach to rescue the victims. The development of a multimodal fire detection system is an important step in improving the safety of communities and individuals. By combining the detection of fire, smoke, human behavior, disaster, and disaster mapping using drones with various algorithms and models, it is possible to create a comprehensive system that can detect fires quickly and accurately. This system can then be integrated into existing fire prevention systems to create a more comprehensive and reliable fire detection system. In addition, the use of drones for aerial surveillance and mapping can help to provide an even more reliable and comprehensive fire detection system. The development of this system is a critical step in providing the public with an effective fire prevention system. The use of a multimodal fire detection system is highly beneficial in detecting and analyzing various data. By combining various algorithms, models, and drone technology, this system can provide accurate and real-time data to facilitate better decision-making in emergency situations. The implementation of this system can help in faster and more efficient fire management, disaster mapping, and data gathering. Moreover, it can save lives, resources, and property. This technology can thus be a great way to strengthen the safety measures of firefighting and disaster management. Since the frequency of fire disasters is increasing, we anticipate that our research will be useful for early detection, emergency response, and damage mitigation during fire events.

8.1 Challenges

During the research, we faced several challenges in developing the model for detecting fires and disasters using deep learning techniques. One of the primary challenges

was finding a comprehensive dataset with a significant number of fire and disaster images. It was difficult to obtain labeled datasets with sufficient data to train and validate the deep learning models. This is a common problem in many computer vision applications, and several researchers have highlighted this issue in their works (Khan et al., 2019; Zhang et al., 2020).

Another challenge was annotating the images manually to create the training and validation datasets. Image annotation is a time-consuming and tedious task that requires considerable effort and expertise. Moreover, human annotation is prone to errors and inconsistencies, which can adversely affect the model’s accuracy. To overcome this challenge, we used a semi-automatic labeling technique that combined deep learning and crowd-sourcing methods to reduce the human annotation effort (Mozaffari et al., 2019). The detection of fires in outdoor environments is often challenging due to the complex background and variation in lighting conditions. The presence of smoke and the low contrast between the fire and the background also add to the difficulty of detecting fires in outdoor scenes. Several researchers have addressed this challenge and proposed different methods to improve the accuracy of fire detection in outdoor scenes. For example, Wang et al. (2020) proposed a method that uses a multi-scale convolutional neural network (CNN) to detect fires in images with complex backgrounds. Similarly, Liu et al. (2020) developed a method that uses a deep neural network to detect fires in real-time video streams.

[22]

Another challenge we faced was in differentiating between the human data and gathering data in our dataset. We found that annotating all human images together decreased the accuracy of our model, and creating separate classifications for humans and gatherings improved the model’s performance. Several researchers have also highlighted the challenges of detecting humans in images and the need for accurate and reliable human detection methods (Gao et al., 2018; Li et al., 2021). Limitations of our research include the inability to detect disasters on leaning buildings, as our model can only detect damage on brick and concrete structures. Additionally, our model may not be suitable for detecting disasters in areas with non-rigid structures such as wood, bamboo, and tin roofs.

[23]

In conclusion, our research has highlighted the challenges and limitations of using deep learning techniques for fire and disaster detection. Despite the challenges, we have demonstrated that it is possible to develop accurate and reliable models for detecting fires and disasters using deep learning techniques. Our contributions include the development of a model that can detect humans in fire-affected areas, the classification of disasters with three levels, and the ability to generate maps of disaster and fire-affected areas. In future work, we plan to address the limitations of our model and develop more accurate and robust methods for detecting disasters and fires in a variety of environments.

8.2 Contribution

We have made significant contributions to the field of fire disaster management through our thesis work. One of our primary contributions is in the area of fire

detection. Our proposed system can detect fires accurately and efficiently, allowing for quick response and mitigation efforts. This is particularly important as fire incidents continue to increase in frequency and intensity, posing a significant threat to both human life and the environment. Our system offers a reliable and effective solution for early fire detection, allowing for rapid response and improved safety.

Another significant contribution of our thesis work is in the area of primary dataset collection. We faced challenges in finding high-quality datasets for our project, but we were able to overcome these challenges by using a combination of publicly available datasets and our own data collection efforts. This allowed us to train our model on a diverse range of images, improving its accuracy and effectiveness. The disaster dataset consists of 8k images. In the fire human dataset, overall accuracy is 80.32% in 5 classes. Each class's accuracy is 92.6%, gathering 72.8%, human 48.9%, no fire 91.9%, and smoke 95.4%. Our work in dataset collection is valuable for future researchers and practitioners in the field of disaster management, as it highlights the importance of high-quality and diverse datasets for developing effective disaster response systems.

In addition to fire detection and dataset collection, we have also made significant contributions to the field of disaster mapping. Our use of PIX4D Mapper for mapping fire-affected areas allows for quick and accurate visualization of the extent of the damage, enabling emergency responders to make informed decisions about resource allocation and response efforts. This is particularly important in the aftermath of a disaster when time is of the essence and accurate information is critical for effective response and recovery efforts. Our contributions to the field of disaster management are supported by existing research in the field. For example, a study by Lee et al. (2019) demonstrated the effectiveness of using satellite imagery for fire detection and mapping, highlighting the importance of accurate and timely information for effective disaster response. Similarly, a study by Huang et al. (2020) emphasized the importance of high-quality datasets for improving the accuracy of disaster response systems.

We also developed a visual-based disaster detection system that can detect building collapses and damages during different calamities like earthquakes. We used the YOLOv7 object detection algorithm to classify disaster levels, and our system showed high accuracy in detecting different levels of damages. Our contribution to the field is the development of an integrated audio-visual-based disaster detection and management system that can detect and alert people in the affected areas. Moreover, we also explored the use of SSL in disaster management to improve the accuracy of our audio-based disaster detection system. We used the wav2vec 2.0 algorithm to pre-train our model on a large corpus of unlabelled audio data. Our SSL approach showed promising results, demonstrating the potential for using SSL in disaster management. However, we acknowledge that there are still some limitations to our research. One of the main challenges we faced was the availability of primary datasets for disaster sounds. We had to spend significant time and effort to collect and annotate suitable datasets, which impacted the accuracy of our model.

Overall, our thesis work has made significant contributions to the field of fire-disaster

management, particularly in the areas of fire detection, dataset collection, and mapping. Our proposed system offers a reliable and efficient solution for early fire detection, improving the safety of both humans and the environment. Our dataset collection efforts highlight the importance of high-quality and diverse datasets for developing effective disaster response systems. Finally, our use of PIX4D Mapper for disaster mapping offers an accurate and efficient way to visualize the extent of damage, enabling emergency responders to make informed decisions about response efforts.

8.3 Limitations

As with any research project, our thesis also has some limitations that need to be acknowledged. In this section, we will discuss the limitations of our thesis. One of the major limitations of our thesis is that it is dependent on the quality and availability of primary datasets. We faced challenges in finding appropriate datasets for our study, and the quality of the datasets we found was not always adequate. This can have a significant impact on the accuracy of our model. As reported by Akhtar and Mian (2018), the performance of object detection models is heavily dependent on the quality of the training data, and the use of low-quality data can result in poor model performance. Another limitation of our study is that the fire detection model is dependent on the type of fire and the environmental conditions. Our model is designed to detect fires in open areas, and it may not be effective in detecting fires in enclosed spaces or in areas with heavy smoke. As reported by Schubert et al. (2019), the accuracy of fire detection models can be impacted by various factors, such as the type of fire, the intensity of the flames, and the presence of smoke. Moreover, our study has limitations in terms of the type of damage that can be detected by the building damage assessment model. The model is designed to detect damage to buildings made of bricks and concrete, but it may not be effective in detecting damage to buildings made of other materials, such as wood, tin, or bamboo. This is a common limitation of building damage assessment models, as reported by Chen et al. (2017). The accuracy of the mapping model is dependent on the quality of the input data. The accuracy of the model can be impacted by factors such as the resolution of the images, the quality of the GPS data, and the presence of obstructions such as trees or buildings. As reported by Tong et al. (2019), the accuracy of mapping models can be improved by using higher-resolution images and incorporating more accurate GPS data. Finally, our study has limitations in terms of the scalability of the models. While our models have shown promising results in detecting fires, assessing building damage, and mapping disaster-affected areas, they may not be easily scalable to larger areas or to real-time disaster situations. As reported by Kanawong et al. (2019), the scalability of disaster response models is a significant challenge, and there is a need for further research in this area.

In conclusion, our thesis has demonstrated promising results in detecting fires, assessing building damage, and mapping disaster-affected areas. However, it is important to acknowledge the limitations of our study, including the dependence on the quality and availability of primary datasets, the limitations in terms of the type of damage that can be detected, the impact of environmental factors on the accuracy

of the fire detection model, the dependence on the quality of input data for the mapping model, and the scalability of the models. We believe that addressing these limitations will be critical for further improving the effectiveness of our models in disaster response situations.

8.4 Future Work

As we continue to improve our thesis model, we have identified several areas of future work that we plan to explore.

One area of focus for future work is enhancing the accuracy and speed of our disaster detection model. While our current model has shown promising results, we believe that there is still room for improvement. We plan to explore the use of other object detection models, such as the CenterNet model, to see if they can provide more accurate results in less time (Zhou et al., 2020). Another area of future work is to expand the scope of our disaster detection model to include other types of disasters, such as floods and landslides. While our current model focuses on building collapses and damage, we believe that there is potential for it to be adapted to detect other types of disasters as well. We plan to explore the use of other types of sensors, such as LIDAR and radar, to detect changes in terrain that may indicate a flood or landslide (Narula et al., 2020).

Additionally, we plan to further explore the use of drones in disaster response. While our current model utilizes drones for data collection and disaster mapping, we believe that there is potential for drones to be used in other ways as well, such as delivering medical supplies and food to areas affected by disasters. We plan to explore the use of machine learning algorithms to optimize drone flight paths and improve the efficiency of drone-based disaster response efforts (Zhao et al., 2020). Finally, we also plan to address some of the limitations of our current model in future work. For example, we acknowledge that our current model is not able to detect damage on certain types of buildings, such as those made of wood or bamboo. We plan to explore the use of other types of sensors, such as thermal imaging cameras, to detect changes in temperature that may indicate a fire or other type of disaster (Wang et al., 2020).

In conclusion, we believe that our thesis has shown promising results in the field of disaster response and detection. However, we also acknowledge that there is still much work to be done in order to further improve the accuracy and efficiency of our model. By exploring the areas of future work outlined above, we hope to make significant strides in improving disaster response efforts and ultimately saving lives.

Bibliography

- [1] W. S. Mseddi, R. Ghali, M. Jmal, and R. Attia, *Fire detection and segmentation using yolov5 and u-net: Semantic scholar*, Jan. 1970. [Online]. Available: <https://www.semanticscholar.org/paper/Fire-Detection-and-Segmentation-using-YOLOv5-and-Mseddi-Ghali/4da6acaa0a6922c78df0c555952aab363ed5bb9b>.
- [2] S. Stramondo, C. Bignami, M. Chini, N. Pierdicca, and A. Tertulliani, "Satellite radar and optical remote sensing for earthquake damage detection: Results from different case studies," *International Journal of Remote Sensing*, vol. 27, no. 20, pp. 4433–4447, 2006.
- [3] M. Touse, J. Sinibaldi, and G. Karunasiri, "Mems directional sound sensor with simultaneous detection of two frequency bands," in *SENSORS, 2010 IEEE*, IEEE, 2010, pp. 2422–2425.
- [4] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [5] J. Tuma, P. Janecka, M. Vala, and L. Richter, "Sound source localization," in *Proceedings of the 13th International Carpathian Control Conference (ICCC)*, IEEE, 2012, pp. 740–743.
- [6] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, pp. 1–11, 2013.
- [7] A. l. o. o. panelSoujanyaPoriaaErikCambriabPersonNewtonHowardcGuang-BinHuangdAmirHussaina, SoujanyaPoriaa, a, *et al.*, *Fusing audio, visual and textual clues for sentiment analysis from multimodal content*, Aug. 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0925231215011297>.
- [8] A. l. o. o. panelSoujanyaPoriaaErikCambriabPersonNewtonHowardcGuang-BinHuangdAmirHussaina, SoujanyaPoriaa, a, *et al.*, *Fusing audio, visual and textual clues for sentiment analysis from multimodal content*, Aug. 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0925231215011297>.
- [9] M. Chen, S. Yang, X. Yi, and D. Wu, "Real-time 3d mapping using a 2d laser scanner and imu-aided visual slam," in *2017 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, IEEE, 2017, pp. 297–302.

- [10] L. D. I. o. Automation, L. Dong, I. o. Automation, *et al.*, *Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018. [Online]. Available: <https://dl.acm.org/doi/10.1109/ICASSP.2018.8462506>.
- [11] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement*, Apr. 2018. [Online]. Available: <https://arxiv.org/abs/1804.02767>.
- [12] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
- [13] V. V, Y. S, and V. B, *3d mapping using lidar*, Apr. 2018. [Online]. Available: <https://www.ijert.org/3d-mapping-using-lidar>.
- [14] E. Goceri, “Challenges and recent solutions for image segmentation in the era of deep learning,” in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 2019, pp. 1–6.
- [15] Z. Kalam, *Interactive chart: Fire incidents on rise in bangladesh*, Mar. 2019. [Online]. Available: <https://www.thedailystar.net/country/news/interactive-chart-fire-incidents-rise-bangladesh-1722880>.
- [16] Z. Kalam, *Interactive chart: Fire incidents on rise in bangladesh*, Mar. 2019. [Online]. Available: <https://www.thedailystar.net/country/news/interactive-chart-fire-incidents-rise-bangladesh-1722880>.
- [17] M. Khairullah, “A novel steganography method using transliteration of bengali text,” *Journal of King Saud University-Computer and Information Sciences*, vol. 31, no. 3, pp. 348–366, 2019.
- [18] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, *Building damage detection in satellite imagery using convolutional neural networks*, Oct. 2019. [Online]. Available: <https://arxiv.org/abs/1910.06444>.
- [19] D. Basu, U. Ghosh, and R. Datta, *Adaptive control plane load balancing in vsdn enabled 5g network*, Jul. 2020. [Online]. Available: <https://arxiv.org/abs/2007.09789>.
- [20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” *arXiv preprint arXiv:2004.10934*, 2020.
- [21] C. Z. T. Inc., C. Zhang, T. Inc., *et al.*, *Model size reduction using frequency based double hashing for recommender systems: Proceedings of the 14th ACM conference on recommender systems*, Sep. 2020. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3383313.3412227>.
- [22] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial intelligence review*, vol. 53, pp. 5455–5516, 2020.
- [23] J. Baek, T. J. Alhindi, Y.-S. Jeong, *et al.*, “Real-time fire detection algorithm based on support vector machine with dynamic time warping kernel function,” *Fire Technology*, pp. 1–25, 2021.
- [24] W. Budiharto, E. Irwansyah, J. S. Suroso, A. Chowanda, H. Ngarianto, and A. A. S. Gunawan, Mar. 2021. [Online]. Available: <https://doi.org/10.1186/s40537-021-00436-8>.

- [25] S. S. Matin and B. Pradhan, *Earthquake-induced building-damage mapping using explainable ai (xai)*, Jun. 2021. [Online]. Available: <https://doi.org/10.3390/s21134489>.
- [26] D. Shin, S. Grover, K. Holstein, and A. Perer, *Characterizing human explanation strategies to inform the design of explainable ai for building damage assessment*, Nov. 2021. [Online]. Available: <https://arxiv.org/abs/2111.02626>.
- [27] T. Ahmad, M. Cavazza, Y. Matsuo, and H. Prendinger, *Detecting human actions in drone images using yolov5 and stochastic gradient boosting*, Sep. 2022. [Online]. Available: <https://doi.org/10.3390/s22187020>.
- [28] G. Bae, I. Budvytis, and R. Cipolla, *Multi-view depth estimation by fusing single-view depth probability with multi-view geometry*, Mar. 2022. [Online]. Available: <https://arxiv.org/abs/2112.08177>.
- [29] N. Dilshad, A. Ullah, J. Kim, and J. Seo, "Locateuav: Unmanned aerial vehicle location estimation via contextual analysis in an iot environment," *IEEE Internet of Things Journal*, 2022.
- [30] H. Du, W. Zhu, K. Peng, and W. Li, "Improved high speed flame detection method based on yolov7," *Open Journal of Applied Sciences*, vol. 12, no. 12, pp. 2004–2018, 2022.
- [31] H. Du, W. Zhu, K. Peng, and W. Li, *Improved high speed flame detection method based on yolov7*, Dec. 2022. [Online]. Available: <https://doi.org/10.4236/ojapps.2022.1212140>.
- [32] *Fire disaster management process: Free essay examples*, Mar. 2022. [Online]. Available: <https://samples.freshessays.com/fire-disaster-management-process.html>.
- [33] T. Kustu and A. Taskin, *Deep learning and stereo vision based detection of post-earthquake fire geolocation for smart cities within the scope of disaster management: İstanbul case*, Dec. 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4295392.
- [34] A. J. Mantau, I. W. Widayat, J.-S. Leu, and M. Köppen, *A human-detection method based on yolov5 and transfer learning using thermal image data from uav perspective for surveillance system*, Oct. 2022. [Online]. Available: <https://doi.org/10.3390/drones6100290>.
- [35] H.-C. Nguyen, T.-H. Nguyen, R. Scherer, and V.-H. Le, *Unified end-to-end yolov5-hr-tcm framework for automatic 2d/3d human pose estimation for real-time applications*, Jul. 2022. [Online]. Available: <https://doi.org/10.3390/s22145419>.
- [36] T. R. Subramaniam, N. N. A. Suhaimi, A. Paizol, and A. H. M. Nor, "Determination of slope stability using (uav) unmanned aerial vehicle," *Multidisciplinary Applied Research and Innovation*, vol. 3, no. 2, pp. 292–301, 2022.
- [37] N. Takhtkeshha, A. Mohammadzadeh, and B. Salehi, *A rapid self-supervised deep-learning-based method for post-earthquake damage detection using uav data (case study: Sarpol-e zahab, iran)*, Dec. 2022. [Online]. Available: <https://doi.org/10.3390/rs15010123>.

- [38] *The worst industrial disasters in bangladesh since 2005*, Jun. 2022. [Online]. Available: <https://www.dhakatribune.com/bangladesh/2022/06/05/the-worst-industrial-disasters-in-bangladesh-since-2005>.
- [39] Z. Wu, R. Xue, and H. Li, “Real-time video fire detection via modified yolov5 network model,” *Fire Technology*, vol. 58, no. 4, pp. 2377–2403, 2022.
- [40] Z. Wu, R. Xue, and H. Li, *Real-time video fire detection via modified yolov5 network model - fire technology*, May 2022. [Online]. Available: <https://link.springer.com/article/10.1007/s10694-022-01260-z>.
- [41] H. Xu, B. Li, and F. Zhong, *Light-yolov5: A lightweight algorithm for improved yolov5 in complex fire scenarios*, Dec. 2022. [Online]. Available: <https://doi.org/10.3390/app122312312>.
- [42] S. G. Zhang, F. Zhang, Y. Ding, and Y. Li, *Swin-yolov5: Research and application of fire and smoke detection algorithm based on yolov5*, Jun. 2022. [Online]. Available: <https://doi.org/10.1155/2022/6081680>.
- [43] [Online]. Available: <http://www1.cs.columbia.edu/~mccollins/6864/slides/asr.pdf>.
- [44] *285,000 fire incidents in bangladesh in two decades*. [Online]. Available: <https://news.priyo.com/e/3663051-285000-fire-incidents-in-Bangladesh-in-two-decades>.
- [45] *Spoken language processing:a guide to theory, algorithm, and system development*. [Online]. Available: <https://dl.acm.org/doi/10.5555/560905>.
- [46] *Spoken language processing:a guide to theory, algorithm, and system development*. [Online]. Available: <https://dl.acm.org/doi/10.5555/560905>.
- [47] D. Sun, *Bahrain executes two shias on terror charges: Daily sun* —. [Online]. Available: <https://www.daily-sun.com/post/410978/Fire-incidents-in-Bangladesh>.