1 INTRODUCTION

In this research project, the author tries to find a novel practical approach to automate the data augmentation pipeline in image classification tasks. The contents of this document consist of the problem, research gap analysis, research challenges, and research strategy that the author wishes to follow over the next few months. Furthermore, the necessary evidence of the problem and previous research interests are also discussed. Finally, the author describes the project's deliverables in the work plan.

2 PROBLEM DOMAIN

Deep convolutional neural networks have succeeded remarkably in many computer vision tasks, such as image classification, image segmentation, and object detection. These networks heavily rely on big data to avoid model overfitting, reducing the model's generalization performance. Improving the generalization ability of modern convolutional neural network models is one of the most difficult challenges because many real-world application domains, such as medical image analysis and bioinformatics, do not have access to big data, and collecting data in such areas is well-known to be costly and labor-intensive (Nanni et al., 2021; Shorten and Khoshgoftaar, 2019).

The term "generalizability" refers to the performance difference of a model when evaluated on already seen data (known as "training data") versus data it has never seen before (known as "testing data"). Models with poor generalizability have overfitted the training data (Shorten and Khoshgoftaar, 2019). The following graphs show what overfitting looks like when visualizing training and validation accuracy over time.

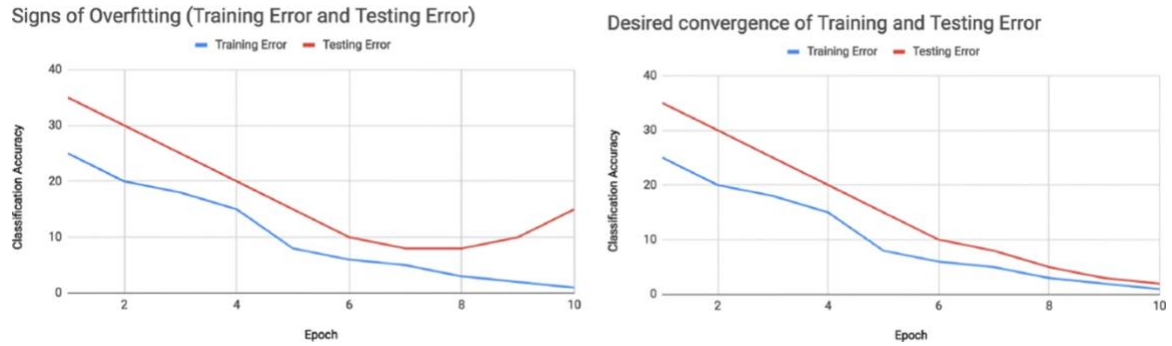

Figure 1: Visualization of model accuracies over training epochs (Shorten and Khoshgoftaar, 2019)

To develop useful deep-learning models, the validation error must continue to decrease with the training error (Shorten and Khoshgoftaar, 2019).

2.1 Data Augmentation

Data augmentation is a data space solution that strikes the heart of the problem of overfitting and improves model generalizability. Data augmentation aims to increase the amount of information extracted from the original dataset by artificially increasing its size through data warping or oversampling. Data warping augmentations transform existing images, so their label is preserved, and oversampling augmentations create synthetic instances and add them to the training set. The augmented data will represent a wider variety of possible data points, thus minimizing the distance between the training and validation sets, which leads to overcoming overfitting and improving model generalizability (Nanni et al., 2021; Shorten and Khoshgoftaar, 2019).

According to Shorten and Khoshgoftaar, image data augmentation, the focus of this research project, can divide into two major categories. Basic image manipulations (such as geometric transformations, color space transformations, kernel filters, random erasing and mixing images) and deep learning approaches (such as adversarial training, neural style transfer, and generative adversarial network-based data augmentation, also known as GAN).

2.2 Automated Data Augmentation

In the image domain, a data augmentation policy refers to a set of image operations used to transform the image data. Data augmentation policies for computer vision tasks have been designed manually, and the best-performing augmentation policies are dataset-specific. For example, on the MNIST dataset, most top-ranked models use elastic distortions, rotation, translation, and scale. On natural image datasets, such as CIFAR-10/CIFAR-100 and ImageNet, image mirroring, random cropping, and color shifting/whitening are more common (Yang et al., 2022). These policies are designed manually, requiring expert knowledge and time, which is highly subjective and error-prone. The traditional trial-and-error approach based on training loss or accuracy can result in extensive, redundant data collection, wasting computational resources and efforts (Yang et al., 2022). To mitigate this problem, a novel direction is to automatically learn the image augmentation policies from the given dataset using automated data augmentation (AutoDA) techniques. AutoDA models aim to find the best data

augmentation policies to maximize model performance gains. Recent research has shown that instead of manually designing the data augmentation policies, directly learning data augmentation policies from the target dataset can significantly improve model performance (Cubuk et al., 2019a).

3 PROBLEM DEFINITION

Data augmentation is an effective technique to avoid overfitting and increasing the generalizability of modern neural network models, which has been widely used in computer vision-based tasks by researchers, students, and industries. However, current data augmentation implementations for computer vision tasks are manually designed. As these methods are developed manually, they require expert knowledge, time, and resources to select the best-performing data augmentation methods for a given dataset. So far, a significant focus of the machine learning and computer vision community has been on engineering better deep neural network architectures. Less attention has been paid to finding better data augmentation methods based on the dataset and nature of the given task (Cubuk et al., 2019a). Motivated by progress in automated machine learning (AutoML), the need for automatically learned data augmentation has been raised recently as an important unsolved problem (Cubuk et al., 2019a).

3.1 Problem Statement

Data augmentation is an essential technique for solving data problems and has been widely used in various computer vision tasks, but it isn't easy to select the optimal augmentation policy when given a specific dataset.

4 RESEARCH MOTIVATION

As mentioned in the above sections, data augmentation is a fundamental task of the computer vision domain. It has been widely used in computer vision-based tasks by researchers, students, and industries. However, it has been proven that standard data augmentation techniques still contain a lot of manual work and need a decent amount of data augmentation domain expertise to achieve the best results. Automating the process of designing data augmentation techniques is a solution for all these shortcomings and will be a huge turning point in the computer vision domain.

5 RELATED WORK

Citation	Summary	Contribution	Limitations
(Cubuk et al., 2019a)	Provide a standard problem formulation to the AutoDA field using Reinforcement Learning (RL) as a key technique and Recurrent Neural Network (RNN) as a policy optimizer.	First Implementation	Computationally infeasible to run for the ordinary user (requires 15000 NVIDIA TITAN 100 GPU hours on the ImageNet dataset)
(Ho et al., 2019)	Improved version of the first implementation (Cubuk et al., 2019a). Used Population Based Training as a key technique and Truncation selection as a policy optimizer.	Reduced the optimal DA policy search time and complexity of the first implementation (Cubuk et al., 2019a).	The optimal DA policy searching and evaluation phase is computationally expensive and time-consuming.
(Lim et al., 2019)	Improved version of the first implementation (Cubuk et al., 2019a). Used Reinforcement Learning (RL) along with density matching as key techniques and Bayesian	Reduced the optimal DA policy search time and complexity of the first implementation (Cubuk et al., 2019a).	The optimal DA policy searching and evaluation phase is computationally expensive and time-consuming.

	Optimization as policy optimizer.		
(Cubuk et al., 2019b)	Proposed a new approach to the AutoDA field by search space (where DA policy consists of) reparameterization and use Grid search as policy optimizer.	Reduced the cost of the search phase. Achieve competitive state-of-art compared to other AutoDA solutions, even in reduced search space.	The optimal DA policy searching and evaluation phase is computationally expensive and time-consuming.
(Li et al., 2020)	A proposed gradient-based approach using an Unbiased gradient estimator as a key technique and Stochastic gradient descent as a policy optimizer.	Reduced search time while achieving competitive accuracy.	The high complexity of implementation.

Table 1: Related Work

6 RESEARCH GAP

After thoroughly reviewing the existing literature, the author identified several limitations of current research works, summarized below.

The major drawback of existing automated data augmentation techniques is that they are highly resource-intensive. For instance, Auto Augment (Cubuk et al., 2019a) requires 15000 NVIDIA Tesla P100 GPU hours on the ImageNet dataset to perform automated image data augmentation tasks. This will limit the widespread applicability of Auto Augment (Cubuk et al., 2019a) in real-world, commercial applications. Hence, it is essential to solving this major drawback of high resource utilization in existing automated data augmentation techniques to use automated data augmentation techniques in practice.

Furthermore, none of the existing approaches considers the class imbalance issue while generating new data points. Class imbalance is a problem that occurs when there is a major imbalance between the number of data points in minority and majority classes (Shorten and Khoshgoftaar, 2019). The imbalanced dataset causes the model to overfit. So, it is essential to generate well-balanced datasets (Yang et al., 2022).

Finally, none of the existing approaches gives the ability to toggle a specific type of data augmentation transformation in different application scenarios. For instance, when a user wants to remove some image data augmentation operations from the search space that are specific to their research work, the ability to toggle a particular type of data augmentation from the search space is essential. Moreover, this allows the obtained augmentation policy to be more tailored to the given dataset and task (Yang et al., 2022).

7 RESEARCH CONTRIBUTION

As mentioned in previous sections, automated image data augmentation is promising because it allows searching for more powerful compositions of image transformations and parameterizations, increasing the generalizability of modern image classifiers.

The biggest challenge with automating data augmentation is to search over the space of image data augmentation policies (compositions of image transformations). However, due to a large number of image data augmentation policies and associated parameters in the search space, this can be prohibitively and computationally expensive by limiting the commercialization of automated data augmentation (Yang et al., 2022).

The main contribution of this research project is to develop a novel computationally affordable framework that explores the search space of image data augmentation policies efficiently and effectively using less computational power, which also finds image augmentation strategies that can outperform human-designed image data augmentation policies. As a result, the proposed novel system will be able to expand the broad applicability of automated image data augmentation techniques in real-world applications.

Due to the nature of this research, it must be noted that contributions to both the problem domain and research domain are similar in this project.

8 RESEARCH CHALLENGES

The main goal of this research project is to address the limitations of the literature and enhance the wide usage of automated data augmentation techniques in real-world applications. Based on the proposed approach, the research challenges can be listed as follows.

1. **Improving efficiency and effectiveness** – Automating data augmentation for image classification tasks is a relatively new concept and has not been fully addressed for real-world applications (Yang et al., 2022). Therefore, it is essential to explore theoretical aspects and other ways of defining automated data augmentation architectures more efficiently and effectively is a challenge.
2. **Improving the accuracy of the image classification model** - The usage of automated data augmentation techniques in practice is limited due to efficiency-related drawbacks. Recent works in automated data augmentation have improved the efficiency of automated data augmentation models, but their accuracy remains a bottleneck (Yang et al., 2022). Model accuracy is the key factor of modern image classifiers (Shorten and Khoshgoftaar, 2019). Thus, improving efficiency while maintaining competitive accuracy will be a challenge.
3. **Data augmentation for sensitive tasks** - With reduced human intervention, generating accurately labeled new data points for sensitive tasks like medical image analysis and bioinformatics can be very difficult to obtain, especially when the dataset is imbalanced and noisy. So, dealing with the sensitive task with an imbalanced and noisy dataset will be another challenge.

9 RESEARCH QUESTIONS

RQ1: What are the latest advancements in the Software Engineering world that can be used to perform automated data augmentation in a practical manner?

RQ2: How to design an automated data augmentation system that can outperform human-designed data augmentation heuristics?

RQ3: What are the potential challenges that may occur while designing an automated data augmentation system, and how to avoid and overcome them?

10 RESEARCH AIM

The aim of this research is to design, develop and evaluate a system that automates the manual process of designing and fine-tuning image data augmentation policies for any given low-diversity dataset with reduced human intervention.

To further elaborate on the aim, this project will produce a system capable of performing data augmentation for a given low-diversity dataset. To achieve that, this system aims to select the best-performing data augmentation policies based on the given low-diversity dataset and will tune the magnitude of chosen policies to improve the diversity of the given dataset. Furthermore, the system and its components will be tested and evaluated against output quality to validate the hypothesis chosen.

11 RESEARCH OBJECTIVES

Objective	Description	Learning Outcomes	Research Questions
Literature Review	To fulfill the following requirements, a survey of the existing literature is conducted. **RO1:** To analyze the existing systems in the AutoDA domain. **RO2:** To identify the limitations, improvements, and research gaps in the problem. **RO3:** To identify the ways of reducing computational power and complexity withholding accuracy levels **RO4:** To identify the technologies, algorithms, frameworks, and other tools required for the development phase of the project. **RO5:** To identify the evaluation criteria and metrics.	LO1, LO3, LO6	RQ1, RQ2
Requirement Analysis	In-depth requirement analysis was performed to,	LO2, LO3, LO5	RQ1, RQ2, RQ3

	RO6: To determine the awareness of the risk of using random DA techniques to perform data augmentation. **RO7:** To gather requirements of an AutoDA framework and understand end-user expectations. **RO8:** To gather academic and industry experts' insights and feedback to improve the proposed system.		
Design	Design the architecture of the proposed AutoDA framework, **RO9:** To design search space where image data augmentation policy consists of. **RO10:** To design DA policy search and evaluation algorithms. **RO11:** To design a method to perform augmentation using identified DA policies.	LO1, LO5	RQ1, RQ2
Development	Developing the proposed AutoDA framework according to the identified design aspects, software and hardware requirements, **RO12:** To develop search space, search algorithm, and evaluation algorithm of the proposed framework. **RO13:** To develop the GUI of the proposed framework. **RO15:** To develop core functionalities of the proposed framework using appropriate software and hardware requirements.	LO1, LO5, LO7	RQ1, RQ2
Testing and Evaluation	Testing and evaluating the proposed AutoDA methods' performance, **RO16:** To create a suitable test plan for functional and unit testing. **RO17:** To benchmark the prototype against accuracy and performance aspects.	LO1, LO7, LO8, LO9	RQ1, RQ2

	RO18: To get feedback from industry and academic experts.		
Publish Findings	To critically evaluate the research, **RO19:** Publish review paper on related works. **RO20:** Publish evaluation and testing results identified.	LO8, LO9	RQ1, RQ2, RQ3

Table 2: Research Objectives

12 PROJECT SCOPE

Based on the project objectives and a review of existing works, the scope is defined as follows, considering the allotted time for this research project.

12.1 In-scope

The scope of the project is as follows:

- An algorithm for preprocessing and preparing any given dataset to work with the proposed system. Existing algorithms will be evaluated and improved at this stage.
- Developing a search space where image data augmentation policies consist of.
- A method to search and evaluate best-fitting data augmentation policies for any dataset.
- A method to perform data augmentation using learned data augmentation policies from any given dataset. Existing algorithms will be evaluated and improved at this stage.
- Web application for presenting and evaluating the proposed AutoDA approach.

12.2 Out-scope

The following are the parts of the project that will not be covered:

- The proposed system will only support image data augmentation in the initial phase.
- The proposed system will only support automating basic image transform data augmentation techniques at the initial phase.
- The proposed framework will not consider transmitting learned data augmentation policies to the other datasets.

12.3 Prototype Diagram

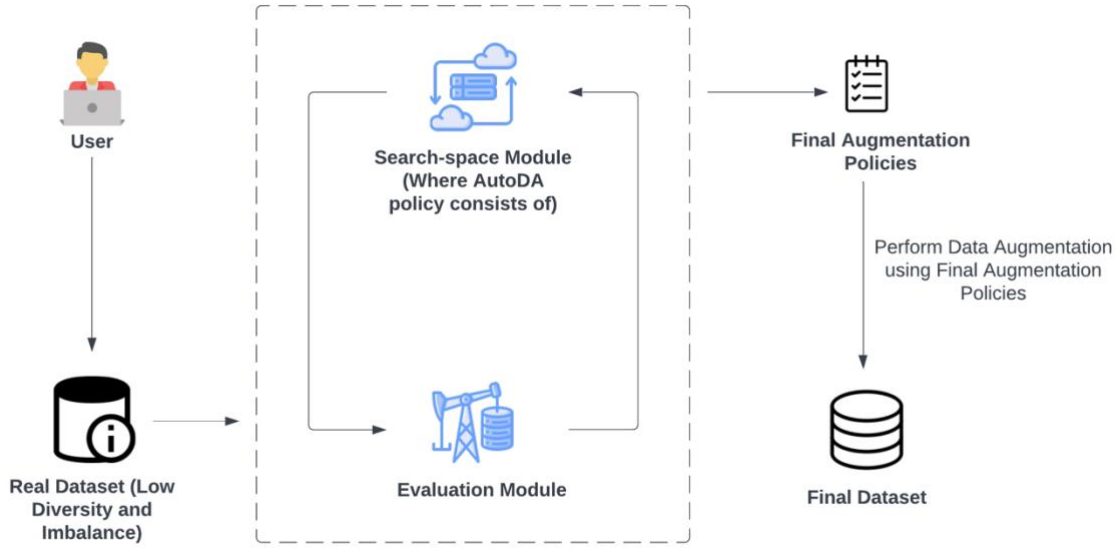

Figure 2: Prototype feature diagram of the research (Self-Composed)

13 PROPOSED METHODOLOGY

13.1 Research Methodology

The quality of any research project is determined by three key factors: scope, cost, and time, all of which must be managed effectively throughout the project's lifespan. Therefore, methodologies are required. The research methodologies for this project were chosen from the predefined Saunders Research Onion Model.

Research Philosophy	Among the pragmatism, positivism, interpretivism, and realism approaches, the **pragmatism** approach was chosen because the author will evaluate and experiment with various methodologies as a combined approach to determine which performs well for achieving the research goal.
Research Approach	From available research approaches, inductive and deductive, the **deductive** approach was chosen because the research aims to extensively apply a combination of existing theories.
Research Strategies	The research strategy refers to how you develop the methodology to answer the research questions. **Surveys**, **experiments**, and **interviews** based on evaluation metrics will be used among the possible candidates.
Research Choice	Among the available options, mono, multi methods, and mixed, the **mixed method** was chosen because, for this research, both qualitative (interviews)

	and quantitative (surveys) strategies were chosen to complement one another.
Time Horizon	The data will be gathered in a single point of this research project (during the evaluation phase). So, among the longitudinal and cross-sectional methods, a **cross-sectional** method was chosen.
Techniques and Procedures	For data gathering and analysis, surveys, interviews, trial and error observations, similar solutions, and literature will be used.

Table 3: Research Methodology

13.2 Development Methodology

Among the many available software lifecycle methodologies, the **prototype** method was chosen because the project will be designed, built, and evaluated until a successful outcome is achieved.

13.2.1 Design Methodology

Since this research focuses on implementing a novel practical AutoDA approach, the author chose **Structured Systems Analysis & Design Method** (SSADM) as the design methodology from the possible approaches of OODA (Object Oriented Design & Analysis) and SSADM.

13.2.2 Evaluation Methodology

A research study relies heavily on evaluation, and proper evaluation highlights the credibility of the research study. The author chose both evaluation metrics and benchmarking evaluation approaches to evaluate the proposed AutoDA framework.

13.2.2.1 Evaluation Metrics

The author chose evaluation metrics from the image classification domain for the quantitative evaluation of the research. They are listed below.

- Classification accuracy
- Precision and recall
- F1 score
- ROC curve and AUC

13.2.3 Benchmarking

Previous research on AutoDA used CIFAR10, CIFAR100, ImageNet-C, and ImageNet datasets for benchmarking. Because of limited resources, the author cannot conduct a

benchmark on ImageNet-C and ImageNet datasets. However, the author will perform a benchmarking for the proposed system using other available datasets.

13.2.4 Solution Methodology

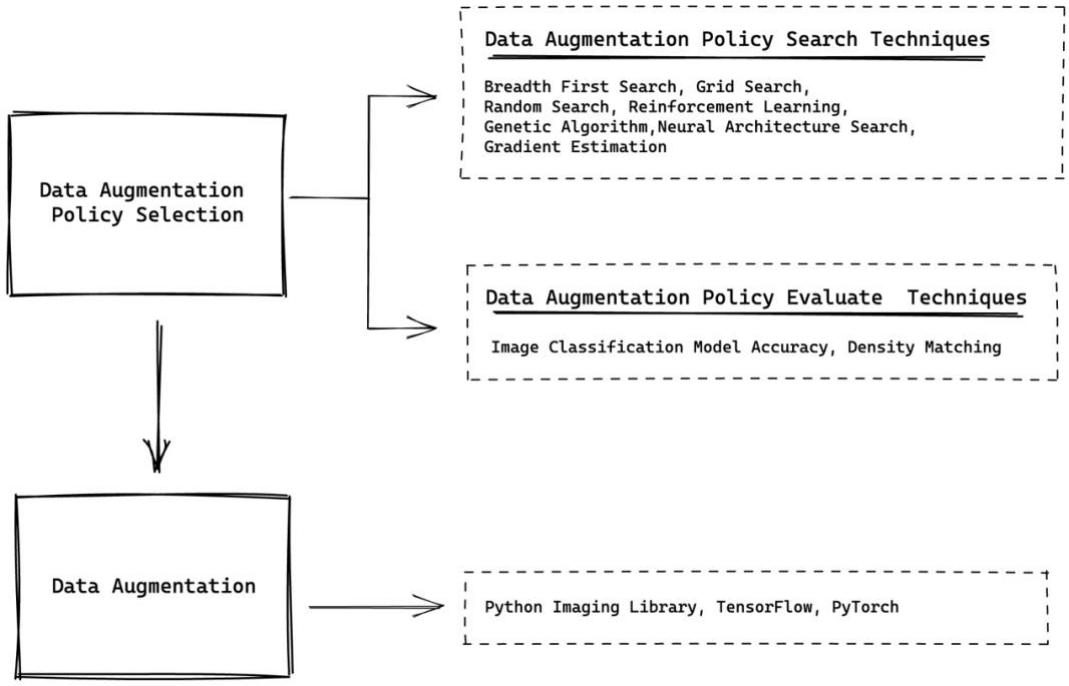

Figure 3: General architecture of AutoDA system (Self-Composed)

Based on the past literature, automated data augmentation systems' general architecture can be divided into two major components.

1. **Data augmentation policy selection** – This component is supposed to select the best-performing data augmentation policies for the given dataset. The search techniques mentioned in the above diagram will be used to find the best-performing data augmentation policies from the search space. Later the selected data augmentation policies will be sampled and ranked using the evaluation techniques mentioned in the above architecture diagram.
2. **Data augmentation using chosen policies** - Top-ranked data augmentation policies from the first component will be used to perform the data augmentation for the given dataset. Existing libraries like the python imaging library (PIL), TensorFlow, and PyTorch will be used in this component.

13.3 Project Management Methodology

Out of the available project management methodologies, the hybrid model **Agile Prince-2** was chosen. Because the author of this research project is a sole developer, and since numerous interim deliverables are expected, the Agile Prince-2 method is ideal.

13.3.1 Resource Requirements

Based on the aforementioned objectives and functionalities, the identified necessary hardware, software, data, and skills to complete this project are as follows.

13.3.1.1 Hardware Requirements

- **Apple M1 Pro (8-core CPU, 14-core GPU, 16-core Neural Engine) processor or above -** To run DA policy search and evaluation engine and perform data augmentation.
- **16GB RAM or above -** To manage large volumes of data.
- **Disk space of 50GB or more -** To store necessary data and application code.

If hardware resources are not sufficient, and since Apple M1 processors are not compatible with all ML/DL frameworks yet, propose using Google Colab free tire version.

13.3.1.2 Software Requirements

- **Operating System -** To run essential programs such as IDEs and other required tools.
- **Python** - The primary programing language used to develop the proposed system.
- **Jupiter NoteBook/PyCharm IDE -** To make the development process more efficient.
- **React Native -** React Native will be used to develop the proposed system's front end.
- **Zotero -** To organize, manage, and back up the relevant research papers.
- **MS Office -** To create the documents which are related to the project.
- **Google Drive -** To back up the project-related files.
- **Git/GitHub -** To version control and backup the application code.

13.3.1.3 Data Requirements

- **Image datasets -** To test and evaluate the proposed system.

13.3.1.4 Skill Requirements

- Knowledge of image data augmentation techniques.
- Knowledge of search and evaluation algorithms.
- Designing skills.
- Creative writing skills.

13.3.2 Risk Management

The risks identified prior to the start of the project are listed below, along with possible mitigation steps.

Risk Item	Severity	Frequency	Mitigation Plan

Lack of domain knowledge	5	3	Seek advice from academic and domain experts, use Stack Overflow, and follow online tutorials and courses.
Insufficient hardware resources	5	4	Use cloud-based solutions such as Google Colab free tier version
Losing ongoing development code	5	2	Use GitHub and external backup of all code.
Losing documentation	5	2	Follow a cloud-first documentation approach
Any unpredictable risk (such as Covid 19, power interruptions, and natural disasters)	3	5	Manage the work on a timetable and try to accomplish daily or weekly goals.

Table 4: Associated Risks And Mitigation

13.4 Schedule

13.4.1 Deliverables

Deliverable	Date
Project Proposal Document - The initial proposal of the research project.	4th November 2022
Literature Review Document - The critical analysis of existing work and solutions.	11th December 2022
Software Requirement Specification - The requirements to be satisfied in the final research prototype.	15th December 2022
System Design Document - A document specifying the architecture of the proposed AutoDA framework based on identified techniques from the literature review.	1st December 2022
Prototype - Prototype of the research which integrates the proposed AutoDA framework.	1st February 2022
Thesis - The final research project documentation discusses the research process, decisions, and findings.	15th March 2022
Review Paper - A review paper reviews existing systems in the AutoDA domain.	1st March 2022

Project Research Paper - A research paper introducing the AutoDA system developed at the end of this project.	1st April 2022

Table 5: Deliverables and Dates

13.4.2 Gantt Chart

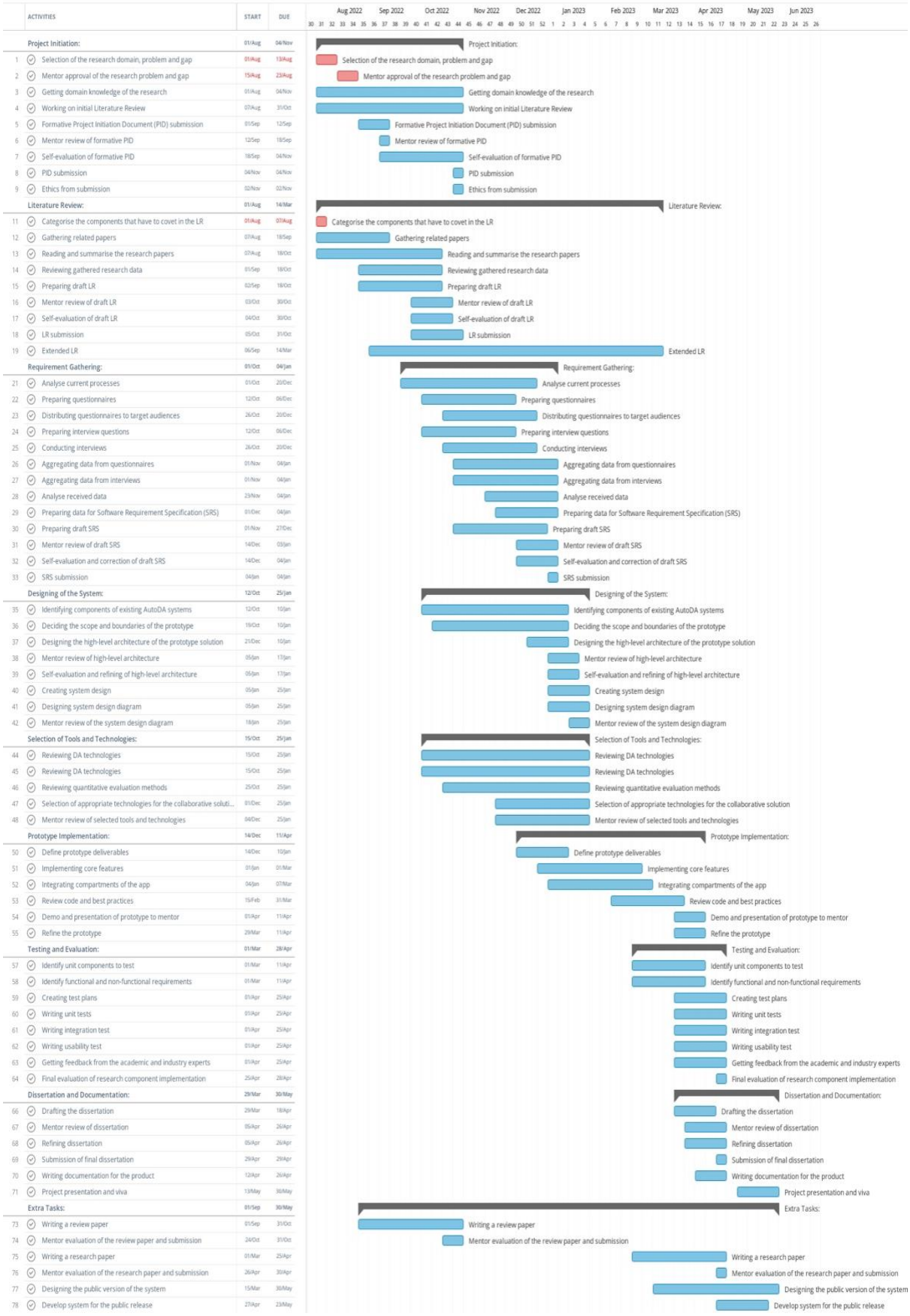

Figure 4: Gantt Chart

14 SUMMARY

This document presented the research project proposal to be conducted in the coming months. At first, an elaboration of the problem domain and problem definition was carried out. Following that, the existing works and the research challenges were thoroughly discussed. Following that, the research objectives and project scope were discussed, and finally, the proposed methodologies were elaborated on.

REFERENCES

Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V., 2019a. AutoAugment: Learning Augmentation Strategies From Data, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 113–123. https://doi.org/10.1109/CVPR.2019.00020

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2019b. RandAugment: Practical automated data augmentation with a reduced search space.

Ho, D., Liang, E., Stoica, I., Abbeel, P., Chen, X., 2019. Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules.

Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A., Seliya, N., 2018. A survey on addressing high-class imbalance in big data. J. Big Data 5, 42. https://doi.org/10.1186/s40537-018-0151-6

Li, P., Liu, X., Xie, X., 2021. Learning Sample-Specific Policies for Sequential Image Augmentation. undefined. https://doi.org/10.48550/arXiv.2205.01491

Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N.M., Yang, Y., 2020. DADA: Differentiable Automatic Data Augmentation.

Lim, S., Kim, I., Kim, T., Kim, C., Kim, S., 2019. Fast AutoAugment.

Nanni, L., Paci, M., Brahnam, S., Lumini, A., 2021. Comparison of Different Image Data Augmentation Approaches. J. Imaging 7, 254. https://doi.org/10.3390/jimaging7120254

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on Image Data Augmentation for Deep Learning. J. Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0

Yang, Z., Sinnott, R.O., Bailey, J., Ke, Q., 2022. A Survey of Automated Data Augmentation Algorithms for Deep Learning-based Image Classication Tasks.

Zhang, R., Liang, Y., Somayajula, S.A., Xie, P., 2021. Improving Differentiable Architecture Search with a Generative Model. https://doi.org/10.48550/arXiv.2112.00171

Zhang, X., Wang, Q., Zhang, J., Zhong, Z., 2019. Adversarial AutoAugment.