MONASH University
Engineering

Unlearning for Medical Image Classification

ENG4702: Final Year Project - Final Report

Author(s): Ibrahim Alfaidi(31276962), Zachary Wilson (31472958), Tom Liefman (29694183), Borui Li(31440398)

Supervisor(s): A/Prof. Mehrtash Tafazzoli Harandi

Date of Submission: 18 October 2024

Project type: Research

1 Executive Summary

This project investigates the integration of machine unlearning techniques into medical AI systems with the goal of enhancing data privacy and reducing model bias. We use a variety of unlearning methodologies to achieve the best results, including the implementation of Saliency Unlearning (SalUn). The motivation behind this research is to address the critical ethical challenges facing AI in healthcare, such as safeguarding patient data and ensuring equitable treatment across diverse patient groups.

SalUn is a state-of-the-art technique in machine unlearning that selectively removes sensitive or irrelevant information from AI models without compromising their diagnostic accuracy. This technique updates only the specific weights associated with the attributes to be forgotten, thereby preserving the overall utility of the model.

Our key findings reveal that even simple forgetting algorithms could result in models with comparable test accuracies to the base models. Saliency unlearning showed mild improvement at times, but was inconsistent. This variability highlights the technique's limited generalizability and underscores the larger challenge of deploying unlearning universally across all medical AI applications.

This itself would work as a stepping stone for future work to streamline the development of medical AI and offer solutions for protecting patients' privacy or bias in model output. This will unleash the full potential of AI in the medical field as this would tackle the privacy concern issue.

Through this work, we were able to establish a base effectiveness of machine unlearning for medical data, and display the future promise and capabilities of this work into the medical industry.

Overall, this research contributes to the ongoing discussion on ethical AI in healthcare by demonstrating the benefits and limitations of applying machine unlearning. It also suggests areas for further research and development, particularly in creating more resource efficient unlearning methods that could be widely implemented to ensure the ethical use of AI across all healthcare settings.

We wish to acknowledge the people of the Kulin Nations, on whose land Monash University Operates. We pay our respects to their Elders, past and present.

Contents

2 Introduction

The Integration of Artificial Intelligence (AI) within the medical field stands as a beacon of hope, primarily because of its remarkable ability to quickly analyze large amounts of data, especially in tasks such as medical image classification which can help with a patient's diagnosis. Nevertheless, the ethical ramifications of this breakthrough must be carefully considered. Foremost, one of the most important issues is how to prevent AI systems from learning about existing discrimination, biases, and social injustices that are embedded within their training data.

Given the highly sensitive nature of medical data, strict precautions must be taken to ensure its security when using AI models. Medical data privacy is extremely important and necessitates extra security measures to prevent breaches and unwanted access. Clinicians and researchers have ethical and legal obligations to patients, to ensure that diagnostic models exhibit both high accuracy and fairness across diverse patient cohorts.

This problem of biases from excessive learning of irrelevant, sensitive attributes during the classification process represents a challenge which can be tackled with unlearning methodologies. In the field of machine learning, unlearning serves as a potent tool which eliminates the influence of specific targets efficiently and effectively and removes associated model capabilities while preserving model accuracy. However, care should be taken while applying unlearning, as many existing Machine unlearning methods often suffer limitations in unlearning accuracy and stability, and introducing a limited unlearning method could potentially lead to an erroneous diagnosis.

To mitigate the risk of accuracy diminution, an innovative technique called Saliency Unlearning (SalUn) comes into consideration. This method, when applied to an unlearning method such as Random Labelling (RL), updates only the weights that are to be unlearned and leaves other weights intact. Because maintaining the model accuracy is imperative, we consider the weight saliency unlearning (SalUn) approach to be effective in the medical domain, due to its robustness against these limitations.

Our proposed project aims to address the biases and privacy issues that can occur from the excessive learning of irrelevant, sensitive attributes during the classification process. Specifically, we will apply the Saliency Unlearning (SalUn) technique, considered to be state-of-the-art within machine unlearning algorithms, and customize this algorithm to be applied to medical imagery for unlearning these attributes, while maintaining its high level of accuracy for its developed usage within natural imagery, aiming to enhance the reliability and integrity of AI in healthcare, fostering trust in AI-assisted diagnostics.

To maximize the impact and reach of our project, the best AI models can be provided to healthcare professionals upon completion. This will ensure that the benefits of Saliency Unlearning (SalUn) are immediately available in diverse medical settings. By delivering a tool tailored for ease of integration and effectiveness, our aim is to set a new standard in medical diagnostics that enhances patient outcomes while adhering to the highest privacy and ethical standards. Our commitment extends to continuous improvement and adaptation of the models to meet evolving medical standards and practices, thus ensuring that our innovations remain at the forefront of technological advancement in healthcare.

3 Aims and Objectives

3.1 Research Question

Our project aims to apply unlearning methodology to medical AI without sacrificing accuracy. In the field of medicine, people are generally more distrustful of AI, resulting in a high bar for acceptance: a phenomena well documented by Henrique and Santos [63].
To remedy this, unlearning methodology should be tried and tested on medical AI, to make way for a less biased, less rigid system that maintains performance. Thereby fostering greater confidence among healthcare providers and patients.

The Research Question: How can unlearning methods be effectively implemented in medical image analysis to enhance data privacy and prevent bias, while maintaining or improving performance?

This question explores the dual aspects of performance and privacy, areas often compromised in current medical AI applications. Through a systematic review of existing data handling techniques and their limitations, this project evaluates unlearning as a potential solution to these pressing challenges. The outcome will likely have substantial implications for the development of bias-free, patient-centric AI tools in healthcare.

3.2 Aims

3.2.1 Conduct unlearning method in medical diagnostic AI model

For this project, we aim to apply various existing models to public medical databases and modify effective machine-learning models to improve medical image classification across various medical imagery. We aim to utilise the Saliency Unlearning technique, considered a highly competitive and robust algorithm within existing AI unlearning especially around image classification. Effectively implementing this with medical imagery enables us to improve information privacy through feature unlearning, which involves removing specific information about a class from a model. Further, we have an ambitious goal of developing medical image analysis through implementing image segmentation across more complex uses, such as segmenting an image for pathology localization.

3.2.2 Enhance the performance of AI in healthcare

The extended aim of the research is to strengthen the performance and reliability of AI in healthcare, the team would research unlearning methods and their capacity to influence AI with different approaches, as well as discover any potential solutions in enhancing medical AI performance such as prediction accuracy by utilizing unlearning methods. Such potential implementation of unlearning methods could enable accurate and precise adjustment to Ai models which would pave the way to a greater breakthrough which will change the world irreversibly.

As solving the ethical concerns of AI would encourage others to use our developed model to diagnose and potentially detect early stage diseases which can’t be seen with the naked eye, as it will be enabled to analyze huge amounts of medical data to identify new disease patterns resulting in an early treatment and saving lifes.

3.3 Objectives

For our project, the objectives are relatively clear due to the specific nature of neural network learning and the use of artificial intelligence in image classification.

To state the objective explicitly, we want to obtain θ_R (retrained model) from the original model θ_o such that θ_{R1} does not make decisions based on sensitive features (x_s) by determining the "salient" weights related to those features, this is class-wise unlearning. Additionally we would like to be able to create another model θ_{R2} to compare the effectiveness of unlearning on a variety of image classification techniques.

Note that a conceptual timeline for our objectives is in section 7.2 This timeline is open to change as adversities are realised, for example, it could be discovered that SalUn is not a desirable approach.

Our 5 main objectives are as follows:

1. To find and investigate a medical AI model's collection of data.
2. To perform SalUn unlearning technique on a chosen model.
3. Develop software for streamlining the unlearning process.
4. To evaluate and optimize the efficiency of the model after SalUn.
5. Compare other approaches and implement image segmentation.

3.3.1 To find and investigate a medical AI model's collection of data

Our first objective is to find a suitable image classification model for medical image classification / diagnostics. We would like to investigate the model before making any changes, to understand the data which helped tune the model's weights, of which allow the model to identify various features and classes and balance these variables for classification, identify the ML algorithm and then identify any data which may contain private information and thus should be unlearned in the case of sample wise unlearning, or identify any features that should be unlearned in the case of class-wise unlearning.

3.3.2 To perform SalUn unlearning technique

We will choose suitable machine unlearning techniques for class-wise and sample-wise unlearning. Note that it is ideal that we develop the capability to do both class-wise and sample-wise unlearning as both are important for developing privacy models. We will then leverage the *saliency unlearning* (SalUn) technique to be used with these MU approaches.It is likely that we will use random labelling (RL) for class-wise unlearning, as this was used in [2] and shown to be effective.

A summary of the SalUn approach combined with RL is as follows:

1. Weight saliency, which used to identify the weights contributing the most to the model output.

$$\mathbf{m}_S = \mathbf{1}(|\nabla_\theta \ell_f(\boldsymbol{\theta}; \mathbb{D}_f)\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_O}| \geq \gamma)\,,$$

$$\boldsymbol{\theta}_u = \mathbf{m}_S \odot (\Delta\boldsymbol{\theta} + \boldsymbol{\theta}_O) + (1 - \mathbf{m}_S) \odot \boldsymbol{\theta}_O$$

Where m_s is the weight saliency map; $\boldsymbol{\Theta}_u$ is the unlearning model, a way of expressing the model to be unlearned by separating the salient weights from the original weights; $\boldsymbol{\theta}_O$ are the weights from a pre-unlearning model; $\ell_f(\boldsymbol{\theta}; \mathbb{D}_f)$ is the forgetting loss, meaning the model weights variable θ under the forgetting dataset $\mathbb{D}_f$;$|\cdot|$ is the absolute value operation; where $1(g \geq \gamma)$ acts as a threshold that if gradient vector will be outputted if it is greater than threshold γ, otherwise the value would be 0; $\nabla_\theta \ell_f(\boldsymbol{\theta}; \mathbb{D}_f)\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_O}|$ is the gradient vector. This method is used to identify the weights which are sensitive to the data/class/concept forgetting.

2. Now we integrate SalUn with RL. RL assigns a random label to an image to be forgotten, then fine tunes the model on the randomly labelled forget set, as well as the remaining data D_r to preserve the model utility, SalUn allows us to only fine tune the salient weights.

$$\textbf{minimise}_{\Delta\theta} \boldsymbol{L}^{(1)}_{\text{SalUn}}(\boldsymbol{\theta}_u) := \mathbf{E}_{(x,\,y)\sim Df,\,y'\neq y}[\ell_{CE}(\boldsymbol{\theta}_u; x, y')] + \alpha \mathrm{E}_{(x,y)\sim Dr}\,[\ell_{CE}(\boldsymbol{\theta}_u; x, y)]$$

ℓ_{CE} is the cross-entropy (CE) loss; y ′ is the random label of the image x different from y; Dr refers to non-forgetting dataset; α > 0 as a regularization parameter; $\mathbf{E}_{(x,\,y)}$ stands for set.

3. Through iterative fine tuning, the model will begin to associate the forgetting concepts with random data points, therefore having it functionally forget about those concepts, as they hold no real meaning after the unlearning process.

The process and results of SalUn method are summarized in Fig1 [2].

$$\textbf{minimise}_{\Delta\theta} \boldsymbol{L}^{(2)}_{\text{SalUn}}(\boldsymbol{\theta}_u) := \mathbf{E}_{(x,\,c)\sim Df,\,t,\,\epsilon\sim N(0,1),c'\neq c}\,[\,\|\,\epsilon_{\theta u}(x_t \mid c')\,\|_2^2] + \beta\ell_{MSE}(\boldsymbol{\theta}_u; \mathcal{D}_r)$$

$_{c'\neq c}$ means c’ is different from c;

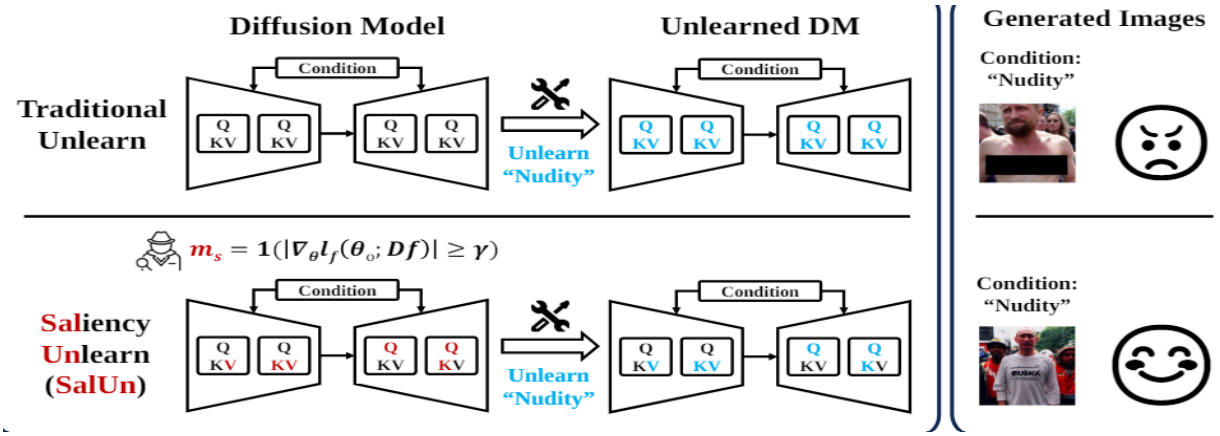


Figure 1. the SalUn process compared to traditional unlearning

3.3.3 Develop software for streamlining the unlearning process

Developing a software to enable the testing of various unlearning techniques, or otherwise improve individual models of unlearning, would largely improve our process and simplify the testing of these models. This would contain a library of models of unlearning techniques and would be able to call upon and train these models with minimal effort. This would then be utilised to compare and analyse these easier. This is a more stretch objective and would aim to simplify our processes if we are able to code and utilise software of this variety.

3.3.4 To evaluate and optimise the efficiency of the model after SalUn

After implementing our unlearning technique we will test and observe it working on a model, and then test the model’s accuracy and analyse the effects of the unlearning on the accuracy of the model. We will then optimize it to ensure the maximum accuracy, as precision in medical diagnostics is of the highest priority.

We will aim to use a full stack evaluation metric as described by Jinghan Jia et al. [1], and is outlined in more detail in section 3.3.2.

3.3.5 Compare other approaches and implement image segmentation

As our stretch goals, we could also utilize our framework to test other tasks, analyse results, and make comparisons across these tasks and to our medical imagery model. This will result in our work being more comprehensive and rigorous and allow us to work towards a more accurate and consistent model.

Additionally, we could try implementing image segmentation to expand our model work and further utilise unlearning to explore various implementations, such as unlearning sensor bias in model types [x]. Image segmentation has been performed within medical imagery already [x], however the implementation of unlearning may guide this process to further improve privacy and other concerns within AI implementation.

4 Literature Review

4.0 Introduction

Machine learning (ML) has become a revolutionary advent in many domains, enabling sophisticated pattern recognition, prediction and decision making. However, as ML systems are deployed in increasingly sensitive and critical applications, the ability to correct bias, retract learned information, and forget data or classes has become imperative, especially when relating to security, privacy and fairness. This has given rise to the field of "machine unlearning" (MU), aimed at developing methodologies to undo or revise the learned knowledge of machine learning models to make them more secure or fair, while maintaining performance. This literature review aims to give a comprehensive overview of the existing research landscape, challenges and future direction of the field of MU.

4.1 Machine Learning

Machine learning is a branch of artificial intelligence that involves a model learning from data without explicit programming. It allows models to recognize patterns existing in the training data and then make predictions or decisions based on new data. Machine learning is used in many domains, from predicting customer preferences in e-commerce to diagnosing diseases in healthcare, the field relevant to our work. The generalisation of machine learning, as well as its potential precision and ability to extract insights from real world data, has made it a popular new technology in industry, research, and personal life.

4.2 The Need For Machine Unlearning

As the prevalence of machine learning applications grows, so too does the demand for the capability of machine unlearning. MU is needed in many different scenarios that relate to individual privacy, ethics in model outputs and model security. This section will provide an overview of the different instances where machine unlearning may be applicable.

Note that removing the desired data points and retraining a model from scratch (known as Retrain) is undesirable due to the immense computational cost. Therefore MU has grown to be an integral area of research.

4.2.1 Privacy

Personal data has become a thing of great value to companies from the domains of banking to hospitality. As a response, many privacy laws have been legislated to protect user's personal data for privacy and security reasons and acknowledge the right for an individual to have their data deleted upon request. Examples of such laws include the following:

- The European Union's General Data Protection Regulation (GDPR) [3], specifically "the right to be forgotten".
- Canada's proposed Consumer Privacy Protection Act (CPPA) [4]
- The California Consumer Privacy Act (CCPA) [5].

However, simply removing the user's data may not be sufficient as machine learning algorithms have also used this information for training, and are able to elicit user data from a trained model [6], potentially resulting in privacy attacks ([7], [8]). Therefore, there is a requirement to have these models forget the influence this data had during the training process.

There has been criticism of these emerging laws including the potential for it to infringe on the freedom of expression [9] and freedom of speech [10], as it can be interpreted as not allowing individuals to criticize someone's past.

Additionally, it has been found that sensitive personal information can be memorised by models such as text-to-image models remembering credit card details [11].

4.2.2 Fairness

There is also a tendency for ML models to exhibit bias and unfairness in their output [12]. For example, an AI model could take into account legally protected characteristics such as race or gender when - for example - a model used for medical diagnostics should be fair to people of different races and free of social bias [13]. This problem can arise due to the training data, for example, a database that has predominantly caucasian people.

Rohit Gandikota et al. [14] experiment with gender and racial debiasing of stable diffusion models and obtain the results shown in Fig 2 and Fig 3.

Figure 2. Gender debiasing

Figure 3. Racial debiasing

4.2.3 Inappropriate Outputs

Another reason MU has become a promising prospect is due to the rise of text-to-image models and the requirement for them to not produce harmful or illegal outputs such as nudity or copyrighted artistic styles [15], as well as gore or violence. Although stable diffusion models have safety filters in place that attempt to prevent the model from generating distasteful images, it has been shown to be less than adequate [16]. MU can be used in these situations to allow the stable diffusion models to "forget" about undesirable or irrelevant aspects. [17]

Rohit Gandikota et al. [14] experiment of manually erasing unwanted and inappropriate outputs versus removing topics from the model is notable in the results shown in Fig4.

Figure 4. Erasing nudity

4.2.4 Adaptability

MU can also be utilized in transfer learning and allow models to adapt / update to changing data as old data becomes outdated or irrelevant [17].

4.2.5 Performance

The impact of bad data can negatively affect the performance of machine learning algorithms. Machine unlearning techniques can also help remedy this by removing the effects that this bad data has on the outputs of the model [18].

4.2.6 Security

MU can repair a model "damaged" by a data poisoning attack, where "bad" data is inserted into the training data by malicious actors to degrade performance [19]. In contention with open source image generation programs and their non-consensual image collection methods, other programs such as Nightshade have arisen to directly poison AI data modelling. [20] MU can repair from these data attacks by effectively removing this data as it is found from certain sources.

MU offers a good defence against data poisoning attacks and other back door attacks [21].

4.3 Existing Problems and Challenges

As MU is a relatively new field of research, there are still many problems that are yet to be solved.

Furthermore, there is often a trade off between unlearning efficiency and model utility, this is a huge limitation on unlearning in general [22], some approaches may also result in catastrophic forgetting [23].

The major challenges in medical image neuron network performance sits within the data domain, such as the lack of labeled datasets are imbalanced thus the network emerges a biased performance.

Shaik et al. provide a comprehensive overview of the current problems in 2024 [24], also Jie Xu et al. in 2024 [25], and Thanh Tam Nguyen et al. in 2022 [26]. This section will summarise the problems discussed in the literature.

4.3.1 Attack Sophistication

ML models are generally susceptible to many sorts of attacks [26]. As MU methods become increasingly sophisticated, so do the potential attacks. Some suggested solutions involve training the model on diverse or adversarial data, making it more robust and resilient to attack [27].

4.3.2 Lack Of Standardisation

There is a general lack of standardisation in the field of MU. Firstly, there is no standard for methodology, making results difficult to reproduce and compare. Secondly, there is no clear standard of evaluation or method to measure the effectiveness, robustness or fairness of a model that has undergone MU.

Some suggested metrics of evaluation are:

- *Unlearning Time Efficiency.*
- *Resource Consumption.*
- *Privacy Preservation.*

Some methods used in the literature to test the effectiveness of unlearning is as follows:

- L^2 *distance* which measures the distance between the unlearned model's parameters and the parameters of an exact retrained model.
- *Feature Injection Test,* proposed by [28] and used in papers such as [29] and [30] . The idea that a model that unlearns a feature should have the associated weights with that feature drop to zero after unlearning. This technique is only applicable to feature-level unlearning.

Jinghan Jia et al. propose a “full stack” evaluation of MU in their work [31], and includes the following:

- *Unlearning Accuracy* which tests the effectiveness of the unlearned model on the unlearned training data. i.e testing how well a stable diffusion model can produce an image of “Elon Musk” after it has been tasked to unlearn “Elon Musk” as a concept.
- *Membership Interference Attack* can be used as an evaluation metric by testing the success rate of an MIA on data that was tasked to be forgotten.
- *Remaining Accuracy* tests how well the model performs post unlearning on the concepts / data that was not tasked to be forgotten from the training data.
- *Testing Accuracy* which tests the model’s ability on new, testing data after the unlearning process.
- *Run Time Efficiency* refers to the computational cost of the unlearning algorithm and is compared to the baseline Retrain approach.

4.3.3 Lack Of Transferability and Interpretability

Some methods of MU may work well with a particular model type or data type, but cannot be transferred to other model types or data types. One approach to mitigating this involves transfer learning, which is a field of ML which allows knowledge gathered from one model or dataset to be applied to another model or dataset [29].

Interpretability refers to the ability to understand the decisions made by the ML model. Understanding the MU algorithms and being able to interpret its effect is essential for building stakeholder trust. Furthermore there is often a tradeoff between accuracy and interpretability.

4.3.4 Lack Of Training Data and Resource Constraints

A lack of training data is a common problem within MU, akin to many domains within ML.

Computational power is also a common resource constraint with many MU algorithms as they commonly require the augmentation or retraining of large neural networks.

4.3.5 Privacy

It has been found that the unlearning process in itself can reveal sensitive information about users [32], and removing one data point can cause others to become more susceptible to membership interference attacks, this phenomenon has been referred to as the “Privacy Onion Effect” by [33] and is defined as the following:

*Removing the “*layer*” of outlier points that are most vulnerable to a privacy attack exposes a new layer of previously-safe points to the same attack.*

A similar phenomenon is the “Streisand Effect”, discussed in other literature [34], where the unlearning effect itself is detectable.

A “Proof of unlearning” is necessary to confirm to the stakeholders that their data has been removed from ML systems, and important for compliance and accountability as privacy laws are increasingly recognised. This is not a trivial task, and it has been shown that such a proof cannot be gleaned from the model parameters , as parameters can be the same regardless of whether a specific data point was involved during the training process [35].

5 Methodology and Methods

5.1 Tested Methods and Approaches

Many techniques have been developed to achieve MU. The naive approach is to delete the unwanted data and retrain the model from scratch, this is commonly referred to as “Retraining”. This technique is heralded as the golden standard for machine unlearning, as it yields the best performance after the unlearning processes. However, this process is extremely computationally expensive, as retraining a model from scratch every time we want to unlearn something is unfeasible. Another naive approach is to simply set weights to 0, or add noise to the weights, but this would come at a detriment to performance. Therefore several other approaches have been suggested, this section will give a brief overview of several of these approaches.

Broadly speaking, there are two approaches to Machine unlearning:

- Exact unlearning, also referred to in the literature as certified unlearning and includes retraining based methods.
- Approximate unlearning.

Exact retraining includes naive retraining and other methods that maintain good model performance, but come at the cost of expensive computation. Approximate unlearning however seeks to augment existing models without a complete retrain, therefore sacrificing performance for computation cost.

Many unlearning approaches address simple learning algorithms such as linear/logistic regression [41], random forests [42] and k-means clustering [43]. This section will focus on methods specifically for convolutional neural networks.

Furthermore, different MU approaches aim to solve different problems. Namely, there is class-wise MU which attempts to get a model to forget a certain class eg. airplane. There is also sample-wise MU, which aims to mitigate the effects of specific data points on the model.

5.1.1 Exact Unlearning Approaches

Bourtoule et al [44] propose an algorithm coined “SISA” or “Sharded, Isolated, Sliced, and Aggregated training” which includes splitting the data into smaller groups, or shards, and training several models on each of these shards in isolation, the final output therefore is the majority vote between all these models. Therefore when a request to forget a data point is received, only the relevant model will need retraining. As shards are smaller than the entire data set, this results in a cheaper computation. Furthermore, these shards are divided into slices which are incrementally introduced during the training process, the parameters of the model are saved at each iteration, therefore when a retrain is required, the algorithm does not need to retrain over the entire shard. This algorithm has a high space complexity as it requires various versions of the model to be saved to memory [45].

Haonan Yan et al. propose a method called “ARCANE” which builds on the SISA approach, but instead of randomly partitioning the training data, the data is partitioned based on class [46]. This approach has improved efficiency compared to SISA.

Chuan Guo et al. propose a method called “Certified removal” [47], which takes influence from differential privacy, adding randomness into the training algorithm to guarantee privacy of the training data. Certified removal involves having multiple models trained, which bounds the maximum divergence of the original model between a partially trained model and a model trained without the now removed data initially. This allows for a comparison point after removing certain data, which allows for an accurate estimate of the effective removal of the data from the model. While this has a much higher initial training cost, which makes this a very slow overall unlearning technique, it maintains a high level of accuracy and efficiency

upon removal of data, and can further allow for easy removal of poisoned data due to the separate training models.

Warnecke et al. offer a solution for unlearning labels which builds on the idea of influence functions [48]. The technique was faster than retrain or SISA and had comparable accuracy.

5.1.2 Approximate Unlearning Approaches

Aditya Golatkar et al. suggest a method referred to as "Fisher forgetting" [49], which utilises a Fisher Information Matrix (FIM) - a way of measuring how much information a variable has about an unknown parameter - to calculate the optimal noise to destroy information. This method, like many methods, draws parallels with differential privacy [50], where noise is introduced to personal information to maintain privacy.

Aditya Golatkar et al. suggest a scrubbing method, referred to as a "one-shot" forgetting algorithm [51] that is obtained based on the Neural Tangent Kernel (NTK). This algorithm improves effectiveness relative to Fisher forgetting - measured by the closeness of activations when compared to a "Retrain" model. However it is computationally expensive due to the computation and storing of huge matrices.

Both "Fisher forgetting" and the "one-shot" scrubbing methods were initial work in the field of forgetting in CNNs, and lack scalability and, although not as bad as a naive Retrain, they are still computationally expensive [52]. Future work would build on these ideas and decrease computational work and increase scalability.

A common method is to leverage the concept of influence functions [53] and is concerned with parts of the model that are highly influenced by the relevant date. This approach is used in work such as that of P. Koh et al in black-box predictions. [54]

Ayush K Tarun et al. introduce a method called "Unlearning by Selective Impair and Repair" or "UNSIR" [52]. This approach considers that an algorithm will try to minimize the loss function during the training phase, therefore, if it is desired to forget what was learnt during this phase, then one would need to maximise a loss function. The algorithm goes through an impair phase, where the model is trained on a subset of data from the original training data which also contains noise, then repairs the weights by training it for a single epoch on retrain data. This algorithm saw improvement in accuracy and scalability compared to previous methods.

Yinjun Wu et al. suggest a method called "DeltaGrad". This model relies on information cached during the training phase and has improved time complexity when compared to the naive approach [55]. This algorithm is specific to stochastic gradient descent based machine learning models and takes inspiration from the idea of "*differentiating the optimization path*".

Aditya Golatkar et al. suggest another approach coined "Mixed-Linear Forgetting (ML-Forgetting)" and is specific to computer vision models [56]. This technique takes advantage of a subset of the training data that we assume will never need to be forgotten called "core data". This could be for example a large, open source database of generic data such as ImageNet. This algorithm will learn two sets of weights, one set trained on the core data, and one trained on the user data. This allows the model to forget all user data simply by setting the user weights to zero. This approach is fast compared to Retrain and is comparable in accuracy.

Zachary Izzo et al. [32] suggests a technique called the "Projective Residual Update" or PRU. This technique utilises *synthetic* data combined with gradient methods to accomplish an approximate parametre update towards forgetting a small group of data points. This technique was found to be faster than influence unlearning under certain conditions such as when the amount of data to remove is small.

Other baseline methods include fine-tuning based approaches, which involves continuing to train a model using corrective data and to initiate catastrophic forgetting the required forget data, an example of this is “Random Labeling” [56].

Other approaches include gradient ascent [57], Amnesiac Machine Unlearning [58], ℓ1-sparse [34], boundary unlearning [59], erased stable diffusion (ESD) [18], and forget me not [60].

Another approach, and the approach that we will be employing in our project is SalUn [2], which has been explained in detail in section 2.2.2.

5.2 Methodology and Approaches

In the current stage of the project, models are being developed in order to ensure that existing models are capable of analysing and successfully classifying various types of medical imagery. The team is attempting to also develop our own model to test alongside existing classification models to provide another point of reference for model and unlearning technique development.

Within developing our model, a reference model is being used to create this model, utilizing the Resnet 18 and testing its base level validity. From here, the model is being adjusted and will continue to be adjusted to improve its accuracy, both before and after the implementation of unlearning techniques. This will then be compared to other existing models, to test the capability of a self-developed model, as well as test more independently the capability of existing models to train on medical datasets.

Resnet-18 is the primary, existing image classification model that we are utilising, with base models being run and tested on the MedMNIST database. We will then use these trained models to implement various unlearning techniques, including and primarily focussing on Salient Unlearning. These tested models will then be objectively compared, using test accuracies and comparing AUC (area under curve) graphs to ensure the overall consistent accuracy of each model. These will then be retested and compared upon the implementation of unlearning techniques.

Testing multiple unlearning techniques on not only the same datasets but various matching image classification models will allow for the most consistent comparison points across the board. By testing on the same datasets, it will ensure that information being forgotten across each learning technique is identical, thus directly comparing their capabilities of removing that specific information. Further, testing these on various image classification models will help identify the more consistent unlearning technique, as some techniques may be directly improved on the type of model it is being enacted upon. Finally, we will ensure that each technique will be tested multiple times on each model, so as to reduce the impact of the random chance accuracy that may be caused after unlearning is tested across each model as models may be able to randomly guess the correct class by chance, of which this technical accuracy can be ideally reduced via repetition.

5.2.1 Difficulties with current Method

There are multiple major difficulties that we have encountered during the testing and development processes. Firstly, some medical datasets have insufficient training, validation and testing data to rigorously and sufficiently test and confirm the capabilities of a model. For example, the RetinaMNIST from the MedMNIST public dataset is an extremely small dataset for training a model, and thus would reach extremely high accuracy on training data whilst having testing results that mirrored that of a completely untrained, and thus randomly deciding, model.

Another major issue we have encountered is the lower computational capabilities of the devices available to the project team. This lower computational power is creating time delays during model training and

testing that, while were generally expected by the team, are longer than originally planned. This is creating time issues for both the overall project, but also time issues in identifying difficulties and failures of techniques being applied.

5.3 Our Method

First, we will train ResNet18 models on MedMNIST data and attempt to replicate the results by [62].

Then, to verify the above techniques can be applied to medical imergary models, we choose three simple unlearning techniques: RL, FT, GA and compare them with and without Salient Unlearning. We also use "Retrain" models to provide a baseline for performance.

MedMNIST offers datasets of different sizes for further analyses. We can use the work done by Fan et al. [2] as reference and a baseline to compare our results to.

To assess the unlearned models, we will test the models' unlearned accuracy (how well the model has forgotten the class), and test accuracy (remaining performance on test data).

Our codebase is available here. All results used a random seed value of 2 for experiment replicability.

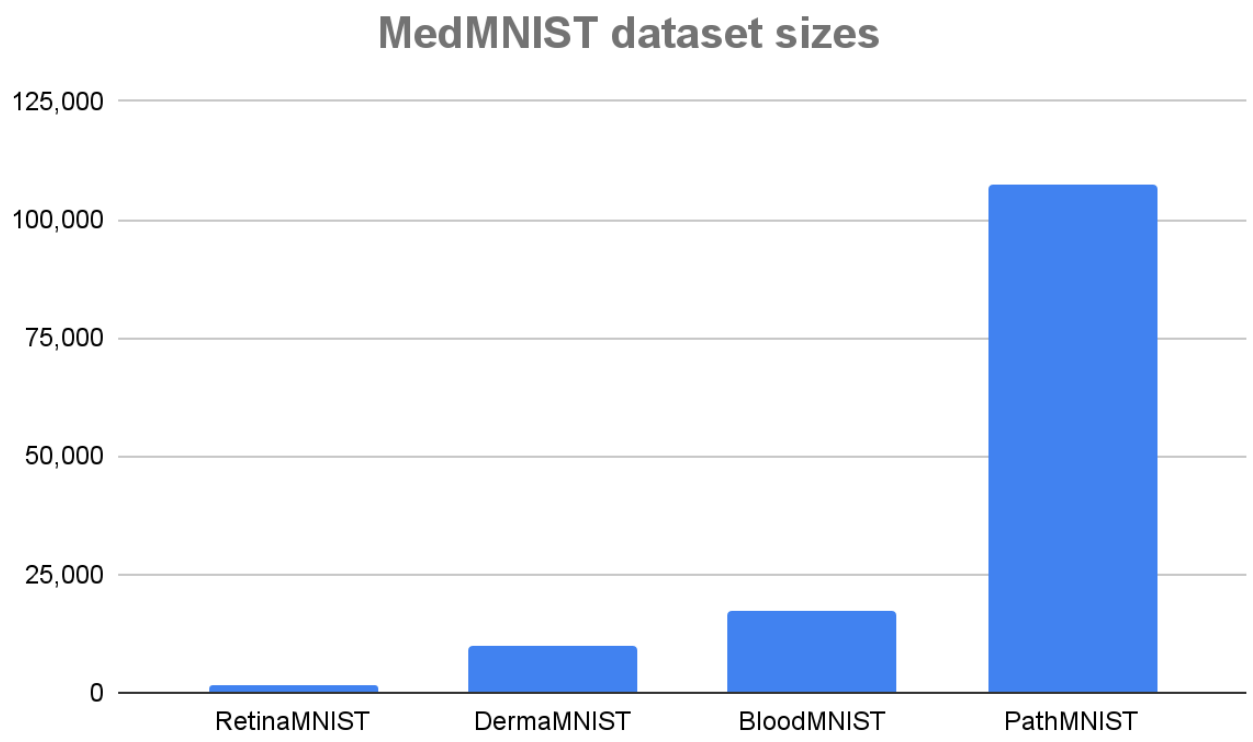


Figure 5: Number of images per dataset from MedMNIST [62]

6 Results and Discussion

6.1 Final results and discussion

	Accuracy (Our Results)	Accuracy (Yang et al.)
RetinaMNIST	59.3%	49.3%
DermaMNIST	87.0%	75.4%
BloodMNIST	98.7%	96.3%
PathMNIST	92.4%	90.9%

Table 1: Base model Accuracies of MedMNIST Datasets [62]

Base models achieved similar accuracies to Yang et al. [62] shown on Table 1, however where they used a learning rate of 0.001, we found that a smaller learning rate of 0.0001 resulted in better models accuracy across the board. This allowed us to utilise higher accuracies throughout our testing to effectively demonstrate the capabilities of unlearning in medical data.

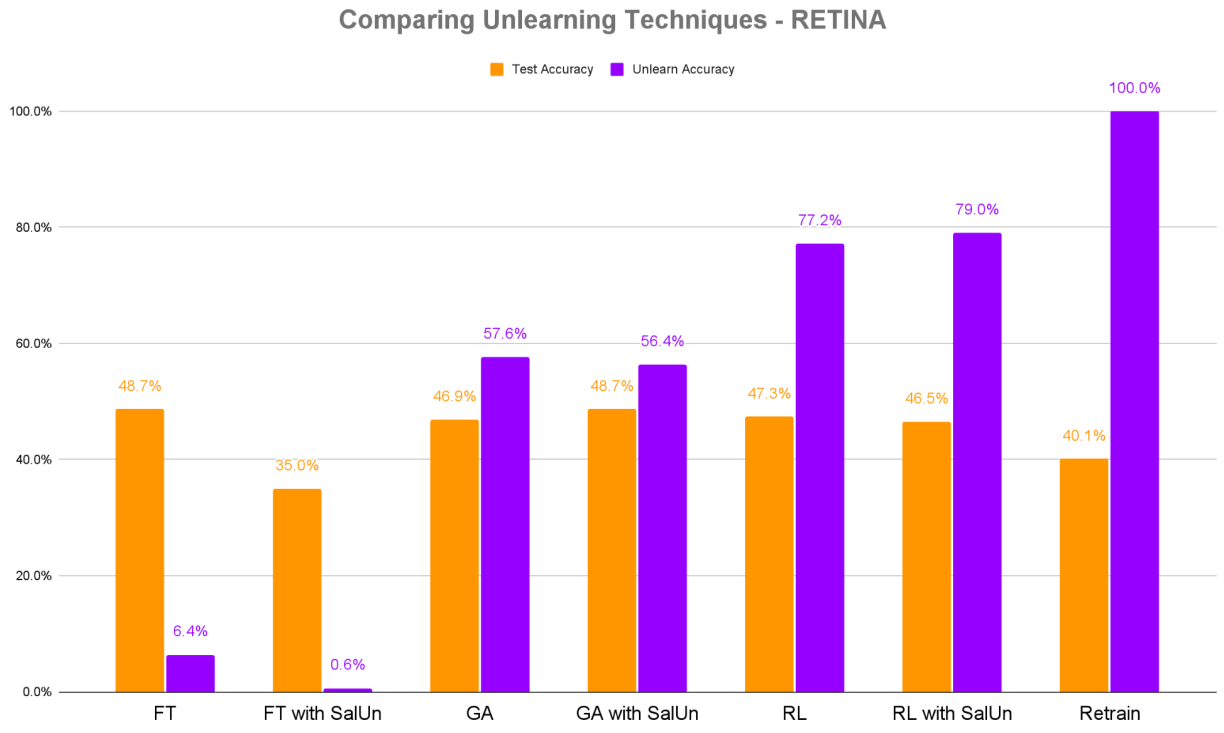


Figure 6: Recorded accuracies on the RetinaMNIST dataset across unlearning techniques

RetinaMNIST was used for initial testing, due to its small size. The results are displayed in Figure 6. The results showed that FT struggled to unlearn the class, but maintained good test accuracy. On the other hand, although not as pronounced GA could easily forget the class, but sacrificed a small amount of test accuracy. These results are in accordance with the literature. A notable result was the poor performance of Retrain. We reason that the small size of the dataset implies that retraining on an even smaller dataset will quickly diminish the performance, however it is the only method seen to successfully forget the class. However, as test accuracy should be our first priority, we note that a model with such a low base accuracy should not be used for medical classification in the case of patient diagnoses.

Two major parameters that were changed for comparison between models was the class to forget and the unlearning technique performed on the model. Maintaining multiple parameters between models was important for ensuring models were easily and directly comparable. For DermaMNIST, we refer to three classes of varying sizes, class 3 makes up 1% of the dataset, class 4 makes up 11%, and class 5 makes up 67%. The model that had forgotten class five performed by far the worse, which did not come as a surprise. Something of note is the slight increase in performance of the class 4 forgotten model when compared to the class 3 forgotten model. It can also be gleaned from the data that salient unlearning might have more of an impact when trying to forget classes that make up a larger portion of the total training data, however more research must be done to verify this.

We concluded that in general, class to forget is not a relevant variable, unless the amount of data is substantial. Therefore we will focus on other variables, namely, different learning techniques across different datasets.

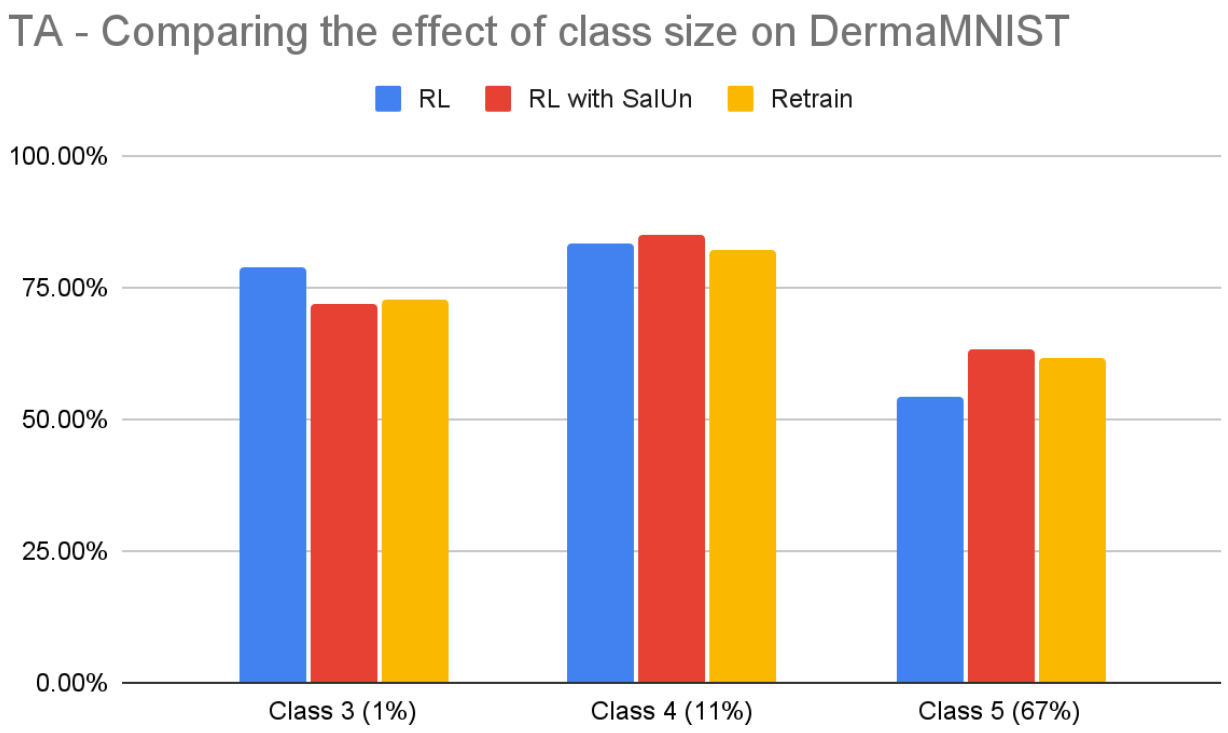


Figure 7: Testing accuracy between different classes of DermMNIST, comparing RL, RL with SalUn to Retrain.

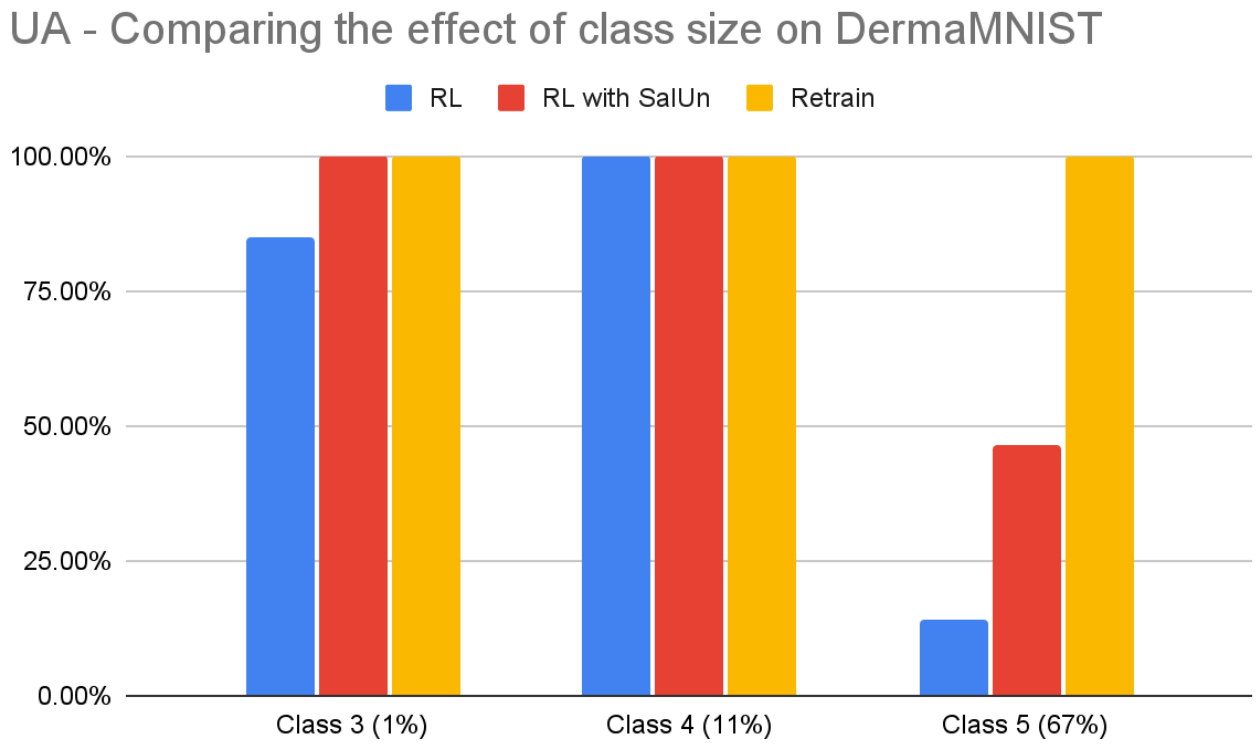


Figure 8: Unlearning accuracy between different classes of DermMNIST, comparing RL, RL with SalUn to Retrain.

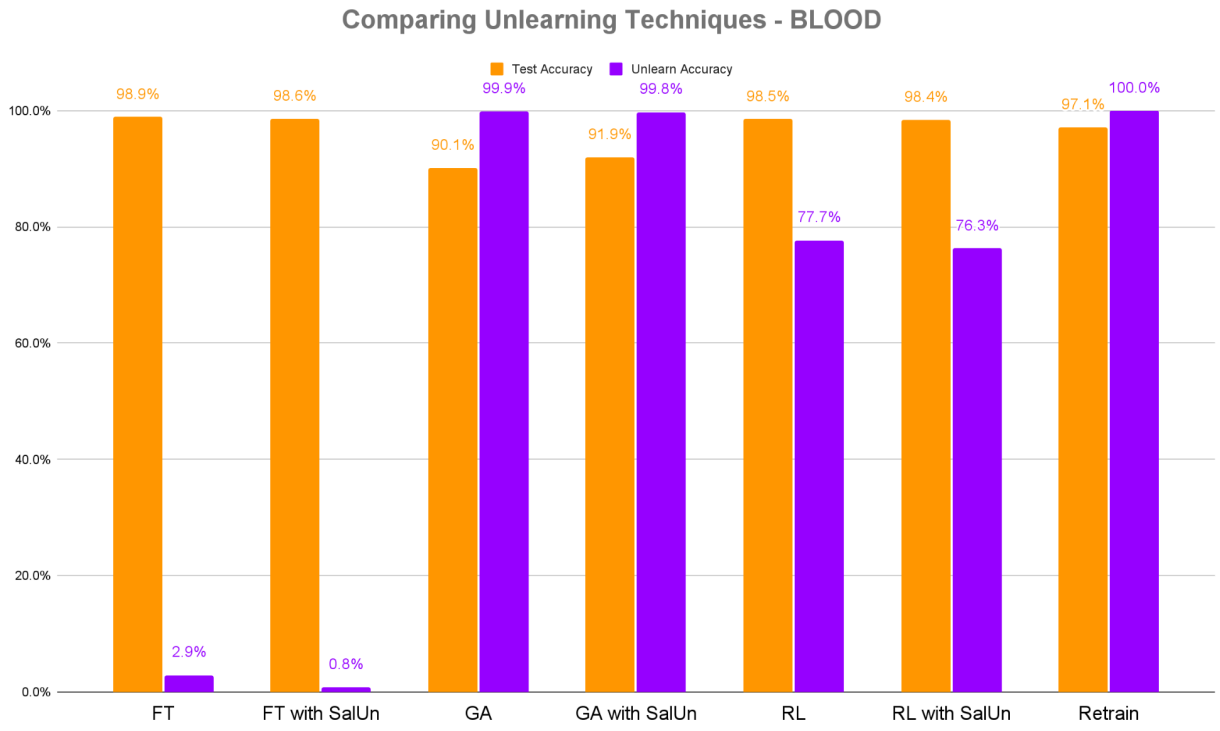


Figure 9: Recorded accuracies on the BloodMNIST dataset across all tested unlearning techniques

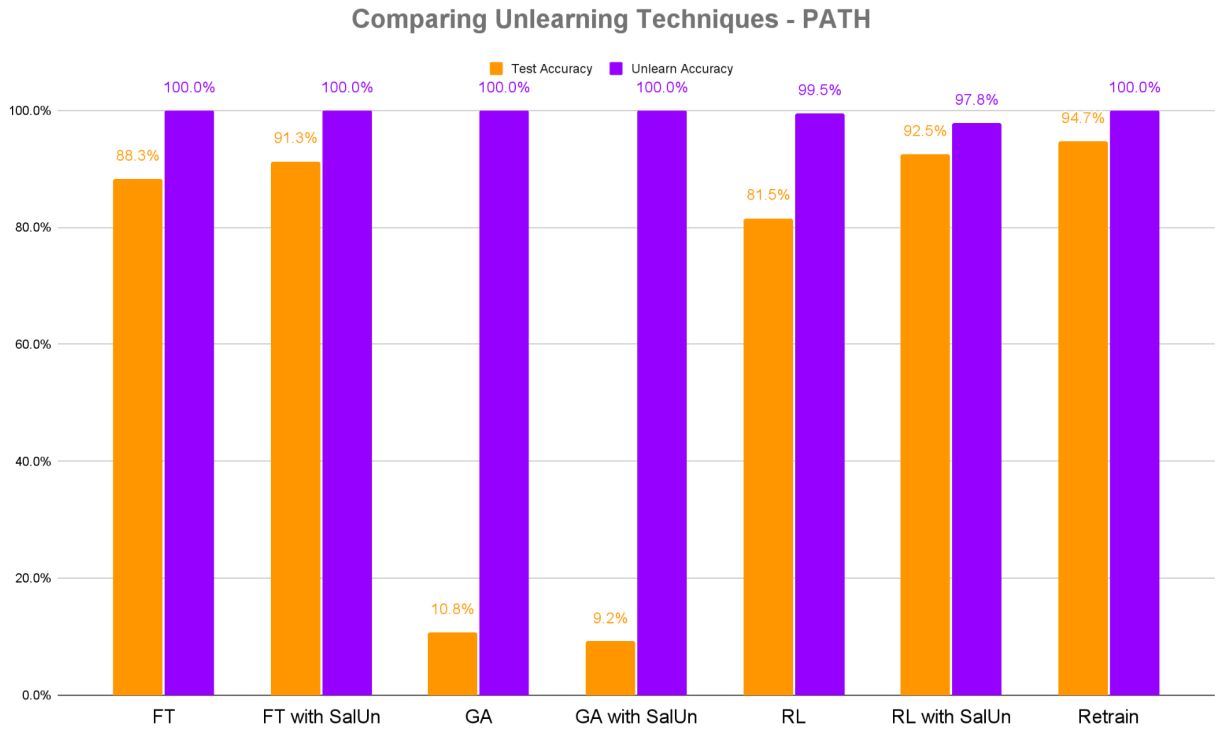


Figure 10: Recorded accuracies on the PathMNIST dataset across unlearning techniques

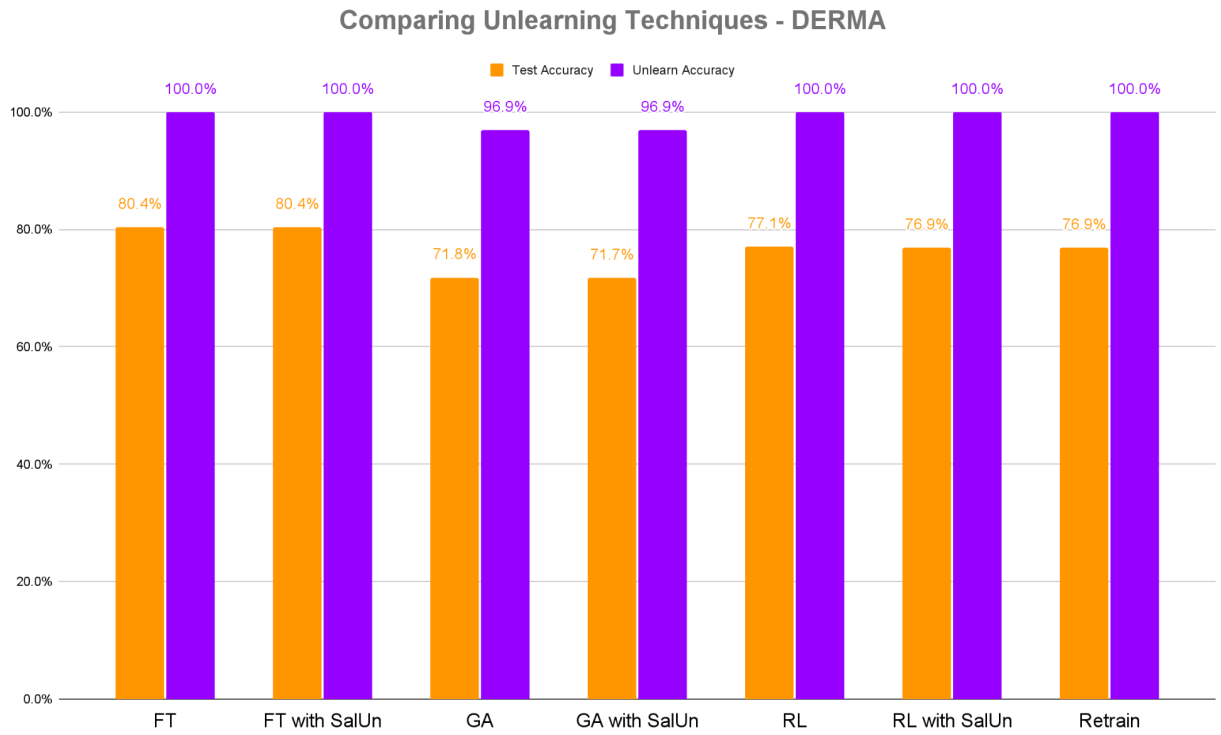


Figure 11: Recorded accuracies on the DermaMNIST dataset across unlearning techniques

Throughout our testing, we found that Salient Unlearning actually had an overall minimal effect in improving either our unlearning or test accuracies. The only significant difference that SalUn made across all of our testing was dramatically improving Random Labelling on the PathMNIST dataset. While this one result was notable, there was not enough evidence throughout the rest of our testing to prove SalUn as a notable and effective addition to the unlearning process.

FT and GA, while sometimes proving optimal, seems to be a hit or miss method, and usually sacrifices test accuracy in the case of GA or unlearn accuracy in the case of FT. RL seems to be the best technique in every dataset we have sampled, as it doesn't struggle to forget classes, but maintains good test accuracy. We compare our results with Fan et al. who used ResNet18 architecture to train on CIFAR-10.

	Our Results		Fan et al.	
	UA	TA	UA	TA
Retrain	100.0%	77.2%	100.0%	92.5%
FT	52.3%	79.0%	31.7%	94.8%
GA	88.4%	58.7%	99.9%	38.2%
RL	88.6%	76.1%	89.3%	94.5%
RL with SalUn	88.3%	78.6%	99.9%	94.5%

Table 2. Unlearning techniques comparison vs baseline performed on CIFAR-10 [62]

PathMNIST and BloodMNIST had the highest base accuracies, as seen in Table 1, however the impact on the models from each unlearning technique varied greatly between the two. Notably, PathMNIST is a much larger dataset than BloodMNIST, however its unlearning accuracy maintained a much higher average across every technique, which made any initial assumptions regarding data size incorrect. This allowed us to conclude that, while datasets may be too small to be sufficiently learned upon, massive increases on the higher end did not seem to improve models further. On the other hand, DermaMNIST was overall the most capable dataset of being unlearned, as every technique was capable of maintaining test accuracy and achieving high unlearn accuracies. Across the board, DermaMNIST was the least impacted by unlearning.

BloodMNIST also ended up with an interesting trend of being difficult to unlearn classes in, most notably with FT failing to unlearn the designated class essentially at all, and RL performing the worst on BloodMNIST as well, with only Retrain being able to fully unlearn the designated class. This was an extremely notable result, however we were not able to determine what specifically caused this difference, as BloodMNIST had no specific difference that would easily explain this phenomena.

Unlearning Method	Unlearn Time (Seconds)
GA	226
FT	716
RL	778
RL with SalUn	743
Retrain	10436

Table 3: Unlearn time for different unlearning Techniques on PathMNIST

Fine Tuning proved an effective unlearning technique initially, but performed extremely poorly on the BloodMNIST dataset. As noted in Table 2, this directly impacted the overall performance of the technique. However, this type of result was an expectation based on the literature, and thus the promising performances on the other datasets came as a surprise, yet ultimately could not allow the technique to promisingly overcome its previous shortcomings.

Gradient Ascent had the exact opposite issue to that of Fine Tuning, with test accuracy being directly sacrificed for rapid yet effective unlearning. This is a much larger issue in our specific circumstance, as model accuracy is a much more important metric, and thus leaves GA as an ineffective technique for our purposes. This was most notable within PathMNIST, as seen in Figure 10, as the test accuracy became akin to a random guess percentage, almost entirely destroying the base model in favour of unlearning the specific class. Gradient Ascent was however capable of achieving its final values at a very rapid rate, but the sacrifice to test accuracy was far too great.

Random Labelling proved our most effective overall unlearning technique, achieving the highest unlearning accuracy across the board while maintaining the majority of the model's test accuracy. Supporting this technique with Salient Unlearning allowed for the model to maintain a higher level of test accuracy, and overall managed to maintain the highest capabilities of all the unlearning techniques excluding Retrain. RL also had a similar run time to that of Fine Tuning, both of which were average, being longer than the speed of GA but significantly faster than the full relearning of the model through Retrain.

Overall, this allowed us to determine RL as our most effective unlearning technique across the board.

6.2 Findings

There were multiple takeaways from our research, with some results less expected than others.

Despite the promising results from the literature of Salient Unlearning, we found that it had a minimal overall impact across our data. This surprised us initially, as the work done within the SalUn paper created the assumption that it would be highly capable of improving our data, yet only had an overall minor difference in our final results on RL. In some cases, when applied to other unlearning techniques, it would even worsen the model, and thus we found it to not be a universally applicable technique. Thus, we found that while Salient Unlearning was capable of minorly improving RL, the use of Salient Unlearning was more nuanced than a simple plug and play into other unlearning models, and has so far proved ineffective in its overall application.

Further, we found that it was not only possible, but practicable, to apply unlearning to various medical datasets. Unlearning was effective in most circumstances, and with the correct parameters was able to be performed with minimal impact on the base model, and in cases such as the PathMNIST dataset, could maintain extremely high test accuracies while effectively removing and forgetting a specific class. While performing well in both unlearning and test accuracy was a challenge for every unlearning technique, with trade offs being a common occurrence, the capabilities of unlearning became apparent but still remained a challenge while maintaining test accuracy.

Comparing various unlearning techniques, as well as being able to run and test Salient Unlearning on a variety of unlearning techniques beyond RL, proved manageable and became a staple process in our testing. Finding the correct parameters became a simple component of our testing as well, and by minimising our parameter differences, we were able to effectively compare the unlearning techniques and simplify the testing process in this aspect as well.

6.3 Limitations and future work

One major limitation of our methodology was the overall consistency of our outputs. No unlearning technique had the same output over different datasets, and thus seemed to be directly impacted by the datasets themselves. A clear and major example of this was Fine Tuning's final results on the BloodMNIST dataset compared to results of the other datasets. Having such a dramatic failure of an unlearning accuracy was quite sudden, and it was extremely unclear what caused this for specifically this dataset compared to the others. This output was also repeatable, so we determined that it was not an error point and was in fact the correct final model. This time of failure of consistency was also seen in our other techniques, such as GA losing a large amount of test accuracy in PathMNIST compared to the other datasets. We were unable to find a specific reason behind these inconsistencies, and though literature originally specified that GA would decrease test accuracy and FT would decrease unlearn accuracy, these values were both inconsistent and dramatic in difference to the expected values based on literature. It is unclear how we may be able to prevent this, but further research may be done about the specifics of unlearning techniques repeatability across datasets and how the type of data available to a model may impact these results.

Another major limitation was the computational cost of running these models. Due to the nature of machine learning, running and creating models for a wide variety of unlearning techniques and datasets had a heavy computational cost on the team. This required the use of high-end Graphics processing units and continuous, strenuous work on these units to ensure that models could be generated and sufficiently tested. Unfortunately, these requirements caused longer computational time on the majority of the team's personal computers, due to the relatively poorer capabilities of the devices. Due to this, we prioritised using a singular computer, which held the highest performance capabilities, and allowed for models to be run

within a reasonable amount of time. Access to higher end computers would allow for models to be run more efficiently, and effectively improve the testing capabilities of the research, especially within the same given timeframe.

This project could be propelled forward by its further application into image segmentation. By being able to segment images down and reduce the amount of external noise that may not be relevant to medical analysis, learning models would be able to further support these analyses and can be more directly aimed at reducing the impact of biases. Further, unlearning of segmented images can allow for models to be more capable of identifying important differences, and aim to improve overall accuracy of the models and develop their capabilities further.

Applying samplewise forgetting to these techniques would further allow models to be able to improve privacy across real world applications. Through samplewise unlearning, patients may request their data be removed from a model at any time, and this data can be selectively removed and completely erased from a model if effectively unlearned. This would help with patient confidentiality and privacy, and give patients and the general public comfort in knowing they have the right to opt in or opt out of the development of further models and data.

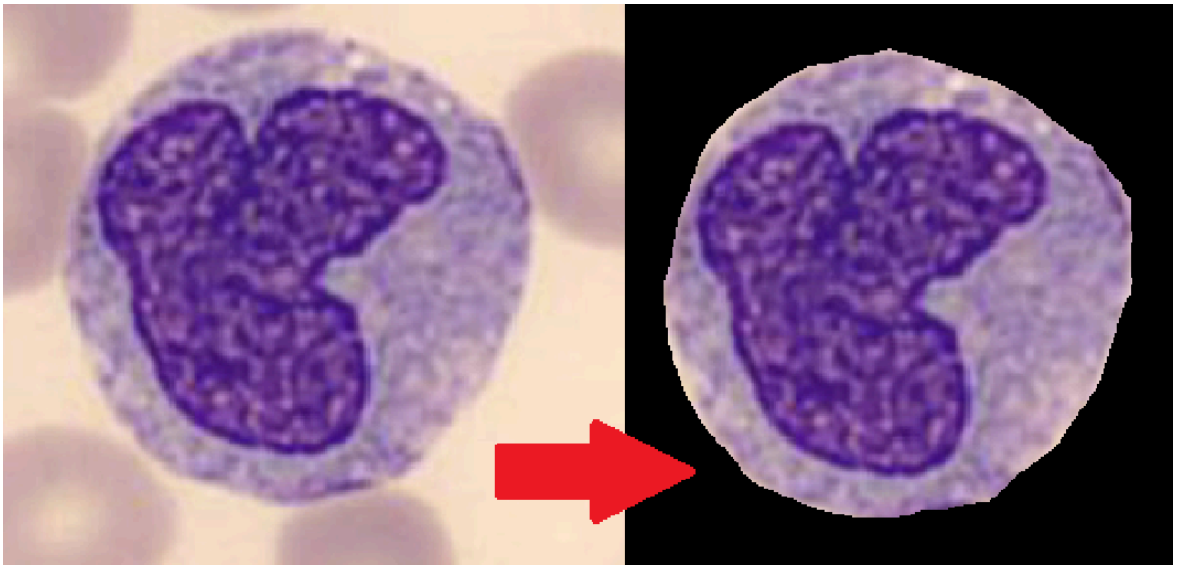


Figure 12: Example of possible future use of image segmentation

Finally, being able to streamline the unlearning process and simplify its usage overall would help for its further use alongside medical professionals. In its current state, the process is geared toward tech-literate and software experienced individuals, with the use of GitHub and a large variety of necessary packages to ensure all the models are capable of being run and effectively used. By implementing a more streamlined, simple process through a future developed software, it would allow for less experienced users who may require the information to utilize and possibly help further progress the models.

7 Conclusion

In this project, we have embarked on a journey to integrate advanced machine unlearning techniques, specifically Saliency Unlearning (SalUn), into the realm of medical diagnostics. Our goal was to enhance the reliability and ethical handling of AI within the healthcare sector by addressing critical challenges associated with data privacy and bias in AI systems.

Our research underscores the potential of AI to transform medical diagnostics through more accurate and privacy-conscious methods. By applying various unlearning techniques, we aimed to refine AI models to disregard irrelevant or sensitive attributes inadvertently captured during their training phase. This approach not only mitigates the risk of perpetuating existing biases but also enhances the overall security of medical data.

The implementation of SalUn has demonstrated promising results in maintaining high diagnostic accuracy while effectively forgetting. This balance is crucial in medical applications where the stakes are exceptionally high. The refined models are poised to offer healthcare professionals tools that are not only technologically advanced but also ethically sound, thereby fostering greater trust in AI-assisted diagnostics.

However, our journey does not end here. The field of AI, especially in sensitive areas like healthcare, demands continuous evolution. Future work will involve further refining unlearning techniques, expanding their applicability, and continuously testing against new data sets to ensure they remain robust against emerging challenges.

By pushing the boundaries of what AI can achieve in medical diagnostics, this project contributes to a future where technology and ethics coalesce to improve patient outcomes significantly. Our ongoing commitment to improvement and adaptation will ensure that the innovations we've spearheaded will continue to lead at the forefront of technological advancements in healthcare.

8 Reflection on Project Management

8.1 Project Scope

The project scope encompasses the following components:

1. Collection of Public Medical Data: Gathering publicly available medical datasets suitable for image classification tasks, ensuring adherence to data privacy and ethical standards.
2. Data Analysis: Conducting comprehensive analysis on the collected medical data to understand its distribution, characteristics, and potential biases.
3. Implementation of State-of-the-Art Unlearning Model: Developing and deploying a cutting-edge unlearning model tailored for medical image classification, capable of adapting to evolving data distributions.
4. Integration of Unlearning Methods: Incorporating various unlearning techniques into the model framework to enable continuous refinement and adaptation.
5. Integration of SalUN Technique: Introducing the SalUN technique into the unlearning process to selectively forget previously learned features, enhancing model adaptability and robustness.
6. Model Testing and Optimization: testing the developed model through testing on diverse datasets, followed by optimization to enhance the performance.
7. Documentation of Methodology and Results: Thorough documentation of the experimental methodology, including model architecture, and training procedures, along with comprehensive reporting of results
8. Integration with Clinical Workflow: Exploring integration possibilities of the developed model with existing clinical workflows, ensuring seamless adoption and utility in real-world medical settings.

9. Ethical Considerations and Bias Mitigation: Addressing ethical considerations surrounding the deployment of AI models in healthcare, including bias mitigation strategies and transparency in decision-making processes.

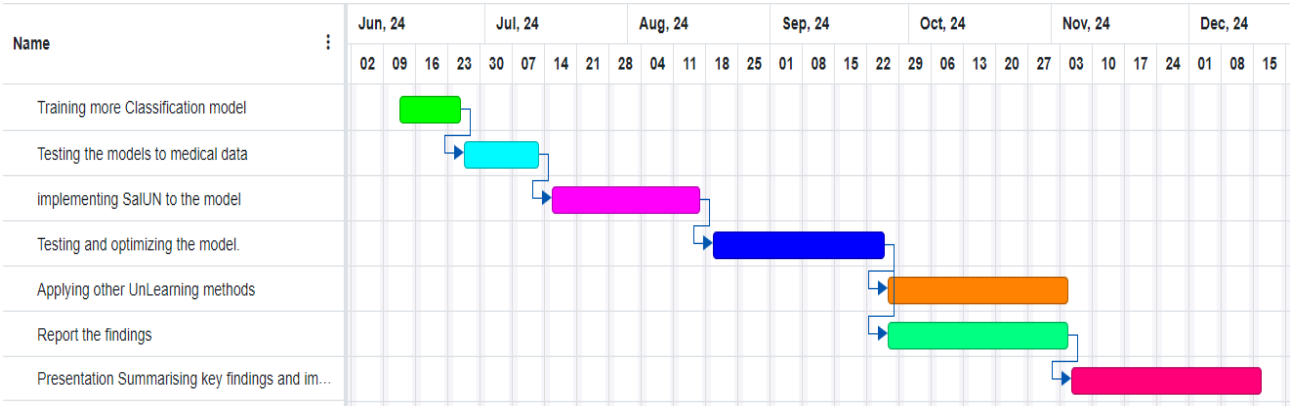
8.1.2 out-of-scope

The project explicitly excludes the following aspects:

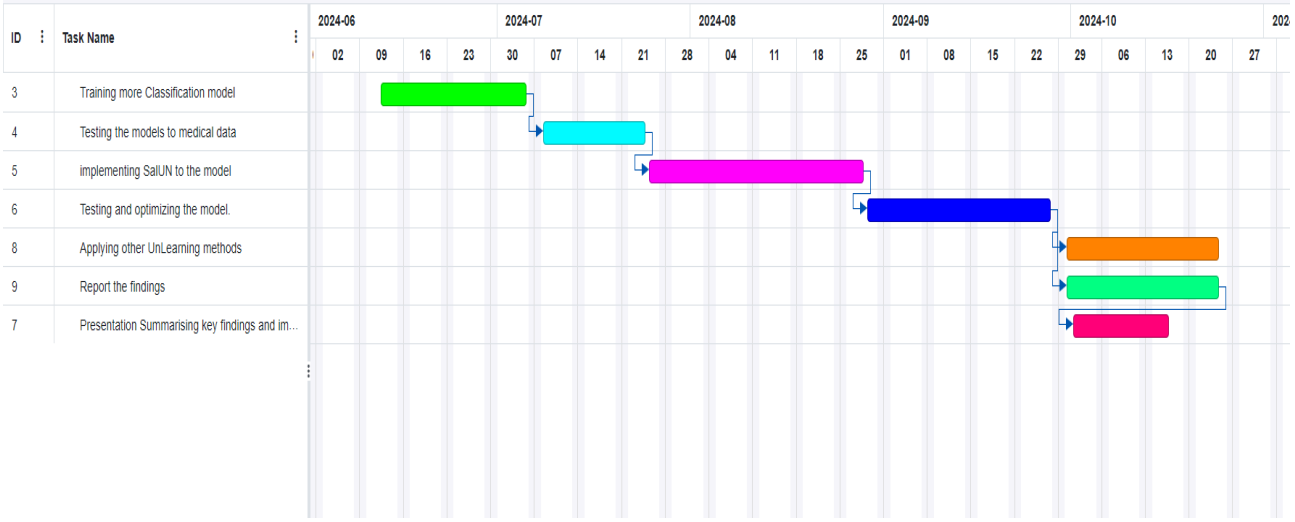
1. Usage of Non-Public Medical Data: Restricting access to confidential medical datasets that are not publicly available.
2. Data Alteration: Avoiding any manipulation of data aimed at influencing the classification model's diagnostic capabilities, ensuring integrity and reliability of the classification process.
3. Development of Hardware Solutions

8.2 Project Plan & Timeline

Initial Gantt Chart - As projected in April



Updated Gantt Chart - As the project has developed, we have finished things well before originally planned, and thus have been able to readjust our plans accordingly.



8.3 Reflection on Project

Throughout the course of the project, our team demonstrated a commendable level of commitment and adaptability, which were crucial in navigating the complexities encountered. Despite facing significant challenges, such as unexpected technical issues and shifts in project scope, the team's ability to pivot and adapt was exemplary.

Team Performance: Our team maintained a high level of communication and collaboration throughout the project. This strong teamwork was instrumental in overcoming the technical issues that arose unexpectedly. The initial stages were challenging due to members' unfamiliarity with specific aspects of the project. However, through dedicated mentorship and collaborative problem-solving, the team enhanced their technical skills and project understanding significantly.

Project Successes: One of the key successes of the project was the ability to adjust the scope and objectives in response to unforeseen technical limitations and resource constraints. This flexibility ensured that we still met the critical goals of the project, albeit through altered pathways. Moreover, the quality of the output remained high, demonstrating the team's resilience and commitment to excellence.

Improvement Strategies for Future Projects: For future endeavors, we plan to implement a more robust planning phase to anticipate potential roadblocks better and allocate resources more effectively. Establishing clear roles from the outset will also be a priority to minimize overlaps and streamline communications. Additionally, incorporating regular review points in the timeline will allow us to make necessary adjustments more dynamically and maintain project momentum.

Changes from the Original Plan:

- **Timeline Adjustments:** Our project timeline did experience some slippage due to the unexpected technical issues and the learning curve associated with new tools. To accommodate this, we extended the project timeline by a week to ensure all tasks and subtasks were thoroughly completed.
- **Scope Modification:** Certain aspects of the original project scope were scaled back after a project review highlighted resource limitation. This strategic decision allowed us to focus more on the core objectives, ensuring depth over breadth in our project deliverables.
- **Pivoting Strategy:** Midway through the project, we pivoted our approach to one of the sub-tasks, shifting from a purely technical solution to a more hybrid approach that incorporated user feedback more directly. This pivot not only enhanced the final product but also increased stakeholder satisfaction.

Despite the deviations from the initial plan and the challenges faced, the project was a valuable learning experience that highlighted the importance of flexibility, thorough planning, and proactive problem-solving. The insights gained from this project will undoubtedly inform our strategies in future projects, aiming for continual improvement and success.

9 References

[1] S. Liu *et al*, "Rethinking Machine Unlearning for Large Language Models," *ArXiv.Org,* 2024. Available: https://www.proquest.com/working-papers/rethinking-machine-unlearning-large-language/docview/2926949850/se-2.

[2] C. Fan *et al*, "SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation," *ArXiv.Org,* 2024. Available: https://www.proquest.com/working-papers/salun-empowering-machine-unlearning-via-gradient/docview/2879448278/se-2.

[3] A. Mantelero, "The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'," Computer Law & Security Review, vol. 29, no. 3, pp. 229–235, 2013.

[4] Howard, R. J. (2024). CPPA welcomes legislation on browser privacy settings. *Cybersecurity Policy Report,* , 1. Retrieved from https://www.proquest.com/trade-journals/cppa-welcomes-legislation-on-browser-privacy/docview/2958054753/se-2

[5] DeLeon, C. (2023). Calif. lawmakers eye bill to extend statute of limitations for CCPA violations. *Cybersecurity Policy Report,* , 1. Retrieved from https://www.proquest.com/trade-journals/calif-lawmakers-eye-bill-extend-statute/docview/2854618782/se-2

[6] A. Sekhari *et al*, "Remember What You Want to Forget: Algorithms for Machine Unlearning," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/remember-what-you-want-forget-algorithms-machine/docview/2498814784/se-2.

[7] N. Carlini *et al*, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," *ArXiv.Org,* 2019. Available: https://www.proquest.com/working-papers/secret-sharer-evaluating-testing-unintended/docview/2071623010/se-2.

[8] K. Leino and M. Fredrikson, "Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/stolen-memories-leveraging-model-memorization/docview/2248796806/se-2.

[9] J. Kline, "A right to be 'forgotten' is a right to censor others; In an ongoing series, Post contributors reflect on a recent European Court of Justice ruling that Internet search providers must remove links to embarrassing information. Should Canadian citizens have a "right to be forgotten"?" *National Post,* 2014. Available: https://www.proquest.com/newspapers/right-be-forgotten-is-censor-others-ongoing/docview/1542603872/se-2.

[10] E. Volokh, "Freedom of speech and information privacy: The troubling implications of a right to stop people from speaking about you," *Stanford Law Rev.,* vol. 52, *(5),* pp. 1049-1124, 2000. Available: https://www.proquest.com/scholarly-journals/freedom-speech-information-privacy-troubling/docview/224082165/se-2.

[11] N. Carlini *et al*, "Extracting Training Data from Diffusion Models," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/extracting-training-data-diffusion-models/docview/2771185010/se-2.

[12] S. Corbett-Davies *et al*, "The Measure and Mismeasure of Fairness," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/measure-mismeasure-fairness/docview/2092745895/se-2.

[13] M. Hardt, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning," *ArXiv.Org,* 2016. Available: https://www.proquest.com/working-papers/equality-opportunity-supervised-learning/docview/2080435426/se-2.

[14] R. Gandikota *et al*, "Unified Concept Editing in Diffusion Models," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/unified-concept-editing-diffusion-models/docview/2858805068/se-2.

[15] R. Gandikota *et al*, "Erasing Concepts from Diffusion Models," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/erasing-concepts-diffusion-models/docview/2786647232/se-2.

[16] J. Rando *et al*, "Red-Teaming the Stable Diffusion Safety Filter," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/red-teaming-stable-diffusion-safety-filter/docview/2724079465/se-2.

[17] P. Schramowski *et al*, "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/safe-latent-diffusion-mitigating-inappropriate/docview/2734835248/se-2.

[18] A. Thudi *et al*, "On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/on-necessity-auditable-algorithmic-definitions/docview/2585639336/se-2.

[19] B. Biggio, B. Nelson and P. Laskov, "Poisoning Attacks against Support Vector Machines," *ArXiv.Org,* 2013. Available: https://www.proquest.com/working-papers/poisoning-attacks-against-support-vector-machines/docview/2085538772/se-2.

[20] M. Heikkilä, "This new data poisoning tool lets artists fight back against generative AI," MIT Technology Review, 23 October 2023. [Online]. Available: https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/ [Accessed 2024]

[21] Y. Liu *et al*, "Backdoor Defense with Machine Unlearning," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/backdoor-defense-with-machine-unlearning/docview/2622690958/se-2.

[22] R. Cummings *et al*, "Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment," *ArXiv.Org,* 2024. Available: https://www.proquest.com/working-papers/advancing-differential-privacy-where-we-are-now/docview/2802174113/se-2.

[23] J. Ye *et al*, "Learning with Recoverable Forgetting," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/learning-with-recoverable-forgetting/docview/2691612042/se-2.

[24] T. Shaik *et al*, "Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy," *ArXiv.Org,* 2024. Available: https://www.proquest.com/working-papers/exploring-landscape-machine-unlearning/docview/2812870591/se-2.

[25] J. Xu *et al*, "Machine Unlearning: Solutions and Challenges," *ArXiv.Org,* 2024. Available: https://www.proquest.com/working-papers/machine-unlearning-solutions-challenges/docview/2850925226/se-2. DOI: https://doi.org/10.1109/TETCI.2024.3379240.

[26] T. N. Thanh *et al*, "A Survey of Machine Unlearning," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/survey-machine-unlearning/docview/2712096049/se-2.

[27] A. Madry *et al*, "Towards Deep Learning Models Resistant to Adversarial Attacks," *ArXiv.Org,* 2019. Available: https://www.proquest.com/working-papers/towards-deep-learning-models-resistant/docview/2076555528/se-2.

[28] Z. Izzo *et al*, "Approximate Data Deletion from Machine Learning Models," *ArXiv.Org,* 2021. Available:

https://www.proquest.com/working-papers/approximate-data-deletion-machine-learning-models/docview/2363574402/se-2.

[29] S. K. Mathivanan *et al*, "Employing deep learning and transfer learning for accurate brain tumor detection," *Scientific Reports (Nature Publisher Group),* vol. 14, *(1),* pp. 7232, 2024. Available: https://www.proquest.com/scholarly-journals/employing-deep-learning-transfer-accurate-brain/docview/3003352408/se-2. DOI: https://doi.org/10.1038/s41598-024-57970-7.

[30] Z. Cao *et al*, "Machine Unlearning Method Based On Projection Residual," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/machine-unlearning-method-based-on-projection/docview/2720676972/se-2.

[31] J. Jia *et al*, "Model Sparsity Can Simplify Machine Unlearning," *ArXiv.Org,* 2024. Available: https://www.proquest.com/working-papers/model-sparsity-can-simplify-machine-unlearning/docview/2799913534/se-2.

[32] M. Chen *et al*, "When Machine Unlearning Jeopardizes Privacy," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/when-machine-unlearning-jeopardizes-privacy/docview/2399006763/se-2.

[33] N. Carlini *et al*, "The Privacy Onion Effect: Memorization is Relative," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/privacy-onion-effect-memorization-is-relative/docview/2679947643/se-2.

[34] R. Chourasia and N. Shah, "Forget Unlearning: Towards True Data-Deletion in Machine Learning," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/forget-unlearning-towards-true-data-deletion/docview/2725745438/se-2.

[35] I. Shumailov *et al*, "Manipulating SGD with Data Ordering Attacks," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/manipulating-sgd-with-data-ordering-attacks/docview/2515931200/se-2.

[36] A. Mahadevan and M. Mathioudakis, "Certifiable Machine Unlearning for Linear Models," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/certifiable-machine-unlearning-linear-models/docview/2546803247/se-2.

[37] J. Brophy and D. Lowd, "Machine Unlearning for Random Forests," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/machine-unlearning-random-forests/docview/2442449570/se-2.

[38] A. Ginart *et al*, "Making AI Forget You: Data Deletion in Machine Learning," *ArXiv.Org,* 2019. Available: https://www.proquest.com/working-papers/making-ai-forget-you-data-deletion-machine/docview/2256346240/se-2

[39] L. Bourtoule *et al*, "Machine Unlearning," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/machine-unlearning/docview/2323284329/se-2.

[40] K. Koch and M. Soll, "No Matter How You Slice It: Machine Unlearning with SISA Comes at the Expense of Minority Classes," *The Institute of Electrical and Electronics Engineers, Inc.(IEEE) Conference Proceedings,* 2023. Available: https://www.proquest.com/conference-papers-proceedings/no-matter-how-you-slice-machine-unlearning-with/docview/2821717644/se-2. DOI: https://doi.org/10.1109/SaTML54575.2023.00047.

[41] A. Mahadevan and M. Mathioudakis, "Certifiable Machine Unlearning for Linear Models," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/certifiable-machine-unlearning-linear-models/docview/2546803247/se-2.

[42] J. Brophy and D. Lowd, "Machine Unlearning for Random Forests," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/machine-unlearning-random-forests/docview/2442449570/se-2.

[43] A. Ginart *et al*, "Making AI Forget You: Data Deletion in Machine Learning," *ArXiv.Org,* 2019. Available: https://www.proquest.com/working-papers/making-ai-forget-you-data-deletion-machine/docview/2256346240/se-2

[44] L. Bourtoule *et al*, "Machine Unlearning," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/machine-unlearning/docview/2323284329/se-2.

[45] K. Koch and M. Soll, "No Matter How You Slice It: Machine Unlearning with SISA Comes at the Expense of Minority Classes," *The Institute of Electrical and Electronics Engineers, Inc.(IEEE) Conference Proceedings,* 2023. Available: https://www.proquest.com/conference-papers-proceedings/no-matter-how-you-slice-machine-unlearning-with/docview/2821717644/se-2. DOI: https://doi.org/10.1109/SaTML54575.2023.00047.

[46] H. Yan *et al*, "ARCANE: An Efficient Architecture for Exact Machine Unlearning", *ijcai.org*, 2022. Available: https://www.ijcai.org/proceedings/2022/0556.pdf

[47] C. Guo *et al*, "Certified Data Removal from Machine Learning Models," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/certified-data-removal-machine-learning-models/docview/2313804933/se-2.

[48] A. Warnecke *et al*, "Machine Unlearning of Features and Labels," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/machine-unlearning-features-labels/docview/2565273597/se-2.

[49] A. Golatkar, A. Achille and S. Soatto, "Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/eternal-sunshine-spotless-net-selective/docview/2314126820/se-2.

[50] C. Dwork, A. Roth, "The Algorithmic Foundations of Differential Privacy", *cis.upenn.edu*, 2014. Available: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

[51] A. Golatkar, A. Achille and S. Soatto, "Forgetting Outside the Box: Scrubbing Deep Networks of Information Accessible from Input-Output Observations," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/forgetting-outside-box-scrubbing-deep-networks/docview/2374915246/se-2.

[52] A. K. Tarun *et al*, "Fast Yet Effective Machine Unlearning," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/fast-yet-effective-machine-unlearning/docview/2598841449/se-2. DOI: https://doi.org/10.1109/TNNLS.2023.3266233.

[53] R. Cook and S. Weisberg, "Residuals and Influence in Regression", *conservancy.umn.edu*, 1982. Available: https://conservancy.umn.edu/handle/11299/37076

[54] W. K. Pang and P. Liang, "Understanding Black-box Predictions via Influence Functions," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/understanding-black-box-predictions-via-influence/docview/2076041884/se-2.

[55] Y. Wu, E. Dobriban and S. B. Davidson, "DeltaGrad: Rapid retraining of machine learning models," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/deltagrad-rapid-retraining-machine-learning/docview/2418458522/se-2.

[56] A. Golatkar *et al*, "Mixed-Privacy Forgetting in Deep Networks," *ArXiv.Org,* 2021. Available: https://www.proquest.com/working-papers/mixed-privacy-forgetting-deep-networks/docview/2473565809/se-2.

[57] A. Thudi *et al*, "Unrolling SGD: Understanding Factors Influencing Machine Unlearning," *ArXiv.Org,* 2022. Available: https://www.proquest.com/working-papers/unrolling-sgd-understanding-factors-influencing/docview/2577596791/se-2.

[58] L. Graves, V. Nagisetty and V. Ganesh, "Amnesiac Machine Learning," *ArXiv.Org,* 2020. Available: https://www.proquest.com/working-papers/amnesiac-machine-learning/docview/2453524097/se-2.

[59] M. Chen *et al*, "Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary," *The Institute of Electrical and Electronics Engineers, Inc.(IEEE) Conference Proceedings,* 2023. Available: https://www.proquest.com/conference-papers-proceedings/boundary-unlearning-rapid-forgetting-deep/docview/2855713796/se-2. DOI: https://doi.org/10.1109/CVPR52729.2023.00750.

[60] E. Zhang *et al*, "Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models," *ArXiv.Org,* 2023. Available: https://www.proquest.com/working-papers/forget-me-not-learning-text-image-diffusion/docview/2793244004/se-2.

[61] Liu R et al, "DeepDRiD: Diabetic Retinopathy-Grading and Image Quality Estimation Challenge," 2022. DOI: 10.1016/j.patter.2022.100512.

[62] J. Yang *et al*, "MedMNIST," *medmnist.com,* 2024. Available: https://medmnist.com/

[63] B. Henrique et al., "Trust in artificial intelligence: Literature review and main path analysis," *Scientific Reports*, vol. 2, no. 2, pp. 100043, 2024. Available: https://www.sciencedirect.com/science/article/pii/S2949882124000033

10 Appendices

Appendix A: Project Risk Assessment

57568	RISK DESCRIPTION	STATUS	TREND	CURRENT	RESIDUAL
	ENG4701_CL_2024_S1_ThomasLiefman_Unlearning For Medical Image Classification	Live		Medium	Not Assessed

RISK TYPE
3. Risk Assessment Template or Framework

RISK OWNER	RISK IDENTIFIED ON	LAST REVIEWED ON	NEXT SCHEDULED REVIEW
THOMAS ABRAHAM LIEFMAN	25/03/2024	25/03/2024	25/03/2027

RISK FACTOR(S)	EXISTING CONTROL(S)	CURRENT	PROPOSED CONTROL(S)	TREATMENT OWNER	DUE DATE	RESIDUAL
Ergonomic risks resulting in back pain, neck pain, and carpal tunnel syndrome.	Control: Be mindful for good posture. Take regular breaks and get the body moving. Control Effectiveness:	Medium	Adjusting monitor height / brightness and chair height to make computer setup more ergonomic.	THOMAS ABRAHAM LIEFMAN	01/04/2024	
Electrocution from power cords.	Control: Protective casing around wiring. Surge protection in power board. Control Effectiveness:	Medium	Make sure cable management is neat.	THOMAS ABRAHAM LIEFMAN	01/04/2024	

Appendix B: Risk Management Plan

Occupational Health and Safety

Due to the nature of our work, the general risks to the group members involved are quite low. The primary risk is the ergonomics of the members due to extended periods of time at computers. Injuries such as long

term back pain and neck damage can occur during long periods of computer usage, and is generally increased due to the extended periods of time while coding and researching for this project. This will be directly managed through continuous monitoring of physical position and ensuring all members stretch and have periodic breaks to reduce strain on the body.

Another safety concern is an extremely low probability of electrocution while performing tasks at a computer. Due to the extremely low chance of this occurring, this is generally not a worry for the team whilst performing their tasks, however it must be considered due to its consequences being quite severe, that being intense injury up to and including death. This is specifically managed through ensuring proper cable management and the use of properly insulated computers and power sources.

The project risk assessment performed using the Monash University Safety and Risk Analysis Hub has determined that the overall risk level of the project is medium, due to the two risks considered above. This assessment can be found in Appendix A.

Project risks

More broadly speaking, there are two major risks to the completion of the project which both work in tangent with each other. The first is in regards to the computational limitations of the group members personal computer or otherwise the machines available to them. These limitations directly affect the timeline of the project, as time can be lost from lower powered machines being used to perform these high-requirement computational tasks, due to the general nature of neural networks.

In a similar vein, there is a general risk of failing to complete tasks within the required project timeline. This acts as a more non-specific risk in terms of the project's tasks itself, and is instead a risk to the group members towards finishing the project within the projected timeline as outlined in section 5.2.

Liability Risks

The nature of working with medical imagery will contain inherent concerns in regards to the confidentiality of all data that is used within the project. As such, the group will ensure that all data used within the development and testing of the neural network will only use publicly available datasets, which will involve entirely de-identified data and as such will remove the risk of confidentiality.

Coinciding with the possible breaching of personal data, all research will be reviewed by the team to ensure no members breach research protocol regarding this information.

Appendix C: Sustainability Plan

The United Nations Sustainable Development Goal 3 aims to ensure healthy lives and promote wellbeing for all at all ages. Our project aims to ensure the continued health and wellbeing of all by improving health outcomes and ensuring all individuals receive sufficient medical treatment. By improving fairness and accuracy within medical image classification, we are able to contribute towards this goal.

This plan evaluates the sustainability implications of our project, with the focus of how they relate to the UN Sustainable Development Goal 3: Good Health and Wellbeing, which attempts to guarantee healthy lives and advance wellbeing for all of us. The research aims to improve the fairness and accuracy of AI medical diagnoses in order to directly contribute to this goal.

1. Impact on Community and Environment:

UN SDG 3 Target 3.d aims to strengthen the capacity of all nations for early warning, risk reduction and management of national and global health risks. Our research focuses on being able to effectively implement an unbiased image classification model to improve the early detection of diseases and thus reduce the risks to the community. This research also allows for easy access to improved medical identification to at-risk nations and communities, which will strengthen the capacity of developing countries through improved medical support.

However, there is an imposed risk of creating an overreliance on technological identification. Over relying on these classification models can lead to worse health outcomes and poorer risk identification if there is no supervision by trained professionals. Medical decisions should be made by medical professionals and these models aim to support these decisions by providing further information and must be cautiously implemented to ensure they are not the only source of information.

Proactive Risk Management: To address potential risks such as data privacy concerns, proactive strategies including strict compliance with data protection laws are implemented, safeguarding community trust and ensuring a sustainable healthcare environment.

2. Positive Impacts:

i.Social and Cultural: Enhances trust in AI diagnostics by ensuring data privacy and reducing biases, promoting equitable healthcare.

ii. Health and Safety: Improves diagnostic accuracy, potentially leading to better patient outcomes and safety.

3. Negative Impacts:

Social: There is a risk of technological reliance potentially reducing the role of human judgment in critical health decisions.

3. Economic, Social, and Cultural Incorporation:

SDG 3 Target 3.8: Achieve universal health coverage, including access to quality essential healthcare services and access to safe, effective, quality, and affordable essential medicines for all. The project lowers barriers to accurate medical diagnostics, a cornerstone of universal health coverage, by reducing bias and improving data handling techniques. This ensures equitable access to healthcare diagnostics across different cultures.

Stakeholder Engagement: Engaging healthcare professionals, patients, and policymakers as stakeholders ensures that the project addresses the varied needs of the community it serves, fostering inclusive health service delivery.

4. Efficiency and Environmental Considerations:

SDG 3 Target 3.9: Reduce the number of deaths and illnesses from hazardous chemicals and air, water, and soil pollution and contamination. By enhancing the efficiency of medical classifications, the project indirectly contributes to environmental sustainability. More efficient data processing reduces the computational load, thus decreasing the energy consumption of healthcare IT systems and contributing to less overall pollution. Sustainable Resource Use: The optimization of algorithm efficiency not only aligns with SDG 12 (Responsible Consumption and Production) by promoting efficient resource use in technology

but also supports sustainable practices within the healthcare sector. Overall, this target is effectively progressed towards by this project, with the aims of improving the early identification of non-communicable diseases and thus improving the prevention and treatment of these diseases. Further, this research will improve the health outcomes of developing countries through an easy to access support point for professionals in struggling areas that need cheap and efficient help. As such, this research will improve and aims to progress the UN SDG 3.

UN SDG 3 Target 3.4 aims to reduce the premature mortality rate of non-communicable diseases through prevention and treatment. Our research targets creating AI models that are capable of identifying diseases and other anomalies that may appear in imagery that is unnoticeable to the naked eye. This directly promotes improving health outcomes and creates progress towards this target.

This will have overall positive impacts on the greater community as improving health outcomes benefits all. However, a major point of concern is ensuring no direct impacts of false negatives whilst utilising this technology. This would directly lead to negative health outcomes within the community and thus is an area of concern that will be closely monitored throughout testing to ensure the positive outcomes from this research.

By addressing the specific targets of SDG 3 through innovative approaches in medical classification, this project not only contributes to enhancing health outcomes but also ensures that these advancements are achieved sustainably and equitably. The project's focus on unlearning biases and improving system efficiency exemplifies a deep commitment to sustainable UN principles, promoting a healthier future for all.

Appendix D: Generative AI Statement

Generative AI use in FYP B (ENG4702)

The responses to this form will need to be copied and put into an appendix in your Final Report.

Email *

zwil0005@student.monash.edu

Name *

Zachary Wilson

Campus

- (●) Clayton
- ( ) Malaysia

Host Department

- ( ) Chemical and Biological Engineering
- ( ) Civil Engineering
- (●) Electrical and Computer Systems Engineeing
- ( ) Materials Science Engineering
- ( ) Mechanical and Aerospace Engineering
- ( ) Software Engineering
- ( ) Robotics and Mechatronics Engineering

Supervisor

A/Prof. Mehrtash Tafazzoli Harandi

This project has been conducted using AI tools *

- ( ) In this assessment, there will be no use of generative artificial intelligence (AI). All content in relation to the assessment task has been produced by the authors.
- ( ) In this assessment, the following generative AI will be used for the purposes nominated in part 2. (Please note: any use of generative AI must be appropriately acknowledged - see Learn HQ)
- ( ) In this assessment, AI writing assistants (e.g., Grammarly, Writesonic, Quillbot, Microsoft Editor) will be the only form of Generative AI used.
- (•) This project involves the development or authoring of Unique Generative AI, Unique operation of commercially available Generative AI OR Unique non-generative AI (Machine Learning, Artificial Neural Network, Logistic Regression, etc.)

Developing AI

In question 1, you answered your project involves the development or authoring of AI. Please choose the specific way you did this.

- ( ) Unique Generative AI
- ( ) Unique operation of commercially available Generative AI
- (•) Unique non-generative AI (Machine Learning, Artificial Neural Network, Logistic Regression, etc.)
- ( ) Other: .....

How did you use this technology?

*

- [x] As a fundamental aspect of the study (i.e., this is a study centred on generative AI, human interaction, associated ethics, etc.)
- [ ] Audio Transcription
- [x] Coding/Scripting
- [ ] For the operation of robotics
- [ ] Generation of novel content - Datasets
- [ ] Generation of novel content - Graphics/Images
- [ ] Generation of novel content - Video
- [ ] Generation of novel content - Writing
- [ ] Idea generation
- [ ] Initial research
- [ ] Machine Language Translation
- [ ] Mathematics
- [ ] Paraphrasing
- [ ] Proofreading
- [ ] Text Analytics
- [ ] Text Summarisation
- [ ] Thematic analysis
- [ ] Visualisation (of data)
- [ ] Writing assistance
- [ ] Other: .....

How was the Generative AI response validated?

We compared results obtained to results of existing literature to ensure that all AI results were consistent before testing our own version further for the project.

Permissions

The use of Generative AI has been discussed with and approved by my academic supervisor. *

- (●) Yes
- ( ) No

End

Thank you for completing this form - your responses will be emailed to you for your Progress Report