FAKULTÄT FÜR
INFORMATIK

Otto-von-Guericke-University Magdeburg

Faculty of Computer Science

Generating Plausible Counterfactual Images using Generative Network

Master Thesis

Author:

Mahantesh Vishvanath Pattadkal

Examiner and Supervisor:

Prof. Ernesto De Luca

2nd Examiner:

Prof. Andreas Nürnburger

Supervisor:

Soumick Chatterjee
Erasmo Purificato

Magdeburg, 23.06.2022

Contents

Acknowledgements 3

Abstract 4

Index of Notation 5

1 Introduction 6
- 1.1 Introduction to Explainable AI (XAI) . . . 7
- 1.2 Introduction to Counterfactual Explanations . . . 9
- 1.3 Motivation . . . 10
- 1.4 Goals . . . 13
- 1.5 Main Contribution . . . 13
- 1.6 Thesis Outline . . . 13

2 Related Work 15
- 2.1 Convolutional Neural Network for Classification . . . 15
- 2.2 Conterfactual Explanation . . . 15
- 2.3 Conterfactual Explanation in different data domains . . . 16
- 2.4 Plausibility of generated counterfactual images . . . 17

3 Fundamentals and Concepts 19
- 3.1 Convolutional Neural Network for Classification . . . 19
- 3.2 Generative Adversarial Network (GAN) . . . 20
  - 3.2.1 Deep Convolutional GAN (DC-GAN) . . . 20
  - 3.2.2 Wasserstein GAN (WGAN) and WGAN with Gradient Penalty (WGAN-GP) . . . 21
- 3.3 GAN based methods for Counterfactual Images . . . 22
  - 3.3.1 Baseline 1: Constrained Minimum Edit Method (C-Min Edit) . . . 24
  - 3.3.2 Baseline 2: PlausIble Exceptionality-based Contrastive Explanations (PIECE) . . . 24
- 3.4 Evaluation Metrics . . . 25
  - 3.4.1 IMI1 . . . 26
  - 3.4.2 MC-Mean and MC-Std . . . 26
  - 3.4.3 U-Net based perceptual loss . . . 26
  - 3.4.4 Fréchet Inception Distance (FID) . . . 27

4 Method 28
- 4.1 Proposed GAN architectitures for high quality images . . . 28
  - 4.1.1 U-Net and Wasserstein GAN (U-Net WGAN) . . . 28
  - 4.1.2 ReconResNet and Wasserstein GAN (ReconResNet WGAN) . . . 29

4.1.3 U-Net and Inception Net Wasserstein GAN (U-Net Inception Net WGAN) . . . 30
4.2 Proposed method for Plausible Counterfactual Images . . . 31
4.2.1 Step 1: Identify the GAN representation of the original image . . . 32
4.2.2 Step 2: Counterfactual Image generation . . . 32

5 Experiments and Evaluation **33**
5.1 Experiment 1: MNIST Handwritten Digits Dataset . . . 33
5.1.1 Dataset . . . 33
5.1.2 Experimental Setup . . . 34
5.1.3 Evaluation on correctly classified instances . . . 37
5.1.4 Evaluation on misclassified instances . . . 40
5.1.5 Journey of original image to counterfactual image . . . 44
5.1.6 Detecting the changes by analysing the transformation of images . . . 45
5.2 Experiment 2: BraTS MRI Dataset . . . 47
5.2.1 Dataset . . . 47
5.2.2 Experimental Setup . . . 48
5.2.3 Experimentation and Evaluation of GANs . . . 49
5.2.4 Evaluation on correctly classified instances . . . 51
5.2.5 Evaluation on misclassified instances . . . 54
5.2.6 Detecting the changes by analysing the transformation of images . . . 56

6 Discussion **60**
6.1 MNIST Dataset . . . 60
6.2 BraTS Dataset . . . 65

7 Conclusion **71**
7.1 Summary . . . 71
7.2 Limitations . . . 71
7.3 Future Work . . . 72

Bibliography **73**

Acknowledgements

Throughout this thesis work, I have received a great deal of guidance, supervision, and support. First and foremost, I would like to express my gratitude to my supervisors Soumick Chatterjee and Erasmo Purificato, for their expert guidance, constant support, encouragement, and patience right from the beginning of the thesis. I would especially like to thank them for accepting the thesis proposal and helping me fine-tune the research questions, and making sure that I do not divert from the topic at hand. The discussions we had throughout the thesis helped in bringing a varied dimension of thoughts and broadened my horizon to tackle and make tough decisions in a limited time interval.

I would also like to extend my gratitude to Prof. Ernesto De Luca for accepting my thesis research topic and taking me under his banner. I sincerely thank him for providing an excellent research environment and all the required resources to complete the thesis. Additionally, I would like to extend special thanks to researcher Eoin M Kenny for sharing his insights on the field of Explainable AI (XAI) and plausible counterfactual generation.

My family has also been a constant source of support for me during these trying times. I always felt calmed and energized after hearing their life advice and casual conversations. Thank you to my sister Manjusha Pattadkal and my friend Anish Singh for continuously encouraging me to finish my thesis. To conclude, I would like to thank my wife Shweta Akki for her continuous support and confidence that enabled me to successfully complete this thesis.

Abstract

Counterfactual Explanation is an Explainable AI (XAI) technique used to study the behaviour of the model in reference to a given instance, which is also termed as local explanations. In an image classification scenario where attributes of images are not known, there are various techniques used to generate counterfactual images, but the problem lies with the plausibility of these images. The generated images have a different prediction, but they appear to be out of the training data distribution. Therefore, the aim of this thesis is to explore possibilities of generating plausible counterfactual images by exploring the applicability of discriminator trained via Generative Adversarial Network (GAN) as a plausibility regulator that might ensure the generated images are closer to the training data distribution. Also, provide the user with a transformation map that be used to attain the counterfactual image from the given image

Index of Notation

In a lot of cases it makes sense to give an overview over your mathematical notation.

Mathematical

x	Point in 3D space
$\overrightarrow{\mathrm{xy}}$	Normalized direction vector from x to y
$\mathbf{v}$	Direction vector in 3D space
$\mathrm{p}_x, \mathbf{v}_x$	x component of point / vector
$\mathbf{v} \cdot \mathbf{w}$	Dot product of vectors $\mathbf{v}$ and $\mathbf{w}$
$(\mathbf{v} \cdot \mathbf{w})^+$	Dot product of vectors $\mathbf{v}$ and $\mathbf{w}$ with negative values clamped to zero
$\mathbf{v} \times \mathbf{w}$	Cross product of vectors $\mathbf{v}$ and $\mathbf{w}$
$\|\|\mathbf{v}\|\|$	Euclidean length of vector $\mathbf{v}$
$\hat{\mathbf{v}}$	Normalized vector $\mathbf{v}$

Quantities & Functions

A	**Area**
ω	**Solid Angle**
ϕ	**Radiant Flux**, light power
I	**Radiant Intensity**, flux density per solid angle
E	**Irradiance**, flux density per area
L	**Radiance**, flux density per area per solid angle
ρ	**Reflectance**, ratio between incoming and outgoing flux
f_r	**BRDF**, function on the relation between irradiance and outgoing radiance

1 Introduction

Classifying things, objects or places in specific categories has always been a pivotal task. Many of the real world use cases and applications rely on the concepts of classification. Classification has found its use cases in all fields and domains ranging from simple to complex applications. Imagine you visit the nearest grocery store, there you can find the staff members classifying the items into different categories and arranging them, this helps the customers to find items they are looking for. Now think of a complex use case, where you visit the hospital and you observe the doctors analysing the MRI scans and classifying them as tumour and non tumour. Before the insurgence of machine learning based applications, all these tasks were done manually, in simple use cases it was done by the humans will little training, but in cases of complex use cases like brain tumour detection, metal wear and tear detection, anomaly detection it was pretty expensive as it required a subject matter expert to classify them and training an individual to become subject matter expert was time taking and expensive at the same time. This gave the rise to usage of machine learning algorithms for the task of classification. These algorithms required data that contains label for each instance in it, this data was used for training the classification algorithms. After training these algorithms, they were able to predict the label of the categories. These machine learning algorithms were quite successful in cases where the data was structured and available as tabular datasets. Soon, these techniques found their applications in domains of unstructured data like text and images. classical machine learning algorithms were replaced by advanced machine learning techniques, this further motivated the research in the field of deep learning. The basic computational unit of a deep learning network is a neuron that takes a certain input and provides a transformed output. Multiple neurons are used together to form a layer, there are numerous layers used in succession to form one deep learning network. Initial research promised groundbreaking results by means of usage of deep learning architectures, soon it gave rise to number of different architectures explored. There were some deep learning architectures that promised great results when used in image classifications, while other set of networks were preferred in text classifications.

These deep learning architectures were expensive to be trained, but in turn they promised results that outperformed the existing techniques, this motivated the various business ranging from retails industry to pharmaceuticals industry to invest in the applications that use deep learning. In this fast-paced era of data science, the focus has been on building classification models that have astounding performance. Complex machine learning and deep learning are the techniques that are leveraged in most scenarios to attain groundbreaking results. These techniques have increased the performance of the classification task in various data domains like text, speech images or tabular data as well, but all of this has come at the cost of lack of interpretability. The models built using these techniques are inherently complex and difficult to interpret or explain, which is why they are termed as black-box. When these models are used in sensitive applications like banking, medical systems, or criminal justice the interpretability of the models takes the front seat:

users of the application have to understand the decision-making process of the models, no matter how accurate the AI model claims to be.

Consider the credit scoring situation, where the bank uses a complex machine learning or deep learning model to decide whether the loan application of the candidate has to be approved or rejected. In earlier phases, these decisions were taken by the committee of the bank, where the committed would analyse all the details of the candidate's application like ongoing loans, income, age or number of family members etc. After considering all these factors the committee would decide the creditworthiness of the candidate and in case if it rejects the loan, the candidate would be informed on why the bank chose to reject their application. Now the bank has aggregated all these past records in form of a tabular dataset and then train a complex machine learning or deep learning model to decide the status of the applications. The model is efficient in imitating the decision making based on the dataset provided for training. Now, if the bank rejects the loan application of the candidate based on the model's decision as shown in Fig 1.1, then it is critical to explain the decision made by the model as the candidate tend to ask questions "Why was I rejected?" or "What can I change in my application to get the loan approved?". Answering these questions is pivotal for both the stakeholders- candidate and bank, the banks can leverage this to showcase transparency of the decision taken by the banks. On the other hand, if these questions are answered the users can attain deep trust into the banking system and apply for the loan by accommodating the changes in their application. This also helps the candidate to remain loyal customer of the bank, as they got their application status quicker, owning to the usage of machine learning techniques and secondly the ease of explanations that boost their confidence in the functioning of the bank.

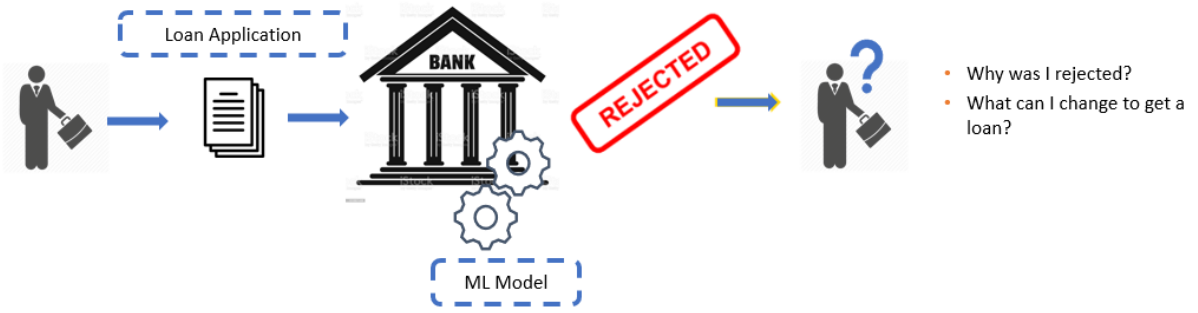


Figure 1.1: Credit Scoring using Machine Learning techniques

1.1 Introduction to Explainable AI (XAI)

The second decade of the twenty first century witnessed sudden rise in the number of applications using machine learning algorithms. The concepts of classification, regression and clustering models found their use cases in multiple domains and industries. Classification models were used for classifying the creditworthiness of the user in banking scenario, the screening rounds of the job decided the results of the recruitment process based on these models, further financial sectors used classification model for identifying fraudulent transactions. Similarly, regression was extensively used to understand complex

relation between market factors and prices of items, clustering techniques were used to develop models that can cluster customer groups and run tailored campaigns on these clusters. Further the advanced machine learning algorithms and deep learning techniques provided capability of building models that can operate in the avenues of unstructured data domains like text and image. This surge in use cases and applications of machine learning also gave rise to the questions being asked regarding these models. Questions like "Why should we trust the model?", "Does the model discriminate?", "Can you explain the decision making of the model?" started stirring discussions among the research communities. Very soon it was realised that the applications involving machine learning are great in mapping input to output, but very few of these algorithms provide transparency. Most of the complex models are built intricately, that even the developer of these models cannot answer the above questions. This motivated to the beginning of a research field termed as "Explainable AI" (XAI).

Explainable AI techniques aim to answer all the questions that deal with the transparency, fairness and trust of the model. The XAI techniques break down the explanations in two parts- global and local explanations. First the global explanations attempt to answer the question "What features are deemed important by the model?". In simple cases, the XAI techniques can leverage the usage of feature importance scores to provide such explanations. There are a few machine learning methods that are built using mathematical modelling like the linear regression model, these models generate weights for the features used in the training, these weights can be indicator in understanding what features are useful in taking decisions. But with the complex models, it is difficult to derive such insights. So in such cases, the XAI techniques use surrogate methods to answer these questions. A simple model is trained over the predictions of the complex model, these simple models provide features importance scores, these scores are used to interpret the feature importance of the complex models. Providing the answers to the above questions, helps the user in understanding what is deemed important by the model. If the application is used in sensitive domain like credit worthiness, it is critical for the application developers to understand that the model is fair across the sensitive features age, sex and religion. So in cases like these the global explanations helps the application developers to assess the fairness of the model before they are used in the real world. When the explanation is derived externally that is without exploiting the inner working of the model it is termed as post hoc explanations [1].

Local explanations attempt to answer the question "What features are important in making decision on the given instance?". Here, the focus of the XAI techniques is to understand the vicinity of the given instance and understand the factors that influence the decision making. Now, understanding why these explanations are important is crucial, the models overall behaviour based on all instances can be understood by global explanations but the model tends to behave differently based on the instance at hand, so it is crucial to answer what is relevant in this case rather than overall behaviour. Many time the user of the application is unhappy with the decision made by the model and tend to ask "What should I do now to get desired outcome?" in such cases it is irrelevant to answer the user based on global explanations, here the explanations should be tailored to user'S use case and hence they should be local explanations.

Consider the earlier case of creditworthiness, when the loan application is rejected the user tend to ask different set of questions and they can be answered by the global and local explanations shown in Fig. 1.2. The global explanations answers what are the most

important factors in rejecting the loan application based on overall instances, while the local explanations answer the user on what needs to be changes to attain desired results. Further the Fig 1.2 also names various XAI techniques used to attain global and local explanations.

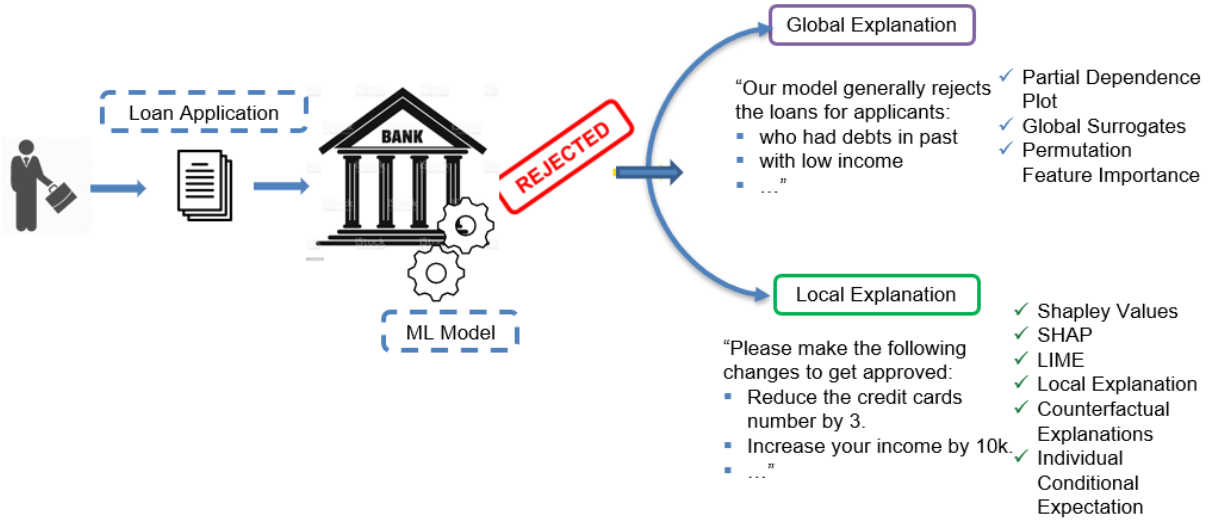


Figure 1.2: Counterfactual Explanations for Credit Scoring application

1.2 Introduction to Counterfactual Explanations

Counterfactual explanations [2] is one of the post-hoc explanation technique that attempts to justify the prediction of the model over a single instance which makes it a local explanation technique. Considering our credit scoring example, when the bank rejects the loan application, here the focus of the counterfactual explanation is to acquire understanding on what needs to be changed to get the loan approved. So, the counterfactual situation would be getting the loan approved, so as seen in Fig 1.3 the counterfactual explanation is understanding what changes are to be made to the profile of the user so that the counterfactual situation can be achieved. As seen in the Fig 1.3, the box highlighted in yellow shows the counterfactual explanation for the given scenario. These Counterfactual Explanations are intuitive and provide reasoning behind the prediction which makes it easier for the humans to comprehend in a lucid way.

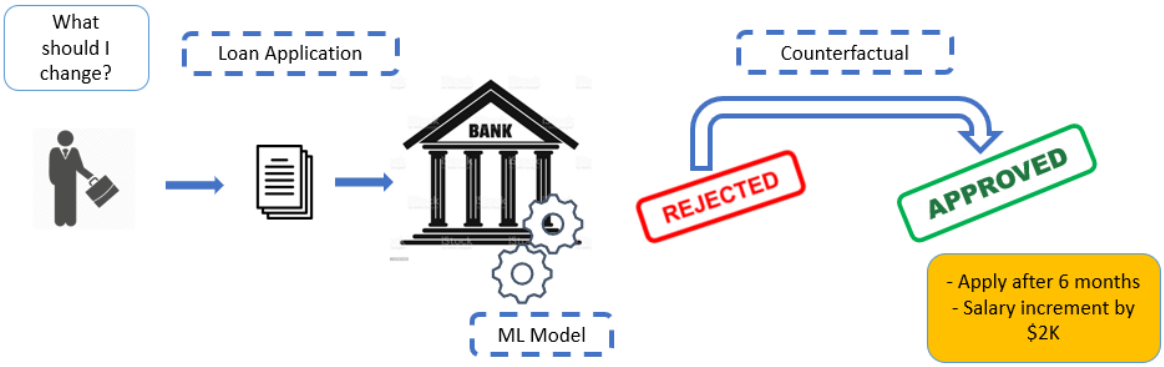


Figure 1.3: Counterfactual Explanations for Credit Scoring application

These techniques are also used in other domain of data like text where the users want to understand what needs to be changed in the text, so that the text classifier alters its prediction. Imagine you have a text classifier that classifies the tone of the text message, if the classifier predicts the text to be rude, it is crucial for the user to understand what needs to be changes in the text to bring down the tone. In applications like these where prediction alone is not enough for user experience, it has become integral to incorporate the explanations along with the trained model. Further, deep learning based applications for image classification are prevalent in all fields, medical practitioners are using it for tumour classification, identification of disjoint in X-Ray, engineers in production industry are using it for metal defect classification. For these applications, the explanation and transparency of the decision taken by the model is super critical which is why counterfactual are leveraged in the image classification domain. The usage of counterfactual answers the question "What should be changed in the given image to attain an alternate prediction?". This has helped the application developers to showcase the efficiency and transparency of the model, the users get to see real time about the changes to be made to the image to attain a certain prediction.

Counterfactual explanations are useful for all the stakeholders involved in the use case, the users attain more clarity in the system. The application developers can use counterfactual for assessing their models against bias and provide fair models to the real world, the business owners experience increased trust from users in their business on account of the transparency provided to the users. Counterfactual explanations have been used to identify bias in image classification scenarios, they also help in understanding regions of the images that are supremely important for a given prediction.

1.3 Motivation

The usage of Explainable AI techniques for interpreting the complex models is on the rise. All the sensitive applications using complex models need explanations to support the decision made by these models. The applications might use standard data structures like tabular data or unstructured data like text and images. Counterfactual Explanation is an XAI method used to provide explanations at an instance level, as in it helps us to

interpret the decision making of the model over a single instance. But these explanations have to be human acceptable, else they

Consider the earlier application of credit scoring, when the user asks for what changes are to be done from their end to get the loan approved, there are 2 counterfactual explanations generated as shown in Fig 1.4. The first explanation makes sense and the user accepts the decision made by the model, now look at the second explanation, if you use the inputs suggested by the second explanation, the complex machine learning model would change the status of the loan application of the user to "Accepted", but it does not convince the user. In other case it removes the trust of the user from the model, because the counterfactual explanations are suggesting the user to make some impossible changes to their profile. These kind of explanations can also take away the trust of the user from the system, which means these techniques would shatter the trust of the user rather than building trust into the system. This is one of the motivation to start exploring methods for deriving plausible counterfactual explanations. Generating counterfactual explanations that are plausible is important, so that the users can derive actual value from these explanations.

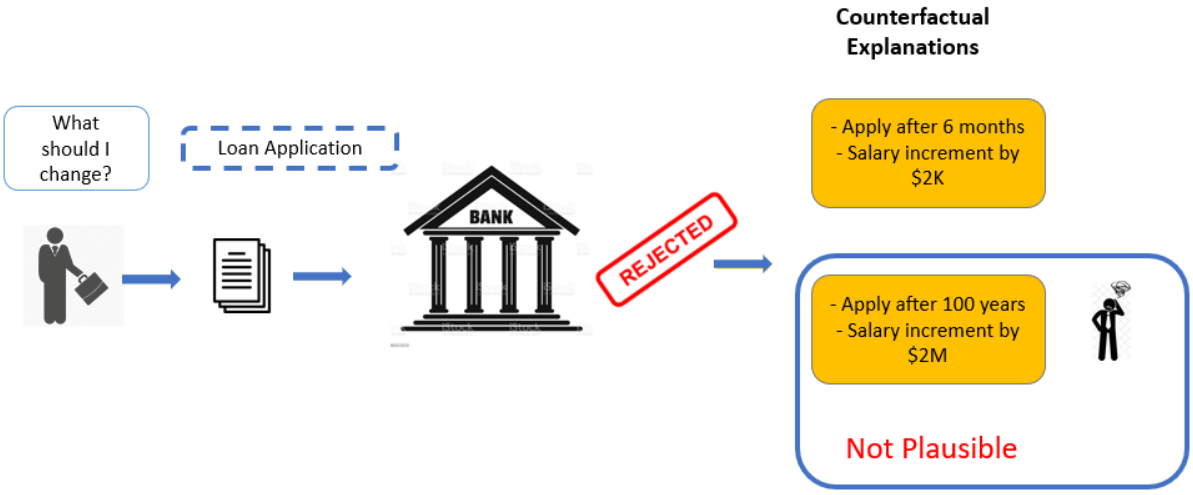


Figure 1.4: Counterfactual Explanations for Credit Scoring application

Consider the application of deep learning for image classification as seen in Fig 1.5, the medical practitioner is provided with a deep learning model, that is trained by a computer scientist on MRI images of the brain. The model has learnt to classify images into "tumour detected" and "without tumour", upon evaluating the model, it is sent as an application to the medical practitioner. Now the medical practitioner is unaware of the underlying techniques used in training the model, this might raise suspicion about the usage of it. The medical practitioner has been trained for years to attain this capacity of classifying MRI scans, so at such times it is important to provide explanations to the user on the decision making science of the model. The global techniques in Explainable AI might produce explanations around what regions are mostly classified as tumour. But the medical practitioner here is looking for an explanation based on the prediction of the model on the current instance of interest. In such cases counterfactual methods produce local explanations that can be consumed by the user. As shown in Fig 1.5, the medical practitioner can query the Explainable AI system regarding the current instance and

ask what should be changed in the given image to alter its prediction. The counterfactual method will generate results based on the user's input, if these results are in line with the user's evaluation of the images, it helps in building trust of the user into the model. Secondly, if the user does not agree with the model's decision these instances can be marked and further used for retraining the model. In such cases if the counterfactual images generated does not really look like brain scans, the medical practitioner would strait away withdraw their interest from these classification model. This acts as a motivation to explore methods of generating plausible counterfactual images so that it assists the user in gaining trust in the system, and in case there is disagreement these plausible counterfactual images can be further used to retrain the model and boost its performance.

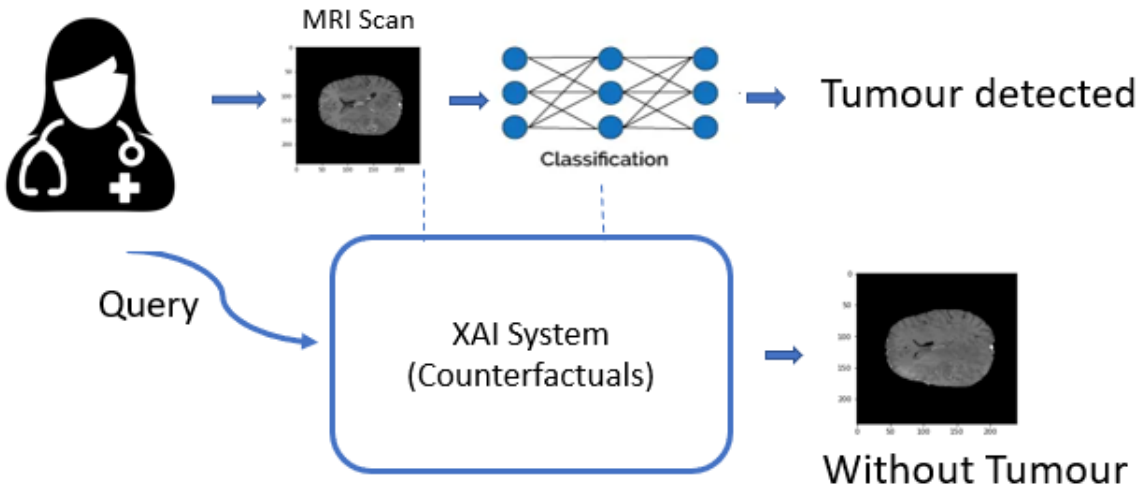


Figure 1.5: Counterfactual Explanations for tumour classification

Further considering the similar sensitive applications like face detection, it is important to evaluate the performance of these models on counterfactual images to understand the bias in the system and make them fair. Even in these applications, it is crucial to generate plausible images so that the fairness of the model can be evaluated on real images rather than unreal instances. Looking at all these applications, the motivation is to explore methods for plausible counterfactual image generation which can be useful to attain trust in the image classification model, provide the feedback to the model in case the user disagrees with the results and also help in fairness evaluation of image classifiers. Therefore we explore the usage of discriminator trained via Generative Adversarial Network (GAN) as a plausibility regulator which would ensure the generated images are closer to the training data distribution. Further,we intend to explore visualisation techniques that can provide us with transformation maps from the original image to counterfactual image, so that it is easier to provide explanations to the user.

1.4 Goals

We will answer the following research questions:

Research Question-1 [**RQ-1**] What is the effect of using a Discriminator trained with Generative Adversarial Network (GAN) as a plausibility regulator in Counterfactual Image generation?

The generated image set using plausibility regulator will be compared with the results of existing methods to explore if the proposed method improves the plausibility(closeness to real image) of the images.

Research Question-2 [**RQ-2**] What are the effective methods of visualisation in attaining the transformations from the given image to the counterfactual image?

The acquired plausible counterfactual images will be used along with the original image to derive the image registration, this will help us to get the transformation map that transforms the original image into counterfactual image.

1.5 Main Contribution

In answering the research questions formulated in 1.4 this thesis contributes the following:

- Provide a novel Generative Adversarial Network architecture to generate high quality images on the MRI dataset in comparison to Deep Convolutional GAN (DCGAN) and Wasserstein (WGAN)
- Provide analysis on the usage of plausibility regulator in counterfactual image generation to attain plausible images in comparison to baseline methods in simple and complex datasets

1.6 Thesis Outline

The rest of this thesis is structured as follows:

- In Chapter 2, we visit some related work on convolutional neural network, Generative Adversarial Network (GAN) and counterfactual image generation techniques
- Chapter 3 deals with fundamental and background concepts, further the baseline are discussed in this chapter
- Chapter 4 introduce the proposed method for generating plausible counterfactual images
- Chapter 5 discusses the datasets used and the experimentation and evaluation of the techniques over these datasets
- Chapter 6 is dedicated in reviewing and discussing evaluation results obtained and analysing the overall utility of the proposed method

- Finally, in Chapter 7, we conclude this thesis with the study results, highlight few limitations and discuss future improvements.

2 Related Work

2.1 Convolutional Neural Network for Classification

Advanced machine learning models and the deep learning based models have provided ground breaking results in terms of classification,segmentation and fraud detection tasks [3, 4, 5]. These tasks are widely used to build applications in variety of domains like healthcare, criminal justice and banking. These domains are sensitive in nature, which is why they demand in depth knowledge on the decision making process of these models. This has encouraged the computer science experts to look for techniques that can help the users to interpret these advanced models. All these techniques are combined and placed under the research area termed as "Explainable AI" [6] . XAI has introduced "post-hoc methods" [7] for interpreting the models, these methods operate on the already trained model and attempts to explain it without tinkering with the internal working of the model. Christopher Molnar in his report [7] categorises these techniques into global and local methods. The global method tend to answer "What features are important in making a decision over majority of data", while the local methods answer "What features are important in decision making on a single instance".

In [8], Marco et all introduce the LIME technique which is used for providing local explanations. This technique uses a sample instance as reference and then considers samples from its neighbourhood to predict the outcome. The prediction on the neighbouring samples is weighted as per the distance from the original sample. This prediction is modelled using a linear model to understand the feature importance of the classifier in the vicinity of the original instance. Further this technique is utilised in image classifiers to highlight the pixels that govern the decision of the classifier. In [9], the authors propose using SHAP based kernels for providing local explanations in cases where the dimensionality is higher eg: text or image classification. SHAP values use the game theory concept of rewarding the features that are primal in determining the outcome.

In Explainable Artificial Intelligence Approaches: A Survey [6], the authors mention the recently introduced post-hoc methods for XAI like - Partial Dependence plots (PDP), Individual Conditional Expectation(ICE), Accumulated Local Effects (ALE). Further the survey mentions "Counterfactual Explanations" which are similar to example based explanations that are part of local explanation.

2.2 Conterfactual Explanation

Judea Pearl in [10] introduced the concept of Structural Causal Model(SCM), where the author defines "counterfactual" as "minimum deviation from reality", these deviations helped them to calculate the probabilities of an outcome based on an action. But for the

scope of this thesis, the definition of "counterfactual" is inline with Shusen Liu et al [11] who define it as minimum change to the input instance to alter the outcome. As it can be understood here that in SCM the focus lies on understanding the outcome of an action, on the contrary our focus is on defining the outcome and then finding an action that helps us achieve it. This definition can be considered inline with "contrastive explanations" [1]. In a machine learning setting, constrastive explanations tend to explain why a particular event occurred as opposed to another. This reasoning is intuitive and human friendly, which is why it is deemed as important by cognitive science. The other case that didn't happen is termed as a "counterfactual" in literal sense, this term is short form of of the phrase "countering the factual".

Explanations that are based on the counterfactual are widely used to explain the prediction of classification models [11, 12, 13]. These explanations can point out the feature changes that could alter the prediction of the model. In a sensitive application like banking, if a advanced machine learning model rejects the loan for a specific applicant. A counterfactual explanation would narrate the applicant on what changes are needed to the applicant's profile to get the loan approved. These explanations provide human-friendly reasoning thereby making the decision making a transparent process. These explanations are also used to evaluate the fairness of the model in such sensitive situations. As Counterfactual Explanations assist in reasoning the decision making process they have earned the stature of being GDPR friendly [14].

2.3 Conterfactual Explanation in different data domains

The definition of Counterfactual remain same in ever domain of data, but the techniques used to calculate the counterfactual instances varies. For tabular data in [2], the authors provide a way to generate faithful counterfactual instances, along with that they introduce the degree of difficulty required to generate a specific counterfactual.

In Interpretable ML book [7], the author suggest that the naive way of searching for a counterfactual is the trial and error method where we vary the features until the desired class is attained by the classifier. Further in [15] Wachter et al, defined it in a mathematical form as a minimising loss that is two fold, firstly reduce the distance between the original instance and the desired instance, secondly reduce the distance between the desired class and predicted class.

In [16], the authors answer the question for a generic text classifier "What changes can be done to the sensitive attribute to alter prediction". Finding answer to these question allows them to evaluate the fairness of the model with respect to sensitive attributes. This is how counterfactuals are also used for evaluating model fairness. In Generate Your [17], the authors use the GPT-2 to generate text samples to test the ML systems. These text are plausible and goal oriented. These text samples can be used to evaluate the text debasing algorithm.

In the image domain, an image can be termed as a counterfactual image, if it is closer to the reference one and has alternate prediction. But the closeness of images can be calculated using various distance measures. Also, there are numerous methods used to identify counterfactual instances. In [13], Yash et all introduce the idea of using distractor image that has an alternate prediction. This image is compared with the original image to identify

the regions in the image that needs to be changed to attain alternate prediction. In other words, the authors provide a specific region in the image that is most relevant in altering the prediction which they term it as "Counterfactual Visual Explanations". Further they use this technique for human teaching, where they explain how an image can be transformed from one class to another using visual explanations.

In [12], the authors leverage the SEDC technique to explain image classifiers. This technique segments the given image in smaller segments using image segmentation algorithm, further it identifies upon removal of which of these irreducible segment or set of segment the prediction of the image is altered. These image segments are produced as visual explanations.

Chun-Hao Chang, in [18] takes one step further in explaining the image classifiers by means of saliency maps. They remove a random patch in the image and in fill it with the generative networks and then identify the significance of that patch in determining the prediction of the given image. They answer the question "What part of the image if not seen by the classifier affects its decision significantly?". The earlier techniques [19] utilised random blurring or adhoc noise to in-fill these patches, thereby creating images that do not lie in data distribution. Generating in-fill images using generative networks helps the classifier to predict on images that are closer to data distribution and this results in attaining useful information about the image regions that are significant for the given prediction.

In "Open Set Learning with Counterfactual Images" [20] by Lawrence Neal, the counterfactual image generation using Generative Adversarial Networks was leveraged to generate samples that are close to data distribution. The aim of the authors was to improve the performance of open set recognition. The classifier was supplied with these generative images along with the real data so as to train it to identify images that do not belong to any of the classes even after close resemblance. Using Counterfactual image generation for open set recognition they outperformed the other traditional methods used for open set recognition.

In [11], Liu et all leveraged the Generative Adversarial Networks to generate image that is close to the given image yet has different prediction. This image was used to identify the minimum change required in the original image to alter its prediction. They termed it as Counterfactual Introspection because this techniques allows us to interpret the decision making of the deep learning model upto certain extent. They also introduce the concept of prototype and criticism, where prototype means the image that is closer to the original one but the prediction of the class is stronger, while criticism means an image that is closer to the original image but has alternate prediction.

2.4 Plausibility of generated counterfactual images

All these methods have provided the research community a set of new lenses for identifying counterfactual images as well as evaluating the models fairness. But these techniques do not consider the plausibility of the generated images. Imagine in the bank loan classification setting, the counterfactual explanation is "Increase age by 100" and "Increase

income by two million dollars", the user of the application will question the decision making of the model. To avoid these problems, it is important to consider plausibility of the identified counterfactual instance.

To attain plausible counterfactual in tabular datasets, one can specify rules and constraints over the feature values while generation of counterfactual class.

Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification, the authors leverage the transformer to generate human like sentences to generate plausible counterfactual explanation for the prediction. They utilise these techniques to explain the text classifier in fin-tech domain.

In "Auditing Deep Learning Classifiers with Realistic Adversarial Examples"[21], the author uses the Counterfactual techniques to audit the model for non experts. As a first objective, Generative Adversarial Network is used to generate images and then certain attributes of the images are varied to generate image closer to the original one. Second objective is to make sure these images belong to the alternate class. Lastly he discriminator is used to determine the plausibility of the generated images. All these objectives are combined to pose this as a multi objective optimisation problem that can be solved using evolutionary algorithms. The success of their work lies in auditing the model on plausible samples that a human could supply to the model rather than assessing the model on non real samples.

In [1], Kenny et all, introduced the method called "PlausIble Exceptionality-based Contrastive Explanations (PIECE),". The authors use the penultimate layer of the classifier to learn the logit distribution of various classes. Upon receiving a query image, the logits are calculated and compared with the distribution to identify the exceptional features, these features are modified to reach the counterfactual class. These modified logits are fed to the GAN to generate the counterfactual image. Further they utilise this technique to identify the semi- factuals as well. Unlike other methods, the counterfactual class need not be specified. This is the state of the art method for generating counterfactual images.

3 Fundamentals and Concepts

3.1 Convolutional Neural Network for Classification

Convolutional Neural Network(CNN) belong to the family of Deep Neural Networks. They have produced groundbreaking results in image classification tasks. Convolution is a mathematical operation where a new function is derived by multiplying two functions. In a deep learning setting, a layer is called as convolutional layer if it performs convolution of image matrix with weight matrices(kernels) to output feature maps. A deep learning architecture having such layers is termed as CNN. CNNs can have numerous convolutional layers. The layers in the beginning are responsible for learning low level features in the image eg edges, shapes and boundaries while the successive convolutional layers learn rather complex features like texture and unique identifiers of the image. The spatial dimension of the image is reduced by pooling operation so that only the important information is retained while performing successive convolution operation. Optionally, dense layers are stacked after the last convolutional layer for mapping the extracted features to the image classes. These dense layers output the logits for each class. The loss is calculated by comparing the logits against the ground-truth. This loss is then back-propagated to train the network weights.

ResNet, Inception Net, Dense Net are the popular CNN architectures that are used for image classification tasks.

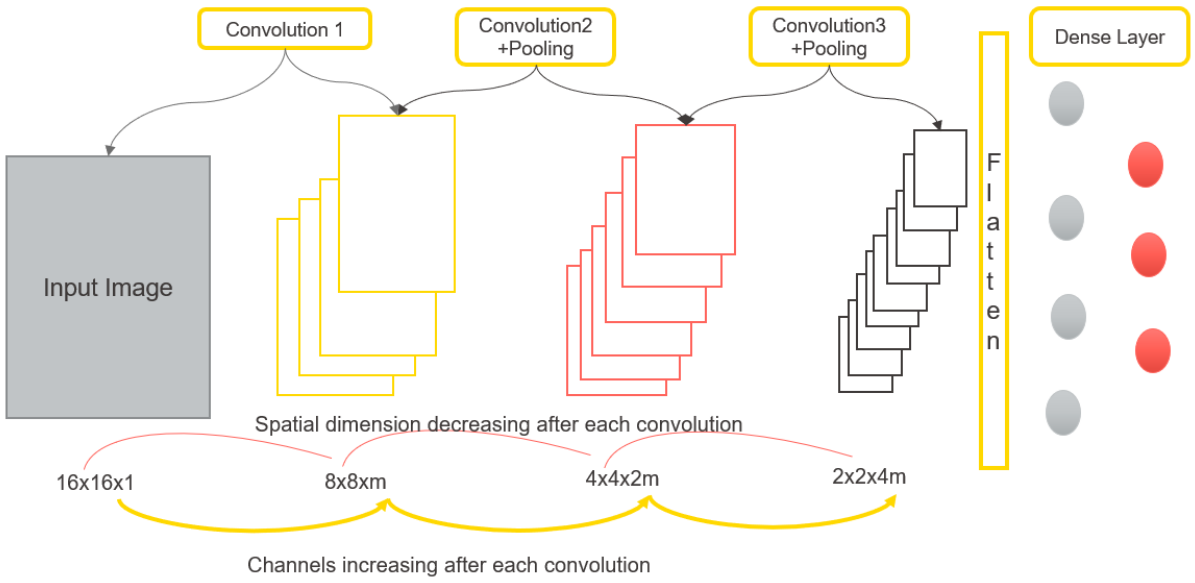


Figure 3.1: Block diagram of general CNN)

3.2 Generative Adversarial Network (GAN)

Generative methods are classified as unsupervised methods because they learn the inherent distribution of the provided samples. These methods can generate new samples that resemble the original ones. Generative Adversarial Network (GAN) is one such generative method that was introduced by [22]. These Networks are comprised of two competing models Generator G and Discriminator D. The Generator is responsible for generating images that are closer to the images from original distribution, while the Discriminator is responsible for identifying whether the sample provided to it came from the generator or the original distribution. In other words, the task of the Generator is to fool the Discriminator, while the task of the Discriminator is to precisely distinguish the images. The competition provides two fold impetus, firstly it encourages Generator to produce images that are close to reality and on the other hand it allows the Discriminator to become stronger in identifying the real samples by focusing on complex features that are traits of the real image. The block diagram of a generic GAN is shown in Fig.4.3.

In mathematical terms, a noise vector z is sampled from the Gaussian distribution $p_z(z)$ and provided as input to the Generator G that learns to map z to the image space $G(z,\theta_g)$. G is differential function and learns the hyper-parameters θ_g during training. The Discriminator D acts as a classifier $D(x,\theta_d)$ that receives the sample x and scores it as $D(x)$, this score denotes the probability that the sample belongs to real distribution. Now both these functions play the mini-max game with the value function V(G,D) as shown in the below equation.

$$\min_{\mathbf{G}} \max_{\mathbf{D}} V(G,D) = E_{x\sim P_{data}(x)}[logD(x)] + E_{z\sim P_z(z)}[log(1-D(G(z))] \tag{3.1}$$

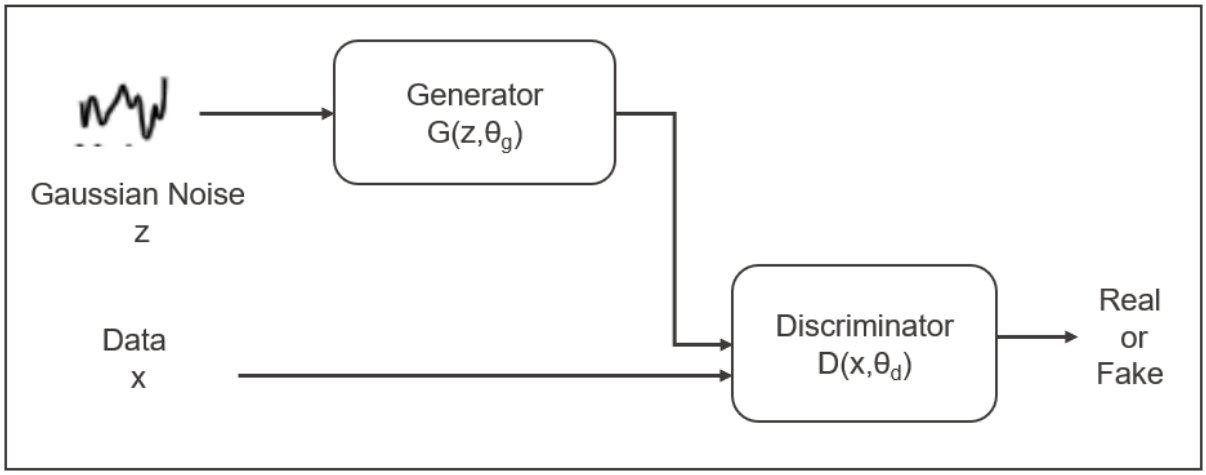


Figure 3.2: Block diagram of Generative Adversarial Nets (GANs)

3.2.1 Deep Convolutional GAN (DC-GAN)

As the name suggests, the Generator and the Discriminator in this network consist of convolutional layers. Layers in the Generator perform transposed convolution and produce

images from a given Gaussian noise vector. Layers in the Discriminator use convolutional blocks to construct a classifier that identifies whether images provided to it are real or fake, the convolutional blocks give the discriminator more power to extract complex features that separate real images from fake images. As a result, DC-GAN produces more realistic images than basic GANs.

The architecture of the DC-GAN is shown in Figure

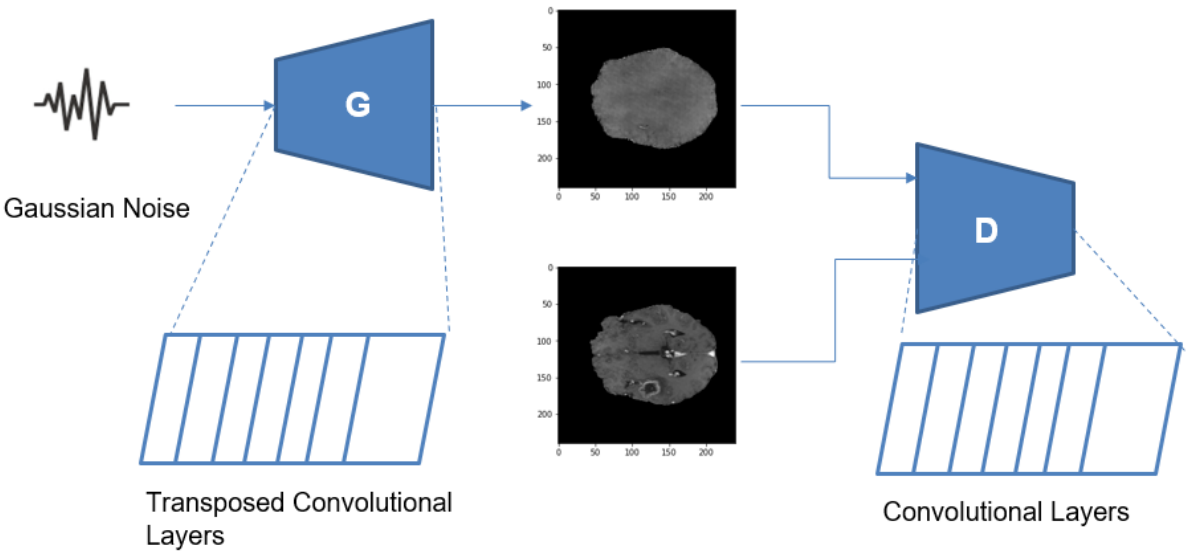


Figure 3.3: Block diagram of Deep Convolutional Generative Adversarial Network (DC-GANs)

3.2.2 Wasserstein GAN (WGAN) and WGAN with Gradient Penalty (WGAN-GP)

It takes a lot of work to train GAN. The model might not converge and modes collapse frequently. As we move forward, incremental improvements are possible or we can pursue a new route to reduce costs. Let us first try to understand the problems in detail. GANs use Jensen Shanon Divergence to calculate the distance between the distribution of real images and the generated images. This learning process is fruitful if the discriminator network is optimal, if the discriminator is bad it does not provide information for the generator network to generate better images, while if the discriminator is perfect it provides almost zero loss and the generator network faces the problem of vanishing gradients. The second problem is mode collapse which means that the generator is generating same type of images and is able to trick the discriminator.

There are various incremental methods performed to improve the training of GANs. One such method is using Wasserstein distance for calculating distance between both distributions. Wasserstein Distance is a measure of the distance separating two probability distributions. It is also known as Earth Mover's distance, or EM distance, because it can be interpreted as the minimum energy cost of moving and transforming a pile of dirt from

one probability distribution to another. Compared to JS divergence, Wasserstein distance can still be meaningful and smooth even when two distributions are located in manifolds of lower dimension without overlapping.

The Wasserstein GAN (WGAN) proposes using Wasserstein distance in the loss calculation and the "discriminator" isn't classifying how to tell fake samples from real ones anymore. Instead, it learns to compute Wasserstein distance using a K-Lipschitz continuous function, this is why it is termed as "critic". When the loss function decreases during training, the Wasserstein distance gets smaller, and the generator model's output gets closer to the real data distribution. To maintain the continuity of the K-Lipschitz function, the WGAN authors offer 2 variants, first they suggest on clipping the weights of the critic after every gradient update and they term this as WGAN and secondly they suggest to clip the gradients after every gradient update and they term it as WGAN gradient penalty (WGAN-GP).

Compared to the original GAN algorithm, the WGAN makes these changes:

- Clamp the critic weights to fixed range after every gradient update
- The loss function will now be derived from the Wasserstein distance, no longer based on the logarithm. "Discriminator" model does not use direct criticism, but assists in estimating Wasserstein metrics for real versus generated data distributions.
- Recommend using RMSProp optimiser on the critic, rather than a momentum based optimiser such as Adam

3.3 GAN based methods for Counterfactual Images

In this section, we look at the various methods that are used for Counterfactual Image generation using GANs. Let us look at the problem definition.

We have an image classifier(C,S) based on CNN, that is trained on a specific train set and it predicts the class of the image I. The architecture of the classifier is shown in Figure 3.4. This classifier could be a binary or multi-class classifier. The first part C of the classifier consists of CNN that extracts features from the image and then these features are flattened to get layer X. Along with X we have a linear network that connects to the Soft max output layer which outputs the probability vector Y which is used to predict the class c of the image. Yc is the vector corresponding to largest probability of class c (predicted class is c). The linear network along with X is termed as S. Now as shown in Figure 3.5,the classifier predicts the class of the image I to be c. The question posed to the proposed system is to identify a plausible counterfactual image I' corresponding to latent input z' that has the classifier prediction of class c'. The proposed system uses GAN that has a Generator G and Discriminator D trained on same train set that was used for training classifier. The latent input (noise vector) to the G is z.

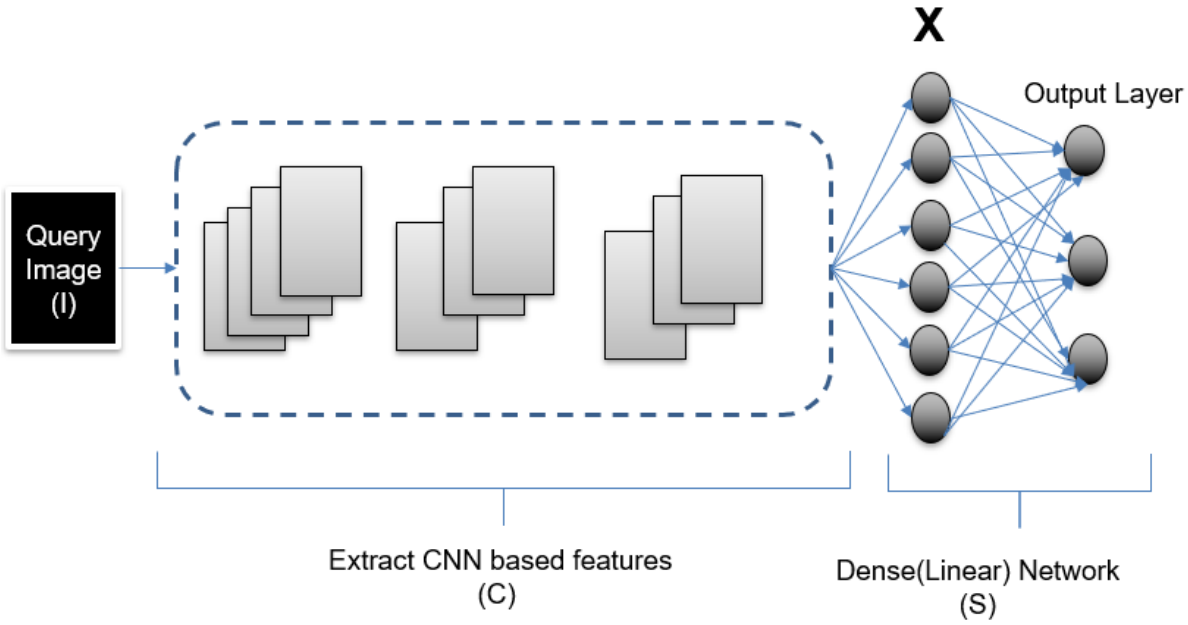


Figure 3.4: CNN in our use case

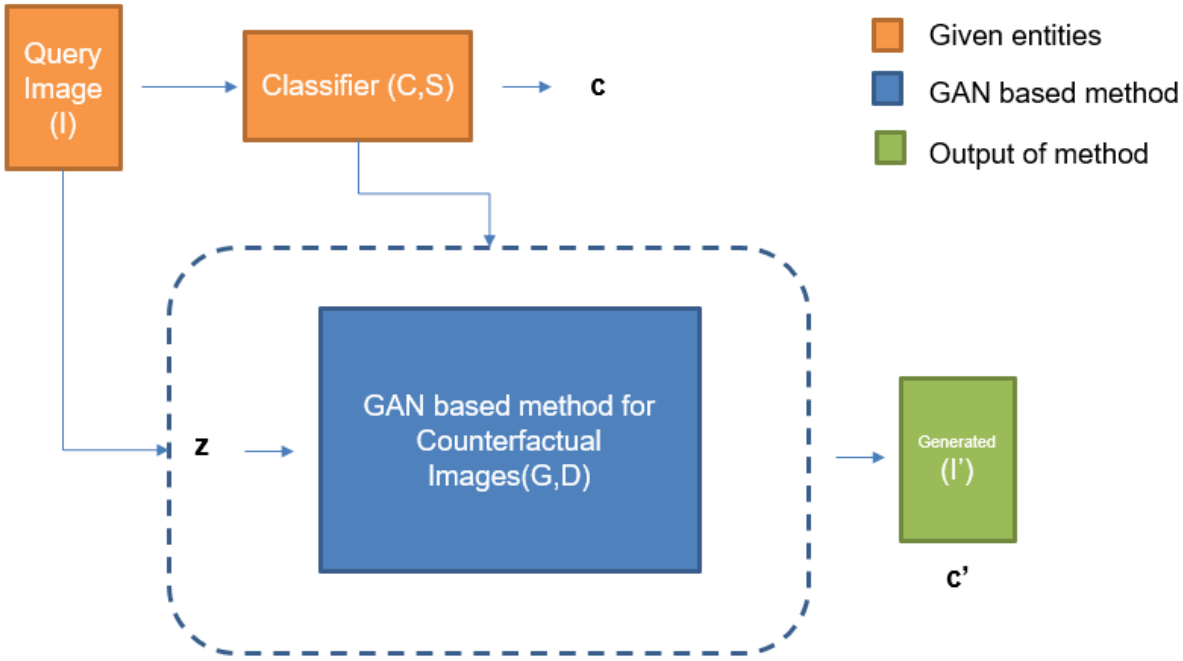


Figure 3.5: Problem definition

We intend to find the z_{org} that corresponds to the given query image I. This z_{org} will prove to be the starting point during the optimization process when we are searching for z'. We use the below given equation to find z_{org}. The term d(G(z), I) measure the distance between the query image and the generated image. Pixel loss, L1 loss or perceptual losses are used as distance function to calculate this distance.

$$z_{org} = \underset{\mathbf{z}}{argmin}\, d(G(z), I) \tag{3.2}$$

3.3.1 Baseline 1: Constrained Minimum Edit Method (C-Min Edit)

This is a modified method derived from [11] paper. Here, our objective is to find an image that is closer to the given image and has the classifier prediction as c'. This is why we have the loss function with two terms, the first term ensures the prediction on generated image is closer to c' while the second term constrains the distance between the images in the latent space to be as small as possible. The λ is annealed from a lower value to higher value to find the optimum z. The optimisation process is immediately stopped when the generated image G(z) attains prediction as c'. As it is evident from the below equation, this method do not consider plausibility of the generated image. This is one of the pioneer technique in finding counterfactual images using GANs, which is why I have chosen this technique to be a baseline-1 method for my analysis.

$$z' = \underset{\mathbf{z}}{argmin}\, \underset{\lambda}{max}\, \lambda||S(C(G(z)) - Y_{c'}||_2^2 + d(z_{org}, z) \tag{3.3}$$

3.3.2 Baseline 2: PlausIble Exceptionality-based Contrastive Explanations (PIECE)

As the name suggests, PIECE tries to modify the exceptional features of the given image and tries to change these feature to their normal value to find a counterfactual image. PIECE exploits the distributional properties of the data to help guarantee plausibility. The method can be broken down as a three step process. (i) Identify the exceptional features (ii) Modify these features (iii) Visualize the resulting image using GAN.

In multiiclass classification scenario there are many options for choosing the he counterfactual class, but PIECE automates this process by using the gradient ascent method to reach the closest class. The equation for finding the counterfactual class is shown in Eq 3.3.2 where the optimisation stops once the closest class is achieved and that class is termed as the counterfactual class.

$$counterfactualclass = \underset{\mathbf{z}}{argmax}\, ||S(C(G(z))) - Y_c||_2^2$$

(3.4)

Suppose that the Query Image I is passed through C and the resulting representation is termed as x. Let us understand the three steps in detail

Identify exceptional features

Here, we need to identify the exceptional features in x, so that we can modify it to x′. We examine the statistical properties of x in training distributions of c′. We pass all the data instances through C and record their representation at X layer. Suppose these

representations consists of n neurons, we learn one distribution for each X_i and for each class in the dataset. So after receiving the vector x corresponding to image I, we compare the values in x with the statistical distributions using the hurdle model with Bernoulli distribution to understand the features that would highly shift the prediction from class c to c'.

Modify exceptional features

The exceptional features identified from the previous steps are classified into positively and negatively affecting. The negatively affecting features are chosen and checked against various conditions and modified only if they bring x' closer to the counterfactual prediction c'. So at the output of this step we get the vector x'.

Visualise the images with GAN

At the last step, we have identified the exceptional features and also modified those that satisfied the criteria. Now we need to visualize the image that corresponds to x' representation after being passed through C. We use the following equation to generate the image using Generator G.

$$z' = \underset{\mathbf{z}}{argmin}\, \underset{\lambda}{max}\, \lambda ||C(G(z) - z'||_2^2 \tag{3.5}$$

The z' obtained from the above equation is fed to the Generator G to achieve $G(z')$ or which is equivalent to I'.

This technique uses distributional properties of the data to introduce plausibility when finding counterfactual images. This is the state-of-the-art method used to generate counterfactual images using GANs, which is why I am using this method as a baseline-2 for my analysis.

3.4 Evaluation Metrics

It is a tricky question to assess the plausibility of generated counterfactual images because we want to assess how likely it is that the given image belongs to the training distribution. When evaluating qualitatively, images can be looked at and compared with images in training data. This can be easier for simpler datasets, however with complex datasets, it can be difficult. In qualitative evaluation, most of the research papers [cited] have computed the reconstruction errors using autoencoder-based metrics and then used the results to demonstrate how close a particular image is to the training distribution. Others have used the Monte Carlo Dropout based metrics to detect images that are out of distribution

3.4.1 IMI1

In [23] the author trains autoencoders for each class in the training data. Now to evaluate using this metric, the autoencoder trained on class c (AE_c)is used to calculate the reconstruction error for that class, while another autoencoder trained on counterfactual class c' ($AE_{c'}$) will be used to calculate the reconstruction error for the counterfactual class. The ratio of reconstruction error for a counterfactual class to reconstruction error for given class is termed as IMI1 and this metric should be lower in value to so that generated counterfactual are plausible. The formula for calculating the IMI1 is shown in equation eq 3.6. The I' represents the generated counterfactual image.

$$IMI1 = \frac{||I' - AE_{c'}(I')||_2^2}{||I' - AE_c(I')||_2^2} \tag{3.6}$$

3.4.2 MC-Mean and MC-Std

The authors of [23] used the Monte Carlo Dropout method for the first time to detect out-of-distribution images and thereby measuring the plausibility of the generated images, using a thousand forward passes and enabling the dropout layers in the classifier to collect the MC Dropout values. MC-Mean is the posterior mean of the MC Dropout on the generated image. In order for the generated counterfactual images to be plausible, the metric needs to be higher in value. The MC-STD can be defined as the posterior standard deviation of the MC Dropout on the generated images. This metric MC-Mean should be higher in value so that the generated counterfactual images are plausible, while the MC-STD should be lower so that it denotes the lesser fluctuations in the MC-Mean values.

3.4.3 U-Net based perceptual loss

The pixel based losses are ineffective in generating high quality images, which is why we observe the surge in the usage of non-pixel loss functions. Pre trained networks are used to derive features from the images and then these feature are used to calculate the loss value. In U-Net based perceptual loss, we consider a pre trained U-Net on the MRI images for tumour segmentation. This U-Net model will act as a pre trained network in our loss calculation. The images are passed to the U-Net and it provides the features at each level of upscale, which means we get four set of features for each image passed through the U-Net. We calculate the overall loss by summing up the loss at each set of the features. The eq. 3.7, shows the mathematical equation to calculate the loss value. I and I' represent the original image and reference image respectively. $UNet_{ln}$ represents the features extracted at the n^{th} level of the pre trained U-Net architecture.

$$PerceptualLoss = ||UNet_{l1}(I) - UNet_{l1}(I')||_2^2 + ... + ||UNet_{l4}(I) - UNet_{l4}(I')||_2^2 \tag{3.7}$$

3.4.4 Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) score was termed by Heusel et al. in [24], it is used for measuring the quality of the images generated by the GAN [25, 26]. The FID score is based on using statistics from the real images and the generated images and then calculating the score. Lower the value of FID points to the high level of quality of the images. The Inception v3 model by Szegedy et al. [27] is used to compute the embeddings of the real images and the generated images. These embeddings are derived from the penultimate layer of the per trained model and the dimension of these embeddings is 2048. The images from both synthetic and real datasets are passed through the model, and the obtained statistics are used to calculate the FID as in Eq. 3.4.4

$$d^2\left((\boldsymbol{\mu}_r, \boldsymbol{C}_r), (\boldsymbol{\mu}_s, \boldsymbol{C}_s)\right) = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_s\|^2 + \mathrm{Tr}\left(\boldsymbol{C}_r + \boldsymbol{C}_s - 2\left(\boldsymbol{C}_r C_s\right)^{1/2}\right) \tag{3.8}$$

In the above **??**, μ_r and μ_s are feature-wise mean for real and synthetic images, C_r and C_s are respective covariance matrices. Tr is the Trace linear algebra operation, which is the sum of main diagonal elements.

4 Method

4.1 Proposed GAN architectitures for high quality images

4.1.1 U-Net and Wasserstein GAN (U-Net WGAN)

U-Net is an extensively used convolutional neural network architecture used for precise image segmentation tasks, it has outperformed several other convolutional models in terms of its performance in biomedical image segmentation. The U-Net can be viewed as a combination of encoder and decoder architecture. The encoder architecture performs convolution and pooling at every stage to contract the image to obtain the most important features, the decoder part uses these features along with the skip connections from the encoding stages to upscale the image using transposed convolution operation. The upscaled image is then compared with the true segmentation mask to calculate the loss and then it is backpropagated to train the U-Net architecture.

In this U-Net WGAN, we propose to replace the generator network with the U-Net architecture to understand if the image quality of the generated images can be improved. The discriminator is the classical deep convolutional network and then we calculate the loss using the Wasserstein distance. The input to this GAN is a Gaussian noise vector with the same dimension as that of the original image. The architecture of the proposed U-Net WGAN is shown in Fig.

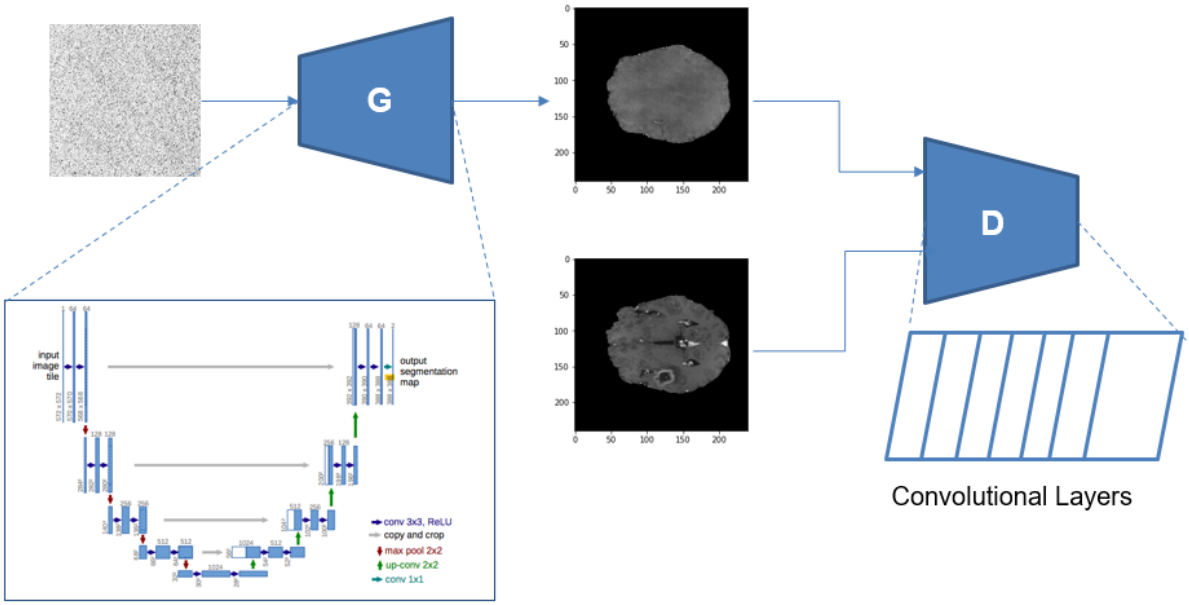


Figure 4.1: Block diagram of U-Net and Wasserstein GAN (U-Net WGAN)

Also, to try the gradient penalty variant of WGAN, we alter the training of this network by clipping the gradients of after every iteration of gradient descent and term it as U-Net and Wasserstein GAN Gradient Penalty (U-Net WGAN-GP)

4.1.2 ReconResNet and Wasserstein GAN (ReconResNet WGAN)

To improve the quality of MR image reconstruction, a deep learning framework called ReconResNet is presented in this paper[28] by Chatterjee et al, this regularised version of ResNet as the network backbone reduces artefacts from the under sampled image, as well as data consistency steps to fusing the network output with data already available from the under sampled k-space to further improve reconstruction quality.

To understand if the quality of the images generated by generator can be improved, we use the ReconResNet as the generator in the GAN architecture, we term it as the ReconResNet Wasserstein GAN (ReconResNet WGAN) as we use the loss function for training it based on the Wasserstein distance. The architecture of this GAN can be seen in Fig.

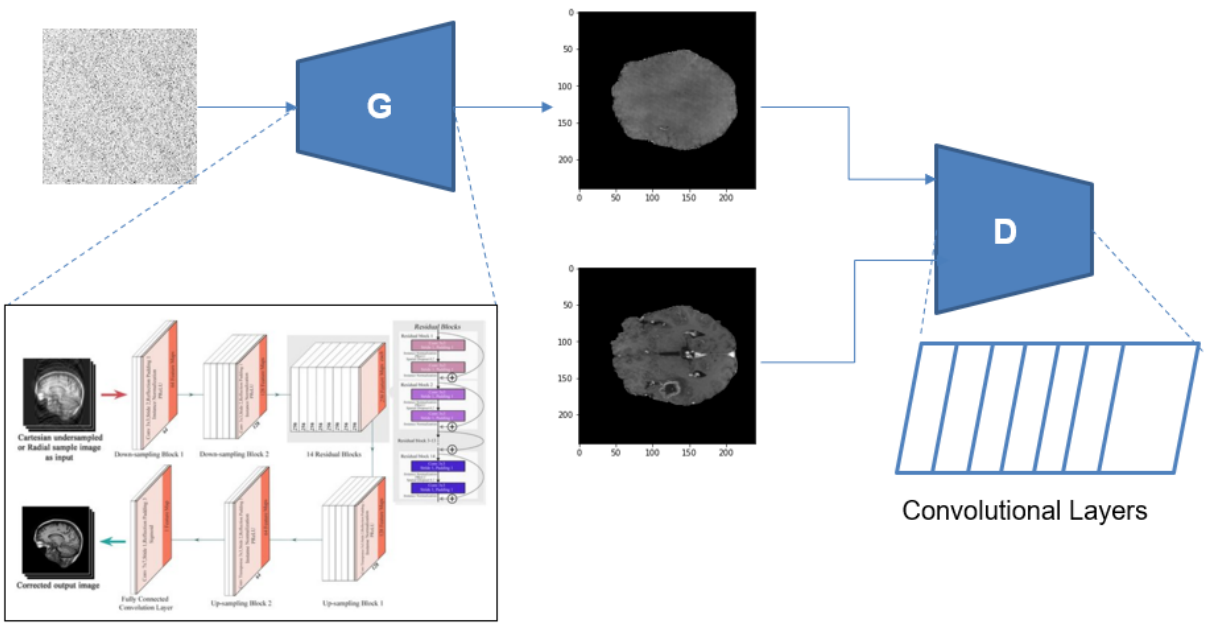


Figure 4.2: Block diagram of ReconResNet Wasserstein GAN(ReconResNet WGAN)

4.1.3 U-Net and Inception Net Wasserstein GAN (U-Net Inception Net WGAN)

In this GAN architecture, we use the U-Net as the Generator network and replace the Discriminator with the Inception Net. The motivation is to make both the networks powerful so that the GAN training can be efficient and this GAN can generate better quality images. We use the loss function based on Wasserstein distance hence we term it as U-Net and Inception Net Wasserstein GAN (U-Net Inception Net WGAN. The architecture can be seen in Fig.

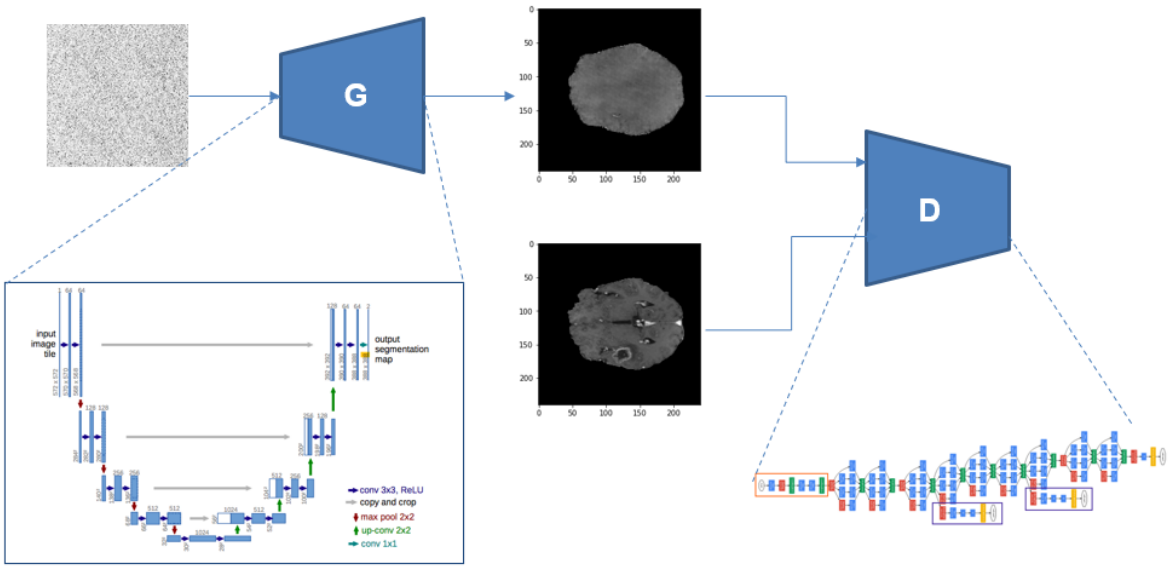


Figure 4.3: Block diagram of U-Net and Inception Net Wasserstein GAN (U-Net Inception Net WGAN)

4.2 Proposed method for Plausible Counterfactual Images

xxx say that we need the info mentioned in fundamentals

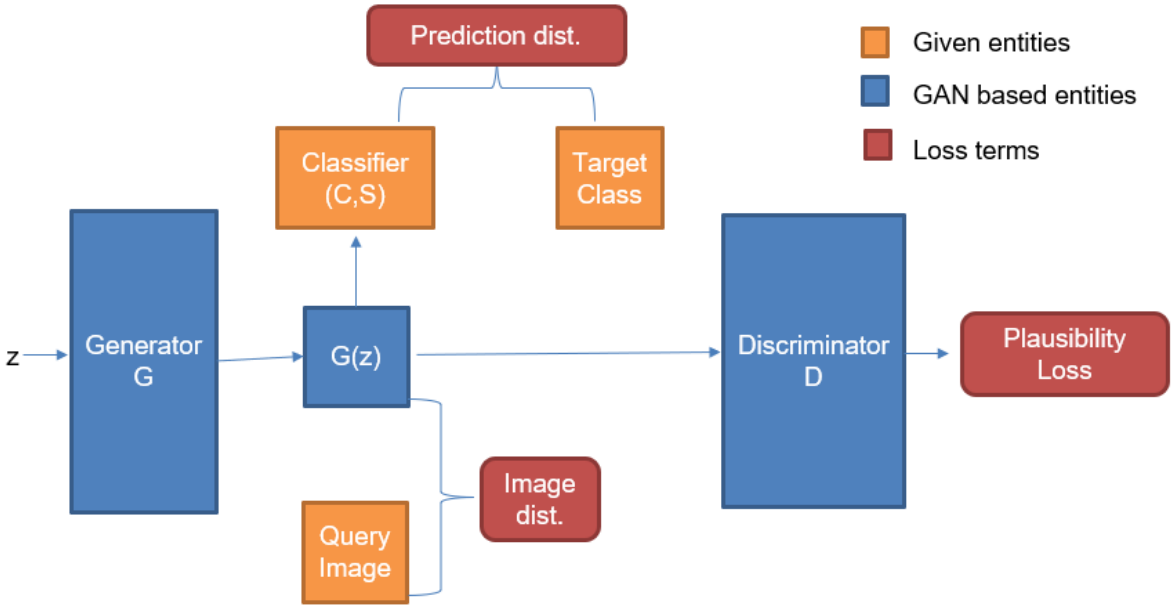


Figure 4.4: Block diagram of Proposed method

In the proposed method , we need the user to specify the counterfactual class c′ and also the plausibility threshold Pl_{thres}. This method uses two steps to attain the counterfactual images.

4.2.1 Step 1: Identify the GAN representation of the original image

The initial step is to identify the latent vector that corresponds to the original image. We term this latent vector as z_{org}, this will be provided as the starting value of latent vector z in the optimisation process for counterfactual generation in step 2.

To compute the z_{org} that that represents an original image I, we define a loss function that has compromises of two terms, first the loss term that defines the reconstruction loss between the original image and the generated image. Secondly the loss term that makes sure that the prediction of the classifier(C,S) on the generated image is close to its prediction on the original image. By optimising the equation 1.5, the z_{org} can be obtained. The lambda value is adjusted to be on higher level so that the priority is given to the first term to generate an image that is close to the original one.

$$z_{org} = \underset{\mathbf{z}}{argmin}\, \underset{\lambda}{max}\, \lambda||G(z)) - I||_2^2 + ||S(C(I) - S(C(G(z)))||_2^2 \tag{4.1}$$

4.2.2 Step 2: Counterfactual Image generation

The focus of the proposed method is to generate plausible counterfactual images, which is why during the optimisation process we also include a plausibility based loss term. This term forces the optimisation process to generate highly plausible images. Unlike baseline-2, we do not use statistical distribution techniques to determine plausibility, we use the Discriminator D as a plausibility scorer, because it can provide a score in between 0-1 whether the given image is real or fake. All the other loss terms are as it is in baseline-1. As shown in Figure 4.4, the orange colour boxes are the entities provided to us, while the blue color boxes are the GAN entities that we have trained. The red colour boxes denote the loss terms, the distance between the prediction of the generated image and the counterfactual prediction forms the first loss term, while the distance between the generated image and the query image forms the second loss term and the last term acts as the plausibility regulator. All these terms are combined to form the below given equation.

$$z' = \underset{\mathbf{z}}{argmin}\, \underset{\lambda}{max}\, \lambda||S(C(G(z))) - Y_{c'}||_2^2 + d(z_{org}, z) + |logD(G(z))| \tag{4.2}$$

The optimisation process stops when the generated image G(z) has attained counterfactual prediction c′ and plausibility score $\mathrm{D(G(z))} > \mathrm{Pl}_{thres}$.

5 Experiments and Evaluation

The baseline as well as the proposed method discussed in the fundamentals and the method section are implemented on two different datasets. The motivation to perform evaluation on two data sets is to analyse the performance of the proposed method on range of datasets. This will help us to understand if the proposed method outperforms the baselines in all types of datasets and would further assist us in evaluating the generalisation of the proposed method. Secondly, if the performance of the proposed method differs on various datasets, we can look into what affects the performance of the proposed method and also it would assist us in understanding the scenarios where the proposed method can be used. Owing to this we have chosen two different datasets for experimentation and evaluation. Firstly we use the MNIST dataset which consists of handwritten digits and we classify it as "simple" dataset as it contains images with simple shapes and curves. Secondly, we use the BraTS MRI dataset that contains MRI scans of brains and we classify it as a "complex" dataset because it consists of complex structures that are minutely intertwined .

Another reason to choose the MNIST dataset is, the images in the dataset are well known and the digits can be understood by everyone, and when we perform qualitative evaluation over teh results from this dataset, we can easily understand the performance of the proposed method by merely comparing the generated images against the baseline methods. imagine we choose an image of digit "0" and we want to generate a counterfactual image with target class as "8". We can look at the generated image and be convinced whether the digit looks like "8" or not. Further we use the BraTS dataset that contains brain MRIs with and without tumour. When we use the BraTS dataset for experimentation, it is difficult to comment on the results of the proposed method as we are not the subject matter expert in detecting tumours. Imagine we consider an image with prediction "tumour deteted" and provide the counterfactual class as "tumour not detected", the proposed method would generate an image without tumour, but the decision on these images has to be taken under the supervision of MRI experts. This is why we have chosen two completely different datasets that will help us in through experimentation and evaluation of the baselines and proposed method.

5.1 Experiment 1: MNIST Handwritten Digits Dataset

5.1.1 Dataset

The MNIST dataset by LeCun et al. (2010), is the well known dataset of handwritten digits. It is widely used to train image classifiers [29]. Each image is made up of 28x28 pixels. The pixel value is between 0-255, denoting the intensity of the pixel where 255 denotes full intensity and 0 denotes dark pixels. It contains 60000 images and has almost

equal number of images for each digits "0", "1", "2", "3", "4" , "5 ", "6", "7", "8", "9". The sample images of the MNIST dataset are shown in fig. **??**.

Figure 5.1: Samples of MNIST Handwritten Digits Dataset [**?**]

5.1.2 Experimental Setup

Pixel values in the MNIST dataset are normalized to be between 0 and 1. Additionally, the dataset is split into train and test sets. The train set contains 50,000 images while the test set contains 10,000 images. For training a classifier that can recognise handwritten digits, a Convolutional Neural Network of 4 Convolutional blocks is used. The Classifier is evaluated over the test set and it achieves 98% test accuracy. As per the concepts discussed in section 3.3 and section 4.2 , we will use GAN-based methods to generate counterfactual images. In order to train a GAN, we employ the Deep Convolutional GAN architecture, where the generator uses a set of convolutional layers to generate a

28x28 image from a one-dimensional Gaussian noise while the discriminator uses the set of convolutional layers to build a classifier.

We use the baseline and the proposed method to to generate counterfactual and to visualise them. First we consider the image for which the prediction is digit "0". We specify the target class from 1 to 9 and show the results for the baseline 1: C-Min Edit method in the first column, while results using the proposed method with plausibility threshold as 0.5 in second column and with threshold as 0.8 in third column as shown in Fig 5.2, 5.3 and 5.4.

The results for the baseline method doesn't seem to be digits from the MNIST dataset in most of the cases, while the images generated by proposed method with plausibility threshold=0.8 seem closer to the actual digits from training data. These results point out that in the given case of the image with prediction of digit "0", the proposed method can be used to generate plausible counterfactual images for most of the target classes ranging from 1-9 in the MNIST dataset. Consider the Fig 5.2, it can be said that the images generated by the baseline hardly qualify as digits as assessed by the human observer, The case where the target class is 1, the baseline generates a blur image , while the proposed method generates something that cannot be qualified as digit "1", but it looks like digit "2" that is still a plausible image. The difference in the baseline and proposed method results is clearly visible in the case where the target class is provided as "3".

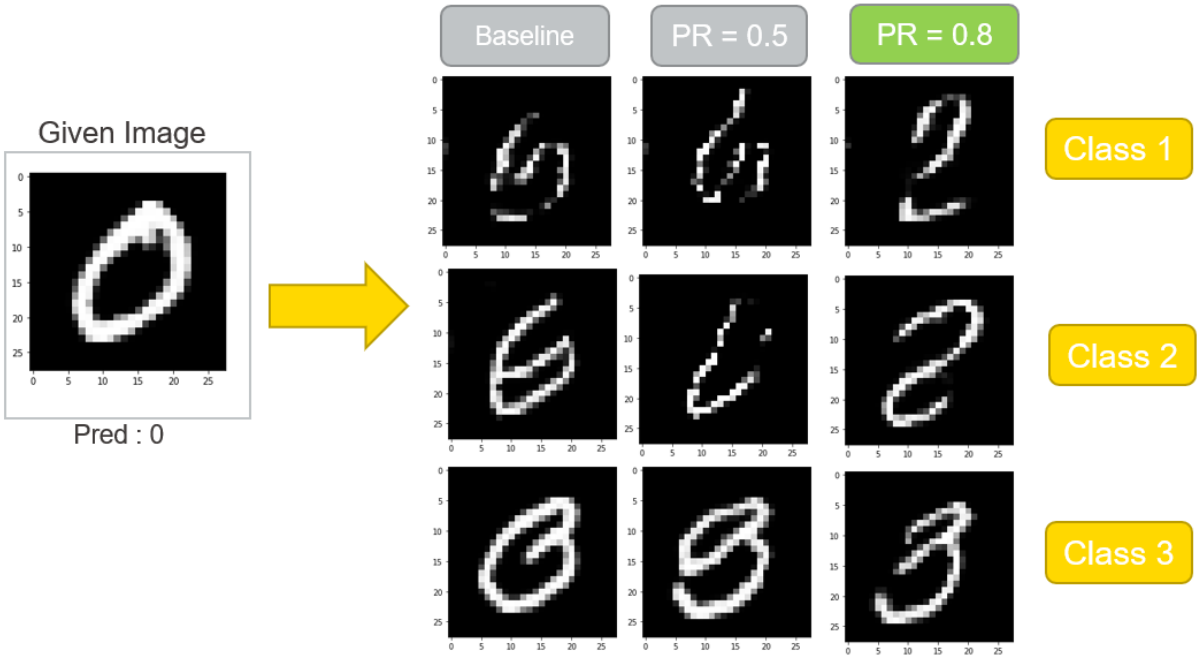


Figure 5.2: Counterfactual Image Generation for target class (1,2,3) given image of digit 0

Consider the Fig 5.3, that shows the counterfactual image generation for the target class between 4-6, for the case with target class as digit "4", the baseline generates an image that cannot be credited as "4" by the human observers, while the proposed method gets the outline correct for it, this example shows that it is indeed difficult to achieve all counterfactual classes with precision using the proposed method, yet the proposed

method outperforms the baseline in most of these cases. In case of counterfactual class of digit "5", the results obtained by the baseline and proposed method seem acceptable.

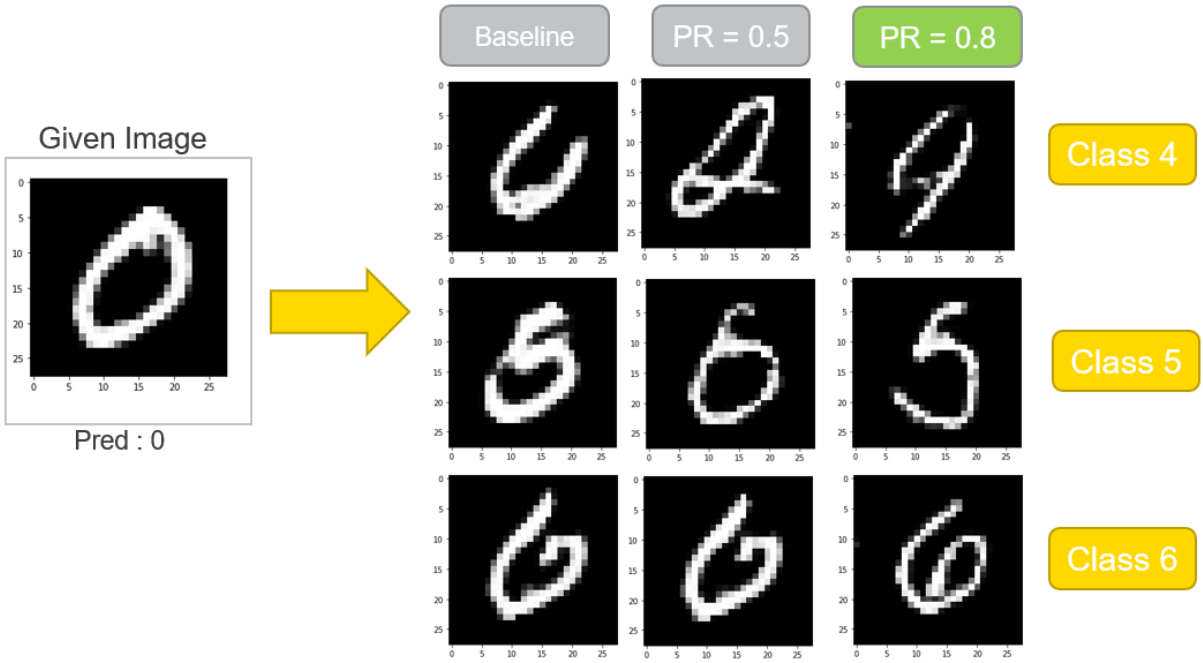


Figure 5.3: Counterfactual Image Generation for target class (4,5,6) given image of digit 0

Consider the Fig 5.4, that shows the counterfactual image generation for the target class between 7-9, in each of these cases the results of proposed method are favourable in comparison to baseline. In case of digit "9", the baseline result cannot be credited as digit "9" by the human evaluator. Overall when we observe the results in Fig 5.2, 5.3, 5.4 it can be asserted that the proposed method outperforms the baseline method in most of cases by human evalaution. Also notice that the images become more plausible as we increase the plausibility threshold.

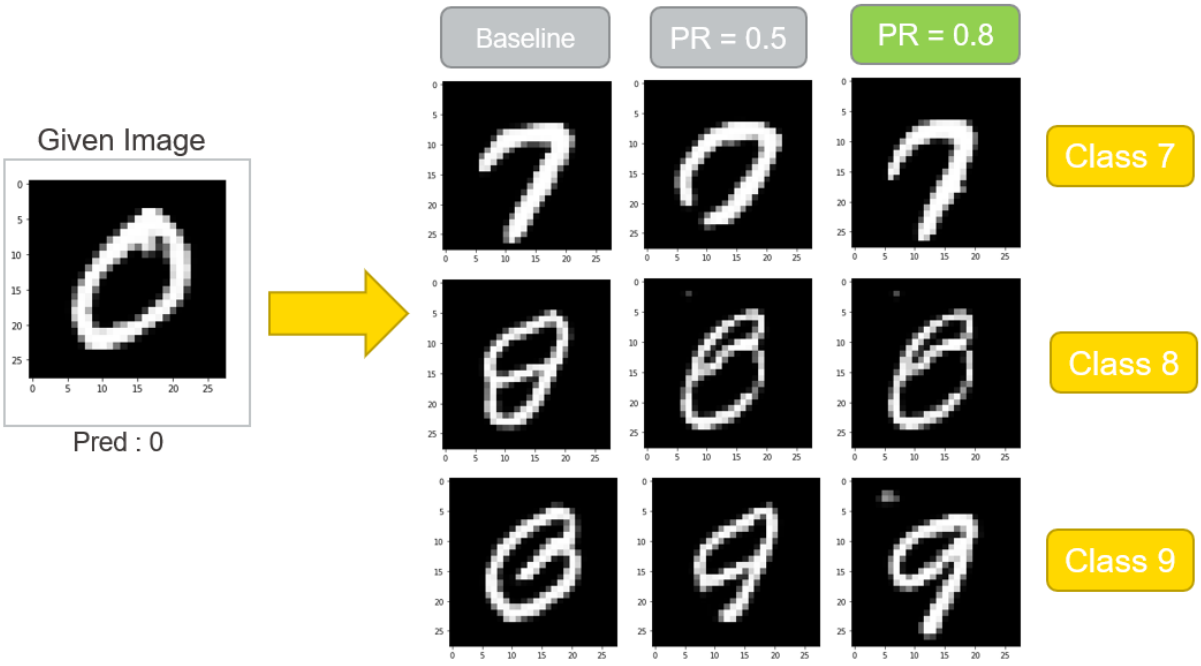


Figure 5.4: Counterfactual Image Generation for target class (7,8,9) given image of digit 0

For the detailed evaluation in terms of quantitative and qualitative, the test set has been divided into two categories - correctly classified and misclassified instances. Using this category-wise evaluation, we will be able to determine how effective the proposed method is in each category. This evaluation will also help us understand the cases that bring out the shortcomings of the proposed method.

5.1.3 Evaluation on correctly classified instances

For the evaluation on this category of data, we have considered six images per class which sums up to 60 images. These images were correctly classified by the trained classifier. In this multiclass classification scenario, there are many options to choose as the counterfactual class. But we use the Eq 3.3.2 to determine the closest class and tag it as the counterfactual class for a particular image. This will help us in assessing the performance of the baselines as well as the proposed method when the original image and counterfactual class are fixed only the techniques used for counterfactual generation are varied.

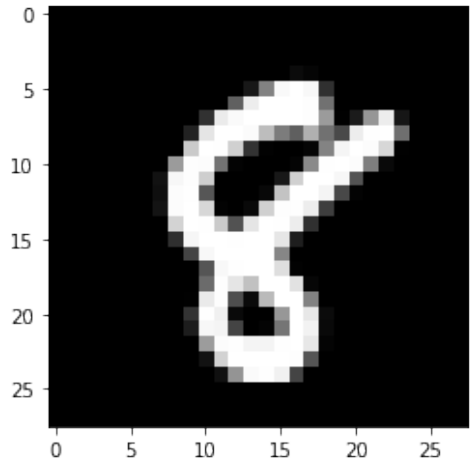
We calculated the values of evaluation metric discussed in section 3.4 for plausibility evaluation on these images to understand if the proposed method is better in comparison to the baselines. These evaluation metrics provides quantitative evaluation of the proposed method.

Table 5.1: Comparison of evaluation metrics for baseline and proposed method for Correctly classified images

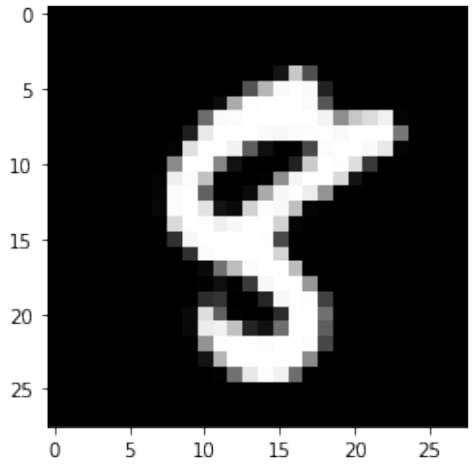
Type	**Method**	**IMI1**	**MC-Mean**	**MC-STD**
Baseline	C-Min Edit	5.2558	0.4013	0.1890
Baseline	PIECE	7.4811	0.1527	**0.0540**
Proposed Method	Plausibility Regulation	**3.2678**	**0.5715**	0.1388

The table 5.1 denotes that for the Correctly classified images, the proposed method is favourable considering the IMI1 score while the scores of MC-Mean and MC-STD favours the second baseline method that uses the PIECE algorithm. To get an overall perspective, we visualise the results so that we understand the qualitative aspect of the evaluation,

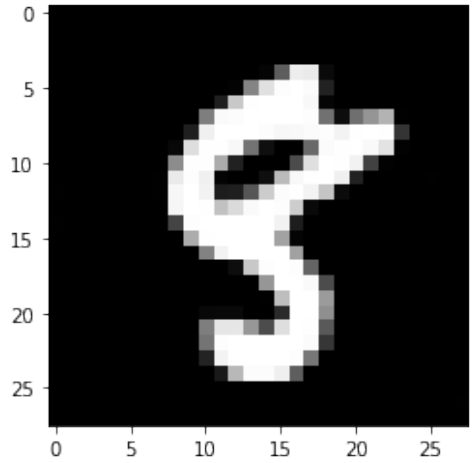
We consider a case where the given image corresponds to digit "8" and has the same prediction by the classifier and we want to generate the counterfactual image with target class as "3". As shown in Fig 5.5, the baseline methods generate the images that does not seem to be digits, while the proposed method generates plausible results. This example is in line with our quantitative evaluation.



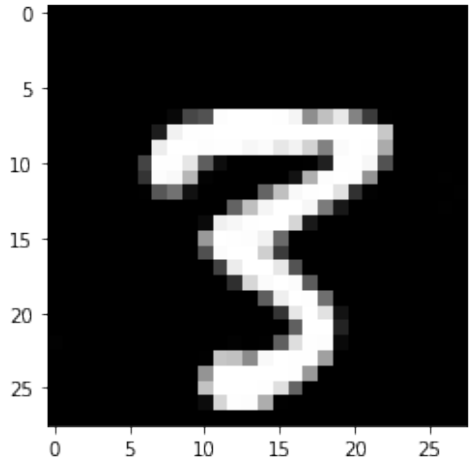
(a) Original Image



(b) Baseline - C-Min Edit Method



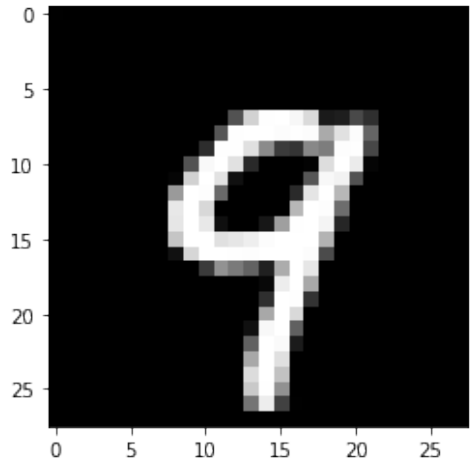
(c) Baseline - PIECE Method



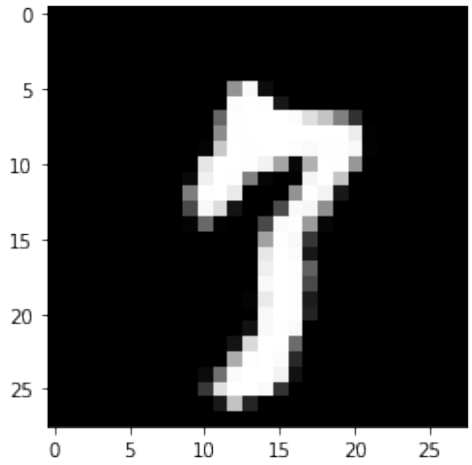
(d) Proposed Method

Figure 5.5: Results of Baseline methods and Proposed methods when generating counterfactual image for original image of digit 8 to target class 3

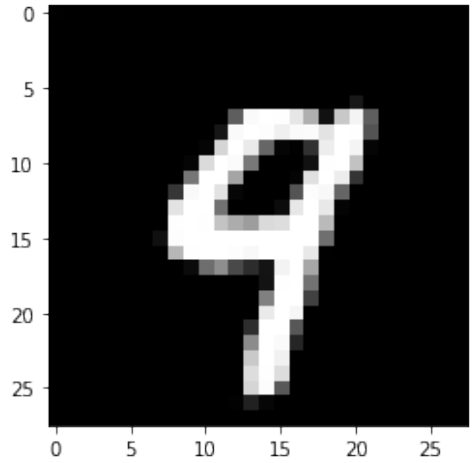
Further qualitative analysis also showed us some cases where the proposed method didn't show promising results. Consider the case shown in Fig 5.5, we have an image with prediction as digit "9" by the classifier. The results produced by Baseline C-Min Edit method are favourable than other methods even though none of the evaluation metrics favoured it.



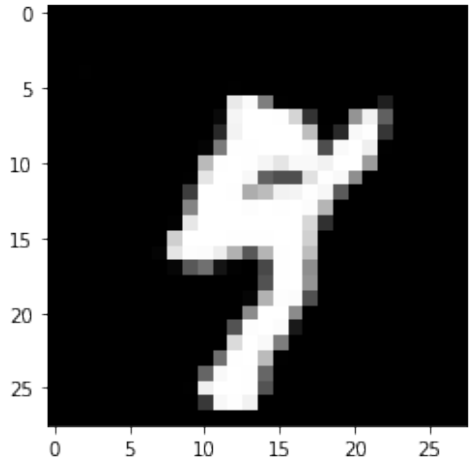
(a) Original Image



(b) Baseline - C-Min Edit Method



(c) Baseline - PIECE Method



(d) Proposed Method

Figure 5.6: Results of Baseline methods and Proposed methods when generating counterfactual image for original image of digit 9 to target class 3

After taking a look at both types of evaluation, we can infer that the proposed method is preferred over the baseline considering the correctly classified instances and this is nacked up by the quantitative and the qualitative analysis discussed above.

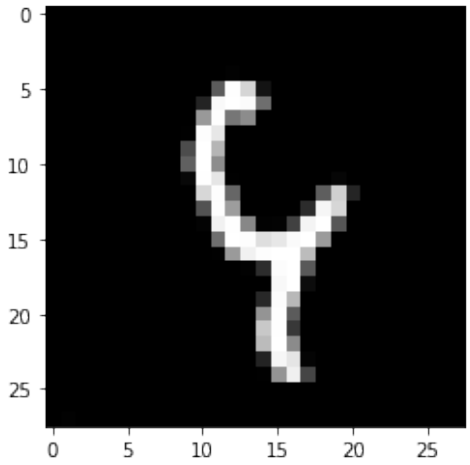
5.1.4 Evaluation on misclassified instances

For the evaluation on this category of data, we have considered 41 images. These images have incorrect classification when compared with their true labels. The counterfactual class for these images is the true label of the given image. Again here We calculated the values of evaluation metric on these images to understand if the proposed method is better in comparison to the baselines.

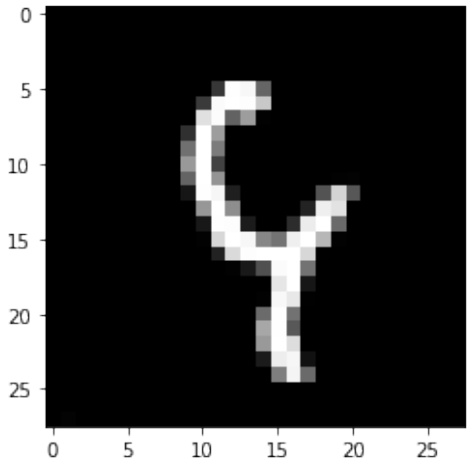
Table 5.2: Comparison of evaluation metrics for baseline and proposed method for misclassified instances

Type	**Method**	**IMI1**	**MC-Mean**	**MC-STD**
Baseline	C-Min Edit	3.8434	0.0993	0.0583
Baseline	PIECE	**2.6309**	**0.6139**	0.1670
Proposed Method	Plausibility Regulation	2.7827	0.6134	**0.1660**

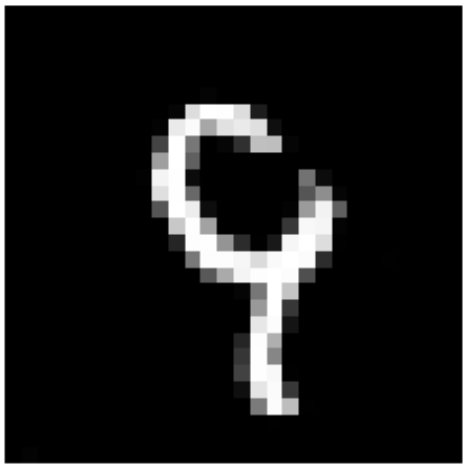
The table 5.2 denotes that the baseline C-Min Edit method is favourable considering the IMI1 score while the scores of MC-Mean and MC-STD favours the proposed method. In order to understand the qualitative aspect of the evaluation, we chose a few cases to visualise the results. As seen in Fig 5.7, we have an original image whose prediction is "4", but the true label is "9", which is why we generate the counterfactual using all the methods by specifying target class as "9". The counterfactual generated by the proposed method seems more plausible in comparison to others by human observation. Further, we chose an image of "8" that was misclassifeid as "3". We generate counterfactual for this image as shown in Fig 5.8. The results generated by PIECE method are much more plausible than the other two methods.



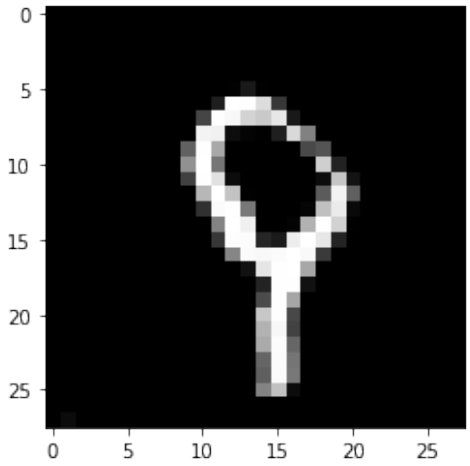
(a) Original Image



(b) Baseline - C-Min Edit Method

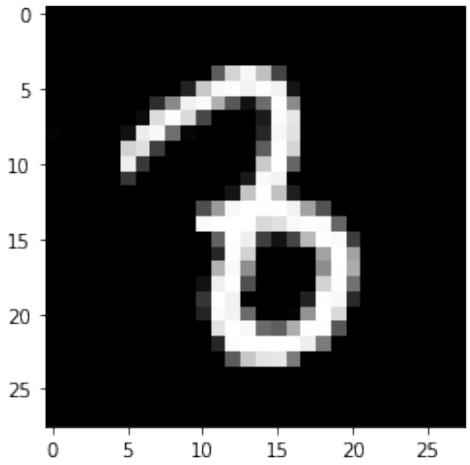


(c) Baseline - PIECE Method

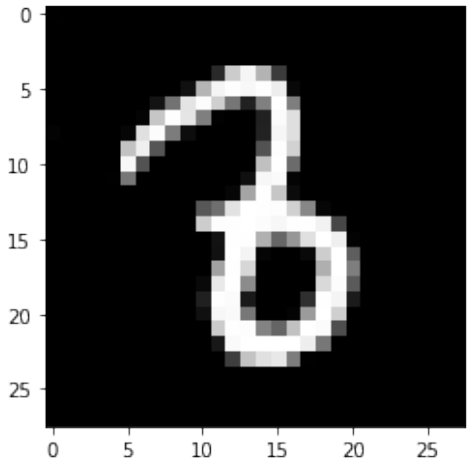


(d) Proposed Method

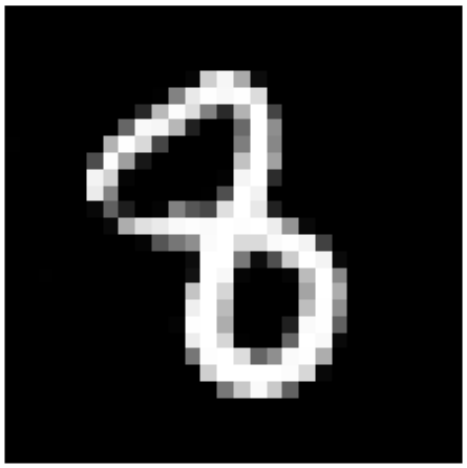
Figure 5.7: Results of Baseline methods and Proposed methods when going from original image of digit "4" to "9"



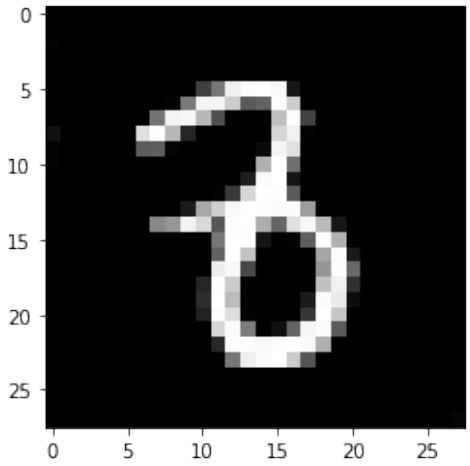
(a) Original Image



(b) Baseline - C-Min Edit Method



(c) Baseline - PIECE Method



(d) Proposed Method

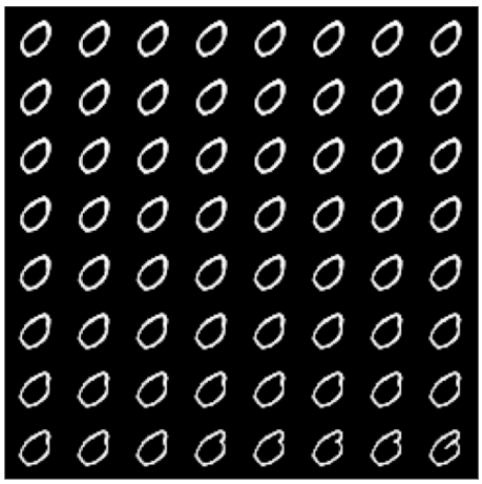
Figure 5.8: Results of Baseline methods and Proposed methods when going from original image of digit "3" to "8"

The qualitative evaluation points out the performance of PIECE is better than the proposed method but in a few cases proposed method outperforms the PIECE results. Considering both the quantitative and the qualitative evaluation, it can be inferred that the PIECE method is favoured while dealing with misclassified instances. On further analysing the misclassified instances we realise that these instances are the boundary line cases where the classifier makes mistakes in classifying them in the right category. The PIECE method uses the class distribution information to move to the counterfactual class and the boundary line cases are very well suited to the use case of PIECE and that is evident in the evaluation. But when we consider the correctly classified instances, these instances are away from the boundary line which is evident from the higher value of prediction probability. This makes it difficult for the PIECE method to alter the pixels in the original image and push it to the counterfactual class. But in cases like these the proposed method is able to reach the counterfactual class because it uses the optimisation function that stops the image generation process only when the particular class is reached.

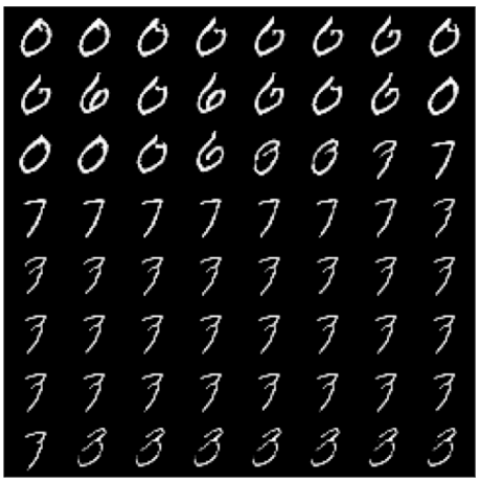
5.1.5 Journey of original image to counterfactual image

After analysing the performance of the proposed method on the different sets of data, here we explore the journey of the images when going from original image to the counterfactual image. The images are captured at every steps of the optimisation to discover how the proposed method differs from the baseline method.

In Fig 5.9, we use the baseline C-Min Edit and the proposed method to understand the difference in the journey of the original image of digit "0" that changes to the counterfactual image of digit "3". As seen in Fig 5.9 (a), the optimisation process stops when the image is predicted as "3" by the classification model. On the contrary, as seen in Fig 5.9 (b) the proposed method stops only when the digit is classified as "3" and the plausibility threshold is achieved. It is also evident that the digit "0" changed to digit "6", "7" and then to "3". These images show the diversion in the optimisation process of these methods



(a) Optimisation steps by baseline C-Min Edit method

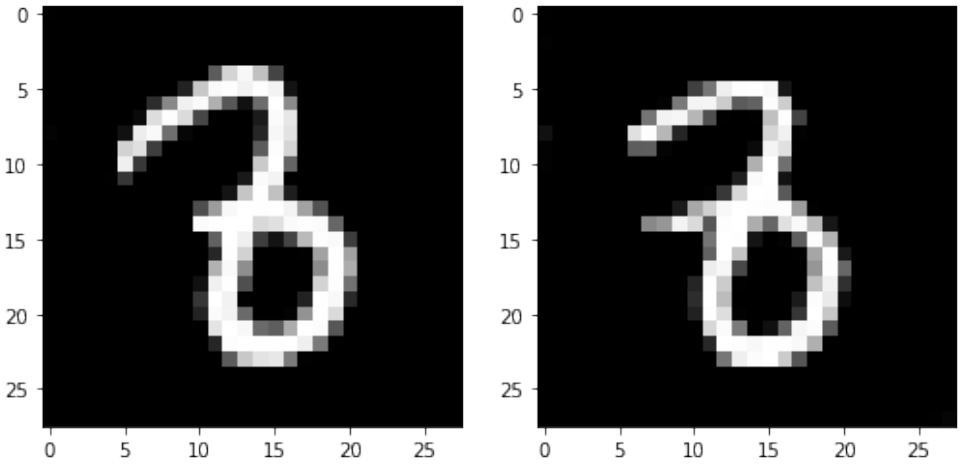


(b) Optimisation steps by proposed method

Figure 5.9: Journey of the original image to counterfactual image

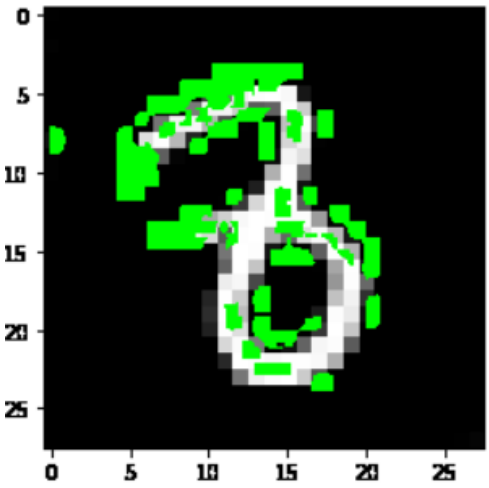
5.1.6 Detecting the changes by analysing the transformation of images

Generating the plausible counterfactuals will help the stakeholders in developing trust into the classifier, in other cases it might also help the application developers to evaluate the fairness of the model or in some cases understand how the model behaves in cases of uncertainty. But, we cannot simply offer the original image along with the counterfactual image side by side to the user to analyse what changed. The user must be provided with transformation maps or diff images so that they understand what set of pixels were altered to attain the counterfactual prediction. In case of MNIST, we decided to explore the pixel difference technique as it provides simple visual explanation interpretable to humans. Consider the case where the original image has the prediction of "3" and the counterfactual image has the prediction of "8" as shown in Fig.5.10 (a) and (b) respectively. The difference image is shown in Fig.5.10 (c). The pixels highlighted in green is the change in the counterfactual image with respect to the original image. In this case it is evident that the green pixels are surrounded around the upper circle of digit "8" which is incomplete in the original image, but in the counterfactual image, the upper circle is attempted to be complete so as to reach the perfect shape of "8". If a human was provided with this example and asked to transform the image from "3" to "8", the human would have taken similar approach. This points in the direction of human interpretability of the pixel difference technique and also assists to build trust of the user in the system as they can relate to the changes denoted by the difference image.



(a) Original Image with prediction of "3"

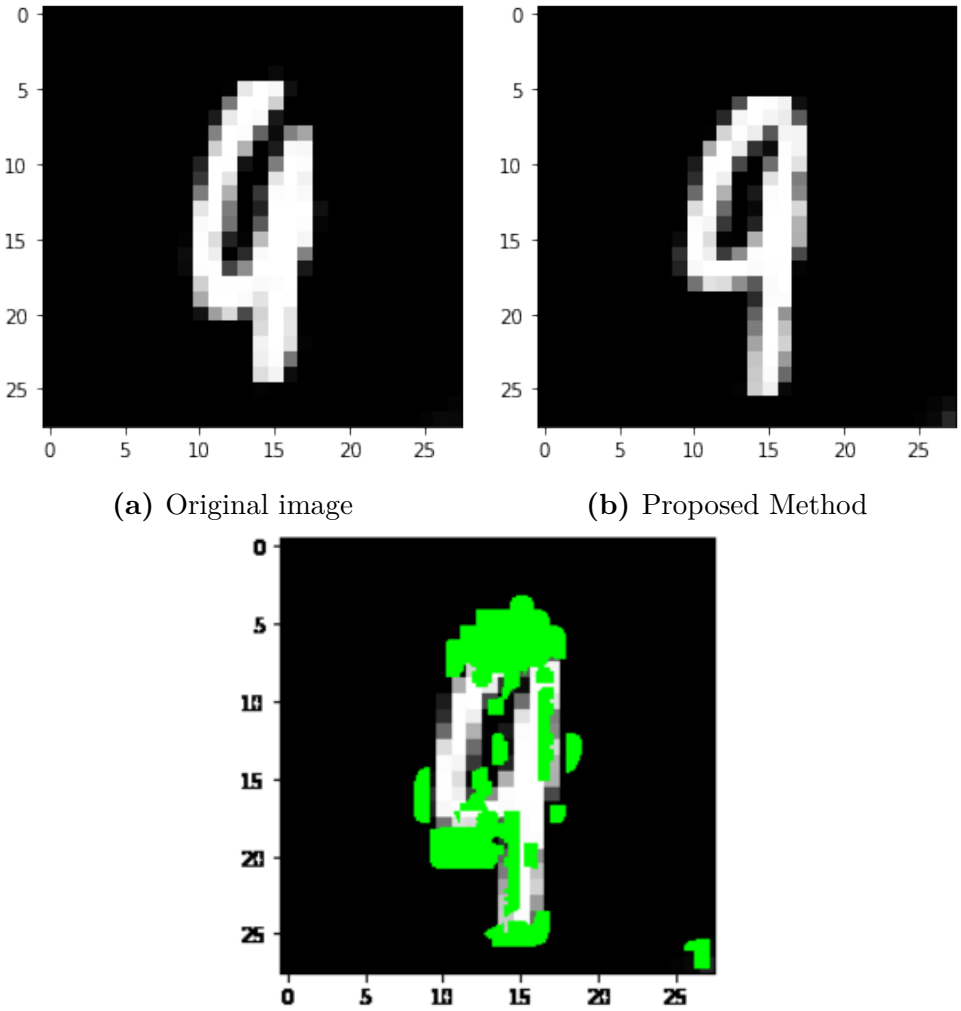
(b) Counterfactual Image with prediction of "8"



(c) Difference Image with respect to original image

Figure 5.10: Detecting the change in images using pixel difference technique

Next, consider the case of unsure digit detection as shown in Fig 5.11(a), the digit can be "4" or "9". In such case the prediction probabilities of these digits are closer to each other. In this case imagine the classifier detects the digit to be "4" and now you want to understand what should be modified to make it "9". We provide the counterfactual class to be "9" and generate the plausible counterfactual using the proposed method as shown in 5.11(b). In cases like these, the pixel difference can highlight the addition or reduction of pixels. As seen in 5.11(c) the pixel difference image shows the green pixels that needs to be added to the original image to make it "9". The green pixels are around the upper part of original image that joins both the curves making it a "9" as interpretable to human.



(a) Original image **(b)** Proposed Method

(c) Difference Image

Figure 5.11: Difference Image

5.2 Experiment 2: BraTS MRI Dataset

5.2.1 Dataset

The BraTS dataset [30] provides a collection of Magnetic Resonance Images(MRI) of the brain for studying the segmentation of Brain tumours. It consists of 220 images of brain with High Grade Gliomas(HCG). These images are available in four modalities T1, T2, Flair and T1Ce sequences. We decided to use the T1Ce modality for our experimentation. The 220 volumes were divided into train, validation and test set. The trainset had 180 volumes while the validation and the test set had 20 volumes each. These volumes were available as the set of 2D slices. The tumour labels were available on the slice level. Each slice had either a label 0, 1 or 2. The label 0 denoted slice with no tumour, while slice with label 1 and 2 denoted presence of tumour of different types. To make this a binary classification task, we selected the slices of label 0 and 2 for this experiment. The slices without any bright pixel values were removed. The pixel values of the slices were were normalised using min-max normalisation based on the value of each slice, this made the

values of the pixels bounded between 0-1. The Fig 5.12 shows the samples of images from the BraTS dataset.

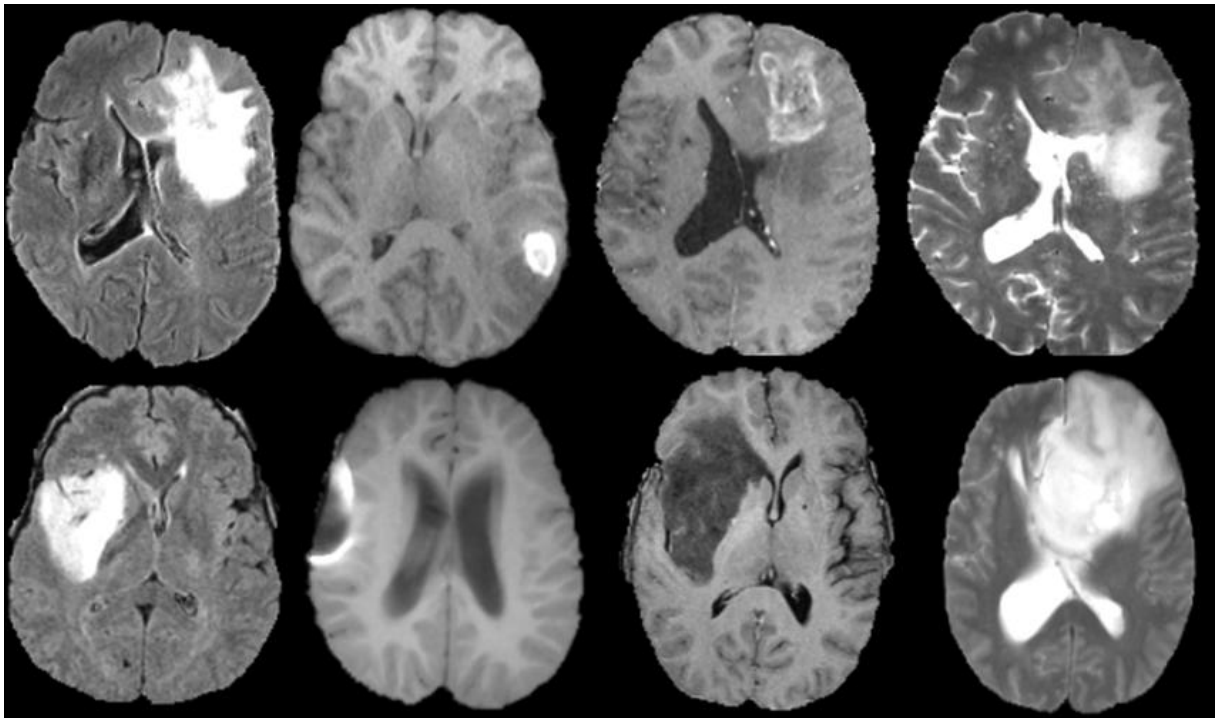


Figure 5.12: Samples of BraTS MRI Dataset [30, 31, 32]

5.2.2 Experimental Setup

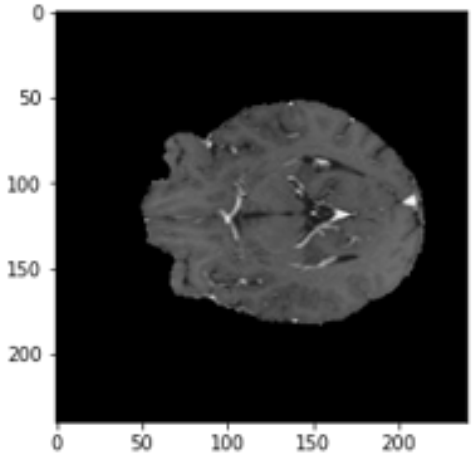
The 2D slices from BraTS dataset were available as 240x240 pixel wide. We intended to train a classifier to identify whether the 2D slice contains tumour or not and based on the prediction of the classifier the slices were classified as "tumour detected" or "no tumour detected". We used the classical Convolutional Neural Network with 4 Convolution blocks to train the classifier over the training set. Upon testing the performance over the test set, the classifier gave an accuracy of 53%, which is almost similar to random guessing. To get a better accuracy we decided to switch to powerful classifiers like VGG Nets and ResNets that have numerous convolutional blocks stacked on top of each other. We implemented the ResNet-101 which uses 101 deep layers of residual network to train the network. ResNets have proven to be effective [33] in comparison to traditional network given their skip connections that solves the vanishing gradient problem in deep neural networks. The ResNet-101 model gave an accuracy of 82% over the test set. This accuracy is better in comparison to the classical neural networks and also it also helps us in classifying the images of tumour and non tumour with some degree of confidence. So we decided to finalise on this classifier for sake of experimentation with counterfactual on BraTS dataset.

5.2.3 Experimentation and Evaluation of GANs

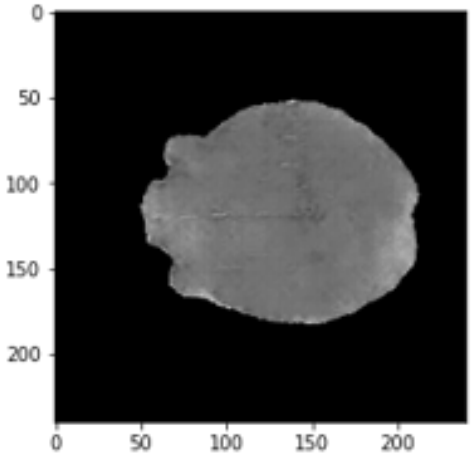
We have termed the BraTS dataset as a complex dataset, because of the images of brain that have complex structures. To generate counterfactual over this dataset, we need a GAN model that can generate these complex images as closer to reality as possible, also the matter of concern here is not only the brain images, but the GAN should also be able to generate the tumour sections precisely.

We decide to train various GAN models to understand their performance over the BraTS model and then choose the suitable GAN for further experimentation. We started with the DC-GAN, where we use the latent vector of 128 dimension as input to the GAN. We trained this network until convergence and noticed that it generated images of the MRI dataset as seen in Fig 5.13(b). But when we compare this image with the original image Fig 5.13(a), it can be inferred that the generated image is mostly blur and DC-GAN lacks the ability to generate complex brain structures in case of this dataset. The problem of DC-GAN is that it faces the problem of vanishing gradients and that is why to tackle this problem, we decided to change the loss function of the GAN, so we decided to train the WGAN as it offers better quality of images by reducing the problem of vanishing gradients. The images generated by WGAN are shown in 5.13(c), these image is better than DC-GAN, yet it is blur and did not contain all the complex structure of the brain.

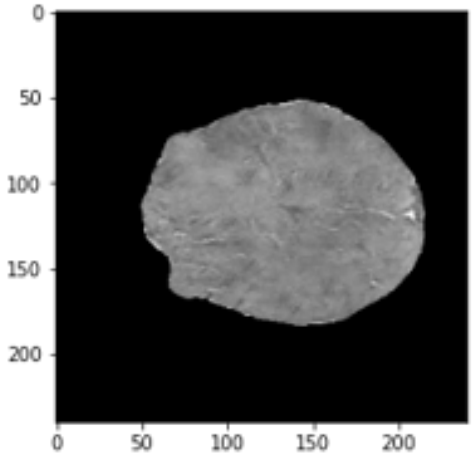
At this point we hypothesised that the GANs were not able to map all these complex structures to the input latent dimension of 128, so we increased the latent dimension to 1000 and then to 2000 for assessing its effect over the quality of images. Even after using higher dimension latent vector, we did not observe noticeable change in the quality of generated images. The authors of WGAN research paper [34] also suggests gradient penalty version of WGAN and mentions the improvement in performance, which is why we trained the WGAN-GP and the results of it can be seen in 5.13(d). The results did not offer any kind of improvement over DC-GAN or WGAN. This motivated us to think in the direction of making the generator architecture in the GAN more powerful along with the increased latent size. The U-Net architecture has proved to be outstanding in biomedical image segmentation [35], further the skipped connections in U-Net helps in reducing the problem of vanishing gradients and also promises improved performance even with smaller sizes of training datasets. This motivated us to use the U-Net as the generator in the WGAN. We term this as U-Net GAN, as seen in 5.13(f). As seen in the U-Net WGAN results the generated image contains complex brain structures upto certain extent and it outperforms all the other GANs. At this point we wanted to explore further powerful architectures like ReconResNet [28] by Chatterjee et al. In this case we replace the U-Net with ReconResNet and the result generated is shown in 5.13(e). The results of ReconResNet WGAN are better than other GANs as it succeeds in generating complex structures upto certain extent yet it does not surpass the performance of the U-Net WGAN. We tried this qualitative comparison on many other images to conclude that the U-Net WGAN outperforms all other GANs tried by us and hence we decided to use it in the proposed method



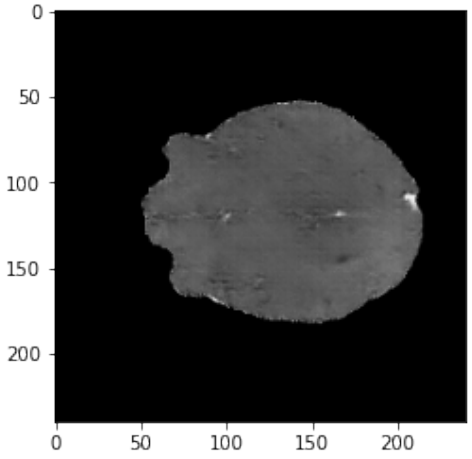
(a) Original Image



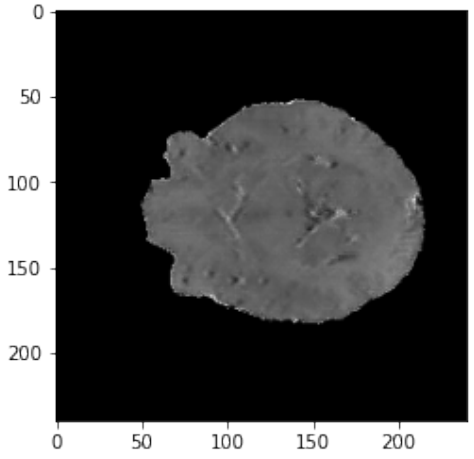
(b) DC-GAN



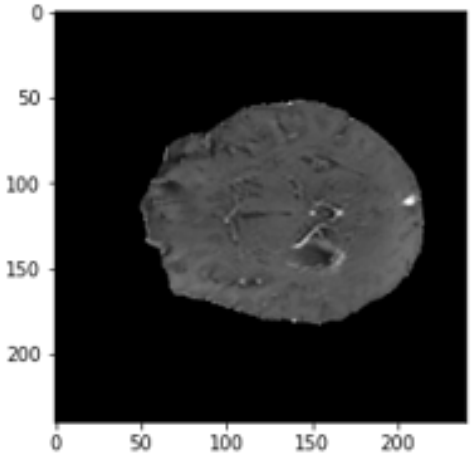
(c) WGAN



(d) WGAN-GP



(e) ReconResnet WGAN



(f) U-Net WGAN

Figure 5.13: Comparison of Original image along with images generated by various GANs

The summary of the experimentation with GAN can be seen in the table 5.3. We calculate the FID scores that point towards the quality of the images generated by the GAN, FID score represents the distance between the original distribution and the distribution of the

generated image. That is why lower FID score represents high quality images. Further we calculate the perceptual loss between the original image and the generated image. We consider test set of 20 volumes and calculate the FID score and the perceptual loss over these images over all the trained GANs to assess the performance of these GANs. The FID score in this case points out the performance of the DC-GAN which promises lower value of the FID in comparision to the other GANs, considering the perceptual loss values, the ReconResNet seems to provide lower values of the loss and is preferred over the other GANs. But looking at the qualitative results of the GAN, we prefer to go ahead with the U-Net WGAN as it generates images that seem closer to the original distribution.

Table 5.3: Comparison of evaluation metrics for baseline and proposed method for Unsure instances

GAN	**Perceptual loss (mean)**	**Perceptual loss (std)**	**FID**
DC-GAN	0.0186	0.0061	**405.62**
WGAN	0.0200	**0.0041**	475.30
WGAN-GP	0.0189	0.0050	490.66
ReconResNet WGAN	**0.0145**	0.0061	506.75
U-Net WGAN	0.0189	0.0067	470.54

We performed a through evaluation on MNIST in the prior section by splitting the test set in 2 parts. Again here we aim to split the test set in two parts, correctly classified instances and misclassified instances. These parts help us in understanding how the proposed method performs in comparison to the baselines in different sets of test data. The baseline2: PIECE method requires the class distribution information about the given dataset, in case of BraTS we could not acquire this information and that is why the evaluation is performed by comparing the performance of baseline 1: C-Min Edit method and proposed method. To attain a complete perspective, we have compared the values of evaluation metrics to get an understanding of the quantitative evaluation while we compare the results by looking at the results side by side to attain the qualitative evaluation of these methods.

5.2.4 Evaluation on correctly classified instances

As the title suggests, we have considered 50 images here, these images are correctly classified by the trained classifier. There are 25 images that have the label "tumour detected" while the remaining 25 images have the label "tumour not detected". This equal distribution of the correctly classified images will help us to understand how the proposed method performs over the balanced test set.

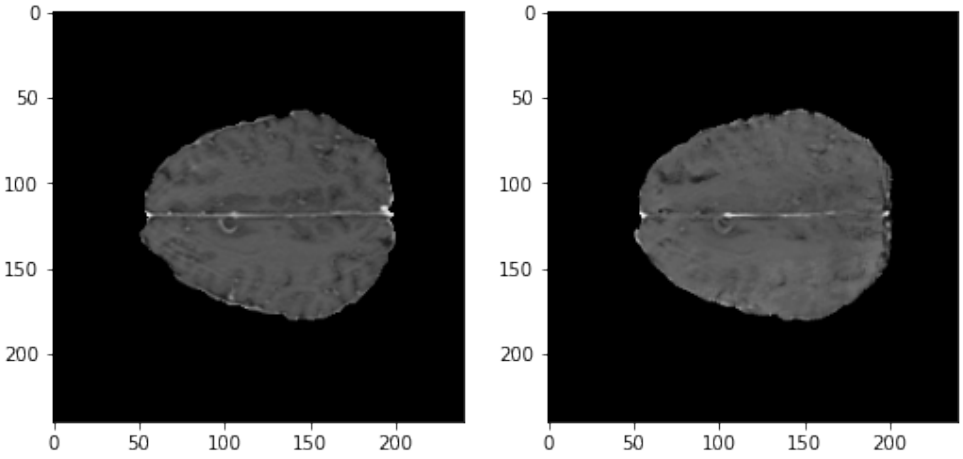
The evaluation metrics were calculated on the results derived from baseline 1: C-Min Edit method and proposed method, the results are documented in the table 5.4. The quantitative evaluation points out that the proposed method is favourable considering the MC-Mean and MC-STD metrics, as we see a significant difference in the values. The IMI1 metric values shows no significant changes in the performance of these two methods. We further perform mannwhitney statistical test on these results for IMI1 and realise that there is no statistical significant difference in the IMI1 values.

Table 5.4: Comparison of evaluation metrics for baseline and proposed method for correctly classified instances

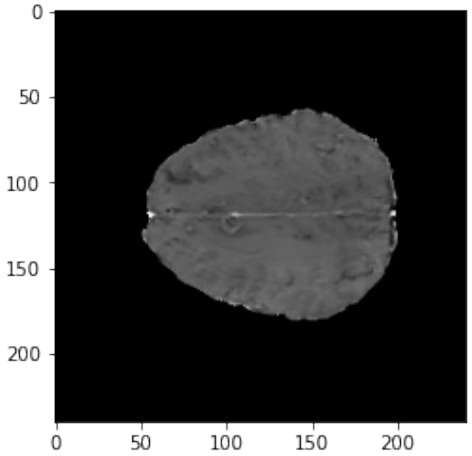
Type	**Method**	**IMI1**	**MC-Mean**	**MC-STD**
Baseline	C-Min Edit	1.006	0.7556	0.1610
Proposed Method	Plausibility Regulation	**1.011**	**0.8052**	**0.1460**

For the qualitative analysis we consider two cases, first a case where the prediction of the classifier on the given image is "no tumour detected" as shown in Fig 5.14(a), we provide this image to the baseline and the proposed method with the counterfactual class as "tumour detected". Looking at the results by baseline in 5.14(b) and 5.14(c), it can be inferred that there is not much difference noticed in the results of these methods by human observation. Also, the results do not seem distinct from the original image. This seems like a case where certain pixels were altered to get the classifier to avert its prediction. Let consider another case where the prediction on the given image is "tumour detected" as shown in Fig 5.15(a), we specify the counterfactual class to be "no tumour detected" and pass on these inputs to the baseline and the proposed method. The results are shown in Fig 5.15(b) and Fig 5.15(c) respectively. In this case we can realise that the tumour like black circular patch in the side of the brain slice has disappeared in the counterfactual image. The counterfactual images from both these methods looks similar and attempt to vanish the tumour like parts from the original image.

Considering the quantitative and qualitative evaluation over the classified instances, it can be inferred that there is no significant improvement in the performance of the proposed method in comparison to baseline.

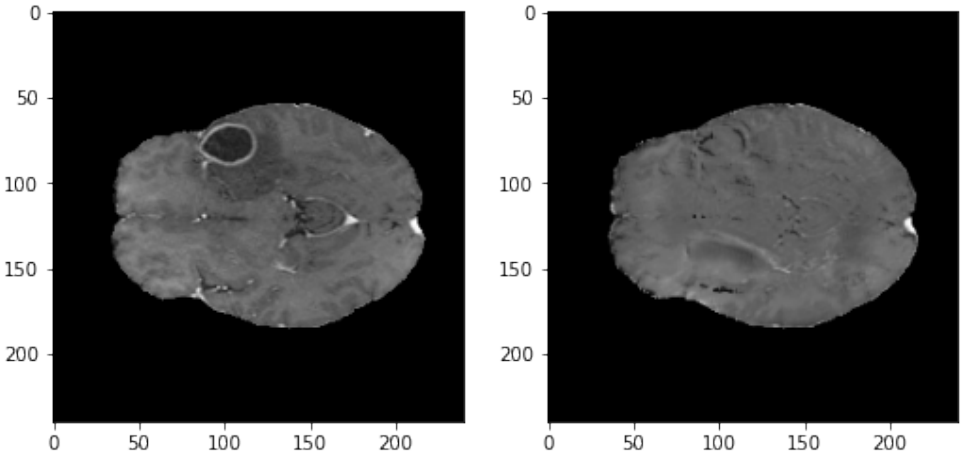


(a) Original image **(b)** Baseline - C-Min Edit Method

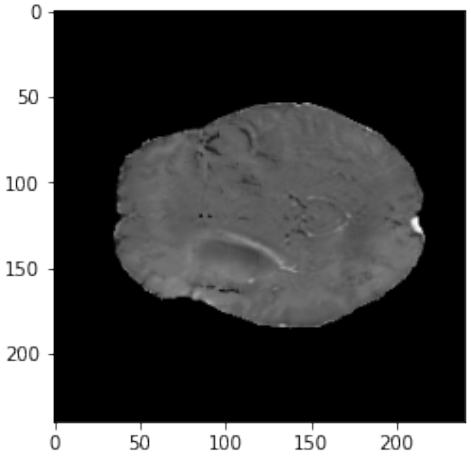


(c) Proposed Method

Figure 5.14: Results of Counterfactual image generation over a given image with prediction "no tumour detected"



(a) Original image **(b)** Baseline - C-Min Edit Method



(c) Proposed Method

Figure 5.15: Results of counterfactual image generation over a given image with prediction "tumour detected"

5.2.5 Evaluation on misclassified instances

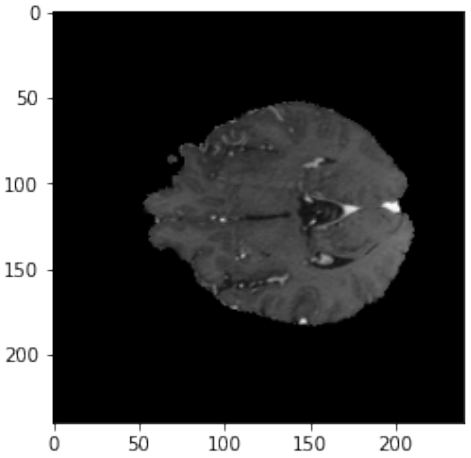
Here, we have considered 50 images as well, these images are misclassified by the trained classifier, which means if the classifier predicts that the given image has "no tumour detected", the true label of the image is "tumour detected". This is why we provide the true label as the counterfactual class in generating the counterfactual images for these instances.

The table 5.5, provides quantitative evaluation over the misclassified instances by providing the values of evaluation metrics over the baseline and the proposed method. The values suggest that the proposed method outperforms the baseline method in terms of MC-Mean and MC-STD metrics, but when we compare the IMI1 score the performance seems similar. We use the mannwhitney statistical test to determine if there is significant difference in the IMI1 scores. The test confirms that there is no significant difference in the IMI1 scores and which is why we can ascertain that the proposed method does outperform the baseline method in terms of IMI1.

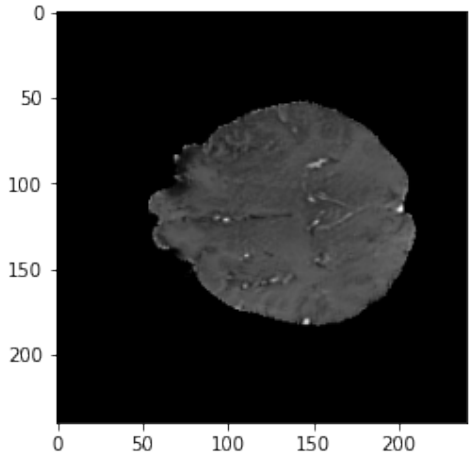
Table 5.5: Comparison of evaluation metrics for baseline and proposed method for misclassified instances

Type	**Method**	**IMI1**	**MC-Mean**	**MC-STD**
Baseline	C-Min Edit	0.972	0.7999	0.1553
Proposed Method	Plausibility Regulation	**0.980**	**0.8423**	**0.1272**

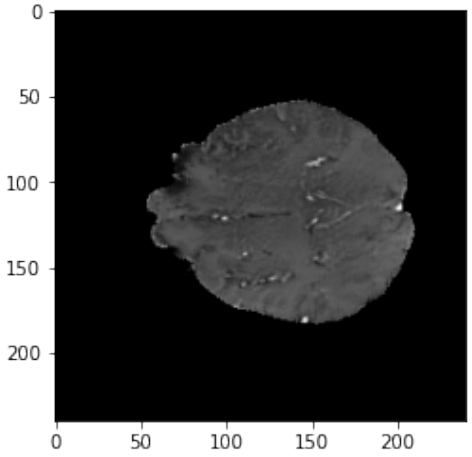
For qualitative evaluation we consider two cases and visualise the results to get an understand the performance of baseline and proposed method. Consider the first case where the image is predicted as "tumour detected" as shown in Fig (a), but the true label is "no tumour detected" and that is why the counterfactual class is "no tumour detected". The results of the baseline and the proposed method are shown in Fig (b) and Fig (c) respectively. The counterfactual results show that the black regions in the original image are blurred to switch the prediction to "no tumour detected".



(a) Original image



(b) Baseline - C-Min Edit Method

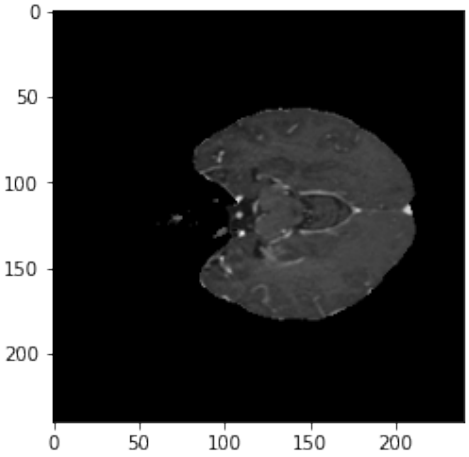


(c) Proposed Method

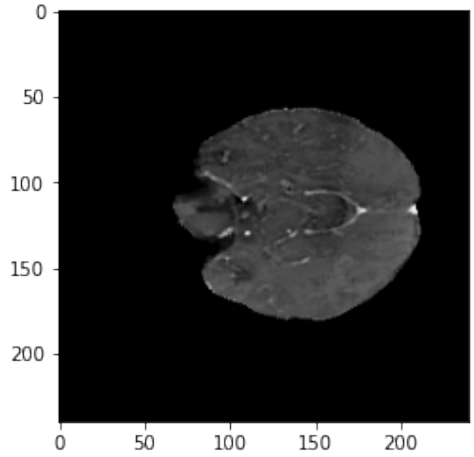
Figure 5.16: Results of counterfactual image generation over a given image with incorrect prediction of "tumour detected" but the true label is "to tumour Detected" and that is why the Counterfactual class is "no tumour detected"

Consider the second case, where the prediction of the classifier for the given image is

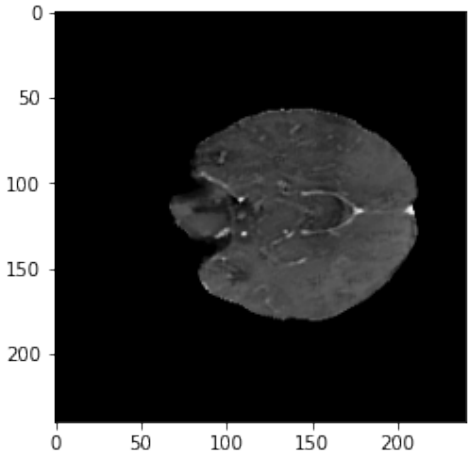
"no tumour detected", but the true label is "tumour detected". We provide this as the counterfactual class to derive the results using the baseline and proposed method as shown in Fig . The results of the proposed method suggests that there is addition of black regions in the original image which inverts the prediction of the image to "tumour detected". Also we see a irregular structure added at the left side of the brain MRI which majorly pushes the prediction to presence of tumour.



(a) Original image



(b) Baseline - C-Min Edit Method



(c) Proposed Method

Figure 5.17: Results of counterfactual image generation over a given image with incorrect prediction of "no tumour detected" but the true label is "tumour detected" and that is why the counterfactual class is "tumour detected"

5.2.6 Detecting the changes by analysing the transformation of images

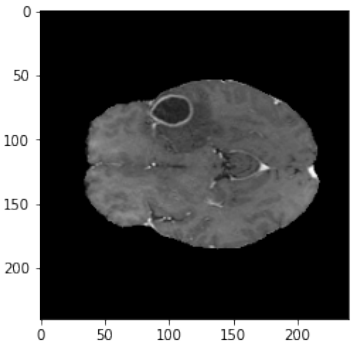
In the tumour classification applications, it is critical to explain the stakeholders on how the classification model makes a decision, the counterfactual image generation methods discussed above provide an overview on how the counterfactual image generation can provide local explanations eg: consider one image that consists of tumour and the counterfactual image generation method outputs an another image that is close to the given

image but has alternate prediction. These explanations can be made better by providing the user with the transformation maps that assist the users in understanding what changed in the image so that the prediction altered. For the BraTS dataset we intended to use the image registration techniques to attain the transformation maps that assist the users in attaining better understanding of how counterfactual images are derived from original image.

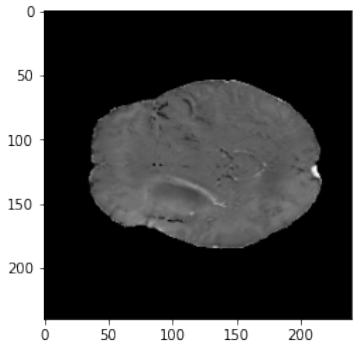
The generated images by the U-Net WGAN are better in comparison to the other GANs we discussed earlier, but these images are not closer to the original images. This impedes the usage of image registration techniques as it cannot accurately map the original image to counterfactual image. This motivated us to look for other techniques that can help the users understand what changed in the image. The simplest concept that we started with was the difference image or mask. We implemented the difference mask by visualising the image that is obtained from subtracting pixels of one image from another. Further, we use the manual thresholding technique to attain the mask that explains the key regions of difference in the counterfactual image when compared to the original image. In Image segmentation we often use Otsu's thersholding technique to segment the image into foreground and background, We utilise this method to derive the explanation mask.

We have considered the case of an image that has the prediction as "tumour detected" and we provide this image to the proposed method given the counterfactual class as "no tumour detected". The results can be seen in Fig 5.18, we showcase the original image as well as the counterfactual image in Fig 5.18(a) and Fig 5.18(b) respectively. Further we visualise the difference mask as shown in Fig 5.18(c), upon further thresholding using manual and Otsu's method as shown in Fig 5.18(d) and Fig 5.18(e) it can be understood that the major changes were noticed in the upper region of the image that had tumour like black region.

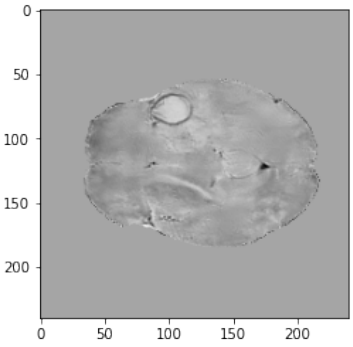
Further, we consider the case of an image that has the prediction as "tumour detected" and it is an incorrect prediction given the true lable and we provide this image to the proposed method given the counterfactual class as "no tumour detected". The results can be seen in Fig 5.19, we showcase the original image as well as the counterfactual image in Fig 5.19(a) and Fig 5.19(b) respectively. Further we visualise the difference mask as shown in Fig 5.19(c), upon further thresholding using manual and Otsu's method as shown in Fig 5.19(d) and Fig 5.19(e) it can be understood that the major changes were noticed in the central region of the image that had tumour like black region and is now replaced by mass like region in the counterfactual image.



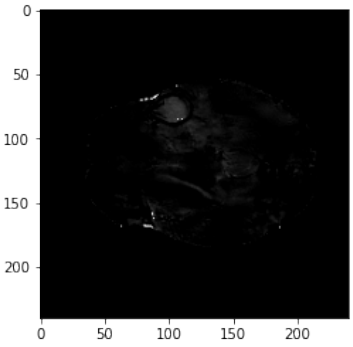
(a) Original image



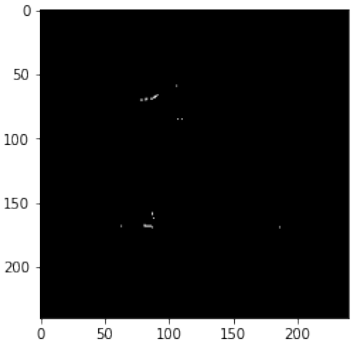
(b) Proposed method



(c) Difference mask

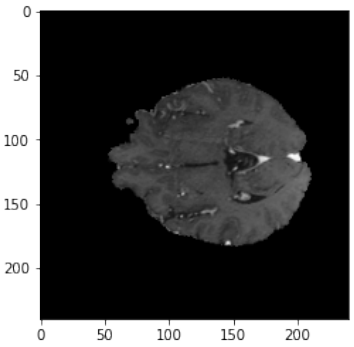


(d) Difference mask with manual threshold

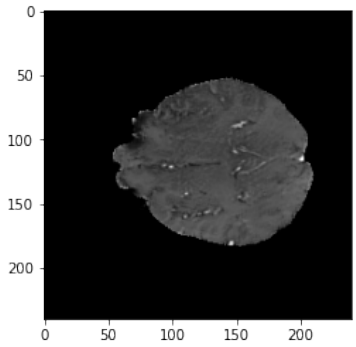


(e) Difference mask with Otsu's threshold

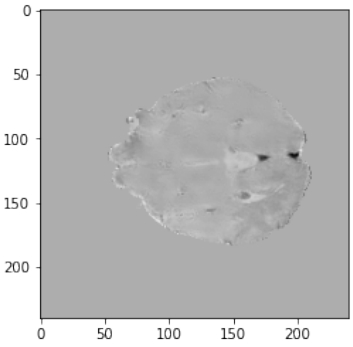
Figure 5.18: Results of difference mask and further thresholding of the mask for image with prediction "tumour detected"



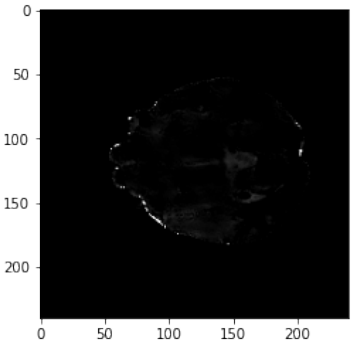
(a) Original image



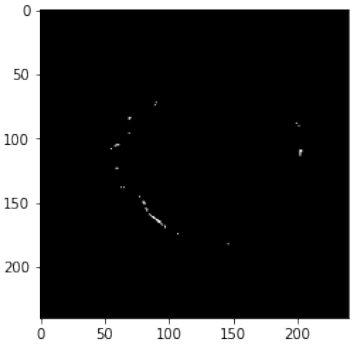
(b) Proposed method



(c) Difference mask



(d) Difference mask with manual threshold



(e) Difference mask with Otsu's threshold

Figure 5.19: Results of difference mask and further thresholding of the mask for image with prediction "tumour detected"

6 Discussion

In this chapter, we discuss the merits and limitations of the proposed method based on the evaluation chapter. We analyse the performance of the proposed method in terms of two different datasets, first MNIST dataset which contains simple shapes and structures and then BraTS dataset that has brain MRI images which contains complex structures. We also consider different set of instances in these datasets to understand how the proposed method performs when the instances are correctly classified or misclassified by the trained classifiers. At last we intend to use this chapter to answer the research questions one by one.

6.1 MNIST Dataset

The MNIST dataset contains of handwritten digits and it is widely used for assessing the performance of counterfactual explanations in images [1, 11]. The study of the performance of the proposed method on MNIST helps us to understand the effectiveness of the method on simple datasets, also it motivates us to further assess the method on complex datasets. The evaluation on MNIST dataset is performed quantitatively and qualitatively, the quantitative evaluation uses various metrics to denote which method outperforms the other, while the qualitative evaluation visualises all the results so that the results from the quantitative evaluation can be validated by observing these visualisations. The qualitative evaluation is subjective yet simpler in the case of MNIST, as the user can look at the generated images and provide feedback whether the image actually looks like a digit.

We performed the quantitative and the qualitative evaluation by splitting the test set into two parts, correctly classified and misclassified instances. The section 5.1.3 shows the evaluation on the correctly classified instances, the quantitative evaluation shows that the proposed method outperforms both the baselines in terms of evaluation metrics IMI1 and MC-Mean as shown in table 5.1. The qualitative evaluation in cases similar to discussed in Fig 5.5 shows that the results of proposed method actually outperform the baselines, then we consider the case as shown in Fig 5.6 where the proposed method did not show promising results in comparison to baseline 1 C-Min Edit method but still efficient as compared the state-of-the-art baseline 2 PIECE method. In case of correctly classified instances we can conclude that the proposed method performs significantly better in comparison to the baselines.

Then we consider the misclassified instances, we perform quantitative evaluation as shown in 5.2 and it point that the baseline 2 PIECE method outperforms the other methods in these set of instances. We visualise the results to understand the qualitative aspect of evaluation, we visualise the first case as shown in Fig 5.7, and the performance of baseline 2 and proposed method is found to be similar. On further analysis we found

many cases as shown in Fig 5.8 which denotes that the PIECE method outperforms the proposed method. It can concluded based on the evaluation of the misclassified instances that the PIECE method outperforms the other methods quantitatively as well as qualitatively.

The outstanding performance of the proposed method in the correctly classified instances answers the [**RQ-1**] , by proving that the usage of discriminator as plausibility regulator in the counterfactual image generation improves the plausibility of the generated images. Further in cases of misclassified instances we see that the baseline 2 PIECE method is favourable, to understand this we looked at the nature of these instances, the correctly classified instances have higher prediction confidence and are away from the boundary, which is why the techniques like PIECE that operate on class distribution parameters cannot generate plausible counterfactual. The proposed method stops the optimisation only when the counterfactual class is reached along with the desired level of plausibility. This makes the proposed method more robust with instances that lie far away from the boundary line. The misclassified instance have lower probability confidence and lie on the boundary line, in such cases PIECE method is able to exploit the distribution properties to generate plausible counterfactual of the target class, while the proposed method stops its optimisation process as soon as the the generated digit crosses decision boundary and have plausibility higher than the threshold. This motivate us to push the plausibility threshold even further, and this shifted the focus of the optimisation process on plausibility entirely and the results were plausible digits but distant apart from the original image and this does not fit into the definition of counterfactual as defined by us in the Chapter 2.

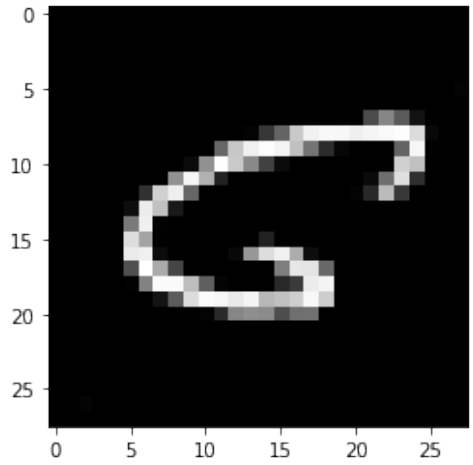
Further, we consider a category of data which we term as "unsure" instances, these instances can belong to both the classes taking into account their shapes. We select the instances that has softmax probability is < 0.8 and also upon human supervision a class cannot be assigned to these images with surety. These instances are different from the misclassified instances, because the prediction on these images is correct but yet they cannot be given a label that is agreed by all the human supervisors. In total we selected 62 such images and wanted to run evaluation on these kind of instances. The counterfactual class for these images is determined by using the gradient ascent method as shown in equation Eq 3.3.2. We used the baselines and the proposed method to attain results on these images. The quantitative evaluation is carried out by calculating the evaluation metrics and the results are shown in table 6.1. The evaluation shows that the baseline 2 PIECE method outperforms the other methods.

Table 6.1: Comparison of evaluation metrics for baseline and proposed method for Unsure instances

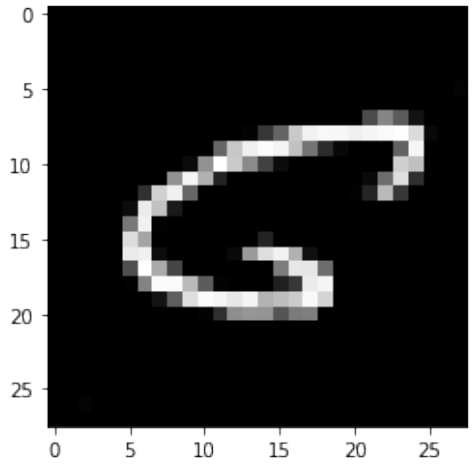
Type	**Method**	**IMI1**	**MC-Mean**	**MC-STD**
Baseline	C-Min Edit	1.0273	0.3560	0.1938
Baseline	PIECE	**1.0505**	**0.4834**	**0.0626**
Proposed Method	Plausibility Regulation	1.0284	0.4468	0.1762

For the qualitative analysis we choose 2 instances. First an image that looks either a "6" or "0". The model predicts it as "0", we generate counterfactual for this image with counterfactual class as "6". The results are shown in Fig. 6.1 and we see minor improvement

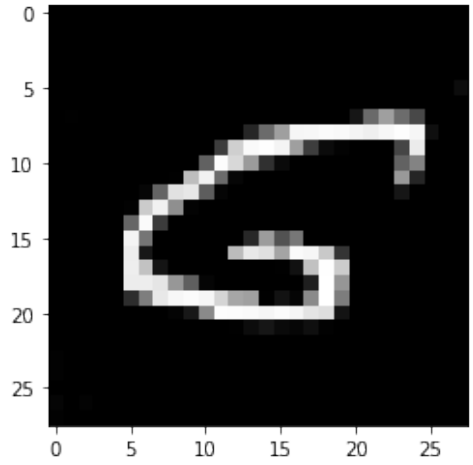
in the results of proposed method as compared to others. Secondly, we consider an image that was predicted as "2" and generate counterfactual with target class "0", as shown in Fig. 6.2, PIECE outperforms the other two methods.



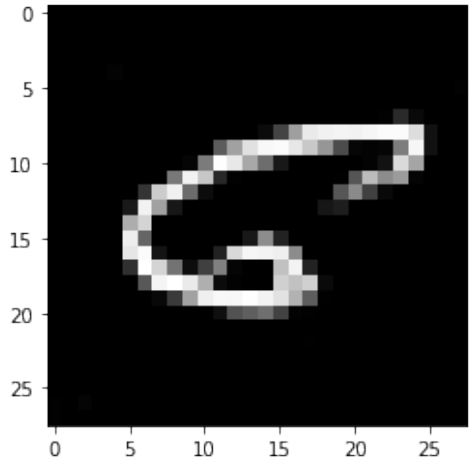
(a) Original Image



(b) Baseline - C-Min Edit Method



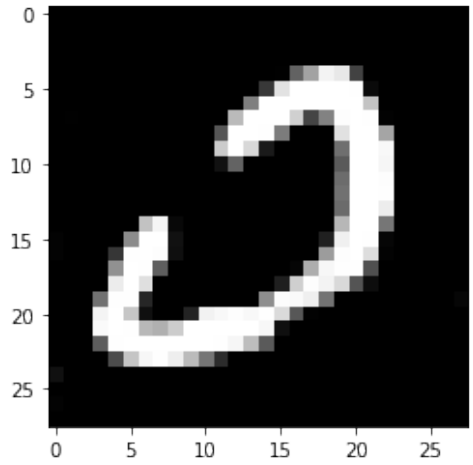
(c) Baseline - PIECE Method



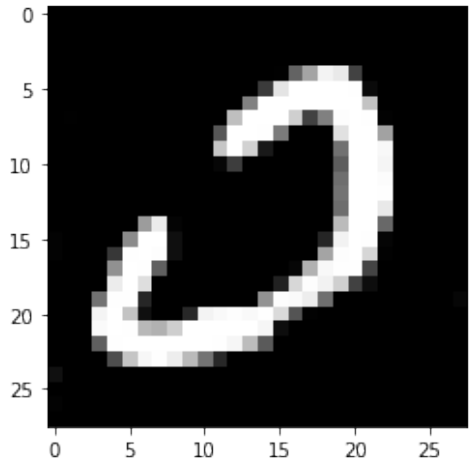
(d) Proposed Method

Figure 6.1: Results of Baseline methods and Proposed methods when going from original image of digit 0 to digit 6

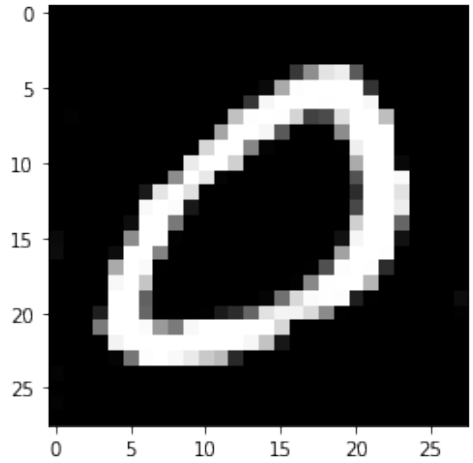
Looking at both the evaluation, it can be inferred that the PIECE methods is favourable in generating counterfactual images considering the unsure instances. These results are similar with the results of misclassified instances and that could be because of the nature of the instances which is being close to the boundary line. This evaluation strengthens our hypothesis that the proposed method is not better than the baselines when the instances of interest are close to boundary line.



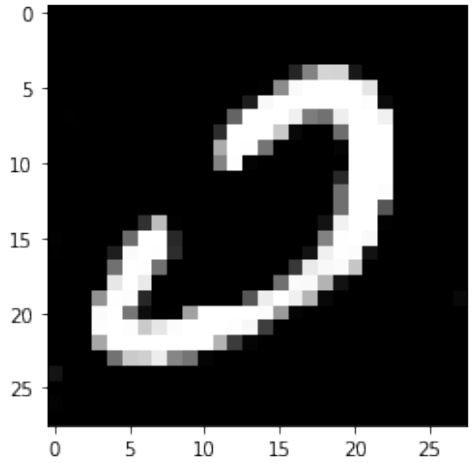
(a) Original Image



(b) Baseline - C-Min Edit Method



(c) Baseline - PIECE Method



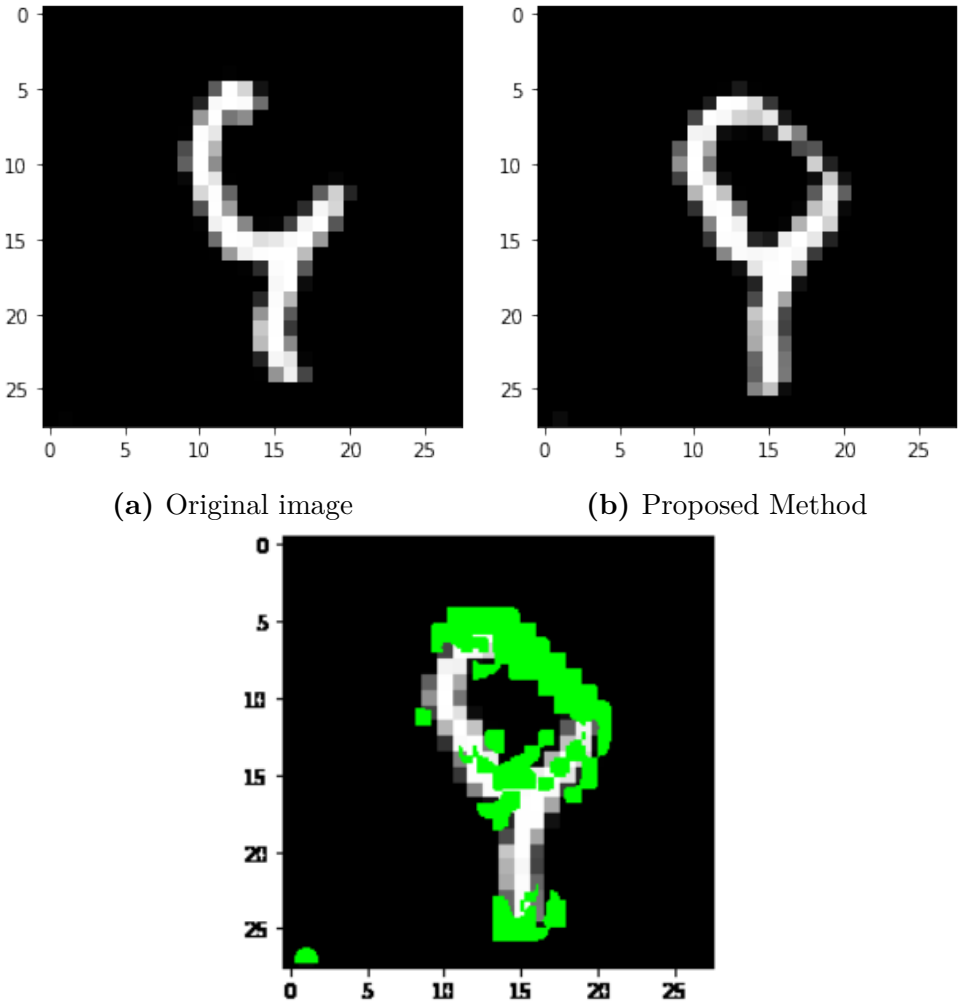
(d) Proposed Method

Figure 6.2: Results of Baseline methods and Proposed methods when going from original image of digit 2 to digit 0

Now to discuss about the capability of the method to generate counterfactual images given any desired class, we noticed that the proposed method and the baseline 1 C-Min Edit method are capable of generating counterfactauls images for any class specified by the user, while the PIECE method can generate a counterfactual image only for the class chosen by the eq. 3.3.2. Consider the case where the original image has prediction of "0" as shown in Fig 5.2, Fig 5.3 and Fig 5.4, we can see the results of baseline 1 and proposed method for the counterfactual when target class is between 1-9, the baseline 2 PIECE fails to generate images for all these classes. So this points out the merit of the proposed method in comparison to PIECE.

Further as stated in the motivation of this thesis, we also aim to use the plausible counterfactual image generation to boost trust of the user in the classifier. The research question **[RQ-2]** was framed to explore the various visualisation techniques that can be used to attain the transformation map that explains the change from the original image to counterfactual image. In section 5.1.5 we visualise the optimisation steps to understand the changes in the original image step by step, these images of the transformation journey

can assist the users in understand what regions of the images were altered at what point of time. To attain a single image that shows the transformation, we exploit the pixel difference map technique and highlight the contours that experienced changes. These pixel difference maps are simple to interpret and can add value to the user. These transformation maps are discussed and shown in section 5.1.6. Consider the user provides the image shown in Fig 6.3(a) which has prediction of digit "4" to the proposed method and wants to understand how this image can be converted to an image that can be classified as digit "9" by the classifier. The user provided the given image and the counterfactual class as "9" to the proposed method. The generated counterfactual image is shown in Fig 6.3(b). We use the pixel difference technique to visualise the transformation. The green pixels in the Fig Fig 6.3(c) shows the region that has undergone changes to transform the original image to digit "9". The user of the proposed method would agree with the transformation needed as the user would have did the same transformation if they were asked to change the digit from "4" to "9". This encourages the user to imbibe their trust in the digit classifier as it behaves in their line of expectation and not only classifies the images, but the counterfactual also provide explanation to the user on what needs to change in the digit to attain desired prediction. At last the final motivation of the thesis is to provide plausible counterfactual images, as in the above case if the counterfactual method would have generated an image that has prediction of "9" but does not actually look like a digit from true distribution, it would have led to swindled trust in the classifier. This shows that the counterfactual image generation alone is not enough in building trust of the users, it is critical to make sure the generated counterfactual images are plausible.



(a) Original image **(b)** Proposed Method

(c) Difference Image

Figure 6.3: Difference Image

6.2 BraTS Dataset

The BraTS dataset [30, 31, 32] is widely used for brain tumour segmentation challenges and studies across various conferences like MICCAI. BraTS dataset has become a standard for comparing the performance of various image segmentation techniques in biomedical applications [36, 37]. In our analysis we classify BraTS dataset as complex one because it consists of complex structures of the brain and tumour sections. After our analysis on MNIST, we wanted to check the generalisation power of the proposed method in cases of complex datasets.

The prerequisite for the proposed method is the GAN training over the training dataset, additionally the images generated from GAN should also be closer to the true distribution and have high quality so that generation of plausible counterfactual can be done. We started with the DC-GAN training, upon fully training this GAN we realised the output to be blur and missing the complex brain structures. This motivated us to look for other powerful GANs that can generate high quality images, we used the WGAN that is known for getting rid of the vanishing gradient problem faced by DC-GAN. Upon completing

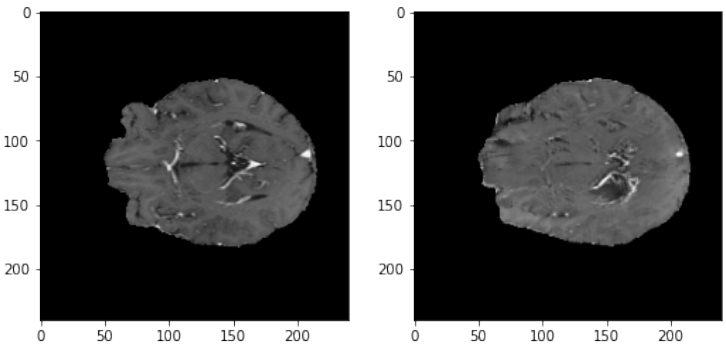
the training of WGAN, we realised the generated images were better than DC-GAN but still they lacked exact structures of the brain. Further, we used the gradient penalty version of the WGAN to attain boost in performance, but the results of it were found to be not satisfying the desired criteria. At this point, we hypothesised that the GAN architectures are unable to map the 240x240 image from a latent vector of size 128, so we used latent size of 2000 and tried DC-GAN and WGAN. This resulted in minor improvement in the quality of the images. Further we hypothesised that the generator network in GAN should be made stronger to attain high quality images. This motivated us to use the U-Net architecture as the generator, as U-Net is widely used in biomedical image segmentation because of its capability to produce outstanding results with the means of skipped connections. The latent vector now changed to 240x240, and it became analogous to image to image translation but we had a noise of 240x240 which will be mapped to the brain image using the stronger network like U-Net. Using the U-Net along with the WGAN concepts, we achieved results that were outstanding in comparison to the pre existing GANs. Further, we experimented with even stronger networks in the generator architecture like the ReconResNet by Chatterjee et. al [28] and couldn't achieve results better than the U-Net WGAN. We evaluated all the GANs and tabulated the scores in table **??**, the quantitaive evaluation points that the DC-GAN is better in comparison to other GANs trained. Further, we perform the qualitative evaluation on these GANs as shown in Fig 5.13 and it is clearly evident that the U-Net WGAN outperforms the other GANs. This is why, we decided to use the U-Net WGAN for further experimentation and evaluation.

For the detailed evaluation, we divided the test set into 2 parts, correct classified and misclassified instances. This will enable us to get an overview of how the performance of the proposed method is affected by the nature of the instances. We perform the quantitative evaluation by calculating the evaluation metrics discussed in Section 3.4 on the test set using the baseline and the proposed method. The results are tabulated in the table 5.4, as the table points out there is no significant improvement realised in the performance of the proposed method in comparison to the baseline. Further we visualise the results to attain qualitative evaluation and see that the results obtained by the baseline and proposed method look similar. The results do show significant difference with the original image to realise that we have achieved the counterfactual. In a few cases as shown in Fig 5.15, we can notice that the black region in the original image that made the prediction as "tumour detected" is completely vanished in the counterfactual image. Again, we perform similar evaluation on the misclassified instances and tabulate the results in the table 5.2, as seen in this table there is no significant difference between the two methods. To answer **[RQ-1]**, in cases of complex dataset we see that the proposed method is on par with the baseline and do not offer any advantage irrespective of the nature of the instances.

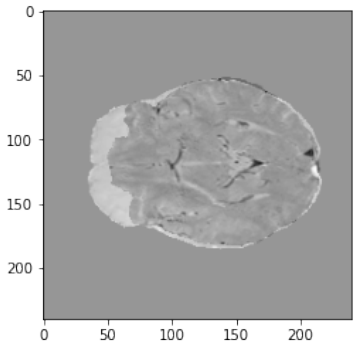
The failure of the proposed method in complex datasets can be attributed to the quality of the images generated, the U-Net WGAN generates better quality images in comparison to the other GANs trained, yet it does not produce realistic images that are closer to the true distribution. To confirm our hypothesis over the quality of the images, we generated overall 10 images each for the label with and without tumour using the U-Net WGAN, this resulted in a test set of 20 images. These images were provided to the MRI experts to attain two kinds of labels, first is the image realistic and secondly which of these images have tumours. The labels "tumour detected" and "no tumour detected" were masked and

then provided to three MRI experts. The first reviewer labelled two images to be realistic and was able to detect 1 image having tumour and provided further comments stating that most of the images do not seem to be brain MRIs as they lack precise complex structures. The second reviewer was able to detect 7 images with tumour and felt that the images require enhanced clarity to be classified as real brain images. The third reviewer was able to label only 3 images as real and discarded the remaining images, the reviewer had further comments stating that the in some images the deformation of the brain is due to a large tumor, but not really a tumor structure. In general the tumor should appear more "roundish" and have a certain distinct boundary to the tissue. All these reviews point in the direction of incapability of the GAN to generate high quality images which further limits the performance of the proposed method.

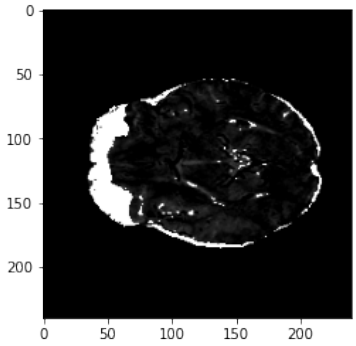
As mentioned in the motivation of this thesis, we intent to explore visualisation techniques to provide the user with the transformation maps or pixel difference maps to provide an idea on what changed while going from original image to counterfactual image.To answer [**RQ-2**] we intended to use the image registration techniques to carve out the transformation lattice, upon trying the image registration techniques, it failed mostly because the original image and the counterfactual image had major changes in the image regions. The counterfactual image had blurs in various regions because of the limitations of U-Net WGAN and this impeded the image registration techniques to derive a transformation lattice. As discussed in section 5.2.6, we utilise the pixel difference maps to visualise the changes, along with these we use the manual thersholding and Otsu's thresholding technique to showcase results as shown in Fig 5.18 and Fig 5.19. The threshold images show the pixel regions that have experienced maximum changes. Similarly, we consider a case as image shown in Fig 6.4 (a), the original image has the prediction "tumour detected", the counterfactual class "no tumour detected" is provided to the proposed method, the generated counterfactual is shown in Fig 6.4 (b), notice that in this case the counterfactual method attempts to complete the brain structure as seen in Fig 6.4 (c) (d) (e). This example was considered to showcase the case where the brain structure itself is altered by the proposed method rather than any modification inside the brain image.



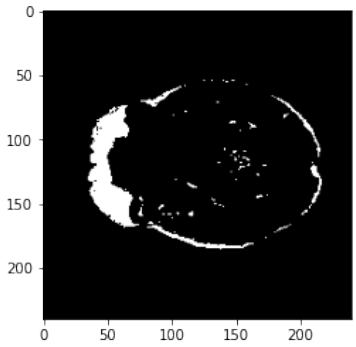
(a) Original image **(b)** Proposed method



(c) Difference mask



(d) Difference mask with manual threshold

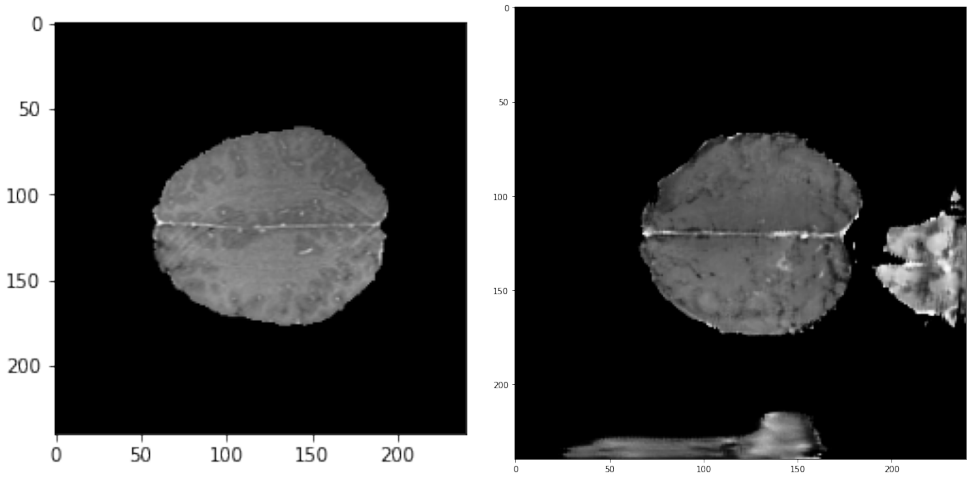


(e) Difference mask with Otsu's threshold

Figure 6.4: Results of proposed method over an image with prediction "tumour detected"

Upon analysing the performance of the proposed method through qualitative evaluation,

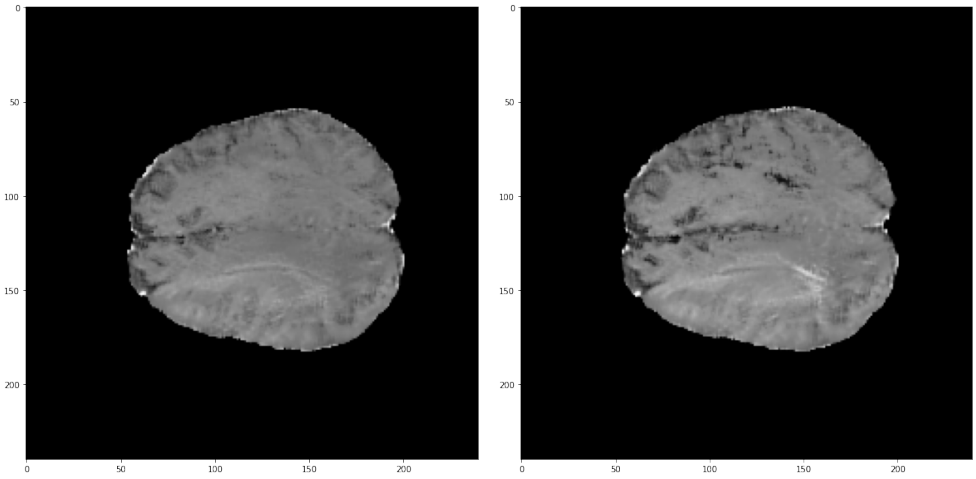
we stumbled upon some uses cases that helps us to comment over the limitations of the proposed method. Consider the case shown in Fig 6.5, the original image was classified by the classifier as "no tumour detected", upon passing this image to the proposed method, it generates some irregular regions outside the brain area and then acquires the label "tumour detected". Looking at these results it appears to be a out of distribution image, which is why it cant be called a plausible counterfactual image. Instances like these points out the limitations of the proposed method and it hints towards using stronger discriminator network while training the GANs as that might reduce the errors arising like these.



(a) Original image **(b)** Proposed Method

Figure 6.5: Counterfactual image generated(right) for the given image with prediction "no tumour detected"(left)

Next we consider the image whose prediction is "no tumour detected" as shown in Fig 6.6 (a), upon passing this image to the proposed method, it generates the counterfactual image shown in Fig 6.6 (b). The counterfactual image adds a few black areas or holes that makes the classifier switch its prediction from "no tumour detected" to "tumour detected". These regions do not have a boundary as required by tumour sections, but based on the GAN training, the proposed methods makes minimum changes to the image that is darkening some pixels and it makes the model switch its prediction. These kind of instances can also be classified as adversarial images, but these images have black region not because of random blurring but because these kind of black regions are seen by the GAN in the training images.



(a) Original image **(b)** Proposed Method

Figure 6.6: Counterfactual image generated (right) for the given image with prediction "no tumour detected"(left)

7 Conclusion

7.1 Summary

The proposed method for plausible counterfactual generation was evaluated on two different datasets, one simple and other complex. The results of the method were over the simple datasets, further the performance of the method is better than the baselines when the instances are away from the boundary line. For understanding the transformation of the original image to counterfactual image we explored the pixel difference method and highlighted the contours in the image that experienced maximum change. Further, we evaluated the performance of the method on the complex dataset and found out that it does not promise significant improvement over the baseline methods. The reason for this can be attributed to the inefficiency of the GAN used in the method, this motivated us to explore various GANs that can produce high quality images and in process of this exploration we provide the novel GAN architecture U-Net WGAN and ReconResNet WGAN that provides high quality images in comparison to the pre existing GANs. We intended to use image registration techniques to derive the transformation lattice, but the inefficiency of the GANs to generate high quality images restricted the usage of these techniques. This motivated us to use simple transformation techniques like pixel difference mask and then further thresholding it with manual and Otsu's method.

7.2 Limitations

In this section, we discuss the limitations observed while conducting experiments for the thesis.

The performance of the proposed method for plausible counterfactual image generation depends heavily on the nature of the dataset, as in the performance improves in case when we consider dataset with images that have simple shapes and structures. This is why we term the performance of the proposed method as dataset dependant. Additionally the performance of the proposed method also banks on the nature of the instances, the performance is promising when the instances are away from the decision boundary. In cases where the instances are closer to decision boundary, the proposed method fails to generate plausible images. The users also need to tune the plausibility threshold to cater to their desired level of plausibility and this can be subjective call. One of the major limitation is the dependence on the performance of the GAN required for the proposed method, the GAN training is often expoensive. In dealing with simple datasets, majority of GANs provide high quality images and they can used to generate counterfactuals. But, considering the complex datasets, it becomes difficult to choose a GAN architecture that can generate high quality image. In our case of complex dataset, most of the pre existing

GANs failed to generated satisfying results, we proposed the U-Net WGAN that promises better image quality in comparison to pre existing GANs. Yet the quality of the images by U-Net WGAN was not closer to the true distribution and had some blurring regions in the generated image. This limitation on the quality of the images harnessed the performance of the proposed method and also restricted us to try out image registration methods to attain transformation maps.

7.3 Future Work

In this section, we discuss some of the avenues that can be explored for further enhancement in terms of plausibility.

- Use advanced Convolutional Neutral Network architectures for the Generator and Discriminator network in GAN.
- Explore the usage of the Discriminator as a plausibility regulator in the counterfactual image generation for RGB dataset
- Explore the usage of the Discriminator as a plausibility regulator in cases of diverse counterfactual generation
- Explore the usage of the Discriminator as a plausibility regulator in the field of Semi-factual

Bibliography

[1] M. T. Keane and E. M. Kenny, "On generating plausible counterfactual and semi-factual explanations for deep learning," 2021.

[2] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," pp. 3126–3132, Association for Computing Machinery, Inc, 4 2020.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 5 2015.

[4] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," vol. 2018-January, pp. 1–6, Institute of Electrical and Electronics Engineers Inc., 3 2018.

[5] S. Misra, S. Thakur, M. Ghosh, and S. K. Saha, "An autoencoder based model for detecting fraudulent credit card transaction," vol. 167, pp. 254–262, Elsevier B.V., 2020.

[6] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," 1 2021.

[7] C. Molnar, "Interpretable machine learning a guide for making black box models explainable."

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," vol. 13-17-August-2016, pp. 1135–1144, Association for Computing Machinery, 8 2016.

[9] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A unified approach to interpreting model predictions."

[10] J. Pearl, "Causal and counterfactual inference," 2019.

[11] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," 7 2019.

[12] T. Vermeire and D. Martens, "Explainable image classification with evidence counterfactual," 4 2020.

[13] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," 4 2019.

[14] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," *AI Magazine*, vol. 38, pp. 50–57, 9 2017.

[15] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, pp. 76–99, 06 2017.

[16] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness."

[17] N. Madaan, I. Padhi, N. Panwar, and D. Saha, "Generate your counterfactuals: Towards controlled counterfactual generation for text," 12 2020.

[18] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," 7 2018.

[19] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *CoRR*, vol. abs/1702.04595, 2017.

[20] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images."

[21] A. Barredo-Arrieta and J. D. Ser, "Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples," 3 2020.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[23] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies," *Artificial Intelligence*, vol. 294, 5 2021.

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017.

[25] Y. Benny, T. Galanti, S. Benaim, and L. Wolf, "Evaluation metrics for conditional image generation," *International Journal of Computer Vision*, vol. 129, pp. 1712–1731, 5 2021.

[26] N.-T. Tran, T.-A. Bui, and N.-M. Cheung, "Improving gan with neighbors embedding and gradient matching."

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[28] S. Chatterjee, M. Breitkopf, C. Sarasaen, H. Yassin, G. Rose, A. Nürnberger, and O. Speck, "ReconResNet: Regularised residual learning for MR image reconstruction of undersampled cartesian and radial data," *Computers in Biology and Medicine*, vol. 143, p. 105321, apr 2022.

[29] S. S. Kadam, A. C. Adamuthe, and A. B. Patil, "Cnn model for image classification on mnist and fashion-mnist dataset," *Journal of scientific research*, vol. 64, pp. 374–384, 2020.

[30] B. S. K.-C. J. F. K. K. J. B. Y. P. N. S. J. W. R. L. L. G. E. W. M. A. T. A. B. A. N. B. P. C. D. C. N. C. J. C. A. D. T. D. H. D. D. C. D. M. D. S. F. J. F. F. G. E. G. B. G. P. G. X. H. A. I. K. J. R. J. N. K. E. L. D. M. J. M. R. P. S. P. D. P. S. R. T. R. S. R. M. S. D. S. L. S. H. S. J. S. C. S. N. S. N. S. G. T. T. T. O. T. N. U. G. V. F. W. M. Y. D. Z. L. Z. B. Z. D. P. M. R. M. V. L. K. Menze BH, Jakab A, "The multimodal brain tumor image segmentation benchmark (brats).," *IEEE Trans Med Imaging. 2015 Oct;34(10):1993-2024.*, vol. 64, 2015.

[31] S. A. B. M. R. M. K. J. F. J. F. K. D. C. Bakas S, Akbari H, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Sci Data. 2017 Sep 5;4:170117*, vol. 64, 2017.

[32] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, M. Prastawa, E. Alberts, J. Lipková, J. B. Freymann, J. S. Kirby, M. Bilello, H. M. Fathallah-Shaykh, R. Wiest, J. Kirschke, B. Wiestler, R. R. Colen, A. Kotrotsou, P. LaMontagne, D. S. Marcus, M. Milchenko, A. Nazeri, M. Weber, A. Mahajan, U. Baid, D. Kwon, M. Agarwal, M. Alam, A. Albiol, A. Albiol, A. Varghese, T. A. Tuan, T. Arbel, A. Avery, P. B., S. Banerjee, T. Batchelder, K. N. Batmanghelich, E. Battistella, M. Bendszus, E. Benson, J. Bernal, G. Biros, M. Cabezas, S. Chandra, Y. Chang, and et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," *CoRR*, vol. abs/1811.02629, 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[36] M. A. Al-masni and D. H. Kim, "Cmm-net: Contextual multi-scale multi-level network for efficient biomedical image segmentation," *Scientific Reports*, vol. 11, 12 2021.

[37] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng, "mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," 6 2022.

[38] C. Nugent, D. Doyle, and P. Cunningham, "Gaining insight through case-based explanation," *Journal of Intelligent Information Systems*, vol. 32, pp. 267–295, 6 2009.

[39] C. R. C. of Computer Education in Colleges Universities, O. T. University, I. E. Society, I. of Electrical, and E. Engineers, *The 14th International Conference on Computer Science and Education (ICCSE 2019) : August 19 -21, Toronto, Canada.*

[40] F. Yang, N. Liu, M. Du, and X. Hu, "Generative counterfactuals for neural networks via attribute-informed perturbation," 1 2021.

[41] A. Sauer and A. Geiger, "Counterfactual generative networks," 1 2021.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.