**FOCUS:** **F**ully **O**ptimized **C**onvolutional **U**Net **S**egmentation for Tumor Detection

Douglas Ta  
*Department of Engineering Education*  
*AI Systems*

*Abstract*—This report was written to showcase the powerful capabilities of using UNET neural network architecture for medical segmentation. With this in mind, the purpose behind exploring the capabilities of UNET is to assist medical professionals with tumor detection. Doing so will ultimately benefit both medical practitioners and patients alike. CNN’s have, in the past, always shown powerful abilities in classification. FOCUS expands on this idea by utilizing the inherent convolutional nature of CNN’s while maintaining a simplistic approach that will be more readily applicable in the field. Some key results to note for this report are that although the performance metrics of the final modal do not reach what should be considered standard in the medical field, that does not mean that FOCUS is not capable of such results but rather more time should be allocated towards the investigation of UNET. Considering the potential of CNNs in the past, FOCUS will surely see more potential uses in the future.

*Keywords*—FOCUS, UNET, medical segmentation, CNN, AI Systems

# Introduction

## Context and Problem Statement

The standard procedure for tumor detection in the medical field is to rely on medical practitioners to carefully examine medical images such as CT scans and identify tumors on an image-to-image basis. This process takes a significant amount of time due to the issue of examining images one-by-one as well as having room for human error due to the small sizes tumors may start off as. This is an important problem to address because the lengthy expenditure of time is both problematic for the medical practitioners with many other time-consuming tasks on their plates as well as for the patients who will want to address any potential tumors as quickly as possible.

## Objective

The goal of FOCUS is to expedite the tumor detection process, enabling medical professionals to make quicker, more informed decisions and ultimately improve patient care. This tool is meant to serve as a diagnostic aid and not a total replacement to medical professionals. FOCUS aims to solve the prior problem by reducing the costly amount of time that is needed to sift through CT scans images. The expected outcome of this project is to help medical experts identify problems with their patients at a much faster rate leading to earlier intervention, better patient outcomes, and reduced healthcare costs.

## Scope and Contributions

The contributions of this paper are towards the advancement of utilizing neural networks such as UNET to assist medical professionals in the field. Although the key results of this paper have not yet met the threshold of what can be considered usable in the medical field, this paper will highlight the potential UNET can have in the field. The novelty that FOCUS brings is not just simply introducing the UNET architecture, but to also showcase how potential technology in the future will need to be maintained and monitored as an actual AI system.

## Report Organization

The organizational structure of this report will be discussed in the following order:

1. Related Works
2. System Design
3. Trustworthiness and Risk Management
4. Evaluation and Results
5. Discussion
6. Future Works and Improvements
7. Conclusions

Further details will be elaborated in each of these sections about how these specific topics relate to FOCUS.

# *Related Works*

Prior to the recent advent of neural networks, the traditional idea of medical segmentation took several different approaches to tackling the issue. These traditional methods included manual segmentation [4], semi-automatic segmentation [3], and automatic segmentation [1]. However, the overall problem with all of these traditional methods was that they relied heavily on prior medical knowledge from the user. These methods also had issues with generalizability as well making it not as efficient in the field [7]. That is why the introduction of CNNs [9] played a powerful role in launching renewed interest in developing more generalizable technology for medical segmentation [2,5,6,8]. CNNs are able to solve many of the problems that traditional medical segmentation had. Medical practitioners do not need to rely as much on their inherent domain of knowledge to create segmentations anymore while also being able to generalize to a much bigger population of people. Although CNNs were able to fix many of the problems traditional methods may have had, at this point there are still some issues that prevent this technology from being used in the field right away. The problem now is the lack of training data to train these deep learning models on as well as creating a lightweight model to perform fast inferencing. This is where UNET comes into play. The goal behind the development of UNET was to address both of these issues. UNET offers a solution to the problem of a lack of training data by utilizing its unique U-shaped architecture to improve generalization faster. It is also a relatively lightweight model itself compared to much bigger CNNs [10]. This UNET architecture is what FOCUS is basing itself off of. Since the goal behind FOCUS is to provide medical professionals with a lightweight system that expediates tumor detection, UNET is the perfect neural network architecture for this system. The difference between FOCUS and neural networks used in previous papers is that none of these tools have reached the level of actual production yet. What makes FOCUS different from other deep learning techniques, is that the focus behind FOCUS is to make a system that not only provides accurate segmentation, but to also make integration into hospitals/clinics a smoother process.

# System Design and Implementation

## System Overview

Before discussing the overall computing architecture of FOCUS, it should be stated that the training and inferencing for FOCUS was done using an A100 GPU. Although it may not necessarily have to be an A100 GPU, for future use of FOCUS a single GPU will be required in order to provide quick inferencing of tumor segmentation. The initial release of FOCUS does not utilize a GPU for inferencing due to limitations on deployment budget. It should also be noted that the amount of storage used throughout the entire process was up to 30GB of data. The majority of the data does not come from the model itself, but rather the size of the CT scans used for training and inferencing. For future FOCUS use, this will have to be a minimum requirement as well due to the large nature of CT scans. Now that the hardware requirements have been discussed, let’s dive deeper into the overall computing architecture of FOCUS.

Let’s go through the system architecture from start to finish. The user will input a CT scan into FOCUS. This CT scan will be preprocessed to fit the UNET model. The model will then perform convolutions onto the preprocessed image otherwise known as down-sampling until a convoluted image of original image is produced. This convoluted image is what will contain the edges used for segmentation. After down-sampling is done, a similar process known as up-sampling is performed on the convoluted image to actually create the segmentation itself. This segmented image is then output to the medical practitioner. That is a basic overview of how FOCUS works. For further visualization of what the UNET architecture looks like refer to Figure 1.

A diagram of a diagram

Description automatically generated

Figure

As for how FOCUS is hosted, this project utilizes cloud-based services such as Streamlit in combination with AWS S3 Buckets to make hosting and storage simple. These services are what allow users to access FOCUS using the Internet. Maintenance and monitoring will also be done by integrating Prometheus and Streamlit to measure the overall system performance and health over time. Further details about how FOCUS implements these services and libraries will be elaborated on in future sections.

## Data Collection and Preprocessing

The data for FOCUS is comprised of 27 CT scans of patient’s upper abdomens obtained from Gong Laboratories which is a UF affiliated research lab. There are no problems with the reliability of the source of data and all the data was thoroughly checked for correct annotations. The data was downloaded from a repository associated with the lab. As for ingestion there are two parts for this: training and deployment. For training, the images were loaded into batches in order to train the most amount of data as efficiently as possible. For deployment, the data is also being ingested in batches as well as ingested in real time in order to get the results for tumor segmentation to the practitioner as soon as possible. The type of data that FOCUS will be using is unstructured because this project will be using CT scans of patient’s bodies with potential tumors. It is important to note that personal patient data was removed prior to any training done. This is to protect the privacy rights of patients as well as prevent any future bias in the deep learning model.

Some challenges that come with handling this sort of data is the type of preprocessing that is necessary in order for the model to handle the data. For example, CT scans may differ in size from hospital to hospital depending on the type of technology each hospital is using. This is a problem that will need to be addressed because FOCUS will need to be able to be integrated into many places with different standards. Another challenge that comes with this data is that people have different body shapes. That means that FOCUS will need to be able to adapt and generalize tumor segmentation onto different people. This can be difficult to achieve due to the fact that there is not an inherently large number of CT scans to be used for training on due to patient privacy. This is different from how normal deep learning training is. Normally, CNNs are trained using at minimum thousands of images in order to improve model generalizability to unique images. That will not be the case for FOCUS or many medical AI systems due to the limitedness of data. That is why in order to tackle this problem, there will need to be a sufficient amount of data augmentations applied to the limited training data to boost performance and generalizability.

Data augmentations played a key role in this project. Especially towards supplementing the very limited training data. The following transformations were applied to the CT scans:

* Foreground Cropping
* Orientation
* Spacing
* Random Spatial Crop
* Random Flipping
* Random Rotation
* Random Intensity Shift

All of these transformations fall under either two purposes. Either to make the image dimensions more uniform or to increase the model’s generalizability to different CT scans. The first three fall under the first purpose of making the dimensions more uniform. The intuition behind creating more uniform images is to make pattern recognition for UNET better. Although humans are able to differentiate obvious patterns easily, it is not necessarily the case for neural networks. This is because different patterns emerge from differently sized images which is why it is necessary to ensure uniformity for the CT scans. To further elaborate, foreground cropping refers to cropping around the tumor area contained within the CT scan. This is necessary because the tumor itself is only a small portion of whole CT scan. The model would not be able to train effectively if not for foreground cropping. This transformation was actually custom built to crop until the bounding box of the tumor itself. For Orientation and Spacing, those transformations are there to make sure that the CT scan is orientated correctly as well as spaced equally in all dimensions. As for the remaining transformations, those are used to increase the generalizability of the model. Random flipping and rotation are self-explanatory, so there is no need to dive deeper into those, but to shed light on random spatial crop and intensity shift may be important. What is meant by random spatial crop is that parts of the already cropped images are sub-sectioned even further to give the model different types of images for training. For further visualization of how random spatial cropping works refer to Figure 2.

A close-up of a brain scan

Description automatically generated

Figure

As for random intensity shift, this refers to slightly tweaking the intensity of the entire image to make sure that the model is not over fitting on the training data.

## Model Development and Evaluation

Before diving into model development, let’s discuss the reasoning behind the choice of model once more. As mentioned in the Related Works section, the traditional methods of medical segmentation have two main issues that UNET solves. The first issue being the reliance of prior human knowledge in the medical field. It is relatively expensive in terms of time for medical practitioners to create the segmentations themselves. Although there are certain traditional methods that claim to “automatically” create segmentations, it still holds true that domain expert knowledge is required [1]. This is where UNET comes into play. Using a UNET neural network architecture, this allows medical practitioners to take a more hands-off approach. This in turn frees up the time normally used for medical segmentation to be directed towards other equally important tasks. The second issue with the traditional medical segmentation methods is the problem with generalizability [7]. Rather than hand-crafting a segmentation from scratch, using a UNET gives medical practitioners more flexibility in what they can input into FOCUS. That is why FOCUS utilizes UNET architecture. Another important point to discuss as well is the size of the model itself. Many convolutional neural networks can get fairly hefty in size. What makes UNET the most optimal choice for this project is because UNET uses less convolutional blocks compared to other convolutional neural networks. Although to counter this point, since UNET is more relatively lightweight compared to state-of-the-art convolutional neural networks, there is a theoretical tradeoff in performance due to the more simplistic nature of UNET. The reason why UNET was still chosen as the model to be used in this project is because smaller models generally mean faster inference times which is an important consideration to have when it comes to actual practice. The faster the inference time, the higher the throughput of patients that are able to receive medical attention. As for how the model was developed, the specific architecture used to create a UNET neural network was already developed by Ronneberger et al. [10]. Let’s dive into the specific details about the architecture itself. There are five principal components to UNET. The first component being the encoder which is comprised of 2D convolutional layers followed by 2D max pooling layers. This pattern is repeated four more times with each layer doubling the number of filters from each layer. This section of the model is what was referred to previously as down-sampling. For my UNET model specifically, this started with 16 filters and ended with 256 filters. The second component is the skip connections. These skip connections are used by the model to directly deliver feature maps from the encoder to the corresponding features maps in the decoder. This is partially the reason why this neural network is named UNET. It is because these skip connections are vital in letting the network understand the original structure of the image before down-sampling, making it a key component for up-sampling. The third component is the decoder. This section of the model is comprised of transposed convolution layers, concatenation blocks from the skip connections, and 2D convolutional layers. This pattern is repeated in a similar fashion to the encoder. This part of the process was previously referred to as up-sampling. The final component of the UNET is the bottleneck. This section of the model is the lowest layer between the encoder and decoder and contains the highest number of filters. This layer is just simply a 2D convolutional layer that connects the encoder and decoder. For further visualization of what UNET looks like refer to Figure 1.

In order to effectively evaluate the performance of FOCUS, the most optimal performance metric will be Mean Dice score. Mean Dice score is a similarity score between the predicted segmentation versus the objective ground truth segmentation provided by medical professionals. This will be a reliable performance metric to compare models during training as well as when maintaining the model’s overall health in the future. For Mean Dice, the scores range from 0 to 1. The closer the score is to 1 determines how similar the predicted segmentation is to the actual segmentation. For instance, a mean dice score of 1 would suggest that the predicted segmentation is identical to the ground truth segmentation. The opposite is true as well meaning a mean dice score of 0 would suggest that the predicted segmentation is nothing similar to the actual ground truth. The benefit of using Mean Dice over a more popular performance metric such as accuracy is that Mean Dice handles class imbalance well. Although the data used for training all contains tumors, in the real world there will be a heavy class imbalance between CT scans that have tumors and do not. This is why Mean Dice is the main metric that is used for FOCUS. To take this natural class imbalance into consideration. However, there are limitations to this metric as well. A limitation that only using Mean Dice could have in the future would be having a high true negative accuracy, but a low true positive accuracy which would make the system unusable. That is why in order to supplement Mean Dice the following metrics were used during inferencing: precision, recall, and F1-score. These are relevant performance metrics because they allow predicted segmentations to be compared in different nuanced light compared to just using Mean Dice. Recall refers to the proportion of actual positive cases that the model correctly identified compared to all of the true positives. Having a high recall would mean that the model correctly identifies CT scans with tumors most of the time. Precision refers to the actual number of positive predictions that were correct in total. This means that having a high precision would mean that mostly all of the ground truth segmented parts were correctly identified as positive by the system. The tradeoff here is that if FOCUS has a high recall, this in turn means that the false positive accuracy would go up as well. It is similar to high precision as well. Usually with higher precision there is a tradeoff of recall. This is worthwhile tradeoff however because it is more important for doctors to identify potential tumors in patients rather than miss any potential tumors. Since this is the case, F1-score is the perfect metric to utilize because it takes into account the harmony of both precision and recall, giving a score that better represents the system’s performance as a whole.

## Deployment Strategy

The main cloud service that was used to deploy FOCUS was Streamlit. What makes Streamlit such a powerful tool to use is the fact that it already has a dedicated interface for hosting Streamlit applications. This means that simply connecting this interface to my GitHub repository allowed for my application to be deployed. Another added benefit to using Streamlit as a host is that secrets are easily implemented through a protected input. This is an essential feature that my application needs considering that FOCUS makes many connections to the S3 bucket located in AWS. Elaborating further on about the S3 bucket created for FOCUS, this bucket stores all of the data and models that will be created by FOCUS. Using the access and secret keys from AWS S3, FOCUS is able to connect to this S3 bucket for both reading and writing files. This is an essential feature for deployment since files cannot be stored locally for two reasons. The first reason being that users may not be able to store the massive amount of data that make up the CT scans. The second reason being that these CT scans are private information, regardless of the fact that personal data was wiped from them. Connecting to this S3 bucket makes deployment, with our goals in mind, much more achievable. The downside to connecting to an external database is that for inferencing this would take up a substantial amount of time to load in everything back and forth. There are several instances within FOCUS that have to utilize either file upload or file download which makes inference time much longer. However, this is a worthwhile tradeoff because connecting FOCUS to this external database is what allows FOCUS to be deployed in the first place.

To go into further detail about the deployment’s environment, a virtual environment is created to store all of the necessary dependencies for FOCUS. This environment then pip installs the requirement.txt file located within the repository into the virtual environment automatically which makes initializing deployment easier. The final reason as to what makes Streamlit a powerful cloud service to utilize for deployment is its integrated continuous deployment pipeline. A continuous deployment pipeline specifically refers to being able to when changes to the GitHub repository immediately reflect to production. This makes development faster and more efficient seeing as there is no middleman to prevent multiple changes to the codebase in the same session. However, the downside of implementing a continuous deployment pipeline is that if there are errors that do not necessarily appear in local, but might appear in deployment. This would be a serious problem. However, with that in mind, this problem is trivial considering that there can be multiple branches included within a GitHub repository. Meaning if changes in production fail, it is easy to rebase main back onto a solid production branch. For a better visualization of what the FOCUS interface appears like, refer to Figure 3 down below. Also to access the deployed FOCUS type <https://douglas-focus.streamlit.app/> into your browser to access FOCUS.

A screenshot of a computer error message

Description automatically generated

Figure

## HCI Considerations

FOCUS is only really intended for use by doctors and other affiliated medical staff. The interface that potential users will interact with is designed to be simple and easy to understand, making it easier for medical practitioners to integrate into their workflow. With this in mind, there are only a few components that users can actually interact with FOCUS. The first component being the input for CT scans. This part of the interface will allow users to upload any of their CT scans into FOCUS. The second component is the evaluation metrics, both being as discussed previously Mean Dice and recall. These metrics will give the user an idea of how confident they can be in the results of the system. This is also why FOCUS is intended to be used as an aid for medical practitioners and not completely replace them. Depending on the performance of the model, the medical practitioners themselves will be able to deem the segmentation created by FOCUS as accurate enough to give to the patient. In the case of a failed segmentation, there is always the medical practitioner to rely on to produce a segmentation themselves in the event of a failure. The third component of the interface will be the output of the system. Here is where the segmentation created by FOCUS will be viewable by the user. For initial deployment, the masks created by the model are shown through a gif in order to fully visualize the magnitude of what the model is predicting to be a tumor. Displaying such information as a gif is what allows the user to see 2D information in a 3D capacity. As for the last component of the interface, there will be a button that will link the user to a feedback survey. This survey is comprised of a Google questionnaire that will ask the user for feedback. The first question will ask if the user agrees or disagrees that FOCUS is a viable product for medical practitioners to use for medical segmentation. The second question will be a text prompt that will allow the user to provide any written comments they may have about the system. This survey will be anonymous and not ask any questions that pertain to breaching user privacy. The answers contents of the survey will also be stored within a Google Drive. Depending on the feedback attained, new features and functionalities can be added at a later time.

# Trustworthiness and Risk Management

## Strategies at Each Stage

There are 3 areas of interest that this section will discuss. Those areas of interest are security, privacy, and ethical compliance. Considering that the data for FOCUS deals with the health of patients, this is a serious consideration to take into account. The reason is that there are many privacy laws that pertain to one’s health which is why this section goes over how the user data is protected and used fairly. The foremost goal after delivering an accurate and expediated tumor segmentation of a patient’s CT scan is to protect confidential user data as well as ensure the model is fair for all users so that users may trust their data with our system. For the initial implementation of FOCUS, user data is being stored inside of an AWS S3 bucket. To access such data, a user would need both the access key as well as the secret key associated with obtaining the right to read and write to that bucket. This ensures that the patient’s data is password protected as well as duo-authentication protected as well. This should suffice for protecting both the security and privacy of any patient’s data. As for new patient data such as the CT scans that doctors would input into FOCUS, those images are secured in the same way. These security implementations should suffice to ensure that patient data is safe and private. As for making sure that the data used is ethically compliant, FOCUS makes sure to adhere to the following standards provided by the International Journal of Cardiovascular Imaging [11]:

1. Imaging is performed by a properly qualified physician.
2. Using protocols that maximize diagnostic yield but minimize patient risk.
3. Primarily benefits the patient and not for the physician or any others.
4. The context of the individual patient’s overall condition and needs for medical diagnostics are taken into account of.

With these standards in mind, FOCUS makes sure to follow all of the listed standards above. It is vital to ensure user privacy and security especially when it comes to handling sensitive information such as CT scans.

## Risk Management Framework

One key risk consideration that needs to be taken into consideration are certain liability risks that naturally come up when medical practitioners rely on technology. For example, if FOCUS had misidentified a tumor cell as an ordinary part of the body, this would potentially make the proprietor of FOCUS and the medical practitioners liable for missing something that is clearly harmful to the patient. The strategy to avoid situations like these from happening is for the model to be trained to skewer towards improving recall which would make missing positive observations any tumors as small as possible. Another important risk that needs to be addressed for FOCUS is the issue of data privacy. FOCUS will need to be trained on real data from patients with tumors. Because this is the case, it is of most extreme importance that no patient’s personal information should be linked to the model that could potentially tie back to the patient. To prevent this from happening the dataset used for training has had the personal information wiped off already. On top of that there is a risk of overfitting the model to the training data due to the relatively small sample size of CT scans available. The plan to mitigate this risk will be to apply data augmentation techniques to the existing training dataset such as random spatial crop to make the data more generalizable. During training, an image from the validation set will be pulled and used to validate the results of each epoch. This is done so that the model can increase its generalizability. Furthermore, there is the risk of integration issues with current medical practitioner workflows. New technology such as this one may be difficult for some practitioners to use at first which is why to mitigate this risk, the plan in the future will be to interview with practitioners to inquire about the exact process in which they go through in order to determine if a patient has a tumor using CT scans. Hopefully, these interviews will glean whether FOCUS will be able to easily integrate into current workflows. This is also why the feedback button was implemented as well. Medical practitioners may anonymously comment on FOCUS giving important feedback towards improving the system. As for the final risk that needs to be taken into consideration is that similar to many AI systems, FOCUS is susceptible to a risk of performance degradation over time. To mitigate this from affecting practitioners and patients alike, monthly inspections of the model’s performance can be done onto the system in order to measure if the effectiveness of FOCUS is still strong or may need maintenance. Further details on monitoring and maintenance will be provided in later sections. Referring to Figure 4, a residual risk assessment has been performed on all of the prior potential risks that FOCUS may have. Looking at this table, it can be seen that the most prevalent risks associated with FOCUS are the liability risk, data privacy, and overfitting on training data. Due to the nature of these risks, these were made the priority of what to address during this process. That is not to say that monitoring for performance degradation over time and smooth integration into workflow were not focused on.

A diagram of impact and impact

Description automatically generated with medium confidence

Figure

# Evaluations and Results

## Performance Metrics

A graph with lines and numbers

Description automatically generated

Figure

A graph with lines and numbers

Description automatically generated

Figure

Training and inferencing metrics were slightly different from each other. For training, only mean dice and loss were used to measure the performance of the system while for inferencing there was mean dice, recall, precision, and F1-score. Refer to Figures 5 and 6 to understand the performance of the UNET during training. Looking at the graphs, it would appear that although there appears to be a slight trend downwards in loss as well as a slight trend upwards in mean dice, the model’s performance appears to stagnate after a while. This can conclude that the model did not perform especially well. The results from inferencing will tell a better story about the model. Now referring to inference metrics, mean dice is a good metric to use because it compares the similarities between two different images and handles class imbalance well. Recall and precision are good metrics to have because they relay information if the model is overfitting on the data. F1-score is used because it represents a harmonic balance between precision and recall. Sometimes it is difficult to measure how the overall system is performing using recall and precision, which is why F1-score was implemented. Using F1-score helps to understand the overall health of the system. Determining an acceptable standard for all of these metrics is vital to setting FOCUS out into production. For this reason, the goal for FOCUS was to achieve a 0.9 score or above across the board on all evaluation metrics. This unfortunately did not become the case as can be seen from Figure 7.

A number and text on a white background

Description automatically generated

Figure

Another important metric that is of importance is taking into account the time it took from start to finish. It would appear that inference time for 2 CT scans was approximately 1.75 mins an image. This is not a bad result considering that segmentation can be done without doctor supervision. To further add on to this, the output of the predicted segmentation was compared to the ground truth segmentation masks to get a better understanding of how the model is predicting these masks. Multiple gifs were ordered side by side with their respective ground truths to make this visualization much easier to review. For a visualization of this refer to Figure 8.

A screenshot of a computer screen

Description automatically generated

Figure

Looking at the results of the predicted masks, it can be seen that FOCUS is over predicting the positive labels for tumors within a CT scan. This could be for a number of reasons such as not enough training data. Model training will have to be further looked upon in the future for FOCUS because as of right now FOCUS is not acceptable by medical standards and does not meet the criteria set in the beginning of this section.

## Monitoring and Feedback

In order to monitor the system’s overall health and performance, Prometheus is used to keep track of the evaluation metric. These statistics are able to be tracked over time by connecting to the 8000 port through a local host. It is mentioned briefly why these metrics are used as evaluation metrics not just for training but for monitoring as well in prior sections, but this section will go into further detail on the matter. Mean Dice is a similarity score that measures the similarity of the predicted segmentation to the actual ground truth. Recall is a score that measures the number of correctly predicted positive pixels that were included in the predicted segmentation compared to the total number of pixels within the ground truth. Precision refers to the total number of correct positive pixels. F1-score is a combination of that. Although in training, the ground truth is always provided, inferencing in the field will not produce a ground truth. However, these metrics may still be used for monitoring. FOCUS plans to set up monthly periodic intervals of inspection in order to verify the health of the system. During these monthly intervals, medical professionals will manually segment CT scans that FOCUS has stored but not necessarily trained on yet to be used as the ground truth. This fixes the limitation of not having a ground truth to use. Another aspect of the monthly periodic inspections is that FOCUS will sample some of the CT scans that were newly introduced during this time period to be retrained on the UNET. This will require medical practitioners to manually segment the CT scans chosen to be used for retraining to provide a ground truth label for these images. The benefit of doing this every periodic inspection is that FOCUS will slowly grow into a model that will be able to generalize to different types of CT scans. Doing this retraining will only make the system more efficient and generalizable to different patients. On the other hand, doing this work of having medical practitioners segment CT scans will add on an extra workload. The benefits of retraining far outweigh the costs of monthly manual segmentation in this instance. Following the standards set out in the beginning of this paper, AI systems that take into account of imaging must be done to benefit the patient as well as try to minimize their risk. Making FOCUS more generalizable will only work to benefit patients. In consideration of a monthly inspection that does not meet the required evaluation metric standards, the system will need to be taken offline and retrained from scratch.

FOCUS plans to take into consideration the feedback collected from the survey that users may fill out. This survey will be accessible from the FOCUS interface through a button that will link the user to a Google survey. This Google survey will ask the user two questions. Whether they agree or disagree that FOCUS is a helpful tool for medical practitioners as well as to provide any written feedback that may be submitted to the FOCUS team for review. Although feedback integration is not automatic, this is actually a benefit to the system. Automatic feedback integration can at times be harmful for AI systems. Especially if user feedback is filled with malicious intent. With a system such as FOCUS that prioritizes user data security and privacy, it is important to not incorporate features and functionalities that may put those values at risk. That is why any feedback from users will be manually sorted through and discussed within the FOCUS team to decide if user feedback should be incorporated into newer versions of FOCUS. For further visualization of what the survey looks like refer to Figure 9.

A screenshot of a survey

Description automatically generated

Figure

## Real-World Testing

Having deployed FOCUS, it can be seen that the Streamlit interface does well to keep the UI simple for user interaction. The feedback button is also useful at taking any written feedback from users that may have questions or comments about FOCUS. Some other observations are that because FOCUS is so easy to use, some of the risks that were assessed in prior sections have been mitigated such as the worry that FOCUS would not be easily integrated into medical practitioners’ workflows. Another observation that was made during deployment is how accessible FOCUS is. As long as there is internet connection, medical practitioners should be able to provide service to their patients. However, some downsides to FOCUS that has been seen is that inference time appears to be taking much longer than anticipated. This is most likely due to the fact that FOCUS has to connect to an external database to access all its data.

# Discussion

Let’s discuss the strengths and weaknesses that FOCUS has to offer. One of the key strengths of FOCUS is that it is able to reliably be accessed from different places as long as internet access is available. This is a strong strength to have because this makes FOCUS easily accessible to different hospitals that might use FOCUS. Not only does this make the system more accessible, but it also helps mitigate the risk of not being able to integrate FOCUS into a normal practitioner’s workflow. Another key strength that FOCUS provides is the easy-to-use interface. Although this may seem like a trivial implementation, time and time again has shown that complex interfaces take longer for users to learn. An easy interface fulfills two objectives. The first objective being easier integration into workflow. The second objective being more difficult to attack. The only input into FOCUS are CT scans. There are no direct text inputs into FOCUS which can be commonly used to attack a system. This simplicity helps achieve both of these objectives. Another strength that FOCUS has is that it is lightweight. The bigger the model generally means the longer the inference time. Since UNET is a lightweight model, inference time is relatively short compared to current state-of-the-art models. This is a big advantage to have considering that the shorter the inference time, the higher the throughput of patients that are able to get their medical scans segmented. One last key strength for FOCUS is its monitoring system. The major problem with AI systems that many companies face is the issue of performance degradation over time. This can be due to the model’s initial training data not being large enough to generalize properly or it could be due to a shifting real-world pattern. For this case specifically, the former is more applicable for FOCUS than the latter. This is because CT scans do not really change in the way that shifting market expectations might. If there are problems with the system, it will most likely be due to the fact that the system did not train on enough training data to be able to segment for certain CT scans with tumors. This is why FOCUS implements a monitoring system utilizing Prometheus and Streamlit. The benefits of having this monitoring system in place doing monthly monitoring intervals is that the system will be able to gradually improve its ability to generalize over time. This turns the risk of performance degradation on its head and makes it a key strength for FOCUS. Now taking a look at the other side of the coin, there are just as many weaknesses to FOCUS as there are strengths. One such weakness is its current segmentation capabilities. As of right now, FOCUS does not meet the evaluation criteria stated in the Performance Metrics section of the paper to be released into the field. As stated previously, the results of the predicted segmentations created by the UNET seem to be vastly overfitting on the training data which is most likely to be directly caused by the lack of training data to begin with. This is a common issue that is prevalent in many deep learning models related to the medical field. Another weakness that is FOCUS is limited to the CPU for inferencing. This is due to the fact that deployment costs did not allow for a single GPU to be used for deployment. This is another reason as to why the model architecture had to be lightweight. In order to balance the cost of having longer inference time with increased model complexity, keeping the model as lightweight as possible was important. Furthermore, FOCUS has only been able to be trained on CT scans. This system does not take into account the various types of medical images that doctors use to deduce if a patient has a malignant tumor. This handicaps FOCUS to a narrow range of uses until further functionality can be implemented into the system.

There were many challenges that stood in the way of the total completion of FOCUS. As stated previously in the limitations, FOCUS does not meet the standard criteria set forth in the Performance Metrics section to be able to be released into production. This was a major obstacle for obvious reasons. Without an accurate model, there is no way for medical professionals to utilize FOCUS and be able to confidently show their patients where the tumor is located in their bodies if there was one. This challenge was tackled in a variety of ways such as applying certain data augmentation techniques to the training data. As mentioned in the Data Collection and Preprocessing section, random flipping, random rotation, random spatial cropping, and random intensity shifting were all techniques that were used to make the model generalize to different images better. Although data augmentation improved results, FOCUS still did not meet the standard criteria to be used in the field. Another solution that was attempted was to add more training data. This also improved the results obtained from the UNET, but again did not lead to desired results for the system. Another challenge that came up during the development of FOCUS was transitioning the model from training in HiperGator to being deployable online. The problem that arose from this challenge was where to host the model developed during training as well as how to make the interface easily usable by users after deployment. These problems were solved after doing research into setting up a Streamlit host. Streamlit has seamless integrations to make AI projects such as FOCUS easily deployable. Using this cloud-based service made deployment less time-consuming compared to developing a front-end interface from scratch. Although Streamlit made abstraction towards deployment easier, at first connecting AWS S3 buckets to Streamlit was a challenge in itself. This is because the way that S3 stores data makes it difficult to read the data being downloaded. This is a deployment only problem because running locally meant that there were places to store the data. However, due to this being a cloud-based deployment, this made reading the data from the bucket difficult. This problem was resolved by doing more research into how to read the bytes provided by S3 and changing the codebase to work around this issue. On top of deployment, the challenge of how to go about preprocessing the CT scans was also an issue. The first-time training UNET did not complete properly due to the exploding gradient problem. This is a common issue within machine and deep learning that was solvable by normalizing the input data to a range of 0 and 1. An important aspect to take note of is that with medical data, there are major variations between intensity ranges between pixels which is why when normalizing it is important to take into consideration that certain ranges should be clipped in order to make deep learning generalization possible. On top of normalization, there was also the issue of not being able to identify tumors during training due to the small size of the tumor compared to the entire CT scan. This was a challenge because the model was not able to learn properly from random spatial cropping over a much larger image compared to the tumors. This problem was solved by creating a bounding box of the tumors for each of the images and cropping out the unnecessary parts of the CT scan. This helped improve model performance during training.

The novelty that FOCUS has comes from its approach to being an applicable system in the field. What many of these state-of-the-art deep learning models focus on is the best accuracy possible which is a good standard to hold themselves to. However, in the real world that is not necessarily the most important factor in determining if a system is successful. For example, latency/inference time is a very important factor when taking into consideration the success of a system. It has been shown that systems that take too long for a user to use will eventually be removed from user workflow. Another important consideration that has been mentioned previously in this paper and is closely tied to latency/inference time would be the size of the model. Making the model as lightweight as possible is a good goal to have in mind when creating real-world applications. This is because smaller models in general have faster inference times and are much easier to store and deploy from an economic standpoint compared to larger models. That is what FOCUS does differently compared to the state-of-the-art models. The development of FOCUS had all these goals in mind in order to build this product. If FOCUS were to be perfected to meet the standard criteria to be used in the medical field, this product would be a huge boon for medical practitioners and patients alike. Hospitals and clinics would not need to invest manpower into processing CT scans for tumors, but could instead invest that time that would be spent sifting through medical images on equally important tasks. Patients would also be able to get the results of CT scans back at a much quicker pace than they normally would even under the circumstance of a lack of manpower because FOCUS would be taking care of most of the segmentation prediction. Overall, the health industry would benefit from having extra time to work with which FOCUS hopes to achieve.

# Future Works and Improvements

Some potential improvements that could be made on FOCUS would be to improve results of the predicted segmentations to meet the standard criteria previously stated in the Performance Metrics section. This is the most pertinent issue that FOCUS has right now. Due to overfitting, the performance metrics used to evaluate the model do not meet the correct standard. Going back and improving upon these results would immediately make FOCUS a product much closer to real deployment. Another potential improvement that would enhance the functionality of FOCUS would be to expand its capabilities for other medical images such as MRI scans and X-rays. Doing this would greatly increase the efficacy of medical professionals’ abilities to identify a malignant tumor within a patient because doctors do not usually base their predictions off of a single type of medical image but instead use a multitude of data and images to reach their conclusions about a patient. This improvement would naturally address one of the limitations that was mentioned in the Discussion section. Taking these improvements into consideration, another improvement that would enhance user experience using FOCUS would be to develop a more aesthetically pleasing interface for users. Although the simplicity of using Streamlit made developing an interface efficient, Streamlit is inherently very limited in what it is able to do as a front-end library. That is why future versions of FOCUS will transition out of Streamlit onto an actual front-end interface using frameworks such as React to keep up with any potential functionalities or features that may be added to FOCUS.

Some areas of further research that would apply to FOCUS could potentially be researching how to make UNET even more lightweight than it is now. There is a decent amount of research already that tries to improve upon the performance of UNET by increasing the complexity of the architecture. There should be potential interest in research directed towards actually reducing the complexity of UNET even further while maintaining acceptable performance metrics similar to the original UNET. This would not only improve FOCUS by making inferencing time shorter, but would overall gear deep learning models to be more applicable to real world applications such as FOCUS. Some other areas of research could potentially be focused on how to create a deep learning model that does not take into consideration the type of images that are input into the data. Something along the lines of being able make generalizations on a variety of images. The theory behind latent space is an already prevailing concept within the AI domain. This concept has the potential to reach the capabilities of being able to ignore image type by reducing the dimensional space of the image to its latent space. This idea could definitely be researched further into the medical domain. The development of this technology would revolutionize the way practitioners examine medical data, potentially being able to provide more than just medical assistance for medical practitioners.

Theoretically, if the following improvements were implemented onto FOCUS, this would solve many of FOCUS’ existing and future challenges and limitations. The evolution of FOCUS could potentially not only resolve the issue of time-consuming medical segmentation for different types of medical images, but could also be used for other downstream tasks such as medical reconstruction which is the idea of enhancing image quality or reconstructing incomplete images. Of course, FOCUS is far from actually be capable of doing such tasks, however given the numerous capabilities of deep learning networks, there may be a chance in the future that FOCUS will be able to leverage these deep learning architectures to be able to improve the speed at which medical practitioners do these kinds of downstream tasks.

# Conclusion

Although the expectations were not quite met for FOCUS, mainly due to the lack of results from the model training, this does not mean that FOCUS does not have the potential to be a useful tool for future medical practitioners alike. With better training in the future, FOCUS can do real good for the medical industry. That being said, it will need a lot of work to get FOCUS up to that point. Especially when it comes to reducing overfitting on the UNET whether that be from an increase in training data or more data augmentations. This project showcases the comprehensive nature of managing an AI System’s lifecycle. All of the steps that it took to complete FOCUS were fundamental not only to building FOCUS, but to also managing future AI systems. Training a model is far from what is required to fully build and maintain a system. Data collection is an important base to start off with building a system because without a solid foundation, the rest of the work completed afterwards will not amount to much. Once training is completed, next comes making sure that FOCUS is compliant with the standards that were set out from the very beginning of the project. Standards such as the ones iterated in the Trustworthiness and Risk Management section are important to take into consideration when building a system. This being especially important for FOCUS due to the private nature of the data from patients. After these risk and trustworthy assessments have taken place, there comes deployment. The final steppingstone towards initial release into production. Streamlit is very much a powerful tool in that regard because it made deployment simple and included different features such as secret protection and a continuous deployment pipeline integration. Finally, once deployment has finished, there comes maintenance and monitoring. An AI system never truly finishes completion, somewhat similar to many normal software applications. What sets apart an AI system from normal software applications is that this cycle is repetitive. An AI engineer must maintain a solid grasp of how their system is fairing. That is why FOCUS, Prometheus and Streamlit are libraries that allow evaluation metrics to be tracked. The periodic monthly inspections are also another part of this. Without these inspections, FOCUS could easily go off the rails by performing poorly in future segmentations which would hurt not only the practitioners’ reputations, but also may endanger patients. To conclude, although FOCUS did not fully meet the expectations set out in the beginning of the project, there is much to improve upon with this system and like what was said previously, an AI system’s lifecycle does not really end. There is definitely a future for FOCUS within the medical industry.

# References

[1] B. Qi, X. Pan, Y. Kang, and H. Li, “Breast mass segmentation in mammography using improved dynamic programming,” IEEE Explore, https://arxiv.org/pdf/1907.10273 (accessed Nov. 29, 2024).

[2] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, “Automatic segmentation and disease classification using cardiac cine mr images,” SpringerLink, https://link.springer.com/chapter/10.1007/978-3-319-75541-0\_11 (accessed Nov. 29, 2024).

[3] A. Vard, A. Monadjemi, K. Jamshidi, and N. Movahhedinia, “Fast texture energy-based image segmentation using directional Walsh–Hadamard transform and parametric active contour models,” Expert Systems with Applications, https://www.sciencedirect.com/science/article/abs/pii/S0957417411004660?via%3Dihub (accessed Nov. 29, 2024).

[4] N. Mudigonda, R. Rangayyan, and J. E. Desautels, “Gradient and texture analysis for the classification of Mammographic Masses,” PubMed, https://pubmed.ncbi.nlm.nih.gov/11131493/ (accessed Nov. 29, 2024).

[5] M. Havaei et al., “Brain tumor segmentation with deep neural networks,” arXiv.org, https://arxiv.org/abs/1505.03540v3 (accessed Nov. 29, 2024).

[6] R. Wang et al., “Medical image segmentation using Deep Learning: A Survey,” arXiv.org, https://arxiv.org/abs/2009.13120 (accessed Nov. 29, 2024).

[7] H. Seo et al., “Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications,” Medical physics, https://pmc.ncbi.nlm.nih.gov/articles/PMC7338207/ (accessed Nov. 29, 2024).

[8] S. S. Yadav and S. M. Jadhav, “Deep convolutional neural network based medical image classification for disease diagnosis - journal of big data,” SpringerOpen, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0276-2 (accessed Nov. 29, 2024).

[9] K. O’Shea and R. Nash, “An introduction to Convolutional Neural Networks,” arXiv.org, https://arxiv.org/abs/1511.08458 (accessed Nov. 29, 2024).

[10] O. Ronneberger, P. Fischer, and T. Brox, “Convolutional Networks for Biomedical Image Segmentation,” arxiv, https://arxiv.org/pdf/1505.04597 (accessed Nov. 29, 2024).

[11] S. Wann et al., “Ethical considerations in CT angiography,” SpringerLink, https://link.springer.com/article/10.1007/s10554-006-9177-5#:~:text=The%20principle%20of%20beneficence%20dictates,limit%20exposure%20to%20potentially%20harmful (accessed Nov. 29, 2024).