

# Capstone Project Proposal



<Andrei – Claudiu Maftai>

## Business Goals

### Project Overview and Goal

What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?

Project name: Detection and classification of tumors

Across the world, it's getting harder and harder to detect and classify tumors due to social factors and the availability of radiologist doctors. Even when detected this process requires focus and time which is quite a luxury these days. The radiologist takes an X-ray or a CT / MR scan and then detect tumor with the "naked eye". This process is difficult and there can be many errors especially when doctors are working under stress and press.

Out of 1000 patients, about 10 % are recalled for additional diagnostic imaging, and of these 10 %, four or five patients are diagnosed with cancer / tumor. These false-positive exams lead to preventable harms, including patient anxiety, benign biopsies, and even worst: unnecessary intervention or treatment.

Furthermore, high false-positive rates significantly contribute to the annual billions of dollars in screening costs. These problems can be solved by emerging Artificial Intelligence (AI) based computer vision technologies, which will reduce the rate of false-positive prediction and improve the efficiency of the radiology departments. The business solution is to improve the effectiveness of radiologists by making the detection part smooth so that they can focus on providing accurate and faster results.

<p><b>Business Case</b></p> <p>Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.</p>	<p>This product will increase patient satisfaction as they don't have to wait for a couple of days to get the report / diagnostic. Furthermore the report will be highly accurate, preventing the patient from having another scanning. This will increase patients trust and increase revenue for medical labs and hospitals, directly and indirectly through customer's satisfaction. The machine learning will reduce the time to process the images and classify them into tumor / no tumor. In case of positives (tumor) will further predict whether it is benign or malignant. The whole process can be quite difficult to detect only by humans and time consuming. This product will help doctors to determine whether the patient needs immediate surgery or if needs treatment.</p>
<p><b>Application of ML/AI</b></p> <p>What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?</p>	<p>This can be done by using Deep Learning Computer Vision, in which my team and I will train our models on X-ray images / CT scans, for different body parts, with tumor and without tumor. The introduction of the Machine learning model will improve the performance of the radiology department in each hospital, and this will directly improve the lives of the radiologist who are working continuously.</p>

## Success Metrics

<p><b>Success Metrics</b></p> <p>What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.</p>	<p>The success metric will be measured by monitoring the increase in revenue of the radiology department and customer satisfaction. The customers include radiologists, patients, and oncologists. The success metric can also be measured by the rating given by a radiologist or oncologist.</p> <p>We want to improve the accuracy of cancer detection and reduce diagnostic's waiting time, which indirectly will help oncologists makes decision faster.</p> <p>Moreover, to establish a baseline value to provide a point of comparison for the product, we could have a comparative look at the statistics of false-positive cases and time between scanning and final diagnostic, before and after the deployment of the product.</p>
---	---

# Data

## Data Acquisition

Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?

Initially, we are going to get five to seven hundred images for every important part of the body (head, chest-abdomen, limbs) and label them either in-house or hire a third party. We will use a platform like Appen to monitor and improve image labeling. For the first model, we will be using only two categories: "Tumor" and "No tumor" and we will try to balance them in order to avoid biased prediction.

Furthermore there will be a radiologist or oncologist into the loop to review the image again, and label the image with the type of tumor (malignant or benign).

After categorizing the images, in case of Tumor detection, will use the images to create a dataset for training a second model to predict whether the tumor is malignant or benign.

Lastly we will train a NLP model using datasets of previous diagnostics and patients' history, in order to generate a preliminary, detailed diagnostic.

## Data Source

Consider the size and source of your data; what biases are built into the data and how might the data be improved?

We estimate using an overall dataset consisting of 800 – 1000 scanings(patients) with an estimated size of 24.5 GB

We will consider collecting some of the X-ray / CT images from hospitals and also make use of publicly available data sets, taking in consideration "Image Augmentation" strategies, which can be used to "trick" the network into thinking we have more training examples at our disposal then we actually have.

For the rest of the data (patient's diagnostics and medical history / images) we will consider contracts with different hospitals so that we can easily get access to new datasets from the real world. In return, we will provide discounts and additional services.

## Choice of Data Labels

What labels did you decide to add to your data? And why did you decide on these labels versus any other option?

Initially, we will keep it simple with only two categories and with time we will add up to six categories.

In case of Tumor, subcategories of benign, malignant, and unknown will be added. We will be using these labels because they precisely arranges our data making it easier for us to normalize the data-set.

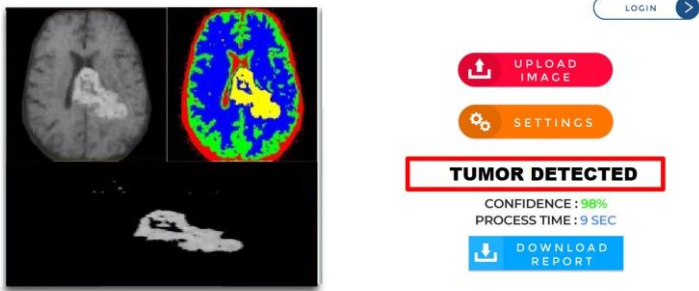
It will be easy to deploy and maintain two categories of the dataset and the model performance will also increase on binary tasks.

The main limitation to this approach is that the model(s) will need a lot of data, so arguably the data collection will be one of the most important parts of the entire process.

# Model

<b>Model Building</b>  How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	<p>Due to privacy concerns and sensitivity of data we want to keep the model private and train it in-house, hiring a team of well prepared machine-learning engineers. This will help us launch a viable product.</p> <p>We will train the dataset on various higher tiers models such as CNN and RNN and then ensemble them to get the best results.</p> <p>This process is iterative and we will keep improving our models with time.</p>
<b>Evaluating Results</b>  Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	<p>In tumor detection, Recall and Precision is the most important metric that we want to improve and subsequently improve overall effectiveness by monitoring the F1 score. Usually we would want to detect all the tumors in all cases. So, we want this to be as high as possible like above 98.0 or 99.0 %.</p>

## Minimum Viable Product (MVP)

<b>Design</b>  What does your minimum viable product look like? Include sketches of your product.	
<b>Use Cases</b>  What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?	<p>The product is designed to be used by one persona: Doctors (radiologists / oncologists)</p> <p>Doctors will access the product via a simple user interface. Our UI will first ask the costumer (i.e. doctor) to add an image. In the “SETTINGS” section, the doctor must select the type of examination (X-ray / CT / MR), the body part and the clinic situation of patient. Program will then process the image and after a few seconds, it will give the result of whether this image contains tumor or not. In case of tumor it will also predict if it is benign or</p>

	<p>malignant. It will also give confidence score and process time.</p> <p>In the end it will generate a report consisting in a preliminary diagnostic, using our pre-trained NLP model. The interface will be user-friendly and it's almost similar to the UI that they are used to in medical professions. We will keep it simple at first so that the users won't need to take hard training.</p>
<p><b>Roll-out</b></p> <p>How will this be adopted? What does the go-to-market plan look like?</p>	<p>We will start small by training the model on two categories and improve our F1 score by preprocessing and model selection.</p> <p>At the end of every examination, the UI will ask the user / doctor to give a rating on the results and the time spent to get those results. This will help us determine success metrics. Overall, the whole product will be deployed hospital's local servers, and users either will be able to connect the database or upload images individually. The go-to-market plan will be as below:</p> <p>Prelaunch – We will conduct market research, test our product enough, be prepared to fulfill orders, offer early birds discount for those willing to test our product, generate awareness and hype.</p> <p>Postlaunch – We will monitor our product performance and keep continuously improving, talk to customers and get their requests and roll out new features, fix bugs if there are any.</p>

## Post-MVP-Deployment

<p><b>Designing for Longevity</b></p> <p>How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?</p>	<p>We will actively monitor the models performance metrics and try to improve them by training it on new images.</p> <p>We will monitor success metrics and customer reviews on our product to provide them with a better interface and product performance.</p> <p>We will keep on improving the interface so that it will take a few clicks to get the results that the doctor wants.</p> <p>We will also improve the quality of image processing, in order to get better results.</p>
---	--

	<p>Active learning will be the main part of this product as we will be getting new data from patients.</p> <p>A/B testing can be implemented in deploying two different models using the same training data to see what they might classify. This would be useful to see if the two models both show the same trends in the data or are showing something different. There would be more confidence given the results if both models show the same results.</p> <p>We will improve the model performance for unknown category by adding humans in the loop such as radiologist or oncologist.</p>
<p><b>Monitor Bias</b></p> <p>How do you plan to monitor or mitigate unwanted bias in your model?</p>	<p>To mitigate bias, we will be collecting data from various hospitals and try to keep our database balanced. This will be done under active learning so our product can predict accurately a new type of tumor. We will also keep humans in the loop to monitor the performance of our product, most probably in house radiologist / oncologist, who will monitor the unknown category and try to label the data to improve the performance.</p>