TecnoCampus
Escola Superior
Politècnica

Centre adscrit a la
upf. Universitat
Pompeu Fabra
Barcelona

Degree in Computer engineering for Management and Information Systems

GENERATIVE ADVERSARIAL NETWORKS AND SCORE – BASED MODELS FOR MEDICAL IMAGING GENERATION

Viability Analysis

Kelly Zhen Zhou

Tutor: Xavi Font Aragones

2022 / 2023

Table of contents

List of Figures

1. Initial Planning

The following chapter presents in detail the definition of the initial planning.

1.1 Task definition

These are the main phases of the project. The steps to reach the main objectives are the following, no steps can be taken without doing the previous one.

1. Information research.
    - 1.1.Research and reading of current papers about Generative Models
    - 1.2.Understanding of Generative Adversarial Models
    - 1.3.Understanding of Energy-based Models
    - 1.4.Research for current best Generative Models.
    - 1.5.Research for a dataset that will fulfill legal requirements.
2. Project definition
    - 2.1.Define project objectives and coverage
    - 2.2.Define project Methodology
    - 2.3.Functional and technical requirements definition
    - 2.4.Budget definition
    - 2.5.Risk definition
    - 2.6.Viability Study
3. Project planning and roadmap
    - 3.1.Timing development phases
    - 3.2.Definition of data and resources management
    - 3.3.Definition of project management framework
    - 3.4.Definition of tasks and dates
4. Project development: This is the most critical section, as the project results depend entirely on what is done during this phase, the hours that will be dedicated should focus on this point.
    - 4.1.Find a dataset

4.2.Clean the provided data: As the data is given in the shape of 4D images, segmentation and reconstruction will be needed.

4.3.Analyze the data: Document the distribution of the data and find exceptions.

4.4.Train the models: GAN models and score-based models.

4.5.Generate an output of images.

4.6.Evaluate the results and compare the input.

5. Risk management

1.2. Time Management

The needs of the software used for management relies on this headline. During the first week of this research, it was agreed with the tutor project the distribution of the meetings to ask questions and review the research progress. The time agreed for these meetings was set to be approximately every two weeks, which would be a rigid model of time distribution and would perfectly fit into Agile Scrum Sprint.

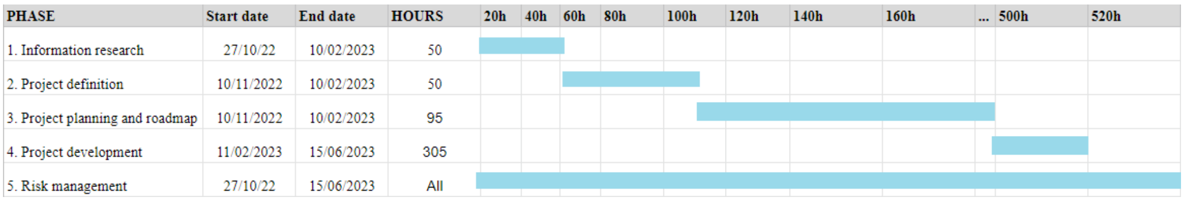
Taking that into account, the work was distributed in Sprints of two weeks and each sprint would aim to achieve a specific goal. As the subject concerning this project demanded approximately twenty ECTS credits from October to July and one ECTS credit equals one working hour, it would mean that the total would be 500 hours and being equally distributed would take 62 hours each month (fifteen hours each week).

It was taken into consideration that the months dedicated to the draft .The beginning of the project definition would have fewer working hours, therefore it was distributed as 20% of the hours before January and 70% after January. This would give us the following time distribution:

- From November first to January first: 100 hours, 12 hours each week. 4 Sprints.
- From January first to June first: 400 hours, 16 hours each week, 6 sprints

Meetings for reviews and questions with the project tutor were set on Fridays at 19:00 every two weeks, that would be the ending date for a sprint. Note that not all scheduled stories should be done in each sprint, as this is a research project the flexibility of deadlines should be considered as new information and findings should be welcomed to improve the research and redirect the comprehension of the approach to the research to its goal. This also follows the Agile framework purpose.

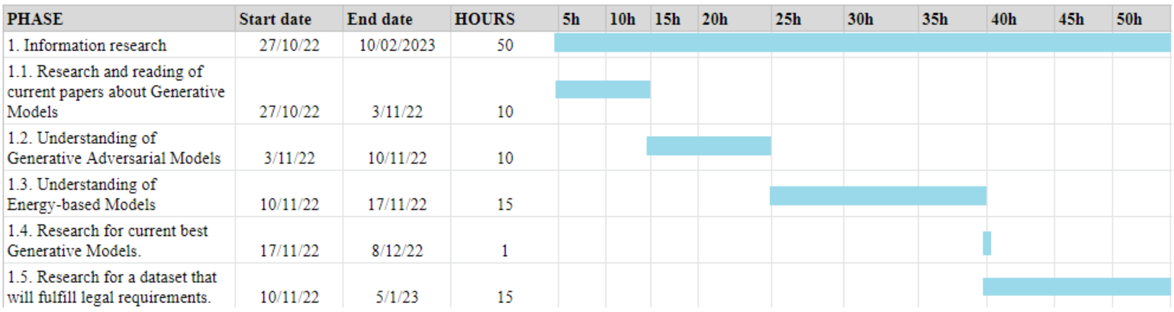
Gantt diagram for main phases:



PHASE	Start date	End date	HOURS	20h	40h	60h	80h	100h	120h	140h	160h	...	500h	520h
1. Information research	27/10/22	10/02/2023	50	▨	▨									
2. Project definition	10/11/2022	10/02/2023	50			▨	▨	▨						
3. Project planning and roadmap	10/11/2022	10/02/2023	95					▨	▨	▨	▨	▨		
4. Project development	11/02/2023	15/06/2023	305										▨	
5. Risk management	27/10/22	15/06/2023	All	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

Fig 1.2.1. Gantt diagram for the main phases of the project .

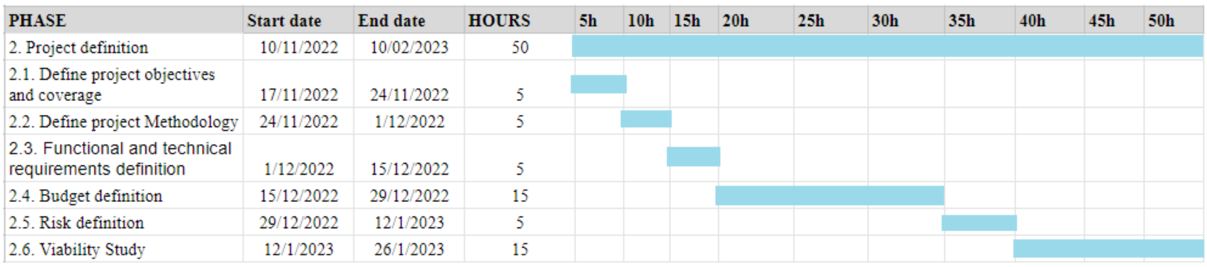
Gantt diagram for first phase: Information research



PHASE	Start date	End date	HOURS	5h	10h	15h	20h	25h	30h	35h	40h	45h	50h
1. Information research	27/10/22	10/02/2023	50	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
1.1. Research and reading of current papers about Generative Models	27/10/22	3/11/22	10	▨	▨								
1.2. Understanding of Generative Adversarial Models	3/11/22	10/11/22	10			▨	▨						
1.3. Understanding of Energy-based Models	10/11/22	17/11/22	15					▨	▨	▨			
1.4. Research for current best Generative Models.	17/11/22	8/12/22	1								▨		
1.5. Research for a dataset that will fulfill legal requirements.	10/11/22	5/1/23	15								▨	▨	▨

Fig 1.2.2. Gantt diagram for the information research phase of the project

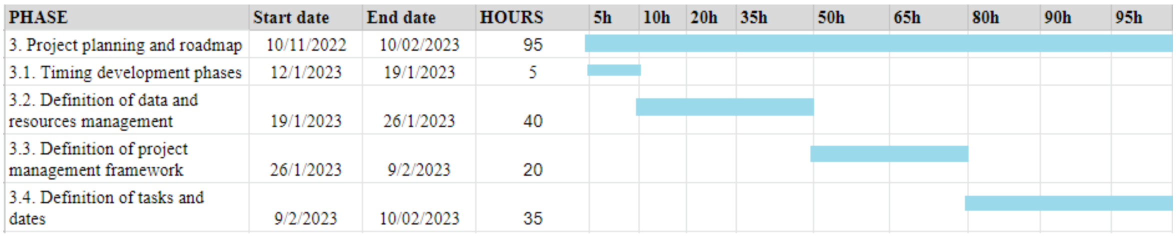
Gantt diagram for second phase: Project definition



PHASE	Start date	End date	HOURS	5h	10h	15h	20h	25h	30h	35h	40h	45h	50h
2. Project definition	10/11/2022	10/02/2023	50	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
2.1. Define project objectives and coverage	17/11/2022	24/11/2022	5	▨									
2.2. Define project Methodology	24/11/2022	1/12/2022	5		▨								
2.3. Functional and technical requirements definition	1/12/2022	15/12/2022	5			▨							
2.4. Budget definition	15/12/2022	29/12/2022	15				▨	▨	▨				
2.5. Risk definition	29/12/2022	12/1/2023	5							▨			
2.6. Viability Study	12/1/2023	26/1/2023	15								▨	▨	▨

Fig 1.2.3. Gantt diagram for the project definition phase

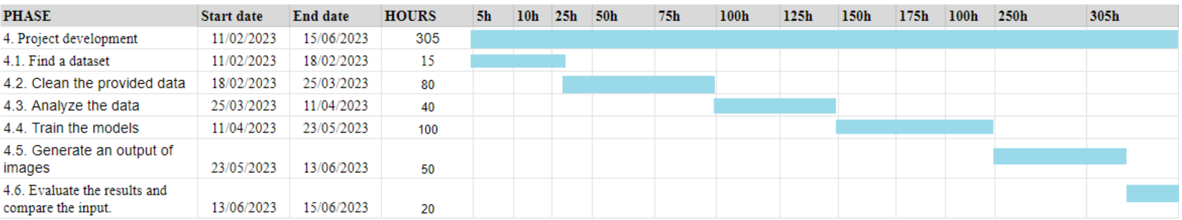
Gantt diagram for third phase: Project planning and roadmap



PHASE	Start date	End date	HOURS	5h	10h	20h	35h	50h	65h	80h	90h	95h
3. Project planning and roadmap	10/11/2022	10/02/2023	95	▨	▨	▨	▨	▨	▨	▨	▨	▨
3.1. Timing development phases	12/1/2023	19/1/2023	5	▨								
3.2. Definition of data and resources management	19/1/2023	26/1/2023	40		▨	▨	▨					
3.3. Definition of project management framework	26/1/2023	9/2/2023	20					▨	▨			
3.4. Definition of tasks and dates	9/2/2023	10/02/2023	35							▨	▨	▨

Fig 1.2.4. Gantt diagram for the project planning and roadmap phase

Gantt diagram for last phase: Project development



PHASE	Start date	End date	HOURS	5h	10h	25h	50h	75h	100h	125h	150h	175h	100h	250h	305h
4. Project development	11/02/2023	15/06/2023	305	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
4.1. Find a dataset	11/02/2023	18/02/2023	15	▨	▨										
4.2. Clean the provided data	18/02/2023	25/03/2023	80			▨	▨	▨							
4.3. Analyze the data	25/03/2023	11/04/2023	40						▨	▨					
4.4. Train the models	11/04/2023	23/05/2023	100								▨	▨	▨		
4.5. Generate an output of images	23/05/2023	13/06/2023	50											▨	▨
4.6. Evaluate the results and compare the input.	13/06/2023	15/06/2023	20												▨

Fig 1.2.5. Gantt diagram of the project development phase.

1.2. Deviations

After the project completion the following deviations took place:

Hour deviations : During the first part of the project the hours (100) were correctly estimated, nevertheless , the hours required for the second part of the distribution were 400 , 16 hours per week, during development 20 hours per week were needed.

Start and End date of each phases advanced as planned until point 4.2, Clean the provided data, which took more than 80 hours with a total of 140 hours. The miscalculation did not suppose an issue in by the end of the phase 4.3 , because the dataset needed were images there was no content to perform a deep analysis.

For phases 4.4 and 4.5 , regarding the training and output generation, time went up to more than 200 hours . This implies that the time needed to correctly train , adjust and evaluate the models was underestimated , the project completed without all of the models showing results and subtracted time for the evaluation of results and conclusions.

2. Budget

The project aims to be a research, therefore all the workers needed for this is limited to up to three researchers. Considering the salary of a machine learning researcher as the average in January 2023, salary is up to 140k euros per year. The project should not take more than a year but 500 working hours which equals to one month, in working hours, two months. The amount to be paid to the researches goes then for one month , going to 11666 euros as gross salary.

Over this number there should be a reduction of 4.70% of social security given that the project is being developed in Spain. There should be at least a payment for the training and a price of 1,55% for unemployment. There is also a percentage of IRPF under any employee around twenty and twenty-five percent giving a tax of 30% per month of SS and IRPF. Therefore, for two months the payment for that employee should be around 14800 euros.

Additional costs include the working space and necessary hardware and software. In the technical requirements most of the software do not need to pay for its use, but in case the batch of data and processing is too big, an additional eleven euros per month may be added for a proper Google Colab plan. This would add twenty-two additional euros to the expense of renting an office with an average price of 250 euros each month.

The total budget calculated with the case of most expenses and given a team of one researcher goes to 15.322 euros. Adding extra researchers to the project may duplicate the need for the budget.

2.1. Economic Viability Analysis

As stated in previous headlines analysis of the market, the project aims to be a research and applications of the results may vary for different clients. There are no products that integrate current studies in this field because it is in development. A client and competence products need to be defined in order to perform a proper economic viability analysis, but analyzing on suppositions the future tendency for the researches around this field is to be released to the market when the results are clear enough to create products based on them.

The use of this project is specifically for computer tomography and limited to the dataset content of carcinogenic cells, but furthermore , models can be trained with different types of medical imaging as long as the training data is correctly prepared.

3. Viability Analysis

3.1. Technical Viability Analysis

Technical requirements were previously listed in header 5.2 . The list of software is open source, therefore there is no real cost for it in exception of paying special payment plans as for Google Colab.

Because most of the software used is in the cloud the major risk is the availability of Google Colab and PyTorch while developing the research. Models may take days to train and the services should keep up with the needs of training , crashing during the trainment may be fatal for the time estimation for developing. Parallel execution of the work may improve the training speed slightly.

Evaluation of technological stack:

- Google Colab
  - Evaluation : Google Colab will ensure a safe and usable work environment with the most used Python libraries already included. The decision on this software relies on the speed for the set up and storage availability. I may be connected to drive folders and heavy datasets are easy to store and access.

    Google Colab will allow access to GPU and TPU which will accelerate training processes.
  - Alternative to Google Colab , Anaconda may be used as a Data Science environment. It works locally and performance depends on the machine As it is local, it is not as easy to share files as Google Colab would.
    Anaconda will provide a more personalized environment but the performance and local
- Google Drive
  - Evaluation : Having files in the cloud will ensure an easy access and share, the institutional email that will be used in Google Drive allows up to 250 GB which will be needed to store the data of this project. As the chapter about the

development indicates, dataset takes up to 1010 subjects and only 200 subjects have a weight of 22 GB that need to be preprocessed.

  - Alternative to Google Drive is to use local storage, which implies the use of local storage. The machine where this is being development had up to 30 GB of free storage and consequently , the download and upload of the data took time and storage resources as it needed to be done in batches of 200 subjects. Alternative storage like OneDrive or Dropbox were also an option, but Drive would adapt better to the rest of software as it allows to be connected with Google Colab.
- Python
  - Python is the programming language that will be used. Python is a general purpose programming language which works with scripting and is used by multiple environments and frameworks within the machine learning stack, is the chosen language. Most machine learning libraries and environments are made in or for python.
  - An alternative R programming language is also an option for data analysis but while R is more focused in statistical analysis and data displaying its usage in the machine learning field in which the project is focused is not as better fit.
- Python Notebooks (Jupyter Notebooks)
  - The .ipynb format provides an easy and clean way to display results and text regarding the study and conclusions.
  - While a regular Python script may be used, when there is the need to present and see results , the notebooks allow an easy read user interface and allow to run code in different blocks.
- PyTorch
  - PyTorch library is a deep learning library that works with tensors. It will be installed to Google Colab with pip3 command and import torch and torchvision libraries to be used in the Python Notebook.
  - The main alternative to PyTorch is TensorFlow , because the current project is being developed in Python and use a full python stack the option chosen is PyTorch. While TensorFlow comes in with the Google Colab environment, during the development it is needed to call TCIA REST API to retrieve data of the subjects. This retriever works with an older version of Pandas (1.5.3 to

current 2.0.0) and needed to force reinstall it. The reinstallation would also lower the numpy version which was needed in Google Collab TensorFlow. Therefore, PyTorch remained as the final decision for the project approach.

- TCIA[11] REST API
    - To retrieve the information and label the subjects, the TCIA (The Cancer Imaging Archive) page provided a REST API to download collections into usable data , between these collection is included the dataset to be use (LIDC-IDRI) . While images are not part of the data retrieved there is all subjects information returned as a list.
    - There are no alternatives for this REST API, it is needed and is the way TCIA allows to work with the data.
- NBIA Data Retriever
    - NBIA Data Retriever is the software provided by TCIA to download collections with .dcm images for each subject. The study of the dataset is included in the development chapter.
    - There are no alternatives for the retriever , the subject collections need to be chosen in the TCIA website and that will provide a manifesto file that initiates the download of the data collection.
- dicom2jpg python library
    - Because the data provided by TCIA , mainly the images, are in DICOM format (.dcm) there is the need of a way to manipulate the images for the data analysis. This library under MIT license allows to convert .dcm images to JPG , PNG and numpy arrays.
- Pickle
    - After the models have been trained there will be a need to storage the trained models. Models take time and resources and training the model each time the code is runned is inefficient. The models are trained once and passed to a Pickle file, then when needed the already trained model will be reloaded and will be used without the need of training again.

3.2. Environmental Viability Analysis

Environmental Viability Analysis is a key point in this project.

The training and use of the models requires great processing power and execution time, since they work with different layers and the results depend enormously on the hours of training. In addition to generating the dataset there is an initial impact coming from the CT scans, from which the computed tomographies are obtained.

Along with these processing demands, the electricity cost of the computers where the project is being worked on has no comparison, but it must also be taken into account, as well as that of the office where the project is being worked on and the set of offices, computers and servers on the part of the project. of the required software that is in the cloud (Google Colab, PyTorch).

The best measure that can be taken for this project is to monitor the training hours, being aware if training has stopped unexpectedly because trying to train the model again would be a waste of time and resources. The training time may also be limited to , the time stated in previous researches that took to train the specific model [1]

3.3. Legal requirements

The main focus of this section is the dataset that will be used for research, since the data is not always public and it must also be adjusted to the needs of the project. In the search for a dataset, interviews were conducted with several radiologists who recommended the Cancer Imaging Archive page, which provides a full breakdown of legal rights of use and needs for each dataset. Because most of these datasets were too large for investigation, another source, The Lung Image Database, was provided.

The legal requirements of this project does take into account Attribution 3.0 Unported (CC BY 3.0) which allows sharing and adapting for the selected dataset and MIT License for library dicom2jpg.

Following the organic law 3/2018, of December 5, Protection of Personal Data and guarantee of digital rights, no image will be directly associated with the names or data of the real patients to whom they belong and the project undertakes to comply with this and avoid at all costs the use of datasets that do not contain data origin specifications.

Because this project is made with free software, there is no need for additional licenses.

3.4. Diversity and genre perspectives

This project is highly focused on human needs that involve diversity. It is a project for the medical area and the results or its execution will not make a gender or racial difference between its patients or participants.

The contribution of this project is intended to be humanitarian, by dedicating itself to solving a problem for people in a health situation that causes them a degree of disability and limits their quality of life.

The development of this project hopes, therefore, to improve the quality of life of those affected by cancer regardless of who or how they are, the health need is not conditioned by any racial, sexual or ethnic factor.

The final user of the project will be experts on the medical field or Machine Learning , the notebooks to read and adapt are public to all and include a README file to understand the execution process of each step , which are also listed and explained in each notebook.