PROJECT PLAN

Module Code: COMP30151

Module name: Project 202021 Full Year

Project Title: COVID-19 vaccine Degradation

Supervisor : Dr Joao De Castro Cardoso Ferreira

NTU ID: N0830182

Name: Dayeeta Das

Module Code: COMP30151
Module Name: Project 202021 Full Year
NTU ID: N0830182
Name: Dayeeta Das

TABLE OF CONTENTS:

Chapter 1 – Introduction

The fight against COVID-19 seems to be almost inevitable. In fact the combined efforts of biochemists, researchers and data scientists all across the world has led towards developing solutions to combat this inconspicuous nemesis. mRNA vaccines have taken a lead towards combating the disease. However, they have their own limitations. The main concern of this project is to develop models and design rules for RNA cell degradations. Machine Learning and Deep Learning will be used on the Eterna Dataset in the development of these models.

1.1 Motivation

The Covid-19 pandemic has caused havoc worldwide. It has led to the loss of millions of lives across the globe (Coronavirus disease (COVID-19), n.d.). In order to combat the SARS-CoV-2 virus there is a necessity to develop a vaccine which is not only fast effective but also can be distributed widely.

1.2 Background

SARS-CoV-2 alike the other coronaviruses has been named so due to the crown-like spikes on the surface. It uses these spikes to invade the human cells (Carnahan and Mishra, n.d.).The virus mainly seeps into the lungs or the respiratory track. On finding its way to the body of the host, the viral RNA becomes a part of the host cell's protein production machinery. It then reprints these viral proteins and RNA marshalling the spread of the disease.

This is where the mRNA vaccines comes into action. The mRNA vaccines basically mimic the natural infection of the virus, but they contain only a short synthetic version of the viral mRNA which encodes only the antigen protein. Since the mRNA used in vaccination cannot become part of the person's chromosomes, they are safe to use [2.].

The Eterna community, led by Professor Rhiju Das, a computational biochemist at Stanford's School of Medicine, brings together scientists and gamers to solve puzzles and invent medicine. Eterna is an online video game platform that challenges players to solve scientific problems such as mRNA design through puzzles. The solutions are synthesized and experimentally tested at Stanford by researchers to gain new insights about RNA molecules (kaggle.com, n.d.).

1.3 Current Challenges

As mentioned previously mRNA vaccines can prove to be the ultimate saviour in this situation but they are currently facing a serious issue. RNA molecules have the tendency to spontaneously degrade. This is a serious limitation--a single cut can render the mRNA vaccine useless. Details of which part of the backbone of a given RNA is the most vulnerable is yet to be discovered. Unfortunately, due to the absence of this piece of information, the present mRNA vaccines needs to be shipped under extreme conditions. This approach is barely feasible as hardly a fraction of the affected population might receive the vaccine. So, there is an urgent need to fortify this process of vaccine creation [3.].

1.4 General Implications

- The models developed in this project will be scored on a 2nd generation of RNA sequences devised by Eterna players.
- The final test sequences are currently being synthesized and experimentally characterized at Stanford University.

This is not the first time that mRNA has been used by bio scientists to scrutinize and manufacture viral proteins. Over the decades, there has been successful cases where RNA molecules were able to render information about a virus and cease its production in the human body.

The successful completion of this project will manifest that machine intelligence can lead towards the design of a vaccine. In the broader aspect of the society, this model can be viewed as a success towards ending one of the lethal pandemics in the history of mankind.

Chapter 2 - Aims and Objectives

2.1 Aim

Create a learning model using Machine Learning(ML) and Deep Learning(DL) that will predict the likely degradation rates at the bases of RNA molecules.

2.2 MoSCoW Analysis

- ❖ Must Haves:
  - A learning model predicting the RNA degradation rates.
  - A generalized model with high accuracy score.
- ❖ Should Haves:
  - Code depicting efficient 'data cleaning' and 'data pruning'.
  - Sufficient visualizations of the training dataset.
  - Coding style showing effective use of Feature Engineering.
  - Usage of different ML and DL models for predicting the score of the training data.
  - An output log.
- ❖ Could Haves:
  - User Interface that allows a programmer to predict the score of the trained dataset using various ML and DL models.
  - Usage of LSTM and GRUs.
- ❖ Would Haves:
  - Converted version of the entire code in Julia.
  - A way to generate a dataset of RNA molecules that has just been created on the Eterna platform.
  - A system for integrating the Eterna gaming platform and the learning model together.
  - An UI for a player at Eterna to check the likely degradation rates of the RNA molecules.

2.3 Primary Objectives

No.	Objectives	Specific	Measurable	Achievable	Relevant	Time-Bound
1.	Describing the Eterna dataset.	A detailed explanation of the different features present in the dataset and how they contribution towards the degradation of the RNA molecules.	Visual representation through graphs and plots can be used to judge the importance of a particular feature in determining the degradation rate.	There are evidences from existing solutions on this problem where the dataset has been thoroughly described. This proves that it is possible to achieve this particular objective.	Data description and graphical visualization will serve to identify the main features needed to be taken into consideration for training the data.	16.11.2020
2.	Creating a 'cleaned' version of the dataset.	Elimination of those features which plays little or no role in	Visualization of the dataset after deleting the features. Graphs	Evidences from previous solutions on the problem proves that it is	Cleaning the dataset will play an important role in determining	18.11.2020

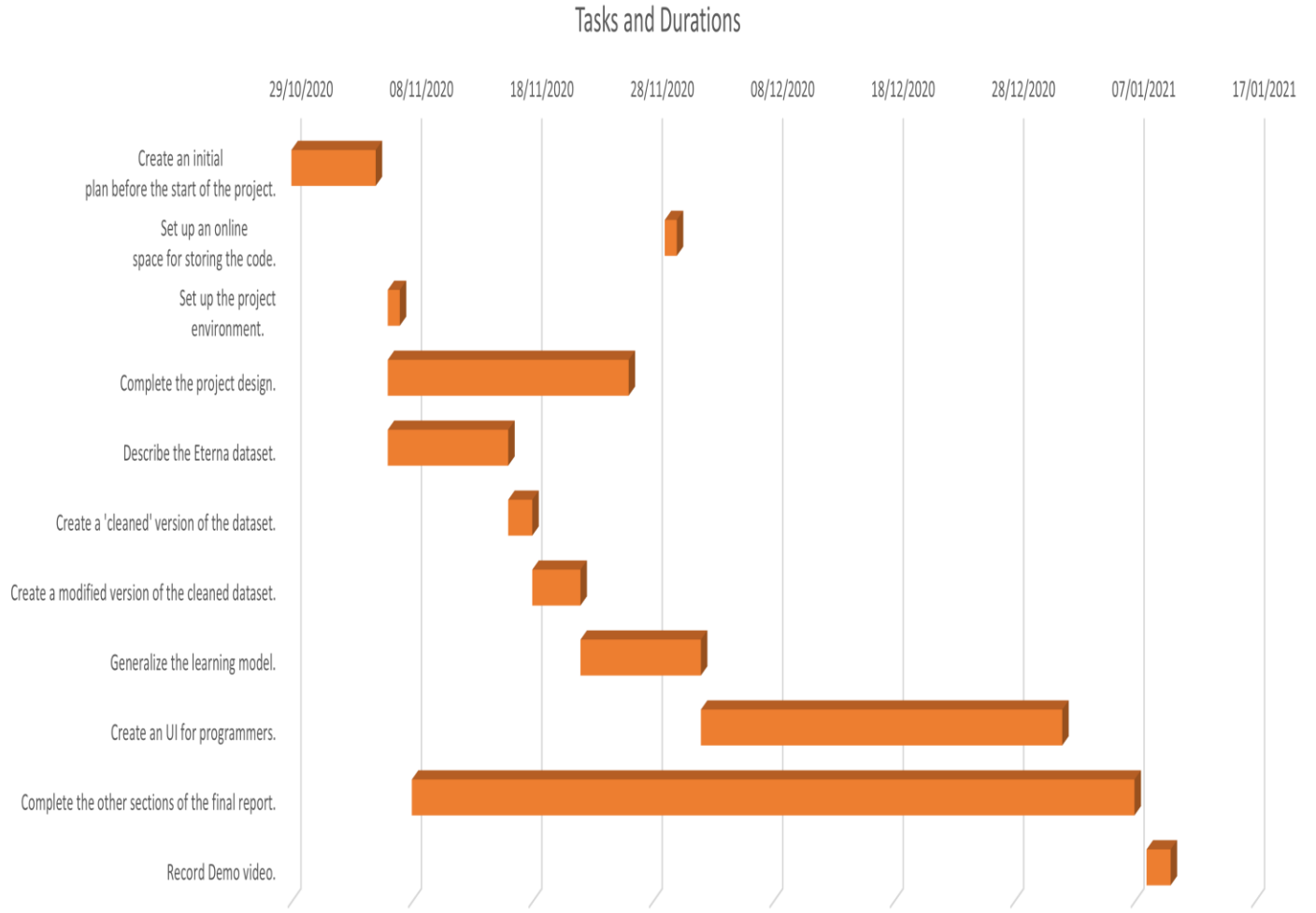
		determining the degradation rate.	can be used to determine the relevance of the current features with respect to cell degradation.	possible to achieve this objective.	the accuracy of the trained model.	
3.	Creating a 'modified' version of the cleaned dataset.	Modifying the structure of the dataset to fit the training model	Line graphs can be used to determine what impact the newly engineered features are having on the target variable.	All the previous solutions to this problem has used a modified version of the dataset for training which shows this objective can be accomplished.	Pruning the dataset plays a very important role in training the data. A perfectly pruned dataset can be used to create a perfect training model. This is what we are trying to achieve here.	22.11.2020
4.	Generalizing the learning model.	Attempting to create a model that will give a high accuracy score when trained on most ML/ DL models.	An average high accuracy score on different ML and DL scoring models will determine that successful creation of a generalized prediction model.	Previous solutions to this problem shows that it is possible to get a generalized model with a high accuracy score on various ML and DL scoring models.	This is the ultimate goal of this particular project.	2.12.2020
5.	Creating an user interface for the programmer.	Developing a UI using widgets in Python. This can enable the programmer to predict the accuracy score on a particular ML/ DL model by just clicking on a button on the screen.	The amount of time it takes for a new user to get used to the environment determines how good/ bad the interface is.	There are not many evidences showing the development of an user interface. Hence, this objective possess a high risk of failure.	This objective steps up the project to a new level. The UI might particularly help a layman. However, it is not essential in determining the main aim of this project.	2.1.2021

Chapter 3 - Tasks and Deliverables

Tasks	Deliverables
1. Create a project plan before the start of the project: 1.1 Research on the project topic chosen. 1.2 Identify the aims and objectives of the project. 1.3 Identify the potential risks of the project. 1.4 Determine a prototype for completing the project. 1.5 Assemble the points mentioned above in a document. 1.6 Get the document reviewed by the supervisor.	Project Planning Document
2. Set up an online space for storing the code.	No Deliverables
3. Set up the project Environment: 3.1 Watch tutorials on how to use Google Colab. 3.2 Download the Eterna dataset from Kaggle. 3.3 Install the packages and libraries necessary.	No Deliverables
4. Complete Project Design: 4.1 Identify the main use cases and classes in the project. 4.2 Create sequence diagram to depict the flow of information among the primary objects in the project. 4.3 Identify the operations and attributes of the objects.	Design section in the final report.
5. Describe the Eterna dataset: 5.1 Peek at the data. 5.2 Check the dimensions of the dataset. 5.3 Check class distribution and correlation among attributes. 5.4 Use Pandas to create graphs illustrating the relationships among the different features.	No Deliverables
6. Create a 'cleaned' version of the dataset: 6.1 Delete columns that contain single values. 6.2 Drop irrelevant columns. 6.3 Delete columns with unique values.	No Deliverables
7. Create a modified version of the cleaned dataset: 7.1 Convert the columns having textual data to numeric values using encoding. 7.2 Generate dummy values for fields with missing data. 7.3 Normalize the dataset.	No Deliverables
8. Generalize the learning model: 8.1 Check the accuracy score across the various ML/DL models. 8.2 Keep training the model till an average high accuracy score is achieved across various ML and DL scoring models.	No Deliverables
9. Create an UI for the programmers: 9.1 Watch tutorials on how to develop GUI on Python platforms like Jupyter.	No Deliverables
10. Complete the other sections of the final report.	Final Report
11. Record a video demonstrating the project.	Demo Video

Module Code: COMP30151
Module Name: Project 202021 Full Year
NTU ID: N0830182
Name: Dayeeta Das

Chapter 4 - Gantt Chart



	Record Demo video.	Complete the other sections of the final report.	Create an UI for programmers.	Generalize the learning model.	Create a modified version of the cleaned dataset.	Create a 'cleaned' version of the dataset.	Describe the Eterna dataset.	Complete the project design.	Set up the project environment.	Set up an online space for storing the code.	Create an initial plan before the start of the project.
Start Date	08/01/2021	08/11/2020	02/12/2020	22/11/2020	18/11/2020	16/11/2020	06/11/2020	06/11/2020	06/11/2020	29/11/2020	29/10/2020
Duration	2	60	30	10	4	2	10	20	1	1	7

Chapter 5 – Resources

5.1 Resources

- Google Colab
- Jupyter Notebook
- Eterna Dataset: Training set and Test set
- Laptop

5.2 Sources

- Google
- Anaconda platform
- Kaggle website
- Eterna gaming platform
- NTU Online Library

Module Code: COMP30151
Module Name: Project 202021 Full Year
NTU ID: N0830182
Name: Dayeeta Das

Chapter 6 - Risks

Risks	Possibility (1-5)	Severity (1-5)	Risk Impact (1-5)	Mitigation Plan
1. Dataset not described or visualized properly.	2	3	3	1.1 In the early stage of the project create graphs taking into account all the features present in the entire dataset. 1.2 Describe each feature including the ones having null values.
2. Dataset ending up with a lot of noise	3	4	5	2.1 Carefully examine the dataset before training. 2.2 Make the training dataset larger 2.3 Resample with different ratios to get rid of imbalanced data.
3. Insufficient implementation of Feature Engineering done on the dataset.	2	5	5	3.1 Identify the indicator and interaction features. 3.2 Perform feature representation. 3.3 Bring in external data.
4. Chances of dataset getting overfitted/ underfitted.	3	4	5	4.1 Use cross validation methods. 4.2 Train learning algorithm iteratively. 4.3 Use regularization techniques. 4.4 Use ensemble methods like bagging and boosting.
5. Risk of losing information due to short term memory	2	5	3	5.1 Use LSTMs and GRUs.
6. Risk of the model not converging to the global maxima.	5	3	4	6.1 Use bigger batch sizes that will create smoother updates by updating the learning rate decay.
7. User Interface too complicated.				7.1 Reduce the number of sections on the interface. 7.2 Create buttons and tabs only wherever necessary.

Chapter 7 - LSEPIs: Legal, Social, Ethical and Professional Issues

I work with due care and diligence, acting in my client or company's best interests at all times. I take personal and collective responsibility for my actions while maintaining discretion and ethical standards (Bcs.org, 2019).

The successful completion of this project does have a significant social impact. It might lead to the production of vaccines that can be supplied to the masses to cure COVID-19. However, it is important to note that the model generated by the end of this project will not be deployed directly for the production of the mRNA vaccines.

This model along with other prediction models devised by data scientists will be tested on a 2nd generation of RNA sequences devised by Eterna players. The model will be scrutinized thoroughly by the members of the Eterna community led by Professor Rhiju Das and his team at Stanford's Medical School. The team might end up taking the best bits from each model to devise another model that will give them the best solution for their problem.

This project aims to help the medical team at Stanford to accelerate their research in developing ways to stabilize the RNA molecules. The model plays a part in the development of the vaccine but does not devise a standalone vaccine itself [3.].

As a member of the BCS I have integrity and show competence, but I know I don't know everything, that's why I continuously learn and grow and never take on tasks that I don't have the skills and resources to complete [4.]. I'm an ambassador for the IT industry and use my voice to help promote it positively to the world. I support my IT colleagues and other members in their growth both personally and professionally [4.].

CHAPTER 8 – BIBLIOGRAPHY

1. Coronavirus disease (COVID-19). (n.d.). [online] Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200805-covid-19-sitrep-198.pdf?sfvrsn=f99d1754_2.
2. Carnahan, R. and Mishra, S. (n.d.). Coronavirus: A new type of vaccine using RNA could help defeat COVID-19. [online] The Conversation. Available at: https://theconversation.com/coronavirus-a-new-type-of-vaccine-using-rna-could-help-defeat-covid-19-133217.
3. kaggle.com. (n.d.). OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction. [online] Available at: https://www.kaggle.com/c/stanford-covid-vaccine [Accessed 6 Nov. 2020].
4. Bcs.org. (2019). BCS Code of Conduct. [online] Available at: https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/.