**ML Project - Midterm Report Draft**

**Cheating Charon: Predicting Excess Mortality Due to Heat Waves in Vulnerable  
Groups in Germany**

**Abstract:**

An abstract should concisely (200-250 words) motivate the problem, describe your aims, describe your contribution, and highlight your main finding(s). Given that your project is still a work-in-progress, it’s OK if ‘your contribution’ and ‘your findings’ are things you’re still working on.

[1](https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(17)30156-0/fulltext#bib1) The Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), the leading international body for the assessment of climate change, has established that anthropogenic emissions of greenhouse gases represent the dominant cause for the warming of the planet. Scenarios of climate conditions depend therefore on current and future trajectories of greenhouse gas emissions, mainly determined by socioeconomic development and climate policies.

[3](https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(17)30156-0/fulltext#bib3)

[2](https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(17)30156-0/fulltext#bib2)

*Predicting the consequences of our changing climate for public health is a crucial area where machine learning can assist policy makers. As the effects of man-made climate change are felt, excess deaths due to extreme temperatures and weather events are expected to rise. However, evidence of this direct impact at a country-level is limited, largely because of complexity of modelling and projecting the epidemiological relationship between populations and their habitats in their totality.*

*As machine learning models and methodologies mature, the scope of their predictive capacities increases. For our project, we have chosen to test and compare predictive machine learning models to anticipate the number of excess deaths in Germany due to heat waves under two distinct emissions scenarios. Our findings highlight the difficulties in generating reliable models and the range in the fit, precision, and robustness of machine learning models in predicting excess mortality. Moreover, our project demonstrates…*

*High-emissions scenarios predict an average increase in surface temperature between 2·6°C and 4·8°C by the end of this century.*

**Keywords:** Heat-related mortality Excess deaths Climate change Projection Adaptation

**Proposed Method:**

Climate change can directly and indirectly impact human health in multiple ways. Increasing food insecurity, the spread of disease vectors, armed conflicts, and extreme weather events can all be attributed to our changing climate. Aside from these indirect effects, the health consequences directly associated with variation in outdoor temperature. For the sake of feasibility, we have chosen to focus on increasing air temperature as a key factor in climate change-related mortality. The public health consequences related to variation in air temperature have been explored in several studies and our project builds on this existing research. In particular, the work of Honda et al in …

[6](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib6), [7](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib7), [8](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib8), [9](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib9), [10](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib10), [11](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib11), [12](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib12), [13](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib13)

Several studies have focused on the health consequences directly associated with variation in outdoor temperature, predicting an increase in heat-related mortality and morbidity, and—when considered—a concomitant decrease in cold-related mortality.

We collected observed daily time series of mean temperature and mortality counts for all causes or non-external causes only, in periods ranging from Jan 1, 1984, to Dec 31, 2015, from various locations across the globe through the Multi-Country Multi-City Collaborative Research Network. We estimated temperature–mortality relationships through a two-stage time series design. We generated current and future daily mean temperature series under four scenarios of climate change, determined by varying trajectories of greenhouse gas emissions, using five general circulation models. We projected excess mortality for cold and heat and their net change in 1990–2099 under each scenario of climate change, assuming no adaptation or population changes.

Impacts on human health can occur through multiple pathways.[4](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib4), [5](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib5) In addition to indirect effects mediated, for instance, by the spread of disease vectors, increase in food insecurity, and migration and conflicts, direct effects are expected from the increase in extreme weather events such as floods, droughts, and heatwaves.[1](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib1), [4](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib4) However, evidence on this direct impact at the global scale is limited. This is mainly due to the complexity of modelling the epidemiological relationships, characterised by differential patterns of non-linear and lagged effects associated with heat and cold, and to limitations of previous location-specific or country-specific assessments to capture the heterogeneity of the risk across different populations and climates.[14](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib14), [15](https://www.sciencedirect.com/science/article/pii/S2542519617301560" \l "bib15) Questions also remain about the extent to which expected decreases in cold-related mortality can offset the increase in deaths caused by heat. These issues make it difficult to draw a comprehensive picture of the direct impact of climate change across regions of the world and under different scenarios. This evidence is nonetheless crucial to develop coordinated and evidence-based climate and [public health policies](https://www.sciencedirect.com/topics/medicine-and-dentistry/public-health-policy" \o "Learn more about public health policies from ScienceDirect's AI-generated Topic Pages).

*Since the fundamental goal of our project is to predict how climate change, in this case defined as increasing temperatures and heat waves, can affect the mortality rate in vulnerable populations within Germany, there are several important variables to consider for our possible machine learning models. In our view, the following list covers the basic necessary data for those models:*

* *Population growth rate*
* *Actual and predicted population under 5 years old*
* *Actual and predicted population 65 years and older*
* *Mortality rate and main causes of death for German population under 5 years old*
* *Mortality rate and main causes of death for German population 65 years or older*
* *Temperature forecasting for Germany for Germany in line with the IPCC high-emissions scenario Representative Concentration Pathway (RCP) also called RCP8.5*
* *Temperature forecasting for Germany in line with low-emissions scenario, also called RCP2.6*

*The project will consider two distinct, yet credible climate scenarios in which the degree to which temperatures increase will vary in its effect on the mortality rate of our target populations. These two scenarios are determined by net CO2 emissions and follow the forecast predictions in the Intergovernmental Panel on Climate Change (IPCC)’s Representative Concentration Pathways (RCP).*

*The pathways describe different climate futures, all of which are considered possible depending on the volume of greenhouse gases (GHG) emitted in the years to come. The two scenarios we have chosen are as follows:*

* *A low-emissions scenario, also called RCP 2.6*
* *A high-emissions scenario, also called RCP 8.5*

*Although there are other RCP trajectories, we have chosen RCP 2.6 as a realistic best-case scenario and RCP 8.5 as an absolute worst-case scenario for the climate. The considerable difference between these two foreseeable scenarios will allow for greater variation in the prediction variable in our models. Furthermore, the contrast between the two outcomes will – we hope – make the consequences of doing nothing to reduce CO2 emissions and mitigate climate change more strikingly obvious.*

*From both a technical and a non-technical standpoint, our project will be a success if we can:*

* *develop machine learning models that can predict the excess deaths in our target groups in line with a baseline projected by the WHO,*
* *we can visually represent the differences in terms of mortality between the two chosen climate scenarios,*
* *if we can compare & contrast the efficacy and interpretability of our chosen ML models*
* *reach high accuracy and precision scores (e.g., ranging from 0.6 to 0.8 in the relevant metric) in our chosen models.*

*Owing to the clearly determined pathways in either scenario, we will not need to predict temperature increases, but rather attempt to predict the outcomes of the temperature increases in terms of mortality rates. As we are considering the effects of two scenarios on two target groups using likely three different models, one key successful outcome of the project will be twelve comparable predictions.*

*We are currently considering using at least two particular machine learning models, namely logistic regression and random forest classifier. If at all possible, we would also like to use a time series model, but need first to investigate our data further before choosing a particular time series model. The models will be trained using the mortality rate data, demographic data, and the mean temperature predictions.*

*The key evaluation metrics for our logistic regression model will be Explained Variance, Mean Squared Error, and the R2 coefficient, whilst the metrics for the random forest classifier will very likely be Prediction, Accuracy, and Recall. However, these metrics may change as we become more accustomed with the requirements of the project.*

*We anticipate that the data processing requirements will very likely vary from model to model, increasing the overall time we spend on this phase of the workflow. However, as our pedagogical goal is to learn as much as possible about the differences between the various models, we believe that this effort will be worthwhile.*

*After the data wrangling process, we will separate all our data into training and testing data, using a 70\% and 30\% split, leaving aside the testing data and working solely with the training data. Our second step in the process will be feature engineering, such as scaling the data to normalize the range of independent variables and performing the necessary transformations.*

*We will also tune the parameters for each specific model referencing their particular guidelines to achieve the highest accuracy and precision scores: solver, penalty, C (regularization strength) for logistic regression, and min\\_samples\\_leaf, n\\_iter, cv for random forest classifier. Ideally, we aim to achieve a score ranging from 0.6 to 0.8 in the key accuracy metrics.*

*Once the parameter tuning testing is done, the remaining 30\% of the testing data will be introduced to the model so it can predict the mortality rate and temperature. This process will be repeated for both emissions scenarios. Our fourth step will be to import GridSearchCV to do cross-validation of the data.*

**Experiments:**

**Data:**

Temperature and meteorological data were obtained from NASA. Mortality data and population projections for Germany were obtained from the Federal Statistical Office of Germany. We also obtained datasets on the causes of death amongst our age groups from the Federal Statistical Office of Germany. This dataset was used to derive a ratio of the citizen who died due to environmental exposure in our historical data to the German population as a whole. The method uses a series of regression equations that quantify the current and historical relationships between mortality and a set of independent variables.

The datasets we used for our project so far include:

* Actual and predicted population of Germany
* Mortality rate and main causes of death for German population
* Temperature forecasting for Germany for Germany in line with the IPCC high-emissions scenario Representative Concentration Pathway (RCP) also called RCP8.5
* Temperature forecasting for Germany in line with low-emissions scenario, also called RCP2.5

We have also had to account for a class imbalance in our original datasets. Originally, we intended to analyze the effects of climate change on young children (under the age of five) and old people (over the age of sixty-five). However, in our data wrangling, we found that there more observations for older people, particularly in the population projection datasets, than young children. To address this imbalance, we sliced our populations into four distinct age groups that cover the entire German population, both actual and predicted. The groups are as follows:

* 0 – 25
* 26 – 50
* 51 – 75
* 76 – 100

Karon\_clean: 2003 – 2021

Karon\_ data: 2022 – 2050

**Datasets:** temperature, mortality, projections, population

* no ratio for the death

*For this project, we will draw on historical research into the effects of heat waves, including the heat wave in Paris in 2003\cite{poumadere2003HeatWave2005}. Additionally, we will use existing research models for a general methodology and baseline for evaluating our approach. We will use data from the Federal Statistical Office of Germany, World Health Organization (WHO), European mortality database, and Eurostat.*

*We will use an Agile model to manage our project work, assigning different tasks according to our respective skill sets and interests, and each team member will be responsible for a particular data collection relevant to the overall project.*

**Evaluation method:**

**Experimental details:**

Linear model

**Results:**

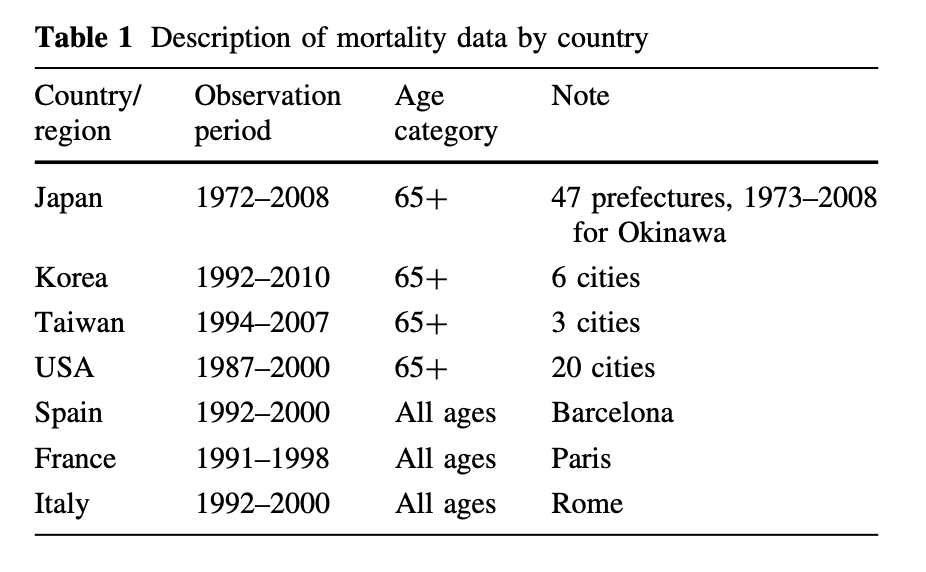
* Comment on quantitative results

**3. Future Work**

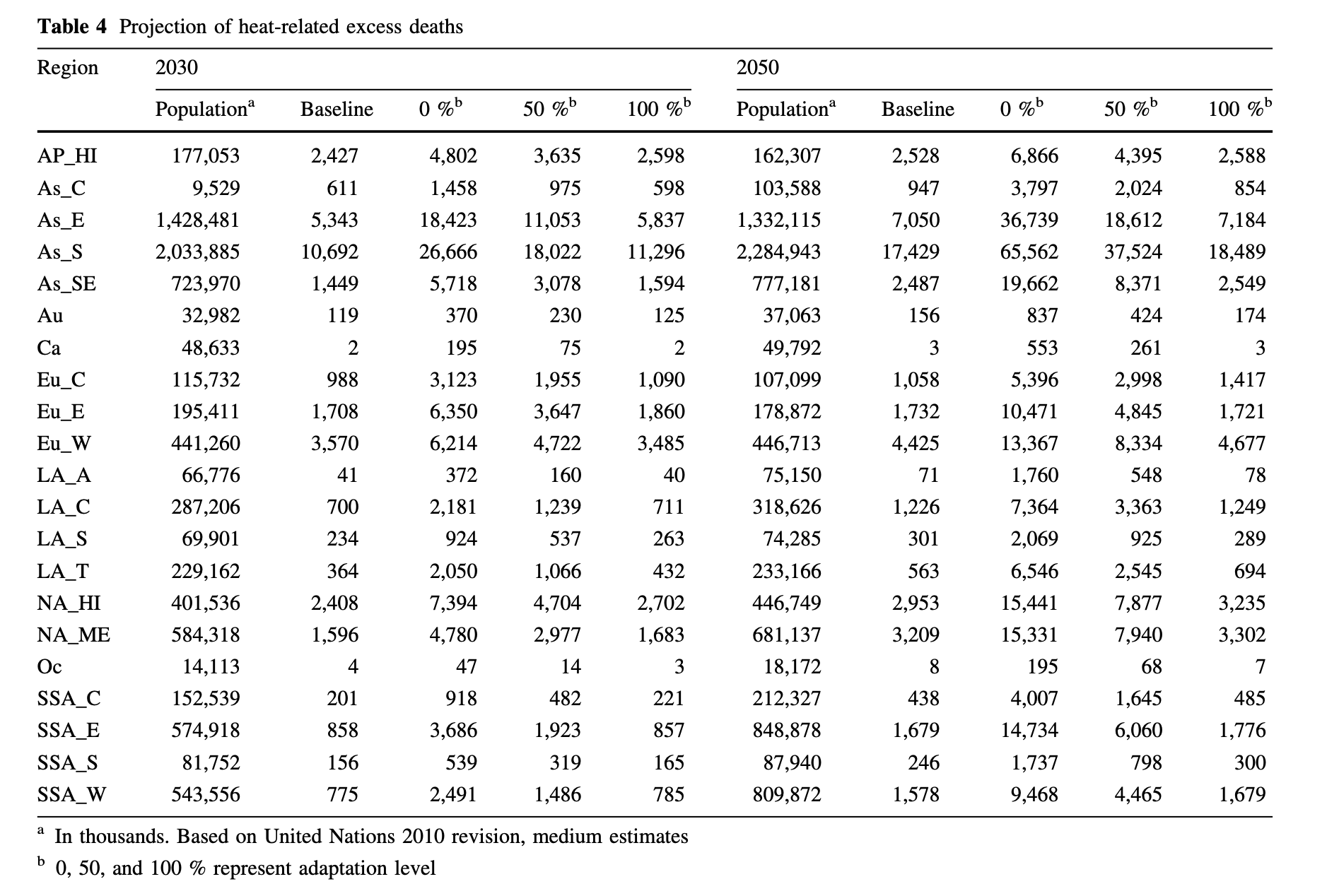
* Got a baseline model, now need to tweak it
* Can we identify hyperparameters that we want to tune?
* How do we intend to improve the fit, precision, and robustness of the proposed model have been improved from the previous projection
* **Differentiating relative risk between age groups** - We need to differentiate the relative risk due to heat temperatures for each population group as elderly people are at greater risk from heat exposure than younger or middle aged people. In Germany’s case, this is particularly important because the relative proporition of elderly people to young or middle aged people is expected to increase considerably in the coming decades, meaning that a commensurate rise in excess mortality should be expected.
* **Adaption** – our model assumes no adaption to increased outdoor temperature on the part of the German population. This is a weakness of our current model.
* *Because there are excess deaths due to heat even at present, the excess number of deaths attributable to climate change is calculated using the ‘‘counterfactual method.’’ The idea is that, with identical annual mortality rate, what is the difference between the number of excess deaths in the baseline case and that in the altered climate case?*

**Ideas:**

* Include a descriptive table of our population and projection data, i.e. this



* Include a breakdown of adaptation, like this:



***From Honda:***

**Problems with Honda’s original method:**

Although the above projection was one of the first projections of climate change impact on heat-related mortality, there were some weaknesses to the method, including (1) the applicability of the OT estimation method to other regions of the world, and (2) that the risk function was categorical. In this paper, we solve problem (2) by using a nonparametric risk function. This should be a substantial improvement, because the average of the tem- peratures in the highest temperature category was actually very close to the lower boundary of the category and huge underestimation was expected for the risk of this category.

**Problem 1: Time lag**

Another concern is the temporal pattern of exposure– mortality effects. This can be regarded in two ways [8]. One is the carry-over effect (lag effect), and the other is mortality displacement or harvesting. The former effect has been considered not to last long for heat-related mortality. In the latter case, heat just kills people who would die in a very short period of time anyway, and avoiding heat would extend the lives of people for only a short period of time. In the case of the 2003 heat wave in Europe, however, there was little evidence of mortality displacement [9]. Here, we address this lag effect using a distributed lag nonlinear model developed by Armstrong [8].

**Method/Calculation:**

The above risk function curve used the relative risk, and we need to estimate the excess number of deaths using this relative risk. Because the reference of the relative risk is the risk at OT, we can convert the relative risks if we can calculate the number of deaths when the temperature is the OT (NDOT). For this purpose, considering that the available information is the annual number of deaths (NAN), we need the following formula:

NDOT 1⁄4 ðNAN=365:25Þ RMOT=RMav;

where RMOT is the relative risk at OT and RMav is the average daily relative risk. (NAN/365.25) yields the daily average number of deaths. We obtained both RMOT and RMav using the Japanese dataset. Based on the Japanese dataset, the average and standard deviation of RMOT/RMav were 0.88 and 0.014, respectively.

**4. References**

* **Honda et al,** Heat-related mortality risk model for climate change impact projection

**EXCESS:**

*Germany’s population, like many in Europe, is ageing rapidly and elderly people (those aged sixty-five and above) are particularly vulnerable to the health effects of heat waves.*

The midterm project report should be 4 pages long (not counting references), and a maximum 10 references}. The report should contain the sections that are already provided in this paper. It forms the basis of the final report with the same structure. Please check out the text in these sections for further information.

Your midterm milestone will be graded on the following criteria:

* Progress: Has the team made good progress on the project? You should have done approximately half of the work of your project.
* As a minimum, your milestone should show that you have setup your data, baseline model code, and evaluation metric, and run experiments to obtain some results (assuming you are doing a typical model-building project). Other than this, `good progress' depends on various factors (e.g., whether your model is implemented from scratch or based on an existing codebase).
* Understanding: Does the milestone show a strong understanding of its problem, tasks, methods, metrics, and research context?
* Writing quality: Does the milestone clearly communicate what you've done and why, providing the requested information, to an appropriate level of detail (given the page limit)?
* You will receive some brief feedback on your milestone. Feedback may contain helpful suggestions for your project (e.g., try a particular method, read a particular paper) and/or warnings about your project plan (e.g., if your plans are too ambitious or not ambitious enough), and how you could improve your technical writing (e.g., adjustments to clarity, level of detail, formatting, use of references).
* \end{itemize}

Here are some other things you can do to improve your technical writing:

* Look carefully at several ML papers to understand their typical structure, writing style, and the usual content of the different sections. Model your writing on these examples.
* Think about the ML papers you've read (for example, the one you summarised for your proposal). Which parts did you find easy to understand and why? Which parts did you find difficult to understand and why? Can you identify any good writing practices that you could use in your technical writing?
* Ask a friend to read through your writing and tell you if is clear. This can be useful even if the friend does not have the relevant technical knowledge.