ML for Healthcare (097248) – Project Report

Hierarchical ML model for ICU mortality rate prediction

Abstract

Clinical decision making is challenging because of pathological complexity, as well as large amounts of heterogeneous data generated by various medical instruments. In recent years, machine learning tools have been developed to aid clinical decision making. Models for early in-hospital mortality prediction in intensive care units (ICUs) especially, could help, for example, with decisions about treatment trajectory and resource allocation.

In this work, we propose a Hierarchical ML model for ICU mortality rate prediction. We believe that the ramification of certain health-conditions can aid the model's ability to predict the mortality rate of ICU patients during their hospital stay. Our approach is a novel take on an existing task – in-hospital mortality prediction, and one we strongly believe can alter the decision making process of ICU care and treatment trajectory.

We evaluate our approach on the task of in-hospital mortality prediction using the F1, precision, recall and AUC measures and while our results do not fully reflect the quality and potential of our work, we strongly believe that our model can support medical experts in their decision making. We know in our brains and hearts that our approach can provide novel insights into the importance of different clinical measurements and the added focus on certain diagnosis, in addition to the predicted outcome results.

Background and Related Work

Unsurprisingly, we find that both mortality prediction and ECG signals classification and analysis, are vastly explored tasks in the decision making field of research when they stand on their own. However, it seems that the combination of the two is still somewhat an unexplored and uncharted territory. We came across the idea while researching for our project's topic after reading and analysing large amounts of papers and articles related to the field of health and medicine. We wanted to deploy a model that resembles a decision tree's way of thinking and the natural thought process of the human mind (differential diagnosis i.e. the process of elimination). Therefore, we designed a hierarchical model that branches according to different patient's diagnosis, such as cardiac conditions, sepsis, pulmonary and etc. Hopefully, this approach will be able to improve the decision making process of caregivers when selecting the treatment path and will decrease the in-hospital mortality rate among ICU patients.

The main paper we will focus on in the scope of this project is the **Early hospital mortality prediction using vital signals**.
The article proposes a method of predicting mortality using features extracted from the heart rate signals of patients admitted to the ICU during the first hours of their stay. The method was tested on the MIMIC-III dataset using different classifiers such as KNN, SVM and Random Forest. Their work is exceptional in its simplicity and is promising especially in patients admitted to the CCU. Therefore, we believe that quantitative, low dimensional features could accurately predict serious cardiac conditions which are beneficial for our hierarchical model.

Another item that is relevant to our project is the following paper – **Towards Understanding ECG Rhythm Classification Using Convolutional Neural Networks and Attention Mappings**. The paper focuses on bridging the gap between machine learning research and delivering direct care to patients at their bedside i.e. providing clinicians with tools to better interpret the classification labels (what kind of ECG waveform) and scores (level of confidence in the classification) generated by the ML tools. The authors present a model that classifies ECG waveforms as either Normal Sinus Rhythm, Atrial Fibrillation, or Other Rhythm using a convolutional neural network (CNN).

The described papers provided us with a good starting point and allowed us to build a good foundation for our model. We based our hierarchical aspect of the model on the first article, we thought that since a closer inspection based on vital signals can lead to better mortality prediction, there is no apparent reason why it should not be alleviated to include more medical conditions and allowing for a more emphasized medical check. The second paper gave us the proper tools to understand the core of our model, gap-bridging between machine learning based research and direct health care decision making.
The use of machine learning tools and models supplied us with the knowledge and ability to handle large quantities of data and produce unthought of conclusions. The strength of ML models is in finding patterns that usually could not be found by any other measures. Potentially, without using the hierarchical structure of diagnosis, the model could find complex connections between different organ systems, somewhat non connected symptoms or rare medical conditions. These complex connections are often not easy or even impossible to find, when limiting yourself to the clinical reasoning of human physicians. We therefore strongly believe that the combination of a ML approach with a hierarchical model can accomplish far more together and provide a break-through in the world of clinical decision making.

Data

In our work we utilized the MIMIC-III dataset, specifically the clinical and waveforms databases it is composed of. MIMIC-III is a large, freely-available database[1], comprising de-identified health-related data associated with over forty thousand patients who stayed in ICU of the Beth Israel Deaconess Medical Center between the years 2001 and 2012. We used a significant sum of the information included in the database, such as demographics, vital sign measurements made at the bedside, laboratory test results, mortality (in the hospital) and etc.
The decision to pick MIMIC-III as our leading and only cohort to base our model on is heavily based on the fact that MIMIC supports a diverse range of analytic studies, specifically, clinical decision-rule improvement and encompasses a diverse and very large population of ICU patients which permits vast examnasion opportunities.

Clinical database The clinical database was collected from two different information systems - CareView and MetaVision. Therefore, it differs in scale, frequency and recordings between the data sources. An important feature we particularly enjoyed in this database, was the appearance of all hospital admissions per patient, for it added a layer of background history we could take advantage of in our learning. We exploited this feature in our analysis, where we calculated the stays average for each patient so that we achieve better understanding of his status. In spite of the bias this averaging method introduces to our model we felt that we needed the high-level outlook of a patient for our task.

Besides the aforementioned pre-processing, we also deployed filtering elements to our patients. We chose to concentrate on adults between the ages of 18 to 89 for our cardiac branch seeing as in our opinion, this age range can properly represent both cardiac and non-cardiac patients while maintaining randomness and avoiding selection bias. Another feature we took into consideration is the ethnicity column, we checked that the field is indeed being utilized and not filled in a somewhat exploit fashion. Finally, we made sure our clinical data is gender-balanced i.e. there is a logical ration between the appearance of males and females in our sample data. By doing so, we successfully avoid the inclusion of gender bias in our model and preserving our model's integrity.

Waveform database The waveforms database contains thousands of recordings of multiple physiologic signals i.e. waveforms, and time series of vital signs collected

[1] Requires a short procedure of credibility checking prior to gaining access to use it.

from patients monitors in ICUs. The recorded signals change among patients according to the medical staff decisions and so does the recordings length.

The size of the waveforms dataset is quite large (approx. 2.4 TB), thus, we implemented an iterative approach for accessing the records. Then, each record that was loaded goes through preprocessing followed by feature extraction. The waveforms we focused on were ECG, Pulse and Heart Rate, all found in the waveform dataset.

preprocessing

Multiple ECG records for one ICU stay Most ICUs stay ECG records are divided into several segments. It is common that the records are not consecutive and rather have gaps between them, meaning that the recording was not continuous. In order to overcome the problems that could rise from concatenating segments that have sampling gaps between them, we followed **Gholinezhadasnefestani 2012** and added synthetic zeroes to account for the missing samples. As we base our results on measurements that occured on the first 4 hours in the ICU stay, this method ensures validity of the time interval of the samples. In addition, we discarded all measurements that took place later than 4 hours from being admitted to the ICU.

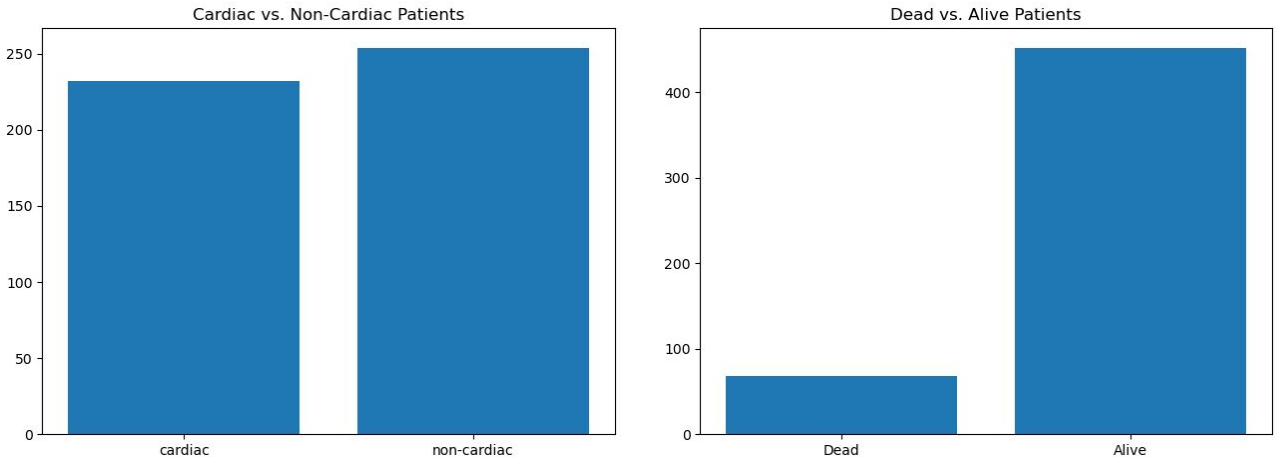
Heart Rate and Pulse Frequency Bias As suggested in **Sadeghi 2018**, we aimed to avoid sampling bias that might occur due to differences in sampling frequency along the different records. Moreover, we implemented oversampling and downsampling procedures to ensure a signal frequency of 1Hz as was described in the aforementioned paper. We tried several methods that to our disappointment took up a lot of execution time and eventually resulted in us deciding to use naive implementations.

For ***oversampling***, we randomly choose an index to add a sample and add the average of the chosen index's value and the following value.
Similarly, for ***downsampling***, we randomly choose an index and merged it with the following index to their average value.
We believe that the frequency shifting schemes we deployed will work well as they do not add much information to the already available information, while correcting some of the sampling bias.

Heart Rate and Pulse Filtering Noisy signals are a common issue of signal processing and might lead to significant bias to the level of false discoveries. We

aimed to avoid this problem using the online available python library *hrv*[2]. We deployed a moving average filter followed by a threshold low pass filter. We meant to use a FIR LPF, as suggested in the Sadeghi 2018 article but due to implementation struggles we decided to follow the aforementioned LPF. The choice of a low pass filter rather than a high pass filter is based on a blog called 'EMS 12-Lead'[3], where it is stated that heart rate and pulse signals suffer more from spikes of high values than low values. The aforementioned was indeed proven to be true in our case, in the contents of both the data and task we deployed.

The following figures are general analysis we pulled on our sampled data to reflect the balance, or lack thereof, in the fields in question. As we can see, the hierarchical branch i.e. the cardiac versus noncardiac partition is almost exact. Yet, the same cannot be said on the in-hospital mortality prediction field, seeing as the number of patients that have passed while admitted in-hospital is drastically lower when compared to patients that have not passed while admitted to the ICU.



Methods

In our work, we focused on two aspects of the MIMIC-III dataset - the clinical database and the waveforms database as mentioned in the previous section. In light of that, throughout most of our work, we took lead in two very different paths simultaneously and prepared them both to the moment where they will be united and hopefully work in perfect harmony to achieve the best results for our model and defined task.

[2] Documentation for hrv library can be found here: https://github.com/rhenanbartels/hrv
[3] A link to said blog post can be found here: http://ems12lead.com/2014/03/10/understanding-ecg-filtering/#gref

Our work consists of two main courses of action – iterative and nonrepetitive. The waveforms feature engineering was done iteratively while the training of the models was executed in a nonrepetitive manner due to the utilization of a feature file that was created once.

Throughout our work we used both a virtual-machine provided by the course staff and the Technion, and our personal computers (PC and laptop) to create our models. Most of the work was done locally on our personal computers and consisted of pre-processing, feature extraction, models training and testing, and etc. the main use of the VM was to query MIMIC-III and extract our data from their server.

The first stage of our work consisted of preprocessing of the data i.e. filtering and cleaning it, and is described in detail in the Data section above. The Waveforms database was used for the hierarchical aspect of our model and the clinical database was used for the mortality rate prediction task.

At first, we wanted to implement the cardiac branch in our hierarchical model with the ECG signals and utilize the heart rate and pulse signals as reinforcements. Due to excessive running time we concluded that we can still manage a decent branching using only our reinforcement signals that are much shorter and therefore do not require a lot of execution time i.e. not including the ECG signals. Nevertheless, we still implemented all necessary analysis and functions for the ECG signals for future work purposes. Another approach we had was the incorporation of deep learning methods to handle the ECG signals and enabling their incorporation into the model, but alas we are not as skilled in the field and could not manage the combinations of signals to our model.

One of the main problems we faced while working with the clinical data is the problematic lab-value entries. As will be discussed later on in the Feature design and engineering sub-section, the lab results are the main basis of our clinical features and thus this issue needed a quick and suitable solution. There is a considerable amount of values that are inconsistent, range-wise, junk-values that are uncertain, values that are given in the 'greater or smaller than' form and values that simply state weather a clinical device used by the patient was calibrated, for example the entry *'cs'* that stands for control solution in glucose testing. Our methods to deal with this missing and uncertain values are as follows, we converted lab results values to their correct measurements when they were wrongly scaled, when the form 'greater/smaller than x' form was used, we replaced it with the value of $(x + 0.1 \cdot x)$,

and lastly, junk-values and device calibrations were erased from the data. In addition, we performed imputation over all missing data fields of the clinical features. We used the normal naive model to actuate the imputation.

As mentioned in the Data section above, sub-section 'heart rate and pulse frequency bias' we experimented with several methods to achieve the best result and avoid the sampling bias. The methods mentioned are FIR, smoothing priors, butterworth, IIR and variations of them. Their utilization seemed logical due to their accessibility, they were already implemented in the scipy and hrv libraries in python, which to us was a relief since we are not experts in dealing and processing of signals. However, these methods proved unfit to our particular necessities, they exceeded in their runtime far above what we were willing to sacrifice and the smoothing priors method seemed to provide too strong smoothing to prove useful and overall disappointing in light of the promise they were sided with. All those led us to the ease of a more naive implementation we came across - threshold LPF which is also implemented in the hrv library.

For the in-hospital mortality prediction aspect of our model, we chose to experiment with the following machine learning models - KNN (for k = 1, ..., 20), random forest, decision trees, SVM. We strongly believed that these models are best equipped to handle our prediction task and our data format. However, after we finished running all the models on our training and test samples, we were proven otherwise. The random forest model achieved the best results and we concluded that we should involve only those.

Class Imbalance and Data Generation

The dataset suffers from severe class imbalance of deceased versus living patients (when considering the in-hospital deaths). Our sampling method puts an emphasis on said imbalance and that is reflected in our sample dataset at its worst (approx. 18% in the filtered data as opposed to 7% in our sample). While we have made several attempts to train models without dealing with this problem, they all resulted in poor performance which led us to the implementation of a data generating process to enrich our data. In order to maintain the unique properties of deceased cardiac/non-cardiac patients, we implemented an imputation scheme that was applied separately for the groups of patients. As the imputation method is not the main concern of our work, we used a simple gaussian data generating process - we randomly sampled features from a multivariate gaussian distribution with a mean vector of feature averages over the groups of patients and a diagonal covariance matrix with the features' variances. We decided not to increase the number of

patients using synthetic patients to more than twice the original size in order to maintain integrity of our work.

Feature design and engineering

In order to establish our primary branch, the cardiac diagnosis branch, we had to obtain the correct labels for our patients. We took on board the SERVICES table in the MIMIC-III clinical database. Said table, describes the service that a patient was admitted under, in two distinct fields - current service and previous service. We labeled a patient as a cardiac patient if sometime during his stays in the ICU he was admitted under a cardiac related service[4]. We filtered out patients that had no record of the services they underwent and attempted to achieve a balanced ratio of cardiac and non-cardiac patients in order to avoid imbalanced classes in our prediction task.

The next rational step was to assert our clinical features for our patients. For this step we mainly used the lab tests and chart measurements (logged at bedside) based on the physionet challenge, while excluding demographics to avoid social bias. These measurements were all numeric and we made them work overtime in the name of our task. Each measurement was taken into account in its original value and more. We also calculated the difference of the recorded value from the normal range of the test[5] while taking into account the sign of the difference i.e. we take into account whether the recorded value exceeds the normal range (positive sign) or the opposite (negative sign). If the recorded value is within the normal range, the value inserted is zero. Moreover, we checked the validity of the received values according to the Glasgow 15 (GCS 15) measure, that states the level of neurological function of an admitted patient. This measure is kept twice per patient - the actual GCS value and its flag form i.e. is it normal or not. If the value measured equals 15 i.e. the patient is considered relatively functioning, the feature will receive a value of 1 and 0, otherwise. Following that, we dealt with missing values in the following manner - we replaced the missing value with the average of the test over all of the ICU stays of the patient while inserting normal noise with the test's variance. In our opinion, this course of action will be better suited for the model and surpasses absolute value. Last but not least, all the calculated features for a patient over his different ICU stays were averaged in order to gain a broader perspective of a patient's medical status.

---

[4] If the value of the prev_service or curr_service was in the following list; {CMED, CSURG, TSURG, VSURG}, a patient was considered cardiac.

[5] The normal ranges we based our calculation on can be found here: https://annualmeeting.acponline.org/sites/default/files/shared/documents/for-meeting-attendees/normal-lab-values.pdf

Later, we worked on the features of the heart rate and pulse signals. The following features are almost identical to the ones in **Sadeghi 2018** and we used them to capture different properties of the signal's distribution. The underlying assumption is that patients with irregular HR and/or Pulse distributions are more likely to suffer from cardiac conditions. We show the mean values of the features in both classes in order to indicate the segregation capability based on the following features;

Feature	Cardiac Patients	Other Patients
Maximum	100.91	105.54
Minimum	64.69	70.85
Mean	83.07	88.28
Median	82.97	88.01
Mode	82.73	87.37
Standard Deviation	4.75	4.95
Variance	33.74	35.83
Range	36.22	34.69
Kurtosis	16.68	5.84
Skewness	0.14	0.172
Average Power	6.71*10^8	8.24*10^8
Autocorrelation	7188.38	8195.01

In a similar fashion of that seen in **Sadeghi 2018**, we too aimed to obtain a measure that represents the signal's behavior using different features. Unlike the paper, instead of using the Energy Spectral Density(ESD), we chose to use the autocorrelation value, which is highly correlated to the article's measure. The said correlation is directly related to the fact that the ESD is the Fourier transform of the autocorrelation biased estimator.

It is important to mention that due to a mistake made at the time of implementation, the signal's power, in the *average power* measure, was calculated instead of the average power i.e. the calculation is missing the deviation by the total sum. We

suspect that some of the problems of the model were caused as a result of this mistake as it introduces bias towards longer signals, seeing as naturally longer signals have a bigger power value.

Results

For the evaluation of our work, we chose the following measures as we have concluded, they fit best for the task at hand and were easy to implement. Our chosen measures – recall, precision, F1 and AUC, handle imbalance data well and were therefore selected.

Non-Hierarchical Models				
Prediction Task	**Precision**	**Recall**	**F1-Measure**	**AUC**
Cardiac vs. Non-cardiac	0.538	0.598	0.565	0.637
In hospital Mortality	0.5	1	0.666	0.967

The cardiac versus the noncardiac prediction has accomplished results that are slightly better than those of a random classifier. While the sampled data was properly balanced, our model failed to segregate between the classes. We hypothesized among ourselves and came up with a combination of reasonable explanations – features extraction was poorly executed in regards of the task, uncertainty of the class labels and the over simplicity of the chosen models. The table above contains the measures of evaluation on non-hierarchical models. We can see that the AUC measure performed best however it might be due to the fact that it is considered optimistic when faced with severely imbalance classification task. As for the precision results, upon manual inspection we came to realize that most patients the model gave a false label for, are diagnosed as cardiac.

Hierarchical Models				
Prediction Task	**Precision**	**Recall**	**F1-Measure**	**AUC**
Cardiac Patients	0.2	1	0.166	0.925
Non-Cardiac Patients	0.333	1	0.25	0.985

As can be seen in the above table, the hierarchical models appear lackluster in comparison to the non-hierarchical models. This results might be another

repercussion of the poorly sampled data (both in size and balance), and under the right conditions will bloom and perform in an incomparable way to the nonhierarchical models. As for the recall measure, we can see that it is impeccable when faced with the mortality classification which is promising as it indicates that our models succeed in deciphering the patterns that lead patients to die in hospitals. However, the precision measure shows that our model, while succeeding, is highly pessimistic and therefore classifies still-living patients as patients that are at the brink of death. This might be due to the fact that our model learns to identify critically-ill patients rather than actually predicting death.

It seems that the non-hierarchical models are able to utilize the heart rate features in their training process and by doing so, they learn more data (processed more features) about the patients which eventually led to them surpassing the hierarchical models that only learned the clinical features.

Regarding the F1 measure, seeing as it is based on the precision and recall measurements, it too is lacking, due to the polarity in the values received. Overall, it can be said that the hierarchical models show promise in the future and might achieve the desired results when it comes to decision making algorithms.

Limitations of the Study

While we have tried our best to present a study that holds no flaws, every good study has its strengths and drawbacks. The limitations of our study include the challenges we faced along the way and causal inference struggles that were unavoidable due to using observational data. We believe these limitations will not hinder us in any way as we still managed to compile a wonderful thought out model and will allow us to grow and improve our presented concept in the future, should we choose to.

Let us first review the held back that originated in the data we worked with. The MIMIC-III dataset is compiled of several databases of different sources as we mentioned in the Data section above, which has led to some decision making regarding the tables that were taken into account in this project. Moreover, there are problematic lab results entries i.e. we had quite a lot of missing/uncertain values which in turn has led us to manually insert mean results for the patients who fell under such values and has undoubtedly introduced bias into our results. Another issue we had with the data is related to our model and the requirements we needed to deploy it. Using both the clinical and the waveforms datasets of MIMIC-III was necessary to accomplish our hierarchical model, but so was labeled data (to perform

the cardiac branching), and unfortunately the ECG and heart-rate data did not come with labels that indicated their validity. We were forced to make adjustments that might have inserted bias into our work. The final limitation concerning the data is the sample size we used, which was relatively small, due to runtime issues and memory we had at our disposal.

Causal inference aspects might be a limitation to us, due to the use of pre-collected data that was not targeted and collected according to our wants and needs. Therefore, some confounders such as ethnicity, insurance and demographics might be confounders. The proof to our project's and proposed model's applicability is not yet full and whole, since there was only a single cohort on which we tested our approach.

Future Work & Discussion

Working on this project was a journey, we had to work in the shadow of a world pandemic, distanced from one another and find a way to combine our strengths and compile a project we can both be proud of.

Early hospital risk of mortality prediction in ICU units is critical due to the need for quick and accurate medical decisions. In our project, we wished to engineer a ML based hierarchical model for mortality rate prediction that will ease the process of treatment selection per patient in time-sensitive cases (ICU patients). Our work has a vast potential to better decision making skills and abilities under pressure for the medical staff, both senior and resident.

We are vastly proud of our model concept and implementation including their flaws (they are our own). We leave behind a revolutionary (we think), hierarchical model for ICU mortality prediction that in the right hands can be groomed to break into the market as a life-saving start-up (might be us, might also accept credit, who knows we'll see where life takes us). Finally, we received the all important answer to one of our main questions, can the notion of a simple ML model aid in our chosen task? And the answer is, it depends on the model (based on our experiments).

In our future work, we plan to apply our proposed method over other intensive care units cohorts, incorporating multiple diagnosis and branches in our hierarchy, in different orders as a means to better understand the cause of mortality and the weight of each diagnosis. The study can also be extended to use a deep learning approach and neural networks for better feature extraction using complex connections (rather than hand-crafted features). Machine learning tools, DL in

particular, are well-suited for the described method as they allow the consideration of large amounts of data while applying well informed and detail-oriented analysis to reach a classification and the appropriate visualization that explains the classification. Another direction is to explore the effect of different smoothing methods for handling the ECG signals for a better analysis and the same can be said regarding the imputation of the data. Due to the imputation method not being at the heart of our model, we used a basic method that satisfies our needs and more advanced methods can be deployed without doubt.

In conclusion, not only have we learned a lot when working on this project, we got to experience genuine hands-on work on real hospital-based data. We achieved great experiments and results that we may not be proud of, but can stand behind on the basis that even Thomas Edison has failed countless times before achieving his greatest accomplishment. As he said, "I have not failed. I've just found 10,000 ways that won't work".

Appendix

For your convenience, the git repository of our project, containing the documentation, can be found here,
https://github.com/benpili/Hierarchical-ML-model-for-ICU-mortality-rate-prediction

References

- Documentation for hrv library (used for the signal's preprocessing) can be found here: https://github.com/rhenanbartels/hrv

- The normal ranges for the lab results we based our calculation on can be found here:
https://annualmeeting.acponline.org/sites/default/files/shared/documents/for-meeting-attendees/normal-lab-values.pdf

- EMS 12-Lead blog:
http://ems12lead.com/2014/03/10/understanding-ecg-filtering/#gref

- Justification to the 0-padding of the signals:
https://www.researchgate.net/profile/Noel_Camilo-Castro/publication/261060425_Atrial_Fibrillation_ECG_Signal_Processing_for_QRST_Cancellation_Zero-Padding_Ver

sus_Time_Alignment/links/54885b090cf289302e3094af/Atrial-Fibrillation-ECG-Signal-Processing-for-QRST-Cancellation-Zero-Padding-Versus-Time-Alignment.pdf

- Johnson, A., Pollard, T., Shen, L. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35