Submission Template for ACM Papers

This is Short Title of the paper, used in page headers

This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word

First Author's Name[[1]](#footnote-1)\*

First author's affiliation, an Institution with a very long name, xxxx@gmail.com

Second Author's Name

Second author's affiliation, possibly the same institution, xxxx@gmail.com

Third Author's Name

Third author's affiliation, possibly the same institution, xxxx@gmail.com

Although there is no distinctive header, this is the abstract. This submission template allows authors to submit their papers for review to an ACM Conference or Journal without any output design specifications incorporated at this point in the process. The ACM manuscript template is a single column document that allows authors to type their content into the pre-existing set of paragraph formatting styles applied to the sample placeholder text here. Throughout the document you will find further instructions on how to format your text. If your conference’s review process will be double-blind: The submitted document should not include author information and should not include acknowledgments, citations or discussion of related work that would make the authorship apparent. Submissions containing author identifying information may be subject to rejection without review. Upon acceptance, the author and affiliation information must be added to your paper.

CCS CONCEPTS • Insert your first CCS term here • Insert your second CCS term here • Insert your third CCS term here

**Additional Keywords and Phrases:** Insert comma delimited author-supplied keyword list, Keyword number 2, Keyword number 3, Keyword number 4

ACM Reference Format:

First Author’s Name, Initials, and Last Name, Second Author’s Name, Initials, and Last Name, and Third Author’s Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock ’18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1. Introduction

The purpose of this work was to examine how external factors affect health care outcomes. The author’s previous efforts submitted for credit to the course (Ding, 2025) examined the ability to predict heart attack survival from specific test results and secondary diagnoses, and influence of socio-economic factors on health outcomes. I wanted to broaden this line of work to look at chronic conditions and socio-economic factors impact health care outcomes. The outcome should help support individuals and governments make decisions that can have positive impacts.

The electronic health records (EHR) in the MIMIC-III database (Johnson et al. 2016) were used as the source of length of stay (LOS) and survival outcomes of patient admissions to the critical care units of the Beth Israel Deaconess Medical Center. The MIMIC-III database includes some socio-economic information about patients, and rich information about the patients’ care during the stay.

Using this information, I demonstrate a workflow that predicts patient outcomes from their interactions with health care services based on their medical status at admission and the available socio-economic information. Patient diagnoses are separated based on their admission diagnosis and their chronic and acute diagnoses during their stay. This information is encoded with the socio-economic information into images for use in convolutional neural network (CNN) models for the prediction of LOS and survival.

People who have chronic problems can use this information about the potential negative effects of chronic illnesses on their health as a motivation to improve. Governments can use this information to determine how to invest in health education and services.

1. Data Preparation

The goals are to predict LOS and survival of patients after admission using the MIMIC-III database. I implemented a number of steps to prepare the data for modeling.

1. Age: MIMIC-III has known issues with the Age information. I truncated ages to values between 0 and 100.
2. LOS is explicitly determined from the admission and discharge date information.
3. ICD-9 diagnosis codes are simplified to the 3-digit main codes.
4. Admissions are limited to emergency and urgent admissions.
   1. ICD-9 Code Simplification

ICD-9 codes provide an alphanumeric code for any patient diagnosis (University of Michigan, 2018). ICD-9 is now out of date and the latest enhanced version is ICD-11. The MIMIC-III database contains ICD-9 codes only. There are 3 digit main categories and hundreds of subcategories within ICD-9. To reduce the number of inputs and ensure there are meaningful amounts of data for each input, I reduce the MIMIC-III diagnosis codes to their 3 digit parent codes. This was performed with an algorithm based on Ding (2025).

I also wanted to consider the combination of chronic conditions (i.e. diabetes, high blood pressure) on the acute condition (i.e. accident, heart attack) that had led to the actual admission. To do this, I used the Chronic Condition Indication mapping available at the United States Government Department of Health and Human Services website (DHS, retrieved 15 April 2025). I applied the same code simplification to this data so I could use it to map the processed MIMIC-III diagnosis codes to chronic or acute diagnoses.

* 1. Data Review

The main table of interest is the patients table. It contains the patient information and associated admission number. The admissions table provides the diagnosis codes associated with the total length of stay for that admission. Figure 1 shows the type of admission and the major diagnosis codes for the admissions.

Graphical user interface

AI-generated content may be incorrect. Chart, histogram

AI-generated content may be incorrect.

Figure 1. Histogram of the types of admissions and the major diagnoses codes for those admission.

In order to keep the data model simpler I selected emergency and urgent admissions. The newborn and electives can be expected to be very different and with elective admissions have little opportunity for intervention to change their outcome.

Now we can look at the primary diagnosis codes that result in admission. The initial model work tried to estimate LOS and survival for all these admissions. I also modeled the LOS and survival for the top 20 codes.

Chart

AI-generated content may be incorrect.

Figure 2. Histogram of the number of admissions in the data set for the top 20 primary ICD-9 diagnosis codes.

The dataframe columns, i.e. the variables for input to the model are: ‘LOS', 'DIAGNOSIS', 'DECEASED', 'chronic\_codes\_list', 'acute\_codes\_list', 'primary\_diag', 'age', 'INSURANCE\_Government', 'INSURANCE\_Medicaid', 'INSURANCE\_Medicare', 'INSURANCE\_Private', 'INSURANCE\_Self Pay', 'MARITAL\_STATUS\_DIVORCED', 'MARITAL\_STATUS\_LIFE PARTNER', 'MARITAL\_STATUS\_MARRIED', 'MARITAL\_STATUS\_SEPARATED', 'MARITAL\_STATUS\_SINGLE', 'MARITAL\_STATUS\_UNKNOWN (DEFAULT)', 'MARITAL\_STATUS\_WIDOWED'.

One dataframe was constructed with all the admissions selected as described previously, and a smaller dataframe was constructed with the same variables but only considering primary admission diagnosis from the top twenty primary diagnoses.

* 1. Data Structure for Modeling

I decided to use CNNs as models because one can encode a lot of information into an image and CNNs are very efficient on this type of data. I had previous experience that showed promise with this method with simpler network structures (Aldrin and Forsyth, 2019).

The data we can get from MIMIC-III includes a selection of ICD-9 codes for chronic and acute diagnoses. The total ICD-9 set is 1000 base codes with many subcodes. A complicating factor is that any admission can have any set of chronic and acute codes.For every patient admission, I encode a binary vector for chronic and acute codes separately. This means that my data input can handle any sets of chronic and acute codes. Any row corresponding to any code diagnosed for the patient is white, and all others black. For example, if code 410 is diagnosed, then the pixel in column 1 row 410 is coded white.

The third vector encodes age and the socioeconomic factors available in MIMIC-III. Rows 0 – 100 encode age, the row with the actual age is white. The next 12 rows are coded to the one-hot variables of insurance and marital status.

Note we could easily add more variables to this structure and still have a small image! Variables that are real valued could be scaled between 0 – 255 and encoded as a grey value on a single pixel. An example of one of the data images is shown below.

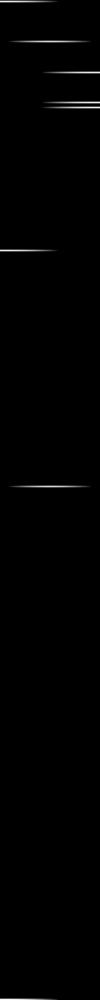


Figure 3. A scaled version of the input image, rotated 90 degrees. Pixels are stretched in the vertical for easy visualization.

1. Modeling
   1. Datasets

A pytorch dataset is constructed by first creating arrays for each patient with the pixel encoding as described in the previous section. A Pytorch tensor is created for each image, and the dataset operation returns the image tensor and the labels for LOS and survival.

* 1. Models

A simple CNN structure is created and used for the model in all the results. It is a typical construction, and the basic elements were derived from other coursework (Krahenbuhl 2025) but are easily found online also.

The code for the model creation and forward function are shown in the table below.

Table 1. Model code.

|  |
| --- |
| class EmergencyMultiTaskCNN(nn.Module):  def \_\_init\_\_(self):  super(EmergencyMultiTaskCNN, self).\_\_init\_\_()   self.conv1 = nn.Conv2d(3, 16, (5,1), stride=(2,1), padding=(2,0))  self.conv2 = nn.Conv2d(16, 32, (5,1), stride=(2,1), padding=(2,0))  self.conv3 = nn.Conv2d(32, 64, (5,1), stride=(2,1), padding=(2,0))  self.conv4 = nn.Conv2d(64, 128, (5,1), stride=(2,1), padding=(2,0))   self.fc1 = nn.Linear(128 \* 63 \* 1, 128)   # Two heads  self.fc\_los = nn.Linear(128, 1) # Predict LOS (regression)  self.fc\_deceased = nn.Linear(128, 1) # Predict Deceased (binary)   def forward(self, x):  x = torch.relu(self.conv1(x))  x = torch.relu(self.conv2(x))  x = torch.relu(self.conv3(x))  x = torch.relu(self.conv4(x))   x = x.view(x.size(0), -1)  x = torch.relu(self.fc1(x))   los = self.fc\_los(x) # Regression output (no activation)  deceased = torch.sigmoid(self.fc\_deceased(x))  return los.squeeze(1), deceased.squeeze(1) |

* 1. Loss Functions

My initial trials included a loss function that was weighted to LOS over survival. Previous work for the course had shown that survival could be predicted with 80 – 90% accuracy.

* 1. Training and Validation Results – All Primary Diagnoses

The results of the model training and validation are shown below for the data set including all the primary diagnoses. Training and validation results are shown in Figure 4 and prediction results in Figure 5.

Chart

AI-generated content may be incorrect.

Figure 4. Training and validation results by training epoch.

A picture containing chart

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

Figure 5. Best model results for prediction of LOS and survival/deceased patient outcomes.

* 1. Training and Results – Top Twenty Primary Diagnoses

The same model training and validation exercise was performed, but only for the top twenty primary diagnoses to determine if this reduction in scope would result in better performance (see Figure 6). In practice, there is no obvious benefit.

Chart

AI-generated content may be incorrect.

Figure 6. Training and validation results by training epoch for the top twenty primary diagnoses.

1. Conclusions

Healthcare is very expensive and a huge contributor to perceived quality of life. Studying EHR’s and associated socio-economic influences on healthcare outcomes could provide actionable ways forward to improve outcomes. However, the incredible range of personal variables and experience creates a huge range of possibilities and outcomes, making predictions on small datasets like MIMIC-III very difficult.

Mapping of variables to inputs that can use different models allows us to explore the range of possibilities of modern machine learning. Trust a “hallucination” issues for LLM models inspire the selection of other ML techniques like the CNNs used in this work.

REFERENCES

1. Ding, Y. (2025). AI In Healthcare: Lecture Series 11. AI 395T, University of Texas.
2. Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. <https://doi.org/10.13026/C2XW26>.
3. INTERNATIONAL CLASSIFICATION OF DISEASES, NINTH REVISION ICD-9 PDF https://simba.isr.umich.edu/restricted/docs/Mortality/icd\_09\_codes.pdf Archived 2019-12-28 at the Wayback Machine
4. The Chronic Condition Indication , <https://hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp>. Retrieved 15 April 2025
5. Aldrin, J. C., & Forsyth, D. S. (2019, May). Demonstration of using signal feature extraction and deep learning neural networks with ultrasonic data for detecting challenging discontinuities in composite panels. In AIP conference proceedings (Vol. 2102, No. 1). AIP Publishing.
6. Krahenbuhl, P. (2025). Deep Learning: Module 3: Deep Networks. AI 394D, University of Texas.

1. \* Place the footnote text for the author (if applicable) here. [↑](#footnote-ref-1)