Citation to the original paper

Zhang, D., Yin, C., Zeng, J. et al. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 20, 280 (2020). https://doi.org/10.1186/s12911-020-01297-6

What is the general problem this work is trying to do?

The paper is trying to use predictive models created using a combination of structured medical data and unstructured clinical notes to predict different risks such as in-hospital mortality, 30-day hospital readmission, and long length of stay prediction.

What is the new specific approach being taken in this work, and what is interesting or innovative about it, in your opinion?

Most models either make use of structured medical data or unstructured clinical notes data but rarely both together. This work seeks to combine both types of data to make a more accurate prediction model.

What are the specific hypotheses from the paper that you plan to verify in your reproduction study?

We wish to verify that combining sequential unstructured clinical notes and structured data from electronic health records (EHR) will result in a better model for prediction than using only one of these at a time.

What are the additional ablations you plan to do, and why are they interesting?

It is difficult to gauge or comment on any interesting additional ablations we might consider at this point. We will have a better understanding of the code once we take a deep dive into the implementation details. Hence, we are leaving this section open to consider at a later stage.

State how you are assured that you have access to the appropriate data.

MIMIC-III database analyzed in the study is available on PhysioNet repository (https://mimic.mit.edu/).

Discuss the computational feasibility of your proposed work.

The work makes use of two different neural network architectures named Fusion-CNN and Fusion-LSTM. Each model mainly consists of 5 parts, static information encoder, temporal signals embedding, sequential notes representation, patient representation, and output layer. The batch size is 64 and max number of epochs is 50. The number of vectors for Doc2Vec model is 200. All this is very feasible to process on an average machine.

State whether you will re-use existing code (and provide a link to that code base) or whether you will implement yourself.

We plan to reuse the existing code as walkthrough, but we might refactor the code and make changes to better understand and implement the paper. The source code is provided for reproducing and is available at https://github.com/onlyzdd/clinical-fusion.