M.Sc. Project Proposal

Souradeep Sen
Department of Computer Science,
University of Exeter
(Dated: April, 2023)

The study aims to compare the performance of a Hybrid Deep Learning (HDL) architecture for predicting mortality and hospitalization in heart failure patients with frailty, against traditional survival analysis techniques. HDL uses unsupervised and supervised layers to estimate prediction probabilities based on low dimensional and contextual/historic features from clinical data. By leveraging longitudinal patient data made available in the form of Electronic Health Records (EHR), a new perspective is sought on the performance of machine learning risk prediction models compared to conventional survival analysis in HF patients.

I. INTRODUCTION

Heart failure (HF) is a complex clinical syndrome that affects millions of individuals worldwide, particularly those with frailty, a common condition in older adults. Predictive models using machine learning (ML) based approaches have shown promise in identifying patients at high risk of adverse outcomes. In this project, we aim to develop and validate predictive models to identify HF patients with frailty who are at increased risk of mortality and hospitalization using EHR data.

II. RESEARCH CONTEXT

The framework of this research is based on previous work in predictive modeling for heart failure patients. Traditional survival analysis has been used extensively to predict mortality in this patient population. Machine learning methods have also been employed, including neural networks and decision trees. However, limited work has been done in predicting mortality and hospitalization in HF patients with frailty, especially using electronic health records (EHR) data.

1. Who is a 'frail' patient?

Our target pool of patients being those suffering from frailty, one first needs to understand what the term 'frail' refers to. Although its meaning may be nebulous to some extent, the eFI (Electronic Frailty Index) is a good clinical approximation of the condition. The eFI is based on the cumulative deficit model of frailty ¡doubt-1¿, which proposes that frailty is the result of a build-up of health deficits or impairments. It is calculated using a set of clinical indicators, such as the presence of chronic conditions, cognitive impairment, and mobility problems, that are commonly documented in EHRs.

2. Investigating Heart Failure

HF can be influenced by lifestyle risk factors, such as diet, exercise, smoking, and alcohol consumption. Clinical factors such as diabetes, hypertension, BMI, cholesterol, and sugar consumption were also identified as potential risk factors for HF. Familial history may not be pertinent for all cases of HF, but may only be relevant for some cases.

Several past papers have addressed predictive modeling for heart failure patients. For instance, Xie et al. (2016) used traditional survival analysis to predict the risk of all-cause mortality and heart failure readmission in patients with HF. Another study by Wang et al. (2018) used machine learning techniques, including neural networks, to predict the risk of 30-day readmission in patients with HF. Meanwhile, Cho et al. (2020) proposed a decision tree-based model for predicting the risk of mortality in HF patients.

Recent studies have also explored the use of deep learning techniques for predicting outcomes in HF patients. For instance, Li et al. (2021) developed a deep learning model to predict the risk of all-cause mortality in patients with HF using EHR data. Many previous studies using machine learning for modeling the risk of HF in patients have focused primarily on aggregate or peak metrics - they have not considered incidences as time-to-event. Longitudinal data, which captures the entire cycle of a patient's diagnosis and treatment, is critical for accurate risk prediction. The use of electronic health records (EHR), such as those available in the Clinical Practice Research Datalink (CPRD), offers an opportunity to access comprehensive longitudinal data. This project will leverage CPRD's rich data source to address this research gap and develop a robust predictive model for hospitalization and mortality in frail heart failure patients.

However, limited work has been done on predictive modeling for HF patients with frailty, particularly using hybrid deep learning and Hidden Markov Models. This project aims to bridge this gap in the literature.

III. AIMS & OBJECTIVES

Clearly stated Research Questions:
Objectives to answer them.
This project aims to develop an efficient and scalable predictive model for frail patients with heart failure, combining deep neural networks with statistical machine learning. The model will enable continuous risk prediction for hospitalization and mortality, addressing a critical need in patient care. The hybrid approach will provide stability and accuracy, enhancing the model's performance compared to traditional survival analysis or neural network models. The results of this project could improve patient outcomes by identifying those at high risk and facilitating targeted interventions. Our ML approach relies on hybrid deep learning, specifically an architecture that combines statistical learning in the lower unsupervised layers and convolutional neural networks on the top layer. Our secondary aim is to compare the hybrid deep learning models with a more mathematically transparent model, Hidden Markov Models (HMMs). We will also assess the trade-off between model performance and interpretability to determine the most suitable model for this task.

IV. DATA & RESOURCES

CPRD (Clinical Research Practice Datalink) data is likely to be used for this project. From a collaboration with the medical school at the University of Exeter, I am being granted access to a set of data contaning EHR records of a vast number of patients, consisting of both HF patients and non-HF patients.

A. Data Sources

1. *Primary Data Source: CPRD Aurum*

The Clinical Practice Research Datalink (CPRD) is a large, longitudinal UK primary care database containing anonymized electronic health records. It includes information on patient demographics, diagnoses, symptoms, prescriptions, and referrals, among other clinical data. The database is widely used for observational research, including epidemiological studies, drug safety monitoring, and healthcare utilization analysis. CPRD is maintained by the Medicines and Healthcare products Regulatory Agency (MHRA) and is available to researchers and industry partners who meet certain criteria and obtain appropriate approvals.

The Clinical Practice Research Datalink (CPRD) Aurum is a database provided by EMIS Health. It includes data for approximately 7 million active patients, and covers consenting practices in England, with consenting practices from Northern Ireland available from 2019. CPRD Aurum is updated monthly, and it is representative of the broader English population in terms of geographical spread, deprivation, age, and gender. However, it does not record data outside of primary care. It is important to be aware of this limitation. To bypass it, the CPRD Aurum database can be linked to datasets such as Hospital Episode Statistics (HES), Death Registration, Cancer data, Mental Health Services Dataset, and Small Area-Level Data.

2. *Contingent Data Source: MIMIC-III*

The MIMIC-III (Medical Information Mart for Intensive Care III) dataset is a large, freely available critical care database that contains de-identified health data from over 40,000 patients admitted to the Beth Israel Deaconess Medical Center ICU between 2001 and 2012. It includes over 2 million ICU admissions and is one of the most comprehensive ICU databases available. The data is diverse, including demographic information, vital signs, laboratory tests, medications, and more. The dataset also includes information about procedures, diagnoses, and outcomes.

The MIMIC-III dataset has been widely used for research in various fields, such as machine learning, clinical decision support, and epidemiology. It is particularly useful for the development and testing of predictive models and algorithms, as well as for investigating clinical phenomena and interventions. The dataset has contributed to numerous studies, and many of its findings have led to improved patient care and clinical decision-making.

The database has been made publicly available to researchers around the world, but access is restricted and requires a formal application process. This is to ensure the privacy and security of patient data, as well as to maintain the integrity of the dataset. Researchers who wish to use the dataset must complete a certification course on data privacy and ethics before gaining access.

In summary, the MIMIC-III dataset is a large, diverse ICU database that has been widely used for research in various fields. It has contributed to numerous studies and has led to improved patient care and clinical decision-making.

B. Coding

There are several possible tools of choice for the project. The ones I plan to use are Python for coding the base unsupervised layers and upper supervised layers as well as encoding contextual and historical data. This is because it has many useful libraries and frameworks at its disposal for numerical computing and neural networks,

namely numpy, scipy, Pytorch and Tensorflow. For the more statistically inclined approach of Hidden Markov Models, I plan to use R as it has native functionalities for building Markov chains as well as pacakges like `brms` and `rstan` for Bayesian simulation methods like Markov Chain Monte Carlo.

V. METHODS & EXPERIMENT DESIGN

A. Model Building

To evaluate the performance of the hybrid deep learning models and Hidden Markov Models developed in this project, we will compare them to traditional survival analysis models such as Cox proportional hazards and Kaplan-Meier curves. These models are widely used in predicting outcomes in heart failure patients, providing a benchmark for our ML-based models. We will assess the performance of our models based on metrics such as accuracy, sensitivity, and specificity. The comparison will allow us to determine whether the ML-based models outperform traditional models, providing insights into the potential of these models for predicting outcomes in frail heart failure patients.

1. Primary Aim: CNN

The idea is to parse through time-series data of individual patient histories and identify signals that can discriminate between HF and non-HF patients. These signals can then be encoded from their high-dimensional space into a latent vector representation in lower dimensions - possible techniques for dimensionality reduction such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) may be used for this purpose. These latent representations can then be fed into supervised layers of a Convolutional Neural Network ¡ref?¿. The output from the network should be an array depicting the persistency of a patient till an even of interest - in our case, time-to-hospitalization or time-to-death. Separate hold-out sets can be reserved for testing purposes. This will be more favourable against a discretized or aggregate output as it will allow more direct comparison with traditional survival analysis like multivariate Cox regression ¡ref?¿.

2. Secondary Aim: HMM

The secondary aim would be to use either the multivariate time series of observed biosignals or their latent vector representation $\{o_{1,\cdots,T}\}$ as observed states in a Hidden Markov Model $\lambda = (A, B, \pi)$ [1]. Algorithms for likelihood estimation (Forward Algorithm ¡ref?¿), decoding (Viterbi Algorithm ¡ref?¿) and learning (Baum-Welch ¡ref?¿) will be used to derive ideal parameters for the model and then extract most probable sequence of hidden states given the observations. The hidden states here can correspond to either a binary set $S = \{1, 2\}$ corresponding to alive and deceased with their individual probabilities at time t allowing us to understand the onset of HF and how it may change over a sequence of observed events.

B. Explainability & Interpretability

In the medical domain, model interpretability is crucial to ensure that the predictions made by the model can be trusted and applied in a clinical setting. Although ML-based models may provide higher accuracy, their black-box nature can make them less interpretable, and therefore cast reasonable doubt on their applicability. Therefore, it is important to assess the trade-off between model performance and model interpretability. One approach to assessing this trade-off is to use model-agnostic methods such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), which provide insights into how the model arrived at its prediction. We will use these methods to explain the predictions made by our models and compare the level of interpretability with traditional survival analysis models. We will also assess the performance of the models against the level of interpretability, enabling us to evaluate the trade-off between model performance and interpretability. The results will provide guidance on the most suitable model for predicting outcomes in frail heart failure patients in a clinical setting.

VI. DATA GOVERNANCE & ETHICS

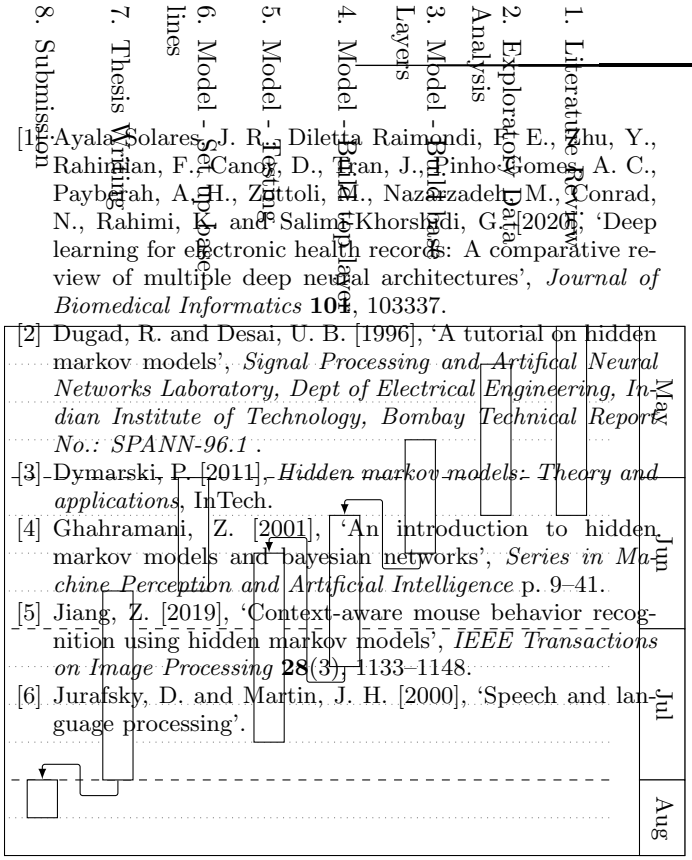
Ethics approval is an essential requirement for this research project as it involves patient data. Patient data is considered sensitive information and approval ensures that this data is handled with care and confidentiality. For the purpose of this project, the medical school at the University of Exeter has consented to share CPRD Aurum data that they have procurred for their own research. Ethical approval is being sought from University of Exeter's management system, WorkTribe. As of writing this proposal however, this is still pending due to unforeseen technical glitches on the University website that is barring me from getting access to the platform. Access to the contingent data source, MIMIC-III is

[1] A and B are respectively the state transition matrix and the observation (emission) matrix

restricted and requires a formal application process to maintain data privacy and integrity. To comply with this, I have undergone the CITI training program for Data or Specimens Only Research.

VII. PROJECT PLAN

The project plan is outlined by the following chart.



VIII. RISK ASSESSMENT

There is a chance of the requested CPRD data not coming through on time. As a contingency, I have gained access to the open-source data MIMIC-III, which requires a thorough training program to be completed.

IX. CONCLUSION

X. REFERENCES

[1] Ayala Solares, J. R., Diletta Raimondi, F. E., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Pinho Gomes, A. C., Payberah, A. H., Zottoli, M., Nazarzadeh, M., Conrad, N., Rahimi, K. and Salimi-Khorshidi, G. [2020], 'Deep learning for electronic health records: A comparative review of multiple deep neural architectures', *Journal of Biomedical Informatics* **101**, 103337.

[2] Dugad, R. and Desai, U. B. [1996], 'A tutorial on hidden markov models', *Signal Processing and Artifical Neural Networks Laboratory, Dept of Electrical Engineering, Indian Institute of Technology, Bombay Technical Report No.: SPANN-96.1* .

[3] Dymarski, P. [2011], *Hidden markov models: Theory and applications*, InTech.

[4] Ghahramani, Z. [2001], 'An introduction to hidden markov models and bayesian networks', *Series in Machine Perception and Artificial Intelligence* p. 9–41.

[5] Jiang, Z. [2019], 'Context-aware mouse behavior recognition using hidden markov models', *IEEE Transactions on Image Processing* **28**(3), 1133–1148.

[6] Jurafsky, D. and Martin, J. H. [2000], 'Speech and language processing'.

[7] Kawamoto, R., Nazir, A., Kameyama, A., Ichinomiya, T., Yamamoto, K., Tamura, S., Yamamoto, M., Hayamizu, S. and Kinosada, Y. [2013], Hidden markov model for analyzing time-series health checkup data, *in* 'MEDINFO 2013', IOS Press, pp. 491–495.

[8] Lorenzoni, G., Sabato, S. S., Lanera, C., Bottigliengo, D., Minto, C., Ocagli, H., De Paolis, P., Gregori, D., Iliceto, S., Pisanò, F. and et al. [2019], 'Comparison of machine learning techniques for prediction of hospitalization in heart failure patients', *Journal of Clinical Medicine* **8**(9), 1298.

[9] Nguyen, P., Tran, T., Wickramasinghe, N. and Venkatesh, S. [2017], 'Deepr: A convolutional net for medical records', *IEEE Journal of Biomedical and Health Informatics* **21**(1), 22–30.

[10] RABINER, L. R. [1990], 'A tutorial on hidden markov models and selected applications in speech recognition', *Readings in Speech Recognition* p. 267–296.

[11] Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G. and Collins, G. S. [2018], 'Minimum sample size for developing a multivariable prediction model: Part ii - binary and time-to-event outcomes', *Statistics in Medicine* **38**(7), 1276–1296.