ENGG4801 Project Proposal

Gareth Booth 44353027

March 2020

1 Topic

The proposed project with supervisor Sally Shrapnel involves using intensive care unit (ICU) data and various machine learning techniques to predict various patient outcomes, such as intubation, morbidity and sepsis. This project will have a special focus on ICU data from patients affected by the current COVID-19 outbreak.

The second part of the project will involve interrogating the machine learning model with explainability techniques to see what the model has learned. Various techniques will be compared, and a thorough investigation will be done into their usefulness and applicability.

2 Background

The recent COVID-19 outbreak has caused great strain on health systems around the world. Many hospitals are recording ICU data in order to get a better understanding of the disease. Using this data to predict some patient outcomes such as mortality, sepsis, or when mechanical ventilation is required (ECMO or intubation) would greatly help doctors. COVID-19 has

also caused a catastrophic shortage of ICU beds in many countries. This highlights the importance of discharging patients as soon as it is safe to do so. Knowing when to extubate patients and discharge them is an important problem that can lead to more optimal bed usage.

Machine learning has shown great potential in the medical space, and has outperformed doctors in some fields such as dermatology [1]. While neural networks can achieve great performance, neural networks are often seen as 'black boxes'. Although machine learning models are capable of learning very complex relationships, it is possible that a model is learning relationships that aren't desirable. If a dataset contains a bias, the model is likely to also learn this bias. Searching the dataset manually for biases is not feasible, due to the amount of data and the presence of unknown biases [2].

Fortunately, there are many explainability techniques that can help interrogate a neural network to see how it's making it's predictions. Explainable models provide much needed transparency, givings users confidence when the explanations match their wisdom. They can also see if the explanation is sensible, and if it isn't they don't have to trust the model for that prediction. They can then start investigating why the model has learned an unexpected relationship. If this is from a bias, targeted data collection can now be done.

Machine learning on ICU data has recently become a popular topic. One of the most widely used datasets is MIMIC-III, a dataset of deidentified ICU data from 53,423 distinct hospital admissions [3]. Due to its size and public availability, it has been widely used for benchmarking in machine learning tasks. The data includes bio-markers, drug administrations, procedures and diagnoses (using International Classification of Diseases).

One difficulty encountered while looking at MIMIC-III data is the presence of duplicate fields, for example there are two fields for white blood cell count that are exactly the same. Another difficulty in using the dataset is the use of two different medical information systems throughout the data collection, CareVue and MetaVision [3].

It has been shown that using models trained on ICU data from one hospital on data from another hospital will have decreased performance in [4]. This study also shows that a large sample size is needed from hospitals when transfer

learning. Ensuring that the model will generalize is especially important in this project, as data will be gathered from ICUs around the world. However, it is worth noting that this method is using logistic regression (LR) models.

2.1 Recurrent Neural Network

There are many ways machine learning can be useful in this space. One of the challenges that would benefit ICU doctors is, given a patients ICU data, predict whether or not the patient will require intubation tomorrow. This will allow them to ensure they have adequate resources allocated when intubation is required. This involves handling time series data.

Traditional neural networks and convolutional neural networks take a fixed size input and produce a fixed size output, doing a fixed number of computations. This makes them unsuitable for unknown amounts time series data. Recurrent neural networks (RNNs) are neural networks with a feedback loop [5]. An RNN consists of a neural network that takes and outputs some data, but also takes some state from the previous application of the neural network as input. This means that an RNN is able to make predictions based on what it's seen before, and is also able to take an arbitrary length of input sequence.

One example use of an RNN is generating sequences. An RNN could take as input a character, and output what it thinks the next character will be. If the output is used as the next input, the network can generate large sequences of output. This RNN could be trained on paragraphs of English text, and will be able to create series of paragraphs [6].

Long short-term memory (LSTM) is an RNN architecture that achieves greater performance than a traditional RNN. It allows models to learn long-term dependencies, i.e. information from many inputs ago [5]. LSTMs take two inputs from the previous application, one is called the "cell state" and the other is the data output. This extra cell state is updated each application by removing unimportant data and adding new relevant data.

However, simple LSTM architectures have a drawback of not being able to

use context from the entire input sequence, and can only make predictions from the internal state and inputs up to that point. For problems like machine translation, this is not desirable [7]. Instead, another commonly used architecture is the seq2seq architecture, which uses an encoder and a decoder network [8]. The encoder is an LSTM network that summarizes the entire input sequence into a single internal state, which is a fixed length vector. It passes on this internal state to a decoder network. The decoder network generates a new sequence by feeding back the output of each LSTM application into the input of the next one. In order to not output an infinite sequence, a special "start" input token is used as input to the decoder, and an "end" token is output from the decoder.

While the encoder-decoder network performs well on small sequences, the summary produced by the encoder network is a bottleneck for longer input sequences [7]. The attention method was created to resolve this problem. Instead of using only the summary of the entire input sequence, attention uses the internal state of the encoder LSTM after every input. The encoder in the original attention paper is a bidirectional RNN, which allows each internal state to have information from the preceding and following input [9].

A neural network is used at each application of the decoder to decide, based on the current decoder internal state, which of the internal states from the encoder will be useful for this output. A 'context vector' is then created, which weights the useful internal data. This context data is added as input into the decoder [7]. In the example of machine translation, in the sentence "I am writing a long report," the words 'writing' and 'report' are expected to attend strongly, whereas 'writing' and 'long' aren't.

Besides removing the information bottleneck, another advantage of using attention over seq2seq is that they can be explainable. The simplest method of making attention models explainable is, for each output, getting how important each feature is from the context vector and making a heatmap (attention map) from all of the outputs. This is a local method, meaning that it provides explanations about a single input.

As an example of how LSTMs can be applied to ICU data is as follows. Each input to an LSTM will be the patients data for the whole day, e.g. all chart

readings, drug administrations etc. The output is the predicted values of their ICU data for the next day. A vanilla neural network can be used to make a prediction as to what the target class will be based on this output data. Alternatively, the output from the model could be a probability that some event occurs tomorrow. While attention was initially developed for machine translation [9], it is also useful in this context. An attention layer can simply be added before the LSTM in the above example. All of the components required to run this example are implemented in the machine learning framework Keras [10].

Some relevant machine learning papers using ICU data and LSTMs are investigated.

In [11], an LSTM with attention is used to predict daily sepsis, myocardial infarction (MI), and vancomycin antibiotic administration with data from MIMIC-III. This model was able to predict next-day vancomycin with a sensitivity of 0.71, and sepsis with a sensitivity of 0.73. However, next-day MI prediction had a poor sensitivity of 0.56, likely due to the lack of strong predictors present in the database. This paper used attention maps to highlight which features were most important for predictions. It finds that the model exploited features that reflect clinician decision-making, which highlights the importance of feature level explainability and that care is required to ensure the model can learn features that are clinically useful. Furthermore, all code is publicly available.

[12] uses data from four ICUs to predict 90-day mortality. It is worth noting that the data has a temporal resolution of 1 hour, which is far higher than most ICU datasets. They use a vanilla LSTM with hyperparameter training. The code for creating and training the model is available. On admission, the model has an AUROC of 0.73, 0.82 after 24 hours, and 0.88 at the time of discharge. The paper is able to produce hourly explanations while data is still being collected. The authors use SHAP to do this.

SHAP is a model agnostic, local interpretability method. It is a unification of six existing methods, and outputs the importance value for each feature based on the contribution that feature makes to the prediction [13]. It is based on Shapley values from co-operative game theory, and has a strong theoretical justification [2]. The SHAP value for a feature is it's compound

effect when interacting with all other features [12]. These local explanations can be combined to make global explanations [14].

In the above paper, SHAP is used to tell if a given feature increases or decreases the probability of mortality. It is used to show that, for example, the presence of a small age in a patient greatly decreases risk, while a higher age increases it [12]. It also shows that the presence of comorbidities greatly increases risk.

The challenging nature of ICU data has also led to the creation of new machine learning techniques. In [15], a new method is proposed to handle data that is irregular in sampling times and features, such as the MIMIC-III dataset. They use Gated Recurrent Unit (GRU) RNNs, which are similar to LSTMs but only pass forward one internal state vector instead of two. Within these GRU RNNs, they use neural ordinary differential equations instead of a neural network, and a Bayesian update network to process the observations. Neural ODEs are used due to their continuous nature, which is well suited for time series data. This means that they don't have to group observations into a single day, which involves downsampling some features and upsampling others. They achieve state-of-the-art results on MIMIC-III. However, the paper does not incorporate explainability. Code is available.

2.2   Random Forests

Another challenge involves predicting an event based on a fixed amount of data. For example, predicting when a patient will require intubation given only hospital admission data. This is a significantly harder challenge, as [11, 12] find that models benefit from access to longer time series data. However, knowing if a patient will require intubation many days in advance would be extremely valuable.

There are many ways to handle a fixed amount of input data. As already shown in [12], LSTMs are able to do this. Another popular method is using random forests. Random forests (RF) and gradient boosted trees achieve state of the art performance in many domains, particularly when features are individually meaningful and in tabular form [14]. Random Forests involve

6

ensembling many decision trees and getting a single aggregated result. Gradient boosting methods are slightly different in the way they're constructed.

There are many global explanation methods for tree based methods. The simplest method is to report the decision path. However, this does not scale for models with multiple trees [14]. Many perturbation based methods exist, where the values of one or more features are perturbed and a change in the final prediction is measured [16]. These have the drawback of being non-deterministic, often providing different outputs for the same input [14].

From the creators of SHAP (and others), TreeExplainer is a local explainability method based on SHAP values [14]. As with SHAP, TreeExplainer has many desirable features, including stability, speed, and fairness regardless of tree depth. A software package is available.

Random forests have been a popular method when using ICU data.

[17] focuses on predicting if patients are ready to be discharged from ICU or not. The authors use data from MIMIC-III and data from another hospital, which allowed them to test generalizability. They use both random forests and logistic classifiers. They used a permutation based method to calculate the relative importance of each feature. This involves permuting the values of the input features and seeing how this change affects the AUROC. Each features importance is slightly different across each dataset, however the top 2 are the same (minimum GCS and airway). The data they use from MIMIC-III is restricted to the MetaVision database, which contains less data than CareVue. In order to address class imbalance (insufficient negative outcomes from discharge), extra data is generated. Data resolution is 4 hours if available, otherwise 36 hours. Missing data is filled in using k-nearest-neighbours.

Mortality is predicted in patients with unplanned extubation in [18]. The paper compares the performance of four different machine learning architectures, including neural networks, logistic regression models and random forests. The study concluded that random forests were the most suitable for this task, having the highest AUROC, precision and recall. However, the dataset only contains data from 341 patients in this study.

[19] uses random forests and gradient boosting to recognize patients at high risk of acute respiratory distress syndrome (ARDS). They use the MIMIC-III database. The paper also uses latent class analysis. This is used to find groups (sub-phenotypes) in their data, which allow them to create 3 separate models, one for each group. This study is relevant due to the fact that COVID-19 can rapidly cause the onset of ARDS [20], and so the list of vitals and measurements from this paper should be used.

3   Aims

This project will feature ICU data from coronavirus patients. Data will be entered by ICU workers in hospitals around the world into UQ's REDCap database (`https://redcap.health.uq.edu.au/`). However, due to time constraints and stretched resources inside ICUs around the world, it is likely that there will be insufficient data for training any machine learning models for some time. Because of this, the MIMIC-III database will also be used.

Given some ICU data, we aim to use various machine learning techniques to predict patient outcomes. These patient outcomes include mortality, sepsis, discharge and ECMO. Another goal will be to see the effectiveness of some treatment methods, such as administration of antiviral medications.

One of the primary goals of this project is to incorporate explainability. This will include an investigation into many existing techniques and an attempt to improve or combine them.

A secondary goal is to create a complete software package combining predictions and explainability that would hypothetically be deployable to doctors.

To promote research in the area and advance the global communities understanding of the virus, it is a goal of this project to make it all open source. The availability of open datasets like MIMIC-III has created much interest in ICU data, and the unique nature of the coronavirus ICU dataset we have access to will provide an interesting challenge for researchers, and will help prepare for outbreaks in the future.

4   Required Resources

Data is a crucial part of this project. Access to the publicly available MIMIC-III dataset involves completing a training course and requesting access [3]. Access to the REDCap dataset will require appropriate ethics approval.

The Python programming language will be used, and a variety of open source machine learning and data visualization packages will be used throughout this project such as Keras [10].

Another key resource is computing power. The MIMIC-III dataset contains 50GB of data. Also, machine learning requires a powerful GPU to accelerate training. Therefore access to a compute cluster will be required throughout the duration of the project. The Wiener cluster (`https://rcc.uq.edu.au/wiener`) has ample high speed memory (384 GB per node) and very powerful GPUs, and will be used for this project. As a backup, access to another cluster, getafix, has been obtained (`http://research.smp.uq.edu.au/computing/getafix.html`).

5   Milestones

5.1   MIMIC-III Data Processing (March)

Creating a dataset from the MIMIC-III database is expected to take some time. This will involve taking all the tables in the MIMIC-III database, extracting relevant data and eventually merging them into one CSV file. The data must be organized into a time series (daily), which makes it easy to get admission only and time series data. As previously noted, the dataset contains many duplicate fields that will have to be found and handled.

This task includes choosing some relevant features for predictions and searching for all of these in the database. Also, the target features must be found e.g. calculating the days to intubation by finding when intubation was ap-

plied, or finding the disease code for sepsis and finding when that event occurred. This task is expected to take only a week because there are some papers that have already done this task and have released code, e.g. [11].

5.2 REDCap data processing (March)

As the schema for the REDCap database is currently known, fields in the dataset can be matched with those in MIMIC-III. Emphasis for this task is placed on ensuring that the format is similar to the processed MIMIC-III data, a time series CSV file with similar features. This will allow both datasets to be used in training interchangeably, which is important as we aren't expecting to have enough REDCap data to do machine learning for a while. As the schema is known already and is quite similar to the expected output from the MIMIC-III data processing step, processing the REDCap data is expected to take a few days.

5.3 Data Investigation (April)

General investigation into the dataset is an important first step. Knowing what data is missing, what units are used etc. is all helpful information. This step can be done with the final dataset and the data dictionary. It may involve creating visualizations for each feature such as bar plots and violin plots.

5.4 Random Forests Investigation (April)

Various random forest methods will be investigated. This will involve random forests and gradient boosted methods. Scikit learn's random forest implementation will be used initially. Gradient boosting methods such as XGBoost will be used, among others.

The models should be created so that it is possible to easily change the

quantity of data that is passed in, for example 1 week of ICU data instead of 1 day. This should be possible by making the input from concatenating the data for each day.

A standard 20% test set and 10% validation set will use used. Hyperparameters for random forests include number of trees, learning rate and interaction depth.

5.5 Random Forests Explainability Investigation (April - May)

Perturbation methods and TreeExplainer will be investigated. Other relevant explainability techniques that are promising will be investigated for this section.

Some sanity checks on the explainability methods will be done. This will include testing their reliability, robustness and model dependence.

5.6 LSTM Investigation (June)

After the data is obtained, an LSTM based network can be trained on either dataset.

Initially, a vanilla LSTM will be trained to predict some of the patient outcomes. Predicting next-day mortality and discharge first will be beneficial as these outcomes are well defined in the database. Training will involve separating the dataset from the feature labels. Feature labels will be shifted to the previous day so that the model will learn to predict next day events.

If time permits, hyperparamter optimization will be done over chosen optimizer and optimizer parameters such as learning rate, neural network depth, batch size, and regularization techniques.

After results are obtained on these two outcomes, more predictions will be

implemented, including sepsis, intubation and ECMO.

After this, attention can be added to the model to see what difference it makes (in accuracy and training time). There are also different variations of attention that can be used.

Training the model to predict multiple days in advance will also be attempted. This can be done by shifting the feature labels by more than one day.

Keras will be used to implement the predictive model, as it is easy to use and runs on the compute clusters.

5.7   LSTM Explainability Investigation (June - August)

Two techniques of interest are SHAP and attention maps. Implementing an attention map technique isn't expected to take long, as Keras has some functionality to get extract data from the attention model. Code is also available for SHAP. Some time will be spent visualizing the explanations using plotting packages.

As with random forests, some sanity checks on these techniques will also be done.

At this stage, the explainability techniques may reveal flaws in the dataset or model. We should then attempt to fix these problems.

At this point, we can see if the LSTM model has learned anything different from the random forests model. We can also compare the LSTM explainability methods, testing if they give similar explanations.

5.8   Open Sourcing

It is difficult to estimate the time required or even determine the feasibility of creating an open database from REDCap. Creating an open ICU dataset has been done in the past, however due to the number of parties involved in creating and managing the data, it might be that it takes months to come to a data sharing agreement. However, open sourcing the project itself is expected to be straightforward, as it involves choosing a licence and ensuring adequate code quality.

5.9   Write Up (September - November)

The final demonstration is due late October, and the final report is due early November. As this report took approximately a week of full time equivalent work, it is expected that the final report (which is typically 60 pages) will take 5 to 6 weeks.

6   Risk Assessment

There are no health and safety risks associated with this project.

There are, however, some significant risks to the project itself. One risk is a data breach. Because the data from the REDCap database, while de-identified, is initially confidential, it is important that the only people who have access to it are the members of the research group. This is a risk because a breach may cause our access to the data to be revoked, and so has a very high severity. However, this risk can be mitigated by using best practices when handling data. One such option is to use UQ's Research Data Manager (RDM) for all identifiable data (`https://research.uq.edu.au/rmbt/uqrdm`). Using RDM will ensure the security of the data. Furthermore, any computers that have access to the data will be password protected and locked when not in use. Some residual risk is left over, as it is still possible

that misconfiguration can cause malicious actors to be able to access the data. However, the likelihood of this is low.

Another risk is insufficient data. In general, machine learning algorithms require at least 5000 data points before achieving acceptable accuracy [21]. It is possible that this amount of data won't be obtained for many months. Fortunately, this is not a severe risk to the project as we can use the MIMIC-III database as a backup.

Another risk to this project is coronavirus itself. Fortunately, as the entire project consists of online work, every aspect of the course can be done online. This greatly reduces the risk of being directly affected by the virus. However, coronavirus could cause long delays in the project due to causing the shutdown of many services. Some services that are critical to the project include access to the internet to collaborate with teammates, and access to UQ servers. Unfortunately, this risk cannot be fully mitigated, as online resources are critical. However, it can be partially mitigated by downloading all required resources

[22] [23] [24]

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115 EP –, Jan 2017. [Online]. Available: https://doi.org/10.1038/nature21056

[2] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" *arXiv.org*, vol. 11797, 2019. [Online]. Available: http://search.proquest.com/docview/2276715881/

[3] A. E. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, 2016.

[4] A. Johnson, T. Pollard, and T. Naumann, "Generalizability of predictive models for intensive care unit patients," *arXiv.org*, 2018. [Online]. Available: http://search.proquest.com/docview/2151577087/

[5] C. Olah, "Understanding lstm networks," Aug 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[6] A. Karpathy, "The unreasonable effectiveness of recurrent neural networks," May 2015. [Online]. Available: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[7] L. Weng, "Attention? attention!" *lilianweng.github.io/lil-log*, 2018. [Online]. Available: http://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

[8] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," vol. 4, no. January. Neural information processing systems foundation, 2014, pp. 3104–3112.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv.org*, 2016. [Online]. Available: http://search.proquest.com/docview/2079082715/

[10] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[11] D. A. Kaji, J. R. Zech, J. S. Kim, S. K. Cho, N. S. Dangayach, A. B. Costa, and E. K. Oermann, "An attention based deep learning model of clinical events in the intensive care unit.(research article)(report)," *PLoS ONE*, vol. 14, no. 2, p. e0211057, 2019.

[12] H.-C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strøm, P. J. Chmura, M. Heimann, L. Dybdahl, L. Spangsege, P. Hulsen, K. Belling, S. Brunak, and A. Perner, "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records," *The Lancet Digital Health*, vol. 2, no. 4, pp. e179–e191, Apr 2020. [Online]. Available: https://doi.org/10.1016/S2589-7500(20)30018-2

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing*

Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[14] S. M. Lundberg, G. G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee, "Explainable AI for trees: From local explanations to global understanding," *CoRR*, vol. abs/1905.04610, 2019. [Online]. Available: http://arxiv.org/abs/1905.04610

[15] E. De Brouwer, J. Simm, A. Arany, and Y. Moreau, "Gru-ode-bayes: Continuous modeling of sporadically-observed time series," *arXiv.org*, 2019. [Online]. Available: http://search.proquest.com/docview/2232265113/

[16] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.

[17] C. J. McWilliams, D. J. Lawson, R. Santos-Rodriguez, I. D. Gilchrist, A. Champneys, T. H. Gould, M. J. Thomas, and C. P. Bourdeaux, "Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from mimic-iii and bristol, uk," *BMJ Open*, vol. 9, no. 3, 2019. [Online]. Available: https://bmjopen.bmj.com/content/9/3/e025925

[18] M. Hsieh, M. Hsieh, C.-M. Chen, C.-C. Hsieh, C.-M. Chao, and C.-C. Lai, "Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units," *Scientific Reports*, vol. 8, 12 2018.

[19] T. Wang, T. Tschampel, E. Apostolova, and T. Velez, "Using latent class analysis to identify ARDS sub-phenotypes for enhanced machine learning predictive performance," *CoRR*, vol. abs/1903.12127, 2019. [Online]. Available: http://arxiv.org/abs/1903.12127

[20] C. Wu, X. Chen, Y. Cai, J. Xia, X. Zhou, S. Xu, H. Huang, L. Zhang, X. Zhou, C. Du, Y. Zhang, J. Song, S. Wang, Y. Chao, Z. Yang, J. Xu, X. Zhou, D. Chen, W. Xiong, L. Xu,

F. Zhou, J. Jiang, C. Bai, J. Zheng, and Y. Song, "Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China," *JAMA Internal Medicine*, 03 2020. [Online]. Available: https://doi.org/10.1001/jamainternmed.2020.0994

[21] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *MIT Press*, 2016.

[22] L. Mica, K. Albrecht, M. Keel, and O. Trentz, "Independent predictors of early death of polytrauma patients: An analysis of 696 patients," *J Trauma Treatment*, vol. 1, 01 2012.

[23] F. Sadaka, C. Ethmane Abou El Maali, M. A. Cytron, K. Fowler, V. M. Javaux, and J. O'Brien, "Predicting mortality of patients with sepsis: A comparison of apache ii and apache iii scoring systems." *Journal of clinical medicine research*, vol. 9, no. 11, pp. 907–910, 2017. [Online]. Available: http://search.proquest.com/docview/1952104244/

[24] S. Serrano and N. Smith, "Is attention interpretable?" *arXiv.org*, 2019. [Online]. Available: http://search.proquest.com/docview/2238247948/