2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

TOWARDS AN EXPLAINABLE MORTALITY PREDICTION MODEL

Jacob R. Epifano[⋆], *Ravi P. Ramachandran*[⋆], *Sharad Patel*[†], *Ghulam Rasool*[⋆]
epifanoj0@students.rowan.edu, ravi@rowan.edu, patel-sharad@cooperhealth.edu, rasool@rowan.edu

[⋆] Rowan University, Department of Electrical and Computer Engineering
[†] Cooper Hospital, Division of Critical Care

ABSTRACT

Influence functions are analytical tools from robust statistics that can help interpret the decisions of black-box machine learning models. Influence functions can be used to attribute changes in the loss function due to small perturbations in the input features. The current work on using influence functions is limited to the features available before the last layer of deep neural networks (DNNs). We extend the influence function approximation to DNNs by computing gradients in an end-to-end manner and relate changes in the loss function to individual input features using an efficient algorithm. We propose an accurate mortality prediction neural network and show the effectiveness of extended influence functions on the eICU dataset. The features chosen by proposed extended influence functions were more like those selected by human experts than those chosen by other traditional methods.

Index Terms— Explainable AI, XAI, Bioinformatics, Influence Functions, Robust Statistics, Mortality Prediction, Explainable deep learning,

1. INTRODUCTION

The black-box nature of deep learning models is limiting their applicability in high-risk areas, including medical diagnosis and treatment planning [1]. Recently, various areas of healthcare have seen an enormous rise in the data quantity [2]. All patient records have moved from paper to the standard Electronic Health Record (EHR), and vital sign data and even waveform data can be downloaded from bedside monitors. Medical diagnosis is especially critical as preliminary studies have shown poor performance and high false alarm rates with the medical community having a negative view of models when used in practice [3, 4].

Jacob R. Epifano is supported by US Department of Education GAANN award P200A180055. Ghulam Rasool was partly supported by NSF OAC-2008690.

In machine learning, the explainability of a model and its performance are generally competing factors [5]. Explainable models, (logistic regression, trees, etc.) are often outperformed by black-box models (neural networks (NNs)). Early work has been done by Ribeiro, et. al who explain the predictions of any classifier using Local Interpretable Model-Agnostic Explanations (LIME) and High Precision Model-Agnostic Explanations (Anchors) [6, 7]. While LIME's local explanation breaks down immediately upon changing the point under test, Anchors resolves this by optimizing for the whole test set. These model agnostic approaches leave much to be desired in terms of explainability, and, therefore, model-specific approaches are preferred especially for NNs. A lot of work in the field of explainable machine learning has been done for image datasets. Specifically, saliency maps are used to propagate classification information from the last layers of Neural Networks to their inputs. For image data specifically this results in a 2D map that is the size of the original input where each pixel's intensity represents how important that pixel is to the classification [8, 9, 10, 11, 12, 13].

In this paper, we investigated analytical techniques known as influence functions which originate from robust statistics. Influence functions allow one to approximate the change that a leave one out (LOO) training scheme would have on parameters [14]. Recently, Koh and Liang showed that influence functions can be used to approximate which training points most effected the loss of a test point and what features were most important for each training point [15]. Since this algorithm is model-specific, it can be leveraged to extract multiple statistics about how the model interprets the data. This makes influence functions an ideal candidate for use in the medical field. One of the limitations of their work is that the gradient is not propagated through the multiple layers of the network [16]. Authors use features from the last layer as an input to a logistic regression model and approximate influence functions [15]. The algorithm cannot provide any information on how the original features affected functions of test loss [15].

The current state-of-the-art for mortality prediction exists in the form of evaluation scores. Acute Physiology and Chronic Health Evaluation (APACHE), and Simplified Acute

Physiology Score (SAPS) are examples of mortality predictors [17, 18]. A simple logistic regression model is fitted to these scores and a binary classifier is built. These methods have been shown to have area under the curve (AUC) from 0.6-0.7 depending on the time of prediction [19]. The prediction time has a huge impact on the utility of these models. Predicting mortality at 48 hours from admittance is not as useful as predicting at the time of admission or after 24 hours. In this work, we focused on predicting mortality at 24 hours after admittance in the ICU.

The main contributions of this paper are threefold.

1. The influence function approximation is extended such that gradients may be calculated in an end-to-end manner from last layer of a network to the first layer.
2. We proposed a new mortality prediction model and compared its performance to the current state-of-the-art.
3. The top features extracted by different prediction models are compared to those chosen by domain experts.

2. APPROACH

2.1. Influence Functions

We consider the problem of predicting binary label y given the feature data $\boldsymbol{x}$. Let $z_i = (\boldsymbol{x}_i, y_i)$, where $i = 1, 2, \cdots, n$, represent n input-output pairs. Building on the work of Koh and Liang, we find approximations for the local feature importance FI_{local} and global feature importance FI_{global} [15].

We start by defining the loss function for a trained model $L(z, \theta) = \sum_{i=1}^{n} L(z_i, \theta)$, where θ represents optimal model parameters. The gradient of the loss function with respect to its parameters θ is given by $\nabla_\theta L(z, \theta) = \nabla_\theta \sum_{i=1}^{n} L(z_i, \theta)$. The change introduced in the parameters θ by removing a training point z is given by [15]:

$$I_{\text{up, params}}(z) = -H_\theta^{-1} \nabla_\theta L(z, \theta), \tag{1}$$

where $H_\theta = \nabla_\theta^2 L(z, \theta)$ represents the Hessian. The effect of removing a training point z on the loss of a particular test point z_{test} is given by [15]:

$$I_{\text{up, loss}}(z, z_{\text{test}}) = -\nabla_\theta L(z_{\text{test}}, \theta)^\top H_\theta^{-1} L(z, \theta). \tag{2}$$

Furthermore, the effect of small changes in the input feature $\boldsymbol{x}$ on the loss at the test point z_{test} is given by [15]:

$$I_{\text{pert, loss}}(z, z_{\text{test}})^\top = -\nabla_\theta L(z_{\text{test}}, \theta)^\top H_\theta^{-1} \nabla_x \nabla_\theta L(z, \theta). \tag{3}$$

It is noted that the components of $I_{\text{up, params}}(z)$, $I_{\text{up, loss}}(z, z_{\text{test}})$, and $I_{\text{pert, loss}}(z, z_{\text{test}})$ can take on positive or negative values. Also, $I_{\text{up, params}}(z)$ provides an approximation for the change in the parameters θ, while $I_{\text{up, loss}}(z, z_{\text{test}})$, and $I_{\text{pert, loss}}(z, z_{\text{test}})$ provide approximate changes in the loss function $L(z, \theta)$. A positive value indicates that removing the training point z or perturbing a particular input $\boldsymbol{x}$ would increase the loss on that test point z_{test}. On the other hand, a negative value would indicate a decrease in the loss function.

2.2. Feature Importance - Local and Global

We define *feature importances* by taking the average of Eq. (3) across all training points in the dataset. Later, when examining the loss of one test point z_{test}, we refer to this as the local feature importance given by:

$$FI_{\text{local}} = \frac{1}{N} \sum_{i=1}^{N} I_{\text{pert, loss}}(z_i, z_{\text{test, point}}). \tag{4}$$

Where N is the number of training points in the dataset. When using the average loss across all test points, we refer to this as the global feature importance:

$$FI_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} I_{\text{pert, loss}}(z_i, z_{\text{test, set}}). \tag{5}$$

2.3. Extended Influence Functions

Koh and Liang freeze all but the top layer of the Neural Net and use the extracted features to train a logistic regression classifier [15] (which breaks the computational graph). The authors assume that the influence functions calculated using the extracted features sufficiently capture all possible changes in the input $\boldsymbol{x}$, which limits the applicability of their framework [15]. We remove this restriction of using the extracted features by keeping the graph intact and provide a methodology for layer-by-layer calculation of the influence functions.

In our settings, we compute the influence functions using Eqs. (4) and (5) with respect to the output layer only rather than with respect to each layer of the Neural Net. This enables an efficient computation of the influence functions. Our approach is inspired by the techniques used in saliency maps [8, 9, 10, 11, 12, 13]. We perform stochastic estimation of the inverse of the Hessian using the LiSSA algorithm [20]. As our approach involves the computation of the Hessian, all operations in the Neural Network must be twice differentiable. Therefore, we used Scaled Exponential Linear Units (SELU) [21] activation functions (as opposed to Rectified Linear Units (ReLU)) and the cross-entropy loss function.

2.4. Synthetic Gaussian Dataset

We tested our extended influence functions, Eqs. (4) and (5), using simulated data. We generated two multivariate Gaussian distributions, each having two dimensions. Both distributions shared the same mean in the first dimension and had different means in the second dimension. Therefore, it should be easy to quantify the feature importance. We split the data

Table 1. AUC scores for both test datasets. The best performing algorithm is shown in bold.

Septic Patients: ROC AUC					
SMOTE Oversampling	SAPS	APACHE IVa	Logistic Regression	XGBoost	Neural Network
No	0.7717	0.7849	0.7985	0.6901	0.7969
Yes	0.7736	0.7864	0.8001	0.7199	**0.8046**
All-Comers Patients: ROC AUC					
No	0.8325	0.8451	0.8463	0.6892	0.8326
Yes	0.8330	0.8457	0.8474	0.8124	**0.8564**

into train and test sets and trained two classifiers, i.e., a logistic regression and a neural network with one hidden layer. We evaluated the correlation between the logistic regression coefficients (ground truth) and feature importance from Eq. (3).

2.5. eICU Collaborative Research Database

The eICU database is a collection of datasets from multiple intensive care units (ICUs) across the United States [22]. We chose the dataset that included first-day laboratory test results as input features ($\boldsymbol{x}$) and patient survival as labels (y). Furthermore, we split the dataset into two groups, (1) septic: dataset of patients who were diagnosed for sepsis only, and (2) all-comers: dataset of patients with all diagnosis. The septic patient data is a sub-set of the all-comers dataset. The decision to split dataset was based on the assumption that the features indicating mortality may differ depending on the diagnosis of the patient, i.e., septic or non-septic. For the septic dataset we have 19379 instances and 28 input features. For the all-comers dataset we have 148532 instances and 20 input features.

Data Pre-processing: We start with a correlation analysis of all features in the datasets. Subsequently, all features with a correlation coefficient greater than 0.9 were dropped from the dataset. The missing data were assumed to be missing at random and features with the number missing data $> 50\%$ of the total data were dropped from the dataset. For the rest of the missing data, we used multiple imputation with chained equations via iterative imputer method available in the ski-kit learn package [23].

Training, Validation, and Testing: We used k-fold cross-validation with $k = 5$ for the performance evaluation of all models. For each split, the training and testing data were standardized using the mean and the standard deviation of the training data. We observed a large class imbalance in our dataset (90%-10% in septic and 95%-5% in all-comers). To explore further and address the potential problem of class imbalance, we trained and tested two sets of models, (1) without introducing any class balancing technique, and (2) performing minority class oversampling using Synthetic Minority Oversampling Technique (SMOTE) [24]. SMOTE performs synthetic sampling using interpolation along the space between a minority instance's k-nearest-neighbors until the number of classes are equal.

In addition to SAPS and APACHE, which were available in the database, we trained three models, i.e., logistic regression, XGBoost, and a Neural Net with one hidden layer using SELU activations. We used nested grid search to find optimal hyperparameters for the logistic regression (C, Penalty, etc.) and XGBoost (max depth, learning rate, etc.) models; however, for the Neural Net, we used Bayesian optimization available in Weights and Biases package [25]. For each patient, we extracted corresponding APACHE and SAPS scores and fit a logistic regression model using the same training and testing scheme as described above for the other three models. We evaluated all five models using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metric. The models scoring highest on the ROC AUC metric were selected to evaluate the test set for each cross-validation loop.

Extraction and Comparison of Important Features: We extracted important features from each model using various techniques and compared these with features manually selected by domain experts. We created a survey that included all features from both datasets and asked ICU clinicians (n=50 for the septic dataset and n=20 for the all-comers dataset) to pick top the 10 features that they believed were the best indicators for patient mortality.

For the logistic regression model, we selected features with the highest absolute value. For XGBoost we calculated SHapley Additive exPlanations (SHAP) values for each feature and selected those with the highest values [26]. For the Neural Net, we used SHAP (via DeepLIFT) and extended influence functions to select important features. To evaluate how each of these methods performed, we found 10 features most often picked by the domain experts and counted how many features were common in the top 10 ranked by the above methods.

3. RESULTS AND DISCUSSION

3.1. Synthetic Gaussian Dataset

We evaluated how the feature importance values can change across test samples. When evaluating the center points of each

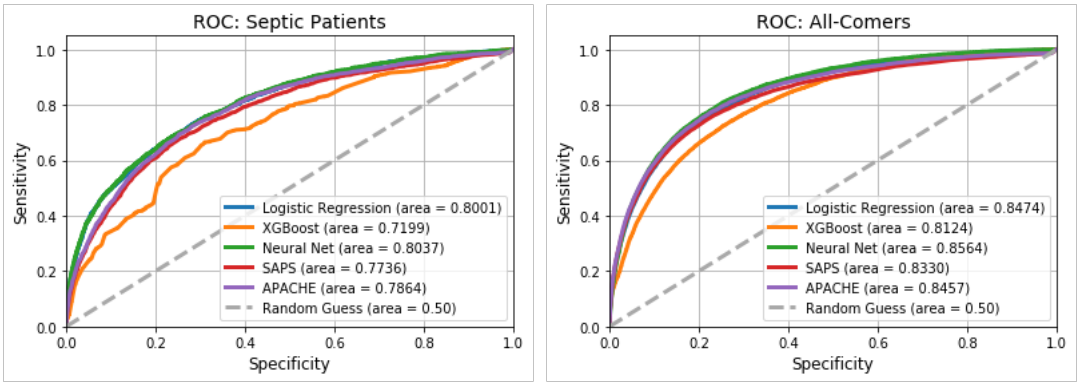

Fig. 1. ROC curves for different models for each test dataset are presented. AUC is calculated and presented in the legend for each model.

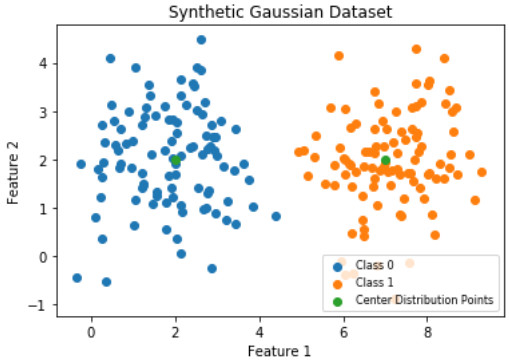

Fig. 2. Synthetic multi-variate Gaussian distributions data used for the evaluation of proposed extended influence functions.

distribution as our test point (Fig. 2) we noticed very different average feature importances that were also different from the logistic regression coefficients (Eq. 4). When the entire test set is used to compute the loss (Eq. 5), however, the feature importance values approach the logistic regression coefficients (the ground truth) with correlation values above 0.95 with $p < 0.05$.

3.2. eICU Database

In Fig. 1, we present ROC curves for both test datasets (septic and all-comers) and for all five models, Neural Net, logistic regression, XGBoost, SAPS, and APACHE. In Table 1 we present AUC scores for both tests datasets and all five models for two cases, i.e., with and without SMOTE oversampling. We note that for both datasets, the Neural Network outperforms all other models and XGBoost is the worst performing model.

If we look at each dataset by itself, the difference in performance between the mortality scores and our models is much better in the septic dataset. It appears that a larger performance boost can be attained by creating models for specific diagnosis groups rather than one model for all diagnosis groups. XGBoost consistently performed worse than all other models in all scenarios. Logistic Regression performed about as well as the traditional mortality scores which is expected due to the predictions of these scores being based on the same raw variables. Neural Nets consistently outperformed all other models when the minority class was oversampled.

In Table 3, we present count numbers for the various models and important features selected by those models for both datasets. We note that XGBoost with SHAP had more in common with the features selected by clinicians. However, XGBoost was the worst predictor of mortality as compared to other models. It is important to note that both SHAP algorithms (XGBoost and the Neural Net) extracted the control variable (patient ID) in their top 10 features. However, this was not the case for the Neural Net with influence functions, which was also the best predictor of patient mortality.

On the sepsis dataset, the runtime for the Neural Net extended influence functions was around 1.5 minutes to calculate global feature importance. However, the runtime for SHAP with Neural Net was around 220 minutes. The computational efficiency of the proposed extended influence functions was even more evident in the all-comers dataset, e.g., SHAP calculations took more than one week on a NVIDIA TITAN V GPU.

We note that there were significant differences between the top ten features chosen by the clinician as compared to those by the algorithms. The features selected by clinician included comorbidities such as metastatic cancer, AIDS, and need for immunosuppression but these were less likely to be in the algorithm feature selection lists. The differences are stark but not necessarily surprising. Physicians being humans will always have biases and blind spots, both of which are more likely with the inherent stressors of the ICU. This study is limited due to the small size of our surveys as well as known variability in clinician decision making for end of life care

Table 2. Feature ranking (FR) of septic and all-comers datasets. RF - Feature Rank (1-highest, 10-lowest), LR - Logistic Regression, XG - XGBoost, and NN - Neural Network.

Septic Dataset									
FR	Survey	LR - no SMOTE	LR - SMOTE	XG SHAP - no SMOTE	XG SHAP - SMOTE	NN SHAP - no SMOTE	NN SHAP - SMOTE	NN IFs - no SMOTE	NN IFs - SMOTE
1	hepaticfailure	CHLORIDE_max	CHLORIDE_max	PT_max	LACTATE_max	gender	SODIUM_max	LACTATE_max	CHLORIDE_max
2	LACTATE_max	BICARBONATE_min	BICARBONATE_min	LACTATE_max	BICARBONATE_min	day1motor	BICARBONATE_min	age	LACTATE_max
3	age	LACTATE_max	LACTATE_max	BICARBONATE_min	day1motor	age	diabetes	PT_max	INR_max
4	immunosuppression	ALBUMIN_min	ALBUMIN_min	BUN_max	BUN_max	BILIRUBIN_max	WBC_max	SODIUM_min	BICARBONATE_min
5	metastaticcancer	SODIUM_max	SODIUM_max	INR_max	INR_max	diabetes	SODIUM_min	SODIUM_max	PT_max
6	CREATININE	day1motor	ANIONGAP_max	PLATELET_min	ALBUMIN_min	patientunitstayid	ANIONGAP_max	BUN_max	SODIUM_max
7	INR	age	age	ALBUMIN_min	age	BICARBONATE_min	day1motor	HEMATOCRIT_min	day1motor
8	GCS-motor	ANIONGAP_max	day1motor	day1motor	BILIRUBIN_max	hepaticfailure	ALBUMIN_min	metastaticcancer	BUN_max
9	BILIRUBIN	BUN_max	SODIUM_min	HEMATOCRIT_min	patientunitstayid	SODIUM_min	BUN_max	INR_max	hepaticfailure
10	leukemia	SODIUM_min	BUN_max	GLUCOSE_min	aids	SODIUM_max	CHLORIDE_max	lymphoma	diabetes
All-Comers Dataset									
1	age	day1motor	day1motor	LACTATE_max	day1motor	ANIONGAP_max	CHLORIDE_max	day1motor	day1motor
2	immunosuppression	BICARBONATE_min	CHLORIDE_max	day1motor	LACTATE_max	gender	PLATELET_min	BICARBONATE_min	LACTATE_max
3	hepaticfailure	ALBUMIN_min	ALBUMIN_min	CREATININE_max	BUN_max	HEMATOCRIT_min	GLUCOSE_min	PT_max	INR_max
4	LACTATE_max	CHLORIDE_max	BICARBONATE_min	ALBUMIN_min	CREATININE_max	GLUCOSE_min	day1motor	ALBUMIN_min	PT_max
5	metastaticcancer	LACTATE_max	age	PT_max	ALBUMIN_min	LACTATE_max	CREATININE_max	CHLORIDE_max	ALBUMIN_min
6	CREATININE	age	LACTATE_max	BUN_max	PT_max	ALBUMIN_min	SODIUM_max	ANIONGAP_max	BICARBONATE_min
7	leukemia	SODIUM_max	SODIUM_max	INR_max	INR_max	age	BICARBONATE_min	CHLORIDE_min	PTT_max
8	PLATELET	ANIONGAP_max	ANIONGAP_max	PLATELET_min	age	BICARBONATE_min	INR_max	gender	CREATININE_max
9	BILIRUBIN	BUN_max	BUN_max	PTT_max	gender	SODIUM_min	PTT_max	SODIUM_min	ANIONGAP_max
10	aids	HEMATOCRIT_min	CREATININE_max	HEMATOCRIT_min	ANIONGAP_max	PTT_max	LACTATE_max	PTT_max	SODIUM_max

[27]. The features selected for each experiment as well as the clinician opinion can be found in Table 2.

Table 3. Common feature count by model and important feature selection algorithm type.

Model with algorithm Type	Septic	All-comers
Logistic regression coefficients	3	2
XGBoost with SHAP	5	3
NN with SHAP	4	2
NN with influence functions	4	2

4. CONCLUSION AND FUTURE WORK

The intersection of physician reasoning and machine learning is beginning to take the center stage. In our study, we attempted to create a more accurate and explainable prediction model using extended influence functions. We compared physician opinion on the top ten comorbidities, labs, and exam findings that would predict ICU mortality to that of various machine learning algorithms. Neural Networks achieved highest AUC score using the first 24-hour hospital admission features as compared to traditional mortality scores such as APACHE, and SAPS, especially in the sepsis dataset. We found that Neural Nets with extended influence functions were superior to the competing algorithms in predictive power, explanation accuracy and computation time. Further work may be able to shed some light on features that have been overlooked by domain experts for certain diagnoses.

5. REFERENCES

[1] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," *arXiv preprint arXiv:1905.05134*, 2019.

[2] Prabha Susy Mathew and Anitha S Pillai, "Big data solutions in healthcare: Problems and perspectives," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE, 2015, pp. 1–6.

[3] Maya Dewan, Naveen Muthu, Eric Shelov, Christopher P Bonafide, Patrick Brady, Daniela Davis, Eric S Kirkendall, Dana Niles, Robert M Sutton, Danielle Traynor, et al., "Performance of a clinical decision support tool to identify picu patients at high risk for clinical deterioration," *Pediatric Critical Care Medicine*, 2019.

[4] Jennifer C Ginestra, Heather M Giannini, William D Schweickert, Laurie Meadows, Michael J Lynch, Kimberly Pavan, Corey J Chivers, Michael Draugelis, Patrick J Donnelly, Barry D Fuchs, et al., "Clinician perception of a machine learning–based early warning system designed to predict severe sepsis and septic shock," *Critical care medicine*, vol. 47, no. 11, pp. 1477–1484, 2019.

[5] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, """ why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Anchors: High-precision model-agnostic explanations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[9] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[10] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[11] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[12] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3145–3153.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," *34th International Conference on Machine Learning, ICML 2017*, vol. 7, pp. 5109–5118, 2017.

[14] R Dennis Cook and Sanford Weisberg, "Characterizations of an empirical influence function for detecting influential cases in regression," *Technometrics*, vol. 22, no. 4, pp. 495–508, 1980.

[15] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1885–1894.

[16] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 27–38.

[17] Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila, "Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients," *Critical care medicine*, vol. 34, no. 5, pp. 1297–1310, 2006.

[18] Rui P Moreno, Philipp GH Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, Jean-Roger Le Gall, et al., "Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission," *Intensive care medicine*, vol. 31, no. 10, pp. 1345–1355, 2005.

[19] Jea Yeon Choi, Jae Ho Jang, Yong Su Lim, Jee Yong Jang, Gun Lee, Hyuk Jun Yang, Jin Seong Cho, and Sung Youl Hyun, "Performance on the apache ii, saps ii, sofa and the ohca score of post-cardiac arrest patients treated with therapeutic hypothermia," *PloS one*, vol. 13, no. 5, 2018.

[20] Naman Agarwal, Brian Bullins, and Elad Hazan, "Second-order stochastic optimization for machine learning in linear time," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4148–4187, 2017.

[21] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, "Self-normalizing neural networks," in *Advances in neural information processing systems*, 2017, pp. 971–980.

[22] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, pp. 180178, 2018.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[25] Lukas Biewald, "Experiment tracking with weights and biases," 2020, Software available from wandb.com.

[26] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[27] Adrienne G Randolph, Mary B Zollo, Marlene J Egger, Gordon H Guyatt, Robert M Nelson, and Gregory L Stidham, "Variability in physician opinion on limiting pediatric life support," *Pediatrics*, vol. 103, no. 4, pp. e46–e46, 1999.