Using MLP partial responses to explain in-hospital mortality in ICU

Ivan Olier
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
ORCID: 0000-0002-5679-7501

Annabel Sansom
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
A.G.Sansom@2015.ljmu.ac.uk

Paulo Lisboa
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
ORCID: 0000-0001-6365-4499

Sandra Ortega-Martorell
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
ORCID: 0000-0001-9927-3209

Abstract— **In this paper we propose to use partial responses derived from an initial multilayer perceptron (MLP) to build an explanatory risk prediction model of in-hospital mortality in intensive care units (ICU). Traditionally, MLPs deliver higher performance than linear models such as multivariate logistic regression (MLR). However, MLPs interlink input variables in such a complex way that is not straightforward to explain how the outcome is influenced by inputs and/or input interactions. In this paper, we hypothesized that in some scenarios, such as when the data noise is significant or when the data is just marginally non-linear, we could find slightly more complex associations by obtaining MLP partial responses. That is, by letting change one variable at the time, while keeping constant the rest. Overall, we found that, although the MLR and MLP in-hospital mortality model performances were equivalent, the MLP could explain non-linear associations that otherwise the MLR had considered non-significant. We considered that, although deeming higher-other interactions as disposable noise could be a strong assumption, building explanatory models based on the MLP partial responses could still be more informative than on MLR.**

Keywords— Interpretable machine learning, multilayer perceptron, neural networks, partial responses, MIMIC-III database.

I. Introduction

The Intensive Care Unit (ICU), which is where severely ill patients are treated in a hospital, is the unit with highest mortality rate in any hospital. Advances in clinical research is seeing an increased survival rate for patients in critical care and determining who is most at risk is at the core of this.

Traditional machine learning (ML) methods such as logistic regression have proved successful for predicting mortality risk in ICU patients [1]. More recently, deep learning (DL) methods have also demonstrated their ability to predict mortality in the field of critical care [2] outperforming more traditional ML methods. However, DL methods lack interpretability in that they cannot explain their predictions, otherwise known as the 'black box' problem [3]. This can be problematic when predicting mortality risk among ICU patients, as identifying the most significant risk factors is essential in saving lives.

The aim of this paper is to produce an interpretable model for predicting mortality risk among ICU patients using data collected during the first 48 hours of ICU stay. A baseline model using logistic regression has been also produced to predict the outcome of mortality using numerous variables and vital sign measurements.

An additional aim is to open the deep learning 'black box' by producing a model that can explain its predictions. We propose calculating the partial response of individual variables, using an initial multilayer perceptron to produce an interpretable model that reveals how each individual variable influences the outcome.

The rest of the paper is structured as follows: the data source and final cohort used for the analysis is explained in the next section; followed by a section that explains the logistic and multilayer perceptron methods, along with the proposed partial responses formulation. Results of the experiments were collated in the Results section, which is followed by the discussion of the results and conclusions.

II. Dataset

A. *Brief description of the data used*

The dataset used in this investigation has been extracted from the MIMIC-III ('Medical Information Mart for Intensive Care') clinical database [4]. MIMIC-III is a large, publicly available database of critically ill patients who stayed in the intensive care units of the Beth Israel Deaconess Medical Centre between 2001 and 2012.

Data includes vital signs measurements, patient demographics, medications, laboratory measurements, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, among others [4]. The variables for our study have been chosen based on a previous publication [5].

B. *Cohort selection*

Abnormal values detected during data exploration were removed, e.g. heart rate measurement below 0. We used one observation per time point (in hours since admission). As the aim of this investigation is to predict in-hospital mortality using information collected during the first 48 hours of ICU stay, any admissions with a length of stay less than 48 hours and recordings taken after 48 hours were removed. This resulted in 13% admissions, 11% subjects and 20% observations being removed. The time frame begins at 1 hour before admission to account for any prior information recorded in the ambulance.

The Glasgow Coma Scale (GCS) scores, which relate to the level of consciousness of patients with acute brain injuries were recoded as per [6]. An additional variable, GCS total score was created by summing the individual scores for each admission per time point.

The mean and standard deviation were calculated for each continuous variable. GCS scores are treated as continuous as they follow an order (ranging from deep coma to fully conscious). For true categorical variables the mode and standard deviation were calculated. Constant variables were used, one value per admission.

Missing values is a common challenge faced when analyzing ICU data as the priority during the first few hours of the stay is stabilizing the patient, hence not all variables are measured during that time. Due to this, another proportion of observations were discarded.

A mortality rate of 11.3% has been observed. It is common to observe a class imbalance in ICU data as the number of patients who die in hospital is relatively less in comparison to those that survive.

The final dataset contains 7529 observations, 25 predictor variables and one binary response (1 = death before discharge, 0 = survival until discharge). The data was split in order to evaluate the performance of the models using multiple techniques and performance metrics – 80% of data was used for training the models and 20% for test.

III. Methods

A. Multivariate Logistic Regression

Multivariate (multiple) logistic regression (MLR) uses the logistic function to model the risk probability of an outcome as a linear combination of its input variables [7]. MLR has the following general formulation:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots . + \beta_p X_p \qquad (1)$$

where β_i are the coefficients. MLR has become the gold standard approach to multivariate explanatory modelling as associations between inputs and output variables can be easily explained from its coefficients.

B. Multilayer Perceptron

Multilayer perceptron (MLP) is a general-purpose type of feedforward neural network. In an MLP, inputs are densely connected via processor units called neurons, each of them applies a (usually) non-linear activation function to linearly combined inputs. An MLP neuron is formulated as follows:

$$y_j = \psi\left(\sum_i w_i x_i + w_0\right) \qquad (2)$$

where ψ is the activation function and w_i the connection weights. Neuron outputs can, subsequently, become the inputs of other neurons. Ultimately, an MLP uses layers of neurons to build non-linear, interlinked, associations between input and output variables. Recently, it was shown that by stacking several layers, MLPs had the ability of learning highly accurate models by finding complex representations of the data. This is what is known as Deep Learning (DL) [8]. However, as a tradeoff, DL models cannot easily explain associations between inputs and outputs, making them less suitable for explanatory modelling.

C. Partial Responses

Partial responses are calculated by feeding one input at a time through the MLP derived above, so that it is possible to determine the contribution of each variable to the log of the response.

In order to model the contribution to the logit using univariate terms, the model is defined as follows:

$$\log(Y) = \varphi(0) + \sum_i \varphi_{i(x_i)} + \varepsilon \qquad (3)$$

Where $\varphi(0)$ is the error that is calculated when all inputs are equal to 0, and $\varphi_i(x_i)$ represents the partial responses of variable i (individually) and ε represents the higher order terms.

The partial responses will be calculated and visualised to assess the contribution on the logit function for each of the predictor variables [9], [10]. The logit function represents mortality risk. A positive contribution indicates increased risk while a negative contribution indicates a decreased risk. The closer the contribution is to 0, the less influence the variable has on the outcome.

IV. Results

A. MLR and MLP results

The multivariate logistic regression model is shown in table 1, including the corresponding 95% confidence intervals. This model was created using the whole dataset, allowing further comparisons. The ROC AUC obtained on the test set was 0.803.

TABLE I. Results of the logistic regression model

	Estimate	**Std error**	**Z value**	**P value**	**Lower 95% CI**	**Upper 95% CI**
Intercept	35.284	3.564	9.901	<1e-4	28.321	42.294
Diastolic BP (Mean)	-0.012	0.010	-1.156	0.248	-0.032	0.008
Diastolic BP (St Dev)	0.038	0.019	1.945	0.052	0.000	0.076
Mean BP (Mean)	-0.021	0.013	-1.651	0.099	-0.045	0.005
Mean BP (St Dev)	-0.031	0.022	-1.424	0.154	-0.075	0.011
Systolic BP (Mean)	0.001	0.004	0.207	0.836	-0.008	0.009
Systolic BP (St Dev)	0.016	0.010	1.589	0.112	-0.004	0.036
GCS Eye (Mean)	-4.691	7.517	-0.624	0.533	-18.083	5.466
GCS Eye (St Dev)	0.283	0.243	1.165	0.244	-0.195	0.758
GCS Motor (Mean)	-4.093	7.518	-0.544	0.586	-17.489	6.065
GCS Motor (St Dev)	0.084	0.222	0.378	0.706	-0.355	0.517
GCS Verbal (Mean)	-4.097	7.519	-0.545	0.586	-17.493	6.060
GCS Verbal (St Dev)	0.367	0.208	1.762	0.078	-0.049	0.772
GCS Total (Mean)	3.920	7.518	0.521	0.602	-6.236	17.314
GCS Total (St Dev)	-0.514	0.215	-2.396	0.017	-0.932	-0.090

	Estimate	Std error	Z value	P value	Lower 95% CI	Upper 95% CI
Glucose (Mean)	-0.001	0.001	-0.880	0.379	-0.003	0.001
Glucose (St Dev)	0.003	0.002	1.840	0.066	0.000	0.006
Heart Rate (Mean)	0.017	0.003	5.833	<1e-4	0.011	0.023
Heart Rate (St Dev)	-0.011	0.009	-1.196	0.232	-0.029	0.007
O2 Saturation (Mean)	-0.108	0.024	-4.490	<1e-4	-0.155	-0.061
O2 Saturation (St Dev)	0.052	0.025	2.103	0.035	0.003	0.100
Respiratory Rate (Mean)	0.083	0.011	7.859	<1e-4	0.063	0.104
Respiratory Rate (St Dev)	0.023	0.024	0.951	0.342	-0.024	0.069
Temperature (Mean)	-0.632	0.071	-8.874	<1e-4	-0.773	-0.493
Temperature (St Dev)	0.223	0.162	1.375	0.169	-0.096	0.541
Weight	-0.010	0.002	-5.272	<1e-4	-0.014	-0.006

The MLP trained on the same data attained a ROC AUC of 0.8102 and a loss of 0.315. The fully connected neural network performed slightly better than the logistic regression model shown in table 1.

B. *Partial Responses*

Figures 1-4 are histograms of the frequency distribution for each variable, the red line represents the mortality risk function. All plots and partial responses are calculated using the full dataset for appropriate comparisons with the logistic regression model in table 1.

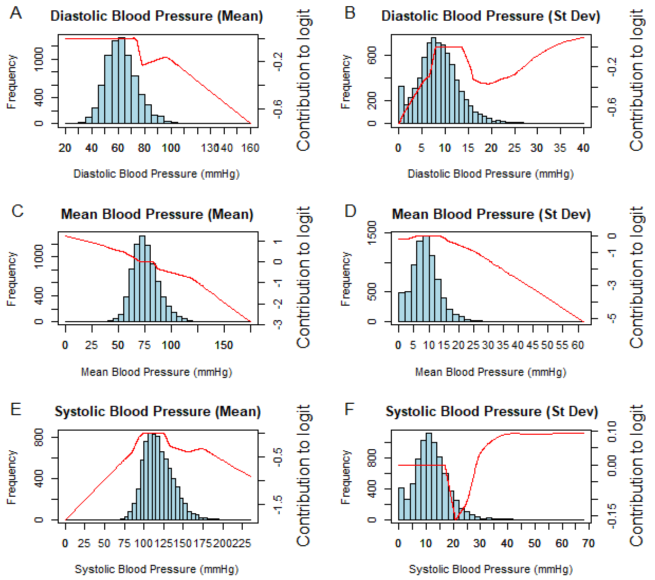

Fig. 1. Blood pressure risk functions

In Fig. 1 we can see there are decreasing risk trends for larger mean values of diastolic (Fig. 1A) and mean (Fig. 1C) BP, the risk increases slightly for diastolic BP values between 80 and 100mmHg. Although the idea that lower BP being associated with higher risk is counterintuitive, it may lead to long term illnesses such as heart failure and cardiac decomposition [11], rather than short term mortality risk. Increased risk trends exist for mean systolic BP (fig 1E) up to 100mmHg where the risk steadily decreases for values above 125 mmHg. Normal systolic BP is between 90 and 140mmHg, suggesting no risk for patients with normal systolic BP as the risk function plateaus at 0.

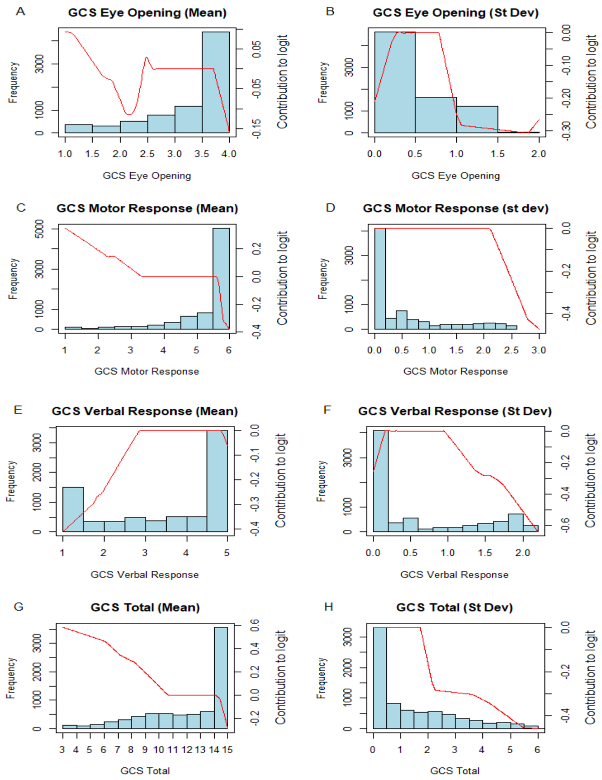

Fig. 2. GCS risk functions

From Fig. 2 there are decreasing risk trends for mean GCS motor (Fig. 2C) and total (Fig. 2G), with GCS total scores below 10 being associated with mortality risk above 0. This supports the findings in [12] as low GCS scores are highly correlated with mortality risk, with patients with GCS total scores below 10 having poor chance of a favorable outcome.

There is an overall decrease in risk as GCS Eye (Fig. 2A) score increases; however, risk increases sharply for scores between 2 and 3. For verbal response (Fig. 2E) there is an overall increase in risk as GCS score increases plateauing at 3 with a slight decrease for a score of 5. The results for GCS Eye (P=0.533) and verbal (P=0.586) mean scores are not significant in the logistic model in table 1 as it cannot identify these increases and decreases in risk. From Fig. 2, these results have non-linear relationships with the outcome, as a change in the predictor variable does not lead to a constant change in the outcome, the MLP can model these non-linear relationships well. The 95% CI's for mean GCS eye and verbal response both crossed the line of no effect in table 1, this indicates that the risk increases and decreases for different GCS scores as shown in Fig. 2.

The changes over all GCS scores for patients during the first 48 hours is represented by the standard deviations in the right-hand panel of Fig. 2. This change could indicate improvements in a patient's condition rather than decline, that would increase the patient's chance of survival. Overall, a larger change in GCS motor score has no effect on mortality risk (Fig. 2D), decreasing for scores above 2.5 although there is not enough data for scores above 2.5 to give reliable results. The logistic regression model shows the result for GCS total standard deviation is slightly significant (P=0.017), showing a clear negative correlation between score and risk (Fig. 2H).

For GCS Eye opening and verbal response standard deviations, there is a non-linear relationship between variable and outcome. For GCS eye standard deviation (Fig. 2B) the risk increases to 0 for scores of 0.3 where it plateaus at 0.7, decreasing sharply to 1 and steadily to 1.7 where it slightly increases. Results for GCS eye standard deviations above 1.5 are not reliable due to the lack of data. For GCS verbal response standard deviation (Fig. 2F), the risk increases to 0 for score of 0.2 where it plateaus to a score of 1 and decreases. All GCS standard deviation scores have little to no impact at increasing mortality risk as the contribution to logit does not exceed 0.

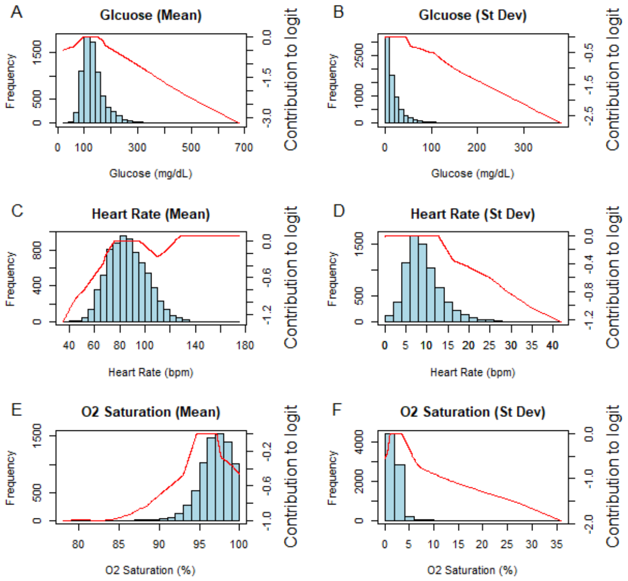

Fig. 3. Glucose, Heart rate and O2 Saturation risk functions

Normal glucose levels are between 80 and 130mg/dL, from Fig. 3 there is a slight increase in risk from 0 to 80mg/dL. Higher mean glucose levels (Fig. 3A) indicate lower mortality risk. There is no risk for patients with normal glucose levels as the contribution to the logit plateaus at 0. The risk decreases after 150mg/dL, although there is not enough data for glucose levels over 200mg/dL to provide reliable results, this supports the evidence in [13] indicating that glucose has little impact on mortality.

The logistic model shows mean glucose decreases mortality risk but not significantly (P=0.379), as it is a linear model it cannot identify the increase in risk from 0 to 80mgdL. The information shown in the plot is more informative as the MLP is very good at modelling non-linear relationships such as this.

Increased heart rate increases mortality risk (Fig. 3C). Heart rate measurements are taken when patients are at rest; therefore, a low resting heart rate may suggest a patient is more physically active and has better cardiovascular health, lowering their risk of mortality. Furthermore, a high resting heart rate may indicate heart failure increasing mortality risk. There is a slight decrease in risk from 100 to 120bpm, this is where the amount of data starts to decrease. Overall, the graph shows some evidence that agrees with the research carried out by Kara, 2016 in section 2.8 that higher heart rate increases mortality risk.

Higher change in heart rate decreases mortality risk (Fig. 3D) after a standard deviation of 15, this could show heart rate increasing or decreasing significantly over 48 hours. Although this result is unreliable as there is a lack of data for standard deviations above 15. Patients with a heart rate standard deviation of less than 15 are at no risk of mortality, as the risk function plateaus at 0.

Normal O2 saturation is between 94 and 98%. Patients with normal O2 saturation are at no risk, the risk decreases for saturations above 98% (Fig. 3E). This result disagrees with [14], suggesting that increased O2 concentrations can lead to oxygen toxicity.

Change in O2 saturation does not significantly increase risk as the contribution does not exceed 0, patients with less change in O2 saturation are at lower risk (Fig. 3F). This result is not reliable as there is limited data for standard deviations more than 5%.

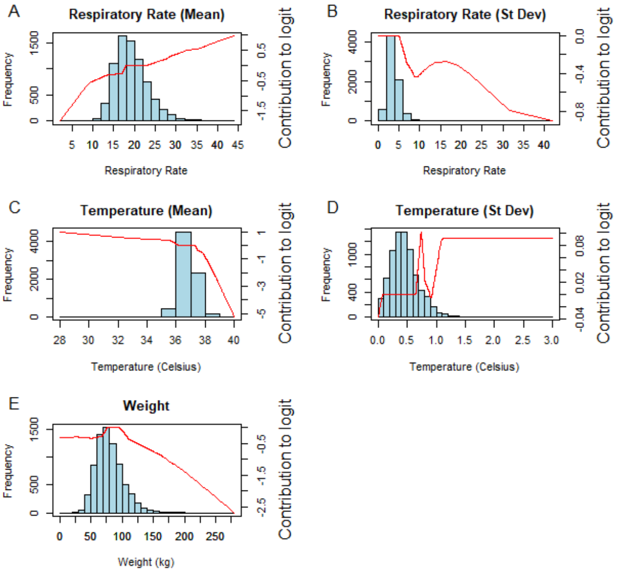

Fig. 4. Respiratory Rate, Temperature and weight risk functions

The results in Fig. 4A show no risk for patients with normal mean respiratory rates between 12 and 25 breaths per minute as the risk function plateaus at 0. Risk increases for respiratory rates exceeding 25 breaths per minute, although this result is unreliable due to the lack of data for respiratory rates above 35.

Overall, larger change in respiratory rate decreases mortality risk. There is limited data for standard deviations above 10 meaning the curve representing an increase and decrease in risk after this point is unreliable (Fig. 4B).

Higher temperature is correlated with lower mortality risk (Fig. 4C), supporting evidence given by [15] that lower temperature increases mortality risk. The result for mean temperature in the logistic model is highly significant (P<0.0001) as there is a clear decrease in risk as temperature increases, the 95% CI in table 1 supports this as it does not cross the line of no effect, containing only negative values.

More change in temperature increases mortality risk (Fig. 4D). It is normal for body temperature to change up to 0.5

degrees throughout the day, anything more than this can indicate other problems. The line plateaus at 0 for standard deviation between 0 and 0.5 showing no risk, peaking at the highest risk of 0.09 for a standard deviation of 0.7 where the risk sharply declines up to standard deviation 1, increasing and plateauing at 0.08 for standard deviations above 1.2. This result appears significant; however, the logistic model result was not significant (P = 0.169) as the risk increases and decreases as temperature changes. The MLP is non-linear and can detect changes in risk for different temperature standard deviation values, showing more informative results.

Lower weight shows the highest risk (Fig. 4E). From low to average weight it is observed that the risk rises slightly, this is where most data is concentrated. The risk is at its lowest for obese patients, the pattern for average to high weight agrees with the 'obesity paradox' explained in section 2.8. The results suggest that low weight increases mortality risk, supporting the evidence given by [16] that mortality rate is significantly higher in underweight patients.

V. Discussion and conclusions

The aim of this paper was to produce an interpretable model for predicting mortality risk among ICU patients using data collected during the first 48 hours of ICU stay. We used multivariate logistic regression and multilayer perceptrons. In one hand, the implementation of MLR models are straightforward, however, it was shown that they sometimes cannot capture all the possible associations between inputs and outputs. This is due to the fact that the MLR output is a linear combination of the inputs. Therefore, if there was a non-linear link, MLR simply would not be able to properly identify it. On the other hand, MLPs have the ability to build non-linear maps, but its downside is interpretability. Hence, explaining associations is not as straightforward as with MLR.

We hypothesized that in some scenarios, such as when the data noise is significant or when the data is just marginally non-linear, we could find slightly more complex associations by obtaining MLP partial responses. That is, by letting change one variable at the time, whilst keeping constant the rest. If the data is not too complex, we could assume that higher order interactions between input variables could be disregarded as noise.

We applied our approach to find in-hospital mortality risk factors in UCI patients. Overall, we found that, although MLR and MLP model performances were equivalent, the MLP could explain non-linear associations that otherwise the MLR had considered non-significant. We considered that, although deeming higher-other interactions as disposable noise could be a strong assumption, building explanatory models based on the MLP partial responses could still be more informative than on MLR.

Although this paper should be seen as preliminary research, we considered we are in the right path towards developing proper interpretable neural networks. Immediate future work will concentrate on allowing for higher-order interactions.

References

[1] Z. Zhao *et al.*, "Prediction model and risk scores of ICU admission and mortality in COVID-19," *PLoS One*, vol. 15, no. 7, p. e0236618, Jul. 2020, doi: 10.1371/journal.pone.0236618.

[2] S. Y. Kim *et al.*, "A deep learning model for real-time mortality prediction in critically ill children," *Crit. Care*, vol. 23, no. 1, p. 279, Aug. 2019, doi: 10.1186/s13054-019-2561-z.

[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[4] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016, doi: 10.1038/sdata.2016.35.

[5] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Sci. Data*, vol. 6, no. 1, pp. 1–18, Dec. 2019, doi: 10.1038/s41597-019-0103-9.

[6] G. Teasdale and B. Jennett, "ASSESSMENT OF COMA AND IMPAIRED CONSCIOUSNESS. A Practical Scale," *Lancet*, vol. 304, no. 7872, pp. 81–84, Jul. 1974, doi: 10.1016/S0140-6736(74)91639-0.

[7] G. James, D. Witten, T. Hastie, and R. Tibishirani, *An Introduction to Statistical Learning*. 2013.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[9] P. J. G. Lisboa, S. Ortega-Martorell, and I. Olier, "Explaining the Neural Network: A Case Study to Model the Incidence of Cervical Cancer," in *Communications in Computer and Information Science*, 2020, vol. 1237 CCIS, pp. 585–598, doi: 10.1007/978-3-030-50146-4_43.

[10] P. J. G. Lisboa, S. Ortega-Martorell, S. Cashman, and I. Olier, "The Partial Response Network: a neural network nomogram," Aug. 2019, Accessed: Sep. 21, 2020. [Online]. Available: http://arxiv.org/abs/1908.05978.

[11] H. A. Christian, "Drug treatment of cardiac decompensation," *J. Am. Med. Assoc.*, vol. 108, no. 1, pp. 44–46, Jan. 1937, doi: 10.1001/jama.1937.92780010002011.

[12] J. Leitgeb *et al.*, "Glasgow Coma Scale score at intensive care unit discharge predicts the 1-year outcome of patients with severe traumatic brain injury," *Eur. J. Trauma Emerg. Surg.*, vol. 39, no. 3, pp. 285–292, Jun. 2013, doi: 10.1007/s00068-013-0269-3.

[13] L. A. Van Vught, R. Holman, E. De Jonge, N. F. De Keizer, and T. Van Der Poll, "Diabetes is not associated with increased 90-day mortality risk in critically ill patients with sepsis," *Crit. Care Med.*, vol. 45, no. 10, pp. e1026–e1035, Oct. 2017, doi: 10.1097/CCM.0000000000002590.

[14] S. Kluge and J. Grensemann, "Methodik," *Dtsch. Arztebl. Int.*, vol. 115, no. 27–28, pp. 455–462, Jul. 2018, doi: 10.3238/arztebl.2018.0455.

[15] H. M. Schell-Chaple *et al.*, "Body temperature and mortality in patients with acute respiratory distress syndrome," *Am. J. Crit. Care*, vol. 24, no. 1, pp. 15–23, Jan. 2015, doi: 10.4037/ajcc2015320.

[16] J. D. Finkielman, O. Gajic, and B. Afessa, "Underweight is independently associated with mortality in post-operative and non-operative patients admitted to the intensive care unit: A retrospective study," *BMC Emerg. Med.*, vol. 4, no. 1, p. 3, Oct. 2004, doi: 10.1186/1471-227X-4-3.