Towards interpretable machine learning for clinical decision support

Bradley Walters
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
B.Walters@2019.ljmu.ac.uk

Sandra Ortega-Martorell
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
S.OrtegaMartorell@ljmu.ac.uk

Ivan Olier
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
I.A.OlierCaparroso@ljmu.ac.uk

Paulo J. G. Lisboa
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, UK
p.j.lisboa@ljmu.ac.uk

***Abstract*— A major challenge in delivering reliable and trustworthy computational intelligence for practical applications in clinical medicine is interpretability. This aspect of machine learning is a major distinguishing factor compared with traditional statistical models for the stratification of patients, which typically use rules or a risk score identified by logistic regression.**

We show how functions of one and two variables can be extracted from pre-trained machine learning models using anchored Analysis of Variance (ANOVA) decompositions. This enables complex interaction terms to be filtered out by aggressive regularisation using the Least Absolute Shrinkage and Selection Operator (LASSO) resulting in a sparse model with comparable or even better performance than the original pre-trained black-box.

Besides being theoretically well-founded, the decomposition of a black-box multivariate probabilistic binary classifier into a General Additive Model (GAM) comprising a linear combination of non-linear functions of one or two variables provides full interpretability. In effect this extends logistic regression into non-linear modelling without the need for manual intervention by way of variable transformations, using the pre-trained model as a seed.

The application of the proposed methodology to existing machine learning models is demonstrated using the Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forests (RF) and Gradient Boosting Machines (GBM), to model a data frame from a well-known benchmark dataset available from Physionet, the Medical Information Mart for Intensive Care (MIMIC-III). Both the classification performance and plausibility of clinical interpretation compare favourably with other state-of-the-art sparse models namely Sparse Additive Models (SAM) and the Explainable Boosting Machine (EBM).

Keywords*— *Interpretability, Generalised Additive Neural Networks, Self-Explaining Neural Networks, Sparse Additive Model, Machine explanation, Multi-Layer Perceptron

I. INTRODUCTION

Artificial intelligence has radically increased the accuracy of inferences made from complex data. However, these algorithms are often difficult to understand by users from other domains and their operation can be opaque. This is of practical importance since models driven by observational data can be difficult to correct for bias and other artifacts that may be present in the data. In clinical decision support, there is a need for reliable and transparent non-linear models whose plausibility can be cross-checked against clinical expertise. Moreover, this requires more than local explanation e.g. by feature attributions. It involves interpretation in the sense the weight that individual input variables have on the response of the model needs to be apparent across the complete range of model inputs [1].

This paper is concerned with clinical decision support where datasets are generally noisy and clinical reasoning often relies on independent effects of individual variables or pairwise interactions of the type routinely modelled by logistic regression using product terms. In fact, it has been claimed that linear models in the medical domain show comparable levels of classification performance to machine learning for tabular data [2]. That is the focus of this paper and we benchmark logistic regression against black-box and glass-box models in a benchmark clinical application to model mortality in the Medical Information Mart for Intensive Care (MIMIC-III).

While rule-based models are used in medical applications [3] it is commonplace to apply logistic regression [2]. Both models are *de facto* standards for interpretability of algorithms for patient stratification and risk prediction, as they are generally accepted by clinicians and used in routine clinical practice. A major driver for clinical take-up is the transparency of the models, in the sense that the flow of information from input to response is transparent and immediately understood. This is a critical aspect of the software development lifecycle in particular the V&V framework which requires not just Verification of the model (is the model built right?) but also Validation (is it the right model?) [4]. The last step involves the plausibility of the algorithm when checked against the domain knowledge of the end-user, in this case, the clinician. And that, in turn, involves interpretability.

A common set of criteria to evaluate interpretability involves the *three "Cs":* completeness – coverage of the explanation for all possible instances; correctness – plausible validity to generate trust among end-users; compactness – using in some sense a minimal set of explanatory rules [5]. All of these criteria are met by sparse General Additive Models (GAMs) provided that the component terms involve only a small number of variables.

1.1 Related work

Formal methods to represent a function of several variables using sub-functions of fewer variables already exist. This can be achieved using a functional Analysis of Variance (ANOVA) Decomposition [6] of which there is an extension that is specific to Support Vector Machines using iterative re-weighing methods [7]. Unlike [6], we use a Dirac measure resulting in the so-called anchored ANOVA decomposition and we focus on classification rather than regression.

Our method relates also to Generalised Additive Neural Networks (GANNs), sometimes also called Self-Explanatory Neural Networks (SENN), which have a long history [8-11]. Models [8-10] have the structure of a GAM when applied to shallow neural networks, similar to our models. However, none of these papers deals with the derivation of the model structure.

More recently, the *interpretML* project introduced the Explainable Boosting Machine (EBM) [12] which does include automatic identification of univariate effects and bivariate interactions. The EBM is a benchmark in our study.

A further benchmark, from mainstream statistics, is Sparse Additive Models [13] which rely on splines rather than distributed processing.

1.2 Novel contribution

We propose a method that applies to any black-box probabilistic binary classifier including the Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forests (RF) and Gradient Boosting Machines (GBM). The first innovative step is to use the ANOVA decomposition applied to a pre-trained black box to elicit the structure of a GAM/GANN/SENN. This step is essential to the method because the component terms in the ANOVA decomposition are required before model selection can take place, since it selects which components are correlated with the target class, with statistical significance. The proposed method is then efficient for selecting not just univariate terms but also bivariate interactions. Having both first and second-order effects present in the model gives it much higher performance while maintaining full interpretability in the sense of the *three "Cs"* above.

Secondly, we show that in the case of the MLP we can refine the interpretable model by mapping the weights of the pre-trained model onto a new model, structured as a GANN/SENN. This model is then further trained to optimise its performance.

Third, we benchmark our method against state-of-the-art interpretable models namely the SAM and EBM, which are examples of the two main classes of sparse models, parametric and non-parametric. Furthermore, we compare the classification performance also against the MLP, SVM, RF and GBM in their original black-box configurations.

II. METHOD

We start with the probability density function of the posterior distribution of class membership, P(C|x), which is the output of a pre-trained probabilistic model. In common with GAMs we focus on the logit, which is calculated from the model output using the inverse of the sigmoid link function.

2.1 ANOVA decomposition

The first novelty of the paper is to apply an anchored ANOVA decomposition [6] to extract component functions of fewer variables from the logit, which is a multivariate function. We call these component functions 'Partial Responses (PR)'. These are orthogonal functions derived by setting the values of all variables except one or two at the anchor value, which in our case is the median of the data. These are numerically the same as the partial dependency plots for the univariate functions only but not for the bivariate functions. In the anchored decomposition the bivariate functions have the property that they are exactly zero when either variable takes the anchor value. In other words, this generalises the property of the usual interaction term in logistic regression $x_1.x_2$, which also vanishes whenever either term is 0. The terms in the anchored ANOVA decomposition form the functional components of an additive model, hence calling them Partial Responses.

We can assume without loss of generality that the median of the data is mapped onto the origin, hence the median point corresponds to a vector with all 0s. Therefore the logit value then takes the value $logit\big(P(C|0)\big)$. Similarly, if all of the variables except x_i are set to their median values, then the corresponding values of $logit\left(P\big(C|(0,..,x_i,..,0)\big)\right)$ represent a function of just that one variable. The same principle applies when only two variables are not *0,* then three, etc.

In the following, eq. (1) is the constant term, eq. (2) calculates the sum of all the terms in the Taylor expansion involving only x_i, so this is a non-linear function, and eq. (3) calculates the terms involving the interaction between x_i and x_j.

$$\varphi(0) = logit\big(P(C|0)\big) \tag{1}$$

$$\varphi_i(x_i) = logit\left(P\big(C|(0,..,x_i,..,0)\big)\right) - \varphi(0) \tag{2}$$

$$\begin{aligned}\varphi_{ij}(x_i,x_j) = logit\left(P\left(C|\big(0,..,x_i,..,x_j,..0\big)\right)\right)\\ - \varphi_i(x_i) - \varphi_j(x_j) - \varphi(0)\end{aligned} \tag{3}$$

The ANOVA decomposition anchored at *0* is as follows:

$$\begin{aligned}logit\big(P(C|x)\big) &= \log\left(\frac{P(C|x)}{1-P(C|x)}\right)\\ &= \varphi(0) + \sum_i \varphi_i(x_i)\\ &+ \sum_{i\neq j} \varphi_{ij}(x_i,x_j) + \cdots\\ &+ \sum_{i_1\neq\cdots\neq i_d} \varphi_{\,i_1\ldots i_d}(x_{i_1},\ldots,x_{i_d})\end{aligned} \tag{4}$$

where the general form of the terms in (4) is a recursive function of nested subsets of the covariate indices $\{i_1,\ldots,i_n\}$:

$$\begin{aligned}&\varphi_{\,i_1\ldots i_n}(x_{i_1},\ldots,x_{i_n})\\ &= logit\left(P\big(C|x_{i_1},\ldots,x_{i_n}\big)\right)\\ &- \sum_{\{i_1\neq\cdots\neq i_{n-1}\}} \varphi_{\,i_1\ldots i_{n-1}}(x_{i_1},\ldots,x_{i_{n-1}}) - \varphi(0)\end{aligned} \tag{5}$$

Note that the decomposition (4) has a *finite* number of terms, each of which represents an infinite summation from the Taylor expansion. Furthermore, the summation (4) *exactly* matches the original $logit\big(P(C|x)\big)$, therefore (4) is not an equation but an *identity*. At this point, there is no approximation, only the original multivariate function broken down into 2^d component functions, all but one of which have fewer variables. The approximation comes in the next step.

2.2 Model selection with the LASSO

The second step in the proposed method is to retain, from the ANOVA decomposition of the logit of the pre-trained classifier, only the terms that involve just one or two non-zero variables. This amounts to truncating the decomposition (4) followed by a re-calibration. To do this, the terms retained from (4) become input variables in a linear model which is the logistic regression Least Absolute Shrinkage and Selection Operator (LASSO) [14]. This is a powerful feature selection method that scales well for a large number of inputs. It uses L_1 regularisation to carry out feature selection, as the coefficients of each input gradually slide towards zero to result in a sparse model.

2.3 Additional step for the MLP

If the original black-box model is an MLP, it is possible to construct a GANN/SENN to replicate the output of the logistic Lasso by replication of the weights from the MLP multiplied by the coefficients of the Lasso. The derivation of the Partial Response Network (PRN) proceeds as follows:

1. Train an MLP for binary classification;
2. Obtained the univariate and bivariate partial responses in Eqs. (1)-(4).
3. Apply the Lasso to the partial responses;
4. Construct a second MLP as a linear combination of the partial responses to replicate the functionality of the Lasso. Each partial response, whether univariate or bivariate, is represented by a modular structure comprising the same number of hidden nodes as the original MLP. The modules are assembled into a single multi-layer structure represented as a GANN, shown in fig. 1.
5. Re-train the resulting multi-layer network by gradient descent. This results in the PRN.
6. Orthogonal Partial Responses can be obtained from the PRN and fed into the Lasso, leading to the PRN-Lasso.

The mapping of the partial responses onto the GANN requires matching the weights and bias terms as follows:

- Univariate partial responses:

$$v_j \rightarrow \beta_i * v_j \quad (6)$$

$$v_0 \rightarrow \beta_i * \Big(v_0 - logit\big(P(C|0)\big)\Big) \quad (7)$$

- Bivariate partial responses comprise two univariate and a bivariate block:

Univariate block weights:

$$v_j \rightarrow (\beta_k - \beta_{kl}) * v_j \quad (8)$$

$$v_0 \rightarrow (\beta_k - \beta_{kl}) * \Big(v_0 - logit\big(P(C|0)\big)\Big) \quad (9)$$

Bivariate block weights:

$$v_j \rightarrow \beta_{kl} * v_j \quad (10)$$

$$v_0 \rightarrow \beta_{kl} * \Big(v_0 - logit\big(P(C|0)\big)\Big) \quad (11)$$

The input weights are the same as for the original, pre-trained MLP; the output weights correspond to the labels in fig 1, and the terms β_k, β_{kl} are the Lasso parameters for each partial response.

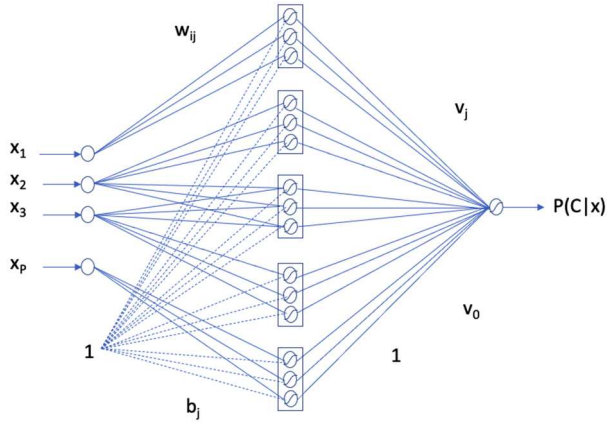


Fig.1 Structure of a Generalised Additive Neural Network (GANN), also known as a Self-Explanatory Neural Network (SENN). Each univariate effect, which we call a partial response, is modelled by a path with a separate block of hidden units. Bivariate terms involve three blocks of hidden units, one for each input and one receiving both inputs. The responses are added to make the input to the output node, i.e. the logit (P(C|x)).

III. EXPERIMENTAL RESULTS

3.1 MIMIC III data

The MIMIC-III clinical database [15] 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. The database is comprehensive and includes vital signs measurements, patient demographics, medications, procedure codes, diagnostic codes, laboratory measurements, imaging reports, hospital length of stay, and survival data, among others [15]. The variables for our study have been chosen based on a previous publication [16].

Outliers were removed during data cleaning e.g. heart rate measurements below 0. We used information from the first 48 hours of the patients being admitted to ICU to model in-hospital mortality, also including the hour before the ICU admission for any prior information recorded in the ambulance.

In the cases where variables were time series, e.g. heart rate, we first calculated hourly means and then extracted the overall mean and standard deviation of each of these variables.

The Glasgow Coma Scale (GCS) scores, which relate to the level of consciousness of patients with acute brain injuries were recorded following a standard clinical protocol [17]. GCS scores are treated as continuous since they are ordered from deep coma, at low scores, to fully conscious for high scores.

Missing values are common in routinely collected clinical data. In this dataset, they were imputed with the same methods as [16] namely using mean values. Patient

records where the level of missingness exceeded 30% were discarded.

The overall mortality rate over the complete dataset is 11.3%. This is used in this study to illustrate how the proposed methods are robust against class imbalance, which is a common feature in clinical datasets. Moreover, our study measures calibration since this is a critical feature for the interpretation of posterior probabilities for patient stratification by risk. We have not extended the study into actual stratification, but calculate the underpinning calibrated risk scores and correlate them with the additive response components for each statistically significant variable and pairwise interaction. To our knowledge, this level of analysis of this dataset is not available in the published literature, and it is also seldom published for clinical datasets generally, even though it is an essential component of performance validation for any probabilistic binary classifier for decision support in a high-stakes application.

The final dataset contains 7,532 observations (ICU patient admissions), 14 predictor variables and one binary response (1 = death before discharge, 0 = alive at discharge).

The study design involves splitting the data into three elements: a training dataset (n=4,519) and a validation dataset (n=1,506) which, together, form the model derivation database. This is used for model estimation and optimisation. However, the performance estimates may be optimistic on the validation dataset. Therefore, there is a third dataset, the test dataset (n=1,507).

For each algorithm, a single model selected to be optimal as described in the next section was taken forward and applied to the out of sample dataset. This provides an unbiased estimate of generalisation performance. This aspect of our study is central to determining how well black boxes perform compared with the baseline models, Logistic Regression, SAM and EBM, and also with the interpretable models derived from pre-trained models.

The data are standardized by an affine transformation that consists of shifting the median to zero and scaling to unit variance.

3.2 Application of ANOVA / Partial Response (PR) Methods

This section benchmarks the classification performance of the PR models against two interpretable models, EBM [12] and SAM [13], as well as three state-of-the-art machine learning algorithms, GBM [18], SVM [19] and RF [20].

For each partial response model, we include two variants labelled 1 & 2 according to the selection of Lasso parameters: 1) best AUC on the validation set and 2) best AUC minus 1 standard deviation, which results in a sparser model.

Our results are shown in Tables I & II, with only mean values of each covariate, and using both the mean and standard deviation. The 95% confidence intervals of the AUC are shown in brackets.

TABLE I. CLASSIFICATION PERFORMANCE FOR MIMIC-III DATA WITH INPUTS AS MEANS ONLY. C: NUMBER OF COMPONENTS

Model	C	Training AUC	Validation AUC	Test AUC
LR	9	0.774 (0.753, 0.796)	0.790 (0.752, 0.827)	0.785 (0.752, 0.818)
SAM	9	0.735 (0.711, 0.758)	0.742 (0.704, 0.780)	0.739 (0.702, 0.775)
EBM	19	0.828 (0.804, 0.850)	0.805 (0.764, 0.847)	0.790 (0.751, 0.829)
Black box models				
SVM	9	0.729 (0.705, 0.752)	0.726 (0.683, 0.768)	0.713 (0.674, 0.752)
RF	9	0.945 (0.935, 0.955)	0.806 (0.771, 0.841)	0.782 (0.747, 0.816)
GBM	9	0.813 (0.793, 0.833)	0.802 (0.767, 0.838)	0.787 (0.753, 0.820)
MLP	9	0.809 (0.786, 0.833)	0.790 (0.747, 0.833)	0.802 (0.763, 0.840)
Partial response models				
prSVM1	34	0.771 (0.747, 0.794)	0.778 (0.737, 0.818)	0.763 (0.727, 0.800)
prSVM2	19	0.755 (0.731, 0.779)	0.769 (0.730, 0.808)	0.755 (0.717, 0.792)
prRF1	43	0.923 (0.913, 0.934)	0.774 (0.735, 0.814)	0.778 (0.743, 0.813)
prRF2	36	0.905 (0.893, 0.917)	0.769 (0.728, 0.809)	0.775 (0.739, 0.811)
prGBM1	10	0.809 (0.789, 0.829)	0.804 (0.769, 0.838)	0.785 (0.751, 0.818)
prGBM2	6	0.786 (0.765, 0.807)	0.795 (0.761, 0.830)	0.768 (0.733, 0.803)
PRN	11	0.795 (0.771, 0.819)	0.791 (0.748, 0.834)	0.805 (0.768, 0.844)
PRN-Lasso	11	0.795 (0.771, 0.819)	0.789 (0.746, 0.832)	0.807 (0.768, 0.845)

TABLE II. CLASSIFICATION PERFORMANCE FOR MIMIC-III DATA WITH MEANS AND STANDARD DEVIATIONS. C: NUMBER OF COMPONENTS

Model	C	Training AUC	Validation AUC	Test AUC
LR	14	0.790 (0.768, 0.812)	0.801 (0.765,0.837)	0.797 (0.763, 0.831)
SAM	14	0.749 (0.726, 0.773)	0.753 (0.716, 0.791)	0.744 (0.706, 0.782)
EBM	24	0.858 (0.837, 0.879)	0.812 (0.771, 0.853)	0.793 (0.754, 0.833)
Black box models				
SVM	14	0.989 (0.982, 0.995)	0.767 (0.724, 0.810)	0.732 (0.691, 0.772)
RF	14	0.960 (0.952, 0.968)	0.814 (0.779, 0.849)	0.797 (0.762, 0.832)
GBM	14	0.827 (0.807, 0.846)	0.805 (0.770, 0.841)	0.791 (0.756, 0.825)
MLP	14	0.828 (0.805, 0.850)	0.810 (0.769, 0.852)	0.815 (0.777, 0.853)
Partial response models				
prSVM1	54	0.830 (0.811, 0.850)	0.797 (0.759, 0.834)	0.794 (0.760, 0.828)
prSVM2	31	0.806 (0.785, 0.827)	0.786 (0.747, 0.825)	0.782 (0.746, 0.818)
prRF1	21	0.855 (0.839, 0.871)	0.770 (0.732, 0.808)	0.770 (0.733, 0.806)
prRF2	16	0.841 (0.824, 0.858)	0.761 (0.723, 0.799)	0.767 (0.731, 0.804)
prGBM1	15	0.817 (0.797, 0.837)	0.811 (0.777, 0.845)	0.783 (0.748, 0.818)
prGBM2	7	0.787 (0.766, 0.809)	0.802 (0.768, 0.836)	0.771 (0.734, 0.807)
PRN	12	0.810 (0.786, 0.833)	0.799 (0.756, 0.841)	0.807 (0.769, 0.845)
PRN-Lasso	12	0.811 (0.787, 0.834)	0.797 (0.755, 0.840)	0.812 (0.774, 0.850)

Most models have comparable performance, but they differ significantly in their interpretability. In particular, while LR is restricted to linear dependence on the covariates it performs well for this dataset. This is explained by the partial responses which show that the dependence on the individual variables is close to linear in the main body of the histogram for that variable. However, the responses saturate either side of it and so flatten out, hence the marginally better performance of some of the partial response models.

The results show that SVM is not ideally suited when non-linearities are weak, with overfitting that results in the prSVM sometimes outperforming the original back box on the out of sample data. In contrast, the GBM generates partial responses that generalise well. The RF model has the highest performance degradation when the ANOVA decomposition is truncated. This is likely because the partial responses for the RF are staggered and not smooth due to the internal structure of the black-box model.

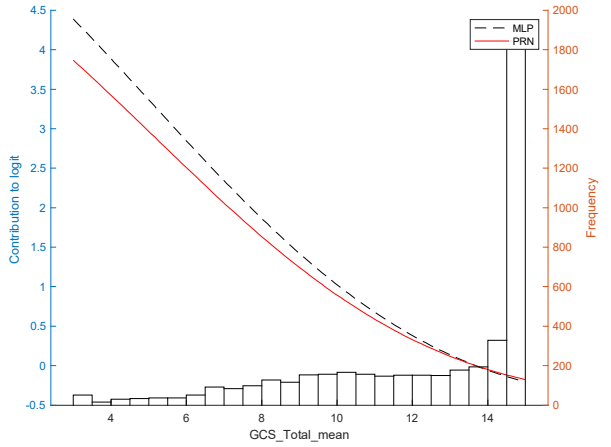


Fig. 2. Example univariate partial responses from the PRN-Lasso model on the MIMIC III data using means only. The Glasgow Coma Scale (GCS) score is the main indicator of consciousness and shows a monotonic decrease in mortality, as expected. The left had scale shows the contribution of this variable to the logit P(C|x), which corresponds directly to the score index $\beta.x$ in logistic regression. The dashed line is the initial partial response after the first iteration of the MLP and the solid line is the result after the second iteration using the GANN/SENN.

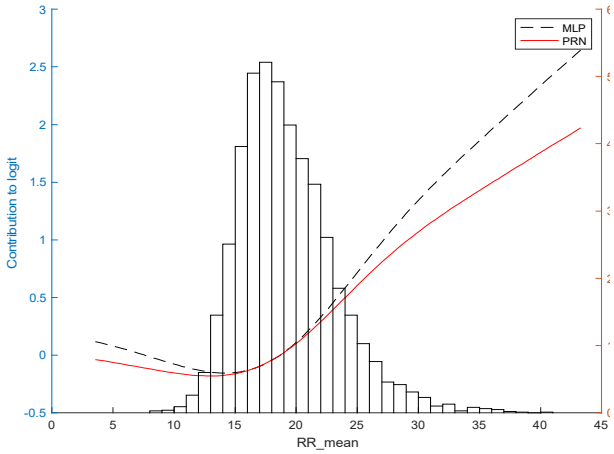


Fig. 3. Another important effect identified in the PRN-Lasso model is the Respiratory Rate (RR). This illustrates the non-linear nature of the partial responses. In this case, mortality increases away from the mean, but more sharply for higher values of RR.

The PRN and PRN-Lasso perform very well for this medical dataset. Figs. 2-4 show three of the component responses that together add to make the logit(P(C|x)) for the PRN-Lasso model. Given a patient profile described by an input vector [GCS, RR, Temperature mean, etc.] the model prediction is calculated as follows:

7. Take the value of each input variable, e.g. GCS, and find its contribution to the logit (P(C|x)). This is the y-axis in fig. 2. For RR it's the y-axis in fig. 3, for Temperature in fig. 4, etc.
8. Include the contributions from all univariate terms and also bivariate terms e.g. z-axis in fig. 5.
9. Add these contributions and also the β_0 from the Lasso. This addition forms the risk index, which is the full logit (P(C|x)).
10. Feed the logit into a sigmoid function, which is indicated by an S in the output node in fig. 1. The result is the predicted posterior probability of class membership, in this case, the probability of mortality, P(C|x).

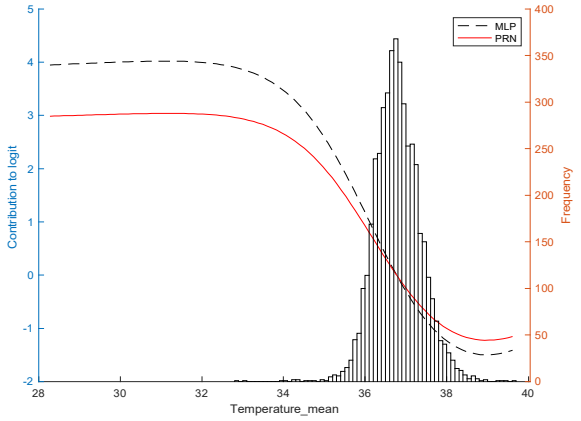
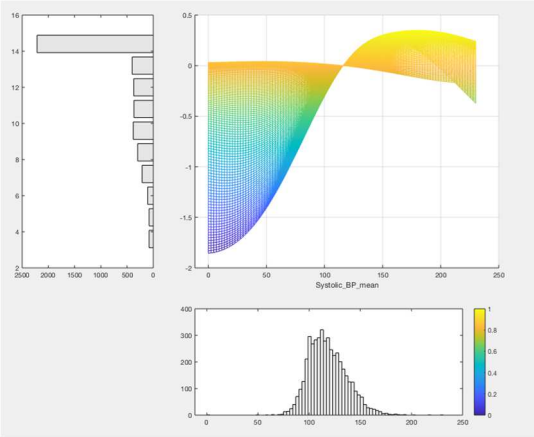


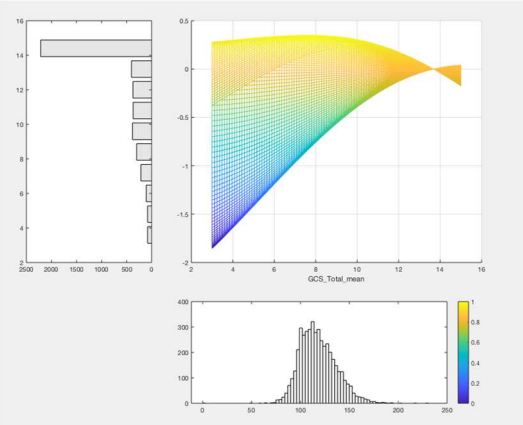
Fig. 4. Mean core temperature also has a statistically significant effect that is quantified in the PRN-Lasso model by the curve shown. Mortality increases for lower temperatures. Note how the curve is approximately linear in the main body of the histogram of temperature values, pointing to why logistic regression does well overall for this dataset.

The predicted probabilities can be compared with the observed occurrence of mortality to produce the calibration curve. This is shown in fig. 6 for the out of sample dataset. Note as well the very good match between the predicted probability of mortality, in the x-axis, and the fraction of predicted cases in the same interval of predicted mortality, shown by the circles.

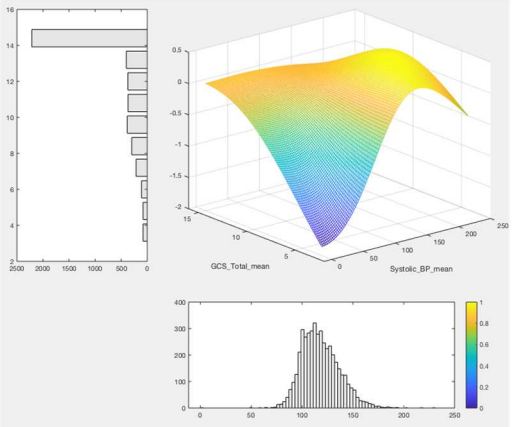
Calibration is vital for clinical applications where the quantitative inference of the output must be numerically accurate. It is very much possible to have poor calibration with excellent classification performance measured by the AUC. This occurs when the predictions are in the right ranking order but their numerical values may be very skewed, for instance, due to class imbalance. Not all classifiers are well-calibrated, but the MLP is, even for extreme imbalances of the order of 1/100, as is the case for instance when predicting event rates for short time intervals in survival modelling.



(a)



(b)



(c)

Fig. 5. Two-way interaction between the GCS score and Systolic Blood Pressure from the PRN-Lasso model. This graphic shows a) & b) views along the main axis to show that the bivariate partial response vanishes along each axis; Note that the axes in the modelled data correspond to the values of the median in the original data, prior to standardisation by median centering and scaling to unit variance. c) a 3D view. This graphic shows that a correction is required to ensure good calibration of the posterior probability for cases where the GCS score and Systolic BP are both low. In common with the other figures of the partial responses, the graphs show histograms of the original variables.

The AUC performance of the proposed approaches is in line with those reported in [16]. Nevertheless, the data structures used and precise experiments are not the same, and we used very simple compression of the time series. The paper makes the comment that "even a model with 0.91 AUC-ROC can make trivial mistakes and there is a lot of room for improvement". We agree and suggest that the interpretability element is helpful to identify the precise weight that each input variable, or pair of input variables, contributes to the prediction, hence finding out what, if anything, misled the model.

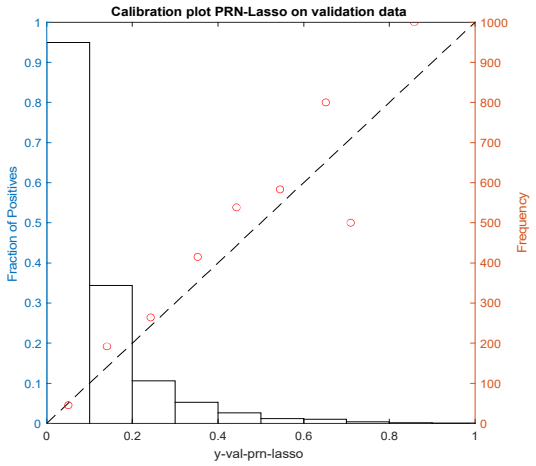


Fig. 6. Calibration of the PRN-Lasso model. This is nearly perfect given the histogram of output predictions, which is heavily skewed to lower values on account of the prevalence of mortality in this dataset being 11.1% for the training data, 10.6% for the validation data and 12.7% for the out of sample dataset. The circles represent the proportion of mortality among the predictions made within the interval of the width of the histogram bar. They are very close to the ideal line for the vast majority of predictions.

The ability to diagnose the model, that is to say, to find out exactly why it was right or not, is important in order to improve it in a controlled manner or, even, to find issues in data collection e.g. artifacts or variables missing from the protocol, or unintended biases in the sampling process. Moreover, this also allows the clinician to integrate the machine learning model, including pre-trained black boxes, into the clinical reasoning process.

Compared with the state of the art, our models perform better than SAM and similarly to EBM. Both of these approaches in principle support univariate and bivariate effects but in SAM the component additive elements are restricted to splines, which can be a limiting factor on performance.

In the case of the EBM, its partial response for the GCS is shown in Fig. 7, alongside the corresponding functions derived from the SVM and GBM algorithms. These plots follow the same trend as Fig. 2, which is the expected decrease in mortality for higher GCS scores due to how it is calculated. Interestingly, while the plot for the prSVM is smooth, it shows a curvature that may be an artifact resulting from the width of the original radial basis functions. A similar effect is present in all of the component functions from this model.

The component function for the GBM is noticeably noisy and for the EBM it is staggered. This may lead to a loss in classification performance compared to a better estimate of the partial response. The example in Fig. 2 is consistent with the smooth interpolation of the curves in Fig. 7 (a) & (c).

TABLE III. COMPONENT FUNCTIONS SELECTED BY THE SPARSE MODELS AND PARTIAL RESPONSE MODELS FOR MIMIC-III DATA WITH INPUTS AS MEANS ONLY.

Model	SAM	EBM	prGBM1	PRN-Lasso
#Components	9	19	10	11
Univariate components				
Diastolic BP mean	✓	✓	✓	✓
Systolic BP mean	✓	✓	✓	
GCS Total mean	✓	✓	✓	✓
Glucose mean	✓	✓	✓	
Heart Rate mean	✓	✓	✓	✓
O2 Saturation mean	✓	✓	✓	✓
Respiratory Rate mean	✓	✓	✓	✓
Temperature mean	✓	✓	✓	✓
Weight	✓	✓	✓	✓
Two-way interactions				
Systolic BP mean X GCS Total mean		✓		✓
GCS Total mean X Heart Rate mean		✓		✓
GCS Total mean X Respiratory Rate mean		✓		✓
GCS Total mean X Temperature mean		✓		✓
O2 Saturation mean X Weight			✓	
Diastolic BP mean X GCS Total mean		✓		
GCS Total mean X Glucose mean		✓		
GCS Total mean X O2 Saturation mean		✓		
GCS Total mean X Weight		✓		
Systolic BP mean X Heart Rate mean		✓		
Heart Rate mean X Temperature mean		✓		

The variables selected by the best performing interpretable models and the benchmark models are listed in Tables III & IV. While we have already noted that SAM supports bivariate effects in principle, we were unable to find any reference to these in the software used. The EBM and prGBM1 models, like the SAM, select all univariate components as well as several bivariate components. Glucose does not appear in any univariate or bivariate term of the PRN-Lasso.

The partial response models are sparser than the EBM, containing a similar number of terms to the SAM, while also including bivariate terms. The bivariate components selected by the PRN-Lasso suggest that they are corrections to the calibration of the GCS Total Mean. In contrast, the EBM utilises more than double the number of bivariate terms.

The additional step for the MLP of mapping the Lasso model onto a SENN and continuing training to result in the PRN model, led to only a small improvement in performance. This is apparent also from the small changes observed in the shape of the partial responses. This indicates that for real-world data sets such as MIMIC the noise present in the data limits performance to the extent that the significant predictive factors are well represented by just the univariate and bivariate terms in the ANOVA decomposition.

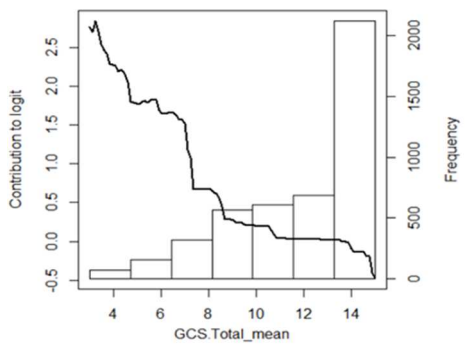
TABLE IV. COMPONENT FUNCTIONS SELECTED BY THE SPARSE MODELS AND PARTIAL RESPONSE MODELS FOR MIMIC-III DATA WITH MEANS AND STANDARD DEVIATIONS.

Model	SAM	EBM	prGBM1	PRN-Lasso
#Components	14	24	15	12
Univariate components				
Diastolic BP mean	✓	✓	✓	
Diastolic BP st dev	✓	✓	✓	
Systolic BP mean	✓	✓	✓	✓
Systolic BP st dev	✓	✓		
GCS Total mean	✓	✓	✓	✓
GCS Total st dev	✓	✓	✓	
Glucose mean	✓	✓	✓	
Glucose st dev	✓	✓	✓	
Heart Rate mean	✓	✓	✓	✓
O2 Saturation mean	✓	✓		✓
O2 Saturation st dev	✓	✓	✓	✓
Respiratory Rate mean	✓	✓	✓	✓
Temperature mean	✓	✓	✓	✓
Weight	✓	✓	✓	✓
Two-way interactions				
Systolic BP mean X GCS Total mean		✓		✓
Systolic BP st dev X GCS Total mean		✓		✓
GCS Total mean X GCS Total st dev		✓		✓
GCS Total mean X Temperature mean		✓		✓
Systolic BP mean X Heart Rate mean			✓	
Systolic BP mean X O2 Saturation st dev			✓	
GCS Total mean X Weight			✓	
GCS Total mean X Glucose st dev		✓		
Diastolic BP st dev X GCS Total mean		✓		
GCS Total mean X Heart Rate mean		✓		
GCS Total mean X O2 Saturation st dev		✓		
Diastolic BP mean X GCS Total mean		✓		
GCS Total mean X Respiratory Rate mean		✓		

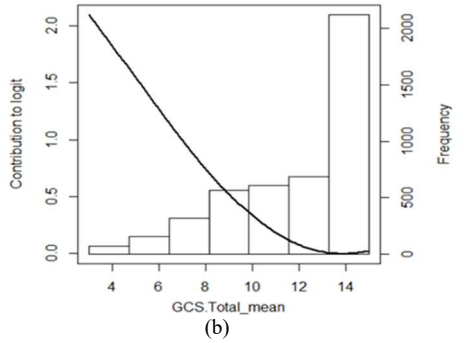
IV. CONCLUSION

We show that it is possible to open any black-box model, including pre-trained models, in cases where significant noise is present, without losing much predictive power, if any, but making the model transparent to the non-linearities in the data.

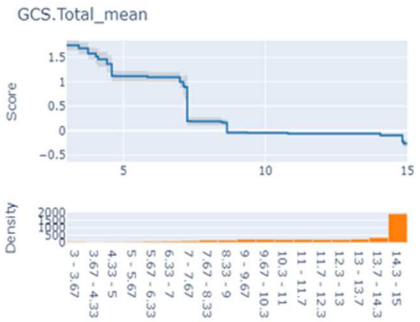
This involves the application of the ANOVA decomposition, anchored on the median of the data followed by the Lasso as a computationally efficient method to derive the structure of a GAM using the component functions derived from the ANOVA. The application of the LASSO to the univariate and bivariate terms in the ANOVA decomposition carries out the functions of model selection and re-calibration, resulting in a globally interpretable model with comparable performance to the original black-box model. In this way, we buck the accuracy/interpretability trade-off for tabular data.



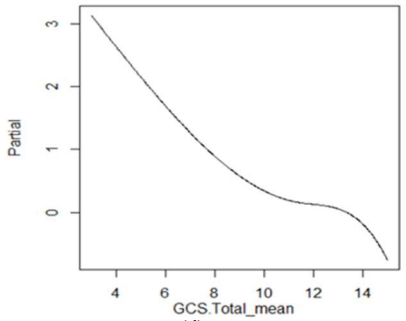
(a)



(b)



(c)



(d)

Fig. 7. Example univariate responses for GCS score from the a) prGBM, b) prSVM, c) EBM and d) SAM models. Similarly to Fig. 2, these plots show a decrease in mortality as the GCS score increases.

Furthermore, the performances of the resulting GAMs compare favourably with state-of-the-art sparse models from the statistical literature, SAM, and from machine learning, the EBM. The derived Partial Responses are consistent across the range of models and have plausible clinical interpretations.

The interpretability of the model by end-users is at the level of nomograms [7]. Nomograms are familiar to clinicians as graphical implementations of logistic regression. GAMs are interpretable in the same way, except that the score for each variable is read from what we call the partial response plot, where the height of the plot directly measures the contribution to the logit, which is the nomogram score for that variable.

REFERENCES

[1] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.

[2] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B.Van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clinical Epidemiology, 10 (2019):12-22.

[3] T. Rögnvaldsson, T.A. Etchells, L. You, D. Garwicz, I. Jarman, P.J.G. Lisboa, How to find simple and accurate rules for viral protease cleavage specificities, BMC Bioinformatics. 10 (2009) 149..

[4] P. J. G. Lisboa, Industrial use of safety-related artificial neural networks, HSE – Health & Safety Executive 327 (2001) 1–36.

[5] A. Backhaus, U. Seiffert, Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size, Neurocomputing 131 (2014) 15–22. doi:10.1016/j.neucom.2013.09.048

[6] G. Hooker, Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables, J. Comput. Graph. Stat. 16 (2007) 709–732.

[7] V. Van Belle, B. Van Calster, S. Van Huffel, J.A.K. Suykens, P. J.G. Lisboa, Explaining Support Vector Machines: A Color Based Nomogram, PLoS One. 11 (2016) e0164568.

[8] W.S. Sarle, Neural Networks and Statistical Models, SAS USers Gr. Int. Conf. (1994).

[9] D.A. de Waal, J. V. du Toit, Automation of Generalized Additive Neural Networks for Predictive Data Mining, Appl. Artif. Intell. 25 (2011) 380–425.

[10] C. Brás-Geraldes, A. Papoila, P. Xufre, Odds ratio function estimation using a generalized additive neural network, Neural Comput. Appl. (2019).

[11] D. Alvarez-Melis, T.S. Jaakkola, Towards Robust Interpretability with Self-Explaining Neural Networks, in: 32nd Conf. Neural Inf. Process. Syst. (NeurIPS 2018), Montréal, Canada, 2018.

[12] Y. Lou, R. Caruana, and J. Gehrke, Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), ACM,:150–158 (2012).

[13] P. Ravikumar, J. Lafferty, H. Liu, L. Wasserman, Sparse additive models, J. R. Stat. Soc. Ser. B (Statistical Methodol. 71 (2009) 1009–1030.

[14] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. R. Stat. Soc. B. 58 (1996) 267–288.

[15] A. E. W. Johnson et al., MIMIC-III, a freely accessible critical care database, Sci. Data, vol. 3, no. 1, 1–9 (2016).

[16] 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, 1–18 (2019).

[17] G. Teasdale and B. Jennett, "ASSESSMENT OF COMA AND IMPAIRED CONSCIOUSNESS. A Practical Scale," Lancet, vol. 304, no. 7872, 81–84 (1974).

[18] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. (2001).

[19] V. Vapnik, Statistical learning theory, 1998.

[20] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–3