Explaining the Neural Network: A Case Study to Model the Incidence of Cervical Cancer

Paulo J. G. Lisboa(✉), Sandra Ortega-Martorell, and Ivan Olier

Department of Applied Mathematics, Liverpool John Moores University,
Liverpool L3 3AF, UK
p.j.lisboa@ljmu.ac.uk

Abstract. Neural networks are frequently applied to medical data. We describe how complex and imbalanced data can be modelled with simple but accurate neural networks that are transparent to the user. In the case of a data set on cervical cancer with 753 observations excluding, missing values, and 32 covariates, with a prevalence of 73 cases (9.69%), we explain how model selection can be applied to the Multi-Layer Perceptron (MLP) by deriving a representation using a General Additive Neural Network.

The model achieves an AUROC of 0.621 CI [0.519,0.721] for predicting positive diagnosis with Schiller's test. This is comparable with the performance obtained by a deep learning network with an AUROC of 0.667 [1]. Instead of using all covariates, the Partial Response Network (PRN) involves just 2 variables, namely the number of years on Hormonal Contraceptives and the number of years using IUD, in a fully explained model. This is consistent with an additive non-linear statistical approach, the Sparse Additive Model [2] which estimates non-linear components in a logistic regression classifier using the backfitting algorithm applied to an ANOVA functional expansion.

This paper shows how the PRN, applied to a challenging classification task, can provide insights into the influential variables, in this case correlated with incidence of cervical cancer, so reducing the number of unnecessary variables to be collected for screening. It does so by exploiting the efficiency of sparse statistical models to select features from an ANOVA decomposition of the MLP, in the process deriving a fully interpretable model.

Keywords: Explainable machine learning · FATE · KDD · Medical decision support · Cervical cancer

1 Introduction

This paper is about explainable neural networks, illustrated by an application of a challenging data set on cervical cancer screening that is available in the UCI repository [3]. The purpose of the paper is to describe a case study of the interpretation of a neural network by exploiting the same ANOVA decomposition that has been used in statistics to infer sparse non-linear functions for probabilistic classifiers [2].

We will show how a shallow network, the Multi-Layer Perceptron (MLP) can be fully explained by formulating it as a General Additive Neural Network (GANN). This methodology has a long history [4]. However, to our knowledge there is no method to

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1237, pp. 585–598, 2020.
https://doi.org/10.1007/978-3-030-50146-4_43

derive the GANN from data, rather a model structure needs to be assumed or hypothesized from experimental data analysis. In this paper we use a mechanistic model to construct the GANN and show that, for tabular data i.e. high-level features that are typical of applications to medical decision support, a transparent and parsimonious model can be obtained, whose predictive performance comparable i.e. well within the confidence interval for the AUROC, with that obtained an alternative, opaque, deep learning neural network applied to the same data set [1].

Fairness, Accountability Transparency and Ethics (FATE) in AI [5] is emerging as a priority research area that relates to the importance of human-centered as a key enabler for practical application in risk-related domains such as clinical practice. Blind spots and bias in models e.g. due to artifacts and spurious correlations hidden in observational data, can undermine the generality of data driven models when they are used to predict for real-world data and this may have legal implications [6].

There different approaches that may be taken to interpret neural networks, in particular. These include derivation of rules to unravel the inner structure of deep learning neural networks [7] and saliency methods [8] to determine the image elements to which the network prediction is most sensitive.

An additional aspect of data modelling that is currently very much understudied is the assessment of the quality of the data. Generative Adversarial Networks have been used to quantify sample quality [9].

Arguably the most generic approach machine explanation is the attribution of feature influence with additive models. A unified framework for this class of models has been articulated [10]. This includes as a special case the approach of Local Interpretable Model Agnostic Explanations (LIME) [11].

However, it is acknowledged in [10] that General Additive Models (GAMs) are the most interpretable because the model is itself the interpretation, and this applies to data at a global level, not just locally.

Recently there has been a resurgence of interest in GAMs [11, 12] in particular through implementations as GANNs. These models sit firmly at the interface between computational intelligence and traditional statistics, since they permit rigorous computation of relevant statistical measures such as odds ratios for the influence of specific effects [12].

A previously proposed framework for the construction of GANNs from MLPs will be applied to carry out model selection and so derive the form of the GANN from a trained MLP. This takes the form of a Partial Response Network (PRN) whose classification performance on multiple benchmarking data sets matches that of deep learning but with much sparser and directly interpretable features [13].

This paper reports a specific case study of the application of PRN to demonstrate how it can interpret the MLP as a GAM, providing complete transparency about the use of the data by the model, without compromising model accuracy as represented by the confidence interval of the AUROC. Our results are compared with those from a state-of-the-art feature selection method for non-linear classification [2].

Moreover, the model selection process itself will generate insights about the structure of the data, illustrating the value of this approach for knowledge discovery in databases (KDD).

2 Data Description

2.1 Data Collection

Cervical cancer is a significant cause of mortality among women both in developed and developing countries world-wide [1]. It is unusual among cancers for being closely associated with contracting the Human Papillomavirus (HPV) [14] which is strongly influenced by sexual activity. This makes cervical cancer one of the most avoidable cancers, through lifestyle factors and by vaccination.

Screening for possible incidence of the cancer is a public health priority, with potential for low-cost screening to be effective. The data set used in this study was acquired for this purpose.

The data were collected from women who attended the Hospital Universitario de Caracas in Caracas [3]. Most of the patients belong to the lowest socioeconomic status, which comprises the population at highest risk. They are all sexually active. Clinical screening includes cytology, a colposcopic assessment with acetic acid and the Schiller test (Lugol's iodine solution). This is the most prevalent diagnostic index and is the choice for the present study.

2.2 Data Pre-processing

The data comprise records from a random sample of patients presenting between 2012 and 2013 (n = 858) [1, 3]. There is a wide age range and a broad set of indicators of sexual activity, several of which overlap in what they measure. Four target variables are reported, including the binary outcome of Schiller's test.

This data set is challenging, first because of severe class imbalance, which is typical in many medical diagnostic applications. The number of positive outcomes in the initial data sample is just 74 cases for Schiller's test, 44 for a standard cytology test and 35 for Hinselmann's test.

Secondly, the data include a range of self-reported behavioural characteristics, where noise levels may be significant. Third, some of the variables were problematic for data analysis. The report of STD: cervical condylomatosis comprises all zero values. STD: vaginal condylomatosis, pelvic inflammatory disease, genital herpes, molluscum contagiosum, AIDS, HIV, Hepatitis B, syphilis and HPV are all populated in <2.5% of all cases. For this reason, these variables were removed from the study as they are unlikely to provide statistical significance in predictive modelling and their low prevalence can cause numerical instabilities for model optimisation.

The number of pregnancies was deemed to be less informative about sexual behaviour than the number of sexual partners, so this was also excluded.

In total 105 rows of data had 20 or more of the 32 covariate values missing. While these values can be imputed, such a large proportion of covariates for individual observations can bias the study, since missingness can be informative. For this reason, these rows were removed from the data.

Among the selected variables, several pairs of covariates measure the same indicator in binary form and as an ordinal count. This applies to variables Smokes, Hormonal Contraceptives, IUD and STDs. Consequently, the initial pool of covariates in this study comprises 9 variables. They are:

- Number of sexual partners;
- Age of first sexual intercourse;
- Years since first sexual intercourse, derived by subtracting the previous covariate from Age;
- Number of years smoking;
- Number of years taking Hormonal Contraceptives
- Number of years using IUDs;
- STD: condylomatosis;
- Number of STDs;
- Number of diagnosed STDs.

The dataset used in this study is a reduced cohort (n = 753) with marginal values summarized in Table 1. The prevalence of missing data in the study sample is now much reduced, especially as the number of pregnancies is not used. The maximum proportion of missing is 4.1% for IUD (years).

Table 1. Summary statistics of the sample population for Cervical Cancer screening. {} indicates a binary variable. [] shows the range of the variable.

Variable	Median [Min, Max]	Missing values
Age	26 [13, 84]	0
Number of sexual partners	2 [1, 28]	14
First sexual intercourse	17 [10, 32]	6
Number of pregnancies	2 [0, 11]	47
Smokes	0 {0, 1}	10
Smokes (years)	0 [0, 37]	10
Smokes (packs/year)	0 [0, 37]	10
Hormonal Contraceptives	1 {0, 1}	13
Hormonal Contraceptives (years)	0.5 [0, 30]	13
IUD	0 {0, 1}	16
IUD (years)	0 [0, 19]	31
STDs	0 {0, 1}	0
STDs (number)	0 [0, 4]	0
STDs: condylomatosis	0 {0, 1}	0
STDs: Number of diagnosis	0 [0, 3]	0

Missing values were imputed with the sample median. The reason for this is that the standardisation used in the following section maps the median value of every covariate to zero, which has the effect of discarding that instance from the gradient descent weight updates, so minimising the impact of unknown information in the training of the MLP.

3 Partial Response Network Methodology

In binary classification, GAMs model the statistical link function appropriate for a Bernoulli error distribution. This is the logit, hence the inverse of the familiar sigmoid function. An appropriate objective function is the equally familiar log-likelihood cost.

In order to control for overfitting of the original MLP, we apply regularisation using Automatic Relevance Determination [15]. This model evaluates the strength of weight decay using a Bayesian estimator, which enables a different weight decay parameter to be used for the fan-out weights linked to each input node. This results in soft model selection, that is to say a modulation of the weight values that compresses towards zero the weights linked to the less informative input variables.

Input variables are divided by the standard deviation and shifted by the median value, so that the median is represented by zero. This is important because in a Taylor expansion of the logit function about the median values, setting an individual variable to the median causes all of the terms involving that variable in the Taylor expansion to vanish. It is then possible to capture much of the most significant terms by systematically setting all bar one covariate to zero, then all but each pair of covariates to zero, and so on.

The MLP response when all but a few variables are zero is called the Partial Response and the GANN obtained by mapping the partial responses onto its weights, forms the Partial Response Network (PRN) [13].

The functional form of the PRN is given by the well-known statistical decomposition of multivariate effects into components with fewer variables, represented by the ANOVA functional model [2] shown in Eq. (1):

$$logit(P(C|x)) \approx \varphi(0) + \sum\nolimits_i \varphi_i(x_i) + \sum\nolimits_{i \neq j} \varphi_{ij}\left(x_i, x_j\right) + O\left(x_i, x_j, x_i\right) \tag{1}$$

where the partial responses $\varphi_k(\bullet)$ are evaluated with all variables held fixed at zero except for one or two indexed as follows:

$$\varphi(0) = logit(P(C|0)) \tag{2}$$

$$\varphi_i(x_i) = logit(P(C|(0, .., x_i, .., 0))) - \varphi(0) \tag{3}$$

$$\varphi_{ij}\left(x_i, x_j\right) = logit\left(P\left(C|\left(0, .., x_i, .., x_j, ..0\right)\right)\right) - \varphi_i(x_i) - \varphi_j\left(x_j\right) - \varphi(0) \tag{4}$$

The derivation of the PRN proceeds as follows:

1. Train an MLP for binary classification;
2. Obtained the univariate and bivariate partial responses in Eqs. (2)–(4).
3. Apply the Lasso to the partial responses;
4. Construct a second MLP as a linear combination of the partial responses so as 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.

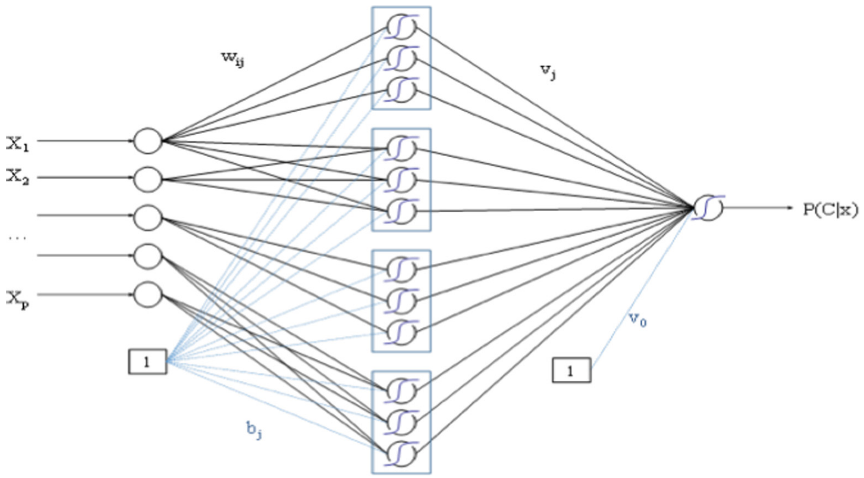

Fig. 1. Representation of the Partial Response Network as General Additive Neural Network (GANN). The weight values are derived from a trained MLP and re-calibrated by further training of the network as a GANN.

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

1. Univariate partial responses

$$v_j \rightarrow v_j * (\beta_k - \beta_{kl}) \tag{5}$$

$$v_0 \rightarrow (v_0 - logit(P(C|0))) * (\beta_k - \beta_{kl}) \tag{6}$$

2. Bivariate partial response

$$v_0 \rightarrow (v_0 - logit(P(C|0))) * (\beta_k - \beta_{kl}) \tag{7}$$

$$v_0 \rightarrow (v_0 - logit(P(C|0))) * (\beta_k - \beta_{kl}) \tag{8}$$

The main limitation of the model as currently used is that it is restricted to univariate effects and bivariate interactions. However, in many medical applications, this is likely to suffice. The method can be extended to higher order interactions but it will generate a combinatorially large number of partial responses.

4 Experimental Results

This section explains how model selection took place and describes the models obtained with the PRN applied to the Cervical Cancer screening data set described in Sect. 2. The variables used in the model are the subset of Table 1 that is listed in 2.2 and the target variable is the outcome of Schiller's diagnostic test for cervical cancer.

Given the low prevalence of positive outcome, 73 out of the 753 cases retained (prevalence = 9.69%) the results presented are all for out-of-sample data using 2-fold cross validation. This choice of number of folds is motivated by the need to retain a meaningful number of events in each fold.

Model selection consisted of an iterative process of removing the least frequently occurring variable or set of variables at each stage in the process. Table 2 shows the frequency of occurrence of each covariate in the partial responses selected by the PRN. It also shows the average AUROC for 10 random starts.

Table 2. Model selection with the PRN applied to the Cervical Cancer screening dataset. φ_i;φ_{ij}: variable present in a univariate/bivariate partial response.

# var	AUC	#Sex partners	Age first sexual Inter	Smokes (Yrs)	Hormonal Contraceptives (Yrs)	IUD (Yr)	# STD	STD: condylomatosis
p = 9	**0.585**							
φ_i		*1*	*1*	*1*	*3*	*1*	*1*	*3*
φ_{ij}		*1*	*6*	*13*	*12*	*7*	*7*	
p = 5	**0.621**							
φ_i		–	–	–	*4*	*4*	*3*	*2*
φ_{ij}		–	–	*16*	*10*	*8*	*5*	*9*
p = 4	**0.593**							
φ_i		–	–	–	*4*	*5*	–	–
φ_{ij}		–	–	*5*	*6*	*5*	*4*	–
p = 3	**0.635**							
φ_i		–	–	–	*9*	*10*	*5*	–
φ_{ij}		–	–	–	*9*	*9*	*2*	–
p = 2	**0.621**							
φ_i		–	–	–	*9*	*10*	–	–
φ_{ij}		–	–	–	*8*	*8*	–	–

The results in Table 1 can be compared with those from a sparse non-linear statistical classifier, the Sparse Additive Model (SAM). This is an additive non-linear model that estimates component functions in an ANOVA decomposition using the backfitting algorithm that is standard for GAMs. It combines that with l_1 regularisation similar to the Lasso [2]. This provides the attractive property of convex optimisation, so that the model only needs to be estimated once.

In contrast, neural network models are not convex and so require multiple estimation. By interpreting the MLP in the form of a GAM with sparse features, the PRN model considerably reduces the variability in classification performance that is typical of the MLP, providing more consistent results.

However, correlations between variables can result in multiple models with very similar predictive power. This is the case for the present data set.

The SAM identified {#Years sexual intercourse; Smokes (years); STDs} for fold 1 and {Hormonal Contraceptives (years); IUD (years); STDs: condylomatosis; STDs} for fold 2 as univariate models; {STDs} for fold 1 and {IUD (years); STDs: condylomatosis; STDs; Number of sexual partners*IUD (years); #Years sexual intercourse*Hormonal Contraceptives (years); STDs: condylomatosis*STDs} when interaction terms were included.

The AUROCs for SAM in 2-fold cross validation are 0.599 and 0.565, respectively.

5 Discussion

The variable subsets extracted with model selection using the PRN model are all consistent with the previously cited work on this data set, and indeed with cervical screening literature.

The iterative process for feature selection applied in the previous section made use of the variability of the MLP under random starts to explore the space of predictive features in the presence of correlated variables. This enable the identification of stable features that could be applied for both folds to build a model with a consistent explanation. These two features are Hormonal Contraceptives (years) and IUD (years).

It cannot be claimed that these are the only predictive variables or indeed the best. However, they are a representative subset that achieves a high predictive model with parsimony, as can be seen from both the size of the derived feature set and high AUROC compared with the SAM.

Equally of interest is the shape of the partial responses and their stability under 2-fold cross validation, shown in Figs. 2, 3, 4 and 5.

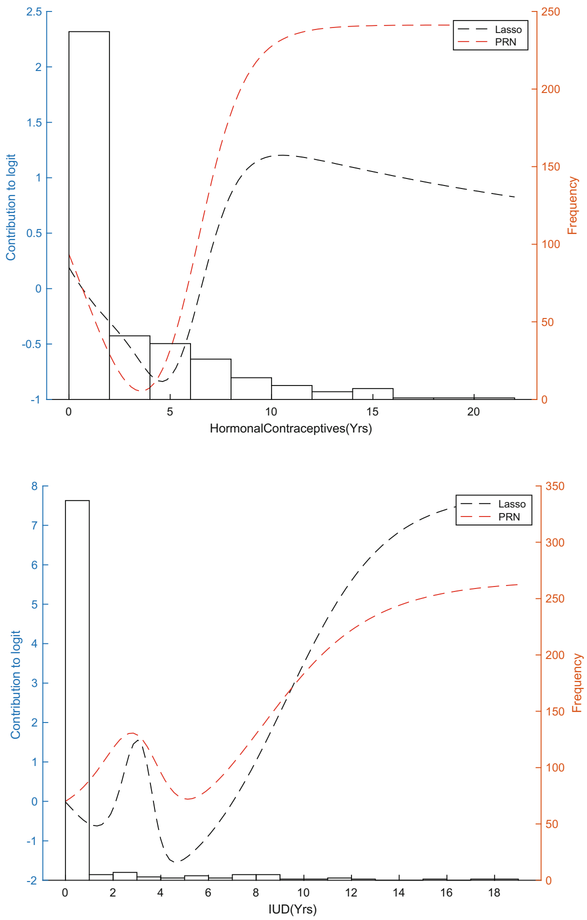

Fig. 2. Two univariate responses identified in the first fold. The abscissa measures the contribution of the individual covariate to the logit response. The histogram represents the empirical distribution of the covariate across the study population. The curves show the response derived from the initial MLP and after re-training with the PRN.

The partial responses are remarkably consistent given the challenges posed by the low prevalence and high noise in the data. Differences are apparent in areas of low data density, which is to be expected. Further work will involve quantifying the uncertainty about these estimates.

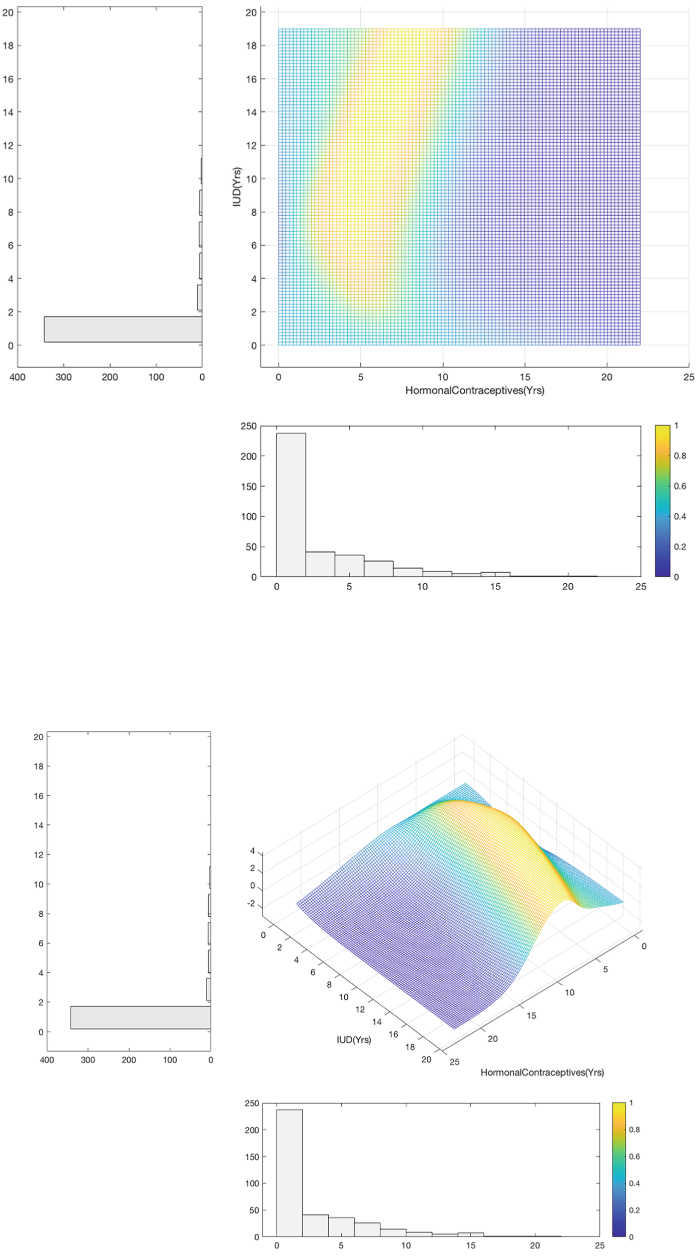

Fig. 3. Bivariate response found to be significant in the first fold of the data. The response is shown as a heat map and as a 3-d surface.

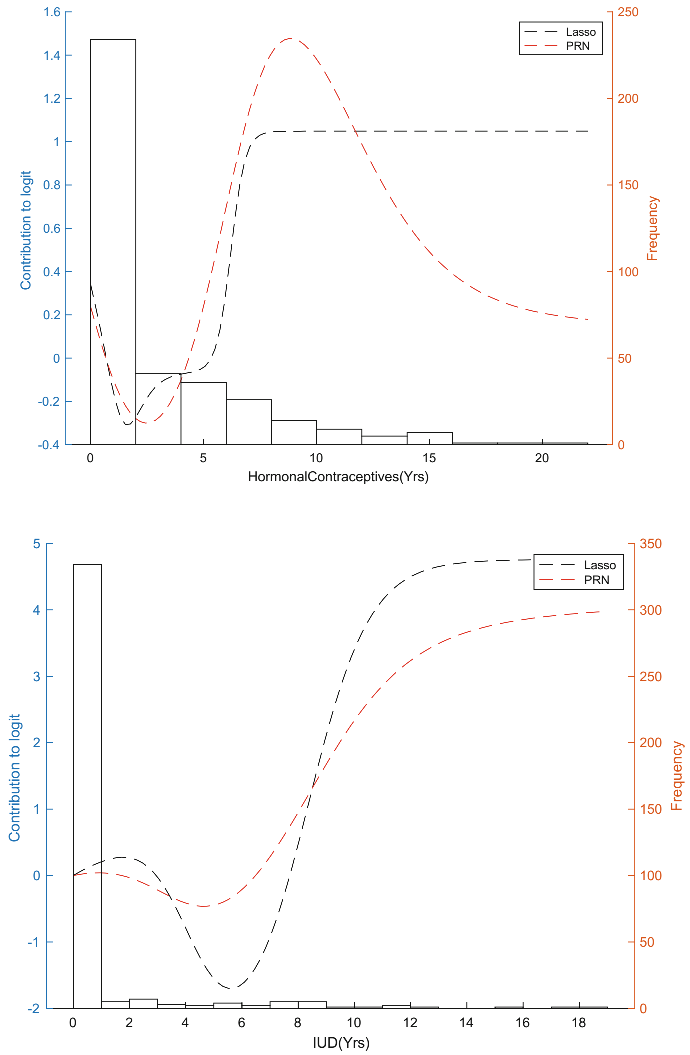

Fig. 4. Two univariate responses identified in the second fold, as in Fig. 2.

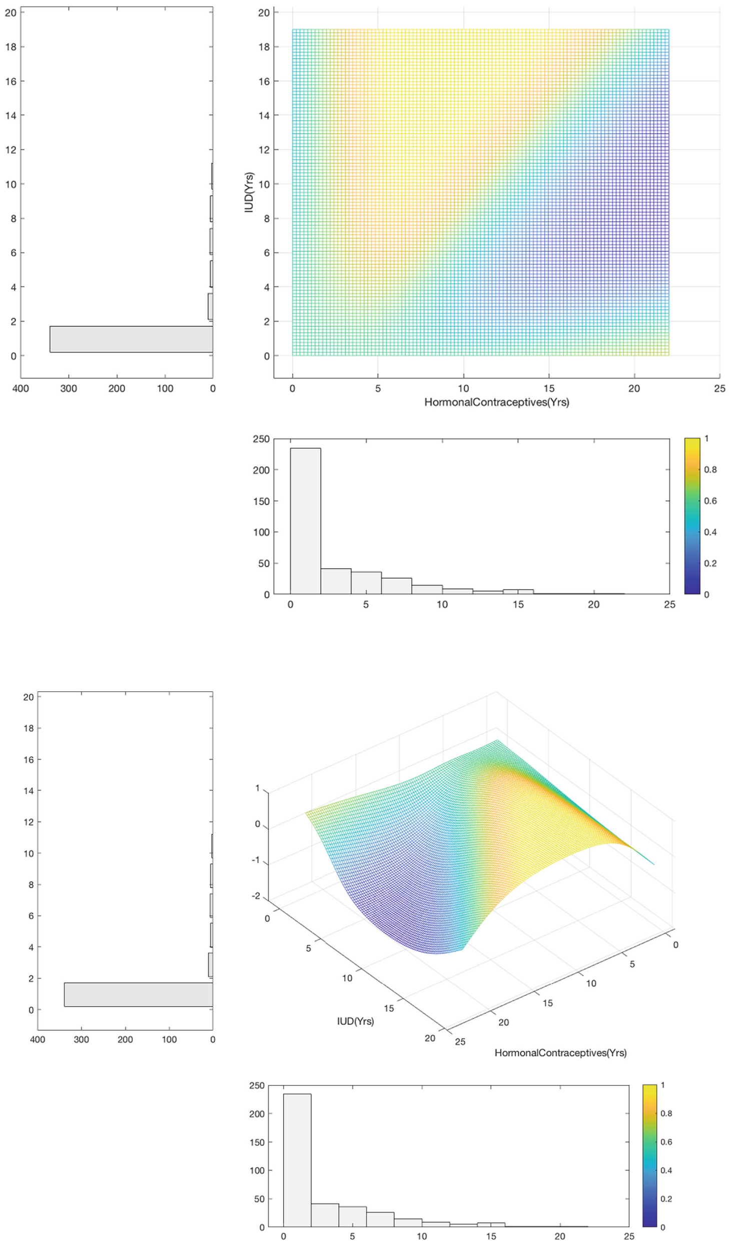

Fig. 5. Bivariate response found to be significant in the second fold of the data, as in Fig. 3.

6 Conclusion

The initial pool of 9 variables contains redundant information. This causes instability in neural network models, as several different models will capture information with similar predictive value. However, an iterative approach to feature selection can produce a stable sparse model.

It is perhaps remarkable how the same predictive information is contained in a small number of covariates compared with the size of the original pool. Bearing in mind that the typical standard deviation of the AUROC is 0.05, making the 95% confidence interval 0.10, the AUROC values for all models listed in Table 2 are comparable. Indeed, the average performance for ten random starts equals that of the best cross-validated model, 0.621 CI [0.519,0.721]. The overall performance figure is also consistent with the deep learning models in [1] and with a statistical approach to non-linear classification with an ANOVA decomposition, the SAM [2].

The main conclusion of this paper is that it is possible to break the black box that is the standard MLP, using it to derive a more interpretable structure as a GANN. Using partial responses is a common way to interpret non-linear statistical models. Here, it is shown that the responses can themselves be used directly in modelling, with little or no compromise in predictive performance.

The result is a small model that explains a large and complex data set in terms of variable dependencies that clinicians can understand and integrate into their reasoning models. Iterative modeling is necessary because of the inherent redundancy in the data set, but the sequence of models obtained is itself informative about the association with outcome for individual and pairs of covariates.

Ultimately, the PRN model shows that it is possible to be sure that the model is right for the right reasons. Moreover, the covariate dependencies provide the ability to diagnose flaws in the data, whether because of sampling bias or artifacts in observational cohorts.

It is concluded that the PRN approach can add significant insight and modelling value to the analysis of tabular data in general, and in particular medical data.

References

1. Fernandes, K., Chicco, D., Cardoso, J.S., Fernandes, J.: Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. PeerJ Comput. Sci. **4**, e154 (2018). https://doi.org/10.7717/peerj-cs.154
2. Ravikumar, P., Lafferty, J., Liu, H., Wasserman, L.: Sparse additive models. J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) **71**(5), 1009–1030 (2009)
3. Fernandes, K., Cardoso, J.S., Fernandes, J.: Transfer learning with partial observability applied to cervical cancer screening. In: Alexandre, L.A., Salvador Sánchez, J., Rodrigues, J. M.F. (eds.) IbPRIA 2017. LNCS, vol. 10255, pp. 243–250. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58838-4_27. Accessed 8 Feb 2020
4. de Waal, D.A., du Toit, J.: Generalized additive models from a neural network perspective. In: Proceedings of the 7th IEEE International Conference on Data Mining, ICDM 2007, Omaha, Nebraska, pp. 265–270. IEEE (2007)

5. Holsting, K., Vaughan, J.W., Daumé III, H., Dudík, M., Wallach, H.: Improving fairness in machine learning systems: what do industry practitioners need? ACM CHI Conference on Human Factors in Computing Systems (2019)
6. Barocas, S., Selbst, A.: 10 Big data's disparate impact. Calif. Law Rev. (2016). http://dx.doi.org/10.15779/Z38BG31
7. Gomez, S., Despraz, J., Pena-Reyes, C.A.: Improving neural network interpretability via rule extraction. In: ICANN 2018. LNCS, vol. 11139, pp. 811–813. Springer, Heidelberg (2018)
8. Borji, A.: Saliency Prediction in the Deep Learning Era: Successes, Limitations, and Future Challenges. IEEE Trans PAMI (2019). arXiv:1810.03716v3
9. Zhou, Z., et al.: Activation Maximization Generative Adversarial Nets arXiv:1703.02000 (2017)
10. Lundberg, S., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774 (2017)
11. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2016, pp. 1135–1144 (2016)
12. Brás-Geraldes, C., Papoila, A., Xufre, P.: Odds ratio function estimation using a generalized additive neural network. Neural Comput. Appl. (8) (2020, in press). https://www.springerprofessional.de/en/neural-computing-and-applications-8-2020/17871790
13. Lisboa, P.J.G., Ortega-Martorell, S., Cashman, S., Olier, I.: The Partial Response Network. arXiv:1908.05978 (2019)
14. Burd, E.M.: Human papillomavirus and cervical cancer. Clin. Microbiol. Rev. **16**(1), 1–17 (2003)
15. MacKay, D.J.C.: The evidence framework applied to classification networks. Neural Comput. **4**, 720–736 (1992)