ISSN: 2366-1313

Interpretable Machine Learning in Healthcare through Generalized Additive Model with Pairwise Interactions (GA2M): Predicting Severe Retinopathy of Prematurity

Dr Sikkakolli Gopikrishna, Bollineni Phanibindu

[1]Associate Professor, Sri Mittapalli College of Engineering, Tummalapalem, Guntur, 522233, India.

[2]UG Sri Mittapalli College of Engineering, Tummalapalem, Guntur, 522233, India.

Abstract:
We have investigated the risk factors that lead to severe retinopathy of prematurity using statistical analysis and logistic regression as a form of generalized additive model (GAM) with pairwise interaction terms (GA2M). In this process, we discuss the trade-off between accuracy and interpretability of these machine learning techniques on clinical data. We also confirm the intuition of expert neonatologists on a few risk factors, such as gender, that were previously deemed as clinically not significant in RoP prediction.

Keywords: interpretability of machine learning in healthcare, generalized additive model, logistic regression, GAM, GA2M, Retinopathy of Prematurity (RoP), neonatology.

I INTRODUCTION

Recent trends in research and medical applications of machine-learning concentrate on the ability to interpret the models. In some applications in healthcare, the it is the accuracy of the models that may be more important than its ability to be interpreted however there are plenty of instances that require interpretability even if accuracy is reduced. A scenario in which the model has excellent accuracy and high interpreability can be achieved by beginning with a basic model like GAM, which is a generalized model (GAM) and after that making it more complex (thus more precise) using GA2M (GAM with pairwise interactions) or starting with a more complex model like XGBoost and trying at interpreting it local using methods such as LIME. In NeurIPS (Neural Information Processing Systems) 2017 one of the top conferences in machines learning, a discussion session was organized by Yann Lecun (head of AI research at Facebook) and Rich Caruana (lead machine learning researcher at Microsoft on applications in healthcare). The session that took place at the conclusion of "Interpretable ML Symposium" highlighted the need for more study of the accuracy as compared to. interpretability in machine learning applications for healthcare.

In this article we will look at an investigation into Severe Retinopathy Prematurity (Severe RoP) that can cause complete or partial blindness in infants if treated. Stevie Wonders is a well-known musician suffering from severe RoP. In the 1950's, doctors recently announced the possibility of oxygen being utilized to protect premature infants. But, they needed longer to understand the fact that an excessive dose of oxygen could cause vessels within the eye to expand in a different way. Infants born prematurely with this condition and excessive oxygen therapy will suffer from permanent blindness. World Health Organization (WHO) estimates that, from the 130 million newborns born each year, about 15 million were born prematurely before 37 weeks of gestation have been completed. About 1 million infants suffer each year from complications associated with preterm birth. A lot of survivors suffer the possibility of a lifelong disability, including hearing and visual issues as well as learning difficulties. Of these, retinopathy is one of the most common. Of premature birth (RoP) is a major and significant reason for disability. Retinopathy of Prematurity (RoP) was first identified by Terry 11 in 1942, as a form of developmental proliferative, vascular, and retinal disorder that occurs in premature retinas of newborns who are not fully vascularized.

Alongside cortical blindness, RoP is one of the most frequent causes of blindness in children around the world. The severity of RoP is increased when the birth weight is lower and shorter gestational weeks, and a variety of risk factors have been identified as contributing to the growth of RoP. If the severity of RoP is not treated the condition can lead to bliness and retinal detachment. This is why it is crucial to identify the problem early and treat the condition appropriately to avoid the development of the condition. International Classification for Retinopathy of Prematurity (ICROP) is used to categorize the progress and severity of the condition. The classification begins with the 1st level being the smallest diagnosis of RoP. It then concludes with 5th level retinal detachment as the most severe result of RoP. The clinical information for our study was gathered through the Newborn Clinic of Zeynep Kamil Woman and Child Diseases Hospital in Istanbul between 2011 and 2014. Each year in Turkey about 150 000 newborns are born with birthweights less than 1500g. They have a higher chance of being found to have severe ROP. We will examine the risk factors believed to trigger or be associated with the presence of severe RoP. As we develop our model of prediction we will be able to observe the degree to which the model is improved when we mix categorical and numerical data and include additional interaction terms.

II. LITERATURE SURVEY

Recent trends in research in the field of machine learning in healthcare are focused on the interpretationability of the models. In some applications in healthcare, the precision of the machine may be more important than its ability to be interpreted however there are numerous instances that require interpretability even if accuracy is reduced. The ideal situation where the model has high precision and high interpretability can be achieved by beginning with a basic model like GAM, which is a generalized additive model (GAM) and later making it more complicated (thus more precise) like GA2M (GAM with pairwise interactions) or by beginning with a complicated model, like XGBoost and then trying in interpreting the model locally using methods such as LIME. In NeurIPS (Neural Information Processing Systems) 2017 one of the top conferences in machines learning, a deliberative session was held between Yann Lecun (head of AI research at Facebook) and Rich Caruana (lead machine learning researcher at Microsoft on applications in healthcare). The discussion session held at the end of "Interpretable ML Symposium" highlighted the need for additional research on the accuracy as compared to. interpretability as a trade-off in machine learning applications for healthcare.

In this article we will look at the case of Severe Retinopathy Prematurity (Severe RoP) that can cause complete or partial blindness in infants if treated. Stevie Wonders is a well-known musician suffering from severe RoP. In the 1950s doctors began to realize the possibility of using oxygen to help premature infants. But, they needed an additional few years to understand that a high dose of oxygen could cause the vessels of the eye to expand in a different way. Children born prematurely and excessive oxygen therapy are likely to suffer for the rest of their lives from blindness. World Health Organization (WHO) estimates that, among the more than 130 million newborns born each year, about 15 million were born prematurely before 37 weeks of gestation have been completed. About 1 million infants die each year as a result of problems associated with premature birth. A lot of survivors suffer an ongoing disability, which includes hearing and visual issues and learning difficulties. One of these is retinopathy of premature birth (RoP) is a major and severe reason for disability. Retinopathy of Prematurity (RoP) was first recognized by Terry 11 in 1942, as a form of developmental proliferative, vascular, and retinal disorder that occurs in retinas of premature infants which have not yet completed vascularization.

In addition to cortical blindness RoP is one of the leading causes of blindness in children around the world. The severity of RoP is increased with a lower birth weight and a shorter gestational week and a variety of risk factors have been identified as contributing to the growth of RoP. If the severity of RoP is not treated the condition can lead to bliness and retinal detachment. This is why it is crucial to identify the problem early and treat the condition appropriately to avoid the development of the condition. International Classification for Retinopathy of Prematurity (ICROP) is used to determine the progression and severity of the condition. The classification starts with the 1st level being the tiniest form of diagnosis for RoP and ends with the 5th level, retinal detachment as the most severe consequence of RoP.

The data we used to be used in our study was gathered from The Newborn Clinic of Zeynep Kamil Woman and Child Diseases Hospital in Istanbul between 2011 and 2014. Each year in Turkey approximately 150 000 newborns are born with birth weights less than 1500g. The newborns with these weights have a greater likelihood of being diagnosed with diagnosed with severe diagnosed with severe. We will study the risk factors considered to be the cause or may be associated with the presence of severe RoP. As we develop our prediction model, we will be able to observe how our model becomes more interpretable by combining categorical and numerical data and include additional interaction terms.

Literature on Interpretability of Machine Learning in Healthcare:

We have first looked into diverse papers and discussions regarding the current developments in machine learning and healthcare. The debate panel discussion at NeurIPS 2017 was especially stimulating. The discussion occurred as part of an "Interpretable ML Symposium" conducted during NeurIPS 2017. We chose to focus on research on interpretability and the interpretability vs. the trade-off between accuracy and interpretability in models of machine learning for healthcare. One of the participants also gave an excellent presentation on their most-cited research paper on the machine learning's interpretability in healthcare at the Allen Institute of Artificial Intelligence. The talk and the accompanying paper provided the direction we needed to steer the conversation of the conference presentation. The paper is listed as the top paper in a study conducted by Harvard and MIT experts in machine learning on healthcare. Omer Gottesman and colleagues have developed reinforcement learning strategies for sepsis in intensive health units. There is a section of recommendations for researchers in this paper, which was especially useful in our research of severe RoP using different methods of machine learning that reinforcement-learning. In general, the data collection and results interpretation sections in this paper were helpful for guiding our research.

As a part of the related works In addition, we must include the related study that explains the in-depth explanation of Generalized Additive Models that have Pairwise interactions. For a more comprehensive analysis of the interpretability and machine learning within healthcare settings, we've been referring to two great studies. Muhammad Aurangzeb Ahmad, et al., discuss in depth the latest developments in interpretability and. the accuracy trade-off in the context of healthcare. Maryzeh Ghassemi, et al. discuss the current possibilities of machine learning in applications in healthcare. There is also a segment on causality as well as a section on statistical inference. Also, we should mention that the NeurIPS 2017, conference, [9] as well as"Interpretable" ML Symposium, "Interpretable ML Symposium" held in conjunction with this conference. [10] We referenced the papers and video presentations on the site of the symposium and conference.

We believe that the interpretability of ML in the field of healthcare is a hot topic and applications are beginning to enter the clinical realm in several countries. B. Previous studies in severe retinopathy of Prematurity The list of references is not complete without the research literature for our case study, which is a severe prematurity-related retinopathy (RoP). Retinopathy among newborns, particularly those who weigh less than 1500 grams, causes blindness when it is at stage 4 or more. So, routine examinations are conducted by neonatologists, nurses as well as eye doctors to make sure that treatment is given when RoP crosses a certain threshold, that is. serious RoP diagnosis.

In Turkey approximately 17 % of babies born with weights of less than 1500 grams are taken to an eye doctor to be diagnosed with Stage 3 of RoP. The eye doctor performs periodic checks to determine if severe RoP is developing and attempts to treat the severe RoP. In 2018 the 5000+ participant study with statistical findings regarding risk factors was conducted by the Turkish Institute of Turkey (TR-ROP) to infants with birth weights less than 1500 grams. Seven risk factors were identified to be statistically significantly associated to the severity of RoP development. These findings were first discovered by using univariate, then multivariate logistic regression. This research forms the basis of future RoP studies in Turkey. Another study, which is which is in the process of being published was done based on the data collected between 2011 and 2014 in Zeynep Kamil Maternity and Child Diseases Hospital. The researchers have released their information (ZK-RoP) in this study to employ interpretable machine-learning methods. In the paper, out of 1066 newborns who had birthweights lower than 2000g, the authors examined the cases of 109 who were that were diagnosed as having severe RoP .

III Methodology

Data Wrangling

We have decided to use both R statistical programming language and SPSS to replicate our results. R is extensively used in biostatistics, and its ML tools are also well documented. SPSS was also used to replicate and cross-check our results with medical doctors who are much more familiar with SPSS than R or Python. As per the recommendations to researchers on our reference paper [3], during the cleaning and wrangling of our data, we worked closely with domain expert neonatologists. Some outlier data was removed after their consultation. Also, we discovered several wrong entries that were obviously the result of typos.

The reference paper on RoP [4] contains over 20 risk factors for RoP. 7 of those risk factors were found to be statistically significant after multivariate logistic regression analysis. We could not create two of these risks factors from our 2011-2014 RoP data; thus used a new set of 13 risk factor, 3 numerical and 10 categorical (binary). These risk factors were chosen among the statistically significant ones after univariate analysis of ZK-RoP data, similar to the approach carried out in TR-RoP study. One of these 13 risk factors is the most critical: the number of days the newborn receives oxygen. Our data from ZK study had 10 columns for various oxygen interventions. 5 of the columns were whether a particular type of intervention was conducted or not. The remaining 5 were for how many days this intervention conducted. After consulting with domain expert neonatologists, we decided to combine 4 of these columns (ignoring 1 column) into a total sum for the days oxygen intervention was used. We eventually decided to turn this numeric value into a categorical variable, whether the child received any oxygen support on mechanical ventilation.

RoP in itself is diagnosed categorically from 0 to 5. Severe RoP is a separate diagnosis, where RoP level 3 diagnosed patients are sent to an opthamologist for further diagnosis. Doctors are interested to predict severe RoP given any type of RoP has been diagnosed already. So we tried to predict 109 severe RoP cases out of 385 RoP cases, not out of the whole sample size of 1066: all newborns under 2000 grams.

IV Proposed Method

As in the original TR-ROP study and also as noted in many papers on biostatistics research, it is important to include not just the p-value, but also the confidence interval (CI) and the odds ratio. Statistical significant is important, but the effect of this statistical significance should be high enough to be considered clinically significant. Thus, we should not exclude values with p-values greater than 0.05. In the SPSS report for binary logistic regression, the beta values next to the confidence interval would be the corresponding "Odds Ratio". As noted in reference paper [3], it is important to work closely with domain experts to interpret causality and all types of correlations. Because TR-RoP study decided to only include 7 statistically significant terms into the multivariate logistic regression, we initially decided to follow their conclusion.

It should be noted however a more complex multivariate regression with more risk factors, including those with p-values greater than 0.05 might lead to a more accurate and more interpretable model. For example, our domain experts suspected that gender could be a clinically significant risk factor whereas it was not included in the TR-ROP study multivariate analysis due p-value higher than 0.05. Thus, we decided to include 13 risk factors in our multivariate logistic regression for the sake of interpretability as well as higher accuracy of the model. Because we know from previous literature on logistic regression that per every 20-30 sample a new variable can be introduced to the logistic regression formula, we constructed the formula with 385/30 ~ 13 variables.

Initially, after running the binary logistic regression with 20 variables (4 numerical, 16 categorical), we plotted the correlation matrix to see which of these risk factors are correlated. If there was a higher correlation than absolute value of 0.3 between any of the risk factors, we only kept one of the risk factors by choosing the one with the higher odds ratio and/or lower p-value. Thus, we tried to keep our regression variables as uncorrelated as possible. Such an approach would also increase interpretability as it would make our risk factors more independent of each other. We used this approach to reduce our variable count to 12. We found that birth weight was highly correlated with gestational week and kept only the gestational week. The comparison criterion between models to calculate the accuracy is particularly important. However, if accuracy was our primary concern, we could use a more complex approach like an artificial neural network. Thus, it is important to note that the investigation of interpretability was our primary goal during this study.

SPSS has a classification table that comes at the report of a binary logistic regression. It can be used to measure the accuracy of the model, including type I and type II errors. For simplicity in R, we decided to use the standard AIC values that are present in R with the summary(model) command. However, in most other intelligent ML research, AUC values are used [2]. It is also known that AIC penalizes complex models with extra variables in comparison to simpler models. Thus, a further discussion might be needed to figure out the best criterion to compare interpretable ML models as we increase complexity and sacrifice interpretability for accuracy

V. RESULT ANALYSIS

We run the multivariate linear regression with 12 risk factors: 2 numerical and 10 binary (categorical) values. We had first run the regression separately with numerical and categorical values. Then we combined the numerical and categorical risk factors in one equation to run the multivariate logistic regression analysis that would predict severe RoP as a binary value.

We should note that our base accuracy rate is 109/385 = 28.3%. If we diagnose all patients as severe RoP, we would make no mistake of type II error, and we would effectively diagnose all real cases of severe RoP. However, our type I error would be maximized. Thus, our overall accuracy will go down. In particular, if we set the cut-off for classification at 0.05 so that no type II error occurs, then only 19 patients would be correctly diagnosed as not having severe RoP. Thus, our overall accuracy will be only slightly better (31%) than the base case of 28.4%. Because our data is skewed, (109 severe RoP vs. 276 non-severe RoP), a case could be argued to use F1 score as a combination of precision or recall. However, given the seriousness of TypeII type error in severe RoP diagnosis, we didn't want to mislead that an increased F1 score would be of better value than a minimal TypeII error. Moving on, we wanted to evaluate how the goodness of fit of our model changed as more covariates and interaction terms were added. To compare the current model versus the null (intercept-only) model, we used the omnibus test as a likelihood-ratio chi-square test (102.654). The significance of this test was 0.000; thus, we concluded that our model was outperforming the null model. Note that our sample size is 385 and the number of our covariates is 12 plus 5 optional interaction terms. Because 385/12 is less than the 40 threshold mentioned in statistical literature for small sample sizes, the goodness of fit between models was measured with AICC (394) instead of AIC (392) as criteria, but we observed no significant difference between these 2 comparison metrics. We observed that including the interaction terms reduced the AICC value by around 1-2%, thus, adding the interaction terms was mainly to increase the interpretability of our model to confirm the pairwise interaction of the suspected risk factors.

There are other metrics to compare models, but we used AIC for this paper due to its simplicity in implementation given our toolset in SPSS and R. Both AIC and BIC penalize complex models, and along with cross-validation, they are commonly used to prevent overfitting. In future, we intend to expand our work with cross-validation in Python using scikit-learn machine learning library. Furthermore, scikit-learn library allows comparison of machine learning models in a simpler fashion compared to SPSS. Currently, standard SPSS supports decision tree analysis using cross-validation, but not with logistic regression. We wanted to know which few interaction terms had most impact on the model. If we had run a full factorial analysis of 12 terms, by adding every possible interaction term, it could take a very long time. Thus, we kept it is simple and only added the pairwise interaction terms: 12*11/2 = 66. Thus, our total number of terms in the regression was 12+66=78. After running this regression, only 5 of the interaction terms had p values less than 0.10. Thus, we combined these 5 pairwise interaction terms with the original 12 risk factors and run a final regression of 17 variables. Thus, we were able to give doctors a more interpretable machine learning model with 5 interaction pairs. These 5 pairs were confirmed by expert neonatologists as risk factors that were most significant that lead to severe RoP. Confirming our results with domain experts, we have shown that a generalized additive model with pairwise interactions was increasing the interpretability of

ISSN: 2366-1313

the model. Our approach was consistent with the research suggestions in the cited papers on interpretability of machine learning in healthcare. [2,3]

The adding of these 5 pairwise interactions increased the accuracy of our model from 33.0 to 33.5 We had chosen the cut-off for 0-1 classification to be 0.05 instead of the 0.5 value, because we wanted to minimize the type II error. We have also not used the ROC metric with various classification thresholds, because classification threshold invariance was not desirable. Minimizing the type II error is critical in healthcare, where a missed diagnosis could lead to full blindness of the patient for lifetime. There was no practical benefits to explore cases where type II error was not zero, but when we relaxed our type II error being zero constraint, we could see a small increase of accuracy, from 75% up to 76.6%, by the adding of most significant pairwise interactions. Thus, we concluded that GA2M approach was helpful mostly with interpretability of the machine learning model by confirming hypothesis of risk factor two-way interactions.

However, comparing our RoP data set with other diseases in healthcare machine learning, it should be noted that the size of our data set is relatively small and also many variables are categorical, rather than numerical. Thus, the accuracy effect of GA2M approach over GAM was rather minimal. Further work should focus on collecting numeric data and turning some of the categorical variables into numeric ones, such as blood transfusion as number of times rather than a binary value. Finally, we should note that further work should be done to compare GAM and GA2M to other explainable machine learning techniques such as decision trees. Based on our decision tree approach with cross-validation in SPSS, which resulted in a higher accuracy rate with 44% and only a few type II errors, perhaps our RoP prediction problem and data was a better fit for other interpretable machine learning techniques. Additional research is needed to compare various interpretable ML techniques and cross-validation on RoP data. Microsoft recently launched an interpretable ML library, which promises both high interpretability and high accuracy, and we hope to use it along with scikit-learn ML library as we expand our explainable ML work based on this paper. .

REFERENCES

[1] R. Caruana, Y. LeCun, The Great AI Debate "Interpretable ML Symposium" as part of NeurIPS - 2017. https://www.youtube.com/watch?v=93Xv8vJ2acI

[2] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1721–1730. ACM, 2015.

[3] O. Gottesman, F. Johansson, J. Meier, J Dent, D. Lee, S. Srinivasan, L. Zhang, Y. Ding, D. Wihl, X. Peng, J. Yao, I. Lage, C. Mosch, L. H. Lehman, M. Komorowski, A. Faisal, L. A. Celi, D. Sontag, and F. Doshi-Velez. Evaluating Reinforcement Learning Algorithms in Observational Health Settings. pp.1-16, 2018. https://arxiv.org/pdf/1805.12298.pdf

[4] A.Y. Bas, N. Demirel, E. Koc, D. Ulubas Isik, I.M. Hirfanoglu, T. Tunc, and TR-ROP Study Group. Incidence, risk factors and severity of

retinopathy of prematurity in Turkey (TR-ROP study): a prospective, multicentre study in 69 neonatal intensive care units. Br J Ophthalmol. 102(12):1711-1716, 2018.

[5] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate Intelligible Models with Pairwise Interactions. KDD2013, August 11–14, 2013, Chicago, Illinois, USA. http://www.cs.cornell.edu/~yinlou/papers/lou-kdd13.pdf

[6] M. Aurangzeb Ahmad, C. Eckert, A. Teredesai, and G. McKelveyet. Interpretable Machine Learning in Healthcare. IEEE Intelligent Informatics Bulletin. Vol.19 (1): pp.1-7, August 2018.

[7] M. Ghassemi, T. Naumanne , P. Schulam, A.L, Beam, and R. Ranganath. Opportunities in Machine Learning for Healthcare. https://arxiv.org/abs/1806.00388

[8] S. Sancak, S. Topçuoğlu, G. Çelik, M. Günay, G. Karatekin. Prematüre Retinopatisi Sıklığı ve Risk Faktörlerinin Değerlendirilmesi. Zeynep Kamil Tip Bülteni;2019;50(1):63-68.

[9] Thirty-first Conference on Neural Information Processing Systems (NIPS2017 or NeurIPS2017) https://nips.cc/Conferences/2017

[10] Interpretable ML Symposium, NIPS 2017 http://interpretable.ml

[11] Terry, T L. "Fibroblastic Overgrowth of Persistent Tunica Vasculosa Lentis in Infants Born Prematurely: II. Report of Cases-Clinical Aspects." Transactions of the American Ophthalmological Society vol.40 (1942): 262-84.

[12] Stone, M. "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, 1977, pp. 44–47.