2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)

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

Tamer Karatekin

Uz.Dr. Selim Sancak, Op. Dr. Gökhan Çelik,
Asst. Prof. Sevilay Topçuoğlu, Prof. Dr. Güner Karatekin
University of Health Sciences,
Zeynep Kamil Maternity and Children's
Training and Research Hospital,
Department of Pediatrics, Division of Neonatology
Istanbul, Turkey

Asst. Prof. Pınar Kırcı
Engineering Sciences Department
Istanbul University-Cerrahpaşa
Istanbul, Turkey

Prof. Dr. Ali Okatan
Computer Engineering Department
Istanbul Gelisim University
Istanbul,Turkey

***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 research trends in healthcare applications of machine learning focus on the interpretability of the models. While in some healthcare applications accuracy of the model is much more important than its interpretability, there are many cases where interpretability is preferred despite loss of accuracy. The ideal scenario where the model has both high accuracy and high interpretability can be achieved either by starting with a simple model such as a generalized additive model (GAM) and then making it more complex (thus more accurate), such as GA2M (GAM with pairwise interactions), or by starting with a complex model, such as XGBoost, and try to interpret it locally with methods like LIME.

During NeurIPS (Neural Information Processing Systems) 2017, one of the leading conferences in the world on machine learning, a debate session was arranged between Yann Lecun (head of AI research at Facebook) and Rich Caruana (lead machine learning researcher at Microsoft on healthcare applications). This session at the end of the "Interpretable ML Symposium" highlighted the need for further case studies on accuracy vs. interpretability trade-off in machine learning for healthcare. [1][9][10]

In this paper, we will investigate a case study around Severe Retinopathy of Prematurity (Severe RoP) that leads to full or limited blindness in newborns if not treated. Stevie Wonders is a well-known musician who suffered from severe RoP. Back in the 1950's, doctors were recently realizing that oxygen could be used to save premature babies. However, it took them additional years to figure out that excess dosage of oxygen would cause vessels in the eye grow irregularly. Premature babies with this condition and excess oxygen treatment would suffer from lifelong blindness.

World Health Organization (WHO) estimates that, out of the 130 million babies born annually, around 15 million babies are born prematurely, before 37 completed weeks of gestation. Approximately 1 million children die each year due to complications of preterm birth. Many of the survivors face a lifetime disability, both visual and hearing problems, as well as learning disabilities. Among these issues, retinopathy of prematurity (RoP) is a leading and serious cause of disability.

Retinopathy of Prematurity (RoP) was first reported by Terry [11] in 1942 as a developmental, vascular, and proliferative retinal disorder occurring in the premature newborns' retinas that have not completed vascularization. Along with cortical blindness, RoP is among the most common causes of childhood blindness in the world. The seriousness of RoP intensifies with lower birth weight and

978-1-7281-2914-3/19/$31.00 ©2019 IEEE
DOI 10.1109/Deep-ML.2019.00020

lower gestational week, and various risk factors are reported relating to the development of RoP. If severe RoP is untreated, it causes bliness with retinal detachment. Thus, it is very important to diagnose early and use appropriate treatment to prevent the progression of the disease.

International Classification for Retinopathy of Prematurity (ICROP) is used to classify the progression and seriousness of the disease. Classification starts with 1st level as the lightest diagnosis of RoP and ends with 5th level, retinal detachment, as the heaviest outcome of RoP.

The clinical data for our analysis was collected by the Newborn Clinic of Zeynep Kamil Woman and Child Diseases Hospital in Istanbul between 2011 and 2014. Every year in Turkey, around 150 000 babies are born with birthweights below 1500g. These newborns have a higher tendency to be diagnosed with severe RoP. We will investigate the risk factors that are thought to cause or to be correlated with severe RoP. As we improve our model of prediction, we will observe how the accuracy of our interpretable model increases as we combine numerical and categorical values and add further interaction terms.

II. Related Works

A. *Literature on Interpretability of Machine Learning in Healthcare*

First, we have investigated various papers and discussions on the popular trends in machine learning for healthcare. The debate session during NeurIPS 2017 was particularly encouraging. [1][9] This debate was part of the "Interpretable ML Symposium" conducted during NeurIPS 2017. [10] Thus, we decided to focus our research efforts on interpretability vs. accuracy trade-off in healthcare machine learning models.

One of the debate participants also had an excellent talk about their most cited paper on interpretability of machine learning in healthcare at Allen Institute of Artificial Intelligence. His talk with the associated paper gave the direction we wanted to direct the focus of our conference presentation. [2]

This paper is cited as top of the list on a recent study done by Harvard and MIT machine learning experts on healthcare. [3] Omer Gottesman, et al, have implemented reinforcement learning methods for sepsis in intensive care units. There is a recommendation section for researchers on this paper that was particularly useful as we investigated this study on severe RoP with machine learning methods other than reinforcement learning. Overall, data gathering and result interpretation sections of this paper have been useful to guide our study.

Among related works, we should also include the complementary study that explain in depth the Generalized Additive Models with Pairwise Interactions. [5,2].

For a more general discussion on interpretability of machine learning in healthcare, we have also referred to two excellent papers. Muhammad Aurangzeb Ahmad, et al, explain in detail the current trends in interpretability vs. accuracy trade-off in healthcare settings. [7]

Maryzeh Ghassemi, et al. [8], talk on current opportunities on machine learning for healthcare applications. They also include a section on causality and about statistical inference.

We should also note the NeurIPS 2017 conference [9] and the "Interpretable ML Symposium" held during this conference. [10] We referred to the papers and their associated video presentations on the website of the conference and symposium.

Overall, we believe interpretability of ML in healthcare is a hot field, and applications are entering clinical usage in many countries.

B. *Previous Studies in Severe Retinopathy of Prematurity*

Our list of references would be incomplete with the disease literature regarding our case study, severe retinopathy of prematurity (RoP). Retinopathy in newborns, especially those born with weights below 1500 grams, leads to blindness if it reaches stage 4 and beyond. Thus, regular checks are made by nurses, neonatolog doctors, and eye doctors to ensure treatment is applied if RoP passes a certain threshold, ie. severe RoP diagnosis.

In Turkey, around 17 % of all newborns with birth weights below 1500 gram are sent to the eye doctor with diagnosis of stage 3 RoP. The eye doctor then does regular checks if severe RoP might develop and attempts to treat severe RoP.

In 2018, a 5000+ patient study along with statistical conclusions on risk factors was conducted in Turkey (TR-ROP) for babies with birth weights below 1500 grams [4] Seven risk factors were found to be strongly and statistically significantly correlated with severe RoP development. These results were reached first using univariate and then multivariate logistic regression techniques. This research constitutes the backbone of further RoP research in Turkey.

A second study, currently in press, has been conducted based on data collected between 2011-2014 in Zeynep Kamil Maternity and Child Diseases Hospital. [8] The researchers shared their data (ZK-RoP) for this paper to apply interpretable machine learning techniques. In that paper, among 1066 newborns with birthweights below 2000g, the authors investigated 109 cases diagnosed with severe RoP.

III. Material and Methods

The material of our study starts with retinopathy of premature (RoP) data between 2011-2014 from Zeynep Kamil Maternity and Child Diseases Hospital. It is a recording of 5000+ patients over 102 variables. (ZK-RoP) We were interested in the same 20 risk factors analysed in TR-RoP study [4] for patients with birth weights below 1500 grams. Because ZK-RoP study worked with newborns below 2000

grams, we also worked with their data of 1066 patients below 2000 grams.

Among 1066 newborns with birthweights below 2000g, 109 cases diagnosed with severe RoP were investigated. However, it must be noted that the important task is to predict those who might develop severe ROP from among those who have already been diagnosed with any type of RoP. Thus, our sample size was reduced to 385. Out of 385 patients diagnosed with any type of RoP, we tried to build a model that would predict the 109 patients diagnosed with severe RoP.

Thus, to minimize type II error and to minimize wrong diagnosis of a disease that could lead to blindness, our baseline accuracy rate was 109/385=28.3%

We run univariate and multivariate logistic regression machine learning algorithms, starting with generalized additive model (GAM) to predict severe RoP based on the same risk factors as in TR-ROP study [4] We also added interaction terms to our multivariate analysis.

We wanted to follow the same methodology as in TR-RoP paper and check their results using a new sent of data, ZK-RoP. Furthermore, we wanted to include accuracy analysis of these predictions, which was missing in the TR-RoP study as it was not focusing for an ML audience.

There are multiple ML algorithms we could use, but we wanted to start with the simplest ones and work our way up to the more complex ones. In this paper, we only focused on the multivariate logistic regression analysis. In future papers, we will conduct analysis with more complex models such as decision trees, k-means, and xgboost.

C. *Abbreviations and Acronyms*

Throughout the rest of this paper, we use RoP to denote retinopathy of prematurity. We used univariate and multivariate logistic regression along with other machine learning (ML) models that may be used with their popular abbreviations, such as Generalized Additive Model (GAM) or Generalized Additive Model with Interaction Pairs (GA2M). Models such as GA3M and GAXM exist, but for our purposes based on previous research, we have kept our model with only the most needed and simple interaction pairs. [2] We will also refer to 2011-2014 RoP study as ZK-RoP study, short for Zeynep Kamil.

We should also note that AIC stands for "Akaike's Information Criteria", a metric developed by the Japanese Statistician, Hirotugu Akaike, in 1970. AIC penalizes the inclusion of additional variables to a model. This penalty error increases when including additional terms. When 2 models are compared, the model with the lowest AIC is the better model. Similarly, Bayesian Information Criterion (BIC) exists, which penalizes additional terms even further.

AIC, BIC, and cross-validation are primarily used to prevent overfitting. It should be noted that asymptotically minimizing the AIC is equivalent to minimizing the cross-validation value. This is true for any model, not just linear models. [12] Therefore, AIC is a popular metric when comparing models.

AUC stands for "Area Under the Curve". AUC is used interchangeably with AUROC, " Area Under the Receiver Operating Characteristic curve". It is also referred as the c-statistic. A ROC curve plots true positive rate (TPR) vs. false positive rate (FPS) at different classification thresholds.

D. *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.

E. *Equations*

Because we are trying to predict a categorical variable, whether the patient is diagnosed with severe RoP or not, we will use logistic regression instead of linear regression. Univariate logistic regression holds only one variable $\ln[Y/(1-Y)]=a+b_1X_1$, while the formula for multivariate logistic regression is as follows: $\ln[Y/(1-Y)]=a+b_1X_1+b_2X_2+b_3X_3+...$

The generalized additive model (GAM) with only 1st order terms is equivalent to the equation above. If we want to add 2nd order terms, we would need to add them as follows: $\ln[Y/(1-Y)]=a+b_1X_1+b_2X_2+b_3X_3+b_{12}X_1X_2+b_{13}X_1X_3+b_{23}X_2X_3+...$

For example, for a GAM with 6 first-order variables $(X_1,X_2,X_3,X_4,X_5,X_6)$, there would be 6*5/2=15 pairwise interaction terms. It is suggested to add only a few and only start with the most influential pairwise interactions. [2,5] Thus, as we make our model more complex, we will add the pairwise interactions one by one.

Note that 3 of the risk factors (features of the model) are numerical values, while the other 10 are categorical values. These 13 risk factors were chosen among a list of 20+ risk factors based on initial univariate analysis. The numerical risk factors are birth weight, birth week, and total days of oxygen given. In the TR-RoP study, the categorical risk factors were late-sepsis, sga, and whether the number of blood transfusions is greater equal than 2. In our analysis of ZK-RoP data, we picked 10 binary risk factors that were statistically significant, including late-sepsis and whether blood transfusions is greater equal to 1. The other 8 risk factors were chosen in consultation with domain expert neonatologists.

When summary(model) command is run in the R programming language, the AIC values are listed to approximate the accuracy of one model over another. SPSS, on the other hand, lists various values, including AIC, as measures for goodness to fit, in other words for accuracy comparison between models. In a journal paper, we are interested in exploring the differences between various of these accuracy measurement tools.

For this conference presentation, we limited our analysis to a comparison of TR-RoP study with ZK-RoP study. Thus, we used multivariate logistic regression techniques with numerical and categorical data on ZK-RoP study. Finally, we added two pair interaction terms to observe how our model has changed. We followed the math and explanations of previous papers on GAM with pairwise interaction. [2] [5].

F. *Considerations on Statistical Values & Best-Fit Model*

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.

IV. Results & Conclusions

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.

Table 1

Binary Logistic Regression without interaction terms (GAM)	Wald Chi-Square	Signifi-cance p-value	Exp(B) ODDS RATIO	95% Confidence Interval for EXP(B)	
				Lower	Upper
gestational week	12.51	.000	.737	.623	.873
mechanical ventilation	4.54	.033	1.016	1.001	1.031
blood transfusion	2.23	.135	1.706	.846	3.438
gender	5.49	.019	1.868	1.107	3.151
late-onset sepsis	.74	.390	1.263	.742	2.149
chorioamnionitis	.16	.690	1.211	.473	3.097
preterm premature rupture of membranes	1.23	.267	1.415	.767	2.609
antenatal steroid therapy	2.89	.089	.616	.353	1.077
respiratory distress syndrome	.50	.480	.772	.377	1.582
dopamin-dobutamin	.44	.505	.815	.446	1.489
necrotizing enterocolitis	1.31	.252	4.565	.339	61.41
intraventricular hemorrhage	2.64	.104	2.029	.865	4.763
constant	10.82	.001	5906		

- GW stands for gestational week, unit is weeks. MV is the total days of oxygen received on mechanical ventilation. Other risk factors are binary.
- It should be noted that when we add 2nd order interaction terms that have most impact on the regression, they affect the odds ratio, significance level, and confidence interval of the previous individual interaction terms. Thus, we included only a table without the interaction terms of the multivariate logistic regression analysis.
- Features were chosen in consultation with neonatologists among 100 recoded variables of clinical data on RoP.

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 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.

Acknowledgment

We would like to thank University of Health Sciences, Turkey, and the Zeynep Kamil Maternity and Child Diseases Hospital for sharing their 4-year long RoP data for this study. We also thank Rich Caruana from Microsoft Research who gave feedback on the paper for future improvements.

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.