Risk-based Assessment of ML-based Medical Devices

Martin Haimerl

Innovation and Research Center Tuttlingen of Furtwangen University – Furtwangen University
Martin.Haimerl@hs-furtwangen.de

Abstract. The rating of risks is a crucial aspect for assessing the performance of medical devices. For machine learning (ML) based systems, this means that an integration of risks into the corresponding metrics should be addressed. The main goal of this paper is to demonstrate the effect when differences in the impact of certain errors is not adequately considered during the development of ML based systems, in particular when they refer to classification problems. An artificial model was utilized to demonstrate the different outcomes when considering different risk ratings. The differences were analyzed visually as well as quantitatively. As a result, a difference of up to 50% was obtained for the total outcome, when a ratio of 4.0 between the types of risks was assumed. This demonstrates that differences in risk impact should be systematically considered and integrated into the associated metric, when assessing the performance of ML based medical devices.

Keywords: Machine Learning; Evaluation; Performance Metrics; Medical Devices; Risk Management.

1 Introduction

The assessment of machine learning (ML) based systems strongly depends on the criteria which are used for the training, validation, and testing of the developed model. These criteria have to provide a score or metric how well the models perform. For supervised learning approaches, the agreement between the ground truth and the values predicted by the model is a core aspect of such metrics. For classification problems, this means the rates of successful assignment to the particular classes. Standard techniques for the assessment of binary classification are accuracy rates, false positive (FP) / false negative (FN) rates, or receiver operator characteristics (ROC curves) (cf. [1] for an overview of common assessment criteria). These techniques refer to the probabilities how often correct respectively incorrect assignments to the classes occur in the training, validation, and testing data sets.

These techniques are well established. However, they are often applied in a very generic way without considering the specific context of its application. In particular for medical applications, such a generic approach has considerable limitations. The impact of different types of errors needs to be considered in an application-specific way. An FN (e.g. missing the presence of a tumor illness) may be more critical than an FN (false alarm for the disease, which can be double-checked subsequently). From the perspective of the patient health, an FN often leads to substantially worse outcomes. Such risks should be considered in a dedicated way when assessing the overall performance of a classifier. This means that not only accuracies and probabilities have to be taken into account but also the impact of the different types of errors. This leads from a primarily probability-based to a risk-based approach for the assessment of ML-based classification systems. Similar approaches can also be found under the names cost curves [2], utility curves [3], or decision curves 4].

For demonstrating the impact of different types of errors, this paper develops a risk-based approach for the assessment of medical devices and analyzes its differences when comparing it to methods using probability-based measures. A parametric model is used for the distributions to systematically analyze the changes in outcome. Assessment scores are developed and analyzed which address individual risks for a single patient as well as aggregate risks representing the overall performance of the device.

2 Materials and Methods

In this paper, a generic setup is used with a classifier F predicting the binary outcome $Y \in \{0,1\}$ from a set of input features X, i.e. the prediction is performed according to $\hat{Y} = F(X)$. This prediction is performed on a set of test data (X_i, Y_i), where Y_i are considered as the ground truth, i.e. the correct classification values for the input values X_i. The classifier is considered to depend on a threshold s, i.e. it predicts a 1 if and only if a certain score value $S = S(X)$ is above the threshold s. Thus, a particular instance of the classifier can be represented by a binary-valued function $F(s, X)$ which includes the threshold s as a parameter. In this paper, we utilize an artificially constructed error distribution to demonstrate the behavior of performance metrics when certain parameters get changed. This means, that we assume that the false positive $FPR(s)$ and false negative rates $FNR(s)$ are given by a parametric function. We use modified Gaussian functions of the form $FPR(s) =$

$(1-s) \cdot \exp\left(\frac{s^2}{\sigma_{FP}}\right)$ and $FNR(s) = s \cdot \exp\left(\frac{(1-s)^2}{\sigma_{FN}}\right)$, for this purpose. The terms $(1-s)$ and s modify the Gaussians in a way that $FPR(1) = FNR(0) = 0$.

As a next step, a risk model is constructed which assigns certain "costs" to the different types of errors FP and FN. These costs reflect the risks or other associated costs which are caused by the particular type of error. Subsequently, they are named risk scores and denoted by R_s. In terms of conditional probabilities $P(\hat{Y}|Y)$, the individualized risk $IR(s)$ can be calculated as the expected risk for a particular individual, i.e. $IR(s) = E(R_s(\mathrm{FP}) + R_s(\mathrm{FN})) = E\left(R_s\left(P(\hat{Y}=1|Y=0)\right) + R_s\left(P(\hat{Y}=0|Y=1)\right)\right)$, where $E(\cdot)$ denotes the expectation value. In this paper, we do not include risks which depend on the threshold levels. Thus, the risk scores boil down to $IR(s) = E(R_s(\mathrm{FP}) + R_s(\mathrm{FN})) = c_{FP} \cdot FPR(s) + c_{FN} \cdot FNR(s)$, where c_{FN} and c_{FR} are constants reflecting the impact of the particular type of error. Further on, not the absolute values but only the relationship between the costs of FP and FN matters. Thus, the value of c_{FP} can be set to 1, without loss of generality. Subsequently, the equation reduces to $IR(s) = E(R_s(\mathrm{FP}) + R_s(\mathrm{FR})) = FPR(s) + c_{FN} \cdot FNR(s)$. Subsequently, c_{FN} is called risk ratio.

So far, all these curves reflect an individual risk, since they only take into account the general error rates for a particular individual as well as the impact such kind of an error has. The analysis does not encounter a situation where the number of positive $(Y=1)$ and negative $(Y=0)$ cases differ and where an aggregate risk score should be used as a reference. The aggregate risk score $AR(s)$ sums up all the individual risks for the given data set (X_i, Y_i). If we again assume an individual risk score of $c_{FP} = 1$ for FP and c_{FN} for FN, this overall risk score is calculated as $AR(s) = FP(s) + c_{FN} \cdot FN(s)$ where $FP(s) = \left|\{i | F(s, X_i) = \hat{Y}_i = 1, Y_i = 0\}\right|$ is the number of false positives and $FN(s) = \left|\{i | F(s, X_i) = \hat{Y}_i = 0, Y_i = 1\}\right|$ the number of false negatives for a fixed threshold s. Again, only the ratio $q = \frac{FN(s)}{FP(s)}$ between $FP(s)$ and $FN(s)$ matters, since we do not focus on absolute levels of risk values but only on relationships between them. This can also be written as $q = \frac{FNR(s)}{FPR(s)}$. Based on this, the aggregate risk score can be calculated as $AR(s) = FPR(s) + q \cdot c_{FN} \cdot FNR(s)$. Contracting q and c_{FN} to a single factor $\tilde{c}_{FN}$, we see that $AR(s)$ has the same form as the individual risk, i.e. $FPR(s) + \tilde{c}_{FN} \cdot FNR(s)$.

3 Results

Based on the described approach, the model was first applied to a test scenario where σ_{FN} and σ_{FP} were both set to 0.3. For the risk ratio c_{FN}, the values were set to $0.25, 0.5, 1.0, 2.0$, and 4.0. The results are shown in Fig. 1. On the left side, the model is shown with the corresponding $FPR(s)$ and $FNR(s)$ values. The diagram on the right side demonstrates the impact of different risk ratios c_{FN} on the overall outcome for the individual risks. The risks associated with both types of errors are balanced out where the curves have their minimum. For $c_{FN} = 1.0$, this is exactly at $s = 0.5$ as indicated by a short line for the black curve. This symmetry appears since the error distributions are symmetric between FPR and FNR. The remaining curves show the situation when the risk ratio c_{FN} takes on the other values $0.25, 0.5, 2.0$, and 4.0. For each curve, the minimum risk is indicated by a short line. The position changes, since it depends on the relationship of the impact for the different types of errors.

Additionally, the dotted lines show the reference $s = 0.5$, where the curve with equal risks between FP and FN, i.e. $c_{FN} = 1.0$, has its minimum. The intersection between the dotted line and the particular curve reflects the risk value which would have been obtained when the optimization would have been performed solely according to the error rates, i.e. without considering the risk factors. It can be seen that in the cases $c_{FN} \neq 1.0$, the minimum lies in a region where the curves considerably decay or increase. This demonstrates a deviation between the $c_{FN} = 1.0$ assumption and the actual risk ratio. This effect is more apparent with increasing difference between the actual c_{FN} and the balanced case $c_{FN} = 1.0$. The same basic model applies to the cases of the aggregate risk. However, the outcomes have to be interpreted according to a replacement of c_{FN} by $\tilde{c}_{FN}$. Thus, the basic analysis can be transferred to the aggregate risk case.

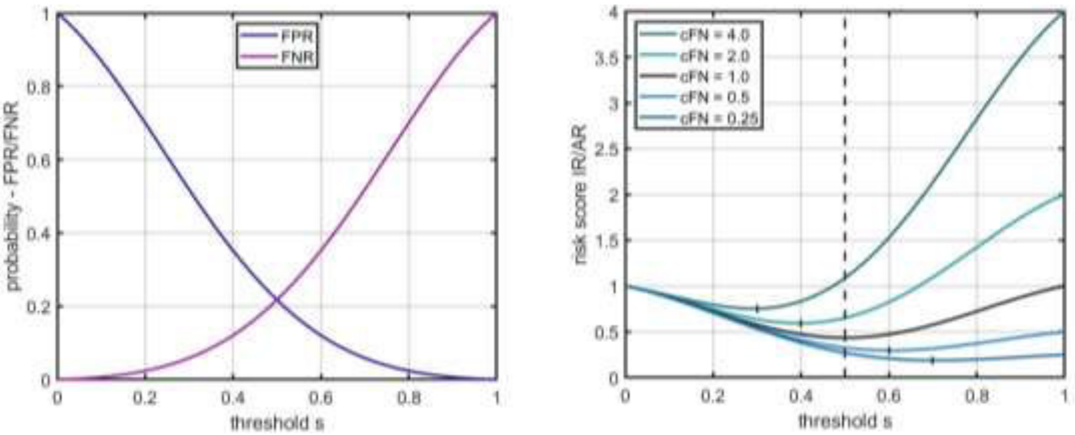

Fig. 1. Left side: Artificial model of error distributions, i.e. $FPR(s)$ and $FNR(s)$ in dependence of the threshold s. The model is based on modified Gaussian functions of the form $FPR(s) = (1-s) \cdot \exp\left(\frac{s^2}{\sigma_{FP}}\right)$ and $FNR(s) = s \cdot \exp\left(\frac{(1-s)^2}{\sigma_{FN}}\right)$, where $\sigma_{FP} = \sigma_{FN} = 0.3$. Right side: Risk scores $IR(s)$ respectively, $AR(s)$ for the same casse, when the risk ratio c_{FN} is varied ($c_{FN} = 0.25, 0.5, 1.0, 2.0,$ and 4.0). The short line shows the minimum for the particular curve. The dotted line represents the reference $s = 0.5$, where the curve with equal risks between FP and FN, i.e. $c_{FN} = 1.0,$ has its minimum.

Fig. 2 shows a comparison when different parameter setting for the artificial model are applied. This includes three scenarios with $\sigma_{FN} = \sigma_{FP} = 0.2, \sigma_{FN} = \sigma_{FP} = 0.3,$ and $\sigma_{FN} = \sigma_{FP} = 0.4.$ On the left side, the corresponding ROC curves are shown to provide an overview about the particular probability-based model performance. In the right diagram, the impact of the particular parameters on the risk-based approach is visualized. It is shown how much higher the resulting risk score would have been, when c_{FN}=1.0 would have been assumed instead of the suited risk ratio. The relative difference exceeds 50% for the case $\sigma_{FN} = \sigma_{FP} = 0.4$ and the risk ratios $c_{FN} = 4.0$ as well as $c_{FN} = 0.25$. On the c_{FN}-axis, a logarithmic scaling was applied since this better reflects that c_{FN} is a ratio parameter. In this logarithmic scaling, all the curves show a symmetric appearance with respect to the $c_{FN} = 1.0$ axis. The difference substantially decays when c_{FN} gets closer to 1.0. Additionally, it can be recognized that the differences decrease slightly when the parameters σ_{FP} / σ_{FN} decrease and subsequently the area under the ROC curve (i.e. the AUC value) increases. The AUC is a standard probability-based measure for assessing the performance of a classifier (see 1).

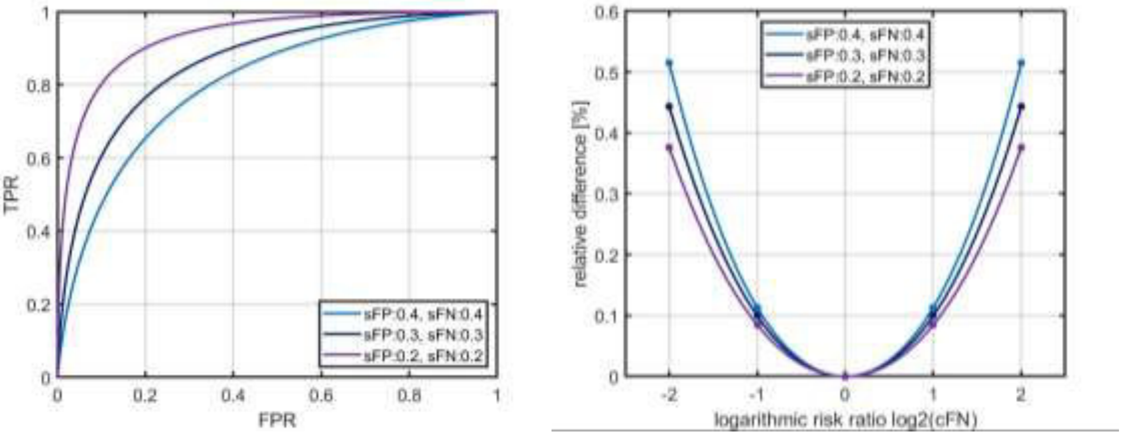

Fig. 2. Variation of the parameters of the artificial model including $\sigma_{FN} = \sigma_{FP} = 0.2, \sigma_{FN} = \sigma_{FP} = 0.3$ (i.e. same case as in Fig. 1), and $\sigma_{FN} = \sigma_{FP} = 0.4$. In the diagram, σ_{FP} is named sFP and σ_{FN} sFN. Left side: ROC curves for the particular cases. Right side: Increase in risk values, when a risk adaption would not have been performed, i.e. the standard probability based threshold $s = 0.5$ would have been used. On the vertical axis, this value is shown as a relative increase in comparison to the true optimum risk value. On the horizontal axis, the used risk ratios c_{FN} respectively $\tilde{c}_{FN}$ are shown. A logarithmic scaling is used for this axis.

4 Discussion

Using an artificial model for the error distribution, this paper demonstrates the relationships between a pure probability-based assessment of ML models on the one hand and risk-based approaches (individual as well as aggregate risks) on the other hand. It was demonstrated that substantial differences occur when risk factors are not addressed adequately. The difference in resulting risk scores goes up to 50% in this simple setting. According to the applicable regulations like the Medical Device Regulation (MDR) and associated standards like ISO 14971, the risks of medical devices should be adequately managed during device development. Following these requirements, appropriate scores should be applied to assess and optimize the outcome of ML-based devices or components. This is not included in standard metrics for ML-based classifiers, like accuracy, FP/FN rates, or also ROC curves. This could only be achieved using risk-based approaches. For this purpose, this paper provides insights regarding the potential approaches as well as the behavior when applying different risk ratios.

An additional question in this context is whether individual or aggregate risks should be used, i.e. whether the risk should be addressed for an individual patient or across the entire population / number of cases. In the latter case, the distribution of the error cases plays a central role. Such an approach is included in ISO 14971, representing the relevant risk management standard for medical devices. Not only the severity but also the likelihood of the hazards / harms for the patient have to be included according to 5. From this perspective, the application of an aggregate risk-based approach is applicable and should be pursued. One further challenge is the assessment of proper "costs" and likelihoods for the particular types of errors in a quantitative way. However, ISO 14971 allows to basically use semi-quantitative approaches for risk management. This means, that probabilities as well as the level of harm (or "costs") may be categorized. Thus, the rating could be addressed and integrated into the models in this way. Adjustments towards true quantitative ratings could be approached during the lifetime of the ML based system, i.e. when enough data is gathered during the operation of the device in real settings. Of course, the rating of different types of errors is strongly application specific and often a balancing of different types of impacts is difficult to justify. It also includes challenging ethical questions. More research in this direction is required to provide proper approaches for achieving an overall optimal result.

This study has some limitations. First of all, the model does not reflect a real case scenario. Thus, future research should include an analysis of the behavior in concrete applications using the actual error distributions. This would also enable to better address the derivation of appropriate risk ratios for the particular applications. Further on, our model includes some simplifications. It assumes that the risk ratios are constant, i.e. do not depend on the threshold s. This may not be adequate in real case scenarios when the severity of the risks may change between clear diagnoses (high values for s) and ambiguous cases (with s in the mid range). For example, a lower rate of pathological substances in a diagnostic test may be associated with less severe courses of an illness. In the current paper, we do only apply the risk scores during threshold selection and not within the model training, i.e. directly in the optimization procedure. Thus, the analysis could be extended in this direction, in the future. This also addresses general steps to approach real case scenarios. For the future of ML-based medical devices, it will be important that a more consequent understanding is achieved how risk factors have to be incorporated in the development of ML-based systems.

5 Conclusion

In summary, this study provides a systematic analysis of the behavior of ML based systems with respect to differences in risk ratings, utilizing an artificial model. It demonstrates that standard metrics have substantial deficiencies when they are applied to ML systems without any adjustments towards the impact of different error types. The systematic integration of risks into the metrics is a crucial point to achieve an appropriate balancing of risk impact. Further steps are necessary to systematically integrate such approaches into the validation of ML based medical devices, in the future. In particular, this is related to the question how to obtain proper ratings for the risks and how to combine performance metrics with risk management requirements in a compliant way with respect to regulatory requirements.

References

1. Tharwat A. Classification assessment methods. Applied Computing and Informatics, 17(1) (2021) 168–192.
2. Drummond C, Holte R. Cost curves: An improved method for visualizing classifier performance. Machine Learning 65 (2006) 95–130.
3. Baker S, Cook N, Vickers A, Kramer B. Using relative utility curves to evaluate risk prediction. Journal of the Royal Statistical Society: Series A. 172(4) (2009) 729–748.
4. Rousson V, Zumbrunn T. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies. BMC Medical Informatics and Decision Making 11 (2011) 45.
5. International Organization of Standards. ISO 14971:2019: Medical devices – Application of risk management to medical devices. (2019).