An Exploration of Keeping Machine Learning Techniques Explainable in the Medical Setting

Abstract

Machine learning, especially advanced techniques, have been able to achieve high accuracy. However, the extent which the accuracy leads to practical use is being doubted. One trend in machine learning literature is then explainable machine learning (or explainable artificial intelligence). With the use of post-hoc analysis and discussion to facilitate understanding of the machine learning techniques, key features, weighted by importance, can be highlighted. In this paper, we discuss some of these efforts in the medical field and explore how we can achieve good explanations without reliance on the most complicated (and hardest to explain techniques).

I. Introduction

With modern computing power, machine learning produced astounding accuracy in predicting results even with datasets without predefined predictor and predicted variable relationships. However, the application of these machine learning algorithms (henceforth models) to real-world applications raises concerns on the incompleteness and soundness of the model. The reaction to these concerns is a greater call for models' explainability [1]–[5]. In this paper, we use the criteria for explainability from Belle and Papantonis [4]: comprehensibility, fidelity, accuracy, scalability, and generality. However, as [4] acknowledges, the criteria were more intuitive in nature rather than measurable. Our discussion below is divided into two main segments: (a) a literature review of the models and explainability achieved in studies published after 2020, and (b) our attempt to derive explainable models using a healthcare dataset.

II. Literature Review

[5] studied hospital mortality with four models (Random Forest [RF], Logistic Regression [LR], Adaptive Boost Classifier [ABC], and Naïve Baynes [NB]) which achieved good accuracy (highest AUC = 0.871 for RF). Using SHapley Additive exPlanations (SHAP) force plot and SHAP individual force plot, [5] provided good explainability with visualizations of feature relevance of all 4 models on a global scale and for two specific examples. **Pro.** Through the SHAP plots, [5] was a good reminder that sometimes models with lower accuracy (NB had the lowest AUC = 0.816) can be more medically sound in some factors than other models with higher accuracy. **Con.** [5] acknowledged that some medical factors required more information for meaningful medical interpretation. With more factors involved, deep learning techniques may be more suitable.

[6] also studied hospital mortality, but with with greater emphasis on how factors are changed over time. [6] used an artificial neural network with a long short-term memory (LSTM). Accuracy varied across the predicted time with AUC ranging from 0.50 to 0.88. [6] then did a SHAP analysis and visualized the feature importance across time with a cluster heat map. **Pro.** [6] is an example of how deep learning techniques can produce explainable prediction results for complicated scenarios like hospital mortality. **Con.** Although deep learning methods can cope with highly complex data, it is not practical to expect hourly data of all medical information [6].

[7] studies the risks of mothers getting Gestational Diabetes Mellitus with LR, support vector machines (SVM), adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost). [7] generated a number of models by using these ML techniques on different list of features. In the best list of features, all 5 models achieved a balance accuracy of around 0.75. [7] then explored that explainability of the models by using LR coefficient and Kenel SHAP scores (for other techniques) to identify global features. Two specific cases were also examined for local explainability. **Pro.** [7] explored models with less features, which though produced lower balance accuracy (varying around 0.55 to 0.61), was most practical to be used in clinical setting. **Con.** [7] acknowledged that even the best models in this paper had lesser features and lower accuracy that other models typically published.

[8] studied the risks of dementia with RF and XGBoosts. The models achieved high accuracies (RF AUC = 0.94, XGBoost AUC = 0.96). Global feature relevance was studied with SHAP scores and individual feature relevance was studied on 4 randomly chosen samples. wo models achieved high accuracy of dementia. **Con.** Although predicting dementia was done accurately, the explanations of feature relevance were inconsistent, limiting practical use. Further, [8] achieved lower accuracy (AUC = 0.51 and 0.63) when studying prevention factors, suggesting a better list of features can be found.

[9] studied emergent health risks to emergency departments with the onset of COVID-19. Using XGBoost, [9] achieved a high AUC of approximately 0.85 between Oct-2019 and Mar-2020. In later months, with COVID-19, AUC dropped to approximately 0.80. The global results were further interpreted with SHAP; [9] found that the features relevant to the results changed with onset of COVID-19. **Pro.** [9] demonstrated a drop in accuracy overtime can be used to call for further analysis, as a data drift may have happened, requiring a different operation plan for practitioners. **Con.** As a data drift has happened, earlier assumptions of the most important features may no longer be valid; hence, when data drifts happens, more features may need to be considered for a better understanding of the data.

[5]–[9] met [4] expectation SHAP analysis is very commonly used to explain feature importance in models. Yet, the assumption of independence of SHAP may have led to some scores being over-estimated [4].

III. Experimental Design

This study explores the negative outcomes of risks factors to prenatal mothers' mental health. Data was collected through a prenatal mental risk screening program, with a quantitative survey designed to categorise people into high and low risk.

The risk factors considered in this survey include common life experiences, psychological anguish, incidents of familial violence, smoking habits, drug use, and difficulties related to transitional periods, designed with the intent to consider the wide perspectives of risks.

A. Data Pre-Processing

The dataset was loaded via the pandas data frame to maintain the table structure. Preliminary descriptive and basic statistical analysis were performed to identify the data characteristics and rectification required. This helps to identify anomalies e.g., outliers or missing data. This will determine the work required to be performed in the initial Extract-Transform-Load (ETL) phase.

No missing data were noted. However, variable naming convention was detected to be inconsistent. Variables had spaces at the end of their names which could lead to errors in data preparation as exact match is required. Therefore, all spaces at the end of the variables names were stripped for formatting purposes.

After the data have been processed, K-Fold was adopted over the commonly adopted train-test split due to the limited 179 observations. This allows the model to train with the entire dataset as compared to imbalance training due to data split variances or unbalance data. Although the data is deemed to be balanced in this instance, adopting the best practice approach, K-Fold will be applied to ensure consistency among the model training as different models may require a different approach, e.g. ensemble learning. A recent study further suggested that k-fold should be used to have the reliable performance evaluation [10].

B. Predictive Model Construction

Three machine learning algorithms were considered for model selection: RF, SVM, and KNN. These algorithms were chosen based on their compatibility with the dataset characteristics and subject matter expertise. Other reasons includes:

1. RF

RF is robust and effective in classification tasks [11]. RF is well-suited for handling binary response variables and provides feature importance measures, which can aid in understanding influential features. The default construction of the RF model involves creating an ensemble of decision trees. We set the number of trees in the forest is set to 100, and defaulted other hyperparameters like the maximum depth of each tree and the number of features considered for splitting.

2. SVM

SVM was chosen for its strong performance with limited datasets and binary response variables. SVM excels in handling high-dimensional data and can capture non-linear relationships using kernel functions [12]. SVM is particularly effective in finding optimal hyperplanes that separate different classes. The SVM model construction utilizes the C-SVC implementation. The default construction value of the regularization parameter C that we use is set to 1, which controls the trade-off between achieving a low training error and allowing more margin violations. The kernel function is set to the radial basis function (RBF), which is suitable for capturing non-linear relationships. The default value of the gamma parameter is set to "auto," which determines the kernel coefficient based on the inverse of the number of features.

3. KNN

KNN, a non-parametric algorithm, was considered due to its suitability for limited datasets. It can effectively handle binary response variables and capture complex relationships without making assumptions about data distribution [13]. The default KNN model construction involves finding the k nearest neighbours in the feature space. The default value of k is set to 5, where the class label of most of the neighbours determines the prediction. Euclidean distance is used by default to compute the proximity between instances.

The next section compares the performance of these models using appropriate evaluation metrics to determine the most effective algorithm for the specific task.

IV. Experimental Results and Discussions

A. Correlation Matrix

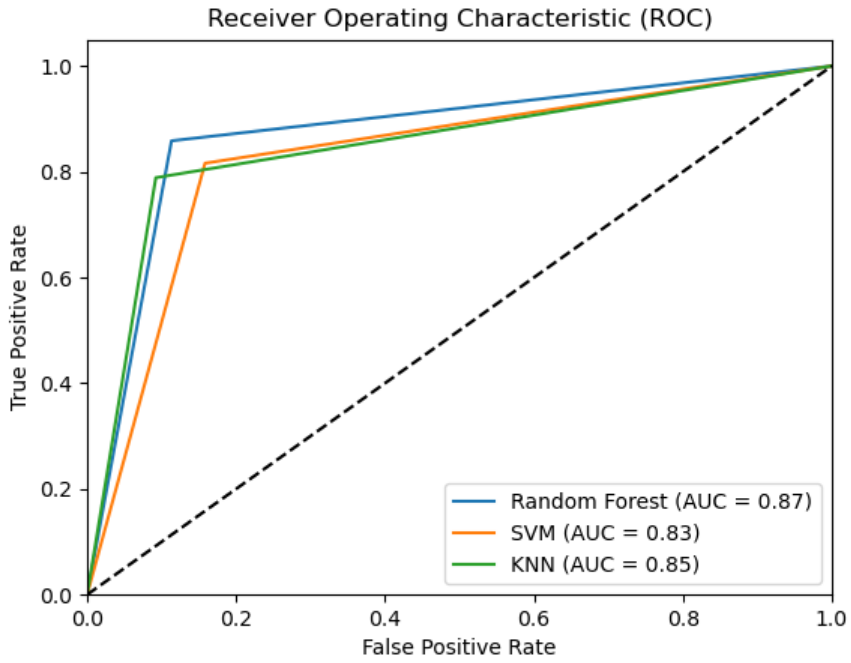
Table 1

Correlation Table	
	Class
Class	1.0000
Q227	0.6334
Q225	0.5375
Q231	0.4044
Q228	0.3792
Q214	0.3469
Q534	0.3270
Q457	0.2714
Q535	0.2650
Q233	0.2633
Q195	0.2608
Q164	0.2555
Q578	0.2445
Q139	0.2445
Q196	0.2429
Q536	0.2230
Q174	0.2218
Q454	0.2215
Q588	0.2187
Q137	0.2073
Q187	0.2068
Q467	0.2064
Q589	0.2057
Q197	0.2015

The correlation matrix (Table 1) reflects the respective variables' correlation to the dependent variable (Class). Q227 ranks the highest at 0.6334 (moderately strong positive correlation). Q225 ranks second at 0.5375 (moderate positive correlation). Notably Q231, Q228, Q214, Q534 have moderately weak positive correlation, while the rest of the variables are deemed to have a relatively weak positive correlation to Class.

B. ROC Graph

Figure 1



The ROC chart (Figure 1) indicates the respective models diagnostic capabilities of a binary classifier, the trade-off between True Positive and False Positive at various intervals. From the above chart, RF has the best performance (highest AUC score) at 0.87 as compared to KNN 0.85 and SVM 0.83. This indicates that the RF algorithm has the highest accuracy in predicting the total number of positive and true positive as compared to the other two algorithm in consideration.

C. Performance evaluation

Table 2

	RF	**SVM**	**KNN**
Accuracy	0.8601	0.8212	0.8324

AUC	0.8730	0.8288	0.8482
Precision	0.8788	0.8504	0.8910
Recall	0.8604	0.8160	0.7887
F1 Score	0.8619	0.8223	0.8269
ERR Score	0.1399	0.1788	0.1676

Notes:

Accuracy – Rate of correct prediction

AUC – Overall performance of a binary classification model

Precision – Ratio of Positive correctly predicted

Recall – Rate of True Positive correctly predicted

F1 Score – Overall performance of Precision and Recall

ERR Score – Misclassification rate

The respective models are evaluated by the above metrics with emphasis on:

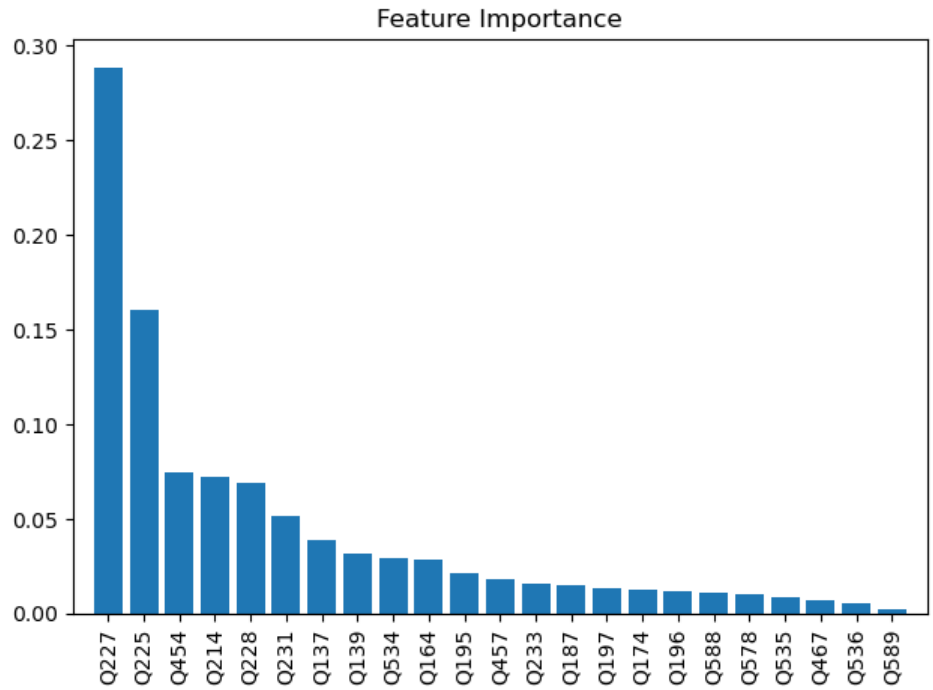
1. Accuracy
2. Recall
3. F1 Score

Measures of accuracies are presented in Table 2. The focus on accuracy and F1 Score is to ensure that the model selected will be able to accurately predict the mental well-being risk of perinatal mothers. This is crucial as it forms the underlying assumption and framework of the classification model.

We have elected to emphasized on recall as it enables the model to collect all positive examples successfully, allowing the model to successfully identify all cases as compared to potentially overlooking cases which could have serious implications. On the contrary, precision is better suited for instances where accurate identification will be required (e.g., surgical intervention).

D. Feature Importance

Figure 2

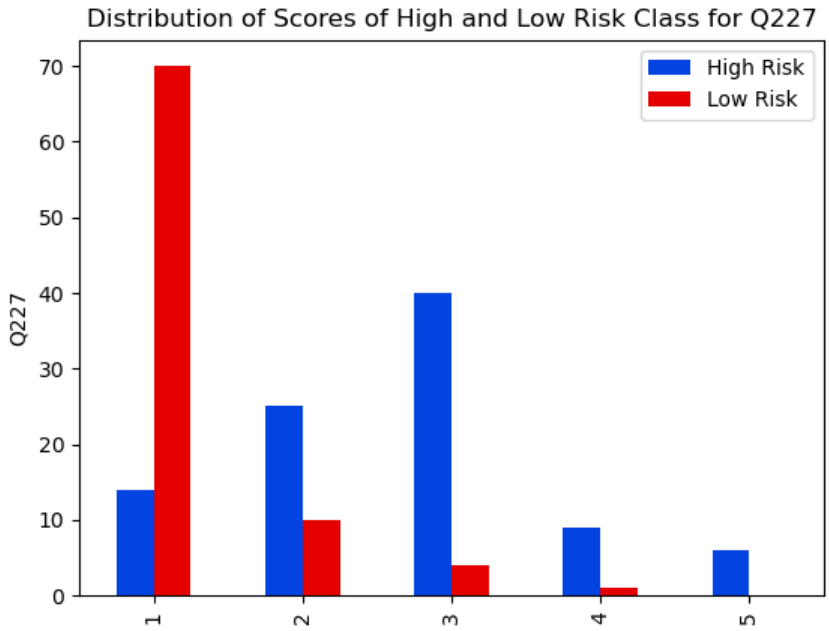


The feature importance (Figure 2) supports the correlation findings with most highly correlated features being reflected in the higher rank of feature importance. With Q227 approximately doubling Q225, and Q225 doubling of the rest of the features. However, there are various features that were weakly correlated (e.g. Q454) that stands out among the feature importance, ranking third.

In the discussion below, we will be delving deeper into the two prominent features specifically.

1) Q227: Individual Self-blame

Figure 3

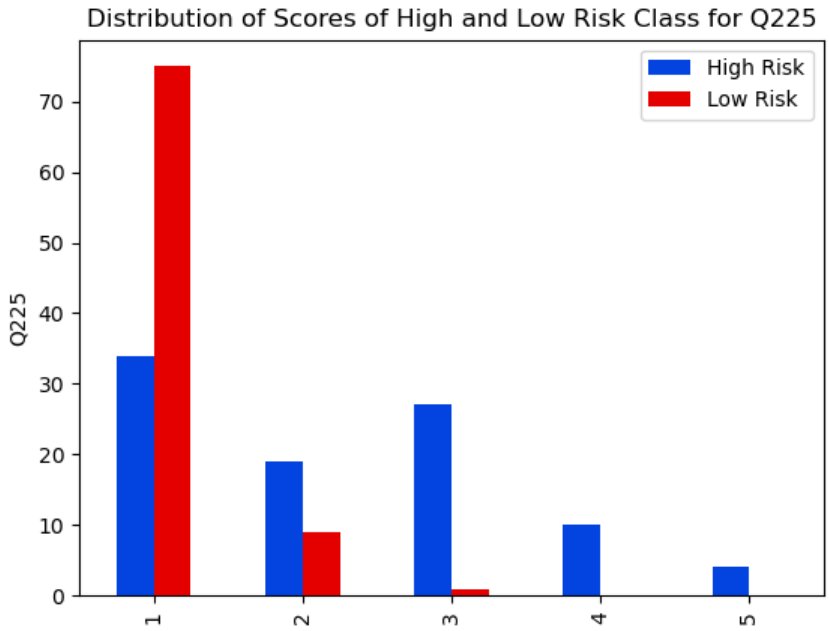


Q227 low-risk participants have a lower score distribution, congregating around the score of '1' (Figure 3). This indicates that participant with low-risk tend not to blame themselves as much. This suggests that individuals' determination of self-accountability contributes to better mental well-being.

However, this is normally distributed for those with high risk. This suggests that accountability blaming score could shift an individual into the high risk category in contrast to being on the other end of the spectrum. In addition, even a slight increase could lead an individual to be high risk as most of the high risk participants have answered with a score of '2' and '3'.

2) Q225: Feeling of Loneliness

Figure 4



Q225 score distribution for low-risk participants indicates that they are able to integrate into their social circle (Figure 4). However, participants with high risk are once again observed to be spread among the scores similar to Q227.

V. Conclusion

A. Summary

In conclusion, the RF would form the best model with an Accuracy = 0.8601, Recall = 0.8604 and F1 score = 0.8619 (Table 2). The approach of the model focuses on Recall as the dataset is centred around a medical study. Therefore, emphasis was placed on diagnosing a catch-all to prevent missing out on high risk patients. In this context, it would be a preventive approach as compared to intervention which would require a higher degree of true positive diagnosis to prevent wrongful medical practice.

The key factors found are Q227 and Q225. Both reflect individual's inner working. Low-risk participants would generally congregate around the lower scores while high risk individuals are spread out with a generic congregation around the middle scores. Therefore, our study suggests that the slightest increase in score of these two features could push an individual to the other risk class instead of possessing a score in the other spectrum.

B. Limitation

The dataset provided consists of only 179 observations and there may be underrepresentation of certain population groups. With limited information on data collection, we had to assume there were no data collection issues such as Hawthorne effects, groupthink, and the influence of social desirability.

In addition, RF of interest as RF operates similar to a black box methodology. This makes it difficult to identify the process and logic behind the model's prediction for full transparency analysis.

This study did not use post-hoc explainability analysis like SHAP due to the small number of features and lack of time as a feature. In a more complex study to reflect how risks may change over time, more complicated machine learning models and inclusion of post-hoc analysis may be required.

C. Future Improvement

For future studies, we recommend:

- Increasing the data size. This would increase accuracy and detect other possible interesting findings that might not have been discovered in the initial dataset provided.

- Using Likert Scale responses instead of binary responses for questions where various insights might be significant. This could potentially reveal stages which are of essence to the dependent variable.

- Variables could be assigned corresponding weights to tune the model (e.g. features selection and sample weighting).

- The research team to include experts for better model tuning.

References

[1] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021, doi: 10.1613/jair.1.12228.

[2] A. F. Cooper, E. Moss, B. Laufer, and H. Nissenbaum, "Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, IEEE, 2022, pp. 864–876. doi: 10.1145/3531146.3533150.

[3] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42200–42216, 2020, doi: 10.1109/ACCESS.2020.2976199.

[4] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Front. Big Data*, vol. 4, p. 688969, 2021, doi: 10.3389/fdata.2021.688969.

[5] E. Stenwig, G. Salvi, P. S. Rossi, and N. K. Skjærvold, "Comparative analysis of explainable machine learning prediction models for hospital mortality," *BMC Med. Res. Methodol.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12874-022-01540-w.

[6] H.-C. Thorsen-Meyer *et al.*, "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records," *Lancet Digit. Health*, vol. 2, no. 4, pp. e179–e191, 2020, doi: 10.1016/S2589-7500(20)30018-2.

[7] Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," *Sci. Rep.*, vol. 12, no. 1, p. 1170, 2022, doi: 10.1038/s41598-022-05112-2.

[8] S. O. Danso, Z. Zeng, G. Muniz-Terrera, and C. W. Ritchie, "Developing an explainable machine learning-based personalised dementia risk prediction model: a transfer learning approach with ensemble learning algorithms," *Front. Big Data*, vol. 4, p. 613047, 2021, doi: 10.3389/fdata.2021.613047.

[9] C. Duckworth *et al.*, "Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19," *Sci. Rep.*, vol. 11, no. 1, p. 23017, 2021, doi: 10.1038/s41598-021-02481-y.

[10] T.-T. Wong and P.-Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.

[11] Y. Zhai and X. Zheng, "Random Forest based Traffic Classification Method In SDN," in *2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB)*, Nov. 2018, pp. 1–5. doi: 10.1109/ICCBB.2018.8756496.

[12] S. jie and H. Wankun, "Experimental Results of Maritime Target Detection Based on SVM Classifier," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, Sep. 2020, pp. 179–182. doi: 10.1109/ICICSP50920.2020.9232038.

[13] H. Xie, D. Liang, Z. Zhang, H. Jin, C. Lu, and Y. Lin, "A Novel Pre-Classification Based kNN Algorithm," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 1269–1275. doi: 10.1109/ICDMW.2016.0182.