Stacked Generalization with an Explainable AI Base Learner in Healthcare

Aubrey Moulton, Abdullah Aman Tutul, Edmund Do
Department of Computer Science and Engineering, Texas A&M University
{amoulton, abdullahaman633, esd}@tamu.edu

Abstract

Artificial intelligence integrated into the healthcare field has a great potential to improve patient outcomes and be a powerful tool to assist doctors. The adoption of artificial intelligence into healthcare is reliant on the interpretability of the results by the doctor and the accuracy of the model. Our goal is to predict the mortality of patients in the hospital intensive care unit using stacked generalization on an explainable artificial intelligence model. Our stacked generalization model has two layers, a base learner and a meta learner. The base learner is an explainable boosting machine that scores features. The meta learner is machine learning classifiers like random forest and support vector machine. The feature scores are input into the meta learner to improve accuracy of classification. The results are a performance improvement of up to 0.148 F1, relative to a standalone explainable boosting machine. Our novel stacked generalization method is generalizable to any healthcare setting with sufficient data in the electronic health records.

Introduction

The healthcare field is ready to adopt artificial intelligence models to more efficiently analyze patient data, but there is no room for error when a patient's life is on the line. Healthcare institutions are increasingly collecting vast amounts of data on patients. Electronic health records have been widely adopted and integrate care across providers. All the patients' data, from vitals, lab results, medical imaging, and every doctor's note is uploaded to this record. This is an extensive amount of data for doctors to analyze and formulate a diagnosis. There is a great potential for machine learning models to assist doctors in their critical work. The machine learning model will be able to more quickly predict patient outcomes and alert the doctors of important information to improve patient care.

Explainable artificial intelligence (XAI) is highly adaptable to the healthcare setting and is in the early stages of being deployed in hospital settings today. The explainable nature of the results enables doctors to practically implement insights and information in care delivery.

Doctors can use explainable AI as a critical tool to analyze large amounts of data, then implement the model insights into their medical care of patients. Explainable AI provides medical professionals with a powerful tool that is complemented by human judgment and professional medical expertise. However there is a significant tradeoff for the explainability of the model, the results are less accurate.

In this paper, we aim to improve the accuracy of the XAI model's prediction by adding stacked generalization. In this study we analyze patient data from a hospital intensive care unit (ICU). Our goal is to predict the patient's mortality while in the ICU. Stacked generalization has two layers. The first layer, the base learner, is the EBM. Training data is inputted into the EBM and the output is feature importance scores for every feature. The second layer, the meta learner, is a machine learning classifier. The feature importance scores are used to train the meta learner. Our hypothesis is that the meta learner will be more expressive if we use feature importance values from EBM to train the meta learner. These results are significant because the results will be both explainable and more accurate. This proposed pipeline will be an excellent tool for doctors to understand the results and also have higher accuracy.

Stacked generalization has shown promising results for clinical outcome prediction. There have been few studies that applied stacked generalization to predict breast cancer survival and malaria severity that are discussed in the literature review. Traditionally, stacked generalization models use the probabilities of the first layer as input to the second layer. In comparison, we are using feature importance scores from EBM to input into the second layer. It is novel to use the XAI as a feature extractor to input into another classifier. Also, there are no studies on using an explainable AI model in the first layer. One existing study used EBM in the second layer, and a logistic regression to define the important features in the first layer. Our innovation is using stacked generalization with an explainable AI model to extract feature importance scores as input for a second classification.

The remainder of this paper is structured as follows, we start with an overview of previous work on explainable artificial intelligence and its applications to healthcare. Next, we describe the method of the two layer model we took in our study. Then, we present the results of the base learner EBM, paired with multiple classifiers. The article concludes with a discussion of results and directions for future work.

Literature Review

The first explainable machine learning model was the Generalized Additive Model (GAM) published by Hastie & Tibshirani (1990) [1]. GAMs blend the properties of additive models and generalized linear models. The model depicts the dependent variable as a sum of univariate models. GAM has the following form,

$$g(E[y]) = \beta_0 + f_j(x_j)$$

The variable g is the link function which is adapted based on the particular problem (classification/regression) and $f_j$is the contribution of feature x_j towards the model prediction. G is a logistic function for our classification task. However, GAM does not permit interactions between features and the performance of GAM is significantly less than other state of the art full complexity machine learning models.

Yin & Rich et al. (2013) [2] proposed adding pairwise feature interactions to the standard GAM model, introducing the updated model as Generalized Additive Models plus Interaction or GA2M. They developed a computationally efficient method named 'FAST' that ranks all possible pairs of features as candidates for the model. The researchers used 10 publicly available datasets (5 regression problems and 5 binary classification problems) and they showed that on average GA2M reduces the RMSE by 34% on the regression problems and it reduces the 0/1 loss by 44% on the classification problems compared to the baseline GAM models.

The explainable boosting machine (EBM) is a faster implementation of GA2M by Nori & Jenkins et al., 2019 [3] . The EBM is considered explainable because of the scores that show the contribution of each feature towards a final prediction in a way that is easily understood by humans. EBM is a generalized additive model of the following form;

$$g(E[y]) = \beta_0 + f_j(x_j) + f_{ij}(x_i, x_j)$$

In the equation, β_0 represents the algorithm specific bias term, f_j and f_{ij} depicts the contribution of feature x_j and pairwise features x_i, x_j towards the model prediction. The variable g is a link function which is an identity function for regression and $g(E[y])$ is assigned as a positive class if $g(E[y]) > 0$, otherwise, it's assigned as a negative class for binary classification. EBM learns the feature functions using bagging and boosting. The model is trained on one feature at a time in a round robin fashion with a very low learning rate so that the feature order does not impact the output of the model. The use of round robin training mitigates the collinearity issue of the features and learns the best feature function f_j for each feature. It provides the state of the art accuracy comparable to random forest and boosted trees along with providing feature explanations in terms of individual and pairwise feature contributions toward the output of the model.

Jiangdong Zhou and Gary et al. (2020) [4] conducted a study to predict patient admission into the hospital ICU using the EBM model. The dataset included 1043 patients, and 19 of them were admitted to ICU. The researchers used a specific version of stacked generalization. The first step was a univariate logistic regression to find the most important features for predicting the likelihood of ICU admission. The second step was to use those features in the EBM model. The results are that the EBM (F1 Score 91.89%, AUC 92.31%) outperforms XGBoost (F1 Score 90.25%, AUC 89.82%), LightGBM (F1 Score 84.18%, AUC 80.40%), Random forest (F1 Score 82.01%, AUC 82.5%), and logistic regression (F1 Score 81.86%, AUC 83.10%). In addition to providing state of the art accuracy, EBM provides explanations for individuals predictions based on tree-based decision systems, which is very much desired for clinical decision making systems. The EBM explained that red blood cells, APTT, sex, age, and white blood cells are the most important risk factors for ICU admission. This study also illustrates that EBM can handle the class imbalance problem. This relates to our research because our dataset is also class

imbalanced. Our predictive variable is the likelihood of mortality for the ICU patients, and there are many more patients that survived than patients that died in the ICU.

Gamal and Sallam et al. (2017) [5] used a stacked generalization method with feature selection to predict the survival of breast cancer patients in the METABRIC dataset. The model had two layers. The researchers used multiple classifiers in the first layer to predict the outcome probabilities. The probabilities were inputted into the second layer model, called the meta-learner. Multiple different feature selection methods were used on the first layer features including backward feature selection, and information gain based feature selection. The classifiers used as meta-learners in the second layer were Support Vector Machine, Gradient Boosting machine, and Random Forest. The researchers achieved the highest accuracy of 80.92% using a Support Vector Machine (SVM) as the meta learner and using information gain based forward feature selection method to select the features of meta learner. They outperformed the best accuracy on this dataset by Chen et al(2013) [6]. This approach is similar to our research because we also used a two layer model. In our approach, we used EBM as the first layer classifier and explored different classifiers like SVM, multilayer perceptron as the meta-learner. The researchers also demonstrated that their feature selection methods chose accurate features that agree with breast cancer experts. In our research, we also consulted with a medical expert whose opinion aligned with the most important features selected by the EBM.

Oguntimilehin and Adetunmbi (2017) [7] used a stacked generalization method to predict the severity of the patients with malaria fever. The data was collected from Adetoyin Hospital, Ado-Ekiti , Nigeria and Afe Babalola University Health Centre, Ado-Ekiti, Nigeria for a period of six months. They used six base-learners in the first layer including PART, REP Tree, J48, Random Tree, RIDOR and JRIP. They used the class probabilities from these base learners, combined them with different arrangement order and used different meta learners (Random Forest and NNGE, Non-Nested Generalized Exemplars) in the second layer. They used 1225 samples as the training set and 408 samples as the test set. They found that each of the meta learners outperformed all the individual base learner accuracy and their results show that the use of NNGE (100% Accuracy) as the meta-learner outperforms the random forest (Accuracy 98.0392%) as the meta learner. Their results demonstrate that the selection of meta learners affects the performance of the model.

Methods

The objective of the work is to improve the performance on the EBM to analyze the tradeoff between explainability and performance. To quantify this tradeoff, we utilize stacked generalization with the EBM classifier as the main classifier and four meta learners: decision tree, random forest, SVM (with a radial basis function kernel), and a multilayer perceptron (MLP).

Dataset

This research was done with the ‘In Hospital Mortality Prediction’ dataset available from Kaggle. This dataset is a subsample from the Medical Information Mart for Intensive Care

(MIMIC-III). MIMIC-III is a large database of forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The patient's confidentiality is maintained. The data is deidentified according to Health Insurance Portability and Accountability Act (HIPAA) standards. Identifying elements such as name, phone number, and address are deleted. Dates of stays were shifted to obscure identifiers. This data set was compiled for researchers and academics to be able to research in healthcare and have reproducible results.

The 'In Hospital Mortality Prediction' dataset includes the data of 1,177 patients collected during their treatment in the ICU. There are 49 feature attributes representing risk factors that lead to mortality and the results of some preliminary and advanced medical tests. There are 11 binary features that indicate whether a patient has a particular condition (e.g. chronic kidney disease). The remaining 38 features are continuous numerical values (e.g. mean blood pressure). The feature we are predicting is mortality, and it is a categorical binary variable indicating whether or not a patient died during treatment. Mortality is coded as 1 if the patient died while in the hospital, if the patient remains alive the value is 0. In this dataset, 1017 of 1177 patients survived their experience in the ICU while 159 died. The imbalance between the classes is a problem that we address in our research methodology.

Stacked Generalization

The explainable boosting machine (EBM) is an explainable additive linear model that finds and uses the contributions of each feature $f_j(x_j)$ to predict the outcome. This being a linear additive model allows for better interpretability of each feature with respect to the outcome. However, it has been observed in the literature that interpretable models often perform worse than models that are considered less interpretable such as SVMs and MLPs [8].

The intuition behind the meta learners is to learn a non-linear link function with the feature importance scores in order to demonstrate the tradeoff between explainability and performance. The feature importance scores are used in one of two ways to train one of the meta learners: feature importance scores combined with the original features and feature importance scores only. The performance of these meta learners is compared to the performance of the EBM trained on the original data.

Pipelines

Pipeline 1: Three-way split

The data set is split into three parts with 60% reserved for training the EBM (A1), 30% for training the meta learner (A2), and 10% for testing (A3). Hyperparameters for the EBM are determined using a grid search with cross validation which remains consistent for each repeated run. After the EBM is trained on A1, the feature importance scores are extracted for each sample in A2. As mentioned above, these scores are either combined with the original features or used standalone to train a meta learner. Similar to the EBM, we use a grid search with cross validation

to determine hyperparameters for the meta learner that remain consistent over all trials. After the meta learner is trained, we evaluate the EBM and the meta learners on A3. This process is repeated 20 times to gauge the average F1 score of the EBM and the meta learners. A t-test is performed to determine whether the EBM or the meta learners perform better.

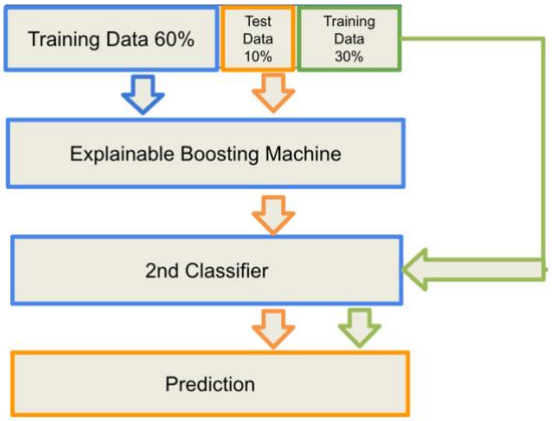

Figure 1: Stacked generalization pipeline for three ways split of data

Pipeline 2: Two-way split

We also use a two part split, 80% for training (B1) and 20% for testing (B2) the EBM and meta learners due to the concern that A2 may be too small for training the meta learner (MLP in particular) in the first pipeline. This pipeline differs in that the meta learners are also trained on features or feature importance scores from B1. Otherwise, the pipeline is unchanged.

Finally, in order to determine the effect of imbalanced data on the performance of the EBM and the meta learners, we take the best performing pipeline and randomly oversampled the minority class in the training set to determine whether the performance of the EBM was limited by class imbalance.

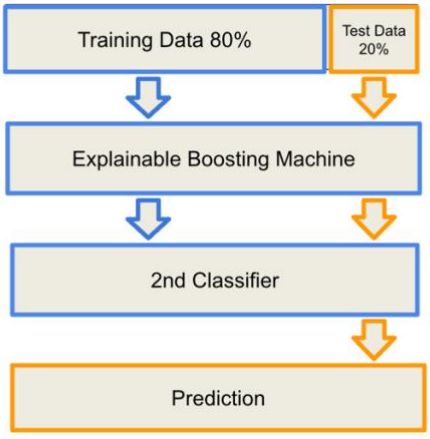

Figure 2: Stacked generalization pipeline for two ways split of data

EBM Explanation

EBM provides the most important features for the prediction of the model. We have checked that the features depicted as important by EBM are also considered as the most relevant risk factors for mortality prediction by medical experts . We built an EBM model based on the 75% dataset with learning rate of 0.01, early stopping rounds of 100 and maximum round of 1000 and we get the following overall feature importance graph. This model achieves 43% accuracy on the 25% test set which was split based on stratified sampling.

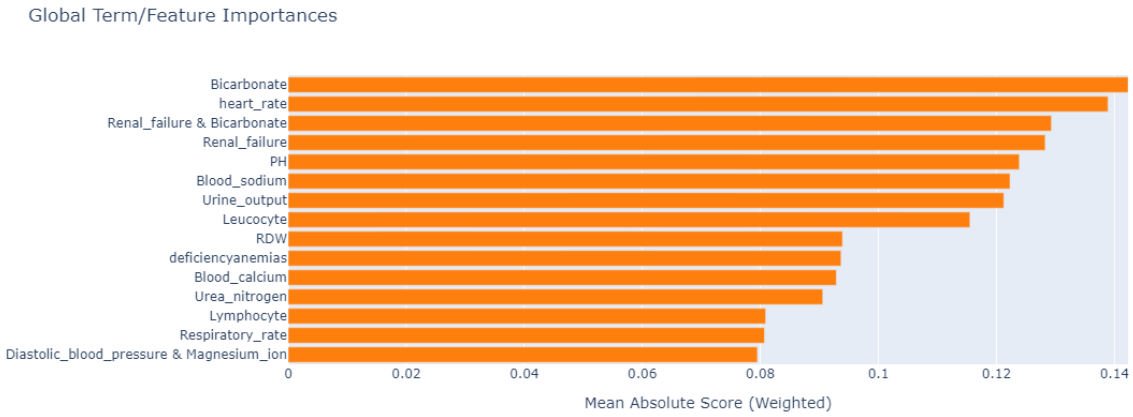

Figure 3: The overall feature importance graph provided by EBM. The features are given in sorted order of feature importance. From this figure, we can see that Bicarbonate is the most important feature for predicting the AI outcome followed by heart rate, pairwise interaction between Renal failure and Bicarbonate etc.

EBM also provides explanations for individual predictions in terms of contribution of each feature towards the model outcome. A sample local explanation graph on a test sample using this EBM model is shown in the following Figure.

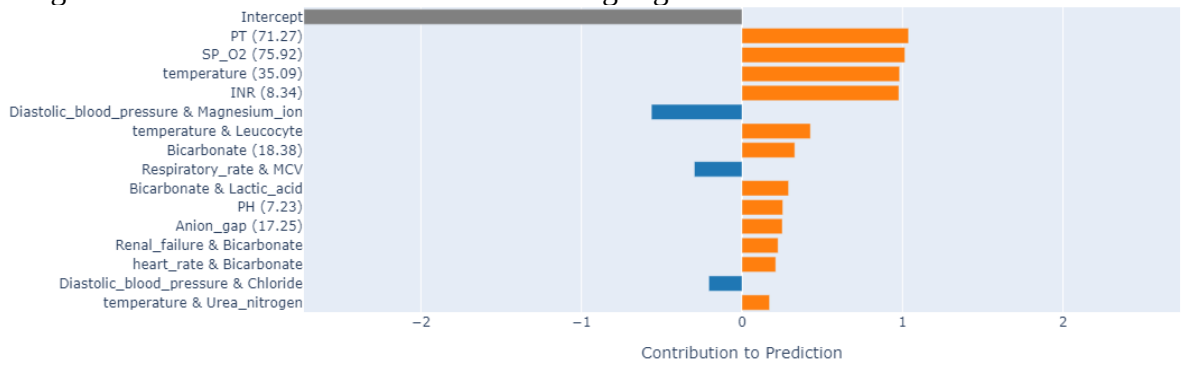

Figure 4: Local explanation of an individual test sample. Here, features are given in terms of sorted order. So, PT is the most important feature for the model prediction for this particular sample followed by SP_O2, temperature etc. The x axis represents the feature contribution scores and the positive contribution depicts positive association between the particular feature and the mortality rate and vice versa. So, we can say that for this particular sample, PT, SP_O2, temperature etc. positively contributed to the mortality rate whereas the INR, pairwise interaction between Respiratory rate and MCV negatively contributes to the mortality rate. The higher the absolute feature importance score, the higher the impact on the AI prediction. We used these feature importance scores as the features for our meta learner.

EBM shows the overall association between each feature and the outcome in terms of global explanation graphs. The global explanation graph based on the same EBM model for BMI is shown in the following figure.

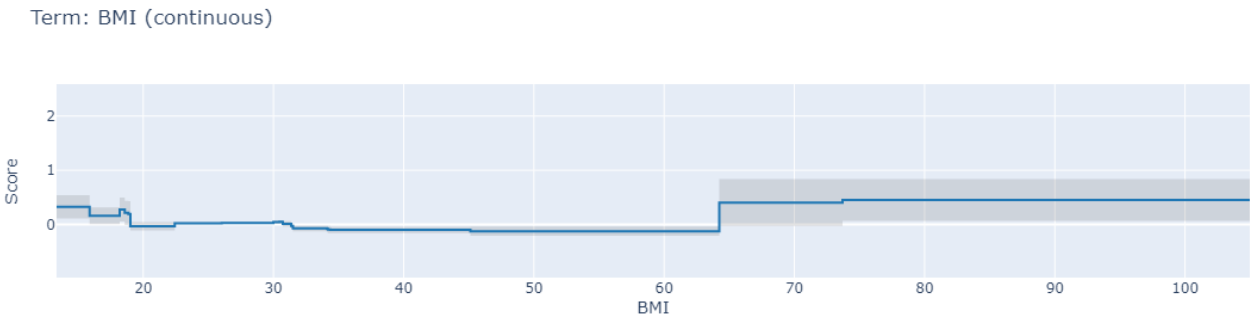

Figure 5: Global explanation graph for BMI feature. X-axis and y-axis depict the BMI feature value and the contribution score of BMI. Positive contribution scores depict positive association between BMI and the outcome (mortality rate), and vice versa. The higher the contribution score the higher the association between BMI and mortality rate. This figure demonstrates that people with high BMI have a higher mortality rate.

Results

From Figure 4 and Table 1, the SVM trained on the feature importance scores is the only meta learner that had a statistically significant performance improvement from the EBM with a higher F1 score on average by 0.148 ($p<0.001$) for three way split. Therefore, the SVM was able to learn a non-linear link function that performed better than the EBM. However, we observe that there is either no difference or worse performance between the performance of the EBM and the Decision Tree, Random Forest, and MLP meta learners trained on the features, features with importance scores, and only the feature importance scores. Across all meta learners, the addition of feature importance scores to the original data did not improve the performance of the meta learner relative to using only the features.

The results for the two part split are similar. From Figure 5 and Table 2, we see that the MLP benefitted from the additional training data in that the MLP meta learner trained on the feature importance scores had a performance improvement of 0.075 ($p=0.002$) relative to the EBM on average. Meanwhile, the difference in performance between the EBM and SVM has diminished to 0.107, but it remains a statistically significant difference ($p<0.001$). This confirms that the size of the training data for the meta learners was small for the MLP. In this pipeline with the two part split, both the SVM and MLP learned a non-linear link function that performed better than the EBM.

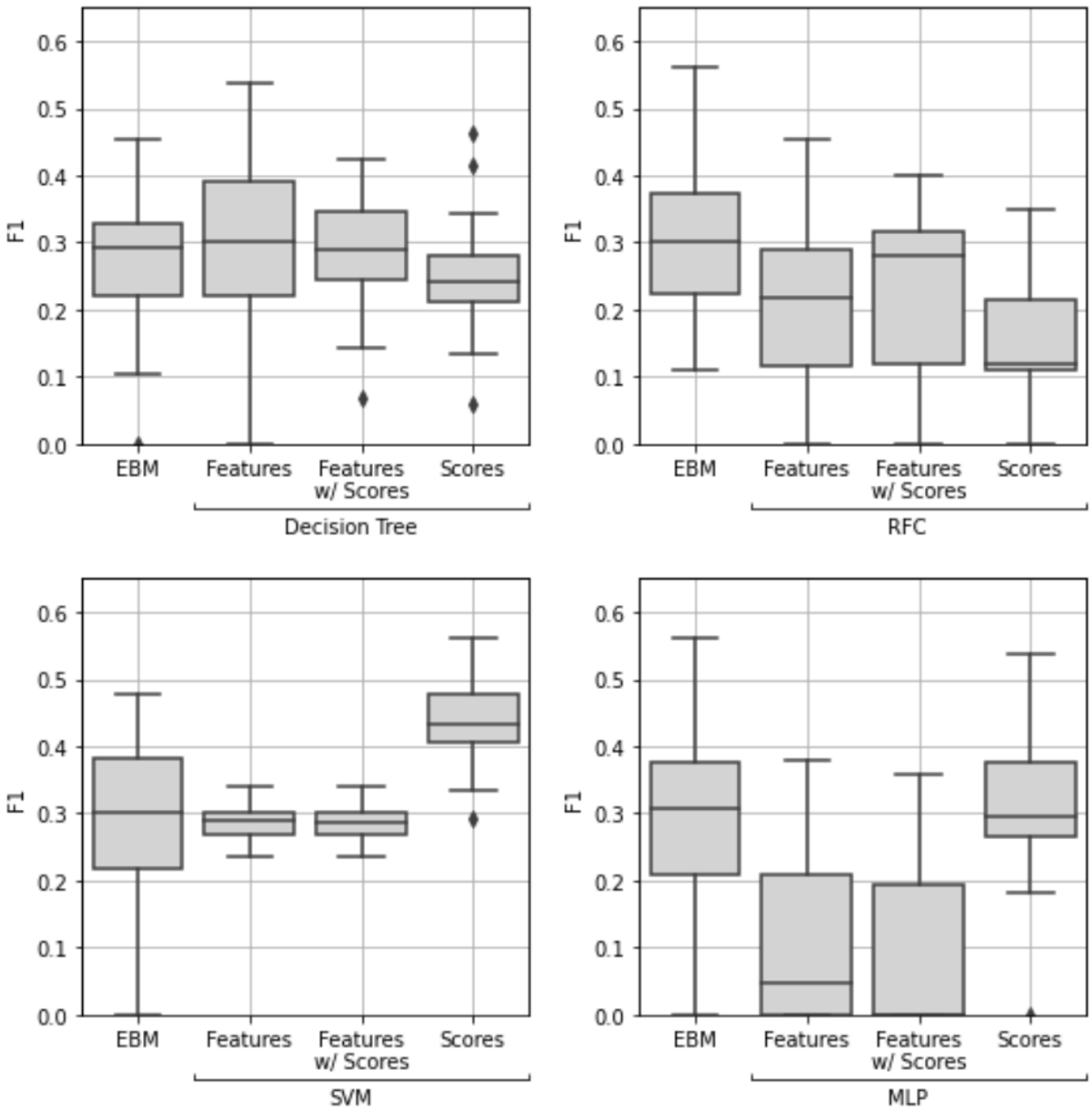

Figure 6: Distribution of F1 scores for EBM and meta learners over 20 trials using a three-part split

Table 1: T-statistic and p-values of pairwise T-tests between distribution of F1 scores of the EBM and each meta learner for a three-part split.

Meta Learner	Features	Features and Scores	Scores
Decision Tree	-0.336 (0.738)	0.613 (0.544)	0.598 (0.554)
Random Forest	2.422 (0.020*)	2.040 (0.048*)	4.526 (<0.001***)
SVM	0.054 (0.957)	0.072 (0.943)	-4.735 (<0.001***)
MLP	4.491 (<0.001***)	4.602 (<0.001***)	-0.333 (0.741)

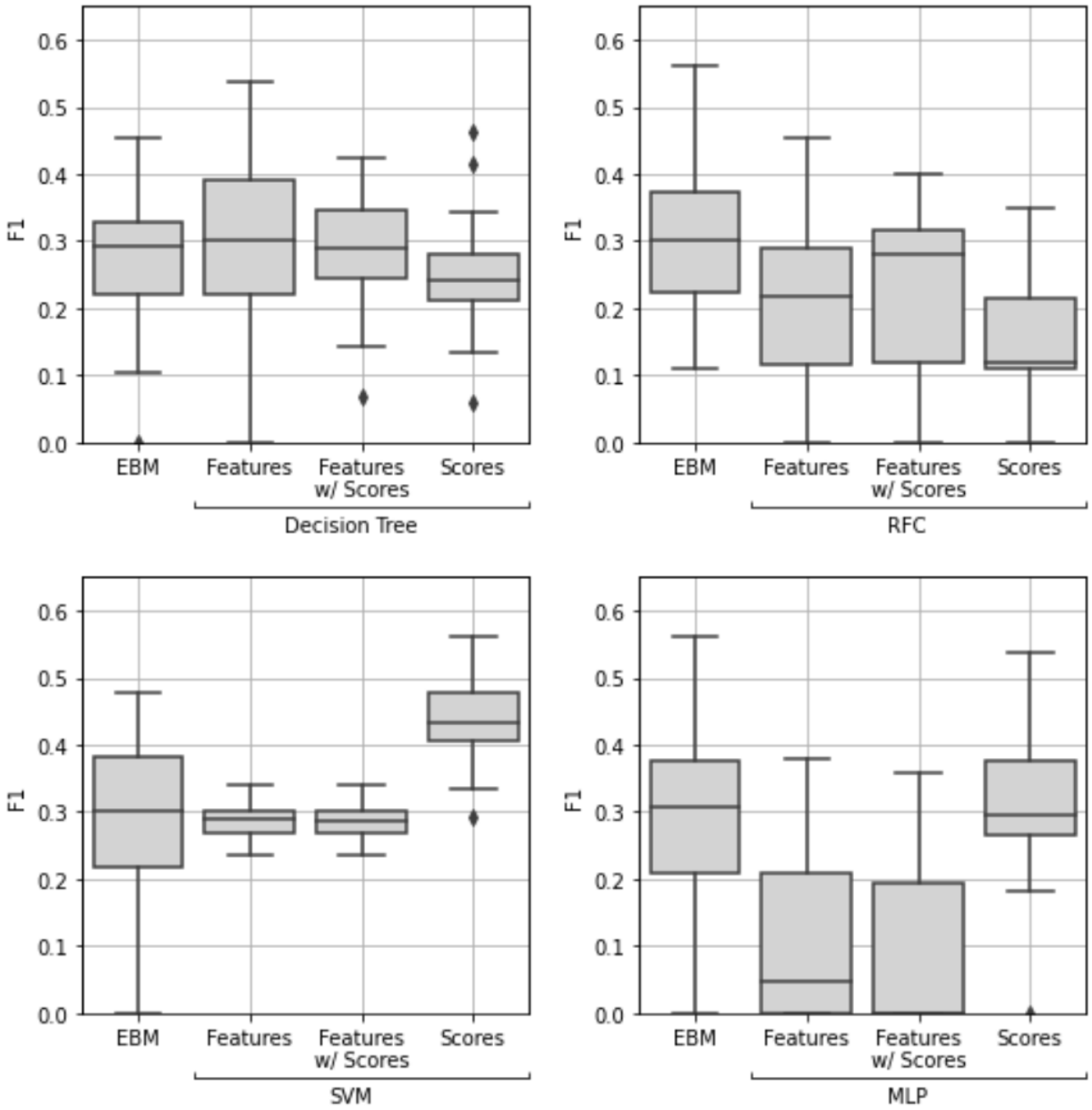

Figure 7: Distribution of F1 scores for EBM and meta learners over 20 trials using a two-part split

Table 2: T-statistic and p-values of pairwise T-tests between distribution of F1 scores of the EBM and each meta learner for a two-part split.

Meta Learner	Features	Features and Scores	Scores
Decision Tree	3.162 (0.003**)	1.412 (0.166)	0.918 (0.364)
Random Forest	4.850 (<0.001***)	2.635 (0.012*)	1.929 (0.061)
SVM	1.574 (0.124)	1.565 (0.126)	-4.958 (<0.001***)
MLP	6.033 (<0.001***)	4.770 (<0.001***)	-3.393 (0.002**)

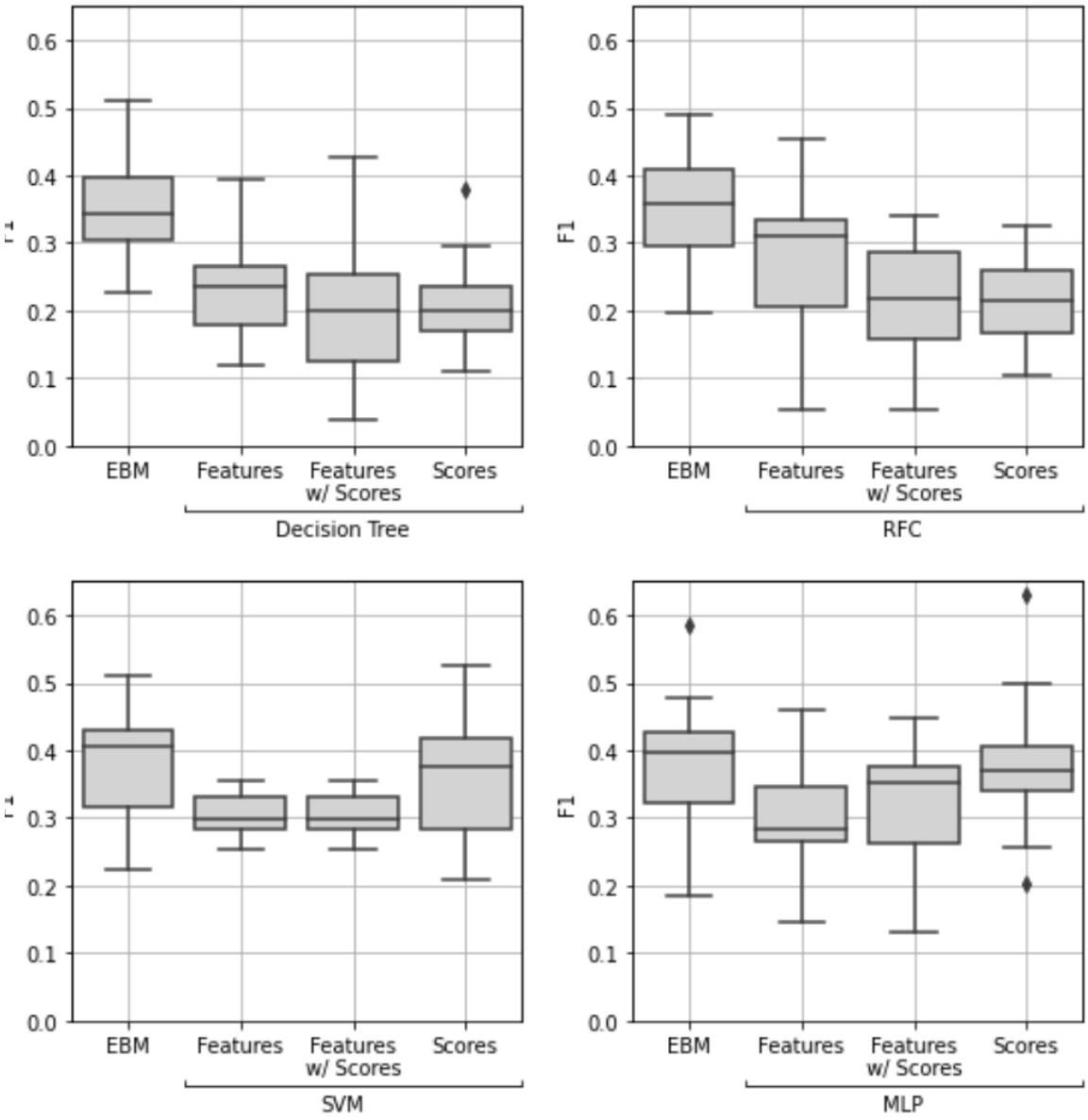

Figure 8: Distribution of F1 scores for EBM and meta learners over 20 trials using a two-part split with oversampling of the minority class in the training set

Table 3: T-statistic and p-values of pairwise T-tests between distribution of F1 scores of the EBM and each meta learner for a two part split with minority class oversampling in the training set.

Meta Learner	Features	Features and Scores	Scores
Decision Tree	5.727 (<0.001***)	5.395 (<0.001***)	6.489 (<0.001***)
Random Forest	2.207 (0.033*)	4.738 (<0.001***)	5.493 (<0.001***)
SVM	4.167 (<0.001***)	4.154 (<0.001***)	0.979 (0.334)
MLP	2.780 (0.008**)	1.988 (0.054)	0.113 (0.911)

Oversampling of the minority class benefited the performance of the standalone EBM. In our oversampling strategy, we oversampled the positive class and made equal number of positive and negative class samples in our training data. From Figure 6 and Table 3, the performance of the SVM and MLP meta learners were on par with the EBM in the oversampling case ($p=0.334$ and $p=0.911$, respectively). This may indicate that a linear link function is sufficient for classifying the data in this case; however, oversampling significantly increased the time to train the EBM. For more complex base models, it would be more efficient to use a meta learner rather than the base model for the best performance.

Discussion

Our first major finding is that the performance of the stacked generalized is highly dependent on the classifier used for the meta learner. We get statistically significant improvement over F1 score when we use SVM as the meta learner compared to using random forest, decision tree, MLP as the meta learner (Fig 4-5, Table 1-2). The MLP result using MLP as the meta learner is not statistically significant when we use three way splits, however, this result is statistically significant when we use two way splits. We do not get any statistically significant results for SVM and MLP as the meta learner classifier for two way splits with oversampling. The training time significantly increases when we use over sampling. The features depicted as important by the EBM model are also considered as important by Domain experts.

Previous studies (Gamal and Sallam et al., 2017, [5]) with stacked generalization and clinical outcome prediction (mortality rate of Breast cancer patient) also found significant improvement using SVM as the meta learner similar to what we found. We did not find any significant improvement using the decision tree and random forest as the meta learner. Part of the reason is that EBM itself learns its decision function through an ensemble of trees for each feature. To get an improvement using a meta learner, the learning function of meta learner should be different in our case compared to the EBM base learner; otherwise, the meta learner will not be able to complement the base learner. We speculate this may be a reason why we might not be getting a significant improvement using a decision tree or random forest as a meta learner over the EBM as a base learner. We found that the MLP provides significant results as a meta learner for two way splits but it does not produce a significant result for three way splits (Table 1 and 2). This indicates that there is not much data to train MLP in a three way split since MLP has a large number of parameters to be trained compared to other models so it needs a large dataset to be well trained. We found that the MLP outperforms the baseline EBM in a two way split. This is reasonable as MLP learns complex decision functions different from the tree based decision function learnt by EBM.

We also found that we get a statistically insignificant difference between the EBM base learner and MLP and SVM as a meta learner when using over sampling. Part of the reason might be that over-sampled positive samples are adding too much redundant information in the learning algorithm of SVM and MLP. Another interesting finding is that adding only the feature scores improves the performance of the meta learner more than adding both features and feature scores

in the meta learner. A potential reason behind this might be due to collinearity issues due to redundant features from adding both features and feature scores.

A limitation of this work is that the EBM is only one type of explainable model. The results here may not apply to other explainable models since other models may be better suited for certain tasks or datasets. Došilović et al. (2018) mention several explainable methods that have achieved comparable results to opaque models in certain problem domains. Additionally, inherently explainable models may not be necessary since post-hoc methods can be applied to opaque models to provide explainability. It avoids the explainability-performance tradeoff entirely since explainability is derived from the opaque model through the extraction of interpretable parameters or approximation of feature contributions (e.g. SHAP scores). As these opaque models potentially have high performance, the use of stacked generalization on these extracted features may yield less improvement than demonstrated in this work.

In conclusion, we were able to improve the accuracy of the XAI model's prediction with stacked generalization depending on which model was used as the meta learner. It is necessary to use explainable artificial intelligence models in healthcare because the results need to interpret and verified by the domain experts. Our model was trained and built on data from patients in the hospital intensive care unit, but it is generalizable to any setting if there is enough data. Electronic health records are a rich source of data to draw upon and there are many healthcare challenges that can benefit from artificial intelligence such as predicting disease, medical imaging identification, and analysis of signals from wearable health monitoring devices. We have exemplified with our study how AI is expanding into areas that were previously thought to be only the province of human experts.

References

[1]Hastie, T. J.; Tibshirani, R. J. (1990). Generalized Additive Models. Chapman & Hall/CRC. ISBN 978-0-412-34390-2.

[2] “Accurate Intelligible Models with Pairwise Interactions”, Yin Lou, Rich Caruana, Johannes Gehrke, Giles Hooker, KDD’13: Proceedings of the 19th ACM SIGKD international conference on Knowledge discovery and data mining, dio: https://doi.org/10.1145/2487575.2487579

[3] “InterpretML: A Unified Framework for Machine Learning Interpretability”, Harsha Nori , Samuel Jenkins, Paul Koch, and Rich Caruana , Microsoft Research, arXiv: 1909.09223v1

[4] “Identifying main and interaction effects of risk factors to predict intensive care admission in patients hospitalized with COVID-19: a retrospective cohort study in Hong Kong”, Jiandong Zhou , Gary Tse, Sharen Lee, Tong Liu, William KK Wu, Zhidong Cao, Daniel Dajun Zeng, Ian Chi Kei Wong, Qingpeng Zhang, and Bernard Man Yung Cheung, medRxiv, doi: https://doi.org/10.1101/2020.06.30.20143651

[5]”A Stacked Generalization Method for Disease Progression Prediction”,Gamal Elkomy, ElSayed Sallam, Sherin Elgokhy, 2017 13th International Computer Engineering Conference (ICENCO), doi:10.1109/ICENCO.2017.8289772

[6] “Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight,” Z. Chen and W. Zhang, PLoS computational biology, vol. 9, no. 3, p. e1002956, 2013

[7] Oguntimilehin, Abiodun, Olusola Adetunmbi, and Innocent Osho. "Towards achieving optimal performance using stacked generalization algorithm: a case study of clinical diagnosis of malaria fever." *Int. Arab J. Inf. Technol.* 16.6 (2019): 1074-1081.

[8] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.