Evolutionary Intelligence (2024) 17:419–428
https://doi.org/10.1007/s12065-021-00663-1

SPECIAL ISSUE

# Diabetic risk prognosis with tree ensembles integrating feature attribution methods

James Hansen[1]

Received: 24 March 2021 / Revised: 20 August 2021 / Accepted: 21 August 2021 / Published online: 18 September 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

**Abstract**
Tree ensemble machine learning models offer particular promise for medical applications because of their ability to handle both continuous and categorical data, their faculty for modeling nonlinear relationships, and ease with which hyperparameters can be adapted to improve performance. Modern methods include Random Forests, XGBoost and LightGBM, which are robust across many areas of diagnosis, prognosis, and medical treatments. Yet a critical limiting factor of ensembles is that they are difficult to interpret due to their complex inner workings. In medicine the ability to explain and interpret a model can be vital for clinical acceptance and trust. Diabetes and cardiovascular disease are two of the main causes of death in the United States. Identifying and predicting these diseases in patients is the first step towards stopping their progression. Utilizing the NHANES diabetes mortality data set, it is shown that the Random Forests ensemble with optimized hyperparameters yields a strong prognosis model. Importantly, conjoining Random Forests with SHapley Additive exPlanations (SHAP) yields reliable interpretability of the contributions and interactions among the features. SHAP results are compared to the recently proposed Agnostic Permutation algorithm.

**Keywords** Machine learning · Medical prognosis · Diabetes · Feature importance

## 1 Introduction

Tree ensemble machine learning models hold particular promise for medical applications because of their ability to handle both continuous and categorical data, their faculty for modeling nonlinear relationships, and ease with which hyperparameters can be adapted to improve performance. Modern methods include Random Forests, XGBoost and LightGBM, which are robust across many areas of diagnosis, prognosis, and medical treatments. These techniques apply either bagging (bootstrap aggregating) or boosting as a way to reduce variance and bias. An example diagnostic application is brain tumor auto-segmentation for magnetic resonance imaging (MRI). Applications to prognosis include risk prediction customized to individual patients and nonlinear risk models with survival trees. Models can also be learned to assist in medical treatment by predicting what the potential effect of a specific treatment might be on a patient. Additionally, tree ensembles can be combined with natural language processing techniques to extract information from radiography reports to assign labels or to establish the basis for a bot for answering medical questions.

A critical limiting factor of ensembles is that they are difficult to interpret due to their complex inner workings. Yet in medicine the ability to explain and interpret a model can be crucial for human acceptance and trust. Accurate and reliable interpretation generates user trust in the model, provides insight into how a model may be improved and supports understanding of the process being modeled [1]. In some applications, simple models (e.g., linear models) have been preferred for their ease of interpretation, even if they yield lower accuracy than more complex models. For example, in Lundberg et al. [2] it was found that while complex machine learning models often provided high prediction accuracy, their application in an actual clinical setting was limited because their predictions were difficult to interpret and hence not actionable. Interpretable methods were preferred because of their ability to explain why a certain prediction was made for a patient, i.e., specific patient characteristics that led to the prediction.

✉ James Hansen
jameshansen858@gmail.com

[1] Marriott School, Brigham Young University, Provo, UT 84602, USA

The motivation for this paper is that the incidence of diabetes continues to rise and has quickly become one of the most prevalent and costly chronic diseases worldwide. Moreover, a close link exists between diabetes and cardiovascular disease, which is the most prevalent cause of morbidity and mortality in diabetic patients [3]. Effective prognosis can contribute to understanding and provide actionable information to clinicians. Using the NHANES epidemiology dataset (CDC Website), our predictive objective is the 10-year risk of death of individuals. This dataset contains relevant features of diabetic patients, as well as their outcomes. This data is known to present a challenging non-linear prediction task [4].

Related research includes Dipnall et al. [5] who used Boosted Regression and imputation data sets from multiple chained regression sequences to identify 21 biomarkers associated with depression. Boiaraskaia [6] studied comparative performance of machine learning algorithms in testing the impact of physical activity on cardiovascular D risk. Lasso Regression, Support Vector Machines and Random Forest classifiers all performed well on large sets of data-driven features, achieving greater than 82% classification accuracy. Single decision trees produced lesser results but yielded the most transparency and interpretability.

Vangeeduram et al. [7] utilized a large-scale dataset from the NHANES corpus to research the performance of a well-known adult diabetes risk self-assessment screener and proposed pediatric clinical screening guidelines for identifying youth with prediabetes and diabetes biomarkers. They then evaluated data-driven machine learning-based classification algorithms, several of which performed significantly better than the pediatric screening guidelines.

Dihn et al. [8] tested several machine-learning algorithms to identify and predict diabetes, with the Information Gain method applied to identify the key variables in predicting diseases. Using deep neural networks, Oh et al. [4] were able to predict depression from health and demographic factors found in both the NHANES and K-NHANES datasets. Their deep learning algorithm was also able to predict depression well on new data set, both cross temporally and cross nationally.

These studies represent medical research where reliable determination of feature importance is fundamental to model trust and actionability. This, and similar requirements in other problem domains, has motivated development of robust methods for for interpreting machine learning models and gauging feature rank [3, 9–13]. This body of research has resulted in several widely implemented methods. Yet, it is shown in Lundberg and Lee [1] that each of these methods can be inconsistent. The features that are most important may not always be given the highest feature importance score. Alternatively, a model can change in a way that relies more on a particular feature, yet the importance estimate assigned to that feature decreases. This inhibits meaningful comparison among features and complicates the objective of conjoining interpretability with high levels of predictability. Moreover, it infers that *gain* and *split count* are not reliable measures of global feature importance, which is important given their common usage.

The contribution of our research is twofold: The first is to design a machine-learning approach that achieves strong results on an important medical problem, as measured by the Concordance Index (C-Index). The second is to provide an empirical comparison of two state-of-the art methods of computing feature importance. The remainder of the paper proceeds as follows. Section 2 summarizes the tree version of SHAP [1] used in this study. Section 3 outlines the recently proposed Agnostic Permutation method [14]. Section 4 describes the methodology of our experiment. Section 5 presents results and analysis. Section 6 provides concluding remarks.

## 2 Decision tree SHAP

SHAP builds on cooperative game theory as expressed in the Shapley value for any feature $\phi_i$:

$$\phi_i = \sum_{S\subseteq M\setminus i} \frac{S|!(|M| - |S| - 1)!}{|M|!}\left[f(S\cup i) - f(S)\right] \tag{1}$$

Here $|M|$ is the total number of features. $S$ represents any subset of features that doesn't include the $i$th feature and $|S|$ is the size of that subset. $\mathrm{f}_{\mathrm{s}}()$ represents the prediction function for the model for the subset $S$.

SHAP values for a tree can be computed by estimating $E[f(x)|xS]$ as shown in Algorithm 1 and then using Eq. (1) where $f_x(S) = E[f(x)|xS]$. In Algorithm 1 $x$ denotes a sample instance, the vectors $a$ and $b$ represent the left and right node indexes for each internal node. The vector $t$ contains the thresholds for each internal node, and $d$ is a vector of indexes of the features used for splitting in internal nodes. The vector $r$ represents the cover of each node (i.e., how many data samples fall in that sub-tree). The weight $w$ measures what proportion of the training samples matching the conditioning set $S$ fall into each leaf.

```
Algorithm 1: Estimating $E[f\ (x) \mid xS\ ]$
procedure $EXPVALUE(x, S, tree = \{v, a,b,t,r,d\})$
    procedure $G(j, w)$
        if $v_j \neq internal$ then
            return $w \cdot v_j$
        else if $d_j \in S$ then
            if $x_j \leq t_j$
                return $G(a_j, w)$
            else
                return $G(b_j, w)$
            end if
        else
            return $G(a_j, wr_{a_j}/r_j + wr_{b_j}/r_j$
        end if
    end if
    end procedure
    return $G(1, 1)$
end procedure
```

Here, $G$ is a function that gets called recursively to walk down the tree starting at the root node. $w$ is the weight given to the predictions of each node. $v$ is the prediction of a leaf node. $r_{a_j}$ and $r_{b_j}$ are the number of data points in the left and right child nodes of node $j$, and $r_j$ denotes the number of data points in node $j$. To clarify, we parse key components below.

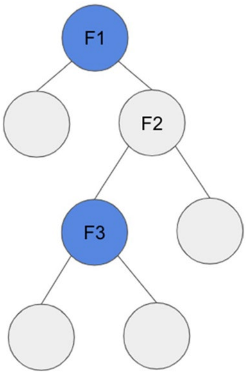
### 2.1 Leaf nodes

The code segment below processes a given leaf node. It accepts the prediction of that leaf node and multiplies it by some weight. The weight is determined by the proportion of training data points that end up reaching that leaf node,

```
if $v_j \neq internal$ then
    return $w \cdot v_j$.
```

### 2.2 Ignoring a feature

This code handles the case where the feature that is used at a node is not included in the subset of features that are allowed for making a split. Here is an example, with F2 not included.



This means that the training samples in both the left and right sub-tree of F2 should be included when making a prediction. In other words, it is supposed that the model was not trained on F2. In this case, the algorithm computes the sum of the weighted predictions from both its left and right subtree.

### 2.3 Including a feature

For cases when the feature at node is within the desired subset of features, then the algorithm can follow the left subtree or the right subtree, whichever path that the input data is assigned to via the split.

```
if $d_j \in S$ then
    if $x_{d_j} \leq t_j$
        return $G(a_j, w)$
    else
        return $G(b_j, w)$
```

After Algorithm 1 computes $f_x = E[f(\mathrm{S}) = E[f(x)|x_S)]$ it can be substituted into Eq. (1) to calculate $\phi_i$. Additive feature attribution methods, which are simple models that are used to explain complex models, can then be applied. In particular

$$g\left(z'\right) = \varphi_0 + \sum\nolimits_{i=1}^{M} \varphi_i z_i' \tag{2}$$

Here the $z_i'$ variables represent the feature being observed $(z_i' = 1)$ or unknown $\left(z_i' = 0\right)$ and the $\varphi_i'$ are the feature attribution values.

Intuitively, think of the tree model as a complex model that we seek to explain with a simple, linear model. Equation (2) specifies that we can select a single datapoint and let the model make a prediction. Its attribution can then be divided among the features based on how important those features are to the model's prediction and based on whether the feature values push the prediction in the positive or negative direction. For example, say that we have trained a model on three features. If it is given no inputs to make a prediction, then its prediction would be the weighted average of all its training samples. For example, suppose that the weighted average of the training labels is 200. That is, if a model is given no feature inputs and asked to make a prediction, it would predict 200, the expected value based on the training labels. Now, assume we give the model a single sample observation for all three features, and the model returns a prediction of 300. The additive feature attribution model might assign feature importance to the three features as follows:

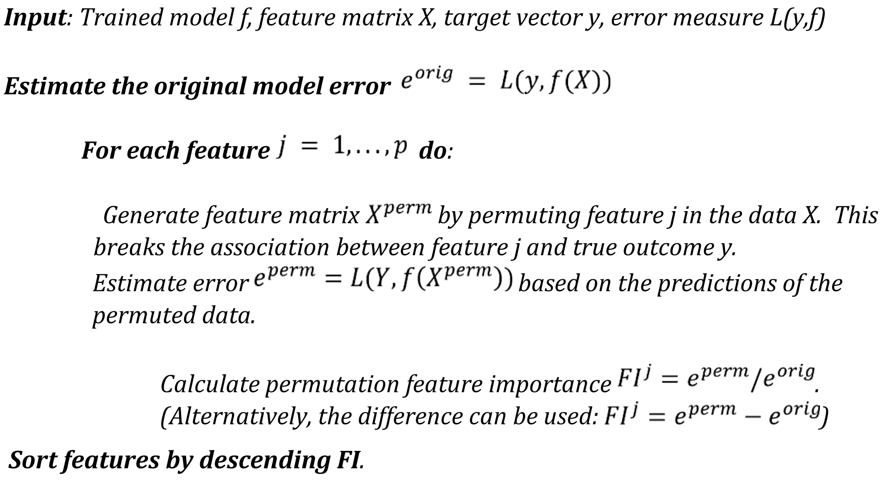
feature 1: +90
feature 2: +70
feature 3: −30

This asserts that feature 1 pushed the model's prediction up by 90, feature 2 pushed the model's prediction up by 70, and feature 3 pushed the model's prediction down by 30. The result is to move from the expected value of 200 to a prediction of 330. In general, the calculated SHAP values are those values that push the model's prediction from the average of the training labels to the model's final prediction. When the SHAP values are summed for all model features the result is the model's prediction.

## 3 Agnostic permutation method

Traditionally, tree-based models have motivated measures of feature importance based on mean decrease in impurity. Impurity is quantified by the splitting criterion of the decision trees (Gini, Entropy, or Mean Squared Error). However, these methods can assign highest importance to features that may not be predictive on unseen data when the model is overfitting. Conversely, permutation-based feature importance is predictive on unseen data. Additionally, impurity-based feature importance for trees is strongly biased and favors high cardinality features (typically numerical) over low cardinality features such as binary factors or categorical variables with a small number of possible categories. Permutation-based feature importance does not exhibit such bias [14].

Historically, permutation feature importance measurement was introduced by Breiman [13] for Random Forests. Fisher et al. [14] modernized this approach by developing a model-agnostic permutation method, which measures the increase in the prediction error of the model after permuting the feature's values and breaks the relationship between the feature and the true outcome. The importance of a feature is measured by calculating the increase in the model's prediction error after permuting the feature. A feature is *important* if shuffling its values increases the model error, inferring that the model relied on the feature for the prediction. Likewise, a feature is *unimportant* if shuffling its values leaves the model error unchanged, indicating that the model ignored the feature for the prediction. Formally,



```
Input: Trained model f, feature matrix X, target vector y, error measure L(y,f)

Estimate the original model error $e^{orig} = L(y, f(X))$

    For each feature $j = 1,\ldots,p$ do:

        Generate feature matrix $X^{perm}$ by permuting feature j in the data X. This
        breaks the association between feature j and true outcome y.
        Estimate error $e^{perm} = L(Y, f(X^{perm}))$ based on the predictions of the
        permuted data.

            Calculate permutation feature importance $FI^{j} = e^{perm}/e^{orig}$.
            (Alternatively, the difference can be used: $FI^{j} = e^{perm} - e^{orig}$)

Sort features by descending FI.
```

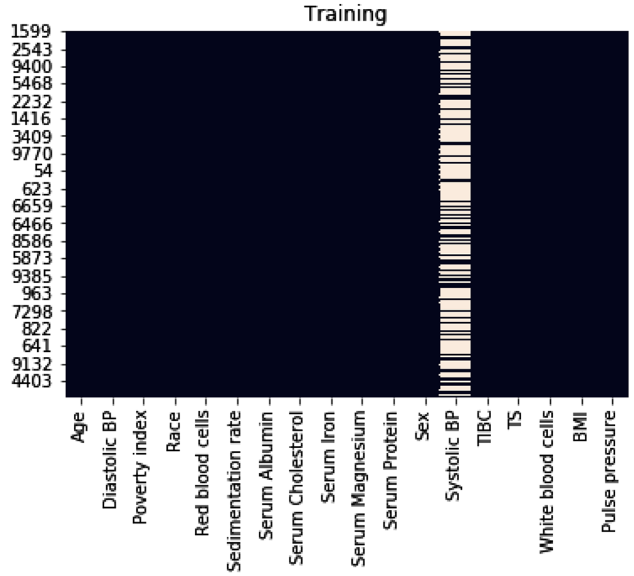


**Fig. 1** Patient feature heat map

In total, the permutation importance measure automatically considers all interactions with other features. Permuting the feature destroys the interaction effects with other features, inferring that the permutation feature importance takes into account both the main feature effect and the interaction effects on model performance. However, this can be limiting because the importance of the interaction between two features is included in the importance measurements of both features. Accordingly, the feature importance values may not add up to the total drop in performance.

## 4 Experimental design

Our experimental design pipelines five components: NHANES diabetes data, the Random Forest algorithm, meta-heuristic hyperparameter search, C-Index performance metric, and feature interpretation by both SHAP and Agnostic Permutation methods. Considering the NHANES diabetic data, an audit of missing values is reflected in the heatmap of Fig. 1. Missing data is a common occurrence in medical data due to a variety of reasons: measuring instrument malfunction, respondents not willing or not able to supply information, or errors in the data collection process. Figure 1 shows each feature as a column, with values that are present shown in black and missing values in a lighter color. From this plot, it can be seen that many values are missing for systolic blood pressure (Systolic BP). A simple solution is to discard all instances containing a missing value in the Systolic BP feature. However, this may eliminate useful information.

Instead, we imputed missing values using stochastic regression imputation. Formally, a feature column is designated as output $y$ and the other feature columns are treated as inputs $X$. A regressor is fit on $(X, y)$ for known $y$. Then, the regressor is used to predict the missing values of $y$. A stochastic modification attempts to preserve the variability of the data by adding an error, or residual, term to each predicted value. The residual term is normally distributed with a mean of zero and a variance equal to the variance of the predictor used for imputing.

### 4.1 Hyperparameter tuning

A common method of improving Random Forest performance is by hyperparameter tuning, which is the task of approximating optimal hyperparameters for a learning algorithm for a particular dataset. While grid search is popular, a potentially more effective way to search for an optimal set of hyperparameters is with a random search estimator. Expressly, a randomized search meta-estimator is an algorithm that trains and evaluates a series of models by taking random draws from a predetermined set of hyperparameter distributions. The algorithm picks the most successful version of the model it's seen after training $N$ different versions of the model with different randomly selected hyperparameter combinations, leaving a model trained on a near-optimal set of hyperparameters. This method has an advantage over grid search in that the algorithm searches over distributions of parameter values rather than predetermined lists of candidate values for each hyperparameter.

### 4.2 Performance metric

We employ the C-Index as a performance metric. The reason for its widespread use in medical studies is its efficacy in measuring the ability of a model to discriminate between different classes by quantifying how often [when considering all pairs of patients $(A, B)$], the model predicts that patient $A$ has a higher risk score than patient $B$, when in the observed data patient $A$ actually died and patient $B$ actually lived. Here, our model is a binary classifier, where each risk score is either 1 (the model predicts that the patient will die) or 0 (the patient will live).

Formally, define *permissible pairs* of patients as pairs where the outcomes are different, *concordant pairs* as permissible pairs where the patient that died had a higher risk score (e.g., the model predicted 1 for the patient that died and 0 for the one that lived), and *ties* as permissible pairs where the risk scores are equal (e.g., the model predicted 1 for both patients and 0 for both patients). The C-Index is then computed as

$$C - Index = \frac{\#concordant\ pairs + 0.5 * \#ties}{\#permissable\, pairs}$$

**Hypothesis 1** A Random Forest ensemble, with parameters optimized by random search, will yield *strong performance* as measured by the C-Index, which has proven an effective measure of how well a model predicts time to an event. S*trong performance* has been assessed as 80 percent or higher [8].

### 4.3 Feature importance

Perhaps most valuable is reliable assignment of feature importance to the results. In addition to enhancing interpretability, feature importance can aid in guiding clinical actions, and evidencing accountability to medical regulations. It is natural that different feature importance algorithms will produce somewhat different results. We seek to test the significance of these differences.

While tree-ensembles lose the natural interpretability of single trees, the recent development of SHAP [1] provides a theoretically sound, cutting-edge method to explain predictions made by models which are too complex to be understandable by humans. Given a prediction made by a machine learning model, SHAP values explain the prediction by quantifying the additive importance of each feature to the prediction. Although it is computationally expensive to compute SHAP values for general black-box models, in the case of trees and forests there exists a fast polynomial-time algorithm. Because of its recent development, we contrast the Agnostic Permutation algorithm [14] with SHAP.

**Hypothesis 2** Reliable assessment of feature importance is fundamental to machine learning applications. Since SHAP [1] provides a unifying theoretical foundation, we hypothesize that it will yield a *s*ignificantly different interpretation of variable importance than the heuristics used in the advanced version of Agnostic Permutation [14]. Other methods have been excluded considering the evidence provided in Sect. 1.

## 5 Results and discussion

After applying multivariate imputation and a randomized search meta-estimator the Random Forests algorithm yields the following results Table 1.

**Table 1** Experiment results

| | |
|---|---|
| Train C-Index | 0.8531 |
| Validation C-Index | 0.8454 |
| Test C-Index | 0.8297 |

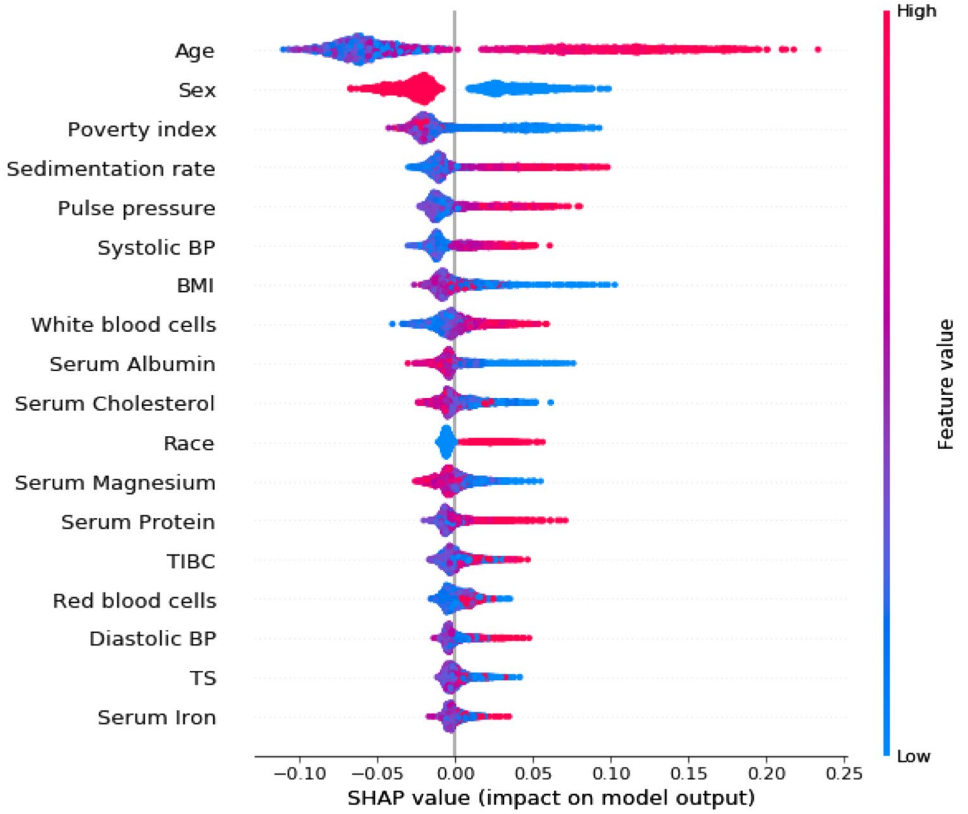


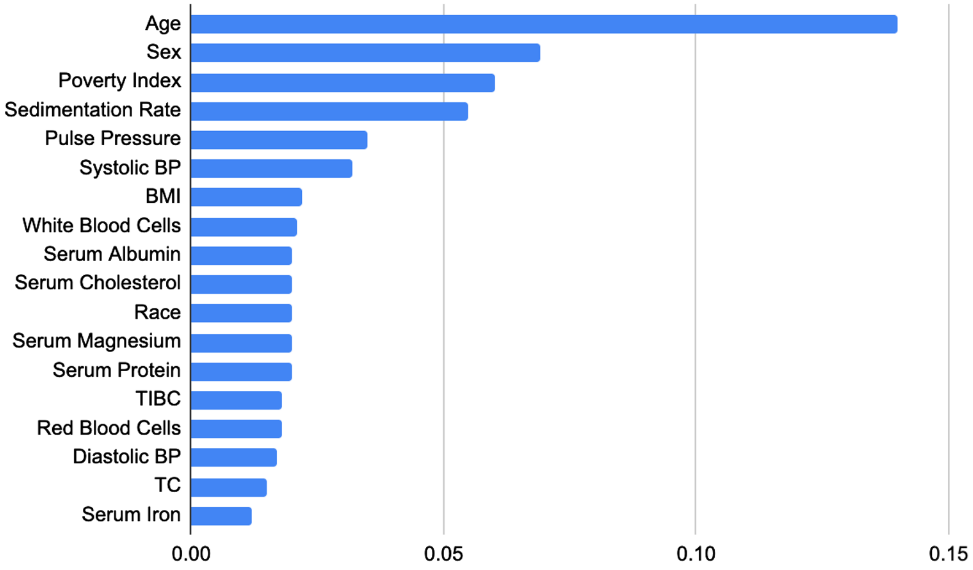
**Fig. 2** Aggregate feature impact

While evaluation of such results can depend on specific objectives, generally C-Index results over 0.80 are indicative of a strong model [8]. Given this constructive result, it is important to interpret what the model has found in the data.

Particularly, we seek to know the underlying importance of the covariate values. Standard feature importance bar charts give a notion of relative importance in the training dataset but they do not represent the range and distribution of impacts that feature has on the model's output; nor how the feature's value relates to its impact. SHAP summary plots leverage individualized feature attributions to convey these aspects of a feature's importance, while remaining visually concise.

Concretely, Fig. 2 sorts features by the sum of SHAP value magnitudes over all samples and uses those values to show the distribution of the impacts each feature has on the model output: color represents the feature value (red high, blue low). For example, it can be seen that being a woman

**Fig. 3** SHAP feature importance



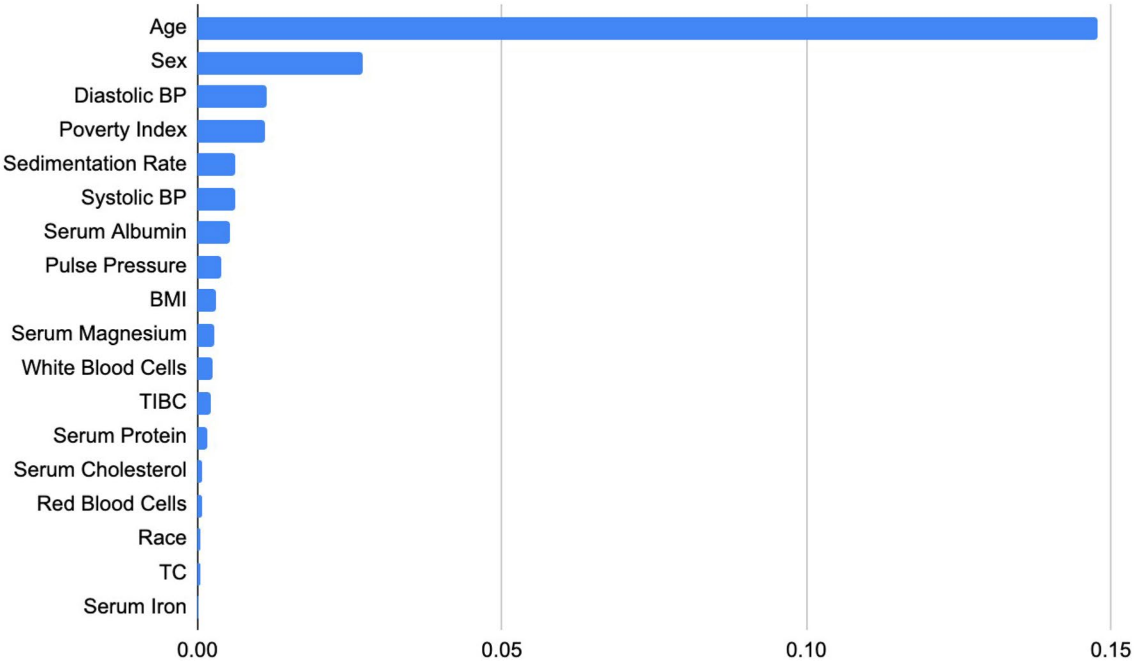


**Fig. 4** Permutation feature importance

(sex = 2.0, as opposed to men for which sex = 1.0) has a negative SHAP value, meaning that it reduces the risk of dying within 10 years. Conversely, high Age and high Systolic Blood Pressure contribute positive SHAP values and are therefore associated with increased mortality. The density of the Age plot shows how common different ages are in the dataset, and the coloring shows a smooth increase in the model's output as age increases. In contrast to Age, Systolic Blood Pressure only has a large impact for a minority of people with high blood pressure. The general trend of long tails reaching to the right, but not to the left, infers that extreme values of these measurements can significantly raise risk of death but cannot significantly lower risk.

We can also examine the mean absolute value of the SHAP values for each feature to obtain stacked bars for multi-class outputs as illustrated in Fig. 3. Here we see respective importance for both types of outcomes. This will be useful in comparing to the results of Agnostic Permutation.

## 5.1 Contrasting agnostic permutation

For comparison we compute feature importance with the Agnostic Permutation [14] method. With this method, the importance of feature *i* is the regular performance of the model minus its performance with the values for feature *i* permuted in the dataset. In this way, one can assess how well a model without that feature would do without having to train a new model for each feature. The resulting feature importance is shown in Fig. 4.

Comparing Figs. 3 and 4 suggests that SHAP and Agnostic Permutation do not produce the same importance distribution for model features. A pairwise t-test confirms this, yielding

$$p = 0.000035$$

Additional perspective is provided in Table 2, which shows the comparative feature importance ordering of SHAP versus Agnostic Permutation. There are consistencies among the top rankings for both methods, such as Age, Sex, and Systolic BP. Conversely, a marked difference is seen in the importance ranking of Diastolic BP. Other notable differences are observed in Pulse Pressure, White Blood Cells, and Serum Cholesterol. These variations in importance may convey different actionable information.

Based upon study outcomes, Hypothesis 1 is supported. Our model yields a C-Index of 0.8297 (> = 0.80).

**Table 2** Feature importance comparison

| Feature | SHAP importance | Permutation importance |
|---|---|---|
| Age | 1 | 1 |
| Sex | 2 | 2 |
| Poverty index | 3 | 4 |
| Sedimentation rate | 4 | 5 |
| Pulse pressure | 5 | 8 |
| Systolic BP | 6 | 6 |
| BMI | 7 | 9 |
| White blood cells | 8 | 11 |
| Serum albumen | 9 | 7 |
| Serum cholesterol | 10 | 14 |
| Race | 11 | 16 |
| Serum magnesium | 12 | 10 |
| Serum protein | 13 | 13 |
| TIBC | 14 | 12 |
| Red blood cells | 15 | 15 |
| Diastolic BP | 16 | 3 |
| TS | 17 | 17 |
| Serum iron | 18 | 18 |

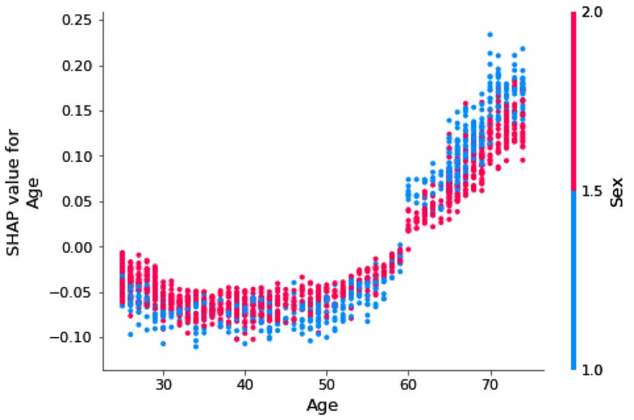


**Fig. 5** SHAP dependence between age and sex

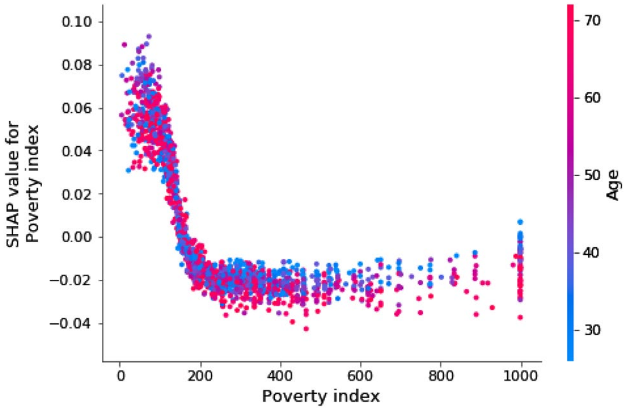


**Fig. 6** SHAP dependence between poverty index and age

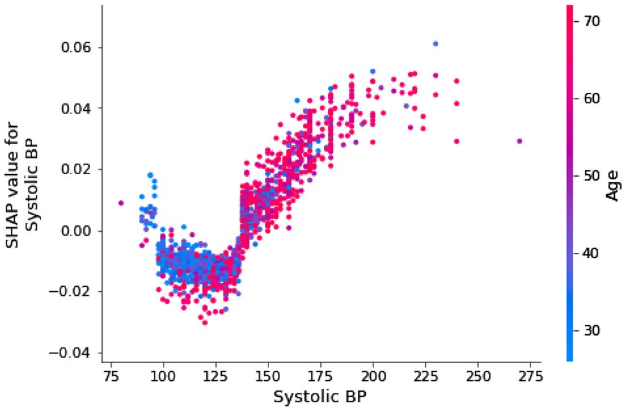


**Fig. 7** SHAP dependence between systolic BP and age

For Hypothesis 2, SHAP and Agnostic Permutation generated significantly different feature importance structures as postulated. Based upon its unifying mathematical theory, we propose that the differences favor SHAP.

### 5.2 Potential actionable dependencies

The SHAP interpretation can probe deeper using dependence plots. These show the SHAP value for a given feature for each data point, and color the points in using the value for another feature. This helps to explain the variation in SHAP value for a single value of the main feature. In particular, it is of interest to explore dependencies among the top contributing features from Figs. 2 and 3. For example, Fig. 5 shows the interaction between Age and Sex.

We see that while Age > 60 is generally riskier (increasing positive SHAP value), being a woman appears to mitigate the impact of Age. This parallels evidence that women generally live longer than men.

Next consider the dependencies between Poverty Index and Age shown in Fig. 6.

Observe that the impact of the Poverty Index transforms quickly from positive to negative SHAP values, and for higher income individuals Age appears to correlate with variation in the impact of the Poverty Index. While the Poverty Index is indirectly associated with 10-year mortality risk, it appears that reducing its underlying causes could be actionable.

Lastly, Fig. 7 indicates that high blood pressure is concerning when a person is young (blue dots associated with high Systolic BP). Positive SHAP values are less surprising as Age increases, as it appears to require time for high blood pressure to lead to fatal complications. Recent reports put an upper bound for normal Systolic BP pressure at 129. Figure 7 suggests that the danger level is about 137, which is considered Stage 1 hypertension. (https://www.health.harvard.edu/blog/new-high-blood-pressure-guidelines-2017111712756)

## 6 Summary and conclusion

This study was motivated by two theses. Firstly, that effective prognosis of diabetic risk can contribute to understanding and provide actionable information to clinicians. Secondly, that accurate interpretation of medical prognosis models can generate user trust in the model, provide insight into how a model may be improved, support understanding of the process being modeled, and provide a data-driven basis for strategic action. Conjoining the robustness of tree ensembles with modern feature importance algorithms provided a schema for testing two hypotheses.

Overall, a novel integration of SHAP with Random Forests yields a strong model and discovers in a consistent way the most important features relevant to predicting 10-year survivability. As this is among early studies of its kind, it may provide a useful platform for investigations of related medical domains.

More broadly, current research is focusing on explanations that improve user understanding and user task performance, with an increasing emphasis on the concept of responsible artificial intelligence (AI) [15]. Responsible AI seeks methodologies for the large-scale implementation of AI methods in real organizations with fairness, model explainability and accountability at its core [16]. In this context, we are currently pursuing research on bias and fairness in medical machine learning, with a focus on fairness and remediation.

## References

1. Lundberg S, Lee S (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4768–4777
2. Lundberg S, Nair B, Vavilala M, Mayumi H, Eisses M, Adams T, Liston D, Low D, Shu-Fang Newman S, Kim J (2017) Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. bioRxiv, 206540
3. Leon B, Maddox B (2015) Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. World J Diabetes 6:1246–1258
4. Oh J, Yun K, Maoz U, Kim T, Chae J (2019) Identifying depression in the national health and nutrition examination survey data using a deep learning algorithm. J Affect Disord 257:623–631
5. Dipnall J, Pasco J, Berk M, Williams S, Dodd S, Jacka F, Meyer D (2016) Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. PLoS One 11(2):e014819511
6. Boiarskaia E (2016) Recognizing cardiovascular disease patterns with machine learning using NHANES accelerometer determined physical activity data. Doctoral dissertation, University of Illinois, Champaign

7. Vangeepuram N, Liu B, Chu P, Wang L, Pandey G (2019) Predicting Youth diabetes risk using NHANES data and machine learning. Sci Rep 11(1):1–9
8. Dinh A, Miertschin M, Mohanty S (2019) A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19:1–15
9. Bach S (2015) Pixel-wise explanations for non-linear classifier decisions by layerwise relevance propagation. PLoS One 10(7):e0130140
10. Ribeiro M, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
11. Saabas A Interpreting random forests. http://blog.datadive.net/interpreting-random-forests/
12. Shrikumar A (2016) Not just a black box: learning important features through propagating activation differences. In: arXiv preprint http://arxiv.org/arXiv:1605.01713.
13. Breiman L (2001) Random forests. Mach Learn 45:5–32
14. Fisher A, Rudin C, Dominici F (2018) Model class reliance: variable importance measures for any machine learning model class, from the “Rashomon perspective.” http://arxiv.org/abs/1801.01489.
15. Gunning D, Aha D (2019) DARPA’s explainable artificial intelligence (XAI) program. AI Mag 40(2):44–58. https://doi.org/10.1609/aimag.v40i2.2850
16. Arrieta A, Diaz N, Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.