Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions

Shuyu Lu, MS[1], Ruoyu Chen, PhD[1,2i], Wei Wei, PhD , Mia Belovsky, BS[3], Xinghua Lu MD, PhD[1]
[1]Dept. Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA;
[2]Computer School, Beijing Information Science & Technology University, Beijing, China
[3]Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA, USA

Abstract

Heart failure (HF) is a major cause of mortality. Accurately monitoring HF progress and adjusting therapies are critical for improving patient outcomes. An experienced cardiologist can make accurate HF stage diagnoses based on combination of symptoms, signs, and lab results from the electronic health records (EHR) of a patient, without directly measuring heart function. We examined whether machine learning models, more specifically the XGBoost model, can accurately predict patient stage based on EHR, and we further applied the SHapley Additive exPlanations (SHAP) framework to identify informative features and their interpretations. Our results indicate that based on structured data from EHR, our models could predict patients' ejection fraction (EF) scores with moderate accuracy. SHAP analyses identified informative features and revealed potential clinical subtypes of HF. Our findings provide insights on how to design computing systems to accurately monitor disease progression of HF patients through continuously mining patients' EHR data.

Introduction

Heart failure, also commonly referred to as congestive cardiac failure, is a clinical syndrome when the heart is unable to pump sufficiently to maintain blood flow to meet the body's needs. According to CDC, HF is one of the most prevalent diseases and with highest mortality rate[1] within the US, with the estimating that 6.2 million adults are affected[1]. When a patient has a clinical encounter, their signs, symptoms, labs and other data is input into the EHR and the treatment plan is tailored to their specific physiologic condition. Thus, accurately monitoring disease states and adjusting the regimen in a timely manner can prevent or slow heart failure progression. The availability of an incoming stream of patient data including (but not limited to) weight and blood pressure readings, symptom monitoring and medication tracking has significantly lowered the mortality rates of heart failure patients, readmission rates, medical expenses, and improved patient satisfaction[2,3]. Accurately monitoring disease progression and timely adjusting regimen to prevent or slow progression would improve HF patients' life expectancy and quality. With prevalent availability of EHR, continuously mining of the EHR by computational agents to accurately detect and report disease progression will likely become a component of long term care of HF patients. In this study, we investigated the feasibility of using computational agents to assess disease stage and characteristic clinical features that provide information about disease and progression.

An objective measurement of heart function is the EF of left ventricle of heart. Broadly classified, there are 2 types of heart failure: heart failure with reduced ejection fraction (HFrEF) and heart failure with preserved ejection fraction (HFpEF). The difference lies in whether the systolic or diastolic capabilities of the left ventricle is affected[4], which is primarily measured by the left ventricle ejection fraction (LVEF or EF) score. The relation between EF score and heart failure is shown in **Table 1**. According to 2016 European Society of Cardiology (ESC) Clinical Practice Guidelines[5], an EF score less than 40% typically suggests HFrEF, while an EF score greater than 50% is often a sign of HFpEF. As a distinct quantitative measure of heart failure stage, therefore, compared to other symptoms of heart failure, EF score is an effective

EF Score	Pumping Ability of the Heart	Level of Heart Failure
50%-70%	Normal	No HF/HFpEF
40%-50%	Slightly below Normal	Slight Symptoms of HF
35%-40%	Moderately below normal	Mild HFrEF
$<$35%	Severely below normal	Severe HFrEF

Table 1: Relation between EF score and HF[5]

[i]This work is completed during Dr. Chen's visit at the University Pittsburgh

variable to use when applying machine learning methods to predict a patient's heart failure status. Currently, EF score is measured using an echocardiogram during inpatient/outpatient visits, but the frequency of echocardiogram can be few and far between in comparison to recordings of other clinical data. Furthermore, encounters of a HF patient with a health system are multi-faceted and occur in multiple settings, e.g., outpatient visits, nurse call, pharmacy visits, etc. All these encounters are usually recorded in the EHR of a comprehensive health system like the University of Pittsburgh Medical Center (UPMC). Therefore, EHR is an ideal resource for exploring indicators for diagnoses and outcomes, and we hypothesize that the continuous mining of EHR to detect disease progress would be of high clinical value in future.

Machine learning models have been increasingly applied to explore medical scenarios related to heart failure[6–8]. In general, interpretable models are preferred over "black box" models in clinical settings. Among modern machine learning models, tree-based models, e.g., the XGBoost[9] is often favored in clinical settings, because they are easy to interpret, capable of handling missing values, and solve overfitting and underfitting problems effectively with the help of regularization methods. However, most earlier studies concentrate on model performance or feature importance[6–8], paying little attention to fully understanding and explaining the predictions with interpretable methods. That is, a human user not only is interested to know what features are informative with respect to a prediction task, but also would like to know how to interpret an observed value of a feature with respect to the prediction task. To this end, machine learning methods have been developed to identify informative features and their interpretations in a model-agnostic fashion[11–16]. For example, a model-agnostic interpretation method, such as SHAP[10], can take a dataset and different prediction models as inputs, apply the models to the data, and subsequently discover the characteristics of data features in each prediction model (thus model-agnostic) different prediction models.

In this paper, we investigated the utility of XGBoost model in predicting heart failure stages, which is represented by EF scores based on structured EHR data. We evaluated informative features and investigated their interpretability and characteristics using the SHAP framework. Finally, based on the characteristics of features and their values, we applied unsupervised cluster learning to discover subtypes (subpopulations) among HF patients. To our best knowledge, few studies attempt to address the same questions as reported here, and we anticipate that our approach lays a foundation for future development of computational agents capable of monitoring HF patient disease progress by continuously mining the EHR data of health systems.

Materials and Methods

Data Collection and Preprocessing

All the EHR data were obtained from UPMC [ii] from 2014 to 2019 and contained clinical conditions of the patients who have been diagnosed with HF before, identified using ICD9/ICD10 codes (I428.* and I50.* respectively) which indicate heart failure. The original dataset consisted of 9 different CSV files that were extracted from the UPMC EHR system, including demographics (DEMO_*), vitals (VL_*), labs (LB_*), medical dispenses (MD_*), medical fills (MF_*), medical orders (MO_*), order results (OR_*), problem lists (PL_*), and diagnoses (DI_*). Each patient had a unique ID, and each file was a collection of data from over 2,000,000 encounters from 60,835 unique patients. In order to build a profile for each patient, we cleaned the raw CSV files based on rules, extracted each patient's record from the CSV files and finally aggregated the information. The workflow is illustrated in **Figure 1**:

The rules for data processing were:

1) For medical fills, medical orders, medical dispenses, problem lists and diagnoses, only keep the drug and disease names that appear over 10,000 times (10,000/2000,000 = 0.5%) in the dataset as valid features.

2) For numerical features like age, BMI, and blood pressure, in order to exclude outliers as much as possible, normalize the values and only keep the ones between 1% and 99% percentile, and values outside this range are then set to the value of 1 percentile (MIN) or 99 percentile (MAX).

3) The medical fills, medical orders, and medical dispenses are mixture of the National Drug Code (NDC) and the Anatomical Therapeutic Chemical (ATC) code. The problem lists and diagnoses are mixture of the ICD-9

[ii] Data acquired through the Health Record Research Request (R3) of University of Pittsburgh under the IRB# PRO18080460

and ICD-10 code. For the sake of consistency, all drugs NDC codes are mapped to ATC codes, and all ICD-9 codes are mapped to ICD-10 codes with the help of Apache Lucene[17].

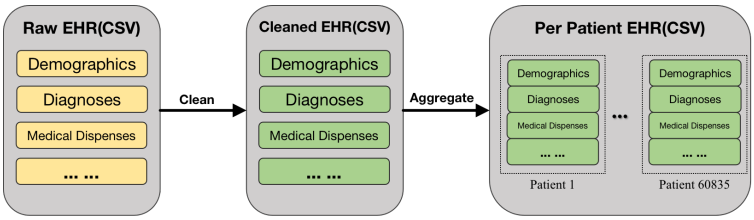


Figure 1: Data clean & aggregate workflow

After data cleaning, 1894 features from the 9 categories were retained and the values were stored tables. Details are available in *Feature Table*[iii].

Training data

To train and test the models, we selected the patients with a valid EF score, which is a predominant indicator for HF diagnosis. The EF score also served as the label in our model because it is an objective measure, easy to interpret, in addition it can seamlessly fit into the SHAP framework.

The EF scores were directly obtained from patients' echocardiogram reports. We extract the rest of structure EHR data of a patient within 45 days of EF measure, and we processed the EHR to derive a feature vector matching an EF measurement. Since each patient can have multiple EF measurement across time, and a patients disease may progress through such time, we treat EF measurements and match features from a patient that were separated by more than 180 days as "independent" cases. Finally, we obtained 130,727 cases. We then split the dataset: 70% for training, 20% for validation and 10% for testing.

XGBoost Model for Regression

We applied XGBoost[9] for our prediction task. XGBoost is an optimized distributed gradient boosting library and has been successfully applied in medical studies[18,19]. Boosting algorithms is to combine weak classifiers together to form a powerful one. XGBoost is a boosting tree model that combines many CART regression tree models.

The regression tree model attempts to predict the original EF score with structured tabular EHR data. We tuned several hyperparameters with the evaluation of 5-folder cross validation for the XGBoost model before reaching the final model. The parameters are tuned with the sklearn.model_selection.GridSearchCV[20] package, with coordinate descent[21] as the strategy. The final tuned parameters are listed in **Table 2**:

Parameter	Value
n_estimators (# of trees)	100
max_depth	3
eta (learning_rate)	0.35
min_child_weight	1
col_sample_by_tree	1
col_sample_by_level	1
subsample	0.85
reg_alpha (L1 regularization)	0
reg_lambda (L2 regularization)	0.5
gamma	0
num_boost_round (# of boosting iterations)	500

Table 2: XGBoost fine-tuned parameters

SHAP for Model-Agnostic Interpretation

SHAP stands for SHapley Additive exPlanations. It is a game theoretic approach to explain the output of a machine learning model[10]. It connects optimal credit allocation with local explanations using the classic Shapley values[22] from game theory and their related extensions. SHAP assigns a unique SHAP value to each feature in a sample for an EF prediction. The SHAP value represents the deviation from the average predicted value for each case prediction brought by each feature. For our work, we applied the SHAP model to firstly generate SHAP values for all our test dataset cases, and then illustrated the SHAP summary plot and SHAP scatter plot using a publicly available SHAP API[23] for a global understanding of our dataset.

T-SNE clustering with SHAP values

T-SNE (t-distributed Stochastic Neighbor Embedding) is a visualization machine learning algorithm based on stochastic neighbor embedding[24]. In particular, it models each high-dimensional object with a 2D or 3D point, with similar

[iii] https://github.com/Frank-LSY/XGB-SHAP-EHR-EF

cases placed as nearby points in t-SNE space.

For our work, after assigning a SHAP value for each feature of a subject, we represented a sample using a feature vector of SHAP values and clustered data points accordingly. We examined whether this approach could reveal different subtypes of HF patients by inspecting the clusters in comparison to clustering samples in original feature space. We applied t-SNE algorithm to map the 1894 features in 2D to visualize the clustering results.

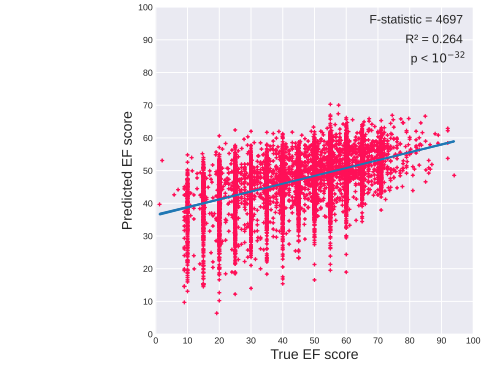
Model implementing details

The experiments were conducted on an on-prem server of 72 of CPU(Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz). The operating system was Ubuntu 18.04.4 LTS. All the codes for the task were written in Python 3.6[25] *(Github Links*[iv]). The XGBoost model was implemented with xgboost 1.1.1[26] and scikit-learn 0.23[27]; the SHAP interpretation was implemented with shap[28]; t-SNE algorithm was implemented with scikit-learn.manifold.TSNE[29], where the perplexity was set to 100.

Results

Predicting performance of XGBoost model

The performance for XGBoost regression is $RMSE = 12.6303 \pm 0.00201$ (95% CI) with 100 random attempts on validation dataset. We evaluated the correlation between the predicted EF and real EF scores (**Figure 2A**), of which $R^2 = 0.264$, with a $p < 10^{-32}$. **Figure 2B**, shows the first tree (the most prominent out of 100 trees), which illustrates what features are utilized by the tree and how it splits samples to make final predictions. For example, the first feature for split is the DEMO_GENDER, which indicates that gender is a very important feature for predicting EF among HF patients. Note that our model only used structured data from the EHR to predict a numeric score with moderate accuracy. Besides, The predictions clearly follows correct trends.



(a) Regression performance

(b) The first regression tree

Figure 2: XGBoost regression tree for predicting EF.

Interpretation with XGBoost and SHAP

The XGBoost model and SHAP evaluate features from different perspectives. The XGBoost uses the *coverage* to reflect the importance of a feature, which denotes the percentage cases in which a feature is utilized during the decision pathway in making the final prediction. On the other hand, SHAP analysis assigns a SHAP value for a feature in each case (i.e., case-specific), which reflects the impact on the feature (measured as deviation from the mean predicted value) when different combinations of features (including or excluding the feature of interest) are used to predict the target value of a case. The important features provided by two methods are shown in **Figure 3**, where **Figure 3A** is the feature importance generated with XGBoost with coverage$\geq$0.01, and **Figure 3B** is the top 20 most important features according to the SHAP analysis.

Interpretation of SHAP scores is as follows. For the SHAP value plot, each row presents the distribution of SHAP values assigned to a feature across all cases. The x-axis denotes the SHAP value, and the unit reflects how much the presence/absence of a feature with a particular value in a case will lead to deviation of predicted EF score (the unit is %) from the mean of prediction values using all possible combinations of feature sets from the case. The pseudo-color of a data point indicates the value of the feature of interest in a case. The further a point deviates from the mean of

[iv] https://github.com/Frank-LSY/XGB-SHAP-EHR-EF

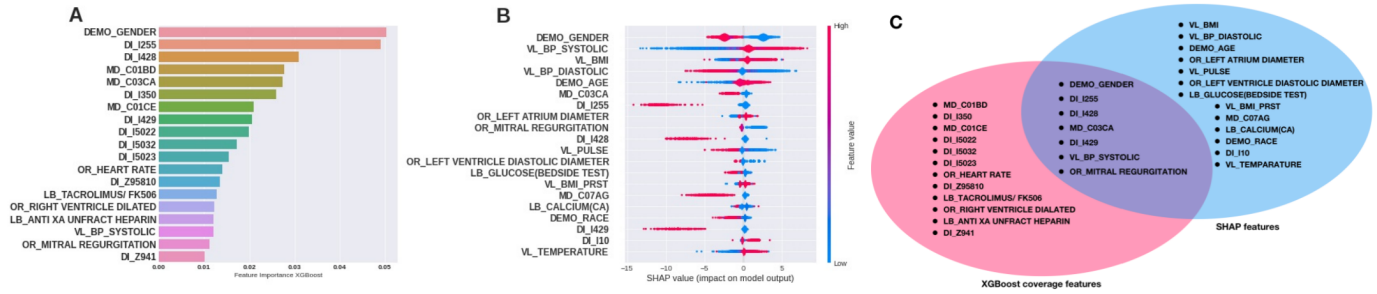


Figure 3: Comparison of informative features identified by XGBoost and SHAP.

predictions (which is 0), the more impact the features have on the prediction in the case. Therefore, a positive SHAP value that is on the right side of the mean on the x-axis, indicates the feature with the value in a case leads to a target value above the average predicted value, and below the average value on the left side. For example, for VL_BMI, the patient's SHAP value is positive when the BMI is higher, and negative when the BMI is lower(except for extreme cases). That is, in the SHAP interpretation for our prediction model, heart failure patients tend to have EF scores higher than average if the BMI value is high and vice versa.

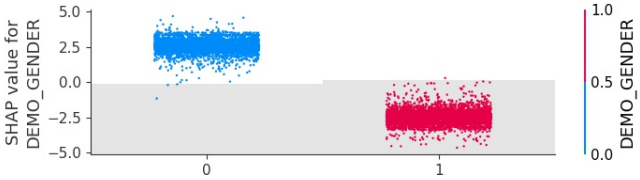


Figure 4: Impact of gender on EF in HF patients.

As shown in **Figure 3C**, the two methods identified a common set of informative features as well as some disjoint features, but the overall ranking of the features is similar. Firstly, the most important features that have highest coverage for XGBoost feature importance and highest SHAP value impact is DEMO_GENDER. If we dive deeper, like what is shown in **Figure 4**, we can see that female patients tend to have about a 5% higher EF score compared to male patients (female is represented as 0 and male as 1).Additionally, both XGBoost and SHAP interpretations treat DI_I255 (Ischemic Cardiomyopathy), DI_I428 (other Cardiomyopathies) and DI_I429 (Cardiomyopathy, unspecified) as critical diagnoses; MD_C03CA (Sulfonamides, plain) as critical medical dispenses, OR_MITRAL REGURGITATION (Mitral valve regurgitation) as critical order results, if presented. According to **Figure 3B**, BP_SYSTOLIC(systolic blood pressure) and BP_DIASTOLIC(diastolic blood pressure) all have relatively high SHAP values, but their contributions to the prediction are quite the opposite. As shown in **Figure 5**, the model tend to assign higher EF score value predictions if patients have higher systolic blood pressures, while assign lower EF score value predictions if patients have higher diastolic blood pressures.

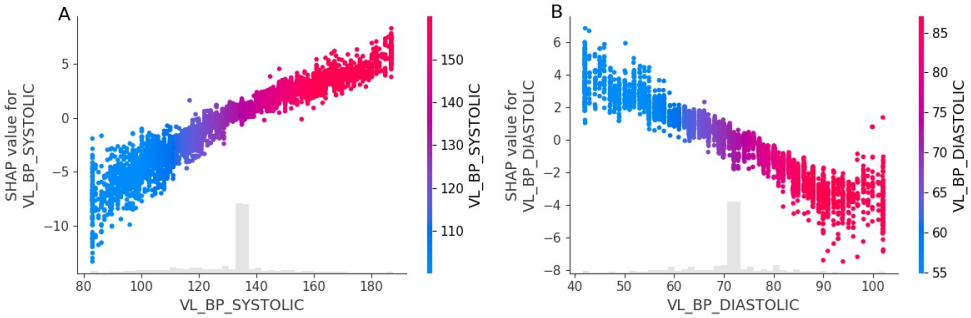
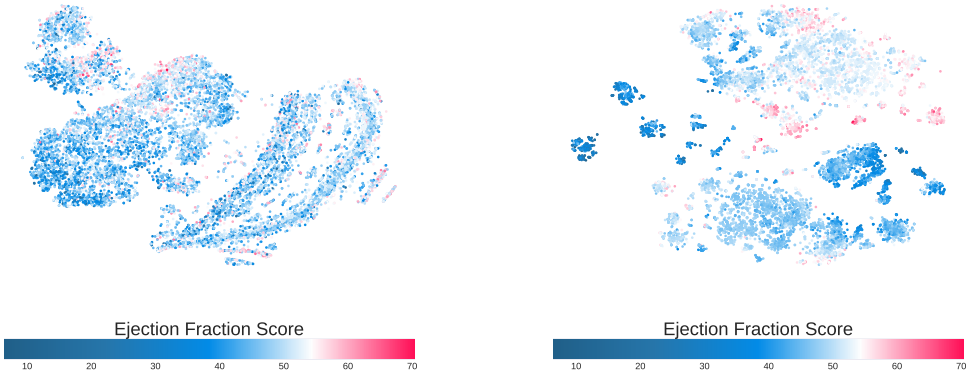


Figure 5: Relationship between blood pressure and EF in HF patients.

Sample clustering analysis with t-SNE and SHAP value

SHAP analysis provides a new perspective for inspecting data points: it shows the case-specific impact of each feature of a data point. This provides us with an opportunity to inspect whether different cases share a common pattern (joint

distribution) of SHAP scores that characterize a subset of cases, which would provide a new perspective to identify patients that may share a common underlying disease mechanism. We applied the t-SNE algorithm to visualize the distribution of samples in the original feature space as well as in SHAP score space. For the former, each case was represented in the original feature space, and for the latter, each case's SHAP values were used as input features, and t-SNE projected the data points from both representation into a 2D space.



(a) t-SNE unsupervised feature value clustering for EF score (b) t-SNE supervised SHAP value clustering for EF score

Figure 6: Clustering HF patients in different feature spaces

Figure 6 shows the general clustering result with original feature values and SHAP values of all features respectively. The color represents EF score for each case. Compared with **Figure 6A**, which is the clustering of samples based on original feature values, clustering based on SHAP values (**Figure 6B**) makes the data points with similar EF values closer on the plot. More specifically, patients with high EFs (HFpEF) appear to be evenly distributed among the patients with reduced EFs (HFrEF) in the t-SNE results based on the original features, whereas HFpEF patients tend to be more clustered together when SHAP feature values were used as inputs. Furthermore, the HFpEF patients form sub-clusters in the SHAP-derived t-SNE space, which indicates that there are distinct combinations of SHAP values (thereby clinical features) that were detectable by the t-SNE algorithm. In summary, representing data points in the SHAP space revealed characteristics of samples that were not detectable in the original data space. We then set out to investigate which features contributed to the sub-clusters in the SHAP-derived t-SNE analysis. For the 10 graphs in this section in **Figure 7**, we plotted each data point in the same position as they were in **Figure 6B**, and we used pseudo-color to illustrate the original values of different features in each sub-plot.

When inspecting the combination of features of different sub-clusters, it is interesting to note that gender appears to a major factor that leads to the division of two large sub-populations (**Figure 7A**). This finding agrees with the finding from XGBoost analysis as mentioned previously: gender is the most important feature to be considered when the XGBoost model trying to predict the EF score. The separation of patients according gender indicates that patients with different genders tend to have distinct combinations of characteristics of other features, which suggests there is a major difference in disease mechanisms of HF patients of different genders.

It can also be seen that mitral regurgitation is present in many patients except few smaller clusters. The presence of mitral regurgitation is usually associated with dilated left ventricles, which is frequently associated with reduced EF (HFrEF). For the rest of sub-clusters, each plot has at least one unique feature that clearly assumes a value different from that of the majority of the patients, reflecting the importance of such a feature in differentiate sub-populations of HF patients. It also provides a more in-depth understanding of the particular patterns of feature combinations and how they may reveal distinct disease mechanisms underlying these subgroups.

Figure 7: Clustering of HF patients (t-SNE) based on combination of SHAP values of different features

Discussion

Findings

In this study, we showed that the XGBoost regression model could be trained to predict EF score (an objective measurement of heart function) with fair performance, only based on structured EHR data. Furthermore, we showed that SHAP analysis revealed information and patterns that could not be easily acquired through analyzing the original features identified by XGBoost. The results are encouraging in that they demonstrated the feasibility of monitoring HF patient disease stage, thereby progression of the disease, of HF patients through mining the EHR data. We anticipate that with more training cases and more information from detailed clinical notes that provide the information of symptoms and signs of HF patients, the accuracy of the predictive model can be further improved and eventually make such a system clinically applicable. Given the dominant impact of HF on human mortality, a small improvement in monitoring patient disease progression can be translated into a significant improvement in overall patient outcomes.

Some of our findings are confirmed by previous studies, which indicate the validity of our approach in discovering patterns and important features. For example, the 5% difference in EF values between men and women was reported in several publications. According to Chung et al.[30] who reviewed 1435 women and 1183 men cMRI, and concluded whether there is heart failure or not, females have a median EF score of 75% and males have a median of 70%, $p < 0.001$. This is consistent with the 5% gender difference in our XGBoost predicted results. Moreover, earlier studies[31–33] also reported that in patients with heart failure of different genders, females are more likely to be diagnosed with HFpEF, whereas males are more likely to be diagnosed with HFrEF. Some other recent findings also demonstrate that there are significant differences in the incidence, prevalence, disease course and pathophysiology of heart failure between biological males and females. Females tend to survive longer after a heart failure diagnosis and more often have diastolic dysfunction, while males have a overall higher incidence of heart failure. Additionally, the macro and microvascular pathology underlying the etiology of the heart failure has been proven to vary between males and females as well[3].

Our study also highlighted trends that have not yet been confirmed in previous publications: the relation between blood pressure and EF score. According to Katsuya et al.[34] among patients with acute heart failure syndromes (AHFS), it has been reported that those with a reduced left ventricular ejection fraction (LVEF) tend to be hypotensive or normotensive, whereas those with a preserved LVEF tend to be hypertensive. Their study which evaluated 4831 patients led to the conclusion that patients with an admission SBP$<$120 mmHg were more likely to have a reduced LVEF than a preserved LVEF. In contrast, patients with an admission SBP $\geq$ 120 mmHg were equally likely to have a preserved or reduced LVEF, indicating that there was no relation between a higher admission SBP and the LVEF. However, we did not find references that suggest the changes in systolic and diastolic blood pressure would have opposite effects on EF score. Such a combination of characteristics in these features are apparent in our analyses. The discovery of such combination patterns, thereby sub-populations of patients, not only prompts further investigation

of potentially distinct underlying disease mechanisms but also suggests that tailored prediction/monitoring models should be developed for different sub-populations (a mixture of expert models), to enhance their performance.

Limitations

This is an early attempt at using ML models to detect the stage of HF in patients. Currently, the model only utilizes the structured data from the EHR, missing a significant amount of information from clinical notes. Besides, the features we selected according to our feature engineering approach are not necessarily matching exactly with heart failure. This is why our XGBoost regression model only achieves fair performance. One future direction should be extracting an informative representation of symptoms and signs associated with HF to enhance the accuracy of the predictions. Another limitation is that the current model does not attempt to model the temporal trajectory of heart function, which is important for targeting earlier interventions in order to improve clinical outcomes. This will be addressed in future studies.

Conclusion

With the XGBoost model, SHAP interpretation and unsupervised clustering visualization, we can 1) predicted EF score from tabular EHR data with decent performance; 2)generated interpretations for both the XGBoost model and dataset; and 3) classified the subgroups of HF. The generated interpretations are consistent with HF diagnosis guidelines and human intuition. This article provides a basic understanding of certain variable associations with heart failure. The model demonstrated the variables such as gender, blood pressure, age, pulse, BMI, some diagnoses (miscellaneous cardiomyopathies), and medications (sulfonamides, alpha and beta blocking agents) all have an impact on heart failure stage. To a large extent, this indicates that the future use of machine learning models to construct clinical decision aids related to heart failure is justifiable and feasible.

Acknowledgements

This work is partially support by NIH grant (R01LM012011) to LX. The authors would like to thank Mr. Zheng Li for discussions.

Author Contributions

Lu, S conceived the study under advice of Lu, X. Lu, S designed and implemented the programs, performed data analysis, and drafted the manuscript. Chen, R contributed to data processing. Wei, W participated in study design and literature review. All authors contributed to writing and editing of the manuscript.

References

1. Virani, S. S., Alonso, A., Benjamin, E. J., et al. (2020). Heart disease and stroke statistics—2020 update: A report from the American Heart Association. In Circulation (Vol. 141, Issue 9, pp. E139–E596). Lippincott Williams and Wilkins. https://doi.org/10.1161/CIR.0000000000000757.
2. Wakefield, B. J., Boren, S. A., Groves, P. S., & Conn, V. S. (2013). Heart failure care management programs: A review of study interventions and meta-analysis of outcomes. In Journal of Cardiovascular Nursing (Vol. 28, Issue 1, pp. 8–19). https://doi.org/10.1097/JCN.0b013e318239f9e1
3. Inamdar, A., Inamdar, A. (2016). Heart Failure: Diagnosis, Management and Utilization. Journal of Clinical Medicine, 5(7), 62. https://doi.org/10.3390/jcm5070062
4. Weinbrenner, S., Langer, T., Scherer, M., et al. (2012). Nationale VersorgungsLeitlinie Chronische Herzinsuffizienz. In Deutsche Medizinische Wochenschrift (Vol. 137, Issue 5, pp. 219–226). https://doi.org/10.1055/s-0031-1292894
5. Ponikowski, P., Voors, A. A., Anker, S. D., et al. (2016). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. In European Heart Journal (Vol. 37, Issue 27, pp. 2129-2200m). Oxford University Press. https://doi.org/10.1093/eurheartj/ehw128
6. Gutman, R., Shalit, U., Caspi, O., Aronson, D. (2020). What drives success in models predicting heart failure outcome? European Heart Journal, 41(Supplement_2). https://doi.org/10.1093/ehjci/ehaa946.3556

7. Frizzell, J. D., Liang, L., Schulte, P. J., et al. (2017). Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. JAMA Cardiology, 2(2), 204–209.
8. Budholiya, K., Shrivastava, S. K., Sharma, V. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2020.10.013
9. Chen, T., & Guestrin, C. (n.d.). XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785
10. Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017-Decem(Section 2), 4766–4775.
11. Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
12. Katuwal, G. J., Chen, R. (2016). Machine learning model interpretability for precision medicine. http://arxiv.org/abs/1610.09045
13. Yang, C., Delcher, C., Shenkman, E., Ranka, S. (n.d.). Predicting 30-day all-cause readmissions from hospital inpatient discharge data.
14. Che, Z., Purushotham, S., Khemani, R., Liu, Y. (2016). Interpretable deep models for ICU outcome prediction. AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2016, 371–380. /pmc/articles/PMC5333206/?report=abstract
15. Luo, G. (2016). Automatically explaining machine learning prediction results: A demonstration on type 2 diabetes risk prediction. Health Information Science and Systems, 4(1), 1–9. https://doi.org/10.1186/s13755-016-0015-4
16. Krause, J., Perer, A., Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models Data Curation View project Interpreting and Visualizing Machine Learning Models View project Interacting with Predictions: Visual Inspection of Black-box Machine Learning. https://doi.org/10.1145/2858036.2858529
17. apache/lucene-solr: Apache Lucene and Solr open-source search software. (n.d.). Retrieved November 28, 2020, from https://github.com/apache/lucene-solr
18. Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, 225–241.
19. Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93–101. https://doi.org/10.1016/j.eswa.2016.04.001
20. sklearn.model_selection.GridSearchCV — scikit-learn 0.23.2 documentation. (n.d.). Retrieved November 29, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
21. Wright, S. J. (n.d.). COORDINATE DESCENT ALGORITHMS FOR LASSO.pdf. 1–27. http://www.optimization-online.org/DB_FILE/2014/12/4679.pdf
22. Shapley, L. S. (2016). 17. A Value for n-Person Games. In Contributions to the Theory of Games (AM-28), Volume II (pp. 307–318). https://doi.org/10.1515/9781400881970-018
23. API Reference — SHAP latest documentation. (n.d.). Retrieved February 25, 2021, from https://shap.readthedocs.io/en/latest/
24. Hinton, G., & Roweis, S. (n.d.). Stochastic Neighbor Embedding.
25. Python Release Python 3.6.6 — Python.org. (n.d.). Retrieved November 29, 2020, from https://www.python.org/downloads/release/python-366/
26. dmlc/xgboost: Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more. Runs on single machine, Hadoop, Spark, Dask, Flink and DataFlow. (n.d.). Retrieved November 29, 2020, from https://github.com/dmlc/xgboost/

27. Release Highlights for scikit-learn 0.23 — scikit-learn 0.23.2 documentation. (n.d.). Retrieved November 29, 2020, from https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_0_23_0.html

28. slundberg/shap: A game theoretic approach to explain the output of any machine learning model. (n.d.). Retrieved November 29, 2020, from https://github.com/slundberg/shap

29. sklearn.manifold.TSNE — scikit-learn 0.23.2 documentation. (n.d.). Retrieved November 29, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

30. Chung, A. K., Das, S. R., Leonard, D., et al. (2006). Women have higher left ventricular ejection fractions than men independent of differences in left ventricular volume: The Dallas heart study. Circulation, 113(12), 1597–1604. https://doi.org/10.1161/CIRCULATIONAHA.105.574400

31. Savarese, G., & D'Amario, D. (2018). Sex differences in heart failure. In Advances in Experimental Medicine and Biology (Vol. 1065, pp. 529–544). Springer New York LLC. https://doi.org/10.1007/978-3-319-77932-4_32

32. Duca, F., Zotter-Tufaro, C., Kammerlander, A. A., Aschauer, S., Binder, C., Mascherbauer, J., & Bonderman, D. (2018). Gender-related differences in heart failure with preserved ejection fraction. Scientific Reports, 8(1), 1080. https://doi.org/10.1038/s41598-018-19507-7

33. Regitz-Zagrosek, V. (2020). Sex and Gender Differences in Heart Failure. International Journal of Heart Failure, 2(3), 157. https://doi.org/10.36628/ijhf.2020.0004

34. K, K., N, S., Y, S., & T, T. (2013). Relationship between systolic blood pressure and preserved or reduced ejection fraction at admission in patients hospitalized for acute heart failure syndromes. International Journal of Cardiology, 168(5). https://doi.org/10.1016/J.IJCARD.2013.07.226