SI 670 Final Project Report

Mingzhou Fu, Michelle Lee, Sanghyun Lee

Introduction

Most heart diseases are associated with and reflected by the sounds that the heart produces. Heart auscultation, defined as listening to the heart sound, is a clinical art that allows the doctor to make accurate diagnoses with the skills obtained after formal training.[1] In order to capture heart sounds, physicians have been using the stethoscope for over a century. While the material and shape of the stethoscope have improved significantly, the overall fundamental design remained relatively constant.[2] One recent technological improvement is the electrification of the stethoscope, which paves the way for a new field of computer-aided auscultation and enables the recording of the sounds. This has opened the door to a number of unique possibilities including telemedicine, which assists clinical decision and education.[3]

A physiologic heart has two heart sounds (S1 and S2), which are caused by turbulence of the closure of the atrioventricular and semilunar valves respectively. When the heart is not functioning properly, additional heart sounds can be detected. These can either be extra heart sounds (S3 and S4) or flow murmurs.[4] Studying additional sounds from auscultation can be vital in the early diagnosis of heart pathologies. According to a previous study, clinical auscultation by a trained cardiologist can provide diagnoses that are relatively consistent with echocardiograph findings.[5]

However, even with advanced technologies, auscultation is still subject to many factors, including inter-listener variability, subjectivity, environment, and training.[6] Furthermore, decreasing emphasis on heart auscultation in medical education caused medical trainees to be less proficient nowadays.[7] In order to support the diagnosis of heart diseases in clinical settings, some studies have proposed to utilize machine learning methods to characterize heart sounds.[8–10]

In our project, we build several machine-learning models using a preprocessed library with heart sounds and patient echo findings collected from Michigan Medicine. The goal was to correctly classify the heart sounds into normal or abnormal labels, which would ultimately help with building auxiliary tools that can be implemented in the diagnosis of heart disease in the clinical settings in the future.

Methods

Our project is a pilot project under the study of “1000 Heart Sounds: Intelligent auscultation to predict cardiovascular disease” developed by the Department of Learning Health Science at the School of Medicine, University of Michigan. The patients enrolled in this study are aged 18 years and older, have outpatient or inpatient encounters at the University of Michigan Health System between 2017 and 2018 along with a scheduled or planned transthoracic echocardiography within 72 hours of collection. Eligible patients are identified via a MiChart report and consented for participation. Until the point of this current report, the study team had collected heart sounds and personal data from 478 patients at Michigan Medicine, and the collection is still ongoing.

In our project, we used a subset of the heart sound library built by Dr. Karandeep’s lab team. The library includes heart sounds recorded from the clinics, electrocardiogram and echocardiography results, and patients’ Electronic Health Records. Patients’ information has been de-identified under the HIPPA regulations. Heart sound recordings have already been pre-processed by other members in the lab, including quality check, signal de-noising, feature extraction, and heart sound segmentation (into S1, systole, S2, diastole) using Hidden Semi-Markov and extended Viterbi algorithm.[11]

We used the pre-processed and segmented heart sound signals to build our classification models. The final and raw dataset has 182 observations (patients) and 120 features extracted from the heart sound signal. The feature list can be found below:

(1-4) zero crossing rate:	S1, systole, S2, diastole
(5-8) duration:	S1, systole, S2, diastole
(9-12) mean:	S1, systole, S2, diastole
(13-16) maximum:	S1, systole, S2, diastole
(17-20) variance:	S1, systole, S2, diastole
(21-24) skewness:	S1, systole, S2, diastole
(25-28) kurtosis:	S1, systole, S2, diastole
(29-32) power:	S1, systole, S2, diastole
(33-36) Shannon entropy:	S1, systole, S2, diastole
(37-40) Bandwidth:	S1, systole, S2, diastole

(41-44) Q-factor:	S1, systole, S2, diastole
(45-96) mean of 13 MFCC coefficients:	S1, systole, S2, diastole
(97-120) mean of 6 wavelet packet transform coefficients:	S1, systole, S2, diastole

Because we have many features with limited number of observations, we decided to do another feature selection step. The feature selection step was conducted in the training set to avoid data leakage and was based on Koehrsen's work in 2018.[12] The features were removed if 1) greater than 60% missing value; 2) single unique value; 3) correlation magnitude greater than 0.90; 4) zero importance after one-hot encoding; 5) cumulative importance of 0.98. After feature selection, we have 39 features left in the training dataset and we implemented the same features in the testing set.

In addition, the distribution of the class (normal, abnormal) was imbalanced. There were almost three times as many normal samples as the abnormal ones. In order to avoid bias introduced by imbalanced class distribution, we upsampled the abnormal ones by randomly selecting replicated records into the dataset.

For model building, we applied multiple classic supervised machine learning algorithms such as regressions, gradient-boosted decision trees, and Support Vector Machines (SVM) to do the classification. Feature normalization and hyperparameter tuning were performed during the model building as appropriate. We also attempt applying Neural Network methods in deep learning as well.

Evaluation and Analysis

Our project goal was to successfully detect abnormal or normal heart sounds and label them correctly. Therefore, we decided to calculate accuracy, precision, and recall using the confusion matrix and classification report from the sklearn library for each model: Random Forest, Decision Tree, Gradient Boosting Decision Tree, Logistic Regression, SVM and Neural Network. Accuracy and precision were important measures. However, we put a much heavier emphasis on looking at the recall metric of the models due to the nature of our project goal (further discussed in the Discussion section).

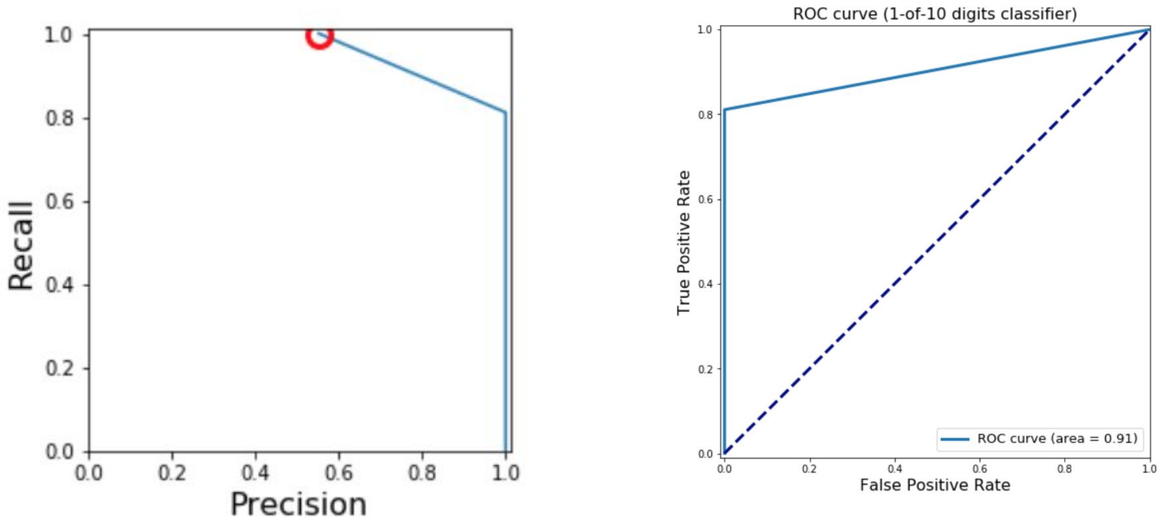
When we first tested our six models, we selected the top 10 or 20 important features from 120 features based on Random Forest's feature importance measure. However, all of our accuracy, recall, and precision performances couldn't go pass the 80% mark. Therefore, we went back and took our time to preprocess our data more thoroughly to select the most important features. We also tuned and selected parameters based on model performance and using GridSearchCV, such as C and penalty parameters of Logistic Regression, kernel, gamma and C parameters of SVM, etc.

Performance summary of different models with the highest recall score:

	Random Forest	Decision Tree	GBDT	Logistic Regression	SVM	Neural Network
Recall	0.84	0.85	0.87	0.70	**0.91**	0.45
Accuracy	0.84	0.85	0.87	0.70	**0.90**	0.55
Precision	0.85	0.85	0.88	0.70	**0.91**	0.20

After carefully selecting the optimal features and tuning the parameters for each model, we were able to improve all of the model performance of accuracy, recall, and precision score. The performance of the classification models such as Random Forest, Decision Tree, and Gradient Boosting Decision Tree were in a range from 84% to 88%, which was a significant improvement. However, the Neural network model using keras surprisingly produced the lowest recall score of 45%. Even though we tried different layers to enhance the performance, we could not find a reason to choose Neural Network to be the right model for detecting abnormal heart sounds.

SVM Precision-Recall and ROC curve



Out of the six models we used to predict the abnormal and normal heart sounds, SVM showed the highest accuracy of 90%, the highest recall of 91%, and the highest precision of 91% compared to the other models. (See figures above for the precision-recall and ROC curve for the SVM performance).

Related work

Based on the literature review, there are a few existing works related to the heart sound classification similar to ours. A study using a deep neural network for the recognition and classification of heart murmurs reaches an accuracy of 97% and a sensitivity (recall) of 93.2%, which is higher than our models.[13] Another study on logistic regression-Hidden Semi Markov Model-based heart sound segmentation reported an F1 score of 95.63 ± 0.85%, which also performs better than ours.[14]

Despite numerous techniques, there are still limitations in those studies. Most of the studies relied on a limited number of sound samples, and some of the heart sounds were artificially created. The modeling buildings and predictions were mostly based on purely clinical diagnosis; in other words, the training sets were trained with clinical diagnoses by physicians. It is unclear if the echocardiographic findings were taken into account during the diagnoses in those studies. In addition, previous studies have not considered other clinical factors, including demographic factors such as age and gender, comorbidities such as diabetes and hypertension, and examination and laboratory factors such as body mass index and blood pressure. These factors could also show a significant impact on the diagnosis of heart diseases.

Discussion and Conclusion

Our models were evaluated with three metrics: accuracy, precision, and recall scores.

The accuracy measure shows the percentage of correct predictions made out of the 67 observations that our test dataset contained. This value essentially indicates the number of observations with heart sounds that were normal in reality that the model predicted to be normal, as well as the heart sounds that were abnormal in reality that the model predicted to be abnormal. While this shows a good overview of how well the model predicts correct labels, the cost of false negatives (i.e. having a misclassified actual positive) is extremely high in our scenario. With the accuracy measure, a patient who has abnormal heart conditions that the model falsely predicts as having normal sounds would not be calculated in the score.

The precision measure shows the model performance for how many actual positives there are for the model's predicted positive cases. In our scenario, it indicates the number of cases that were actually abnormal heart sounds out of all the heart sound cases that the model predicted as abnormal. While this gives us a fair idea of grasping the situation and is important, it also is not the best metric because as said previously, this score does not include patients who have abnormal heart conditions that the model mistakenly predicted as normal. This is a more fatal case with us compared to the case where the model predicts and diagnoses a normal patient to be carrying heart anomalies because the patient can soon find out they are normal after taking more in-depth tests.

Finally, the recall metric is what we deemed as the most important in light of our project goal. This score shows how many of the actual abnormal heart sounds are correctly labeled as having abnormal heart sounds. The higher the score, the more "sick" patients the model is capturing. Correctly being able to identify patients who have abnormal heart sounds that may indicate different kinds of cardiac pathologies is crucial in the medical setting as treatment can be considered quickly.

By doing this project, we learned how important it is to understand the research question and goal of a project. With the same models, there are different evaluation metrics one can obtain. However, it is crucial to fully understand the meaning and implications of each since some may not be the ideal metric to use in a certain scenario. Incorrect understandings may lead to suboptimal conclusions, which may then lead to negatively impacting real people in society.

Observing our high performing models and recall scores that were obtained as the results, we are confident and excited to see the promising future where real-world applications of auxiliary tools get developed for early detection of cardiac pathologies based on heart sounds. If we had more time to work on the project, we would definitely strive for increasing the patient data library. We had a relatively small sample size (182 before upsampling due to imbalanced labels) so more observations of different heart sounds would definitely be beneficial in training our models. In addition, adding more information about the patients, such as their demography and comorbidities, as features may increase in correct labeling of whether they have normal or abnormal heart sounds that may lead to cardiac diseases.

References

1. Guadalajara Boo JF. [Auscultation of the heart: an art on the road to extinction]. *Gac Med Mex*. 2015;151(2):260-265.
2. Kalinauskienė E, Razvadauskas H, Morse DJ, Maxey GE, Naudžiūnas A. A Comparison of Electronic and Traditional Stethoscopes in the Heart Auscultation of Obese Patients. *Medicina (Kaunas)*. 2019;55(4). doi:10.3390/medicina55040094
3. Pyles L, Hemmati P, Pan J, et al. Initial Field Test of a Cloud-Based Cardiac Auscultation System to Determine Murmur Etiology in Rural China. *Pediatr Cardiol*. 2017;38(4):656-662. doi:10.1007/s00246-016-1563-8
4. Schneider M, Kastl S, Binder T. [Auscultation of the heart in the 21st century]. *MMW Fortschr Med*. 2019;161(6):39-42. doi:10.1007/s15006-019-0357-3
5. Patel A, Tomar NS, Bharani A. Utility of physical examination and comparison to echocardiography for cardiac diagnosis. *Indian Heart J*. 2017;69(2):141-145. doi:10.1016/j.ihj.2016.07.020
6. Mangione S, Nieman LZ, Gracely E, Kaye D. The teaching and practice of cardiac auscultation during internal medicine and cardiology training. A nationwide survey. *Ann Intern Med*. 1993;119(1):47-54. doi:10.7326/0003-4819-119-1-199307010-00009
7. Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees. A comparison of diagnostic proficiency. *JAMA*. 1997;278(9):717-722.
8. Palaniappan R, Sundaraj K, Sundaraj S, Huliraj N, Revadi SS. Classification of pulmonary pathology from breath sounds using the wavelet packet transform and an extreme learning machine. *Biomed Tech (Berl)*. 2018;63(4):383-394. doi:10.1515/bmt-2016-0097
9. Coleman W, Weidman-Evans E, Clawson R. Diagnosing and managing mitral regurgitation. *JAAPA*. 2017;30(6):11-14. doi:10.1097/01.JAA.0000516342.41351.6d
10. Kang S, Doroshow R, McConnaughey J, Shekhar R. Automated Identification of Innocent Still's Murmur in Children. *IEEE Trans Biomed Eng*. 2017;64(6):1326-1334. doi:10.1109/TBME.2016.2603787
11. Hidden semi-Markov model. In: *Wikipedia*. ; 2019. https://en.wikipedia.org/w/index.php?title=Hidden_semi-Markov_model&oldid=919316332. Accessed December 9, 2019.
12. Koehrsen W. A Feature Selection Tool for Machine Learning in Python. Medium. https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0. Published June 22, 2018. Accessed December 9, 2019.
13. Dominguez-Morales JP, Jimenez-Fernandez AF, Dominguez-Morales MJ, Jimenez-Moreno G. Deep Neural Networks for the Recognition and Classification of Heart Murmurs Using Neuromorphic Auditory Sensors. *IEEE Trans Biomed Circuits Syst*. 2018;12(1):24-34. doi:10.1109/TBCAS.2017.2751545
14. Logistic Regression-HSMM-based Heart Sound Segmentation v1.0. https://physionet.org/content/hss/1.0/. Accessed December 9, 2019.