Proceedings of the 2022 Design of Medical Devices Conference
DMD2022
April 11-14, 2022, Minneapolis, MN, USA

DMD2022-1074

DEEP UNSUPERVISED CLUSTERING OF SPARSE ECHO DATA TO IDENTIFY PATIENTS FOR IMPLANTATION OF CARDIOVERTER-DEFIBRILLATOR

Moein Enayati, Ph.D.
Mayo Clinic
Rochester, MN

Nasibeh Zanjirani Farahani, Ph.D.
Mayo Clinic
Rochester, MN

Christopher G. Scott, M.S.
Mayo Clinic
Rochester, MN

Johan M. Bos, M.D., Ph.D.
Mayo Clinic
Rochester, MN

Xiaoxi Yao, Ph.D., M.P.H.
Mayo Clinic
Rochester, MN

Che G. Ngufor, Ph.D.
Mayo Clinic
Rochester, MN

Michael J. Ackerman, MD., Ph.D.
Mayo Clinic
Rochester, MN

Adelaide Arruda-Olson, M.D., Ph.D.
Mayo Clinic
Rochester, MN

ABSTRACT

According to the 2020 report of the American Heart Association's Heart & Stroke Statistics report, nearly 1,000 people are dying daily because of sudden out-of-hospital cardiac arrests and unfortunately, their survival rate is as low as 10%. Hypertrophic Cardiomyopathy (HCM), a relatively rare genetic heart disease is one of these diseases but finding the right patient for the implantation of ICD is still a research question. Implantation of cardioverter-defibrillator (ICD) can save the life of some of these patients. Due to the complexity of the identification of HCM patients, financial burdens, and the clinical risks involved in the ICD implantation procedure, HCM patients will go into a monitoring state before reaching the implantation trigger. Our study cohort shows about 82% of HCM deaths, did not have an ICD, which highlights the need to improve the pre-screening algorithms.

In the current paper, we have proposed a new deep learning-based unsupervised clustering technique to facilitate the prioritization of patients to undergo ICD device implantation. This model uses over 900 echocardiographic measurements to find patients who benefit more from the ICD implantation procedure. Our model was trained and tested over 6 years of echo reports collected at Mayo Clinic. This model can be used as a decision support assistant for cardiologists in finding the right HCM patient when decision-making is hard.

Keywords: Deep learning, Sparse Unsupervised Auto-Encoder, Implantable Cardioverter-Defibrillator, Hypertrophy Cardiomyopathy, Echocardiography.

1. INTRODUCTION

Hypertrophic cardiomyopathy (HCM) is a heritable heart disease that may cause sudden death in young adults. Implantation of cardioverter-defibrillator (ICD) is the only effective therapy for the selected high-risk patient population to save their lives.

Currently, the selection of eligible patients for device implantation relies on individual noninvasive risk markers1 recommended by clinical practice guidelines (e.g. refer to Ommen, et al. [1]). These risk factors are collected from personal and family history notes, imaging i.e., echocardiography, and contrast-magnetic resonance reports. Every institution might rely on a different risk stratification strategy to calculate a risk score from risk factors for HCM patients. [2]. Most of the risk scores rely on the C statistic and they are less sensitive in identifying patients requiring ICDs. This would cause these patients to be left vulnerable to arrhythmic sudden death events. Also, studies have shown that only one-third of patients with appropriate ICD therapy had high-risk scores based on recommendations (greater than 6% in 5 years) and half of the patients with ICD had a low-risk score (less than 4% in 5 years). [2] The studies show, statistical models focusing on current risk factors and prognostics variables were unreliable and insensitive for the selection of HCM patients for ICD. Therefore, the need for proposing new strategies for finding the right HCM patient candidates for ICD implantation is crucial. [2, 3] In the Mayo Clinic study cohort, among 94 deaths with the confirmed cause of HCM as their main reason of death, only 17 patients had ICD

[1] The major markers are family history, left ventricular hypertrophy, unexplained syncope, Late gadolinium enhancement (LGE), end-stage HCM and LV apical aneurysm.

(18%) which shows the need for the development of better models.

Focusing on non-traditional models for HCM patients was seen in many recent publications. Authors have used machine learning models to identify hypertrophic cardiomyopathy patients from EHR billing codes and echo reports features. [4, 5]. In another study, magnetic resonance imaging reports were used for developing a natural language processing-based model for the identification of these patients[6]. The recent studies are suggesting deep analysis of the large dataset of HCM patients such as their echocardiography reports to reveal the hidden relationships between different measurements to address the cavities in current clinical decision-making processes.[7] Deep learning methods on HCM related studies have focused on ECG signals [8]. However, none of these prior studies has proposed a model for finding the right candidate for ICD from imaging reports.[1, 9]

In this work, we have focused on Mayo Clinic's longitudinal echocardiographic reports, one of the most complete data sources of HCM patients. This data source has hundreds of measurements with minor to strong linear or nonlinear correlations. The current echo dataset is very sparse, as the sonographer might fill in certain measurements and leave the rest unfilled, based on the manifestation of diseases. [4] To overcome limitations imposed due to the data sparsity, traditional techniques such as K-means and Density-Based Spatial Clustering of Applications with Noise (DBScan) were utilized in the literature, as two of the most popular clustering algorithms in unsupervised machine learning. On the other hand, deep learning has shown promising results in the clinical setting, including medical imaging data sources.[10] and was used in this work.

2. MATERIALS AND METHODS

2.1 Dataset

We have a total of 14,641 echo reports for 5,892 patients in our dataset being acquired from 2014 to 2019. These patients had an average age of 58 ±15 years old and were 62% male. 4,982 of these patients (85% of patients with 11,339 echo reports) had no ICD. From 1,145 patients with ICD, 238 patients' records were excluded as their implant date was not available. As for the remaining 524 patients with ICD, we had a total of 1,092 echo reports that were acquired before the ICD date implantation. The total number of echoes after exclusion was 2,210 reports.

2.2 Data Preprocessing

Overall, 12,431 echoes from 4,744 patients were selected for cleaning and preprocessing. From the available 927 echo measurements, 447 echo measurements for the selected patients were completely null and were removed. 56 columns with non-numeric values were transformed to categorical and 22 columns with variance < 0.05 (mostly constant values) were removed from the dataset, reaching 451 variables.

Data was divided into 75% train (n= 9323) and 25% test (n= 3108) and each group of test and train were separately normalized to the [0-1] range. Also, the null values were filled with 0, as the current implementation of autoencoder (based on Keras) does not handle null values.

3. Experiments and Results

To handle the large number of echo measurements considering their sparsity, we decided to utilize autoencoders to encode the echo information in a limited set of features in the embedded space. These features were then reduced to 2 principal components for the SparsePCA to provide clinicians with a simple visualization and further investigations.

First, the data was trained for autoencoder and resulted in a 32-embedding space. Stacked Sparse Autoencoder (SSAE) was defined as shown in Figure 1. The autoencoder was trained on the training dataset and validated on the test data, for 500 epochs with a batch size of 64. We specifically enabled the sparsity option in the setup configuration to get the best possible results.

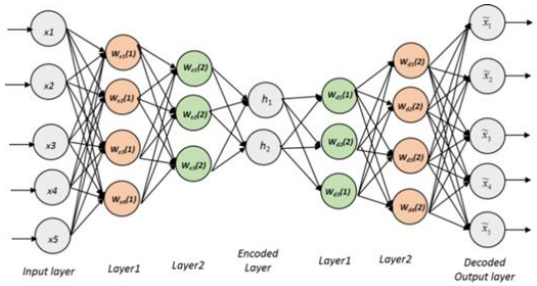

FIGURE 1: OUR STACKED AUTOENCODER WITH A TOTAL OF 5 HIDDEN LAYERS WAS IMPLEMENTED IN KERAS, AND PYTHON 1.3. THIS AUTOENCODER WAS USED TO REDUCE THE HIGH-DIMENSIONAL ECHO DATA TO 32D EMBEDDED SPACE.

Then the trained autoencoder was used to encode the entire dataset and passed through a SparsePCA for dimensionality to reduce their dimensionality to 2 for visualization purposes. We plotted the points in 2D space using the first two principal components (SparsePCA x1, x2) and color-coded the points. As shown in FIGURE 2, the echo data may be clustered into four main segments (G1, G2, G3, and G4).

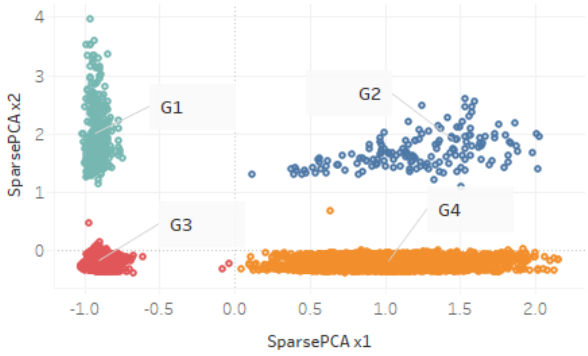

FIGURE 2: AFTER PASSING THE 32 ENCODED FEATURES TO THE SPARSEPCA THE PATIENT DATA CAME OUT IN FOUR DISTINCT CLUSTERS. IN THIS PAPER WE EXPLORED WHICH CLUSTER BENEFITS MORE FROM ICD DEVICE THERAPY.

We then added two target variables to visually investigate the different responses of each group to the installation of ICD devices. This will create four combinations as shown in Figure 3, which still hold the general theme of initial cluster orientations where the right column shows data for patients who had ICD installed, and the left column indicated patients with no ICD. Similarly, the top row shows the patients who are still alive, whereas the lower row depicts the dead patients.

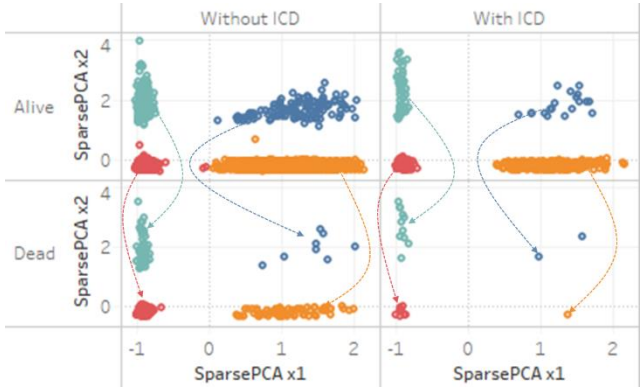

FIGURE 3. INVESTIGATING THE EFFECT OF ICD INSTALLATION (RIGHT VS LEFT COLUMNS), ON THE DEATH OUTCOME OF HCM PATIENTS (UP VS LOWER ROW) FOR EACH OF THE MENTIONED FOUR PATIENT CLUSTERS.

Now by comparing the subplots on the right columns, we can see what ratio of the patients with or without ICD have died. In fact, we can compare each cluster group (G1-4) and compute the death-ratio for each cluster and compare these rates between patients with and without ICD, to investigate which cluster group has taken more advantage of the ICD installation. We have computed these death-ratios as provided in Table I.

TABLE I: EFFECT OF ICD THERAPY ON HCM PATIENTS' DEATH RATIO. G4 HAS THE MOST REDUCTION IN DEATH.

Cluster	Without ICD	With ICD
G1	19.0%	20.4%
G2	6.3%	10.5%
G3	10.1%	4.3%
G4	3.3%	0.4%

Based on the values in Table I, while installing ICD has no significant effect on the death rate of patients in group 1, it clearly has reduced the death rate in groups 3 and 4 (to 42% and 12% of the original rates). At the same time, group 2 has a notable increase in the death rate after ICD implantation, which opens a question about the necessity of ICD for such patients, both clinically and financially.

We have also tested a few echo measurements to find the ones that can be used to identify each of these four clusters groups and found a subset of these measurements can provide at least partial clues for clusters G2 and G3 (the ones with best ICD response). Table II depicts the preliminary list of these measurements that we found to be useful in the identification.

TABLE II: LIST OF SAMPLE ECHO MEASUREMENTS (WALL SCORES) THAT SHOWED SIGNIFICANT CORRESPONDENCE TO UNSUPERVISED CLUSTERS.

No	Measurement Name
1	Basal Left Ventricular Antero-Lateral Wall Score
2	Basal Left Ventricular Anterior Septal Wall Score
3	Basal Left Ventricular Infero-Lateral Wall Score
4	Basal Left Ventricular Anterior Wall Score
5	Apical Left Ventricular Inferior Wall Score
6	Apical Left Ventricular Lateral Wall Score
7	Apical Left Ventricular Anterior Wall Score
8	Apical Left Ventricular Septal Wall Score

Figure 4 provides a sample visualization for one measurement (Basal Left Ventricular Antero-Lateral Wall Score). As shown in this figure, all of the patients in groups 3 and 4 have null values for all the measurements listed above, which needs future investigation.

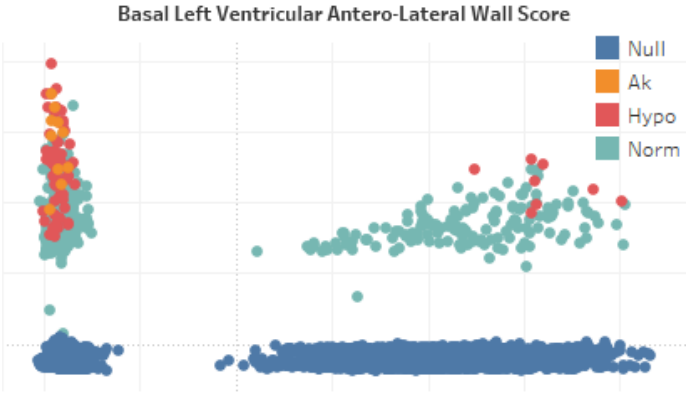

FIGURE 4: SAMPLE ECHO MEASUREMENT WAS USED TO COLOR PATIENT DATA IN ALL FOUR CLUSTERS. THE MEASUREMENT "BASAL LEFT VENTRICULAR ANTERO-LATERAL WALL SCORE" HAS FOUR VALUES INCLUDING NULL. AS SEEN ABOVE, NULL VALUES IN THIS MEASUREMENT CLEARLY CORRESPOND WITH G3 AND G4 CLUSTERS.

4. CONCLUSION

HCM is a genetic disease that could cause sudden cardiac death in young adults. The life of HCM patients with the once untreatable disease now can be saved by implantation of a device named ICD to deliver shocks to patients as needed. The clinical decision for implantation of this device relies on several risk factors that are calculated using mathematical models mainly to identify high-risk patients. The major problem with the longitudinal analysis of echo data is due to the sparsity of old measurements as the technology progresses. In the current paper, we investigated the use of deep autoencoders and SparsePCA to develop a deep similarity index that can assist clinicians in prioritizing the patients to receive ICD devices.

We have identified four visually separable clusters of patients and showed that ICD installation has a significant correlation to the reduced death rate in two of these groups. We have initiated an investigation on the difference between these four groups by evaluating a small subset of echo features and have noticed that some of these features have a clear correlation to the cluster groups. Such correlations can help in the future development of decision support tools to help clinicians identify patients who may benefit more from the ICD installation.

Although HCM is mostly diagnosed in patients with diastolic dysfunction, our experiments showed that systolic dysfunction was an important factor in finding the best ICD patient candidates. The next important factor was the left ventricular wall motion score in different segments of the heart. HCM is a disease that phenotypically often presents with normal wall motion. A wall motion score index of 1 indicates normality. Larger scores reflect more severe degrees of systolic dysfunction and our study also showed that patients with higher wall motion scores are better candidates for ICD implantation which clinically has been of proven benefit.

While we have identified these four cluster groups and showed how each one is affected by ICD installation, further investigations are needed to clarify the clinical relevance of each group and why they have different responses to the ICD therapy.

REFERENCES

[1] S. R. Ommen *et al.*, "2020 AHA/ACC guideline for the diagnosis and treatment of patients with hypertrophic cardiomyopathy: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *Journal of the American College of Cardiology,* vol. 76, no. 25, pp. e159-e240, 2020.

[2] B. J. Maron, E. J. Rowin, and M. S. Maron, "Paradigm of sudden death prevention in hypertrophic cardiomyopathy," *Circulation research,* vol. 125, no. 4, pp. 370-378, 2019.

[3] M. S. Maron *et al.*, "Enhanced American College of Cardiology/American Heart Association Strategy for Prevention of Sudden Cardiac Death in High-Risk Patients With Hypertrophic Cardiomyopathy," *JAMA Cardiology,* vol. 4, no. 7, pp. 644-657, 2019, doi: 10.1001/jamacardio.2019.1391.

[4] N. Z. Farahani *et al.*, "Application of Machine Learning for Detection of Hypertrophic Cardiomyopathy Patients from Echocardiogram Measurements," in *2021 Design of Medical Devices Conference*, 2021, vol. 2021 Design of Medical Devices Conference, V001T02A009, doi: 10.1115/dmd2021-1078. [Online]. Available: https://doi.org/10.1115/DMD2021-1078

[5] N. Z. Farahani, S. P. Arunachalam, D. S. B. Sundaram, K. Pasupathy, M. Enayati, and A. M. Arruda-Olson, "Explanatory Analysis of a Machine Learning Model to Identify Hypertrophic Cardiomyopathy Patients from EHR Using Diagnostic Codes," (in eng), *Proceedings (IEEE Int Conf Bioinformatics Biomed),* vol. 2020, pp. 1932-1937, Dec 2020, doi: https://doi.org/10.1109/bibm49941.2020.9313231.

[6] D. S. B. Sundaram *et al.*, "Natural Language Processing Based Machine Learning Model Using Cardiac MRI Reports to Identify Hypertrophic Cardiomyopathy Patients," in *2021 Design of Medical Devices Conference*, 2021, vol. 2021 Design of Medical Devices Conference, V001T03A005, doi: 10.1115/dmd2021-1076. [Online]. Available: https://doi.org/10.1115/DMD2021-1076

[7] S. Gleeson *et al.*, "ECG-derived spatial QRS-T angle is associated with ICD implantation, mortality and heart failure admissions in patients with LV systolic dysfunction," (in eng), *PLoS One,* vol. 12, no. 3, p. e0171069, 2017, doi: 10.1371/journal.pone.0171069.

[8] K. C. Siontis *et al.*, "Detection of hypertrophic cardiomyopathy by an artificial intelligence electrocardiogram in children and adolescents," *International Journal of Cardiology,* vol. 340, pp. 42-47, 2021/10/01/ 2021, doi: https://doi.org/10.1016/j.ijcard.2021.08.026.

[9] B. J. Gersh *et al.*, "2011 ACCF/AHA Guideline for the Diagnosis and Treatment of Hypertrophic Cardiomyopathy: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Developed in collaboration with the American Association for Thoracic Surgery, American Society of Echocardiography, American Society of Nuclear Cardiology, Heart Failure Society of America, Heart Rhythm Society, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons," *J Am Coll Cardiol,* vol. 58, no. 25, pp. e212-60, Dec 13 2011, doi: 10.1016/j.jacc.2011.06.011.

[10] J. M. Kwon, K. H. Kim, K. H. Jeon, and J. Park, "Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography," (in eng), *Echocardiography,* vol. 36, no. 2, pp. 213-218, Feb 2019, doi: 10.1111/echo.14220.