Proceedings of the 2021 Design of Medical Devices Conference
DMD2021
April 12-15 2021, Minneapolis, MN, USA

DMD2021-1076

NATURAL LANGUAGE PROCESSING BASED MACHINE LEARNING MODEL USING CARDIAC MRI REPORTS TO IDENTIFY HYPERTROPHIC CARDIOMYOPATHY PATIENTS

Divaakar Siva Baala Sundaram
Mayo Clinic
Rochester, MN

Shivaram P. Arunachalam
Mayo Clinic
Rochester, MN

Devanshi N. Damani
Mayo Clinic
Rochester, MN

Nasibeh Z. Farahani
Mayo Clinic
Rochester, MN

Moein Enayati
Mayo Clinic
Rochester, MN

Kalyan S. Pasupathy
Mayo Clinic
Rochester, MN

Adelaide M. Arruda-Olson
Mayo Clinic
Rochester, MN

ABSTRACT

Hypertrophic Cardiomyopathy (HCM) is the most common genetic heart disease in the US and is known to cause sudden death (SCD) in young adults. While significant advancements have been made in HCM diagnosis and management, there is a need to identify HCM cases from electronic health record (EHR) data to develop automated tools based on natural language processing guided machine learning (ML) models for accurate HCM case identification to improve management and reduce adverse outcomes of HCM patients.

Cardiac Magnetic Resonance (CMR) Imaging, plays a significant role in HCM diagnosis and risk stratification. CMR reports, generated by clinician annotation, offer rich data in the form of cardiac measurements as well as narratives describing interpretation and phenotypic description. The purpose of this study is to develop an NLP-based interpretable model utilizing impressions extracted from CMR reports to automatically identify HCM patients. CMR reports of patients with suspected HCM diagnosis between the years 1995 to 2019 were used in this study. Patients were classified into three categories of yes HCM, no HCM and, possible HCM. A random forest (RF) model was developed to predict the performance of both CMR measurements and impression features to identify HCM patients. The RF model yielded an accuracy of 86% (608 features) and 85% (30 features). These results offer promise for accurate identification of HCM patients using CMR reports from EHR for efficient clinical management transforming health care delivery for these patients.

Keywords: hypertrophic cardiomyopathy (HCM), natural language processing (NLP), machine learning, electronic health records (EHR), cardiac MRI

1. INTRODUCTION

Hypertrophic Cardiomyopathy (HCM), is the most common inherited myocardial disease with a prevalence of 1:500 (0.2%) in the general population [1-3]. The diagnosis of HCM is chiefly based on imaging findings, with emphasis on asymmetrical left ventricular hypertrophy without underlying alternative etiologies. [4-7]. CMR imaging provides accurate and in-depth phenotypic profile and information for risk stratification for HCM [3, 8].

HCM diagnosis is effective only at about 13% even using the gold standard CMR imaging, confirming the unmet need to develop better strategies among diverse population [9] A prior study used billing codes for identification of HCM patient cohorts from EHR data - reported limited performance and significant misclassification [10-11]. Comprehensive analysis of CMR reports may also reveal unknown disease correlations, characteristics associated with infrequent genetic causes of HCM, and enable genotype-phenotype association studies [2, 11]. This provides the opportunity to use CMR records to develop accurate models for HCM cohort identification using NLP and machine learning.

A manual abstraction of the CMR data is inefficient, time-consuming, and often an unmanageable task that is more prone to human errors [13]. NLP is an approach that enables software programs to analyze free-form text, detect patterns and derive meaning from human input [13-15]. This enables automatic identification and extraction of information reducing the manual review task. NLP was previously applied in extracting risk factors for SCD of HCM patients including syncope, family history of SCD, and family history of HCM (FH-HCM) [16]. The purpose of this study is to develop an NLP based interpretable machine learning model using categorical concepts from impressions from CMR reports for accurate identification of HCM patients.

2. METHODS

This section describes our method to prepare the raw data, data processing and interpretable machine learning model development steps visualized in Figure 1.

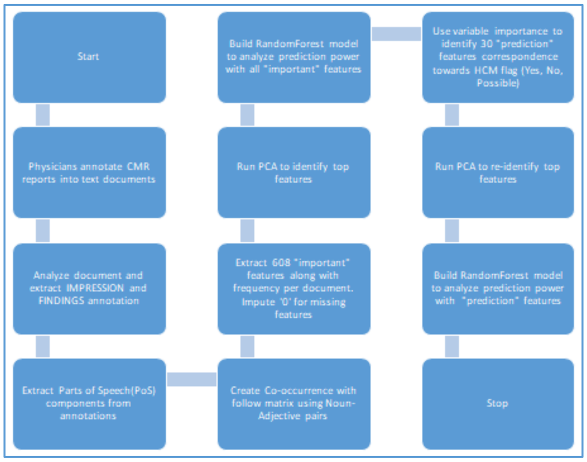

Start
Physicians annotate CMR reports into text documents
Analyze document and extract IMPRESSION and FINDINGS annotation
Extract Parts of Speech(PoS) components from annotations
Create Co-occurrence with follow matrix using Noun-Adjective pairs
Extract 608 "important" features along with frequency per document. Impute '0' for missing features
Run PCA to identify top features
Build RandomForest model to analyze prediction power with all "important" features
Use variable importance to identify 30 "prediction" features correspondence towards HCM flag (Yes, No, Possible)
Run PCA to re-identify top features
Build RandomForest model to analyze prediction power with "prediction" features
Stop

Figure 1: Visualization of data processing steps

A. Data Source

12,372 HCM patients were identified by EHR in cardiovascular registry from 1995 to March 2019 in this study after approval by the Institutional Review Board of the Mayo Clinic. For this study, HCM diagnosis was confirmed by manual review of CMR reports for ground truth. 3801 patients had CMR imaging with associated CMR reports, out of which 2257 patients had HCM mention in the CMR report. HCM cases were annotated into three HCM groups namely, yes, no, and possible. The manual review were annotated into text files under specific headers including INDICATIONS, EXAM, COMPARISON, IMPRESSION, FINDINGS, ADDITIONAL FINDINGS and MEASUREMENTS.

B. Natural Language Processing

An unsupervised NLP based CMR impressions model was developed to seek characteristic HCM text features, which can be used to accurately identify HCM cases. From the IMPRESSION and FINDINGS section of the annotation we extract different parts of speech (UPoS), Figure 2, such as nouns and adjectives to create simple noun phrases and frequency plot (Figure 3). The vector of nouns and adjectives are then analyzed for co-occurrence frequency within same sentence. The co-occurrence plot without follow (Figure 4) shows the highly co-occurring noun and adjective pair within same sentence that may not follow one another, as in, might have additional terms in between the noun-adjective pair and the co-occurrence plot with follow (Figure 5) shows the highly co-occurring noun and adjective pair within same sentence that follow each other without additional terms in between them. Finally, we used co-occurrences pair with follow to extract features. A total of 77,360 features were extracted with length from 1 to 8 and frequency from 1 to 5194 (Figure 6). We used a combination of feature-length (ngram) and frequency to extract the top 608 features (Figure 3) to train our model. Out of the 2257 records, we use 1835 first point of contact records with 608 features each. Missing features were set to zero as part of data imputation.

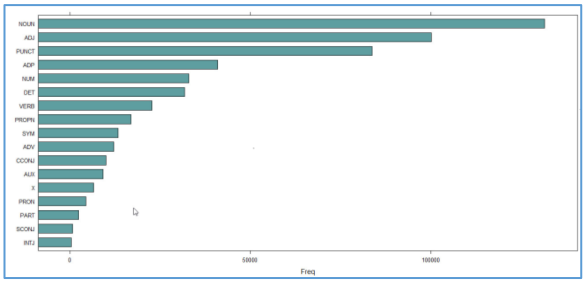


Figure 2: Frequencies of different parts of speech

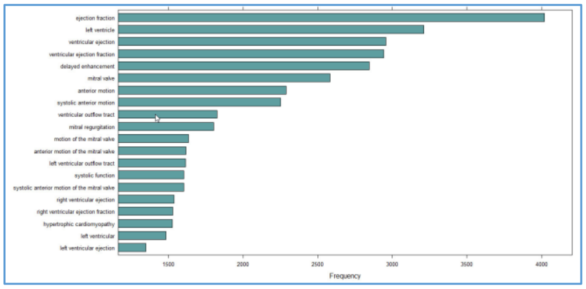


Figure 3: Simple noun phrases used in CMR reports and their frequencies

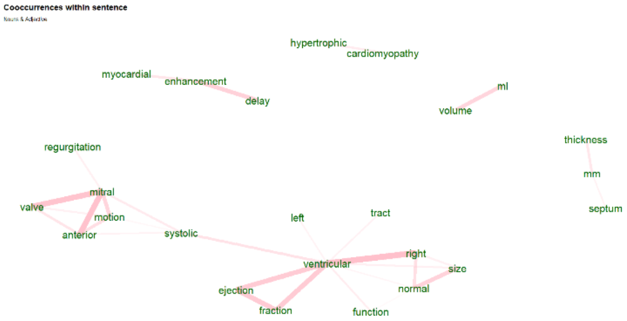


Figure 4: Co-occurrences of nouns and adjectives – No Follow (Not first adjective and second noun)

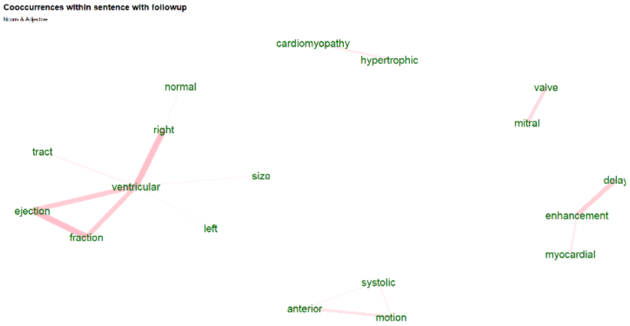
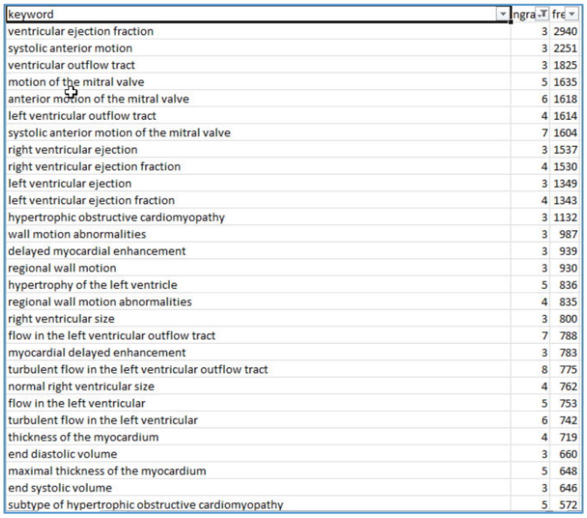


Figure 5: Co-occurrences with follow (adjective then noun) extracted from CMR reports



keyword	ngra	fre
ventricular ejection fraction	3	2940
systolic anterior motion	3	2251
ventricular outflow tract	3	1825
motion of the mitral valve	5	1635
anterior motion of the mitral valve	6	1618
left ventricular outflow tract	4	1614
systolic anterior motion of the mitral valve	7	1604
right ventricular ejection	3	1537
right ventricular ejection fraction	4	1530
left ventricular ejection	3	1349
left ventricular ejection fraction	4	1343
hypertrophic obstructive cardiomyopathy	3	1132
wall motion abnormalities	3	987
delayed myocardial enhancement	3	939
regional wall motion	3	930
hypertrophy of the left ventricle	5	836
regional wall motion abnormalities	4	835
right ventricular size	3	800
flow in the left ventricular outflow tract	7	788
myocardial delayed enhancement	3	783
turbulent flow in the left ventricular outflow tract	8	775
normal right ventricular size	4	762
flow in the left ventricular	5	753
turbulent flow in the left ventricular	6	742
thickness of the myocardium	4	719
end diastolic volume	3	660
maximal thickness of the myocardium	5	648
end systolic volume	3	646
subtype of hypertrophic obstructive cardiomyopathy	5	572

Figure 6: Sample features extracted from CMR reports using Co-occurrences with follow

C. Principal Component Analysis

To understand the most important features in the dataset, a principal component analysis was run where the principal components were in decreasing order of degree of variability in the data. Principal component analysis was ran twice on the dataset. The first principal component analysis was run on the selected 608 features and figure 7 shows that only 56% variability in MRI impressions data can be explained by the first ten principal components. The top 10 features for each of the models are identified and listed below in PCA Table 1.

The second principal component analysis is run on the 30 features selected from the dataset based on variable importance. Figure 8 shows that 57% variability in MRI impressions data can be explained with first two principal components. The top 10 features for each of the models are identified and listed below in PCA Table 2.

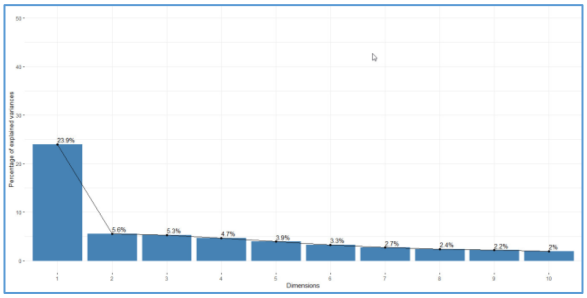


Figure 7: Principal components using selected 608 features

PCA Table 1

PC#	Impression Model
1	systolic_anterior_motion_of_the_mitral_valve
2	anterior_motion_of_the_mitral_valve
3	turbulent_flow_in_the_left_ventricular_outflow_tract
4	motion_of_the_mitral_valve
5	flow_in_the_left_ventricular_outflow_tract
6	left_ventricular_outflow_tract
7	systolic_anterior_motion
8	turbulent_flow_in_the_left_ventricular
9	hypertrophy_of_the_left_ventricle
10	right_ventricular_ejection_fraction

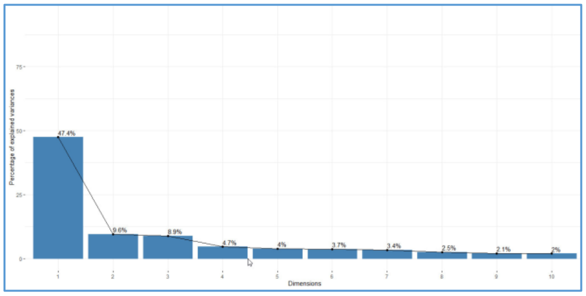


Figure 7: Principal components using selected 30 features

PC#	Impression Model
1	motion_of_the_mitral_valve
2	right_ventricular_ejection_fraction
3	flow_in_the_left_ventricular
4	maximal_thickness_of_the_myocardium
5	ventricular_ejection_fraction
6	anterior_motion
7	thickness_of_the_myocardium
8	hypertrophic_obstructive_cardiomyopathy
9	mitral_regurgitation
10	hypertrophic_cardiomyopathy

PCA Table 2

D. Machine Learning (ML) Model

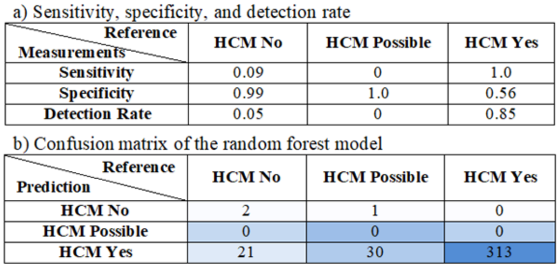
The random forest (RF) model was considered for this study for its known advantages over other machine learning tools through usage of bagging. RF model was trained using the MRI impressions derived from the rich clinical narratives after PCA.

The data set was divided into 80% training and 20% test set, where the model predicted a label based on the highest probability for any test sample. The outcomes for these models were HCM status ("yes," "no", "possible"). Ten-fold cross-validation was applied and the final (best) model was used for evaluating performance.

3. RESULTS

Of the 2257 CMR records with HCM, only 1835 records had usable rich clinical narratives to extract impressions for the model using NLP. The best model from 10-fold cross-validation for the MRI impressions data had an accuracy of 86.7% and kappa 0.13. The test set validation had an accuracy of 85.8%, no information rate (NIR) of 85.3%, p-value [ACC > NIR] of 0.419 and kappa of 0.08. Other parameters for the impression model are listed in Table 1.

TABLE 1: RF MODEL # 1 (608 FEATURES – SUMMARY)



a) Sensitivity, specificity, and detection rate

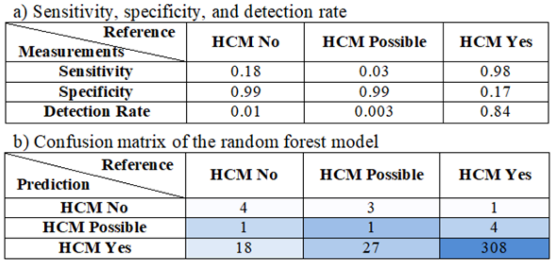
Reference Measurements	HCM No	HCM Possible	HCM Yes
Sensitivity	0.09	0	1.0
Specificity	0.99	1.0	0.56
Detection Rate	0.05	0	0.85

b) Confusion matrix of the random forest model

Reference Prediction	HCM No	HCM Possible	HCM Yes
HCM No	2	1	0
HCM Possible	0	0	0
HCM Yes	21	30	313

A second CMR impressions models were developed with just 30 features, also chosen based on variable importance using VSURF package [17]. The test set validation had an accuracy of 85.3%, no information rate (NIR) of 85.3%, p-value [ACC > NIR] of 0.536 and kappa of 0.17. Other parameters for the impression model are listed in Table 2.

TABLE 2: RF MODEL # 2 (30 FEATURES – SUMMARY)



a) Sensitivity, specificity, and detection rate

Reference Measurements	HCM No	HCM Possible	HCM Yes
Sensitivity	0.18	0.03	0.98
Specificity	0.99	0.99	0.17
Detection Rate	0.01	0.003	0.84

b) Confusion matrix of the random forest model

Reference Prediction	HCM No	HCM Possible	HCM Yes
HCM No	4	3	1
HCM Possible	1	1	4
HCM Yes	18	27	308

4. CONCLUSION AND FUTURE WORK

An NLP based RF model was developed for HCM identification using CMR impressions. The model demonstrated high accuracy between 85-87% in classifying the patients to three classes of HCM Yes, No and Possible. The results indicate the opportunity of deploying these models into real-time clinical decision support systems to enable clinicians to make informed decisions on HCM patient identification and efficient management to improve their lives. The results also provide insights into a complimentary AI solution using CMR reports for augmenting diagnosis using CMR. Future work includes incorporating the care provider team's working hours from the EHR. Further evaluation and comparisons against the state-of-the-art wearable devices are postponed to after COVID, which could validate the accuracy of the work.

REFERENCES

[1] Amano, Y., Kitamura, M., Takano, H., Yanagisawa, F., Tachi, M., Suzuki, Y., & Takayama, M. (2018). Cardiac MR imaging of hypertrophic cardiomyopathy: techniques, findings, and clinical relevance. Magnetic Resonance in Medical Sciences, 17(2), 120.

[2] Kamal, M. U., Riaz, I. B., & Janardhanan, R. (2016). Cardiovascular magnetic resonance imaging in hypertrophic cardiomyopathy: current state of the art. Cardiology journal, 23(3), 250-263.

[3] Miller, R. J., Heidary, S., Pavlovic, A., Schlachter, A., Dash, R., Fleischmann, D., & Yang, P. C. (2019). Defining genotype-phenotype relationships in patients with hypertrophic cardiomyopathy using cardiovascular magnetic resonance imaging. PloS one, 14(6), e0217612.

[4] Houston, B. A., & Stevens, G. R. (2014). Hypertrophic cardiomyopathy: a review. Clinical Medicine Insights: Cardiology, 8, CMC-S15717.

[5] Maron, B. J. (2002). Hypertrophic cardiomyopathy: a systematic review. Jama, 287(10), 1308-1320.

[6] Gersh, B. J., Maron, B. J., Bonow, R. O., Dearani, J. A., Fifer, M. A., Link, M. S., ... & Seidman, C. E. (2011). 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation, 124(24), 2761-2796.

[7] Authors/Task Force members, Elliott, P. M., Anastasakis, A., Borger, M. A., Borggrefe, M., Cecchi, F., & Mahrholdt, H. (2014). 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). European heart journal, 35(39), 2733-2779.

[8] Maron, M. S. (2009). The current and emerging role of cardiovascular magnetic resonance imaging in hypertrophic cardiomyopathy. Journal of cardiovascular translational research, 2(4), 415-425.

[9] B. J. Maron, "Clinical course and management of hypertrophic cardiomyopathy," New England Journal of Medicine, vol. 379, no. 7, pp. 655-668, 2018.

[10] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012; 13(6): 395-405.

[11] M. Pujades-Rodriguez et al., "Identifying an unmet clinical need in hypertrophic cardiomyopathy using national electronic health records," PLoS One, vol. 13, no. 1, p. e0191214, 2018, doi: 10.1371/journal.pone.0191214.

[12] P. Magnusson, A. Palm, E. Branden, and S. Morner, "Misclassification of hypertrophic cardiomyopathy: validation of diagnostic codes," Clin Epidemiol, vol. 9, pp. 403-410, 2017, doi: 10.2147/CLEP.S139300.

[13] Pons E, Braun LM, Hunink MG, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology 2016; 279(2): 329-43.

[14] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009; 42(5): 760-72.

[15] Cai, T., Giannopoulos, A. A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K. K.,. & Mitsouras, D. (2016). Natural language processing technologies in radiology research and clinical applications. Radiographics, 36(1), 176-191.

[16] Moon, S., Liu, S., Scott, C. G., Samudrala, S., Abidian, M. M., Geske, J. B., & Nishimura, R. A. (2019). Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. International journal of medical informatics, 128, 32-38.

[17] Genuer, R., Poggi, J., & Tuleau-Malot, C. (2015). VSURF: An R Package for Variable Selection Using Random Forests. R J., 7, 19.