scientific reports

OPEN

Predicting New York Heart Association (NYHA) heart failure classification from medical student notes following simulated patient encounters

Ishan R. Perera[1,2✉], Taylor Daniels[2], Janella Looney[2], Kimberly Gittings[2] & Frederic A. Rawlins II[2]

Random forest models have demonstrated utility in the determination of New York Heart Association (NYHA) Heart Failure Classifications. This study aims to determine the prediction accuracy of a random forest model to derive NYHA Classification from medical students' free-text history of present illness (HPI). NYHA Classifications established terminology for delineation of various heart failure presentations, this terminology was converted into keywords shared by standardized patients. 649 typed HPIs were de-identified, tokenized, cleaned, and assessed for number of correct keywords, incorrect keywords, and keyword usage. Models were trained using bootstrapped training data and assessed on test data. In testing, the model demonstrated a 0.775% error rate in identifying NYHA II, 26.3% for NYHA III, and 6.90% for NYHA IV. Overall reporting a 0.420% estimated error rate on the bootstrap sample training set and an 8.20% misclassification rate on the testing set. In future applications, developing a method of instantaneous feedback centered around keywords and their importance measures, specifically as determined by the variable importance plot (VIP), may aid students in their determination of NYHA Classifications and improve their lexical density.

Keywords New York Heart Association (NYHA) heart failure classification, Random forest, Machine learning, Simulated patient encounters, Medical education

Electronic Health Records (EHR) are conglomerates of data formats including structured and unstructured data[1–3]. The Edward Via College of Osteopathic Medicine (VCOM) Hospital Integrated Clinical Cases (HICC) course has adapted this EHR format in subjective, objective, assessment, and plan (SOAP) Note documentation, as a critical part of the student assessment. This standard note type is structured; however, each section may contain highly pertinent data entered in an unstructured format. Here, heuristics are important, textual analysis provides a solution to uncovering the hidden data, namely the timely diagnosis of a patient's ailment[3].

Textual analysis has been used for a multitude of applications, such as discovering language patterns, identifying document similarity and dissimilarity, classifying and sorting using clustering methods, and locating unusual behavior or anomalies in textual data[1–6]. Neural networks have outperformed current prediction systems for mortality risk in the Intensive Care Unit (ICU), risk levels in patients with atrial fibrillation, and risk of possible cardiac arrest in patients with heart disease[4,7]. Diagnostically, neural networks have surpassed current systems for identifying and evaluating cardiac arrhythmias, as well as genotypes and phenotypes of cardiovascular diseases[8–10]. Hospitals have been using text analytics and natural language processing for improved efficiency by reducing documentation time[3]. Advanced systems provide treatment recommendations based on textual input directly typed or vocalized into an EMR, known as a Computerized Clinical Decision Support System (CDSS)[11]. Analytical systems have been developed to assist physicians in extracting relevant information from unstructured modalities like free-text, demographics, imaging, and disease trends in high-risk populations like seafarers[2,10]. By having easier access to a more inclusive picture of a patient, these systems can help physicians in their differential diagnosis, while better understanding their patient's needs[2,10].

[1]Edward Via College of Osteopathic Medicine-Virginia Campus, 2265 Kraft Drive, Blacksburg, VA, USA. [2]Center for Simulation & Technology, Edward Via College of Osteopathic Medicine - Virginia Campus, 309 N. Knollwood Drive, Blacksburg, VA 24060, USA. ✉email: iperera@vt.vcom.edu

The New York Heart Association (NYHA) Functional Classification for Heart Failure stratifies patients by severity of subjective symptoms and eligibility determination for clinical trials[12]. Typically these determinations stem from the unstructured textual data within the EMR[1,13,14]. To date, the literature demonstrates the utility of machine-learning methodologies, such as natural language processing (NLP) and decision trees in the identification of NYHA Classification from unstructured data[13–15]. The primary goal of this study is to predict NYHA Classification in the unstructured history of present illness (HPI) section of first year medical students' standardized Block 4 HICC Cardiopulmonary Testing utilizing a random forest model.

Methods

The VCOM Institutional Review Board determined our study exempt from review on September 22, 2022 [1948897-1], as the study does not directly involve human subjects; instead, de-identified pre-collected data is evaluated. All activities were performed under relevant guidelines and regulations, including the Declaration of Helsinki. Informed consent was not required as human subjects were not directly involved.

The data utilized for our analysis was generated by first year Osteopathic Medical Students (OMS1) from all four VCOM campuses (Blacksburg, VA; Spartanburg, SC; Auburn, AL; Monroe, LA) during their Block 4 HICC Cardiopulmonary Testing. Specifically, following their standardized patient examinations on one of four cases: Myocarditis (NYHA II), Familial Hypertrophic Cardiomyopathy (NYHA II), Sarcoidosis Cardiomyopathy (NYHA III), and Ischemic Cardiomyopathy (NYHA IV), students completed their unstructured content entry into a simulated EMR as part of their typical Block 4 testing process. This sampling method was of convenience, due to the pre-existing integration of these processes within our institution. Importantly, standardized patients were instructed on the case background, case presentation, physical exam, and level of clinical impairment based on NYHA Classification. Patients learned the appropriate descriptors to employ in their portrayal of the assigned case for the appropriate NYHA Classification. Post-examination, raw student data was de-identified and overall performance data was excluded. NYHA I Classification descriptors were not used in this series of testing; therefore, the study will only discuss NYHA II through IV. 649 free-text HPIs and assigned standardized cases were collected.

Lexical analysis began with tokenization of responses by splitting paragraphs, sentences, and phrases into individual terms. These responses were then cleaned by the elimination of capitalization, punctuation, abbreviations, and symbols. Inverse Document Frequency was applied to find both important and frequently used words. Keyword lists were generated for each standardized case, each validated by an attending physician. Next, the amount of correct and incorrect keywords used by each student were calculated for their assigned case. The algorithm returned each keyword with a use count, percentage based on the number of correct keywords, and a percentage based on incorrect keywords specific to each individual case. A partial empty matrix was created with select data from responses consisting of the keyword count total, incorrect keyword total, correct keyword percentage, incorrect keyword percentage, and overall keyword usage denoted by either a 0 (did not) or a 1 (did use).

The data was organized into a train-test split 5/8-3/8 prepared at random using 405 entries for training (203 NYHA II, 106 NYHA III, 96 NYHA IV) and 243 entries for testing (129 NYHA II, 57 NYHA III, 57 NYHA IV). To further expand our training set, bootstrap sampling or sample selection with replacement was conducted 10 times, once merged a total of 4,050 entries were generated (2080 NYHA II, 1065 NYHA III, 905 NYHA IV).

Random forest model hyperparameters

Model hyperparameters were as follows: n_estimators, the number of trees in the model was set to 150; max_depth, depth of individual trees, was set to 'none'; min_samples_split, the minimum number of samples to justify a split, was set to 2; min_samples_leaf, minimum number of samples in a leaf node, was set to 1; max_features, the number of features considered when splitting a node, was set to 2, and criterion, the function used to measure quality of a split, was set to the Gini Impurity Score. Additional parameters, including bootstrapping was set to 'true', cp. (complexity parameter) used for minimal cost-complexity pruning set to 0.01, and 5 K-folds with K-1 folds for training (Table 1).

Parameter	Value
n_estimators	150
max_depth	None
min_samples_split	2
min_samples_leaf	1
max_features	2
criterion	Gini Impurity Score
bootstrap	True
cp.	0.01
K-folds	5

Table 1. Random forest model parameters. 6 hyperparameters and 3 additional parameters are provided.

Keywords	Column 2	Column 3	Column 4
*muscle aches	uncomfortable	six	weeks
no sob	6	two months	*flutter
*marked	*unable	*cold	*chair
*increased	comfortable	*cough	*worse
*gym	*dull	*pass out	*rest
*swelling	2	months	two
*most	*some	two weeks	*tired
*any	six months	*activity	sob

Table 2. Keyword list. All 32 keywords are listed, 19 unique keywords indicated by (*).

Case	Average correct keywords	Average incorrect keywords
NYHA II	8.05	1.24
NYHA III	8.53	0.816
NYHA IV	7.73	0.662
Total	8.09 (+/- 2.84)	0.995 (+/- 0.998)

Table 3. Keyword averages. The average correct and incorrect keywords used by students by case. Totals are reported with a standard deviation.

Statistical analysis

Discussed data falls under the category of parametric data. Specifically, Poisson Distributed Variables, such as the correct word count range and wrong word count range; Binary Distributed Variables, such as the presence or lack of keywords; Categorical Nominal Variables, such as NYHA Classification. As appropriate, variables were summarized as mean and standard deviation (SD). Multiple comparisons were not made, thus corrections were not indicated. All percentages were rounded to three significant figures, while binary data was reported as assessed. Final model comparison was made by the misclassification rate, described as the sum of total false positives and total false negatives all divided by the total observations. Additional metrics, including the sensitivity, specificity, precision, recall, accuracy, F1 score, and Matthews correlation coefficient (MCC) will be reported; however, in the medical education setting we assume all errors to be equally important. Thus, the low resource and easily interpretable metric, misclassification rate, was selected.

Analyses were conducted using R-Studio 2024.12.0 + 467 and Python version 3.11.

Results

Keywords

57 initial keywords were identified, 25 with zero variability indicated a lack of use by any student regardless of the case or intended classification. These 25 variables were dropped to reduce dimensionality as they didn't provide any additional information for the models to improve upon. Additionally, highly correlated variables, such as the phrase "six months" were included as complete phrases and split on white-space as "six" and "months". 11 initial keywords became 14 after the split for Myocarditis (NYHA II), 6 became 9 after the split for Familial Hypertrophic Cardiomyopathy (NYHA II), 7 became 10 with the split for Sarcoidosis Cardiomyopathy (NYHA III), and 8 became 10 with the split for Ischemic Cardiomyopathy (NYHA IV). In total, 32 individual keywords were assembled, 19 of them unique (Table 2).

For those students who received a NYHA II case, the average correct keywords used were 8.05 out of the maximum 14 possible correct keywords, while the average incorrect were 1.24. In the NYHA III case, the average correct keywords were 8.53 out of 10, while the incorrect average was 0.816. For the NYHA IV case the average correct keywords were 7.73 out of 10, while the incorrect average was 0.662. Overall, the average correct keywords for all cases was 8.09 (SD +/- 2.84) and 0.995 (SD +/- 0.998) for incorrect words (Table 3).

A variable importance plot (VIP) was generated to demonstrate the importance of a given variable in regard to the overall integrity of the model; in other words, removing a variable from the model would result in an appreciable proportion of observations to be misclassified (Fig. 1). Figure 1a demonstrates the Mean Decrease Accuracy plot, which expresses the accuracy loss suffered by the model by excluding a particular variable. The correct word variable is the most important factor for Mean Decrease Accuracy. Similarly, Fig. 1b demonstrates the Mean Decrease Gini plot, which is a representation of the per variable contribution to node homogeneity, a measure of node purity. Here, the keyword "months" is most important for Mean Decrease Gini, likely due to the artificially assigned symptom duration within each of the set cases. These variables are presented in descending importance from top to bottom in each graph (Fig. 1).

Random forests

The generated Random Forest model consisted of 150 trees. Traditionally, increasing the trees in the forest imbues the model with more accuracy; however, the mean squared error in our model stabilizes around 50 and
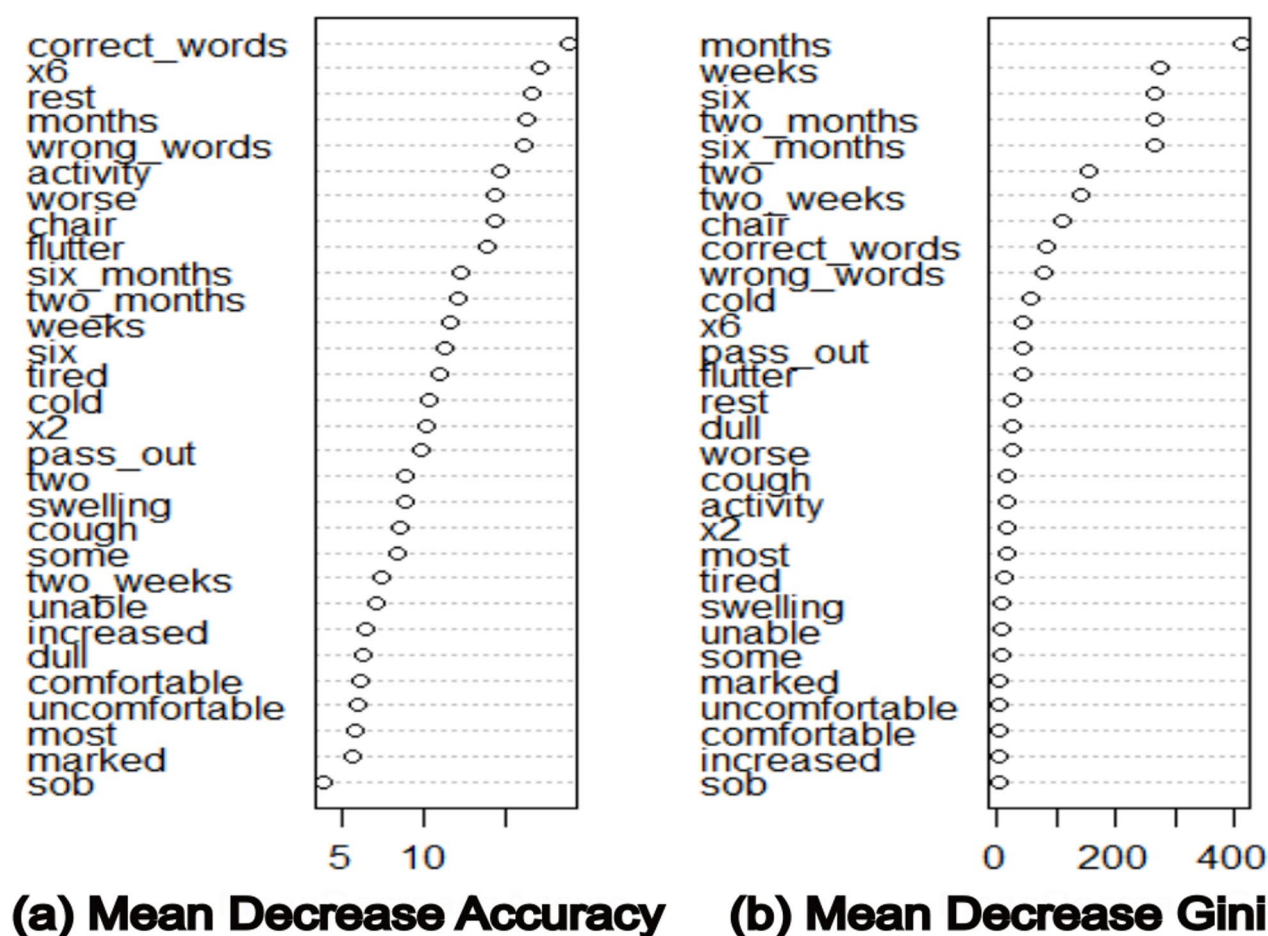
Fig. 1. Variable Importance Plot. (**a**) Mean Decrease Accuracy variable importance plot and (**b**) Mean Decrease Gini variable importance plot.

150 (Fig. 2). The lines in the figure represent each of the NYHA Classifications: red represents NYHA II, green NYHA III, blue NYHA IV, and black represents the data used strictly for testing. During training the model reported a 0% error rate in the identification of NYHA II cases, 1.01% error rate for NYHA III, and 0.552% error rate for NYHA IV (Table 4). When the model was assessed on the test data we found a 0.775% error rate for NYHA II, 26.3% for NYHA III, and 6.90% for NYHA IV (Table 5).

Figure 3 demonstrates a multidimensional scaling proximities plot, while Fig. 4 is focused on the convergence point of these axes. The plots illustrate model performance regarding the testing data set. From the graph we can see that there are three distinct lines representing each of the three classification groups, spaced equally in the XYZ coordinate plane, thus equally spaced within three-dimensions (Fig. 3). The Spatial distance amongst the groups indicates that the random forest model has clearly identified the NYHA classifications. The model performed accurately based solely on the keywords used by students, correct word percentage, and wrong word percentage. There is slight overlap between groups, observations from other classifications that bleed into the other group clusters, indicating misclassification (Fig. 4). The final model demonstrated a 0.420% misclassification rate on the bootstrap training set and an 8.20% misclassification rate on the testing set. Additional metrics include, sensitivity of 97.7%, specificity of 85.1%, precision of 88.1%, recall of 97.0%, accuracy of 91.8%, F1 score of 92.7%, and MCC of 0.840 (Table 6).

Discussion

This study led to the development of a 150-tree random forest model for the identification of NYHA Heart Failure Classifications II-IV in the unstructured textual data of medical student HPI entries rather than typical structured forms[15]. The model performed well with an 8.20% misclassification rate. Model development began
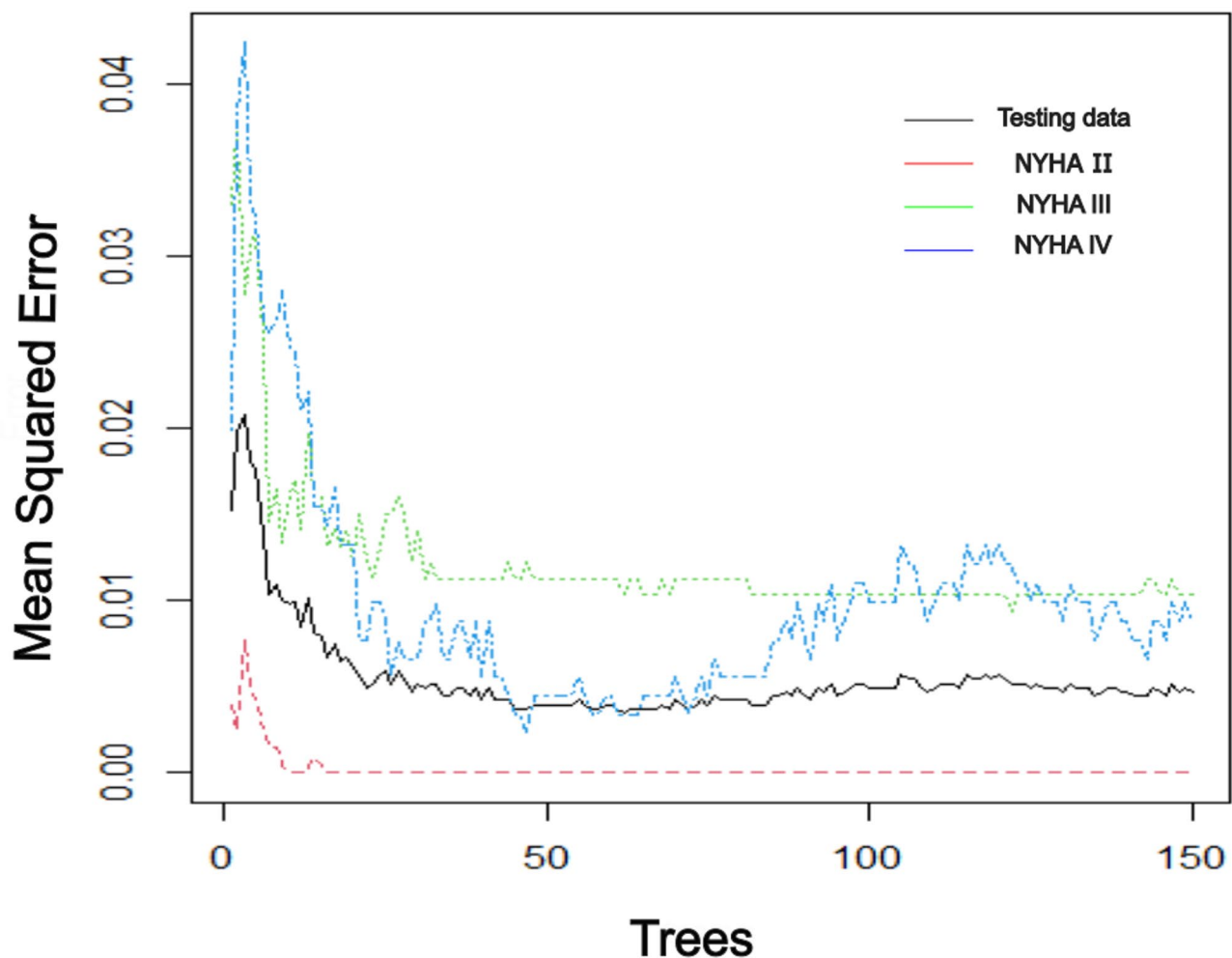
Fig. 2. Trees vs. error. Demonstrates the flux of mean squared error in respect to increasing the number of trees in the Random Forest model.

NYHA Classification	NYHA II Determined	NYHA III Determined	NYHA IV Determined	Error
NYHA II Expected	2080	0	0	0
NYHA III Expected	0	1054	**11**	0.0101
NYHA IV Expected	0	**8**	897	0.00552

Table 4. Random forests training set output. The expected counts for each case in comparison to the count determined by the random forests model using the training data set. Model misclassification counts are bolded.

NYHA Classification	NYHA II Determined	NYHA III Determined	NYHA IV Determined	Error
NYHA II Expected	126	**3**	0	0.00775
NYHA III Expected	**12**	45	0	0.263
NYHA IV Expected	**5**	0	52	0.0690

Table 5. Random forests test set output. The expected counts for each case in comparison to the count determined by the random forests model using the testing data set. Model misclassification counts are bolded.
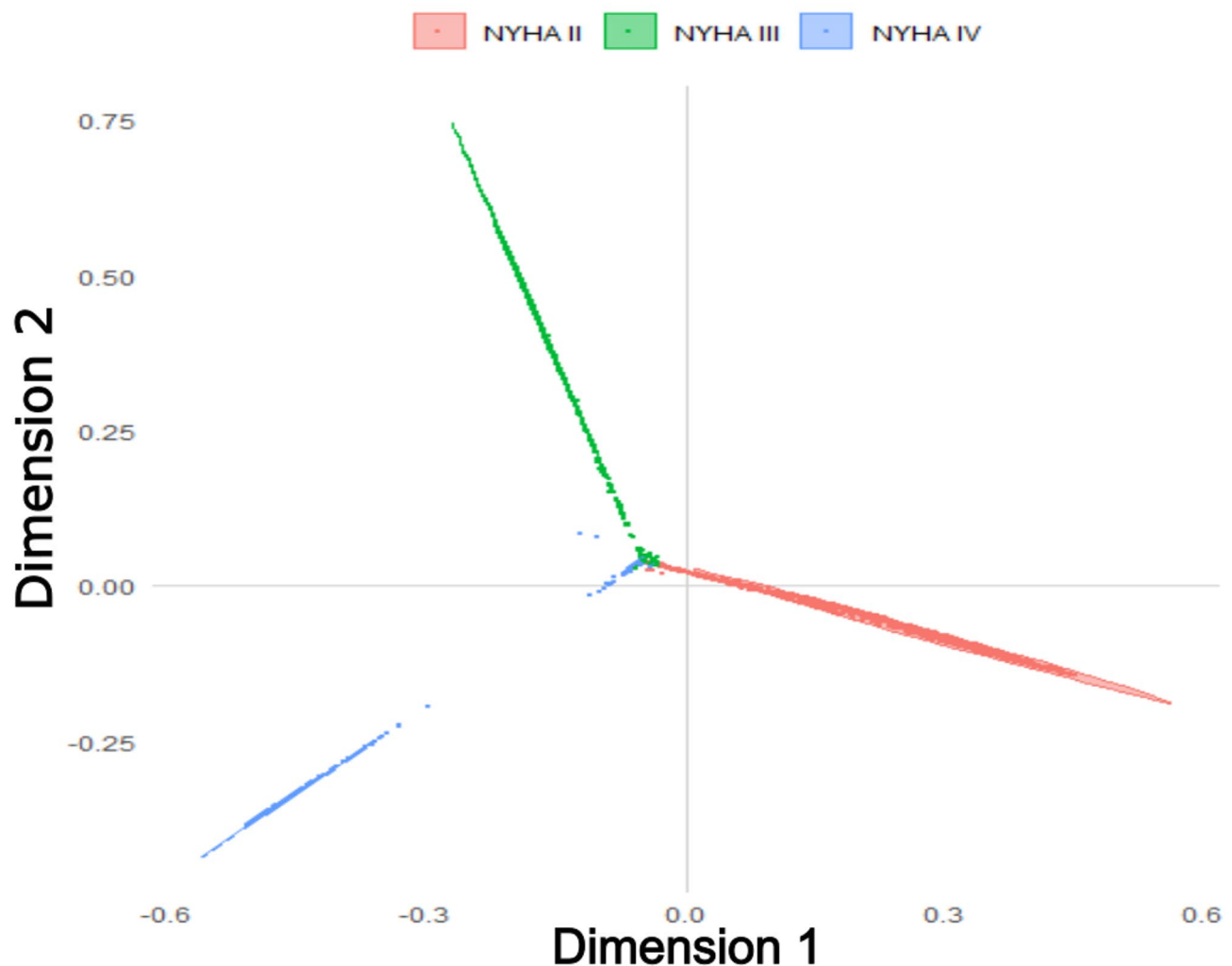
Multidimensional Scaling Proximities Plot

Fig. 3. Multidimensional scaling proximity plot.

with lexical analysis and keyword development (Table 2). Significantly, in the medical education setting, we administer the scenario in which students extract pre-written case information, these pre-written scenarios hinge on the use of very targeted keywords. The students then report their findings in a structured SOAP note format, with much of their effort concentrated in the unstructured HPI. This section, a clinical summary portraying the relevant details for the current illness, is comparable to the verbal presentation one may provide to a medical team.

Model characteristics were paramount in method selection including execution time, cost, accuracy, and resource consumption. Comparable studies utilized deep learning NLP models[13], rule-based NLP models[14], and decision trees[15]. Random forest models performed similarly with less computational overhead[15]. Novel ensemble learning techniques, such as those presented for k-nearest neighbors (kNN)[16], extended neighborhood rule (ExNRule)[17], and pruning strategies for random forest models[18], may have precluded the selection of a standard random forest approach; however, random forest models closely paralleled ensemble learning performance[18] with overall shorter execution times[17]. As our model was expected to provide instantaneous feedback upon text file submission, limiting execution times was predominant. Limited funding further precluded our team from utilizing other model types, as large language models (LLMs) typically function on a cost-per-token system and the graphics processing unit (GPU) market was suboptimal. Importantly, LLMs and other machine learning (ML) methods can be trained to outperform random forest models; however, they require more training samples and consume more computational power, thus necessitating costly GPU upgrades or managed LLM providers.

Misclassification rate provides a simple interpretation of accuracy without delving into specific false positive or false negative determinations. This fidelity is necessary in disease prediction and screening; however, in the medical education setting class determination and keyword evaluation is sufficient. Overall performance is
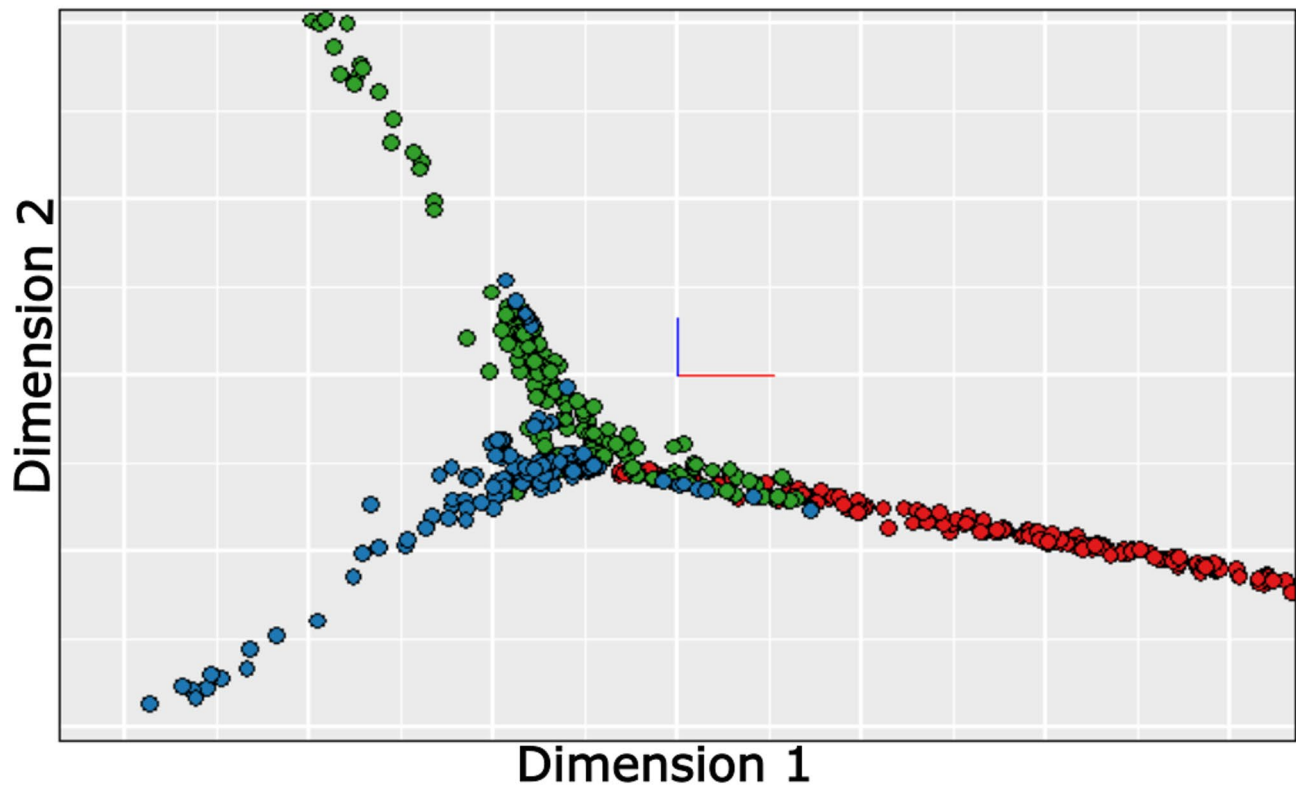
Multidimensional Scaling Proximities Plot

Fig. 4. Zoomed multidimensional scaling proximity plot.

Metrics	Value
Misclassification rate	0.0820
Sensitivity	0.977
Specificity	0.851
Precision	0.881
Recall	0.970
Accuracy	0.918
F1 score	0.927
MCC	0.840

Table 6. Random forest model metrics. MCC or Matthews correlation coefficient.

comparable to the literature, including the disproportionate misclassification of NYHA II and III cases[13–15], which underscores the challenges of relying on keywords[19]. Additionally, the model suffers from a paucity of training data, as noted previously 203 of the original cases were NYHA II, 106 NYHA III, and just 96 NYHA IV. Many comparable studies with better accuracy either included a larger initial dataset or a larger fold cross validation[13–15].

Determining the optimal tree count is an important aspect of balancing the computational strain of a random forest model and directly affects accuracy. As demonstrated in Fig. 3, a maximum default of 150 trees was selected; however, 50–75 trees performed similarly. This de-escalation would result in a significant diminution in the computational costs of the model due to the reduction in variables required for prediction[20].

Keywords are critical for deciphering unstructured textual data; however, traditionally keywords solely aided a random forest model in its determinations. Our study begins to assess the use of these keywords in each case, building towards the development of efficient, yet accurate case presentations or medical decision making[3]. Similar computational models have been used in the academic setting to support standardized scoring platforms for medical education[9]. However, none have been poised to provide lexical feedback to students.

Here, we demonstrated an average of 8.09 correct keywords in comparison to 0.995 incorrect keywords used by students throughout the cases. This demonstrates a tendency for the students to heuristically identify

keywords (Table 2), while self-assigning keyword importance. The students are likely demonstrating an ability to surmise the veiled VIP (Fig. 1) through their studies, preparation materials, and standardized patient visits. Notably, the development of a VIP could better assist educators in developing a more lexically-based assessment system that gives weighted credit to the use of keywords with a higher mean decrease accuracy; therefore, increasing model accuracy and maximizing the effectiveness of the medical case presentation.

The limitations of the study stem from the initially limited sample size, unbalanced distribution of NYHA Classification cases, and the lack of NYHA I cases. Additionally, the available equipment was rudimentary, limiting our bootstrap iterations to 10 and our fold cross validation to 5. Improvements in our available equipment would have further increased the training and testing sample sizes, and allowed for the employment of a novel ensemble learning technique.

Conclusion

Random forest models can be applied with acceptable accuracy in the determination of NYHA Classifications based solely on the free-text HPI from medical student standardized patient simulated cases. Our model demonstrated a comparably low 0.420% and 8.20% misclassification rate, for training data and testing data respectively. These scores would likely improve with an increased volume of training cases and utilization of more purpose-built computational equipment. Continued improvement of this model would comfortably supplement an educator in assessing medical students on their performance in simulated cases. Future implementations of these findings may lead to the development of an instantaneous feedback system for students to receive constructive critique regarding their medical decision making and overall medical lexical density.

Data availability

The textual data analysed during the study are available in the following GitHub repository: https://github.com/tDaniels1701/Random-Forest-HPI-Classification.git.

References

1. Zhang, R. et al. Discovering and identifying new York heart association classification from electronic health records. *BMC Med. Inf. Decis. Mak.* **18** (Suppl 2), 48 (2018).
2. Amir, A. et al. AALIM: a cardiac clinical decision support system powered by advanced multi-modal analytics. *Stud. Health Technol. Inf.* **160** (Pt 2), 846–850 (2010).
3. Elbattah, M., Arnaud, É., Gignon, M. & Dequen, G. The role of text analytics in healthcare: a review of recent developments and applications. *Int. Conf. Health Inf.* **5**, 825–832 (2021).
4. Krishnan, G. S. & Kamath, S. S. A supervised learning approach for Icu mortality prediction based on unstructured electrocardiogram text reports. *Nat. Lang. Process. Inf. Syst.* **10859**, 126–134 (2018).
5. Sugamiya, Y., Otani, T., Nakadate, R. & Takanishi, A. Construction of automatic scoring system to support objective evaluation of clinical skills in medical education. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2019**, 4177–4181 (2019).
6. Chintalapudi, N., Battineni, G., Canio, M. D., Sagaro, G. G. & Amenta, F. Text mining with sentiment analysis on seafarers' medical documents. *Int. J. Inf. Manag Data Insights*. **1** (1), 1–9 (2021).
7. Hill, N. R. et al. Predicting atrial fibrillation in primary care using machine learning. *PLoS ONE.* **14** (11), e0224582 (2019).
8. Krittanawong, C., Zhang, H., Wang, Z., Aydar, M. & Kitai, T. Artificial intelligence in precision cardiovascular medicine. *J. Am. Coll. Cardiol.* **69** (21), 2657–2664 (2017).
9. Savalia, S. & Emamian, V. Cardiac arrhythmia classification by multi-layer perceptron and Convolution neural networks. *Bioeng. (Basel).* **5** (2), 35 (2018).
10. Li, J., Si, Y., Xu, T. & Jiang, S. Deep convolutional neural network based Ecg classification system using information fusion and one-hot encoding techniques. *Math. Probl. Eng.* **7354081**, 1-10 (2018).
11. Sutton, R. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* **3**, 17 (2020).
12. Caraballo, C. et al. Clinical implications of the new York heart association classification. *J. Am. Heart Assoc.* **8** (23), e014240 (2019).
13. Adejumo, P. et al. Natural Language processing of clinical Documentation to assess functional status in patients with heart failure. *JAMA Netw. Open.* **7** (11), e2443925 (2024).
14. Zhang, R. et al. Automatic methods to extract new York heart association classification from clinical notes. *Proc. (IEEE Int. Conf. Bioinf. Biomed).* **1296-1299**, 2017 (2017).
15. Jandy, K. & Weichbroth, P. A machine learning approach to classifying new York heart association (NYHA) heart failure. *Sci. Rep.* **14** (1), 11496 (2024).
16. Gul, N., Mashwani, W. K., Aamir, M., Aldahmani, S. & Khan, Z. Optimal model selection for k-nearest neighbors ensemble via sub-bagging and sub-sampling with feature weighting. *Alexandria Eng. J.* **72**, 157–168 (2023).
17. Ali, A. et al. A k nearest neighbour ensemble via extended neighbourhood rule and feature subsets. *Pattern Recognit.* **142**, 109641 (2023).
18. Khan, Z., Ali, A., Khan, D. M. & Aldahmani, S. Regularized ensemble learning for prediction and risk factors assessment of students at risk in the post-COVID era. *Sci. Rep.* **14** (1), 16200 (2024).
19. Raphael, C. et al. Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. *Heart* **93** (4), 476–482 (2007).
20. Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. How many trees in a random forest? *Mach. Learn. Data Min. Pattern Recognit. (MLDM).* **7376**, 154–168 (2012).

Acknowledgements

Author contributions

I.R.P. and T.D. conceived the study. I.R.P., T.D., F.A.R., J.L., and K.G. collected and produced the data. T.D. gen-

erated the model and performed the statistical analysis. I.R.P. wrote the manuscript. I.R.P. and T.D. prepared the figures and tables. I.R.P., T.D., F.A.R., J.L., and K.G. critically revised and approved the final manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to I.R.P.

Reprints and permissions information is available at www.nature.com/reprints.

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