Patient Classification Based on Symptoms Using Machine Learning Algorithms Supporting Hospital Admission

Khoa Dang Dang Le[1], Huong Hoang Luong[2], and Hai Thanh Nguyen[3(✉)]

[1] Hospital Information System Team, Vietnam Posts and Telecommunications Group, Tien Giang, Vietnam
[2] FPT University, Can Tho, Vietnam
[3] College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam
nthai.cit@ctu.edu.vn

Abstract. Overcrowding in receiving patients, medical examinations, and treatment for hospital admission is common at most hospitals in Vietnam. Receiving and classifying patients is the first step in a medical facility's medical examination and treatment process. Therefore, overcrowding at the regular admission stage has become a complex problem to solve. This work proposes a patient classification scheme representing the text to speed up the patient input flow in hospital admission. First, the Bag of words approach has been built to represent the text as a vector exhibiting the frequency of words in the text. The data used for the evaluation were collected from March 2016 to March 2021 at My Tho City Medical Center - Tien Giang - Vietnam, including 230,479 clinic symptom samples from admissions and discharge office, outpatient department, accident, and Emergency Department. Among learning approaches used in the paper, Logistic Regression reached an accuracy of 79.1% for stratifying patients into ten common diseases in Vietnam. Besides, we have deployed a model explanation technique, Locally Interpretable Model-Agnostic Explanations (LIME), to provide valuable features in disease classification tasks. The experimental results are expected to suggest and classify the patient flow automatically in the hospital admission stage and discharge office to perform the patient flow in the clinics at the hospitals.

Keywords: Hospital admission · Bag of words · Explanation · Clinic symptom · Disease classification

1 Introduction

Receiving and flowing patients into hospital clinics is a rather complicated matter. In [1], the authors stated that "long waiting times became an important

P. Cong Vinh and N. Huu Nhan (Eds.): ICTCC 2021, LNICST 408, pp. 40–50, 2021.
https://doi.org/10.1007/978-3-030-92942-8_4

part of the perception of a general *health care crisis*". For most hospitals in Vietnam, patients have to wait for a long time[1] for hospital admission procedures. Although some solutions have been proposed to reduce and pre-arrange appointments with the doctors'[2], almost all people still tend to crow at the hospital for their health care services. In hospitals, the idea of designing arrangements and suggestions for flowing patients to the medical clinics accurately and quickly helps to increase work efficiency. Usually, in the morning in most hospitals, patients wait for reservations, especially in outpatient departments, accident, and emergency departments, receiving and flowing patients into appropriate hospital clinics to help patients avoid fatigue and minimize conflicts between patients and between patients and medical staff. Receiving patients and flowing patients to clinics in the hospital to diagnose diseases corresponding to signs and clinical symptoms helps avoid wasting time on medical examination and treatment of doctors and patients. Because if the patient's flow is not accurate, the doctor must operate to transfer the patient's room, which is both laborious and time-consuming for the doctor and the patient.

In recent years, the development of machine learning in health care has made significant progress. As a result, machine learning applications have a wide application area, especially supporting and comprehensive care of people's health. For example, numerous applications are related to personal health monitoring [2–8] to reduce disease risk, early detection of modern diseases such as cancer, cardiovascular disease, help reduce costs and prolong life, help patients comply with medication, monitor disease progression. Machine learning is applied in critical areas of medicine: in clinical decisions, electronic health records, diagnostics, medical robots, personalized medicine, medical examination, and treatment management. The progressive application of machine learning in healthcare can improve access to healthcare and the quality of healthcare.

In this work, to develop an automatic patient classification based on symptoms in the text to speed up admission at hospitals, we propose leveraging the Bag of words (BOW) model primarily to generate word features from the text. After converting the text into BOW, we have deployed Term Frequency - Inverse Document Frequency (TF-IDF) [9] to determine the importance of a word in the text to extract the features which are input for machine learning algorithms to perform the classification tasks. Then, we use machine learning algorithms to diagnose ten diseases based on the set of symptoms. After obtaining the learning from models, we have proceeded with an explanation model technique, Local Interpretable Model-agnostic Explanations (LIME) [10], to extract insights in data, including keywords for disease diagnosis. Finally, the experiments are evaluated on the dataset, including 33,024 health record samples at the Health Center of My Tho City - Tien Giang - Vietnam, with ten common diseases in Vietnam.

[1] http://baodongnai.com.vn/xahoi/201702/thoi-gian-cho-kham-chua-benh-con-qua-lau-2781882/index.htm, accessed on 01 May 2021.

[2] https://nhandan.vn/tin-tuc-y-te/thi-diem-dat-lich-truc-tuyen-rut-ngan-nua-thoi-gian-kham-chua-benh--643125/, accessed on 10 May 2021.

The rest of this article is as follows. Section 2, we introduce some related work. Section 3, we represent the data for the experiment. Section 4 presents and illustrates our proposed method while Sect. 5 reveals, evaluate the experimental results. The conclusions are rendered in Sect. 6.

2 Related Work

One of the challenges in hospitals today is to make the right decisions to flow patients from the admissions and discharge office to the clinics in the hospital and reduce patient waiting times. Therefore, using machine learning algorithms in the analysis and support of medical staff to perform patient flowing brings more satisfactory results. Furthermore, advances in machine learning algorithms and big data have allowed artificial intelligence to replace humans gradually.

The scientists have proposed a vast amount of research on proposing early diagnosis, improving patient flow in hospitals, and achieving positive results. For example, the study in [11] proposed to improve streaming patients in a hospital emergency department.

Numerous studies have attracted patient classification tasks based on symptoms. In [12], the authors performed quantitative exploration of symptoms using Chi-square test, association with LASSO for analysis to conduct the symptoms of COVID-19 patients. The study in [13] analyzed the Parent-Reported Symptoms to propose efficient treatment for children. The authors in [14] evaluated and experimented on voice features for the Dysphonia-based classification of Parkinson's Disease. The authors in [15] provided a living with uncertainty method that mapped the transition from pre-diagnosis to a diagnosis of dementia. The work in [16] deployed the (TF/IDF), Bag of words (BOW) to analyze symptoms on COVID-19 clinical text data. Finally, the work in [17] applied standard text categorization methods for exploring patient text. The proposed aimed to predict treatment outcomes in Internet-based cognitive-behavioral therapy.

With the BOW and TF/IDF achievements, our work has deployed these approaches to analyze the patient clinical symptoms of 10 diseases in the text at the Health Center of My Tho City - Tien Giang, Vietnam. We combine the obtained vector with meaningful attributes in the patient's medical examination and treatment data to build a model to classify patients into common diseases for hospital admission.

3 Dataset Description

This research examined the PatientAdmission dataset, with 230,479 samples containing the information: age, gender, clinical symptoms of the patient to propose building a machine learning model, which supports the implementation of automatic classification diseases and decision to flow patients to hospital clinics. Data were collected from March 2016 to March 2021 from the Medical Center of My Tho City - Tien Giang - VietNam, from the admissions and discharge office, outpatient department, accident, emergency department, and related reports.

Table 1. Information on the considered dataset with the number of samples

No.	Disease name	Total of samples
1	Neoplasms	16271
2	Endocrine, Nutritional and metabolic diseases	38672
3	Diseases of the eye and adnexa	18443
4	Diseases of the circulatory system	37782
5	Diseases of the respiratory system	41888
6	Diseases of skin and subcutaneous tissue	7044
7	Diseases of the musculoskeletal system and connective tissue	35427
8	Diseases of the genitourinary system B212	17503
9	Pregnancy, childbirth and the puerperium	3666
10	Injury, poisoning and certain other consequences of external causes	13783

Data are manually or semi-automatically retrieved patient information (the QRCode in the health insurance card). The dataset includes patient's age (AGE field), patient's gender 1 Male, 0 Female (SEX field), and CLINICAL SYMPTOMS field. Clinical symptoms were noted by the medical staff who received the patient when the patient was declared at the admission and discharge office. The patient's clinical symptom data may contain information about physical fitness, abnormal vital signs, symptoms, and manifestations before and during hospital arrival. ID_DISEASES field is about a patient's type disease that is recognized after being examined, treated by a doctor, and identified with the patient's ICD10- disease code, a group of diseases that includes many diseases, a type disease that includes many groups of diseases. In the table of data, there are ten considered common type diseases which were recorded in Vietnam hospitals (exhibited in Table 1).

4 Methods

The overall architecture for proposed hospital admissions is exhibited in Fig. 1. The data are extracted from medical information systems[3] at the Health Center of My Tho City - Tien Giang, Vietnam and divided into training and test set. Feature engineering is performed with the BOW technique on the training set to provide a word vector containing a dictionary that includes words extracted from clinical symptoms and signs described in pre-diagnosis records when the patients are admitted into the hospital. Then, we use the TF-IDF [9] method to calculate the value for features which are words included in the BOW extracted from the training data. Finally, we use the training set is fetched into learning algorithms to build a model for evaluating the test set. Details of methods are introduced in the following sections.

[3] https://hotrovnpthis.wordpress.com/, accessed on 01 May 2021.

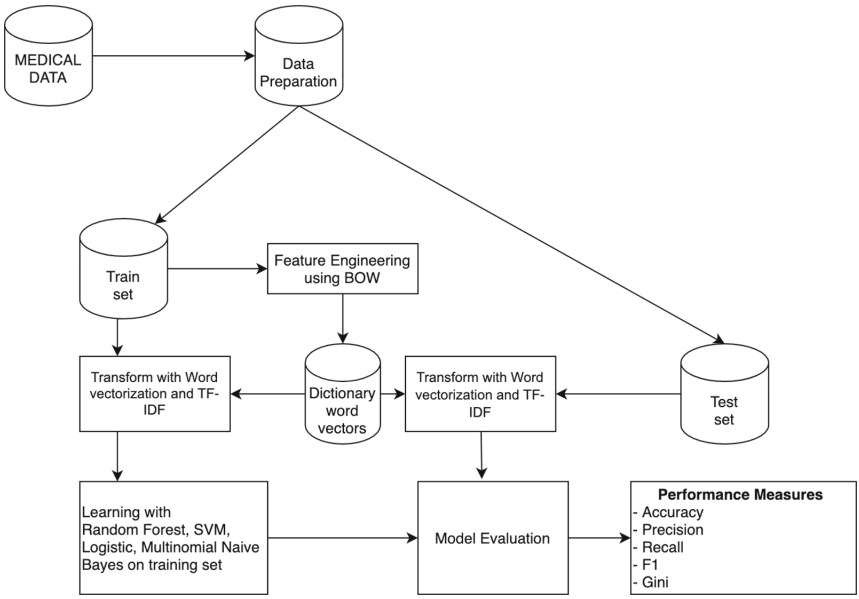

Fig. 1. The proposed architecture for patient classification based on symptoms description

The BOW model is a wording technique implemented in natural language processing. More specifically, text representation is the Bag of words. In this model, BOW disregarding grammar but keeping multiplicity [1]. Let us consider the following example:

(1) Jong has a bad headache. Ronaldo also has a bad headache.
(2) Nadal is going to the doctor for a headache

Considering these texts, we see that a list of words is generated as follows for each text:

"Jong ", "has ","a", "bad", "headache", "Ronaldo ", "also"

"Nadal","is", "going", "to", "the", "doctor", "for", "a", "headache"

Representing each BOW as a JSON object as follows:

B1 = "Jong":1,"has":2,"a":2,"bad":2,"headache":2,"Ronaldo":1,"also":1

B2 = "Nadal":1,"is":1,"going":1,"to":1,"the":1,"doctor":1,"for":1,"a":1, "headache":1

In order to represent text, each word can be represented as the key of the BOW model, while it corresponds to the value of the number of occurrences and the order of the words that can be ignored:

"Jong":1,"has":2,"Ronaldo":1,"also":1,"a":2,"bad":2,"headache":2 also B1.

When the dataset is a few hundred documents, we can see that the dictionary can range from a few tens of thousands of words to a few hundred thousand words. Therefore, the number of dimensions obtained after transforming BOW

is vast. Some machine learning models like Naive Bayes handle it inefficiently. To overcome, we can use the method of reducing the number of data dimensions. The reduction method can select the most important words to distinguish one text from another or the dimensionality reduction method. However, this reduction step often causes information loss, reducing the accuracy of the classifier after implementation.

We transformed the text with Term Frequency - Inverse Document Frequency (TF-IDF) which is considered as a robust numerical statistic tool to reveal essential words in a text document[9].

After applying such these approaches with BOW and TF-IDF, we fetch transformed into learning algorithms such as Random Forest, Support Vector Machines, Logistic Regression, and Multinomial Naive Bayes with hyper-parameters as presented in Table 2. In comparison with various learning methods, we use accuracy, precision, recall, f1-score, and Gini score.

Table 2. Information of hyper-parameters use in the training phase

Machine learning algorithm	Hyper-parameters
Logistic Regression	C: 1.0, penalty: l1
SVM	C: 10, gama: 0.0001, kernel: rbf
Decision Tree	Criterion: gini, max depth: 450
Multinomial Naive Bayes	Default value: Laplace/Lidstone smoothing value of 1.0
Random Forest	Criterion: gini, max depth: 450, max features: auto, 750 trees

5 Experimental Results

5.1 Setting

This study used a PC Windows 10 Home Single Language 64 bit with a CPU of 28 cores Intel(R) Xeon(R) CPU E5 - 2680 v4 @ 2.4 GHz and 96 GB of RAM that is a sufficiently responsive hardware environment to test the proposed algorithm. The experiment code for all the models is available through python that supports several libraries, such as Sklearn [18], Matplotlib [19], and Pandas [20,21].

5.2 Disease Classification Task with Various Algorithms

In Table 3, we compare performance among considered classifiers. As observed, Logistic Regression reveals the best result while Multinomial Naive Bayes exhibits the worst one.

In the above analysis, using the Logistic Regression algorithm to achieve the best results with measures Accuracy = 79.1%, Precision = 79.9%, Recall = 79.1%, F1 = 79.5%, Gini = 92.1% and time 159 s. Multinomial Naive Bayes algorithm

Table 3. The disease classification performance performed by various learning approaches. The **bold** results are the best results among the considered learning algorithms

Method	ACC (%)	Precision (%)	Recall (%)	F1 (%)	Gini(%)	Time (s)
RandomForest	78.6	79.3	78.6	78.9	89.7	9009
SVM	78.0	78.7	78.0	78.3	88.8	237
Decision Tree	74.0	74.9	74.0	74.4	70.6	481
Multinomial Naive Bayes	72.7	75.1	72.7	73.9	88.7	46
Logistic Regression	**79.1**	**79.9**	**79.1**	**79.5**	**92.1**	159

achieved the lowest results with the measures Accuracy = 72.7%, Precision = 75.1%, Recall = 72.7%, F1 = 72.5%, Gini = 88.7% and time 246 s.

The algorithms SVM, Random Forest gives results with a reasonably high Accuracy measure of more than 78%. The algorithm's Decision Tree gives results with the measures accuracy = 74%.

5.3 Analysis of Disease Symptoms

Some interesting analyzes from the dataset when using the LIME technique to analyze important words affecting type disease classification performed by Logistic Regression shown in Figs. 2 and 3.

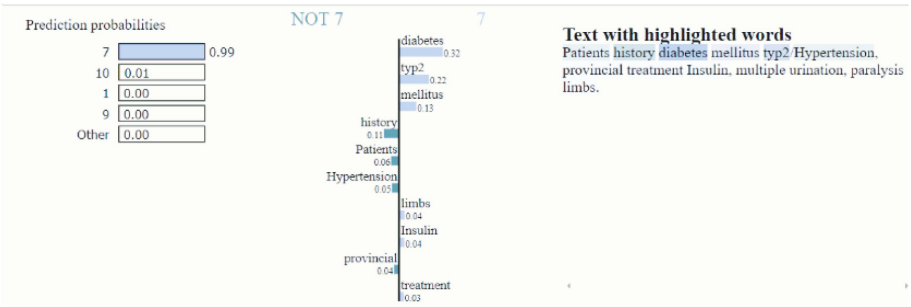

Fig. 2. Use LIME to Explain the prediction of Logistic Regression and display the words symptoms description that can be key word to describe the disease

We extract the essential features in the above classification model, important information that dramatically affects the patient classification process is shown in the patient's blood pressure and temperature measurements, this shows the possibility of when measuring body temperature, measuring blood pressure when asking the patient at the admissions and discharge office, the model predicts the possibility of having the disease.

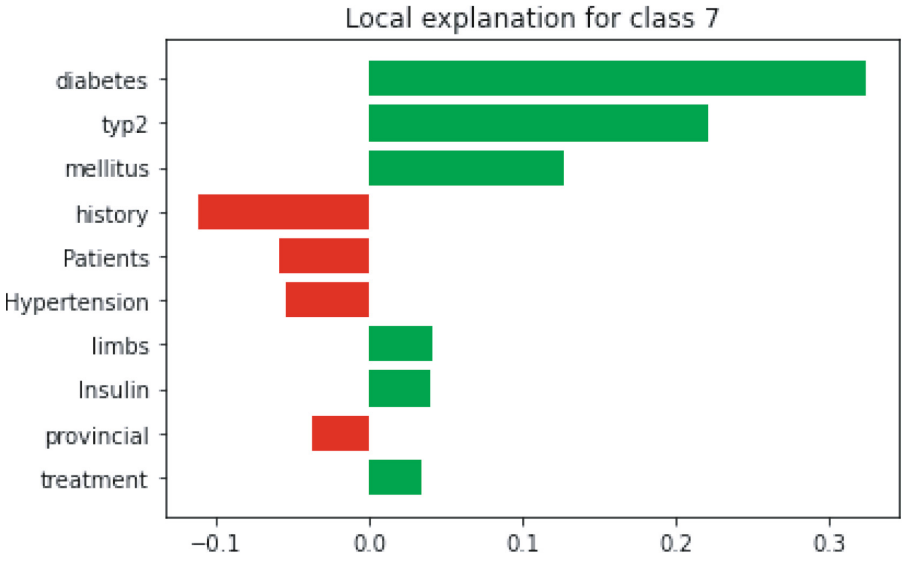

Fig. 3. Use LIME to Explain compare the importance of words affecting data

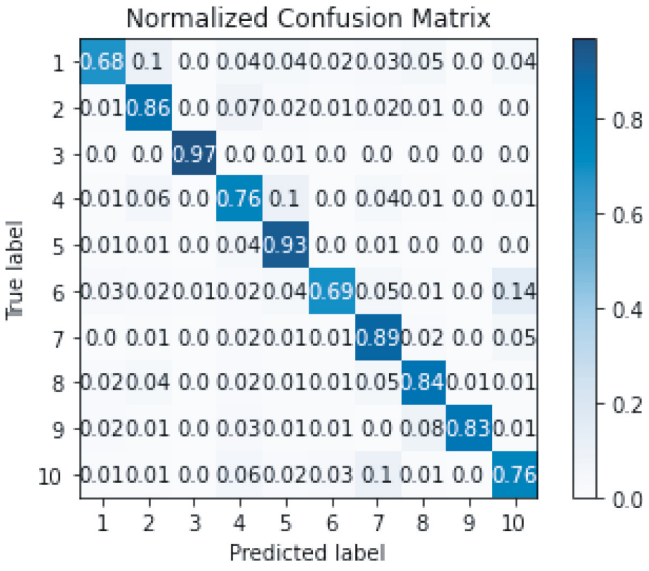

Fig. 4. Confusion matrix of prediction results with Logistic Regression

We reveal the confusion matrix in Fig. 4 of Logistic Regression, where we obtain the best results. Because we have an imbalanced dataset, the classes are not represented equally. The model is perfect for classes: 3 and 5 with more than

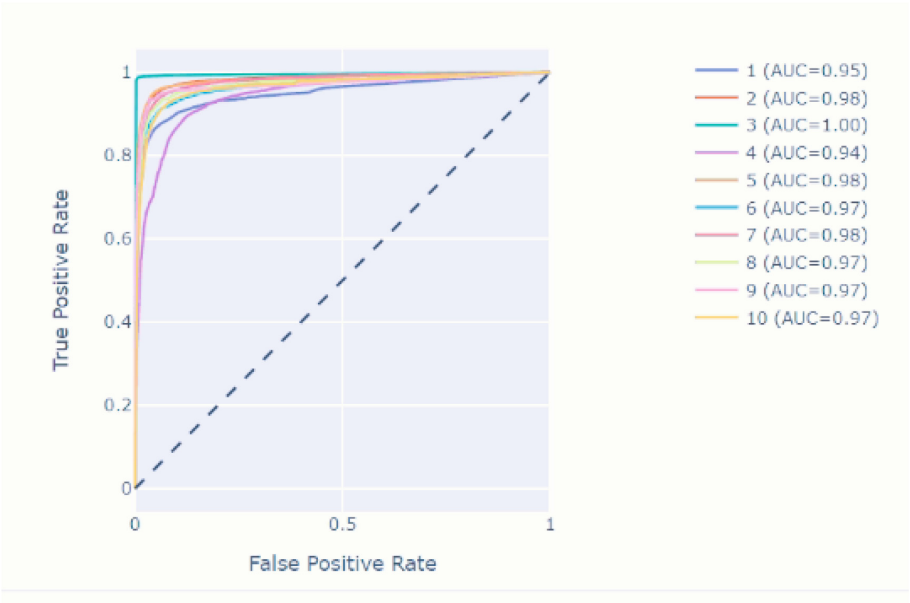

Fig. 5. ROC curve

90% accuracy, and suitable for classes: 2, 7, 8, 9 with more than 80% accuracy, not suitable for class 1 with 68% accuracy (the indexes of classes correspond to No. in Table 1).

Figure 5 illustrates experimental results of Logistic Regression in another metric, ROC curve. The curve of 10 diseases is above the diagonal, demonstrating that the model is good.

6 Conclusion

We have presented a method to solve automatic patient input flowing. Our research is based on a combination of representing text by the BOW model and data classification algorithms. The BOW model is built quickly to represent the text as a vector of occurrence frequency of words in the text with the number of dimensions is very large. Incorporating other patient-specific information allows efficient classification of this data set. The proposed approach provides an automated way of patient flow in the hospital to improve healthcare in the future.

Acknowledgement. To conduct this study, our team would like to express our gratitude to the Hospital Information System team in Vietnam Posts and Telecommunications Group, Tien Giang province, Vietnam, for providing valuable data for our work.

References

1. Ringard, Å., Hagen, T.P.: Are waiting times for hospital admissions affected by patients' choices and mobility? BMC Health Serv. Res. **11**(1) (2011). https://doi.org/10.1186/1472-6963-11-170
2. Sajedi, S.O., Liang, X.: Uncertainty-assisted deep vision structural health monitoring. Comput.-Aided Civ. Infrastruct. Eng. **36**(2), 126–142 (2020). https://doi.org/10.1111/mice.12580
3. Valsalan, P., Baomar, T.A.B., Baabood, A.H.O.: IOT based health monitoring system. J. Critic. Rev. **7**(04), 739–743 (2020). https://doi.org/10.31838/jcr.07.04.137
4. Dong, C.Z., Catbas, F.N.: A review of computer vision-based structural health monitoring at local and global levels. Struct. Health Monit. **20**(2), 692–743 (2020). https://doi.org/10.1177/1475921720935585
5. Nasiri, S., Khosravani, M.R.: Progress and challenges in fabrication of wearable sensors for health monitoring. Sens. Actuat. A Phys. **312**, 112105 (2020), https://doi.org/10.1016/j.sna.2020.112105
6. Li, C., Sun, L., Xu, Z., Wu, X., Liang, T., Shi, W.: Experimental investigation and error analysis of high precision FBG displacement sensor for structural health monitoring. Int. J. Struct. Stab. Dyn. **20**(06), 2040011 (2020). https://doi.org/10.1142/s0219455420400118
7. Kim, J., et al.: Self-charging wearables for continuous health monitoring. Nano Energy **79**, 105419 (2021), https://doi.org/10.1016/j.nanoen.2020.105419
8. Chen, Z., Sheng, H., Xia, Y., Wang, W., He, J.: A comprehensive review on blade tip timing-based health monitoring: status and future. Mech. Syst. Sig. Process. **149**, 107330 (2021), https://doi.org/10.1016/j.ymssp.2020.107330
9. Uther, W., et al.: TF-IDF. In: Encyclopedia of Machine Learning, pp. 986–987. Springer, Boston (2011). https://doi.org/10.1007/978-0-387-30164-8_832
10. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (2016)
11. Medeiros, D.J., Swenson, E., DeFlitch, C.: Improving patient flow in a hospital emergency department. In: 2008 Winter Simulation Conference, pp. 1526–1531 (2008)
12. Qu, G., et al.: A quantitative exploration of symptoms in COVID-19 patients: an observational cohort study. Int. J. Med. Sci. **18**(4), 1082–1095 (2021). https://doi.org/10.7150/ijms.53596
13. Molloy, M.A., et al.: Parent-reported symptoms and perceived effectiveness of treatment in children hospitalized with advanced heart disease. J. Pediatr. (2021). https://doi.org/10.1016/j.jpeds.2021.06.077
14. Goyal, J., Khandnor, P., Aseri, T.C.: A comparative analysis of machine learning classifiers for dysphonia-based classification of parkinson's disease. Int. J. Data Sci. Anal. **11**(1), 69–83 (2020). https://doi.org/10.1007/s41060-020-00234-0
15. Campbell, S., et al.: Living with uncertainty: mapping the transition from prediagnosis to a diagnosis of dementia. J. Aging Stud. **37**, 40–47 (2016). https://doi.org/10.1016/j.jaging.2016.03.001
16. Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R., Rouf, N., Mohi Ud Din, M.: Machine learning based approaches for detecting COVID-19 using clinical text data. Int. J. Inf. Technol. **12**(3), 731–739 (2020). https://doi.org/10.1007/s41870-020-00495-9

17. Gogoulou, E., Boman, M., Ben Abdesslem, F., Hentati Isacsson, N., Kaldo, V., Sahlgren, M.: Predicting treatment outcome from patient texts: the case of Internet-based cognitive behavioural therapy. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 575–580. Association for Computational Linguistics, April 2021. https://aclanthology.org/2021.eacl-main.46
18. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
19. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007)
20. Pandas Development Team: pandas-dev/pandas: Pandas, February 2020. https://doi.org/10.5281/zenodo.3509134
21. Wes McKinney: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010)