

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/351956430>

Using Patient Descriptions of 20 Most Common Diseases in Text Classification for Evidence-based Medicine

Conference Paper · July 2021

DOI: 10.1109/MAJICC53071.2021.9526252

CITATIONS

2

READS

294

3 authors:



Samiullah Shaikh

Mohammad Ali Jinnah University

1 PUBLICATION 2 CITATIONS

SEE PROFILE



Muhammad Yaseen Khan

Alexa Translations

29 PUBLICATIONS 150 CITATIONS

SEE PROFILE




Muhammad Suffian Nizami


Università degli Studi di Urbino "Carlo Bo"


20 PUBLICATIONS 66 CITATIONS

SEE PROFILE

Using Patient Descriptions of 20 Most Common Diseases in Text Classification for Evidence-based Medicine

Samiullah Shaikh
Center for Language Computing,
Mohammad Ali Jinnah University
Karachi, Sindh, Pakistan
 0000-0001-8611-7389

Muhammad Yaseen Khan
Center for Language Computing,
Mohammad Ali Jinnah University,
and R&D, Love For Data,
Karachi, Sindh, Pakistan
 0000-0002-9049-8492

Muhammad Suffian Nizami
Dept. of Pure and Applied Science,
University of Urbino 'Carlo Bo'
Urbino, PU, Italy
 0000-0002-1946-285X

Abstract—Evidence-based Medicine (EBM) reflects a combination of clinical expertise, patient's values, and best available evidence in the decision-making process related to healthcare. In EBM, the medical professional prescribe medicine based on information from previous medical records (which is available in textual format). This information is often used in clinical practice and recently proved to be very useful in predicting diseases with computational approaches. This paper presents an extensive dataset of 11.8K patient descriptions of the most common 20 diseases, and contribute to their classification through unpre-tentious supervised machine learning techniques. After rigorous experiments under the Monte Carlo method, we found Random Forest Trees (RFT) outperformed all algorithms by achieving the overall highest accuracy of 83%, followed by Linear-Support Vector Machines (SVM) with 81% accuracy.

Index Terms—Disease Classification, Evidence-based Medicine, Machine Learning, Natural Language Processing, Patient Descriptions

I. INTRODUCTION

Many medical interventions are associated with significant adverse effects and patients may be harmed if they receive the wrong intervention or do not receive the right intervention. The first reason is that most physicians and machines/robots are not able to formulate a valid question because of the change in the terms due to the many variables. The second possible reason can be the less awareness of technology to doctors i.e. how to search/retrieve the information from the corpus? [1,2]; and use these descriptions with computational approaches for disease classification.

In the field of health sciences, Evidence-Based Medicine (EBM) emerge in the early 1990s from the work of British epidemiologists such as Archie Cochrane within the discipline of clinical epidemics [3]. Around 1991, the term made its first appearance in the medical literature. However, it was not until the succeeding year that the concept was first fully articulated by the EBM working group [4]; and the they claimed that the EBM is a new way to change the profession of medicine [5].

Beyond the limits of clinical treatments, we have observed from the literature, EBM has a great impact on the medical research and pharmaceutical industry [6,7]. In the world of today, surprisingly, the EBM has been redefined as a factor that strengthens and enhances the effectiveness of medical professionals. The authors in [8] argued that the EBM has compelled the medical profession against patients, pharmacists and health consumers because it reinforces the character of medical practice. They have also maintained that the development of EBM had allowed the medical profession to control medicine and maintain professional dominance.

The history of a patient and narrations in form of patient descriptions spoken or written by the doctor and patient respectively are very important for the doctor (or the machine/robot) to further search and suggest the medication strategy from the large medical corpus or using the own skillset based on experience. NLP can do helpful things for the EBM. Recent research in information retrieval w.r.t medical/clinical data has focused on the diagnoses and clinical interrogation for information and other data recovery tools to support practitioners. Regarding EBM, the very first thing NLP can help doctors/machines is the formulation of the query/strategy which needs the semantic extraction or information extraction from the sentences uttered/written by the patient. The other contribution from the domain of NLP and text classification could be to classify the patient description into a specific disease. The correct/true information searched or retrieved by the doctor/machine depends on three main factors, first the correctness of the formulation of query, second how much accurately disease was classified? and the third as understanding developed by the doctor/machine from the patient description (history) [9,10]. From these three factors, we have targeted the contribution of NLP in the second factor which is the classification of disease.

This paper contributes the textual classification of the patient descriptions for the most common 20 diseases, employing a non-overlapped dataset of 11.8 thousand patient descriptions comprising of 2.1 million tokens. Since, it is an ongoing study on the direction set by the authors in [11], Where the

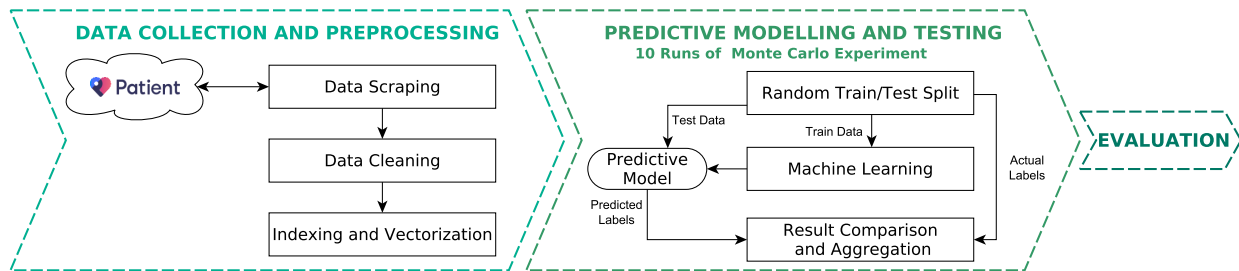


Fig. 1. Overall scheme of presented work.

authors suffered through the small dataset of 690 patients and they performed the disease classification on only 6 different diseases. In this paper, the supervised machine learning algorithms are used on the manually cleaned data. Although the widely considered NLP preprocessing pipeline (lemmatization, stemming etc.) and other NLP feature extraction techniques are omitted in this work, the presented work unpretentiously shows the promising statistics (for the overall system, and the individual class/disease-level) in term of accuracy and F_1 -measure for the 20 most common diseases.

The rest of the paper is organized in five sections; where §II provides literature review of the related work, §III discusses data and methodology, and §IV presents the insights into the results followed by a conclusion and future work in the end.

II. RELATED WORK

The globalization of information processes in all areas of knowledge, and in particular, in medicine, has posed qualitatively new problems in choosing a solution for a doctor, healthcare organizer, and patient [12–14]. EBM provides a solution for this. It is in the focus of aforesaid practitioners and general public [12,13]. The practise of evidence-based medicine means combining individual clinical experience with the best available independent clinical evidence from systematic research [15].

Without individual clinical experience, practical decisions are significantly influenced by evidence obtained even from impeccably conducted studies that may be inadequate for an individual patient [16]. On the other hand, making practical decisions without taking into account independent practical decisions can also harm the patient [15].

In the 1920s, Ronald Fisher first introduced the principles of statistical planning and analysis of experimental studies. After World War II, thanks to the work of Austin Bradford Hill and his followers, British epidemiologists Richard Doll and Archibald Cochrane, this area of science began to have a significant impact on clinical practice and public health [17]. Finally, at the end of the 20th century, thanks to the joint efforts of more than fifty specialists, primarily from McMaster University of Canada, as well as from other universities and institutions in different countries, the basic principles of evidence-based medicine were formulated [18]. By the last quarter of the twentieth century, a situation had developed where, every 5 years, the amount of medical information

doubled, and experts did not have time to get to know it for use in everyday practice.

In the field of NLP and machine learning (ML) a significant literature found on EBM. An open source system ‘RobotReviewer’ was developed by Marshal *et al.* [19] with the notion to support EBM with data. Similarly, ML-based system ‘Abstrackr’ [20] to support the easy access to EBM-based resources and an enterprise level analytics of EBM [21]. The need and requirement of EBM is stressed with aim of right knowledge and practice in NLP and ML domains [22–24].

International Disease Classification proposed [25] with the scientific and actionable measures on the classification of obesity into sub-diseases. Zhao *et al.* [26] showed, along with a research group including experts of EBM, work on the recommendation of the respiratory patients of COVID-19 by classifying the patients on symptoms.

Kim *et al.* [27] presented an automatic classification of the sentences in one thousand abstracts were made by manually annotating the labels with an accuracy of 80% by conditional random fields. Mollá *et al.* [28] made one of the good contributions in EBM with extraction and summarising of relevant information from abstracts relating to different diseases.

In the previous work, Suffian *et al.* [11] attempted to classify the diseases from textual patient descriptions. Firstly, the authors faced the problem of a small dataset (i.e., only 690 patient descriptions), secondly, only 6 diseases were classified. They employed a keyword extraction approach for feature extraction from the patient descriptions, they also tested the different combinations of features against a disease prediction and run four machine learning classifiers ultimately getting some promising results with SVM.

III. DATA AND METHODOLOGY

In this work, we deal with the most common 20 distinct diseases that are common and majorly concerned with the course of daily life; and which are to involve with a history (or the passage of time) before diagnosis. However, we maintain that the course of the presented work should not be confused with the impact of epidemics such as plague, chicken guinea, and most current COVID-19 *et cetera*. Furthermore, to produce baseline results on their classification we kept the proposed methodology very simple and straightforward. The overall scheme of the experimental setup is shown in figure 1,

TABLE I
DISEASES, NUMBER OF PATIENT DESCRIPTIONS, AND AMOUNT OF
TOKENS THEREIN.

#	Disease	Number of Posts	Number of Tokens
1	Acne	180 (1.52%)	2,561 (0.12%)
2	Angina	198 (1.67%)	38,456 (1.76%)
3	Appendicitis	122 (1.03%)	25,496 (1.17%)
4	Arthritis	297 (2.51%)	44,712 (2.05%)
5	B12 Deficiency	448 (3.79%)	93,458 (4.29%)
6	Cancer	619 (5.23%)	114,998 (5.28%)
7	Cataract	1,091 (9.22%)	198,821 (9.12%)
8	Conjunctivitis	1,247 (10.55%)	198,832 (9.12%)
9	Diabetes	535 (4.52%)	83,713 (3.84%)
10	Headache	707 (5.98%)	167,496 (7.69%)
11	Heart Attack	1,108 (9.37%)	211,385 (9.7%)
12	Hepatitis	79 (0.67%)	11,879 (0.55%)
13	Hernia	822 (6.95%)	140,087 (6.43%)
14	Hypertension	888 (7.51%)	138,991 (6.38%)
15	Otitis-Media	705 (5.96%)	141,215 (6.48%)
16	Piles	1,034 (8.74%)	243,237 (11.16%)
17	Renal-Failure	631 (5.33%)	98,446 (4.52%)
18	Stroke and Tia	231 (1.95%)	47,511 (2.18%)
19	Urinary-Tract-Infection	591 (5%)	112,707 (5.17%)
20	Urticarial-Rash	296 (2.5%)	65,032 (2.98%)
Total		11,829	2,179,033

where the discussion on every component is provided in the subsequent subsections.

A. Data: Collection, Cleaning, Statistical Distribution, and Visual Insights

We scrap patient descriptions from a public forum available online at patient.info¹ where people from all around the globe share their health problems which are near-accurately categorized into specific sections. In this regard, we used Python-based web-scraping package, namely, **BeautifulSoup**². A generic script is written for the three different tasks, which covers scraping of *disease group*, followed by *disease names* that exist therein the group, and lastly all of the individual *posts* relating to the diseases. A separate file with Comma-Separated-View (CSV) format is created for every disease which contains all of the posts for it. Besides it, these CSV files keep three things which are: *Title* i.e. a short topic of the post, *Link* i.e. a web-address of the post, and *Discussion* i.e. a textual description of disease, as of a medium length document.

As a result, we succeeded in totaling 11,829 distinct posts for the the most common 20 diseases; where the size of the corpus is ≈ 2.18 million tokens. We name this dataset **PI-PD-20**³. On average, every disease contains ≈ 591 posts with ≈ 109 thousand tokens. The details of the number of posts for the individual disease and the count of tokens are given in table I; we can see the maximum number of posts (i.e., > 1000) are for cataract, conjunctivitis, heart attack, and

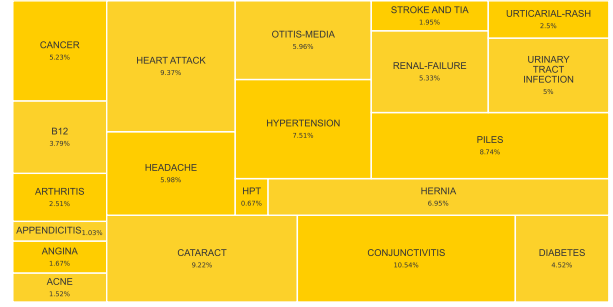


Fig. 2. Proportion of the diseases w.r.t the number of records.

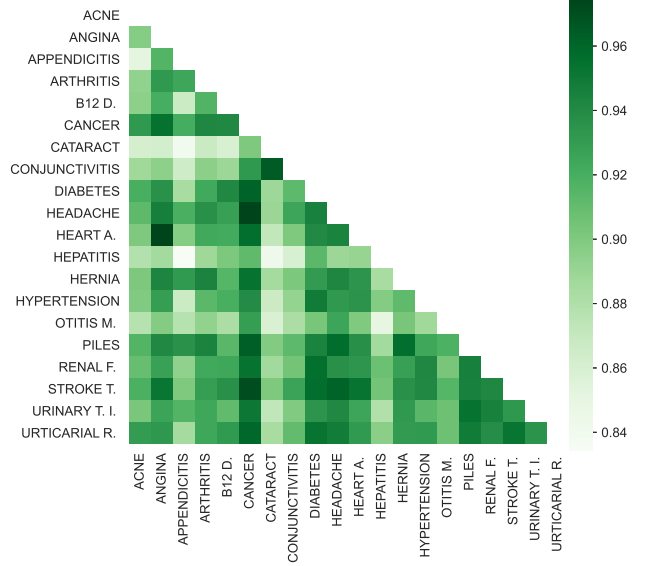


Fig. 3. Heat-map showing the (cosine) similarity among the disease descriptions. The darker shade indicates the higher similarity.

piles. The disease with a lower number of posts (i.e., < 200) includes acne, angina, appendicitis, and hepatitis. Figure 2 shows a bird-view of the proportions of diseases in the dataset; in which we present a squared-proportions chart instead of a pie chart due to the lack of space. We maintain that the figures reported in the table I are after committing the duplicate removal process; alongside this, we also perform manual cleaning of extra white-spaces, unwanted JavaScript/JQuery snippets, and special characters.

Figure 3 shows the insights into the overlapping of content in disease descriptions. We concatenated all patient descriptions against the diseases and then computed cosine similarity among them in a pair-wise manner [29–31]. We can see that diseases like ‘angina’ and ‘heart attack’ have got the higher similarity—which appears natural in a layman’s term that both of the diseases are related to the chest-pain and heart; and in an equal manner, ‘cataract’ and ‘conjunctivitis’ have got high similarity—both of them are relating to the problem with eyes. For ‘hypertension’, we figured the major similarity out is with ‘diabetes’, which also confirms the seminal findings of [32]. We also see the pairs of irrelevant diseases with due lower similarity, e.g. (‘otitis media’–‘hepatitis’) in which otitis media

¹<https://patient.info/forums>

²<https://www.crummy.com/software/BeautifulSoup/>

³**PI-PD-20** is available at <https://github.com/muhammadyaseenkhan/ebm-patient-descriptions-dataset/tree/main/PI-PD-20>

is related to ears whereas hepatitis is related to the liver.

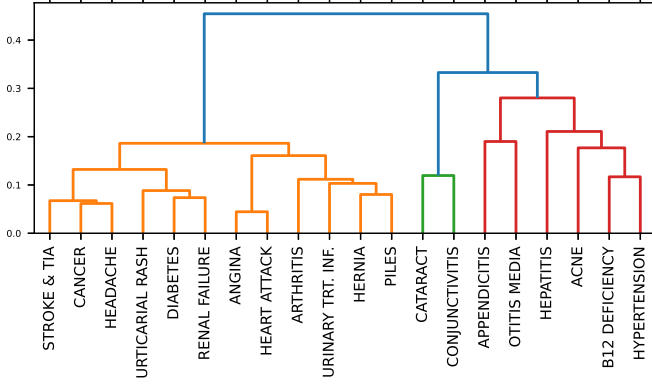


Fig. 4. Dendrogram showing the hierarchical clustering of diseases; constructed with the similarity matrix produced in figure 3.

Figure 4 shows the agglomerative hierarchical clustering of diseases; developed with the complete-linkage clustering method [33,34]. We see that the ‘angina’ and ‘heart attack’ are close to be set in the same cluster, so with the eye diseases: ‘cataract’ and ‘conjunctivitis’. Though most of the abdominal diseases (urticarial rash, renal failure, urinary tract infection, hernia, and piles) are not clustered very early but shown in the same (orange) group; ‘B12 deficiency’ causes stomach/gastrointestinal problems—which is, as well, highly correlated with the ‘hypertension’ [35–37]—appeared together.

B. Preprocessing, Term Indexing, and Vectorization

For the classification task, we need to provide the input in a vector format to the ML algorithms. Thus, we are required to transform the text documents into vectors, typically in a high-dimensional vector space. We can think of every document is transformed into the vector of the same length, i.e. equal to the number of distinct terms in the corpus⁴. We, for producing baseline and prefatory results in this paper, put the vectorization process in a very simple way with a binary or Boolean representation of uni-grams that are present in the very document. In the same context, we usually see stop words removal, stemming, and lemmatization in preprocessing a text document which is also not considered for the experiments dealt with in this paper.

C. Supervised Learning and Predictive Analytics

In a supervised learning-based classification system, documents or text with already tagged class labels are set as an input to the ML algorithm; with the given information it targets to learn the underlying information and patterns from the data to generate a predictive model; to mean it numerically, consider $f: \mathcal{X} \rightarrow \mathcal{Y}$; where f is a learning function that aims

⁴The vectorization can be made with multiple approaches, for which we can include the term weighing (like TF-IDF) and counting approaches; similarly, the vector can be consisting of word pairs and word sequences, technically called as n -grams. For more details on term indexing, and generalized approach for word and character-level n -gram vectorization please see [31,38].

to map output $\mathcal{Y} \in \{c_1, c_2, \dots, c_n\}$ on the given input feature vector $\mathcal{X} = \{x_1, x_2, \dots, x_{n'}\}$ [39]. Likewise Suffian *et al.* [11], in our case, the problem is a Multi-Class Classification Problem (MCCP); which means there are more than 2 classes in the output universe i.e., $\mathcal{Y} = \{\text{names of the diseases}\} \Leftrightarrow |\mathcal{Y}| > 2$; however, the predictive model has to return only one label which has got the highest probability of prediction. The phase in which learning for the predictive model is made is also called the training phase, and the phase where we test the performance of the generated model is called the testing phase. We set the train–test split ratio as per Pareto principal [38,40] and stratified manner.

We set the Monte Carlo Method (MCM) for the model assessment. The MCM functions with the shuffling of data and then splitting of dataset into training and testing parts, when a predictive model is generated employing training data, we evaluate it on test data followed by saving resulting Confusion Matrix. Thus, in k runs (provided that $k \geq 2$), we keep aggregating CM; when k runs are completed, in last, we average out the evaluating metrics.

As a baseline experiment, we have used 4 different ML algorithms. We have employed Python-based library, namely **scikit-learn**⁵ in the ML tasks. The listing of algorithms used in the experiment with brief discussion is given below.

- 1) *Decision Tree* (DT) with Iterative Dichotomizer-3 [41] as a functioning algorithm. It generates a tree by calculating the entropy for every attribute/feature and splits its branches such that the attribute with the lowest entropy has to get the maximum depth [42,43]. In most of the cases entropy subsumes Information Gain as well [41], which has to behave contrasting to the entropy.
- 2) *Random Forest Trees* (RFT) [44,45], which is the ensemble learning approach in classification, that aims to generate more than one DT from random samples and then consolidate the prediction made through each of the decision tree [46]. We ensemble 200 estimators in RFT; however rest of the parameters are in default setting.
- 3) *Support Vector Machine* (SVM) [47], which is the linear model for classification that targets to fit a hyper-plane in vector-space for separating vectors into distinct parts such that the marginal space separating them is maximum. **scikit-learn** employed a lightweight SVM library namely, LIBSVM [48], for performing classification; it, by default, assumes radial basis function kernel [49]. Alongside it, we also tested LINEAR SVM that is given, by default, in Stochastic Gradient Descent (SGD).
- 4) *Naïve Bayes* (NB), which is a likelihood based probabilistic classifying function. It is based on the Bayes’ theorem but holds the assumption that all attributes are conditionally independent [38,43,50–53].

D. Evaluation Metrics and Criteria

The metrics considered for evaluating the predictive model are basic. These metrics are itemized in the following text; for

⁵<https://scikit-learn.org/>

TABLE II

MODEL CONFUSION MATRIX FOR MULTI-CLASS CLASSIFICATION PROBLEM. IN ROW AND COLUMN WITH HEAD \mathcal{S}_i , VARIABLE n GENERALIZES \mathbf{T} . AND \mathbf{F} .

		Predicted Labels				\mathcal{S}_r
		$class_1$	$class_2$	\dots	$class_c$	
Actual Labels	$class_1$	\mathbf{T}_{11}	\mathbf{F}_{12}	\dots	\mathbf{F}_{1c}	$\mathcal{S}_{1+} = \sum_{i=1}^c n_{1i}$
	$class_2$	\mathbf{F}_{21}	\mathbf{T}_{22}	\dots	\mathbf{F}_{2c}	$\mathcal{S}_{2+} = \sum_{i=1}^c n_{2i}$
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	$class_c$	\mathbf{F}_{c1}	\mathbf{F}_{c2}	\dots	\mathbf{T}_{cc}	$\mathcal{S}_{c+} = \sum_{i=1}^c n_{ci}$
\mathcal{S}_c		$\mathcal{S}_{+1} = \sum_{i=1}^c n_{i1}$	$\mathcal{S}_{+2} = \sum_{i=1}^c n_{i2}$	\dots	$\mathcal{S}_{+c} = \sum_{i=1}^c n_{ic}$	$\mathcal{N} = \sum_{i=1}^c \mathcal{S}_{i+} = \sum_{i=1}^c \mathcal{S}_{+i}$

their definitions and derivations, consider the Confusion Matrix (CM), as shown in table II, corresponding to the number of items coinciding actual labels and the predicted labels for c classes; wherein the table, T_{ii} informs True-Predictions or no-conflicts i.e. actual and predicted labels are same for class i ; whereas, F_{ij} shows False-Predictions or misclassifications of items in class j w.r.t their actual class i ; $F_{ij} \neq F_{ji}$.

- Precision (PR). It is the positive predictive value, which means the right potential of a predictive system to predict true-positives from the number of item it predicts as positives. To mean it mathematically, the PR for class i , and whole system will be:

$$PR_i = \frac{T_{ii}}{\mathcal{S}_{+i}}; \quad PR = \frac{1}{c} \sum_{i=1}^c \frac{T_{ii}}{\mathcal{S}_{+i}} \quad (1)$$

- Recall (RC). It is the true positive rate, which means the right potential of a predictive system to predict true-positives from the number of item that are actually positive. Mathematically, the RC for class i , and whole system will be:

$$RC_i = \frac{T_{ii}}{\mathcal{S}_{+i}}; \quad RC = \frac{1}{c} \sum_{i=1}^c \frac{T_{ii}}{\mathcal{S}_{+i}} \quad (2)$$

- F_1 -Measure (F_1). It is the harmonic mean of PR and RC [54,55]. In comparison to the arithmetic mean, it is a strict measure as it has got the propensity of leaning towards the global minima of the input numbers. Mathematically, using equations 1 and 2, the F_1 for class i , and whole system will be calculated as:

$$F_{1i} = 2 \cdot \frac{PR_i \cdot RC_i}{PR_i + RC_i}; \quad F_1 = \frac{2}{c} \sum_{i=1}^c \frac{PR_i \cdot RC_i}{PR_i + RC_i} \quad (3)$$

- Accuracy (ACC). It quantifies overall success of the predictive system. Mathematically, the ACC of the whole system will be:

$$ACC = \frac{1}{\mathcal{N}} \sum_{i=1}^c T_{ii} \quad (4)$$

Since we set the ML experimental setup in a MCM, therefore, we report average scores of evaluation metrics; and, since the

dataset is imbalanced, therefore, we reported weighted scores of the metrics as well [54]. In practice, we are interested in bigger numbers on the left diagonal of CM, which is shown in red colour, indicating the coinciding numbers for True-Predictions. In the same context, for EBM, we are more interested in high RC; however, we cannot neglect the lower PC as well, thus, the ML algorithm which, besides the F_1 measure, attains high score in both metrics would be called champion of disease classification.

IV. RESULTS AND DISCUSSION

In this section, we report the performance of the proposed methodology in three parts; of which, the first one is related to the overall performance of ML algorithms, the second part is related to the analysis of the class-level performance of ML algorithms, followed by analysis on misclassifications w.r.t the most optimal algorithm thereof.

TABLE III
AVERAGED PERFORMANCE OF ML ALGORITHMS.

Algorithm	Type	PR	RC	F_1	ACC
NB	macro	68%	55%	56%	73%
	weighted	74%	73%	70%	
DT	macro	73%	70%	71%	73%
	weighted	73%	73%	73%	
RFT	macro	87%	75%	78%	83%
	weighted	84%	83%	82%	
LIBSVM (RBF)	macro	73%	54%	57%	68%
	weighted	74%	68%	67%	
SGD (LIN. SVM)	macro	78%	78%	78%	81%
	weighted	81%	81%	81%	
Stacked Average	macro	75.8%	66.4%	68%	75.6%
	weighted	77.2%	75.6%	74.6%	

The overall performance of the system is reported in table III. We can see that the algorithms RFT and SGD (i.e. similar to linear SVM) have secured approximately same results (83% and 81% respectively) with a marginal difference of 2%. However, we put forward that RFT is the most optimal algorithm w.r.t the overall accuracy of the system, and also for gaining the highest weighted PR and RC both.

TABLE IV
REPORT OF PRECISION, RECALL, AND F_1 -MEASURE FOR THE CLASSIFICATION OF DISEASES.

Diseases	NB			DT			RFT			SVM (RBF)			SGD (LIN. SVM)		
	PR	RC	F_1	PR	RC	F_1	PR	RC	F_1	PR	RC	F_1	PR	RC	F_1
Acne	1	.03	.05	.68	.62	.65	.93	.62	.75	1	.17	.3	.77	.75	.76
Angina	0	0	0	.39	.28	.33	.83	.13	.22	0	0	0	.41	.44	.43
Appendicitis	0	0	0	1	.87	.93	.96	.87	.91	1	.4	.57	.81	.87	.84
Arthritis	.9	.16	.27	.52	.57	.55	.84	.64	.73	.68	.3	.42	.63	.7	.66
B12 Deficiency	.93	.84	.88	.79	.81	.8	.93	.9	.91	.98	.73	.84	.92	.91	.92
Cancer	.89	.38	.53	.63	.52	.57	.92	.46	.61	.57	.39	.46	.66	.69	.67
Cataract	.8	.87	.83	.8	.82	.81	.83	.91	.87	.73	.86	.79	.81	.92	.86
Conjunctivitis	.7	.79	.74	.74	.73	.73	.79	.83	.81	.75	.77	.76	.85	.79	.82
Diabetes	.87	.47	.61	.73	.73	.73	.77	.75	.76	.84	.48	.61	.81	.81	.81
Headache	.58	.89	.7	.64	.71	.67	.7	.87	.77	.71	.69	.7	.67	.77	.72
Heart Attack	.47	.9	.62	.63	.73	.68	.69	.9	.78	.51	.83	.63	.6	.82	.69
Hepatitis	.5	.06	.11	.41	.53	.46	1	.29	.45	0	0	0	.64	.41	.5
Hernia	.81	.91	.85	.85	.83	.84	.92	.91	.91	.93	.76	.84	.95	.88	.91
Hypertension	.81	.79	.8	.7	.74	.72	.8	.83	.81	.72	.79	.75	.87	.75	.81
Otitis Media	.97	.85	.91	.83	.81	.82	.9	.88	.89	.97	.68	.8	.9	.63	.74
Piles	.71	.98	.82	.82	.78	.8	.84	.97	.9	.41	.95	.58	.92	.9	.91
Renal Failure	.84	.69	.76	.7	.71	.71	.87	.83	.85	.82	.56	.67	.9	.78	.84
Stroke & Tia	0	0	0	.59	.6	.59	.92	.53	.68	1	.18	.3	.82	.71	.76
Urinary Trt. Inf.	.94	.87	.9	.81	.77	.79	.88	.93	.9	.97	.65	.78	.87	.82	.85
Urticarial Rash	.97	.6	.74	.92	.79	.85	.96	.88	.92	1	.51	.67	.93	.95	.94

Table IV reveals the performance of ML algorithms for the individual diseases. (See the cells with red shade) ‘Angina’ and ‘hepatitis’ appear to be very critical of all for classification. ‘B12 deficiency’, ‘hernia’ ‘urinary tract infection’, ‘cataract’, and ‘otitis media’ are the diseases that have secured approximately good F_1 score by all algorithms. NB and SVM with RBF are the only two algorithms who failed to produce any positive result on ‘angina’, ‘appendicitis’, ‘hepatitis’, and ‘stroke & tia’. We can get the idea on the due mistake of such failure is based on a high similarity between ‘angina’ and ‘heart attack’ documents. The cells with the green shade show the highest F_1 score achieved by the very algorithm for a particular disease, in case of a tie, we selected F_1 score where the RC is higher. The cells with the yellow shade inform on subsequent low RC though the yielded PR is 1; except for ‘appendicitis’.

Figure 5 presents insights into the misclassifications. The most significant confusion is between ‘angina’+‘stroke’ and ‘heart attack’, which respectively loses more than 56% and 20% of the documents in the wrong prediction. Followed by ‘heart attack’, ‘Acne’ stands as a confusing disease between ‘conjunctivitis’+‘piles’; we can surmise that the term relating to *colours* (red stool, pale/yellow and red/pink colour of eyes) are the key subjects for such misclassifications. We maintain misclassification of ‘hepatitis’ similar to the case of ‘acne’.

In comparison to the previous work [11], misclassification between ‘hypertension’ and ‘diabetes’ is reduced to 1.1% from 5%. misclassification of ‘diabetes’ with heart-related diseases (4% i.e., as in [11]) is approximately equal 4.01% (aggregating ‘heart attack’= 1.5% and ‘angina’= 2.6%).

V. CONCLUSION AND FUTURE WORK

In the world of medicine, mostly with any situation, we aim for the collective good of living beings. In this regard, we presented a solution by preparing ML data from patient descriptions. We conclude that in the absence of visual clues

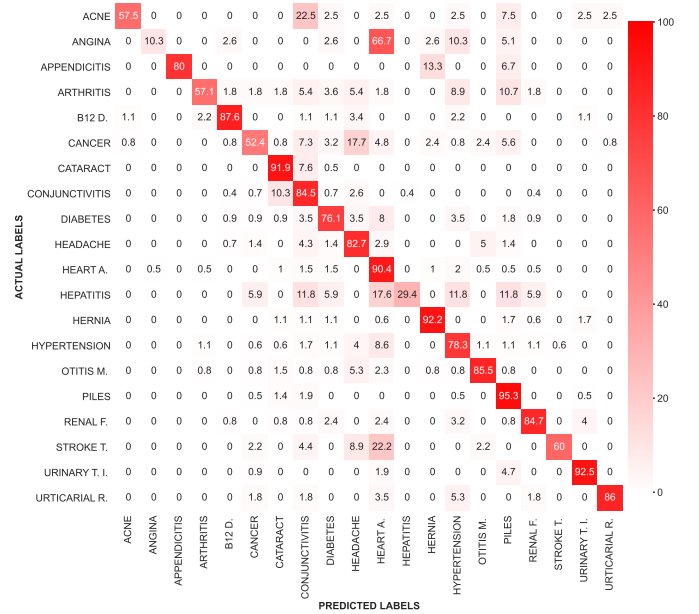


Fig. 5. Confusion matrix for RFT.

(medical images/x-rays, CT-scans *et cetera*), these patient descriptions can be utilized for the classification of diseases. We showed, an unpretentious solution for the problem under study, without sophisticated preprocessing and feature engineering steps, with a fair accuracy of 83%, alongside attaining a considerably competent precision of 84% and recall of 83% with RFT with 200 estimators. These baseline statistics welcomes the attempts for further improvements.

In future, we intend to expand this experiment for pandemic related diseases and on the causality of common diseases. We are hopeful that the dataset constructed in this paper, **PI-PD-20**, will be beneficial for EBM studies.

REFERENCES

- [1] A. La Caze, "The role of basic science in evidence-based medicine," *Biology & Philosophy*, vol. 26, no. 1, pp. 81–98, 2011.
- [2] A. Mauro, "Medicalization: current concept and future directions in a bionic society," *Mens sana monographs*, vol. 10, no. 1, p. 122, 2012.
- [3] A. Stavrou, D. Challoumas, and G. Dimitrakakis, "Archibald cochrane (1909–1988): the father of evidence-based medicine," *Interactive cardiovascular and thoracic surgery*, vol. 18, no. 1, pp. 121–124, 2014.
- [4] I. Masic, M. Miokovic, and B. Muhamedagic, "Evidence based medicine—new approaches and challenges," *Acta Informatica Medica*, vol. 16, no. 4, p. 219, 2008.
- [5] P. Batchelor, "The legal and ethical implication of evidence-based clinical guidelines for clinicians," *Evidence-Based Dentistry*, vol. 2, no. 1, pp. 5–6, 2000.
- [6] S. Timmermans and M. Berg, "The gold standard: the challenge of evidence-based medicine and standardization in health care temple university press," *Philadelphia PA*, 2003.
- [7] K. Dickersin, S. E. Straus, and L. A. Bero, "Evidence based medicine: increasing, not dictating, choice," *BmJ*, vol. 334, no. suppl 1, pp. s10–s10, 2007.
- [8] Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. J. Cimino, J. Ely, and H. Yu, "Askhermes: An online question answering system for complex clinical questions," *Journal of biomedical informatics*, vol. 44, no. 2, pp. 277–288, 2011.
- [9] P. E. Marik, "'less is more': The new paradigm in critical care," in *Evidence-Based Critical Care*. Springer, 2015, pp. 7–11.
- [10] I. Mikisek, *Evidence Based Management: gesundheitsförderliche Führung*. Springer-Verlag, 2015.
- [11] M. Suffian, M. Y. Khan, and S. Wasi, "Developing disease classification system based on keyword extraction and supervised learning," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 599–605, 2018.
- [12] K. Galbraith, A. Ward, and C. Heneghan, "A real-world approach to evidence-based medicine in general practice: a competency framework derived from a systematic review and delphi process," *BMC medical education*, vol. 17, no. 1, pp. 1–15, 2017.
- [13] K. Prasad, *Fundamentals of evidence based medicine*. Springer, 2013.
- [14] K. Svore, L. Vanderwende, and C. Burges, "Enhancing single-document summarization by combining ranknet and third-party sources," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 448–457.
- [15] A. Morabia, "Pierre-charles-alexandre louis and the evaluation of bloodletting," *Journal of the Royal Society of Medicine*, vol. 99, no. 3, pp. 158–160, 2006.
- [16] R. Fletcher, S. Fletcher, and E. Vagner, "Clinical epidemiology: basics of evidence based medicine," *Trans. from English]. Moscow: Media-Sfera*, 1998.
- [17] A. L. Zimmerman, "Evidence-based medicine: a short history of a modern medical movement," *AMA Journal of Ethics*, vol. 15, no. 1, pp. 71–76, 2013.
- [18] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.
- [19] I. J. Marshall, J. Kuiper, E. Banner, and B. C. Wallace, "Automating biomedical evidence synthesis: Robotreviewer," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2017. NIH Public Access, 2017, p. 7.
- [20] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos, "Deploying an interactive machine learning system in an evidence-based practice center: abstrackr," in *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, 2012, pp. 819–824.
- [21] M. Puppala, T. He, S. Chen, R. Ogunti, X. Yu, F. Li, R. Jackson, and S. T. C. Wong, "Meteor: An enterprise health informatics environment to support evidence-based medicine," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2776–2786, 2015.
- [22] M. Alsawas, F. Alahdab, N. Asi, D. C. Li, Z. Wang, and M. H. Murad, "Natural language processing: use in ebm and a guide for appraisal," *BMJ Evidence-Based Medicine*, vol. 21, no. 4, pp. 136–138, 2016.
- [23] M. M. Al-Jefri, R. Evans, P. Ghezzi, and G. Uchyigit, "Using machine learning for automatic identification of evidence-based health information on the web," in *Proceedings of the 2017 International Conference on Digital Health*, 2017, pp. 167–174.
- [24] I. A. Scott, "Machine learning and evidence-based medicine," 2018.
- [25] W. T. Garvey and J. I. Mechanick, "Proposal for a scientifically correct and medically actionable disease classification system (icd) for obesity," *Obesity*, vol. 28, no. 3, pp. 484–492, 2020.
- [26] H.-M. Zhao, Y.-X. Xie, C. Wang *et al.*, "Recommendations for respiratory rehabilitation in adults with coronavirus disease 2019," *Chinese medical journal*, vol. 133, no. 13, pp. 1595–1602, 2020.
- [27] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, vol. 12, no. 2. BioMed Central, 2011, pp. 1–10.
- [28] D. Mollá, M. E. Santiago-Martínez, A. Sarker, and C. Paris, "A corpus for research in text processing for evidence based medicine," *Language Resources and Evaluation*, vol. 50, no. 4, pp. 705–727, 2016.
- [29] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.
- [30] A. Singhal *et al.*, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

- [31] M. Y. Khan and M. S. Nizami, "Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled dataset for urdu sentiment analysis," in *2020 International Conference on Information Science and Communication Technology (ICISCT)*. IEEE, 2020, pp. 1–15.
- [32] M. V. Williams, D. W. Baker, R. M. Parker, and J. R. Nurss, "Relationship of functional health literacy to patients' knowledge of their chronic disease: a study of patients with hypertension and diabetes," *Archives of internal medicine*, vol. 158, no. 2, pp. 166–172, 1998.
- [33] B. S. Everitt, S. Landau, and M. Leese, "Cluster analysis arnold," *A member of the Hodder Headline Group, London*, pp. 429–438, 2001.
- [34] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining text data*. Springer, 2012, pp. 77–128.
- [35] J. Bosch, J. G. Abraldes, A. Berzigotti, and J. C. Garcia-Pagan, "Portal hypertension and gastrointestinal bleeding," in *Seminars in liver disease*, vol. 28, no. 01. © Thieme Medical Publishers, 2008, pp. 003–025.
- [36] S. George, P. Reichardt, T. Lechner, S. Li, D. Cohen, and G. Demetri, "Hypertension as a potential biomarker of efficacy in patients with gastrointestinal stromal tumor treated with sunitinib," *Annals of oncology*, vol. 23, no. 12, pp. 3180–3187, 2012.
- [37] Z. Zhu, S. Xiong, and D. Liu, "The gastrointestinal tract: an initial organ of metabolic hypertension?" *Cellular Physiology and Biochemistry*, vol. 38, no. 5, pp. 1681–1694, 2016.
- [38] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.
- [39] M. Y. Khan and K. N. Junejo, "Exerting 2d-space of sentiment lexicons with machine learning techniques: A hybrid approach for sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0110672>
- [40] V. Pareto, *Cours d'économie politique*. Librairie Droz, 1964, vol. 1.
- [41] J. Quinlan, "Induction of decision trees. mach. learn.," 1986.
- [42] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [43] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [44] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.
- [45] —, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
- [46] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [47] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [48] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.
- [49] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and testing low-degree polynomial data mappings via linear svm," *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.
- [50] D. J. Hand and K. Yu, "Idiot's bayes—not so stupid after all?" *International statistical review*, vol. 69, no. 3, pp. 385–398, 2001.
- [51] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002, pp. 79–86.
- [52] M. Y. Khan, S. M. Emaduddin, and K. N. Junejo, "Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach," in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, 2017, pp. 242–249.
- [53] W. Muhammad, M. Mushtaq, K. N. Junejo, and M. Y. Khan, "Sentiment analysis of product reviews in the absence of labelled data using supervised learning approaches," *Malaysian Journal of Computer Science*, vol. 33, no. 2, pp. 118–132, 2020.
- [54] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th Conference on Message Understanding*, ser. MUC4 '92. USA: Association for Computational Linguistics, 1992, p. 22–29.
- [55] N. Jardine and C. J. van Rijsbergen, "The use of hierarchic clustering in information retrieval," *Information storage and retrieval*, vol. 7, no. 5, pp. 217–240, 1971.