Diagnosis of Diseases from Medical Check-up Test Reports Using OCR Technology with BoW and AdaBoost algorithms

Conference Paper · June 2019

DOI: 10.1109/IEC47844.2019.8950605

2 authors:

Wisam Abdulaziz
Ishik University
4 PUBLICATIONS **1** CITATION
SEE PROFILE

Musa M.Ameen
Tishk International University
8 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Diagnosis of Diseases from Medical Check-up Test Reports Using OCR Technology with BoW and AdaBoost algorithms

Wisam A. Qader
Computer Engineering Dep
Tishk International University
Erbil,Iraq
wisam.softeng@gmail.com

Musa M. Ameen
Computer Engineering Dep
Tishk International University
Erbil, Iraq
musa.m.amin@gmail.com

***Abstract*—This research introduces an approach to diagnose diseases from medical check-up test reports. The proposed approach is produced from Optical Character Recognition (OCR) technology to convert the hard copy test reports into editable textual data, Bag of Words (BoW) model as feature selection algorithm, Naïve Bayes as classification algorithm, and AdaBoost technique to enhance the performance of the Naïve Bayes classifier. The performance of the proposed approach is very good in terms of validity and can be used in diagnosing of diseases from medical check-up test reports. The proposed approach is trained on dedicated trained partitions of multiple medical datasets, and then tested on the testing sets partitioned from the original datasets. The proposed algorithm is compared with the Support Vector Machine (SVM), Naïve Bayes (NB), Decision Table (DT), and k-Nearest Neighbors (k-NN) classifiers, in which all the algorithms are tested on the same datasets. The proposed algorithm showed higher accuracy than the other four classifiers. So, the proposed approach which is the combination of BoW with AdaBoost technique is used to predict the name of the diseases from the medical check-up test reports. After that, an image as an example of the disease will be presented as well with the name of the disease to the physician and the patient. The image presentation is very important for the patients, because they may not familiar with the medical terms and disease names. Finally, the proposed approach can be used in the medical area because of its good performance and showing validated results after it is tested.**

Keywords*—*Natural Language Processing, Optical Character Recognition, Bag-of-Words, AdaBoost, Support Vector Machine, Decision Table, Naïve Bayes, k-Nearest Neighbors

I. INTRODUCTION

Diagnosis of diseases is considered as one of the most important, complex and required issues in the medical system for reducing the false detection rate of the diagnosis of diseases [5] with the hope to effectively guarantee the diagnosis of diseases from the medical test reports, and the best matching medication for the patients. In fact, it is significant in health sectors to set long, medium and short-term strategies to determine concrete systems, so that to give a higher rate of accuracy for diagnosing of diseases by the medical diagnostic centers. Thus, it is perfectly safe to state that these concrete systems of diagnosis will tip and aim at reaching closer and faster to ultimate health goals within the scope and domain of the medical framework.

It is worth to mention that, discovering new approaches to improve the medical system for the health sectors needs a very good performance, and it can be done by the cooperation of the physicians with the system engineers to identify the key features for diagnosis of diseases from medical reports easier. It is obvious to mention that highly qualified and professional medical systems can be considered as one of the major assets for improving the quality level in health sectors. Therefore, it is very crucial to convey systems for analyzing medical test results, assessment in a professional way to comprehend the accuracy of the system, performance in terms of pros and cons to make sure for analyzing and diagnosing of diseases. The proposed approach or model can be summarized by the following steps:

- A captured or scanned copy of the medical check-up test report is inserted into the proposed model, it is being converted from the image into textual data using OCR.
- Pre-processing for the textual data will be done to prepare it for feature selection.
- The BoW will be used to create bags for each disease type according to their features.
- The AdaBoost technique will be used for enhancing the classification performance of the medical check-up tests.
- An image of the related disease will be displayed to the physician and the patient.

Ease of use, raising the accuracy and showing the sample image to the patients to understand easier are considered as the highlighted advantages of the proposed approach. And the disadvantage of the proposed approach is that it needs too much data to train itself and it is not a simple task, then it can be used in real life.

As a matter of fact, analyzing features for diagnosing diseases with perfect accuracy is viewed as tremendously important. The performance of classification is the most important measure to evaluate the classifiers. By taking the result of the proposed approach under test to reach the main goal, the help and contribution of specialties of the field are highly noticed.

Consequently, medical diagnosis is a very sensitive task, since if the result of the diagnosis was incorrect it causes using wrong medications and causes a very terrible result, since an error in the prediction of the disease causes a

978-1-7281-4377-4/19/$31.00 ©2019IEEE

disobey about the life of people. It is true to say that the process of diagnosis of diseases in medical check-up tests is a vigorous issue at both theory and practical levels, so it is deemed as a research subject for the health sector. So, we need some models with capabilities to reduce the rate of incorrect diagnosis of diseases. For those reasons, this project was designed to show a very good performance and display the disease name and a sample image in a way that is very simple for the patient to understand.

This approach is developed to give results that affect the improvement of the diagnosis of diseases. In addition to the accuracy, the time needed for the report to for analyzing is also considered, wherein [1] is proved that the data processing of medical area using computerized systems is 40 times faster than the one is done manually, and also it will be easier to be understood by the patients.

Designing an approach with good performance allows the model to guarantee that works well for the specified domain [2]. For that reason, a novel approach is presented to diagnose diseases from medical check-up test reports.

Over the last five decades, a great number of researches have been done on character recognition techniques. Reference [3] mentioned the OCR as a successful technology for transformation of printed text into editable text. They are mentioning that OCR is very popular and beneficial technology in various fields to be used in different kinds of applications. They also mentioned that the accuracy rate of the OCR is directly affected by the quality of the images. And sometimes it is hard to extract text from printed texts or images, because of the color, complexity of the background, the orientation of the text on the image, .etc.

BoW algorithm for the annotation of medical images to be performed automatically is offered in [4], it was used to reduce the dimensionality of the text that is based on the concept of "term frequency - inverse document frequency" (tf-idf), and reduced the co-occurrence between terms. The medical report images which are used were containing text describing medical images. They were interested to extract all the relevant terms that were containing medical concepts. They used the BoW model to recover the vector of the features. Finally, they got to the conclusion that the medical diagnosis classification has to cover feature selection too, because combining feature selection technique with machine learning algorithms as feature extraction technique produces an output with a higher accuracy rate.

The researchers in [5] were developed an algorithm for extraction of mammographic infection information from free-text mammography reports, the algorithm was a rule-based Natural Language Processing (NLP) system. The system could extract four mammographic findings: architectural distortion, calcification, asymmetry, and mass. The annotations of status and anatomical locations were associated with each extracted findings from the system through association rules. After excluding unnecessary findings, confirmed extracted findings were summarized. A set of hundred reports were manually checked for the purpose of evaluating the accuracy of the proposed NLP system. The system could correctly mention the findings from 96 samples out of the 100 reports. According to the results, the system successfully extracted helpful information from medical mammography reports.

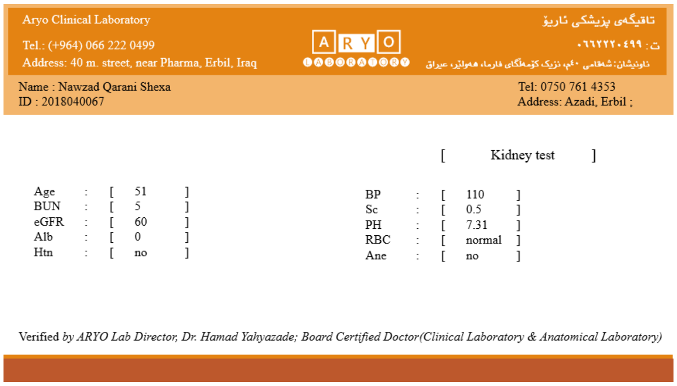
The authors of [6] proposed a hybrid technique algorithm which is derived from the combination of AdaBoost and fuzzy methods to analyze medical data and diagnosis of diseases. In which AdaBoost is used as a general method to improve the performance of the learning algorithms, and the fuzzy method used as base classifier, the combination of both could provide higher accuracy rate in medical data classification and diagnosis of diseases. They tested the proposed algorithm against four medical datasets for cardiovascular, epileptic seizure detection, hepatitis and Parkinson diseases. The average accuracy rate from the four datasets was 95.8%, and it was a validated result, because of the complexity of classification of the medical data.

In this work, the combination of BoW with AdaBoost techniques are used, where the Naïve Bayes algorithm is used as the base classifier for the AdaBoost to classify novel instances. As a result, the proposed algorithm showed a validated result, since the combination of those two techniques is often shows great accuracies.

This paper is organized as follows: Section II describes the methods and materials used in the proposed approach. Section III shows the experimental results and discusses them briefly. Finally, some points are concluded in section IV.

II. Methods and Materials

This research falls into the field of machine learning, which is a very hot field and is getting common to be used in the medical domain. This project is used to diagnose diseases from medical check-up test reports that contain tests with their results, such as Heartbeat rate is 84, Hypertension is abnormal, Diabetes mellitus exists, ...etc. as shown in Fig. 1. The aim of this research is to find a novel approach for diagnosing diseases from medical check-up test reports with a high and validated accuracy rate.



Aryo Clinical Laboratory
Tel.: (+964) 066 222 0499
Address: 40 m. street, near Pharma, Erbil, Iraq

تاقیگەی پزیشکی ئاریۆ
ت: ٠٦٦٢٢٢٠٤٩٩
ناونیشان: شەقامی ٤٠م، نزیک کۆمەڵگای فارما، هەولێر، عیراق

Name : Nawzad Qarani Shexa
ID : 2018040067

Tel: 0750 761 4353
Address: Azadi, Erbil ;

[Kidney test]

Age	: [51]	BP	: [110]
BUN	: [5]	Sc	: [0.5]
eGFR	: [60]	PH	: [7.31]
Alb	: [0]	RBC	: [normal]
Htn	: [no]	Ane	: [no]

Verified *by ARYO Lab Director, Dr. Hamad Yahyazade; Board Certified Doctor(Clinical Laboratory & Anatomical Laboratory)*

Fig. 1: sample of medical check-up test report

There are many different NLP techniques for transforming handwritten text or text on papers into editable textual data, in this research Tesseract OCR is used to perform the task of extracting text from medical check-up test reports, because it is a very powerful tool and simple to use to perform that task. And BoW is selected to be used as feature section algorithm among various algorithms with the aim to simplify the datasets before classifying their data in the upcoming steps to produce results with higher accuracy. And there are also many different machine learning classification algorithms to be used for classification purposes, among them four powerful classifiers have been selected to be compared and choosing the one with the best performance in medical data analysis domain to be used as the base classifier with the AdaBoost technique. The four classifiers are SVM from the Functions family, k-NN from the Lazy, DT from the Rules, and NB from the Bayes family.

The SVM, k-NN, DT, and NB classifiers are tested twice. Once without adopting the BoW, and once after applying the BoW on the datasets. Raising of the accuracy have been noticed after the adoption of the BoW as they are compared with the same algorithms before applying the BoW. Then AdaBoost technique is applied on all the four classifiers individually after the adoption of the BoW, and again raising the performance rate is noticed in terms of correctly classified instances when they are compared to the classifiers as they showed their performances before using AdaBoost. This research showed that the proposed approach could be used as a new model in the machine learning field and medical area. So, the combination of the BoW and AdaBoost has been selected to be trained against the datasets, then to be used for the testing phase with the aim to diagnose diseases from the medical check-up test reports.

The final step in the proposed approach, after diagnosing the name of the disease and display to the physician and the patient, is displaying a sample image to them to understand the case easier by the patients with the name of the disease on the top of the image. The image is not the infected organ image of the patient, it is just a sample to have an idea on the disease. Finally, the disease name with the sample image will be the output from the proposed approach as shown in Fig. 2. And the structural model of the proposed approach is shown in Fig. 3.

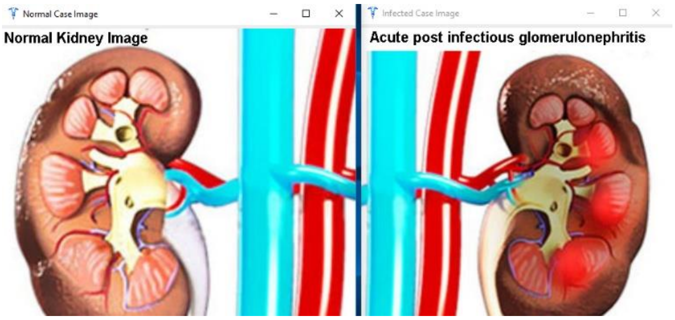


Fig. 2: A sample of normal and infected kidney images as the output of the proposed approach

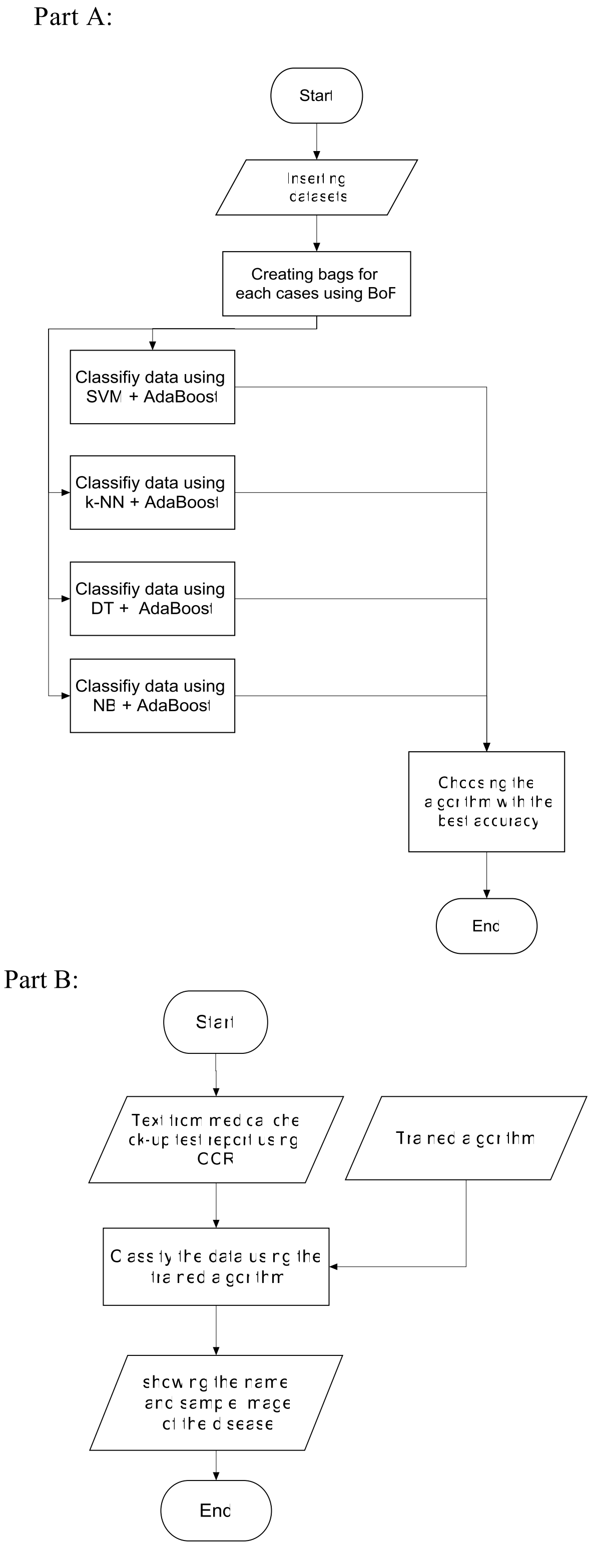

Part A:
Start
Inserting datasets
Creating bags for each cases using BoF
Classifiy data using SVM + AdaBoost
Classifiy data using k-NN + AdaBoost
Classifiy data using DT + AdaBoost
Classifiy data using NB + AdaBoost
Choosing the algorithm with the best accuracy
End
Part B:
Start
Text from medical check-up test report using OCR
Trained algorithm
Classify the data using the trained algorithm
showing the name and sample image of the disease
End

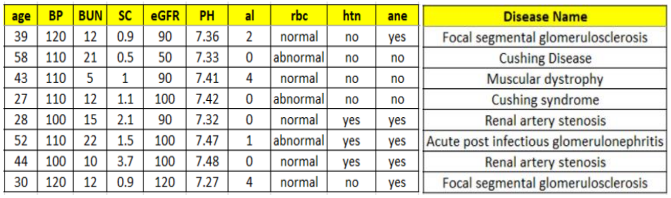
Fig. 3: Shows the methodology of the proposed model. Part A shows the methodology of the training phase, and part B shows the methodology of the testing phase

The materials and the methods used in this work are the followings:

A.Performance Datasets

Three different medical check-up test datasets of patients were collected from different sources, each contains different medical tests and values for each of the tests, but they do not contain any disease names, for that purpose, with the help of physicians the disease names were added to each record for all the three datasets as shown in the last column in Table I. The first dataset contains heart disease tests, obtained from the University of California, Irvine (UCI) database, which is one of the world's greatest databases for scientific datasets. The dataset contains medical records; each record contains 14 different attributes for 303 different patients which have been collected by four researchers from four different medical centers [7]. The second dataset contains Kidney disease tests, obtained from Medya Diagnostic Center (MDC) in Erbil, Iraq, which is a high qualified diagnostic center in the Middle East. The MDC dataset contains 10 different tests for 576 patients. And the last dataset contains Thoracic disease tests, also obtained from the UCI database, the dataset contains 17 different tests for 470 patients which have been collected by two researchers from Wroclaw University [8]. Finally, 70% from each of the three datasets were partitioned as the training set and 30% as the testing set. The number of samples, training, and testing set instances of each dataset are shown in Table II.

TABLE I: A PIECE FROM THE MDC DATASET.



age	BP	BUN	SC	eGFR	PH	al	rbc	htn	ane	Disease Name
39	120	12	0.9	90	7.36	2	normal	no	yes	Focal segmental glomerulosclerosis
58	110	21	0.5	50	7.33	0	abnormal	no	no	Cushing Disease
43	110	5	1	90	7.41	4	normal	no	no	Muscular dystrophy
27	110	12	1.1	100	7.42	0	abnormal	no	no	Cushing syndrome
28	100	15	2.1	90	7.32	0	normal	yes	yes	Renal artery stenosis
52	110	22	1.5	100	7.47	1	abnormal	yes	yes	Acute post infectious glomerulonephritis
44	100	10	3.7	100	7.48	0	normal	yes	yes	Renal artery stenosis
30	120	12	0.9	120	7.27	4	normal	no	yes	Focal segmental glomerulosclerosis

TABLE II: NUMBER OF THE SAMPLES, TRAINING, AND TESTING SET INSTANCES OF EACH DATASET

Dataset name	Total no. of samples	No. of training samples	No. of testing samples
UCI (heart)	303	212	91
MDC (kidney)	576	403	173
UCI (thoracic)	470	329	141

B.Feature Selection

In machine learning, feature selection is the process of selecting a subset of relevant variables or features to be used in model construction. It is also known as variable subset selection, and attribute selection [9]. Feature selection methods work to decrease and remove irrelevant features in the sets which have no impact on the output. It enhances the data for the next step, thus the classification of new instances will be more accurate [10].

In the proposed approach, the BoW is chosen to be used as feature selection algorithm, since it has a great capability for selecting features. The algorithm creates bags for each instance types, in this work bags are created for each disease types for all the three datasets used.

C. Bag-of-Words (BoW)

It also called Bag of Features (BoF) and vector space model [11]. In this technique, a text in a document is demonstrated as the bag (multiset) of its words (or bag of its features). The order of the word and the grammar does not affect the output, but the number of occurrences affects. The BoW model is also used widely in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and computer vision areas. This model is commonly used in applications using document classification algorithms where the occurrence of each word is used as a feature for the classifier training [12].

D. Classification of the Data

The classifiers used in this work are the followings:

1)Support Vector Machines (SVM)

The SVM is a Machine Learning classification algorithm, and it is generally identified as a separating hyperplane. An optimal hyperplane is the output of the algorithm that categorizes the entities. In two-dimensional plane the algorithm separating hyperplane is a line that separates the space into two parts, each class type will be in each side [13]. The SVM finds a separating hyperplane in an N-dimensional space, where N is the number of class types. The main goal of SVM is to find a separating hyperplane with the maximum margin, since for separating classes of entries, there are more than one possible separating hyperplanes to do the task of categorization.

2) k-Nearest Neighbors (k-NN)

The k-NN is a non-parametric and instance-based learning algorithm and is commonly used for classification and regression [15]. In the feature space, the input consists of k closest training examples for both the classification and regression. It flows into lazy learning or instance-based learning. The k-NN algorithm is one of the simplest and common machine learning algorithms. It is a very useful technique, and it can be used to assign weight to each object with the help of its neighbor objects. This way the nearer neighbors have more impact on the average than the others which are more distant. For example, the weighting scheme of each point is done by giving each neighbor a weight which is equal to 1/d, where d is the distance to each neighbor from that object.

3) Decision Table (DT)

It is a brief visual representation for identifying which actions to be performing according to the given conditions. The output of the Decision Table is a set of actions. Each of the decisions communicates to a variable, or verify whose reasonable values are listed among the conditions, and the actions are operations to be performed. Rather than the four basic quadrant structure, decision tables widely vary according to the conditions and actions represented. Sometimes it uses simple true and false values to represent the conditions, sometimes it uses numbers for the conditions, and few times it uses probabilistic representations for conditions [14].

4) Naïve Bayes (NB)

It is discovered and falls into the community of text retrieval in 1960 [12]. It could remain as a baseline technique in the domain of text categorization. In machine learning, NB classifiers are a group or family of "probabilistic classifiers" that are simple and based on applying the theory of Bayes with assumptions between the features that are strong and independent. It is popular to be used as a solution to the problem of document categorization to decide that to which category it belongs, by using word frequencies as the features for the categories. NB classifiers are scalable as a high level; in learning problems it requires a number of parameters linear with the number of features (variables). One of the common fields that it deals to work with, is automatic medical diagnosis applications [16].

5) AdaBoost

AdaBoost is shortened for Adaptive Boosting. It is a meta-algorithm of machine learning, which is invented by Schapire and Freund [17]. In AdaBoost, the base algorithm will be boosted or improved in an iterative manner by selecting the training set depending on the accuracy of the previous training iteration. And at any iteration, the accuracy weight of the trained algorithm depends on the best accuracy which is achieved in the previous iterations. AdaBoost is more sensitive than the other learning algorithms when encountering outliers and noisy data. The boosting algorithm of the Schapire and Freund which is called AdaBoost was the first algorithm in the field of machine learning. Nowadays, it is one of the most studied and widely used algorithms that is used in a huge number of applications in different fields [18].

III. Results and Discussion

Classification is done for all the three datasets after splitting them into 70% as the training set and 30% as the testing set, then using four different machine learning classification algorithms; SVM, k-NN, NB and DT to classify the data.

Firstly, SVM, k-NN, NB and DT algorithms are separately tested against the three datasets without employing the BoW as feature selection algorithm. Table III shows the complete accuracy rate in terms of Correctly Classified Instances (CCI) for each algorithm. Then, SVM, k-NN, NB and DT algorithms are separately tested against the three datasets after employing the BoW as feature selection algorithm. Table IV shows the complete accuracy rate for each classifier after selecting the features using the BoW algorithm. As it can be seen in Tables III and IV, the average of the accuracy rates of classification for all the three datasets after selecting the features using BoW for SVM is raising from 0.869 to 0.952, k-NN from 0.864 to 0.959, NB from 0.836 to 0.896 and DT from 0.806 to 0.865. According to the results recorded, the results are raised by %7.4 ratio after employment of BoW as it is compared to the same results before employing the BoW. So, BoW algorithm will be used as the feature selection algorithm in the proposed approach.

TABLE III: ACCURACY OF THE CLASSIFIERS BEFORE EMPLOYING BOW

Dataset name	SVM	k-NN	NB	DT
UCI (heart)	0.878	0.843	0.778	0.753
MDC (kidney)	0.923	0.929	0.900	0.878
UCI (thoracic)	0.807	0.821	0.830	0.788
Average	**0.869**	**0.864**	**0.836**	**0.806**

TABLE IV: ACCURACY OF THE CLASSIFIERS AFTER EMPLOYING BOW

Dataset name	SVM	k-NN	NB	DT
UCI (heart)	0.905	0.913	0.860	0.816
MDC (kidney)	0.988	0.985	0.919	0.898
UCI (thoracic)	0.963	0.981	0.910	0.883
Average	**0.952**	**0.959**	**0.896**	**0.865**

In this research, the combination of the BoW with the AdaBoost techniques is used for enhancing the performance of the algorithms to obtain better results. The goal behind doing this is to select the algorithm with the best performance among each of the specified records. Table V shows the results from the combination of the BoW with the AdaBoost technique and each of the SVM, k-NN, NB, and DT algorithms are used as base classifiers for the AdaBoost.

TABLE V: ACCURACY OF BOOSTED CLASSIFIERS USING ADABOOST AFTER EMPLOYING BOW

Dataset name	SVM	k-NN	NB	DT
UCI (heart)	0.947	0.904	0.965	0.904
MDC (kidney)	0.982	0.982	0.982	0.952
UCI (thoracic)	0.972	1	0.990	0.944
Average	**0.967**	**0.962**	**0.979**	**0.933**

As a result, AdaBoost technique with the base classifiers each at a time are used to classify the data after using BoW for selecting the features, Table V shows the results of boosted of all the four classifiers after using BoW as feature selection algorithm. Comparison between the accuracy of the algorithms is done, and the algorithm with the highest accuracy is chosen with the aim to be used for classifying the medical data, and diagnosis of diseases from medical check-up test reports in the proposed approach. According to the results of the four classifiers after boosting as shown in Table V, the NB algorithm that is used as the base classifier for the AdaBoost technique after employment of BoW is chosen as the algorithm with the highest accuracy which is 0.979.

Finally, after inserting the medical check-up test reports into the proposed model and converting them into the textual format by using OCR, and some processing is done onto the output of OCR by removing all the information located on the medical reports except the test names and test values. The test names and their results are analyzed and processed according to the proposed model that is trained against the three datasets. Finally, the name of the disease, a normal kidney image and a related image that describes the disease case will be displayed to the physician and patient as shown in Fig. 2.

The researchers in [19] proposed an algorithm to diagnose heart diseases from medical records of 920 patients, in which they made a comparison between three algorithms, the best was Bagging with Naïve Bayes algorithm with the accuracy rate 0.940. As a comparison made between the model proposed in the mentioned research with the one proposed in this research, we can reach a decision, in which the proposed model in this research with the worst accuracy rate which is 0.965 when it is tested against UCI (heart) dataset, outperforms its counterpart one with the accuracy rate 0.940.

IV. Conclusion

Medical data classification is a sensitive field of study. It is related to human's life. Therefore, the studies done in this field need to be specific and should have very good accuracy rates. It is obvious that further studies and developments are required to be performed in this domain. The proposed approach in this study takes the medical check-up test report as input and extracts the text on the medical reports using OCR, performs feature selection using the BoW algorithm and performs classification using AdaBoost technique and NB as the base classifier. The proposed approach is tested against three different datasets. The datasets contained different medical test results for patients.

Experimental results showed that using the BoW with the AdaBoost technique improves the results and increases the accuracy rate as mentioned previously in the results and discussion section. The result is obtained when it is trained and tested against three different datasets that contained medical tests. So, finally, this paper concludes to develop an approach by using the combination of BoW as feature selection algorithm and AdaBoost technique and NB as the base classifier to obtain accurate results for diagnosing diseases from medical check-up test reports.

Another perspective of future work is introduced during the development of the proposed approach. While working on that case, we observed that there are no limitations to develop approaches with better accuracy rates. So, different approaches with high accuracy rates can be proposed for future works.

Acknowledgment

We are highly thankful for all the sources we have gotten data and instructions even a statement and especially for Dr. Abdullah Chalaby and Dr. Aso Dizayi for helping us in medical issues and the MDC and UCI for providing the datasets of medical tests that were used in this research.

References

[1] R. S. Evans, "Electronic health records: then, now, and in the future," Yearbook of medical informatics 25.S 01 (2016): S48-S61.

[2] Zhou, Zhi-Hua, "Ensemble methods: foundations and algorithms," Chapman and Hall/CRC, 2012.

[3] V.Aggarwal, Jajoria, S. and Sood, A., "Text Retrieval from Scanned Forms Using Optical Character Recognition," In Sensors and Image Processing (pp. 207-216). Springer, Singapore, 2018.

[4] R. Bouslimi, , and Jalel Akaichi, "New approach for automatic medical image annotation using the bag-of-words model," 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2015.

[5] H. Gao., Bowles, E.J.A., Carrell, D. and Buist, D.S., "Using natural language processing to extract mammographic findings," Journal of biomedical informatics 54 (2015): 77-84.

[6] K.A Abuhasel, A. M. Iliyasu, and Chastine Fatichah, "A combined AdaBoost and NEWFM technique for medical data classification," Information science and applications. Springer, Berlin, Heidelberg, 2015, 801-809.

[7] Blake, Catherine L., and Christopher J. Merz, "UCI repository of machine learning databases, 1998," (1998).

[8] M. Zięba, Tomczak, J.M., Lubicz, M. and Świątek, J., "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," Applied soft computing 14 (2014): 99-108.

[9] Y. Saeys, I. Inza, and Pedro Larrañaga, "A review of feature selection techniques in bioinformatics," bioinformatics23.19 (2007): 2507-2517.

[10] M.L. Bermingham, Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P. and Haley, C.S., "Application of high-dimensional feature selection: evaluation for genomic prediction in man," Scientific reports 5 (2015): 10312.

[11] J. Polpinij, and Aditya K. Ghose, "An ontology-based sentiment classification methodology for online consumer reviews," 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Vol. 1. IEEE, 2008.

[12] Russell, Stuart J., and Peter Norvig, "Artificial intelligence: a modern approach, " Malaysia, Pearson Education Limited, 2016.

[13] N. Guenther, and M. Schonlau, "Support vector machines," The Stata Journal 16.4 (2016): 917-937.

[14] R. Kohavi, "The power of decision tables," European conference on machine learning, Springer, Berlin, Heidelberg, 1995.

[15] N.S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician 46.3 (1992): 175-185.

[16] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[17] R.E. Schapire, "Explaining adaboost," Empirical inference, Springer, Berlin, Heidelberg, 2013. 37-52.

[18] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," Journal-Japanese Society For Artificial Intelligence 14.771-780 (1999): 1612.

[19] Tu, My Chau, D. Shin, and D. Shin, "A comparative study of medical data classification methods based on decision tree and bagging algorithms," 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. IEEE, 2009