DRUG DISCOVERY WITH HIGH ACCURACY USING HYBRID TEXT MINING THROUGH IVR(INSIGNIFICANT VALUE REMOVAL) AND NATURAL LANGUAGE PROCESSING

Neha Sharma
Mtech Scholar
Sri Sai College of engineering and technology

Deepak Kumar
Assistant Professor
Sri Sai College of engineering and technology

ABSTRACT

Drug discovery using text mining is the antidote for laying the stone for accurate classification of diabetic disease. The objective of this study is to uncover diabetic based on medicines prescribed to patients using the applications of natural language processing and text mining. To this end, proposed work is partitioned into phases. First phase, pre-processing yield purified data by eliminating any abnormalities or noise using IVR(insignificant value removal) mechanism. Natural language processing then plays critical part on pre-processed data by extracting meaningful words and eliminating stop words. Next phase i.e text mining forms clusters. Earlier work kept the hold rate at 0.2 but in proposed work hold out rate is changed to 0.3 and classification accuracy is greatly improved. Result in terms of sensitivity , specificity and F-score shows improvement by 2%.

Method Used

Entire simulation is divided into primarily two phases.

1. Preprocessing: At this stage resizing operation is performed since input layer of the network requires predefined size. 20% hold out rate is maintained for testing data.
2. Text parsing and Classification: Multi class text parsing is used to extract critical and non critical segments from within the training dataset. After extracting the features, classification is performed on the test data.

Parameters

Parameters used for optimization includes classification accuracy.

Simulation and Result

Simulation is conducted in MATLAB using image processing and neural network toolbox. The proposed mechanism shows improvement in terms of classification accuracy by the margin of 3% which is significant, thereby enhancing the recognition rate.

Keywords: Diabetes, Deep learning, Accuracy.

I. INTRODUCTION

Diabetic condition is a serious and broadly spread disease all over the world. It is the commonest reason of legal blindness and kidney failure in the working-age population of created nations[1]. Diabetic condition happens when diabetes harms the blood vessels inside the body, leaking blood and fluids into the tissue. This liquid fluid produces microaneurysms, hemorrhages, hard exudates, and cotton wool spots (a.k.a., soft exudates)[2]. Diabetic condition is a noiseless disease and may just be perceived by patients when changes in the vessels have advanced to a level where treatment becomes difficult and even impossible.

The critical approach of filtering of data set used to discover normal and abnormal patterns from the database is step by step analysis process. Filtering of data set is the process of extraction of useful information from large database. The fetched information must be converted into user understandable form for future use. Mining approaches used at different places vary according to size and complexity of problem in hand.[3] Mining approaches useful for detecting patterns from the database includes web, text, sequential and temporal mining. Step by step analysis process is the process of discovering patterns that are frequent within database. [4]The

interest in pattern mining grown due to its ability to discover the hidden patterns within the database, that are useful for the users and cannot be extracted manually. Patterns category discovery is vital for successful interpretation of the disease.

The step by step analysis process finds out frequent pattern from the sequence database. [5]The well-known pattern mining methods are utilized for web-log analysis, medical record analysis and disease prediction. It identifies strong symptom/disease correlations which can be valuable information for the diagnosis and preventive medicine.

The expanding number of diabetic condition cases overall requires to strengthen the creation of instruments to determine diabetic condition. Programmed recognition of diabetic condition will save time and efforts. In this manner, S.rubhini et al[6]. proposed a technique for programmed discovery of microaneurysms in retinal fundus pictures. Maher et al [7] already assessed a choice based emotionally supportive network for programmed screening of non-proliferative diabetic condition. Truth be told, support vector machines were utilized by Maher et al. [4] in the computerized analysis of non-proliferative diabetic condition. A few picture pre-preparing strategies have additionally been proposed keeping in mind the end goal to distinguish diabetic condition in [8]–[10]. However, regardless of all these past works, mechanized discovery of diabetic condition still remains a field for development [11].

Hence, this paper proposes another computerized handling of retinal images with a specific end goal to help individuals recognize diabetic condition in advance. The end goal of proposed literature is to achieve desired level of classification accuracy by reducing noise levels from within the 1-3 levels of non-proliferative dataset where '1' indicates mild DR, '2' indicates moderate DR and '3' indicates severe or proliferative DR. Noise handling through Gaussian filtering is used at pre-processing stage. Gaussian filter is capable of handling noise at edges and also considered the best filters in time domain. Resizing operation is done at preprocessing stage thus, ensuring uniformity along the input layer for faster operation. MSVM gives multiclass segmentation and classification operation. As an output, we acquired a most extreme affectability of 94.6% and estimation capacity value of 93.8%. Heartiness as for changes in the parameters of the calculation has also been evaluated.

Details of proposed system are discussed in the 4 section. Rest of the paper is organized as under: section 2 gives literature survey of techniques used to detect DR, section 3 gives the phases associated with proposed system, section 4 gives the performance analysis, section 5 gives conclusion and future scope followed by the references at the end.

2 Literature survey

This section provides in depth the techniques used to detect diabetic patients at an early stage and suggest appropriate actions. [9]uses the contrast enhancement, morphological filtering and segmentation technique to detect hard exudes from the various input image. The system utilizes Contrast Limited Adaptive Histogram Equalization (CLAHE) technique to enhance the image and top hat transform to enhance the blood vessels. After that filtering is to be done and pattern recognition techniques are utilized to recognize the diseases. In [12] SVM classifier based system is utilized in to diagnose DR affected patient. It also uses test fundus images to contribute to SVM classifiers.

Technique discussed by [13] provides mechanism to detect diabetic retinopathy by the use of Deep learning. Considerably large dataset is used for this purpose. Data driven artificially intelligent deep learning mechanism is used to derive training and testing images for distinguishing normal image from DR image. However, mechanism used lacks preprocessing mechanism including noise handling within the automatic detection of DR and multiclass classification. Several other techniques of data mining in healthcare are surveyed by [14].

Alzahrani ,proposed data mining method for disease prediction[15] for this purpose sequential data mining is used in order to accomplish this data preprocessing mechanism . After applying preprocessing mechanism the attributes will be analyzed this will be done using passes on medical data. The first pass determines whether support for each disease is present or not at the end of this phase the frequent disease within the database will be identified, a counter will be maintained to count the occurrence of each disease within the dataset. Next phase determines the second sequence of diseases present within the dataset. The overall process yields the diseases which can cause the occurrence of other diseases. The disease resulting in another disease is termed as candidate generation. And for declaring that it is generated from the previous level Pruning is used.

Algorithms associated with data mining provides the filtering mechanism[16], [17] to ensure the better classification of result. By analyzing discussed techniques, best possible technique can be selected for future enhancement. Low cost medical image processing mechanism is proposed by [18]. Field programmable field array is merged along with the processor for analyzing the complex diseases like DR. [19] Utilizes microneurysm segmentation that is automatically done by using mathematical morphology. It does not work on predefined set of directives. Sensitivity and specificity can also be further improved .[11] uses multiclass SVM classifier[20], [21] that assure classification phase thus ensuring unwavered analysis of human observer and also presents supervised classifier that uses testing sets to obtain results. [22] classifies the

DR stages using automated system having feature classifier. It detects the disease by extracting features using image processing method and classifies them accordingly. [23] SVM and MDA methodologies are surveyed in this paper and also their utilization for detecting diabetic patients are explained.

Kunjir,et al., proposed multiclass Naïve Bayes algorithm that is used for prediction of particular disease. This is downloaded from UCI repository work. The proposed system can help doctors to take clinical decisions where traditional decision support system fails, J47 algorithm is also used for proving the worth of study of accuracy in diabetes disease, breast cancer and diabetes approaches 83% by using this approach[24]. Thus accuracy requires improvment in future.

Alamanda, et al., proposed sequence pattern mining in order to detect the time duration used for promotion .The sequence or pattern is checked within the database. The weight of each sequence in each database is achieved from the interval of the successive element in the sequence and the mining is performed on the basis of weight considering time interval[25]. Time interval based pattern is used in this case. In preprocessing missing values are not considered.

Ghosh,et al. [26],proposed a technique that extract sequential patterns from hypotensive patient groups. These patterns are further utilized to inform medical decisions and randomized clinical trials. It further extended by including various clinical features and also include some sequential patterns. It also does not considered missing value during the preprocessing phase.

Zhang,et al., proposed a technique named ConSgen that are used to identify the contiguous sequential generator and also minimize the redundant patterns, It utilizes the divide conquer technique to find the sequential generator with contiguous constraints[27]. But it does not consider the gapped alignments and also not discovered the binding sites.

M. Zihayat et al., identified a problem of top –k utility based regulation pattern which is used to find out meaning in biology. Firstly proposed a utility model called TU-SEQ which is used to find top –K high utility gene regulation sequential patterns[28]. It is considering the relation between the various patterns and interactions in biological studies.

Y.song et al.[29], proposed a mining technique that are used to reduce the complexity and cost of the data storage. It divides chunks into separate parts and regression analysis is to be done to analyze the trial variable and samples dataset. But it does not considered separate chunks for feature analysis and separate storage reservoir also not utilized.

Al-khasawneh ,proposed an application that utilizes the data mining technique to predict the diabetes disease[30]. Also it guides the patient to take treatment at early stage. But is completely dependent upon patient input and does not considered predefined dataset values. It does not utilize the missing value that is essential to predict diseases.

Abbasghorbani et al., analyzed various pattern mining techniques and the features of all the algorithms. It introduced various minimizing support counting which is used for minimizing search space[31]. We have generated small search space which will include earlier candidate sequence pruning then database is analyzed and compression technique is used to analyze.

The comparative analysis of the literature is also presented in this section.

Table 1: Comparative analysis of literature for diabetic detection

Author	Technique	Parameters	Merit	Demerit
Bulsara et al. 2011	Field programmable array for DR detection	Accuracy	This mechanism is fast and efficient and can handle small dataset easily	Large dataset handling leads to poor classification accuracy
Paranjpe & Kakatkar 2013	Contrast and Segmentation mechanism (CLAHE)	Classification Accuracy	Improved mechanism of segmentation by eliminating noise and hence better classification accuracy	No pre-processing mechanism for handling stop words
S.rubhini et al 2014	Eigen value based classification	Sensitivity , accuracy	It is discovered microaneurysms in retinal fundus pictures.	Has not handle segmented data efficiently
Jothi et al. 2015	Analysis of Data mining approach	Classification Accuracy	Relevant literature of Deep learning	No new mechanism to

			for disease detection is conducted	handle large dataset is proposed
Zhang,et al.2016	ConSgen	Sensitivity , accuracy, specificity	It utilizes the divide conquer technique to find the sequential generator with contiguous constraints	It does not consider the gapped alignments and also not discovered the binding sites.
M. Zihayat et al., 2017	a utility model called TU-SEQ	F-score, Classification accuracy	It is considering the relation between the various patterns and interactions in biological studies.	It does not handle complex dataset efficiently.
Ramya 2018	SVM classifier for Diabetic retinopathy	Classification Accuracy, Specificity, sensitivity	SVM classifier based on two hyperplanes for DR detection	Multiple problems due to DR cannot be detected
Zhou et al. 2018	Deep Learning for DR detection	Accuracy, Sensitivity, F-Score	Deep learning for heavy data can be used efficiently for better classification accuracy	Smaller dataset cannot be handled

The parameters used in various literatures differ in characteristics. The parameters used in existing literature differ in length and accuracy. In addition some literatures do not used relevant parameters that could test relevance accuracy of the techniques.

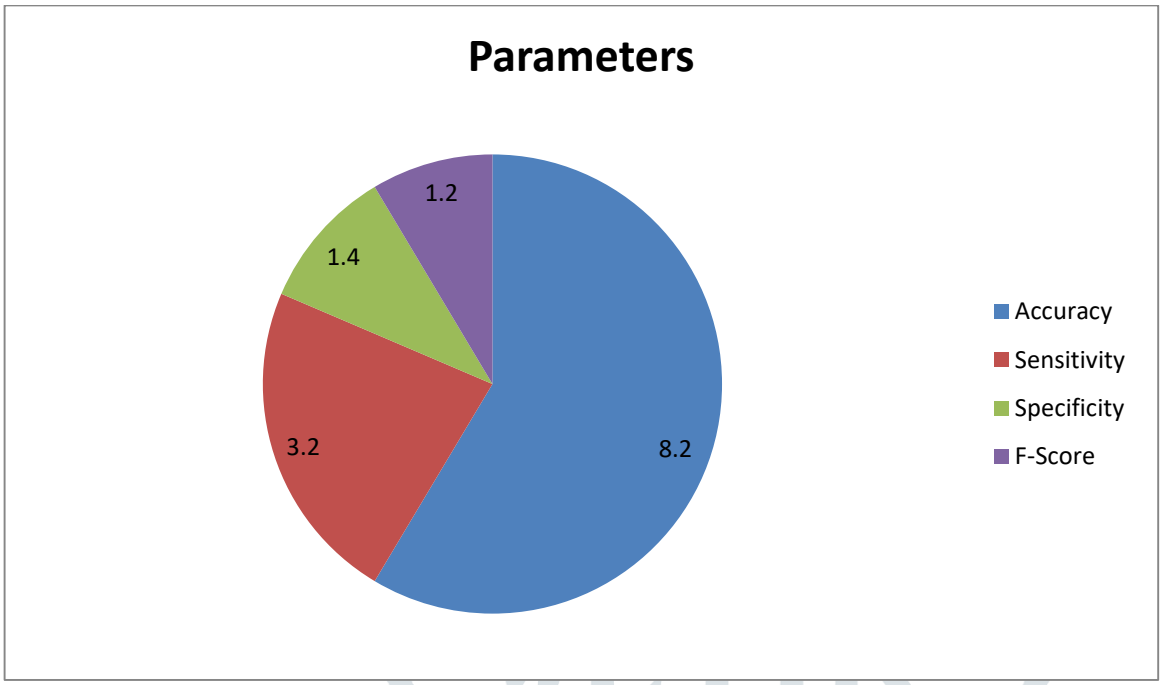
Table 2: Comparison of parameters used in the existing literatures

Author	Accuracy	Specificity	Sensitivity	F-Score
Paranjpe & Kakatkar 2013	✓	X	X	X
Ramya 2018	✓	✓	✓	X
Zhou et al. 2018	✓	X	✓	✓
Jothi et al. 2015	✓	X	X	X
Bulsara et al. 2011	✓	X	X	X

From the comparative analysis it is clear that in most of the literatures Specificity and F-Score parameters are not considered. Hence classification accuracy shows deviation in the analysis of multiple datasets.

The parametric comparison in terms of plot is given in figure 1

Figure 1: Comparison of parameters used in existing literatures for Diabetic detection



In the proposed system we try to use almost all the parameters for validating the output generated through the proposed system.

3. Proposed system

PRE-PROCESSING

Preprocessing mechanism used in this literature contains noise handling in terms of missing data along with resizing operation. Noise handling is done using most significant value replacement from dataset. This filter is capable of handling missing data along with smoothening operation. Equation 1 gives the operation of filtering along with smoothening.

$$G_{Smoothened_{image}} = \frac{1}{2\pi\alpha^2} e^{(a^2+b^2)\frac{1}{2a^2}}$$

Equation 1: Gaussian Filtering

'α' is standard deviation, 'a' is distance from horizontal axes and 'b' is a distance of origin from vertical axes.

After handling noise, resizing operation is done. Resizing is done to present the uniform data to the input layer. Resizing is done using equation 2.

$$Resized_G = Resize\left(G_{Smoothened_{image}}, [x\ y]\right)$$

Equation 2: Resized image

This resized image set obtained is passed to the network for further processing.

III C) TRAINING

Training operation begins by receiving the image set from the pre-processing phase. To create a new network, proposed mechanism used inbuilt layers with input layer accepting images of XxY with 3 channels. Training parameters is defined using training option command using deep learning toolbox. Train network function is used to finally train the network. Flow of network definition is given as under

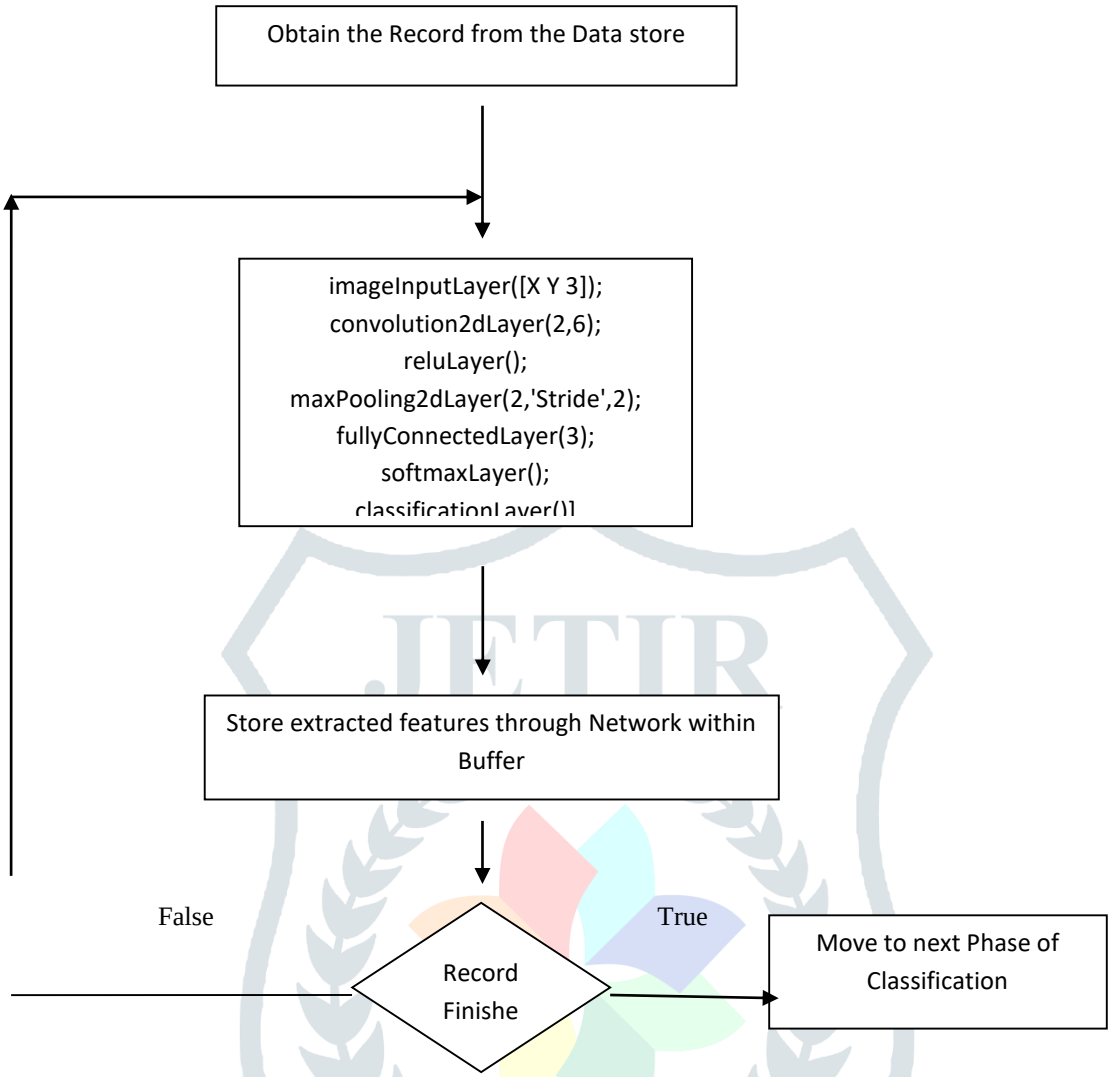


Figure 2: Flow of Training Process

III D) CLASSIFICATION

The classification is done using text parser with kernel function that combines the features to contribute to a significant improvement in accuracy. For instance, in the task of dependency parsing, it would be hard to confirm a correct dependency relation with only a single set of features from either a head or its modifier. Rather, dependency relations should be determined by at least information from both of two phrases.

4. Result

In this research work, there are three measures used. Correctly classified instances are properly classified by any classification technique.

Accuracy is calculated by an exact value.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The mentioned rule for the accuracy calculation the above mentioned formula is used with TN = True Negative.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Sensitivity determines the amount of true positive values predicted. True positive rates should be high for better classification accuracy.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Specificity is the metric determining amount of negative values prediction through the technique specified. Since this metric deals with the negative result hance its value in percentage must be reduced.

The performance of the system is analyzed by the use of parameters such as accuracy, specificity and sensitivity.
Result in terms of classification accuracy by determining amount of true positive and true negative values is given through figure 3

Table 3 Classification accuracy of existing and proposed system

Disease Predicted	With 0.2 Hold out rate(%)	With 0.3 Hold out rate(%)
Level 1 Diabetic(Mild)	85	95
Level 2 Diabetic(Moderate)	85	95
Level 3 Diabetic(High)	86	91

The plot for Table 3 is given in figure 3.

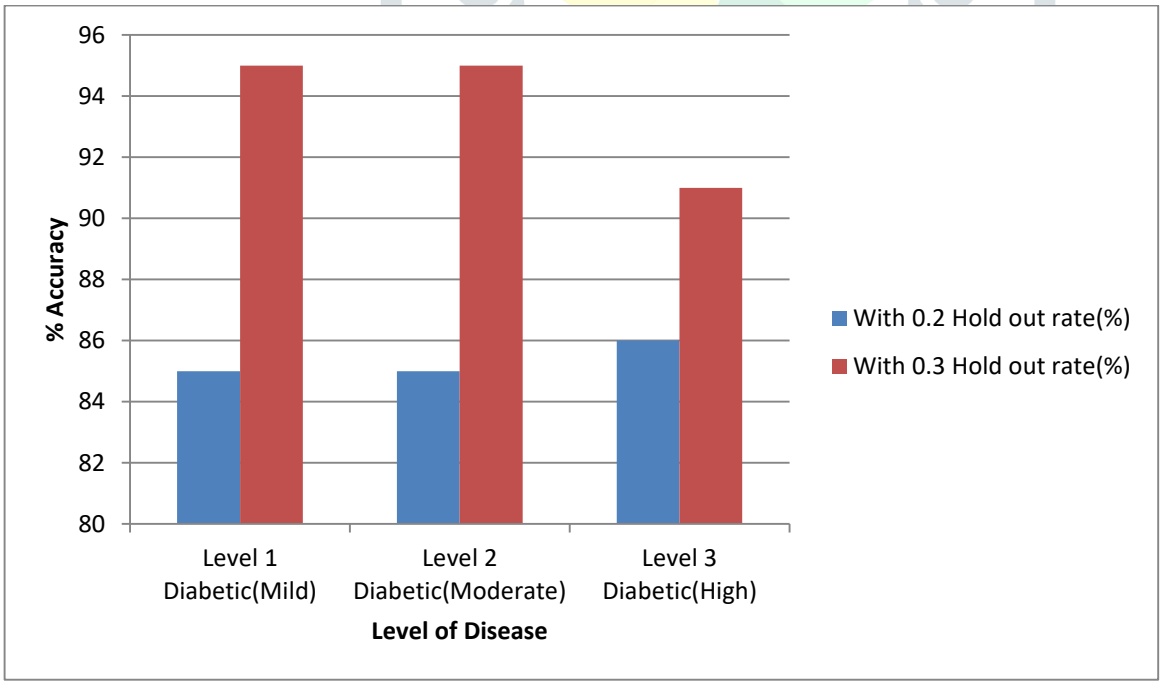


Figure 3: Classification accuracy

The sensitivity of proposed system is better as compared to existing system. The primary reason for the same is better hold out rate and removal of insignificant values from the dataset.

Table 4: Sensitivity Prediction through existing and proposed literature

Disease Predicted	With 0.2 Hold out rate(%)	With 0.3 Hold out rate(%)
Level 1 Diabetic(Mild)	84	92
Level 2 Diabetic(Moderate)	87	97
Level 3 Diabetic(High)	87	96

Sensitivity is a measure of true positive predicted by the employed system. Higher values of sensitivity indicate better performance of the proposed system. This is also indicated through figure 4.

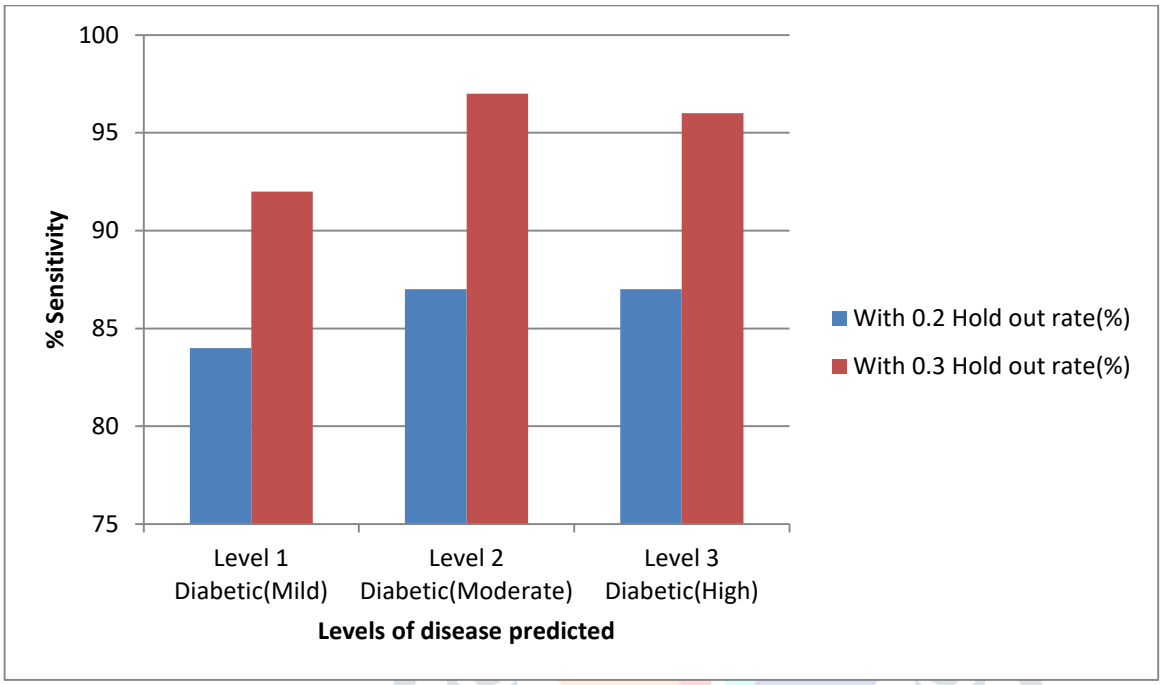


Figure 4: Plots of sensitivity against existing system

The specificity that is a measure of true negative values must be less to prove worth of the study. The observed specificity values are significantly lower and hence better classification accuracy is obtained through the proposed system. Specificity is indicated through table 5

Table 5: Comparison of Specificity

Disease Predicted	With 0.2 Hold out rate(%)	With 0.3 Hold out rate(%)
Level 1 Diabetic(Mild)	84	80
Level 2 Diabetic(Moderate)	80	78
Level 3 Diabetic(High)	87	84

There is a significant difference between existing system without insignificant removal mechanism and with insignificant value removal mechanism. in addition better hold out rate also enhanced classification accuracy. This is indicated through figure 5

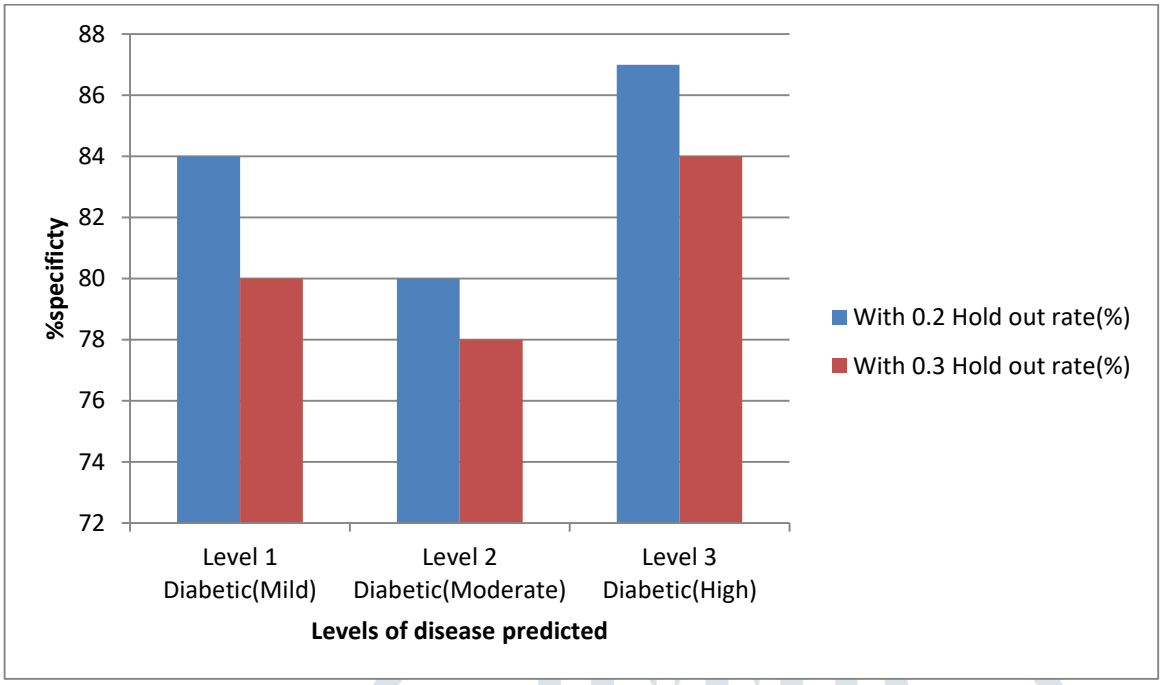


Figure 5: Specificity comparison of existing and proposed system

The medicine detection and prediction is given though accurate classification, result in terms of plots is given as under Result comparison in terms of accuracy, sensitivity and specificity are given as under

Table 6: Aggregate Result with 0.3 and 0.2 hold out rates

Disease Predicted	Parameters	Existing (%)	Proposed (%)
Level 1 Diabetic(Mild)	Accuracy Specificity Sensitivity	85 84 84	95 80 92
Level 2 Diabetic(Moderate)	Accuracy Specificity Sensitivity	85 80 87	95 78 97
Level 3 Diabetic(Severe)	Accuracy Specificity Sensitivity	86 87 87	91 84 96

Classification accuracy of proposed system appears to be more as compared to existing techniques. Multiple class prediction mechanism showing higher accuracy proving the worth of study.

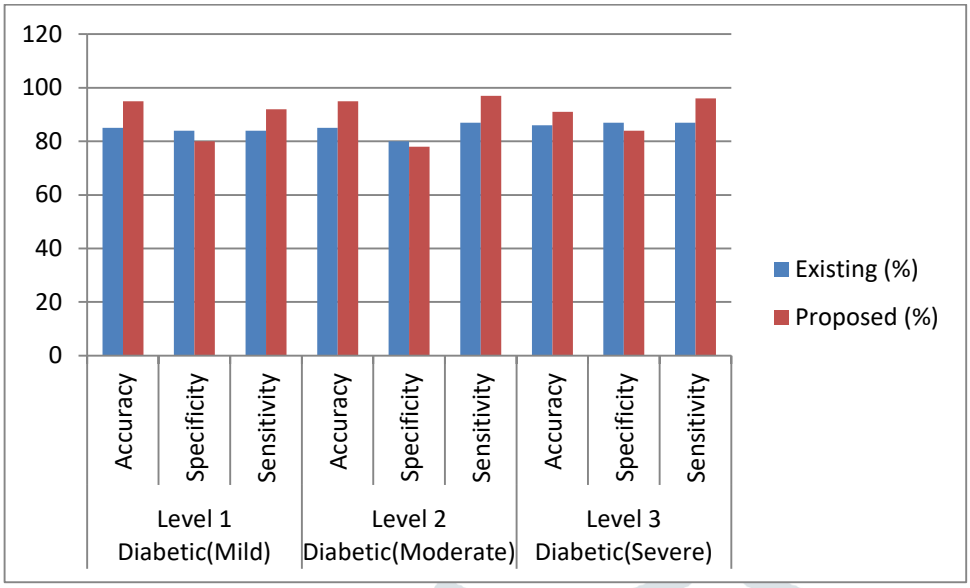


Figure 6: Aggregate result of existing and proposed system

Results and performance analysis as indicated through the plot shows that proposed algorithm with increased holdout rate and insignificant value removal algorithm yield better result.

5. Conclusion

The proposed methodology includes data driven novel algorithm using text parsing of dataset. The performance indicators accuracy, specificity, sensitivity, precision, error rate are calculated for the given dataset. Accusation beside with a proper data preprocessing technique can get better the accuracy of the classifier. The function of data normalization had noticeable impact on categorization performance and considerably enhanced the performance of proposed methodology. To improve the overall accuracy, it is necessary to use more data set with large number of attributes and use the best feature selection method in future. Future works may also include hybrid classification models by combining some of the data mining technique.

REFERENCES

[1] R. Gargeya and T. Leng, "Automated Identi fi cation of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, pp. 1–8, 2017.

[2] M. S. Haleem, L. Han, J. Van Hemert, B. Li, and A. Fleming, "Retinal Area Detector from Scanning Laser Ophthalmoscope (SLO) Images for Diagnosing Retinal Diseases," vol. 2194, no. MARCH, 2014.

[3] J. Hu and A. Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets," *Pattern Recognit.*, vol. 40, no. 11, pp. 3317–3324, 2007.

[4] J. Lee, U. Yun, and G. Lee, "Analyzing of incremental high utility pattern mining based on tree structures," *Human-centric Comput. Inf. Sci.*, 2017.

[5] J. W. Huang, C. Y. Tseng, J. C. Ou, and M. S. Chen, "A general model for sequential pattern mining with a progressive database," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1153–1167, 2008.

[6] S. S. Rubini and A. Kunthavai, "Diabetic retinopathy detection based on eigenvalues of the hessian matrix," *Procedia Comput. Sci.*, vol. 47, no. C, pp. 311–318, 2014.

[7] R. Maher and M. Dhopeshwarkar, "Automatic detection Non-proliferative Diabetic Retinopathy using image processing techniques," *J. Eng. Res. Appl.*, vol. 6, no. 1, pp. 122–127, 2016.

[8] S. M. Student, C. Landran, and G. Kaur, "Review on: Detection of Diabetic Retinopathy using SVM and MDA," *Int. J. Comput. Appl.*, vol. 117, no. 19, pp. 975–8887, 2015.

[9] M. J. Paranjpe and P. M. N. Kakatkar, "Automated Diabetic Retinopathy Severity Classification using Support Vector Machine," *Int. J. Res. Sci. Technol.*, no. 3, pp. 86–91, 2013.

[10] V. Ramya, "SVM Based Detection for Diabetic Retinopathy," *IEEE*, vol. V, no. I, pp. 11–13, 2018.

[11] P. Adarsh and D. Jeyakumari, "Multiclass svm-based automated diagnosis of diabetic retinopathy," *Int. Conf. Commun. Signal Process.*, pp. 206–210, 2013.

[12] V. Ramya, "SVM Based Detection for Diabetic Retinopathy," *Iccad*, vol. V, no. I, pp. 11–13, 2018.

[13] L. Zhou, Y. Zhao, J. Yang, Q. Yu, and X. Xu, "Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images," *IET Image Process.*, vol. 12, no. 4, pp. 563–571, 2018.

[14] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," *Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.

[15] M. Y. Alzahrani, "Discovering Sequential Patterns from Medical Datasets," 2016.

[16] M. Saini, "A Hybrid Filtering Techniques for Noise Removal in Color Images," *IEEE*, vol. 5, no. 3, pp. 172–178, 2015.

[17] Y. Ma, D. Lin, B. Zhang, Q. Liu, and J. Gu, "A Novel Algorithm of Image Gaussian Noise Filtering based on PCNN Time Matrix," in *2007 IEEE International Conference on Signal Processing and Communications*, 2007, pp. 1499–1502.

[18] V. Bulsara, S. Bothra, P. Sharma, and K. M. M. Rao, "Low Cost Medical Image Processing System for Rural / Semi Urban Healthcare," *IEEE Access*, pp. 724–728, 2011.

[19] S. Shetty, K. B. Kari, and J. A. Rathod, "Detection of Diabetic Retinopathy Using Support Vector Machine (SVM)," *IEEE 17th Int. Conf. Parallel Distrib. Syst.*, vol. 23, no. 6, pp. 207–211, 2016.

[20] "A Modified Median Filter for the Removal of Impulse Noise Based on the Support Vector Machines."

[21] P. Naraei, V. Street, V. Street, and V. Street, "Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data," *IEEE Access*, no. December, pp. 848–852, 2016.

[22] I. Ntroduction, "Diabetic Retinopathy Classification using SVM Classifier," *ACM Comput. Surv.*, vol. 6, no. 7, pp. 7–11, 2017.

[23] S. M. Student, C. Landran, and G. Kaur, "Review on: Detection of Diabetic Retinopathy using SVM and MDA," *Int. J. Comput. Appl.*, vol. 117, no. 19, pp. 975–8887, 2015.

[24] A. Kunjir, H. Sawant, and N. F. Shaikh, "Data mining and visualization for prediction of multiple diseases in healthcare," *Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017*, pp. 329–334, 2017.

[25] S. Alamanda, S. Pabboju, and N. Gugulothu, "An Approach to Mine Time Interval Based Weighted Sequential Patterns in Sequence Databases," *2017 13th Int. Conf. Signal-Image Technol. Internet-Based Syst.*, pp. 29–34, 2017.

[26] S. Ghosh, M. Feng, H. Nguyen, and J. Li, "Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 5, pp. 1416–1426, 2015.

[27] J. Zhang, Y. Wang, C. Zhang, and Y. Shi, "Mining contiguous sequential generators in biological sequences," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 855–867, 2016.

[28] M. Zihayat, H. Davoudi, and A. An, "Top-k utility-based gene regulation sequential pattern discovery," *Proc. - 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016*, pp. 266–273, 2017.

[29] Y. Song, G. Alatorre, N. Mandagere, and A. Singh, "Storage Mining : Where IT Management Meets Big Data Analytics," *IEEE Access*, pp. 2–3, 2013.

[30] A. Al-khasawneh, P. Al-hussein, A. Ii, and I. Technology, "A Method for Classification Using Data Mining Technique for Diabetes : A Study of Health Care Information System," *IEEE Access*, vol. 10, no. September, pp. 1–23, 2015.

[31] S. Abbasghorbani and R. Tavoli, "Survey on Sequential Pattern Mining Algorithms," *2015 2nd Int. Conf. Knowledge-Based Eng. Innov.*, pp. 1153–1164, 2015.