An Efficient Hybrid Classifier to Improve the Health Prediction using Data Mining

Narender Kaur[1], Vivek Gupta[2], Surabhi Kataria[3]
[1]*Student, Department of Computer Science and Engineering,*
[23]*Assisant Professor, Department of Computer Science and Engineering*
[123]*Rajasthan Technical University, Kota, Rajasthan, India*

Abstract- In today"s era, data analytics and machine learning have been widely used in many verticals, especially in the medical sector. Diabetes is the primary issue of global health problems in every country. In this paper, an early diabetic"s prediction model is proposed which identifies the diabetic at the very early stage so that appropriate actions can be taken by the doctors as well as patients. An early diabetics prediction model is designed by generating the patterns using the KDD process. The focus of this papers is intended to address the challenge of improving the prediction model to predict the diabetic status of patients and providing a timely response in predicting the disease. The performance evaluation of the proposed model is compared with well-known classifiers. Our hybrid classifier was evaluated on various parameters which have higher accuracy than others.

I. INTRODUCTION

In recent years, diabetes has been affected over 300 million people worldwide. As per the WHO report, it is expected to rise to over 380 million by 2025. The United States has been named it the fifth deadliest disease which has no imminent cure in sight. Data analytics and machine learning play an important role in the medical field to predict various diseases at early stages. Knowledge Data Discovery (KDD) process can be used to study and analyze the health conditions of diabetic patients. Healthcare data contains different variable types and missing values those make the data-set huge, heterogeneous and complex for further processing. The health-care data-set it first pre-processed and then transformed for applying various data mining techniques. Data-mining techniques are applied to diagnose the type of diabetes and its level of severity for each patient. Data mining process helps in extracting the useful information from large databases of diabetes available in the hospitals and medical dispensaries. It involves computational process, classification, clustering, machine learning, statistical techniques, and discovering patterns to make decisions in the future. The proposed model consists of following steps:

1. Develop an application domain: In this step, the prior knowledge which is essential for diabetes are collected. Further, the prediction goal is set from the discovered knowledge.
2. Data-set Creation: diabetes-related data- set is chosen which includes relevant variables (attributes) and data points (examples) to be used in performing discovery tasks. Since the medical data-set can have more information, the desired subset is obtained by querying the existing data-set.
3. Data cleaning and pre-processing. This step deals with noise and missing values in the data-set and accounts for time sequence information and known changes.
4. Data reduction and projection. In this step, dimension reduction and transformation methods are applied to the attributes of the data-set.
5. Choosing data miner: Here the data miner is chosen against the goals defined in Step 1 by using a particular DM method, such as classification, regression, clustering, etc.
6. Choosing the data mining algorithm: The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
7. Data mining. This step deals with representational and generates patterns such as classification rules, decision trees, regression models, trends, etc.
8. Interpreting mined patterns; Analyst performs visualization of the extracted patterns and models from the previous step.

Further, visualization is performed on the data obtained from the extracted models.

9. Consolidating discovered knowledge: Finally, the discovered knowledge is incorporated into the performance system. The documentation and reports are checked. It also resolved the potential conflicts with previously believed knowledge. In our paper diabetes in humans is classified as true- positive(diabetic patient) or true negative (non-diabetic patient). The detection of diabetes is considered as a classification problem of our proposed hybrid classifier and it is evaluated against various state-of- the-art classifiers i.e. ZeroR, PART, J48, Random Tree, REP Tree, Naive Bayes, simple Logistic, and SGD. Our proposed hybrid classifier improves the accuracy of classification. This paper is organized as follows. In section 2, related work is presented. the statement is given. Section 3, discusses some recent review of similar work in the data mining field. Section 4 discusses experiments and results with which the diabetes dataset is mined. Section 5 concludes the paper.

II. LITERATURE SURVEY

Jothi et al. [1] presented the role of data mining and healthcare to design a reliable early detection system. In regard, they surveyed various method, algorithms and results of peer-reviewed papers. and methods. They also discussed their results and evaluation methods.

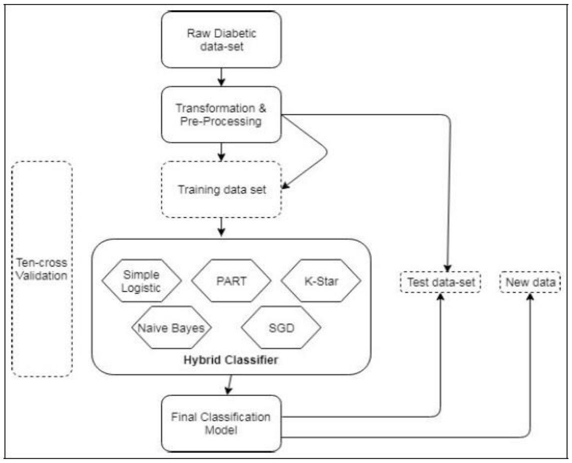
Zhang et al. [2] focused on the marketed drugs and clinical candidates for new indications in diabetes treatment by mining clinical „omics‟ data. They analyzed data from genome-wide association studies (GWAS), proteomics and metabolomics s studies and revealed a total of 992 proteins as potential anti-diabetic targets in human. The findings indicated that „omics‟ data mining based drug repositioning is a potentially powerful tool for discovering novel anti-diabetic indications from marketed drugs and clinical candidates.

Iyer et al. [3] provided solutions for diagnosing the disease by analyzing the patterns found in the data. They used Decision Tree and Naive Bayes algorithms for employing classification. Finally, their research proposed a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.

Gonccalves et al. [4] studied and compared peer-reviewed papers that developed systems for making prediction and decision support in the clinical area. Dewan and Sharma [5] proposed an efficient hybrid genetic algorithm with the backpropagation technique approach for predicting heart diseases prediction. In order to achieve a correct and cost-effective treatment, computer-based and support systems were developed for making a good decision. Finally, they developed a prototype to determine and extract unknown knowledge (patterns and relations) related to heart disease by using a heart disease database record. It was capable to process queries for detecting heart disease and thus assist medical practitioners in making smart clinical decisions. Gandhi el al. [6] provided the details about various techniques of knowledge abstraction by using data mining methods for predicting heart diseases. they analyzed various data mining methods such as Naive Bayes, Neural network, Decision tree algorithm on the medical-data sets. Rojas el al. [7] conducted a literature review to explore the role of data mining in health-care. They covered and analyzed 74 papers with their associated case studies. Finally, they identified, the most emerging topics that (i) provided a useful overview of the present work scenario in this field; (ii) helped researchers in mining algorithms, techniques, tools, methodologies and approaches for their own applications; and (iii) highlighted the use of process mining in order to improve the health-care system. Li et al. [8] designed a constrained-based accurate and robust prediction models for the health-care. The performance of the proposed model was demonstrated by using the record type-2 diabetes accumulated from multiple sources from all fifty states in the U.S. Zhang et al. [9] used cloud and big data analytics technologies for designing a cyber- physical system for patient-centric health-care applications and services. It has a unified data collection layer, a data management layer for distributed storage and parallel computing, and a data-oriented service layer. Their results showed the model enhanced the performance of the health care system by providing smart health-care applications and services. Komi et al. [10] compared various data mining techniques for predicting diabetes. Their experimental result showed that the Artificial Neural Network provided the highest accuracy than other techniques. Kavakiotis et al. [11] presented a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes. Their review presented a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management for the researchers. with the first category appearing to be the most popular. Their findings showed the usefulness of extracting valuable knowledge for setting new hypotheses, targeting, deeper understanding and further investigation in the field of data mining in the health-care sector.

III. PROPOSAL APPROACH

This section discusses the proposed algorithm. Traditional classification techniques use a single approach for classification. However, using a single method for classification may generate local optima. In this paper, we proposed hybrid approaches. While anomaly learning approaches can achieve good accuracy but the false positive rate is still high. This work is a combination of various best classification techniques that aim at maintaining the high accuracy while reducing the false positive rate as much as possible.



IV. EXPERIMENTAL SETUP

The experimental setup for the proposed framework is simulated on the java based toolkit i.e. WEKA 3.8.2. The data-set related to diabetes is collected from the UCI library which has eight relevant attributes along with one class attribute. It has 768 instances and two classes with 500 tested-negative and 268 tested positives. In order to balance the data-set Synthetic Minority Oversampling Technique is used. An unsupervised filter called "ReplaceMissingValues" is used that replaces all missing values for nominal and numeric attributes in the diabetic data-set with the modes and means from the training data. Finally, various classification techniques are applied to the diabetic data- set are their results are recorded. Subsequently, our hybrid classifier is designed that showed the highest accuracy among all the classifiers. The hybrid classifier of the proposed framework is trained by using ten-fold cross-validation. The evaluation criteria for the proposed framework is based on two-class classification by computing accuracy, precision for each classified class to examine the performance individually. The performance of classifications can be summarized with the use of a confusion matrix. It provides an idea to choose a better classification model not only on the basis of classification accuracy but also on the types of error made by the model.

V. PERFORMANCE COMPARISONS OF ALGORITHMS

The calculate the performance of any classifier the best metric is used almost all researchers are Classification accuracy. It is the ratio of truly classified samples and a total number of samples. As shown in the equation below:

Classification Accuracy = Truly classified samples/total number of samples

Both Sensitivity and Specificity are the metrics that can also be used to calculate the performance of classification algorithms. Sensitivity is the ratio of true positive (TP) samples and the total number of samples in a single row of confusion matrix as shown in below equation.

Sensitivity = True positive/(True Positive + False Negative)

Specificity is a ration of the true negative samples to the sum of both true and false negative samples. Specificity = True Negative / (True Negative + False Positive)

VII. RESULT AND DISCUSSION

The results obtained with the hybrid classifier of the proposed framework are the best among all classification algorithms, hence improves the accuracy and precision. The comparison among all classification technique is shown in Table.

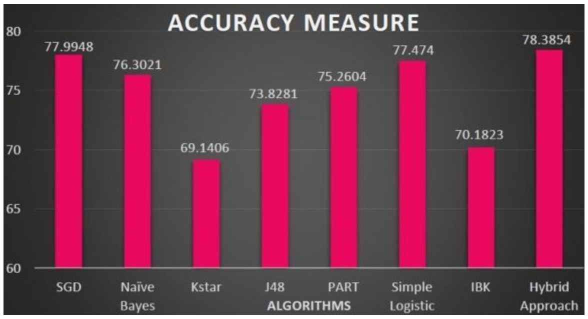
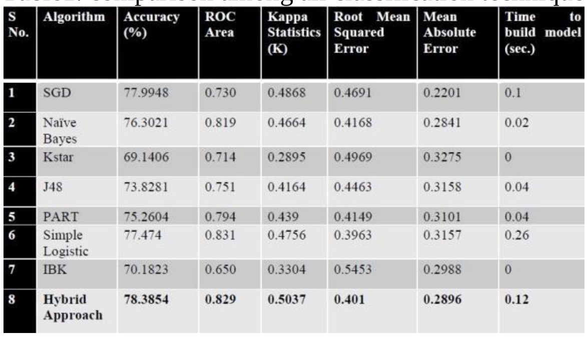


Fig.1: Accuracy Measure of Classifiers

Table1. comparison among all classification technique



S No.	Algorithm	Accuracy (%)	ROC Area	Kappa Statistics (K)	Root Mean Squared Error	Mean Absolute Error	Time to build model (sec.)
1	SGD	77.9948	0.730	0.4868	0.4691	0.2201	0.1
2	Naïve Bayes	76.3021	0.819	0.4664	0.4168	0.2841	0.02
3	Kstar	69.1406	0.714	0.2895	0.4969	0.3275	0
4	J48	73.8281	0.751	0.4164	0.4463	0.3158	0.04
5	PART	75.2604	0.794	0.439	0.4149	0.3101	0.04
6	Simple Logistic	77.474	0.831	0.4756	0.3963	0.3157	0.26
7	IBK	70.1823	0.650	0.3304	0.5453	0.2988	0
8	**Hybrid Approach**	**78.3854**	**0.829**	**0.5037**	**0.401**	**0.2896**	**0.12**

VI. CONCLUSION

Data mining provides several benefits in the prediction of diabetes at an early stage. In the realization of classification, building an accurate prediction model is one of the key challenges. In this paper, an efficient hybrid classifier is used effectively for improving the health prediction of diabetes. The data- set for diabetes is chosen from the UCI repository. The KDD process is implemented to gain the knowledge from the raw data. In our experimental section, a Java-based toolkit calledWeka is used for simulating the KDD process on the data-set. Various classifiers those are most common and used by researchers are used to compare the accuracy of our proposed hybrid classifier. Our hybrid classifier achieves 78.39% accuracy. The results show the effectiveness of the proposed framework. In future, the real-time data regarding diabetes can be obtained using IoTs and processed by using Big Data analytics techniques in the cloud computing environment.

VII. REFERENCES

[1]. N. Jothi, W. Husain et al., Procedia Comput. Sci. **72**, 306 (2015).

[2]. M. Zhang, H. Luo, Z. Xi, and E. Rogaeva, PloS one **10**, e0126082 (2015).

[3]. A. Iyer, S. Jayalalitha, and R. Sumbaly, arXiv preprint arXiv:1502.03774(2015).

[4]. J. Gonçalves, B. M. Faria, L. P. Reis, V. Carvalho, and Á. Rocha, "Data mining and electronic devices applied to quality of life related to health data," in "Information Systems and Technologies (CISTI), 2015 10thIberian Conference on," (IEEE, 2015), pp. 1–4.

[5]. A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," in "Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on," (IEEE, 2015), pp. 704–706.

[6]. M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," in "Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on," (IEEE, 2015), pp. 520–525.

[7]. E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, J. biomedical informatics **61**, 224 (2016).

[8]. Y. Li, C. Bai, and C. K. Reddy, Inf. sciences **330**, 245 (2016).

[9]. Y. Zhang, M. Qiu, C.-W.Tsai, M. M. Hassan, and A. Alamri, IEEE Syst. J. **11**, 88 (2017).

[10]. M. Komi, J. Li, Y. Zhai, and X. Zhang, "Application of data mining methods in diabetes prediction," in "Image, Vision and Computing (ICIVC), 2017 2nd International Conference on," (IEEE, 2017), pp. 1006– 1010.

[11]. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, Comput. structural biotechnology journal 15, 104 (2017).