International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Big Data and Reality Mining in Healthcare Smart Prediction of Clinical Disease Using Decision Tree Classifier

Iman Mmohammad Alqahtani[1], Ebtesam Shadadi[2], Latifah Alamer[3]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: The Healthcare system is the most essential and intersecting research field. Implementing effective technology in the healthcare system is a boon for the human community. Recently, the need for medical advancement has turned huge attention to healthcare practices. Healthcare practices mainly require healthcare data that comprises patient data, treatment data, and resource management data-daily, the amount of healthcare data increases, making the accuracy and classification more complex. Data mining is the most superior technology for handling those healthcare data effectively. This paper proposes an artificial intelligence with a J48 classification algorithm. This proposed mechanism works intelligently in discovering the hidden patterns of the data and enhances classification accuracy. It is applicable for handling various disease datasets, which include heart diseases, diabetes data, etc. The result from the proposed mechanism improves the accuracy of disease prediction, like whether the disease impacts the patient or not. The comparison proves the proposed system's accuracy efficiency is carried out with random forest, naive Bayes, and k-means. The performance factors for comparison are correctly classified instances, accuracy, sensitivity, and specificity. The simulation outcome shows that the proposed J48 is more efficient in achieving the diagnosis accuracy than the others.

Keywords: *Healthcare data, Classifications, Data mining, Data Accuracy, Artificial Intelligence, J48*

1. Introduction

Disease diagnosis is the combination of medical and technical terms directly related to human life. For an expert physician, diagnosing the patient's disease with accurate pathological data is simpler. In a country, every citizen's health is vital because it reflects the growth of the country. Healthcare diagnosis in a human is a complex process that includes diagnosis, disease prevention, injury, treatment, and other physical and mental impairments. The healthcare industry receives massive data such as administrative reports, electronic medical records, and other benchmarking findings [1,2]. The Healthcare domain has become the top research-intensive with vast public funds. The evolution of computers and their algorithms maximize the utilization of computer tools; it also has massive future inventions. The combination of healthcare and computing is known as health informatics. This evolution mode enhances the health care system's performance in terms of time, quality, and cost [3].

Data mining is the best technology for handling massive amounts of healthcare data. These data are of various formats and come from different resources, making mining significantly complicated. The physician uses these patient-related data for useful analysis of the treatments. Past treatments, comparing the causes and adverse effects, give a standard disease guideline. As a result, the physician can deliver effective healthcare services at an affordable cost and time. Most hospitals and healthcare organizations use data mining for making patient-related decisions. Data mining is advantageous in observing the data's hidden patterns and determining patient preferences and current and future needs [4].

Medical data are precious; this makes dealing with data mining data as most challenging and leading research. The advancement of data mining in recent days has multiple milestones, including several healthcare applications. Its significant contribution is pattern recognition, disease diagnosing intrusion detection, and procurement methods [5,6]. Most data mining systems comprise two learning tasks such as supervised and unsupervised learning. The supervised learning system is developed for situation-based prediction on the class labels. The unsupervised learning model does not contain any labels which distribute the data to get more information about that data. The most popular technique in unsupervised data mining is clustering groups with unlabelled data. Clustering is mainly used for the grouping of data based on the similarities among themselves. Artificial Intelligence (AI) is an efficient mechanism that enhances the system with more sensitivity and accuracy. AI is used as a subfield in machine learning to predict medical data [7] better. This paper uses AI with J48 to achieve better prediction than the existing mechanism.

This work is organized as follows: Section 2 represents the related works, Section 3 defines the proposed methodology, and Section 4 defines the result and discussions. Finally, Section 5 defines the conclusion.

[1]*King Khalid University, Abha, Kingdom of Saudi Arabia,*
Email: ealqahtani@kku.edu.sa
[2]*College of Computer Science and Information Technology,*
Jazan University, Jazan, Kingdom of Saudi Arabia,
Email: ashedadi@jazanu
[3]*College of Computer Science and Information Technology,*
Jazan University, Jazan, Kingdom of Saudi Arabia,
[3]*Email:laalamer@jazanu.edu.sa*

2. Related Work

V. Jackins et al. [8] present Artificial Intelligence with random forest and Naive Bayes classification algorithm. This system is developed for predicting the life-threading diseases of humans, like heart disease datasets, diabetes, and cancer datasets. This system's main advantage is minimizing the dataset's unwanted information and estimating the diseases. This approach gives effective results compared to a random forest and Gaussian Naïve Bayes. Godwin Ogbuabor et al. [9] contribute his clustering work using Silhouette score values. There are multiple clustering algorithms available but no clarity on which clustering algorithm will fit the medical dataset. To reduce the gap, the author proposed a clustering algorithm using Silhouette score values, which is suitable for any medical dataset.

Sampaul TGA et al. [10] & Vimal S et al. [11] discussed different data mining machine learning algorithms used to identify and predict the disease diagnosis. Most existing works are based on the Bayesian network, decision tree, and regression. Vimal et al. discussed the importance of data mining in disease analysis because multiple tests based on different scenarios need to be done to predict disease. The data mining minimizes the work with those existing data and comparison results. As a result, it yields a massive improvement in time saving and performance than the traditional method.

Rasha et al. [12] proposed an advanced IoT-Assisted Healthcare Monitoring System. This work introduces a rapid adoption of cloud computing to improve data processing performance and accessibility in the cloud environment. Initially, Improved Pigeon Optimization (IPO) algorithm is implemented to enhance the prediction rate. Next, a Backtracking Search-Based Deep Neural Network (BS-DNN) is employed to classify healthcare datasets.

Haiou Tang et al. [13] discussed the Bayesian inference-based feature selection algorithms and random forest-based feature selection algorithms for enhancing abnormal data detection in health data. Next, the author used the local importance degree of degree calculation to overcome the drawbacks of the original algorithm.

I de & M. B. Filho et al. [14] introduced an IoT-assisted healthcare platform for ICU patients during COVID-19. The healthcare platform is specially designed for monitoring the patient during emergency scenarios. To monitor COVID-19 patients, wearable and Unobtrusive sensors are added to the system. The results of this research pave the way for the incorporation of machine learning algorithms for identifying risk factors and taking prompt appropriate action to increase treatment efficacy.

B. Ç. Uslu et al. [15] discussed the effects of IoT-supported smart healthcare systems. In this work, the author discusses various existing mechanisms' architecture and their challenges and optimization factors. As a result, the author stated various drawbacks and aspects need to be removed from the smart healthcare system.

L. Greco et al. [16] developed a smart healthcare platform using the combination of cloud and IoT devices. The main motto of this work is to propose an advanced smart healthcare system using artificial intelligence (AI) through wearable sensors. The author discussed various health monitoring techniques in healthcare systems with IoT techniques. Finally, the author concludes that AI and machine learning is significant for proposing an enhanced cloud-based healthcare system.

Liu et al. [17] examined the medical record contents and their characteristics to propose the efficiency of the preprocessing technique by enhancing the characteristics. Next, the author applied this technique to a coronary heart disease dataset, and the data analysis outcome was noticeably enhanced. The original physical examination dataset cannot be used directly for data analysis and information mining of diseases due to several factors like data missing, redundant data, and aberrant data. Several preprocessing techniques are proposed for various reasons to better utilize the valuable information found in physical examination data.

C. Qiu et al. [18] employed SVM to achieve maximum classification ability and accuracy. It was able to perform clinical diagnostics for sarcoidosis and tuberculosis with success. Li et al. [19] applied artificial neural networks (ANN) with magnetic resonance imaging (MRI) for auxiliary detection of DMD in children by relieving the pain that occurred during the conventional diagnosis and detection methods.

3. Proposed Methodology

In this paper, we applied J48 with Artificial Intelligence to achieve the highest classification accuracy than the existing system. The proposed methodology initially comprises several processes, beginning with data preprocessing. The proposed architecture is described in figure 1. The detailed descriptions of the proposed system and its working mechanism is described in the below sections.

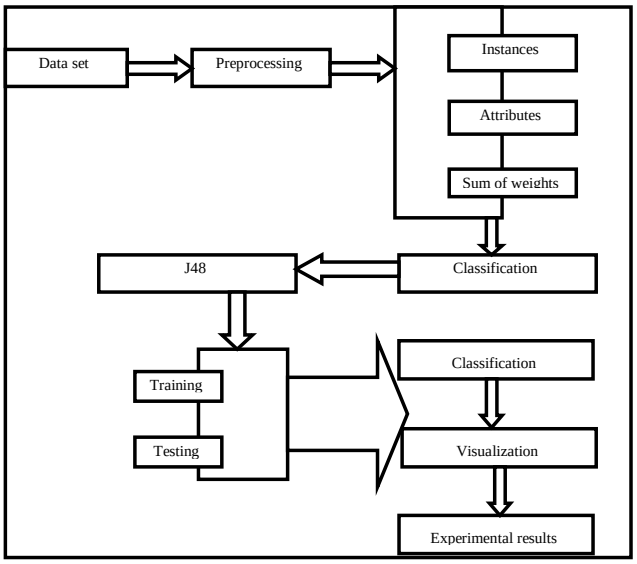


Figure 1. Proposed Architecture

3.1.1 Data preprocessing

Data Processing is the most challenging and problematic task in machine learning, especially computational biology [2]. This complexity is because of the medical dataset's noisy, duplicate, unnecessary, and irregular information. The two essential terms in data preprocessing are data creation and filtering stages. The main advantage of preprocessing is; that it reduces the total executing time. Data preprocessing consists of feature extraction, standardization, feature selection, filtering, transformation, Instance determination, and so forth. The data preprocessing resulting dataset is the last training set.

3.1.2 Data mining Tools

Data mining consists of various data mining tools, with which WEKA is the industry's most popular data mining suite. This popularity is why WEKA results in higher accuracy, is open-

source and is free of cost. WEKA allows the developer to change the algorithm's source according to the needs. WEKA has additional features such as ease of implementation, flexibility, and ease of understanding, requiring minimum coding knowledge. It allows re-implementations of traditional data mining algorithms. Among these, the most consistent algorithm is comprised of C4.5 and also known as J48. Compared to SAS Enterprise Miner, WEKA is effective because Enterprise Miner contains only a graphical user interface (GUI), making the robotized tests more complex. For the researchers, it is challenging to compute multiple variations on analysis. But WEKA consists of various operation modes, and creating experiments in WEKA is easy.

J48 is the extraction of Quinlan's C4.5 algorithms, which creates the C4.5 decision tree. Initially, the dataset is split into multiple subsets, which are the base for the decision. J48 results in standardized data, and the data split are based on the attributes. The standardized data gained using the attributes are summarized and utilized for further process. J48 returns the minor subsets and stops the split when the similar class in all instances. The expected class estimation is taken by the J48 and develops the decision nodes. J48 decision tree works with the missing data attributes, specific characteristics, and wavering attribute costs. Here, pruning is applied to expand the accuracy (Venkatesan, 2015).

3.2 Limitations of J48 Algorithm

J48 is one of the most used algorithms in the industry, but it also has some shortcomings, and a few of them are mentioned below;

3.2.1. Empty Branches

In the J48 algorithm, the essential step is constructing a tree with a significant value. But in this work, we have attained multiple nodes with zero values near that value. These values will not be helpful for creating a class for the classification task.

3.2.2 Insignificant Branches

The total distinct attributes taken to build a decision tree will result in the same quantity of potential division. But not all the divisions are valid for the

3.2.3 Over Fitting

The algorithm results in the information with extraordinary attributes here evolves over Fitting. The invalid nodes with the least examples are referred to as fragmentations, and these fragmentations cause process distribution. Generally, the J48 algorithm tree growth is deep enough to classify the training examples correctly.

The proposed system overcomes the fittings and frees the data from noisy information. But in some cases, the training examples are over-fits with the noisy data. In decision tree learning, the overfits can be bypassed through two they are:

- In the training data, the maximum point of accurate classification is noted if the tree growth reaches that point to stop the tree growth.
- The only solution for the over-fit problem is post-pruning the tree once the training data get over-fitted.

In this research, two to reduce the input space of data, two tools are used: the Entropy of Information Theory and Correlation Coefficient. Here the experimental work is carried out with dengue medical data. The Java-based machine learning tool WEKA describes a detailed explanation of the datasets.

4. Experimental results

In this work, each algorithm's accuracy is determined based on the testing data set passed on each training model of the algorithms. The average accuracy is measured based on the three sets of training data. The accuracy measures are described below;

4.1 Dataset

The dataset for the experiments consists of various disease data, including coronary heart disease, diabetes, etc. The wearable devices and prediction data are used to collect the datasets.

4.2 Comparative parameters

Accuracy

The accuracy is measured by the ratio of the number of correct assessments to the total assessments. The process begins with image extraction and then is compared with the complete dataset using the below-given expression. Accuracy percentage (%) is determined by two main factors: data quality and errors.

$$Accuracy = \frac{(TN+TP)}{(TN+TP+FN+FP)} \quad (1)$$

Where TN-True Negative, TP-True Positive, FP-False positive, and FN-False Negative.

Sensitivity

The dataset's true positives and negatives are then added to calculate the sensitivity. The count of true positive to the added estimation of true positive and false negative ratio gives the sensitivity. From the obtained results, an amount of positive measures is declared. The sensitivity is estimated in percentage (%), which based on the below-given expressions;

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (2)$$

Specificity

Modifications and their impact on prediction from the original dataset determine the specificity. In other words, specificity is the progression of the proposed work. It is calculated in percentage (%) based on the correctly perceived negative measures. It can be shown as the number of negative assessments to the summation of true negative and false positive assessments. The measuring of specificity is expressed as below;

$$Specificity = \frac{TN}{(TN+FP)} \quad (3)$$

Table 1. Classified parameters

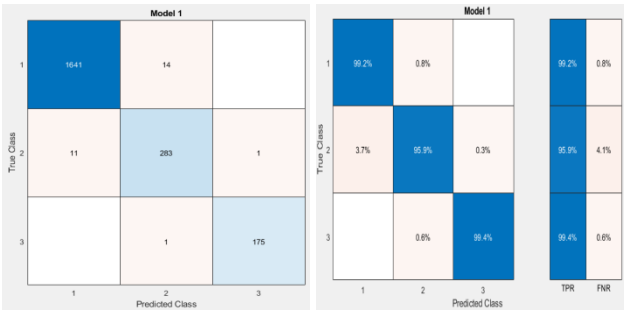
Correctly Classified Instances	2113 99.3885
Incorrectly Classified Instances	13 0.6115 %
Kappa statistic	0.9833
Mean absolute error	0.0078
Root mean squared error	0.0624
Relative absolute error	3.1745 %
Root relative squared error	17.8273 %
Total Number of Instances	2126

Table 2. Detailed accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.988	0.015	0.996	0.988	0.976	0.997	0.985	0.997	1
0.973	0.003	0.983	0.973	0.978	0.974	0.992	0.980	2
0.994	0.001	0.994	0.994	0.994	0.994	1.000	0.995	3
0.994	0.012	0.994	0.994	0.994	0.984	0.995	0.995	

4.3 Confusion Matrix

A confusion matrix is nothing but a quick progression of the predictions on the classification problem. There are two types of forecasts such as correct predictions and the second one is incorrect predictions. In each class, count values and broken down are used to calculate the correct and incorrect predictions. The best thing, along with describing the error, it also states what kind of error occurred.



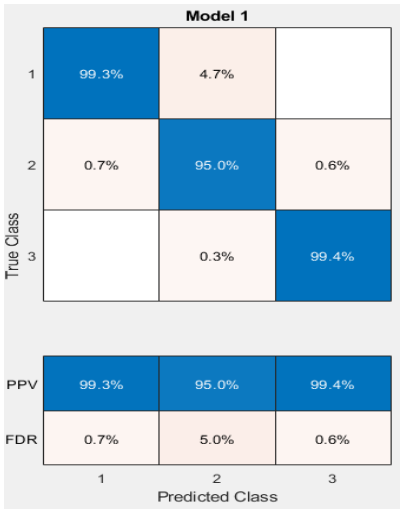
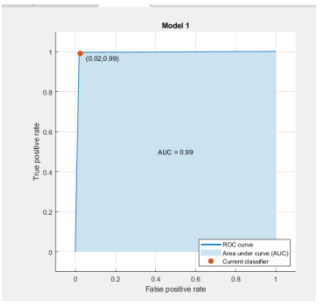


Figure 2. Confusion matrix

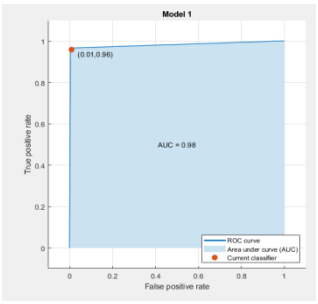
4.4 ROC Analysis

The Receiver Operating Characteristic (ROC) analysis [5][9] [13] provides the best way of measuring the accuracy level of the classifier performance wholly and independently. ROC analysis is defined with two classes: the true-positive rate (TPR) on the y-axis and the false-positive rate (FPR) on the x-axis. This representation is simple to define and accountably reasonable. For each classifier, these two classes are plotted with the point. A "Curve" is gained, which shows the extremely multiple derived classifiers are obtained and segmented. The connection between the two classes is by representing different weights. The lower TPR and/or higher FPR is because of the high cost for any class distribution and cost matrix. The classifiers below the convex hull are discarded by indicating the classifiers at points (0,0) and (1,1), respectively. This indication determines the prediction of negative and positive with the default classifiers.

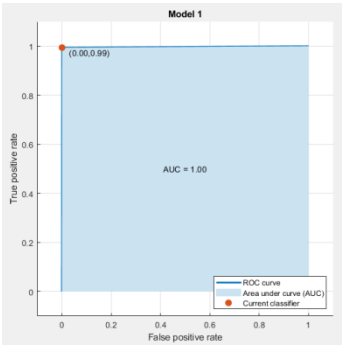
The works [5][9] explain ROC analysis, and here we are not representing the ROC space. But the maximization of correct predictions and the false class show the preciseness of incorrect predictions. This choice is not utilized for more than two classes. In this case, the false-negative rate (FNR) and the FPR are proposed by the (0,1) and (1,0), respectively. The (0,1) defines the classifier that classifies anything as negative, and (1,0) defines the classifier that classifies anything as positive. The points (0,1), (1,0), and (1,1) execute the curve with a new curve known as Area Under the ROC Curve (AUC). Completing Area Above the ROC Curve (AAC) is a better way to minimize depreciation. The AAC, along with AUC, gives the technical statement of ROC.



(a) AUC-0.99



(b) AUC-0.98



(C) AUC – 1.00

Figure 3. Confusion matrix

4.5 Parallel Coordinates Plot

Parallel Coordinates Plots are mainly used for the comparison of multiple variables which are associated with each other. This kind of plotting effectively plots the numerical data and multivariate such by comparing the products containing the same attributes. Parallel Coordinates Plot contains its axis, which has various scales and is positioned parallel. Each variable in the axis contains multiple measurement units, and the axes scales are normalized to have uniform scales. The line series's values across the axes are plotted, and all the plotted points are connected. The connection order determines the reader's understandability of the data. This connection among the adjacent variables is simpler to recognize than the non-adjacent variables. These axes are reordered to observe the correlations or determine patterns across the variables. In the Coordinates Plots, parallel downside becomes over-cluttered, which means they are not useful because of very data-dense. "Brushing" is an interactivity technique for overcoming data sense issues. Brushing determines the collection of lines or selected lines that are fading out all the others. This mechanism separates or filters the noises in the data.

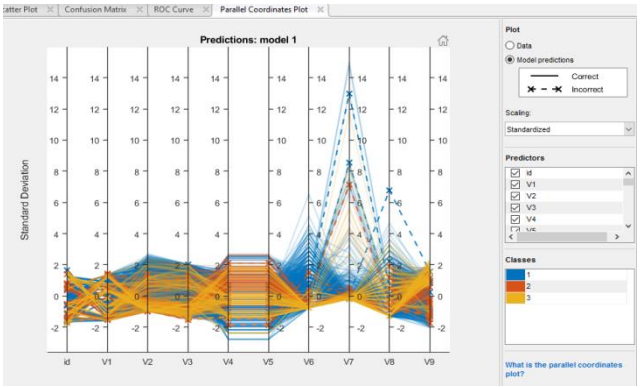


Figure 4. Parallel Coordinates Plot

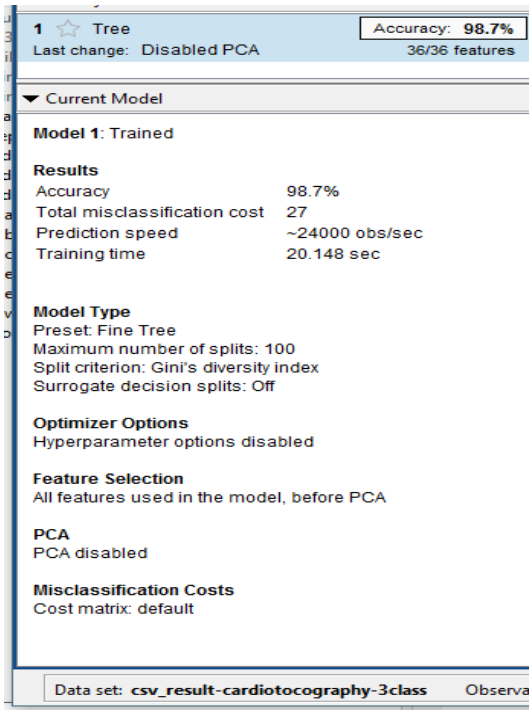


Figure 5. Accuracy simulation

Table 3. Result obtained by proposed J48

Correctly Classified Instances	98.7%
Accuracy	98.88%
Sensitivity	98%
Specificity	98.4%

4.6 Attributes

The dataset taken for this observation contains 2126 fetal cardiotocographs (CTGs). These CTGs are automatically processed and are categorized by three expert obstetricians respective to the diagnostic features. The classification comprises morphologic patterns (A, B, C. …) and fetal state (N, S, P). This dataset can be applied for either 3-class or 10-class experiments.

Table 4. Classified attributes

LB - FHR baseline (beats per minute)	
AC -	# of accelerations per second
FM	#of fetal movements per second
UC	#of uterine contractions per second
DL	# of light decelerations per second
DS -	# of severe decelerations per second
DP	- # of prolonged decelerations per second
ASTV	percentage of time with abnormal short-term variability
MSTV	mean value of short-term variability
ALTV	percentage of time with abnormal long-term variability
MLTV	mean value of long-term variability

Width	width of FHR histogram
Min	minimum of FHR histogram
Max	Maximum of FHR histogram
Nmax	- # of histogram peaks
Nzeros	- # of histogram zeros
Mode	- histogram mode
Mean -	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency
CLASS	- FHR pattern class code (1 to 10)
NSP -	fetal state class code (N=normal; S=suspect; P=pathologic)

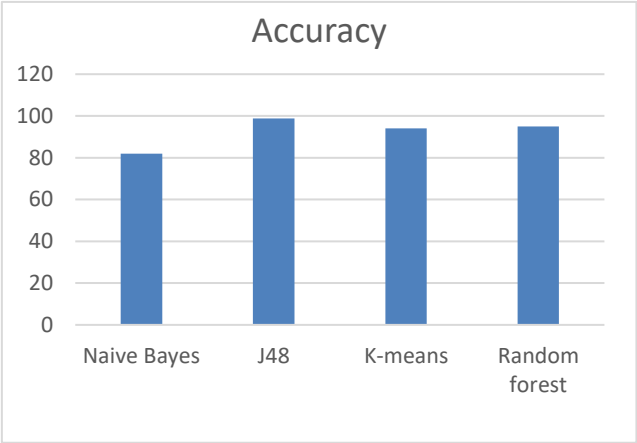


Figure 6. Accuracy Comparisons

Table 3 represents the result obtained by the proposed J48, in which the accuracy is 98.88%. This obtained result is compared with the existing Navies Bayes, K-means, and Random Forest.

The above figure 10 illustrates the accuracy level obtained by each algorithm. The X-axis determines the algorithms, and the Y-axis determines the accuracy percentage obtained. The proposed J48 achieves 98.88% accuracy, whereas Naive Bayes with 82.07%, K-means with 91.84%, and Random forest with 92.01%. The above results prove that the proposed J48 is far better than the others in accuracy.

Conclusion:

In this paper, discussed the implementation of J48 with the WEKA tool to observe accuracy. The main motto of this study is to enhance the accuracy level in medical diagnosis. To achieve this, the 2126 fetal cardiotocograms dataset is taken for observation in this work. In this proposed method preprocessing is initially progressed followed by determination, feature extraction, transformation, standardization and so forth are accomplished. The implementation of J48 results is obtained and analyzed with the comparative parameters, including Accuracy, Sensitivity, and Specificity. This observation is further continued with Receiver Operating Characteristic (ROC) analysis and Parallel Coordinates Plots analysis. Finally, a comparison work is executed to examine the proposed system's overall performance. The proposed system is compared with Navies Bayes, K-means, and Random Forest on the aspect of accuracy. The proposed J48 achieves greater accuracy of 98.88 %, which is more efficient than the others.

Reference:

[1] Vikas Mongia , Dr. Deepak Mehta,"Feature Selection and Diagnose of Healthcare Issues using Classification Algorithms", International journal of Advance Research in Science and Engineering, IJARSE,Vloume 6 Nov 2017, ISSN: 2319-8354.

[2] H.S. Hota, Seema Dewangan,"Classification of Health Care Data Using Machine Learning Technique", International Journal of Engineering Science Invention ISSN (Online): 2319 – 6734, Volume 5 Issue 9 September 2016.

[3] Shortliffe, EH., Perrault, LE., (Eds.). Medical informatics: Computer applications in health care and biomedicine (2nd Edition). New York: Springer, 2000.

[4] Anand Sharma, Vibhakar Mansotra, ―Emerging Applications of Data Mining for Healthcare Management - A Critical Review ǁ, 2014 International Conference on Computing for Sustainable Global Development (INDIACom), DOI: 10.1109/IndiaCom.2014.6828163.

[5] Priyadarshini R, Dash N, Mishra R (2014) A novel approach to predict diabetes mellitus using modifed extreme learning machine. In: International Conference on Electronics and Communication Systems (ICECS), 2014, pp 1–5 3.

[6] Sankaranarayanan S, Perumal TP (2014) Diabetic prognosis through data mining methods and techniques. In: International Conference on Intelligent Computing Applications, 2014, pp 162–166.

[7] Dahiwade D, Patle G, Meshram E (2019) Designing disease prediction model using machine learning approach. In: Third IEEE International Conference on Computing Methodologies and Communication (ICCMC), 2019.

[8] V. Jackins, S. Vimal, M. Kaliappan & Mi Young Lee, "AI-based smart prediction of clinical disease using random forest classifer and Naive Bayes", Springler October 2020, The Journal of Supercomputing https://doi.org/10.1007/s11227-020-03481-x

[9] Godwin Ogbuabor and Ugwoke, F, "CLUSTERING ALGORITHM FOR A HEALTHCARE DATASET USING SILHOUETTE SCORE VALUE", International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 2, April 2018

[10] Sampaul TGA, Robinson YH, Julie EG, Shanmuganathan V, Nam Y, Rho S (2020) Diabetic retinopathy diagnostics from retinal images based on deep convolutional networks. Preprints. https://doi.org/10.20944/preprints202005.0493.v1

[11] Vimal S et al (2020) Deep learning-based decision-making with WoT for smart city development. In: Jain A, Crespo R, Khari M (eds) Smart innovation of web of things, CRC Press, Boca Raton, pp 51–62. https://doi.org/10.1201/9780429298462

[12] Rasha M Abd El-Aziz, Rayan Alanazi , Osama R Shahin , Ahmed Elhadad , Amr Abozeid , Ahmed I Taloba and Riyad Alshalabi, "An Effective Data Science Technique for IoT-Assisted Healthcare Monitoring System with a Rapid Adoption of Cloud Computing", Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 7425846, 9 pages https://doi.org/10.1155/2022/7425846

[13] Haiou Tang, "Intelligent Processing and Classification of Multisource Health Big Data from the Perspective of Physical and Medical Integration", Hindawi Scientific Programming Volume 2022, Article ID 5799354, 11 pages https://doi.org/10.1155/2022/5799354

[14] I. de M. B. Filho, G. Aquino, R. S. Malaquias, G. Girao, and S. R. M. Melo, "An IoT-based healthcare platform for patients in ICU beds during the COVID-19 outbreak," IEEE Access , vol. 9, pp. 27262–27277, 2021.

[15] B. Ç. Uslu, E. Okay, and E. Dursun, "Analysis of factors affecting IoT-based smart hospital design," Journal of Cloud Computing, vol. 9, no. 1, p. 67, 2020.

[16] L. Greco, G. Percannella, P. Ritrovato, F. Tortorella, and M. Vento, "Trends in IoT based solutions for health care: moving AI to the edge," Pattern Recognition Letters, vol. 135, pp. 346–353, 2020.

[17] S. Liu, L. Zhang, Y. Long, Y. Long, and M. Xu, "A new urban vitality analysis and evaluation framework based on human activity modeling using multi-source big data," ISPRS International Journal of Geo-Information, vol. 9, no. 11, p. 617, 2020

[18] C. Qiu, M. Schmitt, L. Mou, P. Ghamisi, and X. Zhu, "Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets," Remote Sensing, vol. 10, no. 10, Article ID 1572, 2018.

[19] Y. Li, G. Wen, Y. Hu et al., "Multi-source Seq2seq guided by knowledge for Chinese healthcare consultation," Journal of Biomedical Informatics, vol. 117, Article ID 103727, 2021.