ISSN 2581 – 3781

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2022;7(2):71-76

RESEARCH ARTICLE

Impact of Classification Algorithms on Cardiotocography Dataset for Fetal State Prediction

D. Sudharson[1], K. Vignesh[2], Aman Kumar Dubey[2], K. Mukilan[3], M. K. Nanda Kizore[3], S. Abishek Ram[3]

[1]*Department of AI and DS, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India,* [2]*Department of MBA, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India,* [3]*Department of AI and DS, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India*

Received on: 12/02/2022; Revised on: 29/03/2022; Accepted on: 30/04/2022

ABSTRACT

Monitoring of fetal heart rate and fetal health is done by cardiotocography (CTG). Obstetricians can observe CTG records and make life-saving decisions. The ability to go throh all the data points is fairly challenging. One possible solution is to use clinical decision making systems. The selection of these systems is made possible by choosing the best classifier, in this paper we compare four simple classifiers (K Nearest Neighbors, Decision Tree, Support Vector Machine, Naive Bayes). To improve accuracy, the dataset is split based on “Outlier Removal” and “Feature Selection”.

Key words: Cardiotocography, Classification, Decision tree, Feature selection, Outlier removal

INTRODUCTION

About 295 000 women die globally due to “Maternal mortality”. Sub-Saharan Africa and Southern Asia accounted for rohly 86% of the estimated global maternal deaths in 2017. This shows the inequality in health facilities across the world. The deaths could be prevented by observing the cardiotocography (CTG) records and taking timely actions during pregnancy.

CTG records fetal heart rate (FHR) and uterine contractions (UC) during pregnancy using an ultrasound transducer which is placed on the mother’s abdomen, this method is done to check fetal well-being typically in the third trimester (27–40 weeks). Babies at risk of hypoxia (lack of oxygen) are mostly monitored using CTG to avoid death and long-term disablement due to lack of oxygen during the time of delivery. The data in CTG are interpreted to ensure precise prediction of fetal well-being and prepare the mother for delivery. Obstetricians manually observe FHR patterns during this process by CTG. Manually viewing CTG recordings is daunting and challenging. Optimal classifier-based clinical decision-making (CDM) systems offer a viable solution for finding patterns in CTG datasets. Four classifiers were used in this work: K Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), and Naive Bayes (NB). These classifiers were used to predict a target variable NSP that indicates fetal status (N = Normal, S = Suspicious, and P = Pathological). Optimal classifiers for CDM systems have been proposed.

Address for correspondence:
Dr. D. Sudharson
E-mail: sudharsondorai.ads@kct.ac.in

LITERATURE SURVEY

The electronic monitoring of FHR was introduced way back in 1987, but due to a lack of proper records, the development of clinical decision-making (CDM) algorithms was not developed. CDM algorithms and systems were gradually improved using data mining techniques thereby predicting the state of the fetus based on uterine constraints and accelerations of the fetus per second. Research shows that clustering medical data have improved predictability.[1] The CTG dataset which was used in the research was used along with Sisporto which does automatic analysis of CTG tracings, which was introduced by Bernades.[2,3] Comparative analysis done with four classifiers on the CTG dataset with 21 attributes and a reduced CTG showed that the latter produces the best results.[4,5] On the surface, ANN classifiers seem to provide good average performance but in individual runs, the performance varies wildly.[6,7]

Another method to improve the prediction accuracy of a model is to introduce data sample filtering (DSF). DSF was effective with a large number of samples but the computational time required to do DSF is not sustainable.[8] Balanced dataset radically improves the performance of classifiers[9,10] and the viability of algorithms such as ANN, SVM, Logistic Regression, and KNN shows that Random Forest gives the best possible result.[11] CART algorithms in general along with boosting and bagging provide a good enoh classifier.[12,13] Preprocessed data when fit on several high-performing machine learning models indicate low estimated test error.[7,14]

FRAMEWORK

Using reduced features from the CTG dataset which include seven features: "AC", "UC", "ASTV", "mSTV", "ALTV", "mLTV", and "Mean".[12] These features were used alongside the complete dataset to compare the results. Our research is to determine the best baseline machine learning algorithm to employ in CDM systems for classification based on CTG records.

The CTG dataset consists of 2126 CTG collected in the Maternity and Gynecological Clinic. It is secondary data downloaded from UCI Machine Learning Repository.[15] There are 22 main features extracted automatically and processed by SisPorto. The features are tabulated in Figure 1 along with the description of NSP in Figure 2.

Feature description

The Figure 1 shows the name and description of the features outlined in the CTG dataset. There

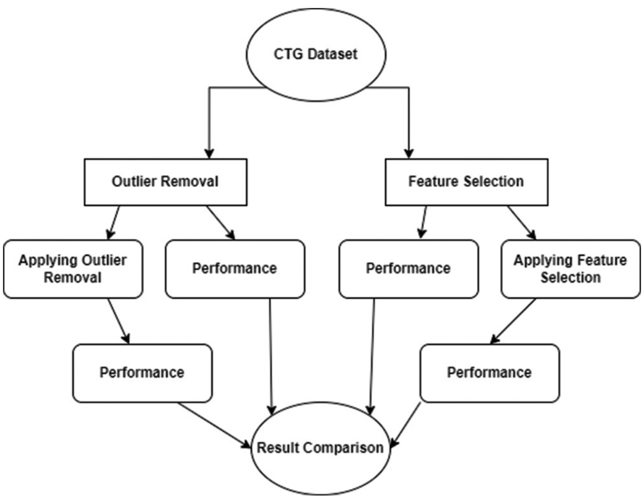

Figure 1: Classification methods splitting based on outliers and features

are 22 features and one target variable. These are discussed in brief in the upcoming sections.

Target description

The following research methodologies were used on the CTG Dataset [Figure 3].

Data cleaning

One of the most important steps in building a machine learning model is to input data. This step is important because the cleaner and more relevant the data, the easier it is for the model to predict the output. If the data is clean and well-defined, even a simple algorithm will give the best results.

Removing unwanted columns

This is one of the simple steps, data duplicates and irrelevant data are removed in the step.

1	LB	FHR baseline value(beats per minute)	12	DR	repetitive decelerations
2	AC	accelaerations(SisPorto)	13	Width	histogram width
3	FM	foetal movement(SisPorto)	14	Min	low freq. of histogram
4	UC	uterine contractions(Sisporto)	15	Max	high freq. of histogram
5	ASTV	percentage of time with abnormal short term variability (SisPorto)	16	Nmax	number of histogram peaks
6	mSTV	mean value of short term variability(SisPorto)	17	Nzeros	number of histogram zeros
7	ALTV	Percentage of time with abnornmal long term variability(SisPorto)	18	Mode	histogram mode
8	mLTV	mean value of long term variability(SisPorto)	19	Mean	histogram mean
9	DL	light decelerations	20	Median	histogram median
10	DS	severe decelerations	21	Variance	histogram variance
11	DP	prolongued decelerations	22	Tendency	histogram tendency:-1-left assymetric; 0-symmetric; 1 =right assymetric

Figure 2: Attribute information of cardiotocography dataset

NSP value	Description
1	Normal
2	Suspect
3	Pathologic

Figure 3: Target information of cardiotocography dataset

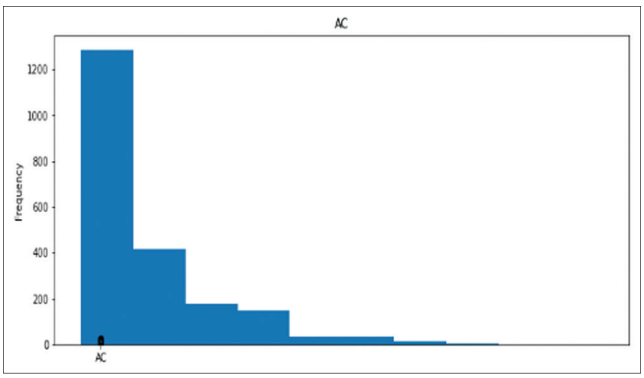

Figure 4: Univariate analysis on "AC"

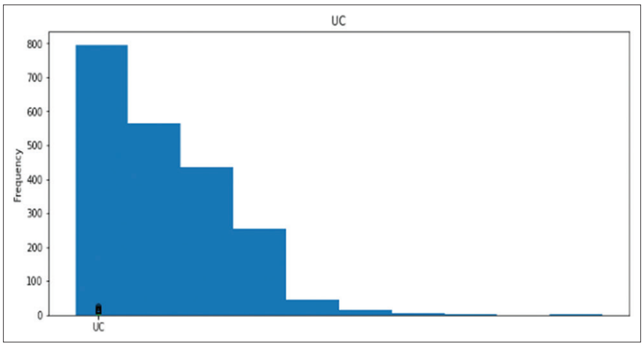

Figure 5: Univariate analysis on "UC"

Keeping this type of data affects our results greatly.

Checking for duplicates

An in-built function called duplicated.sum to check the total number of duplicate values. The function drop_duplicates() helps to remove all the duplicate values.

Checking for missing values

To find the missing values in a dataset, the function "notnull()" and "isnull()" can be used. However, this dataset does not have any missing values.

Checking for outliers

It is also an equally important step because few outliers can deviate the result so much. We deployed multiple functions to visualize and also remove these outliers.

Exploratory data analysis

Exploratory data analysis is used for understanding data more deeply before employing our model. We often use data visualization for this. This helps to identify the main characteristics of the dataset. Exploratory data analysis can be classified into two main types.

Univariate data analysis

In this method, only one variable is used for data analysis. This does not give us a relational understanding of two variables.

Bivariate data analysis

This is used to find the relational understanding between two variables of the dataset. Correlation plot is an often used graphical tool for this method.

Feature engineering and selection

Using the results from the above analysis feature selection and engineering are done. This includes both dropping unwanted features and also creating a new feature using the other variables for achieving more efficiency. The selection of relevant features is called feature selection.

Test train split

It is used to divide the dataset for training and testing of the model. The main purpose is to find the performance of the machine learning model on new data.

Model building

It is the prime step compared to all other steps. The best algorithm is chosen and deployed. The algorithm is chosen based on the dataset and the objective purpose of building it. Furthermore, multiple algorithms can be used and compared to find the best fit for the dataset.

a) KNN: KNN is a supervised machine learning algorithm that uses only the nearby data points for classification.
b) DT: It is one of the most popular classification algorithms. It uses a tree-like structure for classification.
c) SVM: The SVM draws a hyperplane between the two types of data. It can only perform binary classification.
d) NB: A simple Bayesian model that assumes independence between features.

Model validation and evaluation

The integral part of developing a process is called model evaluation it finds the best model which represents our data. The performance of a model is explained by evaluation metrics, the metrics provide important aspects like classification report, precision,f1-score, recall, and also the accuracy of the model.

Precision = (TruePositives_1 + TruePositives_2)/([TruePositives_1 + TruePositives_2] + [FalsePositives_1 + FalsePositives_2])

Recall = TruePositives/(TruePositives + FalseNegatives)

F1 Score = 2 * (Precision * Recall)/(Precision + Recall)

IMPLEMENTATION

The dataset was further split in the basis of:

1. Outlier removal (2003 data instances)
2. Feature selection (2126 data instances)

In the "Outlier Removal" dataset, 123 data instances are removed from the CTG dataset. In the subsequent "Feature Selection" dataset, the number of features is reduced from 22 to 7 with respect to the Figures 4 and 5.

Bivariate analysis

Bivariate analysis shows few important relationships between the features.

- The value of "NSP" is heavily dependent on the value of "AC". The following Figures 6 and 7 shows their correlation.
- The value of "NSP" is dependent on the value of "UC", if "UC" value is <15 then the fetus is at risk.

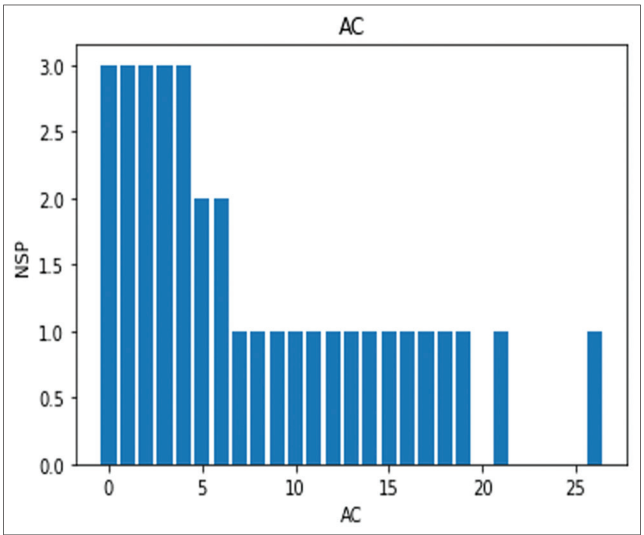

Figure 6: Bivariate analysis between "AC" and "NSP"

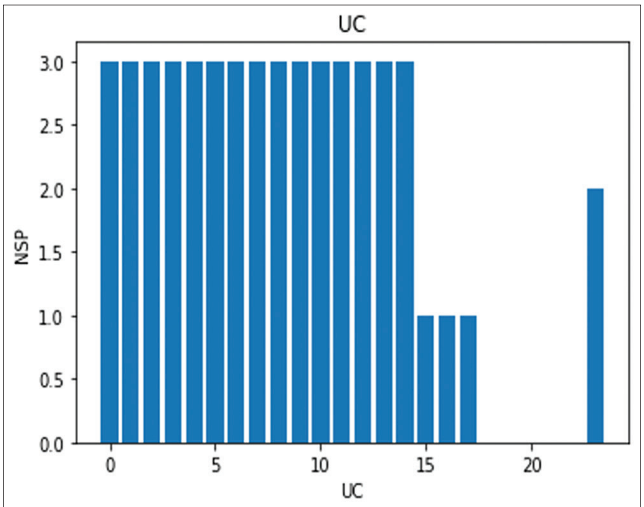

Figure 7: Bivariate analysis between "UC" and "NSP"

RESULTS AND DISCUSSION

The result section is split into two subsections. The first section deals with the result of Outlier removal, and the second section deals with feature selection. The results are arranged in chronological order showing the order of experiments.

Outlier removal

The findings generated from the experimentation done on the CTG dataset are shown in the Figure 8. About 123 instances were removed from the dataset based on the IQR outlier detection method. The performance measure of changing Outlier removal is shown in the Figure 8.

The figure shows the marginal improvements in the accuracy after the removal of Outliers in KNN, DT, and SVM. NB's prediction accuracy seems to reduce after the removal of Outliers. The graphical representation indicates that DT is the best classifier among them.

Feature selection

The list of important features selected from previous related works was taken as a separate group along with a dataset containing all the features. The comparison of these features and the performance measures is recorded in the Figure 9.

The Figure 9 shows comparable results in all the classifiers. Feature selection dataset has only seven features but its performance is on comparable levels with models trained on the whole CTG dataset with 22 features. These research findings indicate that with limited computational resources and with proper feature selection, models trained with few features provide similar results as State Of The Art (SOTA) models.

The consolidated results indicate the best baseline machine learning algorithm to employ for the classification of CTG records is the DT model. It performs much better than the other models in the group. This research also indicates that most of the models perform better when outliers are removed and performance is comparable with feature selection.

	Before Outlier Removal	After outlier Removal
KNN	0.90961	0.90961
Decision Tree	0.91248	0.91535
SVM	0.85796	0.86513
Naive Bayes	0.85222	0.82783

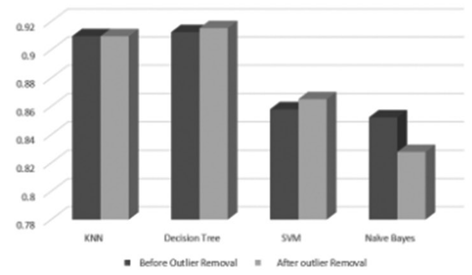

Figure 8: Graphical representation of results before and after "Outlier Removal"

	Before Feature Selection	After Feature Selection
KNN	0.90961	0.90961
Decision Tree	0.91391	0.90817
SVM	0.85796	0.86513
NB	0.85222	0.82783

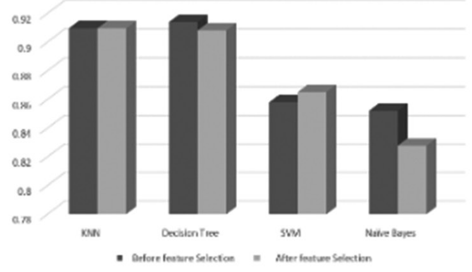

Figure 9: Graphical representation of results before and after "Feature Selection"

CONCLUSION AND FUTURE WORK

The primary objective of this research paper is to employ multiple classification algorithms on the same dataset and to compare the results of the individual algorithms. The comparison between the four different classification algorithms (KNN, Naive Bayes, DT, SVM) is done with precision, recall, and F1-score.

Based on the overall comparison using these classification reports, the DT has the best score among all these four algorithms. The comparison shows that the DT provides better results than all the other algorithms. In the future, other feature selection methods and classification methods can be used in the CTG dataset. This could increase the performance of the future models.

REFERENCES

1. Prabha D, Sudharson D. A novel machine learning approach for software reliability growth modelling with pareto distribution function. Soft Comput 2019;23:8379-87.
2. Eason G, Noble B, Sneddon IN. On Certain Integrals of Lipschitz-Hankel Type Involving Products of Bessel Functions. Vol. A247. London: The Philosophical Transactions of the Royal Society; 1955. p. 529-51.
3. Prabha D, Sudharson D. Hybrid software reliability model with Pareto distribution and ant colony optimization (PD-ACO). Int J Intell Unmanned Syst 2020;8:129-40.
4. Maxwell JC. A Treatise on Electricity and Magnetism. 3rd ed., Vol. 2. Oxford: Clarendon; 1892. p. 68-73.
5. Sudharson D, Divya P, Rajan DP, Ratheeshkumar AM. Performance analysis of Enhanced Adaboost Framework in Multifacet medical dataset. NVEO-Natural Volatiles and Essential Oils J Nveo, 2021;8:1752-6.
6. Sudharson D, Prabha D. Improved EM algorithm in software reliability growth models. Int J Powertrains 2020;9:186-99.
7. Arunkumar B, Sudharson D, Balachander K. A novel approach for boundary line detection using iot during tennis matches, Adv Electr Inf Commun Technol Life Appl 2020;13:243-6.
8. Jacobs S, Bean CP. Fine particles, thin films and exchange anisotropy. In: Rado GT, Suhl H, editors. Magnetism. Vol. 3. New York: Academic; 1963. p. 271-350.
9. Govindaraju S, Kumar AM, Sudharson D, Divya P. A novel AI and RF tutored student locating system via unsupervised dataset. Turk J Physiother Rehabil 2021;32:882-7.
10. Ratheeshkumar M, Sudharson D, Divya P, Saravanan A, Nithiyashree VK, Srinithi J. A PD ANN machine learning framework for reliability optimization in application software. In: Conference 2022 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India. 2022. p. 1-4.
11. Rizvi SWA, Singh VK, Khan RA, Fuzzy logic based software reliability quantification framework: early-stage perspective (FLSRQF). Procedia Computer Science, 2016;89:359-68.
12. Divya P, Sudharson DP, Rajan DP, Ratheeshkumar AM. Performance analysis of enhanced adaboost framework in multifacet medical dataset. Nat Volatiles Essent oils J 2021;8:1752-6.
13. Suganya K, Nethra K, Dhanyaa SS, Sudharson D. Data communication using cryptography encryption. Asian J

Comput Sci Eng 2022;7:53-7.
14. Sudharson DS, Kumar BA, Rani K. An overview of cloud scheduling algorithms. Vidyabharati Int Interdiscip Res J 2021;7:2778-82.
15. Rizvi SW, Singh VK, Khan RA. Fuzzy logic based software reliability quantification framework: Early stage perspective (FLSRQF). Procedia Comput Sci 2016;89:359-68.