

An Innovative Penalty based Heart Disease Prediction system using Novel Random Forest over Logistic Regression Classifier Algorithm

P. Prasanna Sai Teja, Veeramani T

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

ABSTRACT

Aim: The main goal of the research is see how accurately predicting heart disease by Logistic Regression (LR) and Novel Random Forest(RF) Classifications. **Materials and Methods:** Novel Random forest appealed on a heart dataset which consists of 200 records A framework for predicting heart disease in the medical field has been proposed and developed to compare the RF with a LR classifier. The sample size was calculated to be 55 for each group with 80% G performance. The sample size was calculated using a Clinlcalc analysis with Alpha and Beta values of 0.05 and 0.5, pretest performance of 80%, and enrollment rate of 1. The Accuracy of the classifier was Evaluated and Recorded. **Results:** The LR produces 89.0% in predicting the heart disease on the data set used whereas the Novel Random forest classifier predicts the same at the rate of 95.46% of the time with a statistically significant difference between the two groups ($P=0.03$; $P<0.05$) with confidence interval 95%. **Conclusion:** RF is better compared with LR in terms of both precision and accuracy.

Key words

Novel Random Forest, Logistic Regression, Data Mining, Blood pressure, Pulse rate, Heart Disease, Classification.

Imprint

P. Prasanna Sai Teja, Veeramani T. An Innovative Penalty based Heart Disease Prediction system using Novel Random Forest over Logistic Regression Classifier Algorithm. *Cardiometry*; Special issue No. 25; December 2022; p. 1477-1482; DOI: 10.18137/cardiometry.2022.25.14771482; Available from: <http://www.cardiometry.net/issues/no25-december-2022/innovative-penalty-based-heart-disease>

INTRODUCTION

Heart is a significant piece of our human body. Blood pressure, Cholesterol, Pulse rate, blood vessels, chest pain are the major reasons for heart disease. The capacity of the heart isn't done as expected, it will influence other human body parts moreover. Some dangerous components of coronary illness are Family history, High blood pressure, Pulse rate, Cholesterol, Age, poor diet, smoking. At the point when blood vessels are overstretched, the danger level of the blood vessels are expanded. This boosts the blood pressure and Pulse rate. (Sowmiya and Sumitra 2017; Sharma and Agarwal 2019; Micheletti 2019) As a result, it's critical to provide an accurate and timely diagnosis, which is a difficult challenge for doctors. They concentrate on using data mining technology to produce cost-cutting techniques. (Repaka, Ravikanti, and Franklin 2019). The proposed study helps the medical practitioners in diagnosing heart disease in an accurate way, which assists in identifying high risk of heart attacks (Saw et al. 2020).

RF algorithm is used in various healthcare and medical industries to improve the predictability of heart disease. (Kasabe and Trinity college of engineering and research 2020; 2020 *International Conference on Computer Communication and Informatics: January 22-24, 2020, Coimbatore, India* 2020) proposed a novel approach for defining significant features using Machine Learning Methods Accuracy (80%) of Heart-Disease Prediction. (Li et al. 2020) proposed a feature selection algorithm with classifier Logistic Regression for designing the higher level intelligent system to predict heart disease. (Mohan, Thirumalai, and Srivastava 2019) proposed an Algorithm heart disease prediction accuracy level of 83% through the novel random forest. (Fitriyani et al. 2020) By achieving 89% to predict heart disease, we proposed density-based spatial grouping of noise and XG gain applications performed by other models. Our team has extensive knowledge and research experience that has translate into high quality publications (Chellapa et al. 2020; Lavanya, Kannan, and Arivalagan 2021; Raj R, D, and S 2020; Shilpa-Jain et al. 2021; S, R, and P 2021; Ramadoss, Padmanaban, and Subramanian 2022; Wu et al. 2020; Kalidoss, Umapathy, and Rani Thirunavukkarasu 2021; Kaja et al. 2020; Antink et al. 2020; Paul et al. 2020; Malaikolundhan et al. 2020)

The limitation of the paper is that several works have demonstrated that the performance of Logistic Regression is poor and provides less accuracy in prediction of heart disease. A study by (Abdar et al. 2015) compares the accuracy of various mining classification algorithms in predicting heart disease. It's critical to evaluate and compare the many categorization Algorithms offer higher Accuracy. As a result, the goal of the research is to compare the Accuracy of LR and RF Algorithms in predicting heart disease.

MATERIALS AND METHODS

Experimental work was performed at the Machine Learning Institute of Computer Science and Engineering at the Saveetha School of Engineering in the Saveetha Institute of Medical and Technical Sciences in Chennai. The research paper was submitted with 200 records from the cardiac dataset. Two groups were evaluated to determine the accuracy of heart disease prediction. To increase accuracy, each group went through a total of ten iterations. The data was obtained from the Kaggle website. There are 200 rows and 14 columns in the data set. Resting blood pressure, pulse rate, chest discomfort, serum cholesterol, fasting blood glucose, heart rate, and other essential features are used in experimental design..("Website" n.d.).

Sample size was evaluated to be 55 for each group using 80% G power. In (Haq et al. 2018), the Cleveland Heart Disease data-set used with a size of 303 patients, 76 characteristics. Sample size was calculated from clinical analysis with Alpha and Beta values of 0.05 and 0.5, 95% Confidence, 80% pretest power, and a registration rate of 1.

Logistic Regression Algorithm

LR is a Machine Learning Algorithm that can be used for both Classification and Regression tasks. In this study, we trained logistic regression using the svc class in the scikit-learn library. Download the data-set heart.csv and load the dataset. The data-set is divided into a Training data-set (80%) and a Test data-set (20%). Then a logistic regression classifier is generated based on the training set. we used the Kernel Parameter value. The Test Data Set is considered by Training data-set. LR is evaluated and the accuracy is calculated.

Pseudocode for Logistic Regression Algorithm

Input: Heart disease DataSet

Output: Accuracy

1. Download data-set.
2. Randomly split the data-sets. in training (80%) and test (20%).
3. Gini is the parameter.
4. Analyze the dataset by varying dependent and independent variables
5. RF predicts the outcome in a categorical variable.
6. Finally predicts the possibility of an event using the log function.

Novel Random Forest Classifier Algorithm

The Random Classifier class from the Sklearn Ensemble package is used in this research. Use Criterian as a parameter. The parameter value is «Gini». The data-set is randomly divided as Training (80%) and Testing (20%). To predict results, we randomly selected samples and collected decision trees for each sample. We voted for each predicted result and chose the one with the most votes as the final result. The algorithm uses the RF.

Pseudocode for Novel Random Forest Classifier Algorithm

Input: Heart Disease data-set

Output: Accuracy

1. Download data-set
2. Randomly split the data-sets. in training (80%) and test (20%).
3. Gini is a parameter.
4. Use RF classifiers to build a decision tree and forecast the outcome.
5. Voting done from every sample.
6. The results of the most popular predictions were selected as the final results.

The proposed work has been tested on Google Colab. The Hardware and Software required for working experience contains CPU i7, 1TB HDD, 8GB RAM, Windows OS, Python: Colab / Jupyter. First, the data-set is split into two parts, a Training data-set and a test data-set. Algorithm is then tested on the training set and test set. Training and test data-set are updated 10 times depending on the size of test dataset.

Statistical Analysis:

In addition to experimental analysis, the Social Science Statistics Package was used to estimate study work statistically (SPSS). Mean, SD, and mean of the SD error calculated using the analysis. To compare pa-

rameters of groups, we used a variable known as t-test. Age, Gender, Chest discomfort, Blood Pressure, Pulse rate, Cholesterol, Blood glucose, ECG results, Heart rate, old peaks, gradients, segment, and goals are the independent factors in the analysis (Mohan, Thirumalai, and Srivastava). 2019). Precision and precision are dependent variables that have an impact on the outcome. 80 % G power The sample size is computed using the ClinCalc analysis, with Alpha and Beta values of 0.05 and 0.5, a Confidence level of 95%, a pretest performance of 80%, and a registration rate of 80 %, respectively.

RESULTS

Table 1 inferred the results of comparing the accuracy and accuracy of LR and RF over 10 iterations with an accuracy of LR 89.0% and RF 95.46%.

Table 1

Correlations for Study. Data collection from N=10 sample datasets for Novel Random Forest Algorithm(95.46%) Compared with Logistic Regression(89.0%) using target variable as Independent variable.

Sample No.	Test Size	RF	LR
1	784	95.46	89.0
2	774	95.12	88.88
3	764	95.09	88.78
4	660	94.3	88.68
5	640	94.12	88.32
6	620	94.05	87.85
7	610	93.29	87.53
8	600	93.23	87.31
9	540	93.17	87.19
10	510	93.12	87.11

Table 2 inferred the statistical analysis of LR and RF with different test datasets. The LR and RF mod-

el Accuracy of mean value has 94.09% and 88.06 as well as precision has 59.0600 and 81.4900. The Standard deviation RF and LR accuracy and precision(0.24507,1.31540,6.46275,4.59938)has a significantly greater precision than LR. The RF and LR standard error Mean of accuracy and Precision (0.17283,0.82738, 2.04370,1.45445) algorithm outperforms the LR technique in terms of performance.

Table 3 Both groups' statistical analysis of a statistically significant difference between the groups (P=0.03; P<0.05) with confidence interval 95% with two groups, there is no discernible difference, hence RF is better than LR.

Table 2

Statistical Analysis of Mean, SD and Standard Error of Accuracy of LR and RF algorithms. There is a statistically significant difference in accuracy values between the algorithms. RF has better accuracy(88.06%) when compared to accuracy (94.09%) than LR.

GROUP	N	Mean	Std.Deviation(SD)	Std.Error Mean
ACCURACY				
LR	10	94.09	0.24507	0.17283
RF	10	88.06	1.31540	0.82738
PRECISION				
LR	10	59.0600	6.46275	2.04370
RF	10	81.4900	4.59938	1.45445

From Fig. 1 and Fig.2, it is inferred that the ROC graph inferes performance of LR at various classification thresholds.

DISCUSSION

RF is superior to the LR in predicting heart disease by considering the accuracy. However, the average error of RF seems to be better than the average error of LR. Prediction of heart disease is a major issue in Healthcare industries. Experimental work was

Table 3

Comparison of the Significance level for LR and RF algorithms with value $p < 0.05$. Both LR and RF is a statistically significant difference between the two groups ($p=0.03$; $p<0.05$) with confidence interval 95%.

	Levene's Test for Equality of Variance		T-test for Equality of Means						
	F	Sig	t	df	Sig(2-tailed)	Mean Difference	Std.Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Accuracy	9.2	0.03	16.571	18	.000	12.983	0.8876	11.9284	13.7834
Precision	3.992	.061	-8.942 -8.942	18 16.255	.000 .000	-22.43 -22.43	2.50841 2.50841	-27.6999 -27.7408	-17.1600 -17.1191

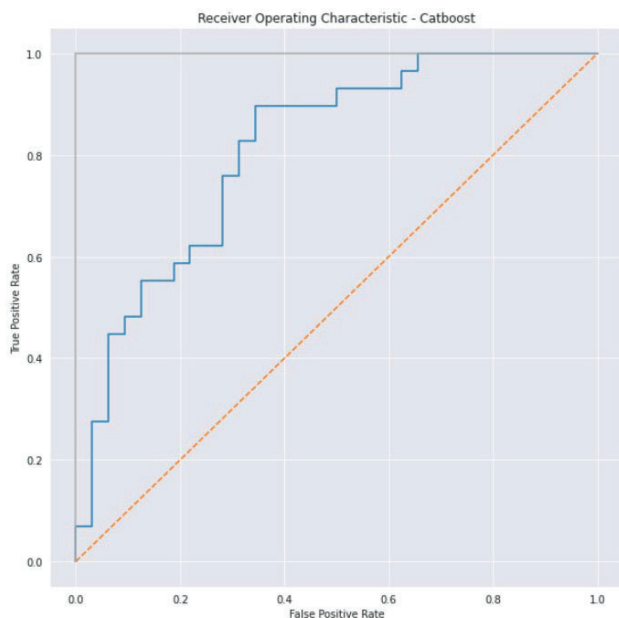


Fig. 1. Receiving Operating characteristic (ROC) Curve for Logistic Regression

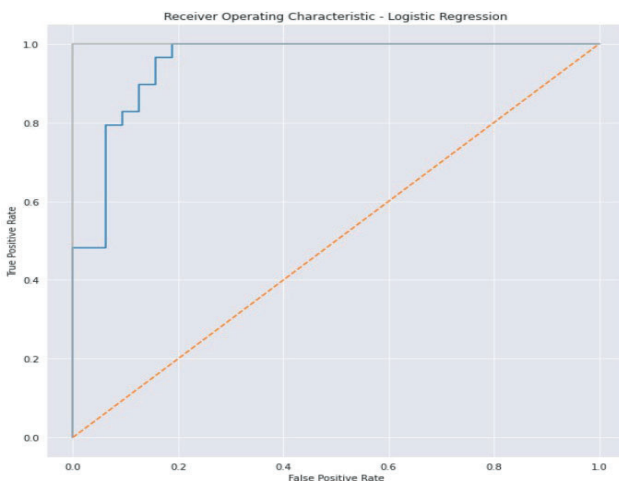


Fig. 2. Receiving Operating characteristic (ROC) Curve for RF

done among two groups Logistic Regression and RF by varying test size. From Experimental results (Fig. 3 and Fig. 4) done in Google colab, the accuracy and precision of RF is 95.46% and 89.0%. This depicts that RF is better than Logistic Regression. The various parameters like Accuracy F-measures are also compared. From the SPSS graph conveys Random forest Classifier performs better in terms of accuracy (84.7%) and precision (81.4%) compared with the Logistic Regression algorithm. Fig. 3 depicts the mean error of RF is found to little higher than Logistic Regression, which has to be minimized.

Our university is committed to conducting research that is both high-quality and evidence-based

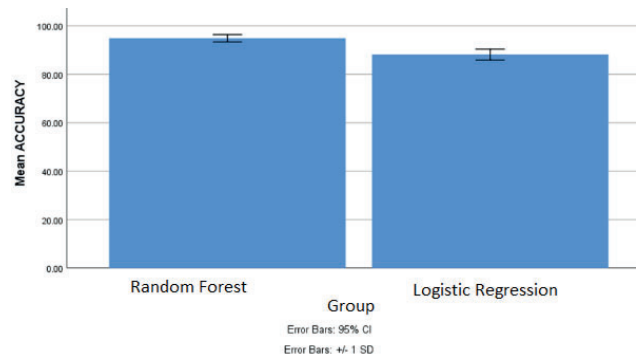


Fig. 3. Bar graph analysis of RF and LR algorithm. Graphical representation shows the mean efficiency of 94.09% and 88.06% for the proposed LR algorithm and RF algorithm respectively. X-axis: LR and RF algorithm Y-axis: Mean precision \pm 1 SD.

significant success in this regard.. ((Vijayashree Priyadharsini 2019; Ezhilarasan, Apoorva, and Ashok Vardhan 2019; Ramesh et al. 2018; Mathew et al. 2020; Sridharan et al. 2019; Pc, Marimuthu, and Devadoss 2018; Ramadurai et al. 2019)). We desire that this study provides to the wealthy records of the area.

Despite the fact that the results of the study are statistically and experimentally superior, Data sets, Accuracy evaluation cannot yield a good result. In addition, mean error in RF looks to be larger than in LR. It would be preferable if the mean error could be lowered somewhat. However, optimization algorithm techniques can be used to increase accuracy while lowering costs of the standard deviation of the task. Feature choice procedures may be used earlier than category to grow the accuracy of classifier category. As a result, we can leverage FS techniques to minimize calculation time and enhance classification accuracy of classifiers.

CONCLUSION

RF Classifier is a Technique used for averaging Accuracy and Precision. This work conveys Accuracy and Precision for Heart Disease Prediction Using RF is better than the LR. The results show that RF performs much better than LR in properly predicting heart disease, however the mean error is a bit greater than with Logistic Regression. As a result, it is concluded that the RF classifier produces appropriate accuracy and precision when compared to the LR classifier.

DECLARATION

Conflict of interests

This manuscript has no conflicts of interest.

Author Contribution

Data acquisition, Data Analysis, Algorithm Framing, Implementation, and article authoring are all tasks that need to be completed were all done by author P. Prasanna SaiTeja. Dr Veeramani was involved in workflow design, coaching, and manuscript revision.

Acknowledgements

We are here to thank Saveetha School of Engineering and Saveetha Institute of Medical and Technical Sciences (Saveetha University) for their ongoing support and facilities in carrying out this research.

Funding

The following company helped us accomplish the study by offering financial support.

1. Sri Cube Innovations Pvt Ltd, Vijayawada.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Abdar, Moloud, Sharareh R. Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto, and Goli Arji. 2015. "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases." *International Journal of Electrical and Computer Engineering (IJECE)*. <https://doi.org/10.11591/ijece.v5i6.pp1569-1576>.
2. Antink, Christoph Hoog, Joana Carlos Mesquita Ferreira, Michael Paul, Simon Lyra, Konrad Heimann, Srinivasa Karthik, Jayaraj Joseph, et al. 2020. "Fast Body Part Segmentation and Tracking of Neonatal Video Data Using Deep Learning." *Medical & Biological Engineering & Computing* 58 (12): 3049–61.
3. Chellapa, L. R., S. Rajeshkumar, M. I. Arumugham, and S. R. Samuel. 2020. "Biogenic Nanoselenium Synthesis and Evaluation of Its Antimicrobial, Antioxidant Activity and Toxicity." *Bioinspired Biomimetic and Nanobiomaterials*, July, 1–6.
4. Ezhilarasan, Devaraj, Velluru S. Apoorva, and Nandhigam Ashok Vardhan. 2019. "Syzygium Cumini Extract Induced Reactive Oxygen Species-Mediated Apoptosis in Human Oral Squamous Carcinoma Cells." *Journal of Oral Pathology & Medicine: Official Publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology* 48 (2): 115–21.
5. Fitriyani, Norma Latif, Muhammad Syafrudin, Ganjar Alfian, and Jongtae Rhee. 2020. "HDPM: An

Effective Heart Disease Prediction Model for a Clinical Decision Support System." *IEEE Access*. <https://doi.org/10.1109/access.2020.3010511>.

6. Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. 2018. "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms." *Mobile Information Systems*. <https://doi.org/10.1155/2018/3860146>.
7. Kaja, Rekha, Anandh Vaiyapuri, Mohamed Sherif Sirajudeen, Hariraja Muthusamy, Radhakrishnan Unnikrishnan, Mohamed Waly, Samuel Sundar Doss Devaraj, Mohamed Kotb Seyam, and Gopal Nambi S. 2020. "Biofeedback Flutter Device for Managing the Symptoms of Patients with COPD." *Technology and Health Care: Official Journal of the European Society for Engineering and Medicine* 28 (5): 477–85.
8. Kalidoss, Ramji, Snehalatha Umapathy, and Usha Rani Thirunavukkarasu. 2021. "A Breathalyzer for the Assessment of Chronic Kidney Disease Patients' Breathprint: Breath Flow Dynamic Simulation on the Measurement Chamber and Experimental Investigation." *Biomedical Signal Processing and Control* 70 (September): 103060.
9. Kasabe, Riddhi, and Trinity college of engineering and research. 2020. "Heart Disease Prediction Using Machine Learning." *International Journal of Engineering Research and*. <https://doi.org/10.17577/ijert-v9is080128>.
10. Lavanya, M., P. Muthu Kannan, and M. Arivalagan. 2021. "Lung Cancer Diagnosis and Staging Using Firefly Algorithm Fuzzy C-Means Segmentation and Support Vector Machine Classification of Lung Nodules." *International Journal of Biomedical Engineering and Technology* 37 (2): 185.
11. Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor. 2020. "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare." *IEEE Access*. <https://doi.org/10.1109/access.2020.3001149>.
12. Malaikolundhan, Harikrishna, Gowsik Mookkan, Gunasekaran Krishnamoorthi, Nirosha Matheswaran, Murad Alsawalha, Vishnu Priya Veeraraghavan, Surapaneni Krishna Mohan, and Aiting Di. 2020. "Anticarcinogenic Effect of Gold Nanoparticles Synthesized from Albizia Lebbeck on HCT-116 Colon Cancer Cell Lines." *Artificial Cells, Nanomedicine, and Biotechnology* 48 (1): 1206–13.
13. Mathew, M. G., S. R. Samuel, A. J. Soni, and K. B. Roopa. 2020. "Evaluation of Adhesion of Streptococcus

- Mutans, Plaque Accumulation on Zirconia and Stainless Steel Crowns, and Surrounding Gingival Inflammation in Primary ...” *Clinical Oral Investigations*. <https://link.springer.com/article/10.1007/s00784-020-03204-9>.
14. Micheletti, Angelo. 2019. “Congenital Heart Disease Classification, Epidemiology, Diagnosis, Treatment, and Outcome.” *Congenital Heart Disease*. https://doi.org/10.1007/978-3-319-78423-6_1.
 15. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. 2019. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.” *IEEE Access*. <https://doi.org/10.1109/access.2019.2923707>.
 16. Paul, M., S. Karthik, J. Joseph, M. Sivaprakasam, J. Kumutha, S. Leonhardt, and C. Hoog Antink. 2020. “Non-Contact Sensing of Neonatal Pulse Rate Using Camera-Based Imaging: A Clinical Feasibility Study.” *Physiological Measurement* 41 (2): 024001.
 17. Pc, J., T. Marimuthu, and P. Devadoss. 2018. “Prevalence and Measurement of Anterior Loop of the Mandibular Canal Using CBCT: A Cross Sectional Study.” *Clinical Implant Dentistry and Related Research*. <https://europepmc.org/article/med/29624863>.
 18. Raj R, Kathiswar, Ezhilarasan D, and Rajeshkumar S. 2020. “ β -Sitosterol-Assisted Silver Nanoparticles Activates Nrf2 and Triggers Mitochondrial Apoptosis via Oxidative Stress in Human Hepatocellular Cancer Cell Line.” *Journal of Biomedical Materials Research. Part A* 108 (9): 1899–1908.
 19. Ramadoss, Ramya, Rajashree Padmanaban, and Balakumar Subramanian. 2022. “Role of Bioglass in Enamel Remineralization: Existing Strategies and Future Prospects-A Narrative Review.” *Journal of Biomedical Materials Research. Part B, Applied Biomaterials* 110 (1): 45–66.
 20. Ramadurai, Neeraja, Deepa Gurunathan, A. Victor Samuel, Emg Subramanian, and Steven J. L. Rodrigues. 2019. “Effectiveness of 2% Articaine as an Anesthetic Agent in Children: Randomized Controlled Trial.” *Clinical Oral Investigations* 23 (9): 3543–50.
 21. Ramesh, Asha, Sheeja Varghese, Nadathur D. Jayakumar, and Sankari Malaippan. 2018. “Comparative Estimation of Sulfiredoxin Levels between Chronic Periodontitis and Healthy Patients – A Case-Control Study.” *Journal of Periodontology* 89 (10): 1241–48.
 22. Repaka, Anjan Nikhil, Sai Deepak Ravikanti, and Ramya G. Franklin. 2019. “Design And Implementing Heart Disease Prediction Using Naives Bayesian.” 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). <https://doi.org/10.1109/icoei.2019.8862604>.
 23. Saw, Montu, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, and Nidhi Lal. 2020. “Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning.” 2020 International Conference on Computer Communication and Informatics (ICCCI). <https://doi.org/10.1109/iccci48352.2020.9104210>.
 24. Sharma, Himanshu, and Rohit Agarwal. 2019. “An Intelligent Diagnosis System for Prediction of Heart Disease Risk Based on Feature Selection and Ensemble Classification Techniques.” *Journal of Advanced Research in Dynamical and Control Systems*. <https://doi.org/10.5373/jardcs/v11sp10/20192856>.
 25. Shilpa-Jain, D. P., Jogikalmat Krithikadatta, Dinesh Kowsky, and Velmurugan Natanasabapathy. 2021. “Effect of Cervical Lesion Centered Access Cavity Restored with Short Glass Fibre Reinforced Resin Composites on Fracture Resistance in Human Mandibular Premolars-an in Vitro Study.” *Journal of the Mechanical Behavior of Biomedical Materials* 122 (October): 104654.
 26. Sowmiya, C., and P. Sumitra. 2017. “Analytical Study of Heart Disease Diagnosis Using Classification Techniques.” 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). <https://doi.org/10.1109/itcosp.2017.8303115>.
 27. Sridharan, Gokul, Pratibha Ramani, Sangeeta Pantankar, and Rajagopalan Vijayaraghavan. 2019. “Evaluation of Salivary Metabolomics in Oral Leukoplakia and Oral Squamous Cell Carcinoma.” *Journal of Oral Pathology & Medicine: Official Publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology* 48 (4): 299–306.
 28. S, Sudha, Kalpana R, and Soundararajan P. 2021. “Quantification of Sweat Urea in Diabetes Using Electro-Optical Technique.” *Physiological Measurement* 42 (9). <https://doi.org/10.1088/1361-6579/ac1d3a>.
 29. Vijayashree Priyadharsini, Jayaseelan. 2019. “In Silico Validation of the Non-Antibiotic Drugs Acetaminophen and Ibuprofen as Antibacterial Agents against Red Complex Pathogens.” *Journal of Periodontology* 90 (12): 1441–48.
 30. “Website.” n.d. Accessed March 19, 2021. <https://www.kaggle.com/nareshbhat/health-care-dataset-on-heart-attack-possibility>.
 31. Wu, Shuang, Shanmugam Rajeshkumar, Malini Madasamy, and Vanaja Mahendran. 2020. “Green Synthesis of Copper Nanoparticles Using Cissus Vitiginea and Its Antioxidant and Antibacterial Activity against Urinary Tract Infection Pathogens.” *Artificial Cells, Nanomedicine, and Biotechnology* 48 (1): 1153–58.