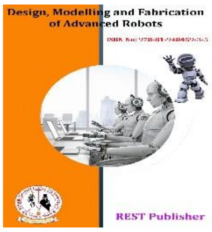
DOI: https://doi.org/10.46632/dmfar/1/2/9

Design, Modelling and Fabrication of Advanced Robots
Vol: 1(2), 2022
REST Publisher; ISBN: 978-81-948459-3-5
Website: http://restpublisher.com/book-series/dmfar/


Design, Modelling and Fabrication of Advanced Robots
ISBN No: 978-81-948459-3-5
REST Publisher

Comparative Analysis: Personalized Random Forest Algorithm versus Gaussian Naive Bayes Algorithm for Diabetes Mellitus Classification in Humans

N. Penchalaiah
College of Engineering and Technology, Gudur, Andhra Pradesh 524101, India.
*Corresponding Author Email: *pench75@gmail.com*

Abstract: In this research, a new Random Forest algorithm is evaluated against a Gaussian Naive Bayes algorithm to enhance the accuracy and precision of diabetes mellitus classification in individuals. The study employs the Pima Indian Diabetes Dataset sourced from Kaggle, a machine learning repository, featuring eight attributes used as input variables for the classifiers. Each classifier is tested with 20 samples per group. An assumed G power of 0.8 guides the study. The results indicate that the novel Random Forest algorithm achieved 91% accuracy and 85% precision, while the Naive Bayes method reached 78% accuracy and 84% precision. A significance level of $p < 0.05$ is employed. Notably, this investigation underscores the superiority of the Novel Random Forest method over the Naive Bayes algorithm for the accurate categorization of diabetes mellitus within the dataset.

Keywords: Gaussian Naive Bayes classifier, Novel Random Forest algorithm, Diabetes mellitus, Blood glucose level, Machine learning, Classification of diabetes

1. INTRODUCTION

Diabetes, commonly known as diabetes mellitus, is a major public health concern. Early disease identification can help save precious human resources. Large datasets in the form of clinical patient records and pathological examination reports are available from a variety of medical sources and can be utilized in real-world applications. Both micro and macro vascular damage can occur as a result of diabetes problems. Diabetes doubles or quadruples the risk of cardiovascular disease. As a result, diabetes is one of the top causes of disease-related death worldwide [1]. A deficiency in insulin secretion and activity causes the metabolic disease diabetes mellitus. Hyperglycemia (blood glucose concentration >180mg/DL) and hypoglycemia (blood glucose concentration 70mg/DL) result as a result [2]. Diabetic hyperglycemia damages, malfunctions, and fails organs such as the kidneys, eyes, nerves, blood vessels, and core [3]. The study's goal is to increase classification accuracy and precision in order to detect diabetes mellitus earlier and prevent diabetic complications [4]. The study's applications were in the medical field, such as diabetes control and artificial pancreas systems [5-6]. In the last five years, several research papers on Random Forest and Naive Bayes have been published. In Google Scholar, 446 research articles were published, while in Science Direct, 56 research articles were published. Machine learning algorithms for diabetic disease have been studied extensively in recent years. Random Forest method and Naive Bayes provided the author with classification accuracy of 73.04 percent and 76.69 percent, respectively, and classification precision of 72.7 percent and 76.1 percent. This author primarily concentrated on using BLE-sensor devices, smartphones, and machine learning-based algorithms to forecast diabetes and blood glucose levels [7]. This author predicts diabetes using several algorithms in order to avoid type 2 diabetes risk factors. The accuracy of the Random Forest algorithm and Naive Bayes is 77 percent and 79 percent, respectively [8]. To predict diabetes mellitus, data mining techniques and machine learning methodologies were applied. This unique Random Forest method was a huge hit and is still in use today. The majority of the data was derived from clinical datasets [9]. All of these ways are effective, according to the findings of the experiments. The topic concludes with some key objectives for future Bayesian network classifier research [10]. Our university is dedicated to doing high-quality, evidence-based research and has achieved success in a number of areas [11-15]. The limitations of a research paper are using less data, the number of subjects considered for detection is less. To overcome these limitations, large datasets are used for detection of diabetes as the proposed work. Random forest had performed better than Naive Bayes algorithm. However, in the proposed work, Novel Random Forest algorithm performance was analyzed using a large dataset collected from Pima Indians dataset.

2. MATERIALS AND METHODS

The review setting was at Saveetha School of Engineering utilizing the Google co-lab online recreation instrument. The DSP research center climate is at the branch of Biomedical designing. Two gatherings were contrasted and an example size of 20 allocated for each gathering. The clinical application is utilized to decide the base example size esteem. Reproductions of the gatherings depended on the python design with pre-test G power 80% utilized for testing [16]. In bunch 1, example readiness was taken as a Gaussian Naive Bayes classifier strategy. In bunch 2, example planning was taken as a Random Forest calculation for the purpose of testing. For bunch 1, 20 examples prepared utilizing GNB and for bunch 2, 20 examples prepared utilizing RF. WUXIA screen with a goal of 1920*1080 pixels with arrangement of tenth Gen, i5, 8GB RAM, 552 SDD and Google co-lab programming with additional items expected for the purpose of testing. In bunch 1 example arrangement is done by taking a dataset from the Kaggle. Load the dataset into the google colab. Ascertain the exactness and accuracy utilizing various cycles. Take the 20 examples from every exactness and accuracy and product those values into the Excel sheet. Make the charts as indicated by exactness and accuracy independently. In bunch 2 example arrangement is done by taking a dataset from the Kaggle. Load the dataset into the google colab. Work out the exactness and accuracy utilizing different iterations. Take the 20 examples from every exactness and accuracy and product that qualities into the Excel sheet. Make the diagrams as per exactness and accuracy independently. The Diabetes dataset gathered from Kaggle and UCI should be handled prior to applying it to the AI model. The handled dataset is given for preparing and testing. The preprocessed dataset with highlights is given as contribution to the RF and Naive Bayes classifier. From the all-out example size 75% of the information is given for preparing and the leftover 25% is given for testing. In the Pima Indians, diabetes dataset is gathered from the Kaggle with a sum of 2000 records. The records were gathered from test tests in the National foundation of diabetes and stomach related and kidney infections. The example size was determined involving clincalc.com with alpha worth as 0.05, 95% certainty time frame with a mean exactness of gathering 1 76.69% and gathering 2 as 73.04% [7].

3. NAIVE BAYES ALGORITHM

The Naive Bayes model is not difficult to direct. The dataset was characterized utilizing a guileless bayes classifier and is just subject to one component in a class and not on some other element. For instance, assuming the natural product is red, has a distance across of around 2 inches, and is round in shape, the natural product is probably going to be a tomato. Notwithstanding the reality every one of the characteristics are free, they all add to the natural product's ID, which is the reason it was given the name Naive.

4. RANDOM FOREST ALGORITHM

Arbitrary Forest is an administered AI calculation. It structures like a tree. A gathering of choice trees will fabricate a Random Forest and the choice trees are consolidated to obtain more exact outcomes. The larger number of trees will build the exactness of the dataset. In the Random Forest calculation, the preparation time is extremely less contrasted and different calculations. It will give high exactness in any event, for the huge dataset. Irregular Forest adds to the model's haphazardness while also encouraging the trees. When splitting a hub, it looks for the best component among an irregular collection of parts, rather than looking for the main section. As a result, there is a diverse selection that, in general, produces an unsurpassed model [17]. The following steps can be used to illustrate the working process:

Step 1: Pick K data points at random from the training collection.

Step 2: Build decision trees for the data points you've chosen (subsets).

Step 3: Decide on the number N for the decision trees you want to build.

Step 4: Find the predictions of each decision tree for new data points, and allocate the new data points to the group with the most votes.

The results are partitioned into two classifications: sound and undesirable. These two gatherings are addressed by a number, with 0 addressing solid and 1 addressing unfortunate. Subsequently, many trees are inherent similar way, forming a timberland of trees that helps right order.

5. STATISTICAL ANALYSIS

Measurable Package for the Social Sciences [18] and google colab programming were utilized. The Pima Indians' diabetes dataset contains 8 ascribes in like manner. Normal credits are Age, pregnancy, debilitated thickness, weight record, glucose, pulse esteem, insulin, diabetes family work, class variable. The datasets were imported and broke down utilizing a free example t-test. The age characteristic segment is taken out for preparing and testing of the model as it doesn't contribute a lot of data and arrangement [19]. Insulin, weight file, glucose, circulatory strain esteem, diabetes family work are the free factors. Utilizing SPSS and beginning diagrams was acquired. The autonomous example test was performed to track down the measurable importance between the gatherings.

6. RESULTS AND DISCUSSION

Utilizing the dataset, the proposed RF calculation is superior to Naive Bayes calculation in correlation with all presentation measurements. Accordingly, the exactness of proposed work is higher than the current one. Table 1 addresses the measurable elements separated from the information for preparing the learning calculation. The measurable highlights removed are mean, standard deviation, least, 25% quantile, half quantile, 75% quantile and greatest. Table 2 saw that RF classifiers have acquired a superior order exactness pace of 91% and accuracy pace of 92% than Naive Bayes classifiers which have exactness and accuracy as 78% and 84% individually. The example size of gathering 1 is 20 and the example size of gathering 2 is 20. Table 3 addresses correlation of gathering insights for RF and NB bunches in SPSS programming RF calculation gives mean (0.8710), standard deviation (0.02989) and standard blunder mean (0.00668) for 20 examples. Table 3 shows an autonomous example t-test giving importance, mean distinction and 95% certainty time frame contrast of the exactness and accuracy. There is a huge distinction between the two gatherings p<0.05. Figure 1 shows the correlation of precision for RF and NB calculations on the diabetes dataset. RF accomplished a precision of 91% and NB accomplished an exactness of 78% in the diabetes dataset. Figure 2 shows the correlation of accuracy for RF and NB calculations on diabetes dataset. RF accomplished an accuracy of 92% and NB accomplished a precision of 84% in the diabetes dataset. Figure 3 shows the correlation of mean exactness and accuracy upsides of RF and NB calculations as a bar graph. In that when analyzed two calculations RF and NB Random Forest performs preferable exactness and accuracy over Naive Bayes calculation. In this review, it was seen that RF has all the earmarks of being superior to Naive Bayes with an exactness of 91% and an accuracy of 92% (p< 0.05). In this investigation, the presentation of Novel RF and NB is broken down in characterizing diabetes mellitus illness from the dataset acquired from Kaggle. The proposed work implies that the RF classifier performs better order contrasted with the NB classifier. The creators have accomplished, in the wake of computing the conviction factor it will get the worth of the assurance of the class chose on the estimation of the Random Forest. After the execution of 100 clinical records information taken from RSUD Bendan Pekalongan, separated into 2 that is 70 preparation information and 30 testing information, the degree of precision of the aftereffects of the conclusion framework is just about as much as 100 percent exact as per the finding of the master. (Waty et al. [20] by involving the strategy for innocent bayes as a technique to deal with information on the patient's conclusion. Test consequences of this framework demonstrate that the framework can foresee the kind of diabetes in patients, from how much information as much as 200 patient information, with a result that is the type of Diabetes Without Complications, Diabetes Type II and Normal yet gotten the most minimal precision rating of 39% and the worth of the greatest exactness of 80%. In this review [21], the proposed technique gives high exactness a precision worth of 90.36% and choice Stump gave less precision than others by giving 83.72% precision [22]. From these outcomes, it tends to be presumed that the analysis calculation Random Forest and K-Nearest Neighbor have the very degree of precision that is equivalent to 91 % and the mistake rate by 9%, however the information arrangement process shortcoming conclusion on both the calculations have contrasts. [23] in this paper, proposed a clever thought for diagnosing diabetes utilizing a RF classifier. Results acquired show RF beats with the most noteworthy precision contrasted and different calculations [24]. As the constraint of this review, the changes can be made in the AI calculations to acquire a superior arrangement exactness rate (91%) and examining size can likewise be expanded in our review. NB neglects to perform when the objective classes are covering. It requires some investment to prepare when the dataset size is enormous. RF isn't computationally productive when the dataset is very large. Our group has broad information and exploration experience that has convert into great distributions [25-30]. In the eventual fate of this work RF classifier exactness can be improved by adding more information in the preparation sets, utilizing multiclass RF, and by consolidating it with other ANN techniques. Likewise, multi-faceted information can be changed over into twofold information to further develop exactness. The PC supported demonstrative (CAD) framework can be created in the future to assist clinicians with foreseeing diabetes mellitus.

TABLE 1. Statistical features for diabetes dataset - sample.

	Pregnancies	**Glucose**	**Blood pressure**	**Skin thickness**	**Insulin**	**Body mass index**	**Diabetes pedigree function**	**Class variable**
count	2000.00000	2000.00	2000.00	2000.000	2000.0	2000.0	2000.0000	2000.00
mean	3.703500	59.4960	69.1455	20.93500	80.254	3.1930	0.470930	0.34200
std	3.306063	30.5576	19.1883	16.10324	111.18	8.1499	0.323553	0.47449
min	0.000000	0.00000	0.00000	0.000000	0.0000	0.0000	0.078000	0.00000
25%	1.000000	37.0000	63.0000	0.000000	0.0000	27.375	0.244000	0.00000
50%	3.000000	55.0000	72.0000	23.00000	40.000	32.300	0.376000	0.00000
75%	6.000000	79.0000	80.0000	32.00000	130.00	36.800	0.624000	1.00000
max	17.000000	135.000	122.000	110.0000	744.00	80.600	2.42000	1.00000

TABLE 2. Comparison of RF and NB classifiers in terms of accuracy and precision. RF has more accuracy (91%) and precision (92%) than NB which has accuracy (78%) and precision (84%).

Classifiers	**Accuracy (%)**	**Precision (%)**
RF	91	92
NB	78	84

TABLE 3. Comparison of group statistics for RF and NB groups in SPSS software RF algorithm provides mean (0.8710), standard deviation (0.02989) and standard error mean (0.00668) for 20 samples.

		N	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	RF	20	0.8710	0.02989	0.00668
	NB	20	0.7630	0.01081	0.00242
Precision	RF	20	0.8870	0.03080	0.00689
	NB	20	0.7920	0.03286	0.00735

TABLE 4. Independent sample test providing significance, mean difference and 95% confidence interval of the difference of the accuracy and precision.

Independent Samples Test										
		T-test for equality of means								
		Levene's Test							95% Confidence Interval of the Difference	
		F	sig.	t	df	Sig. (2-tailed)	Mean difference	Std.error difference	lower	upper
Accuracy	Equal variances assumed	27.12	.000	15.19	38	0.00	0.10800	0.00711	0.093	0.1223
	Equal variances not assumed			15.19	23.88	0.00	0.17250	0.00536	0.093	0.1226
Precision	Equal variances assumed	0.688	.022	9.433	38	0.00	0.01007	0.01007	0.074	0.1153
	Equal variances not assumed			9.433	37.84	0.00	0.02100	0.00624	0.008	0.0336

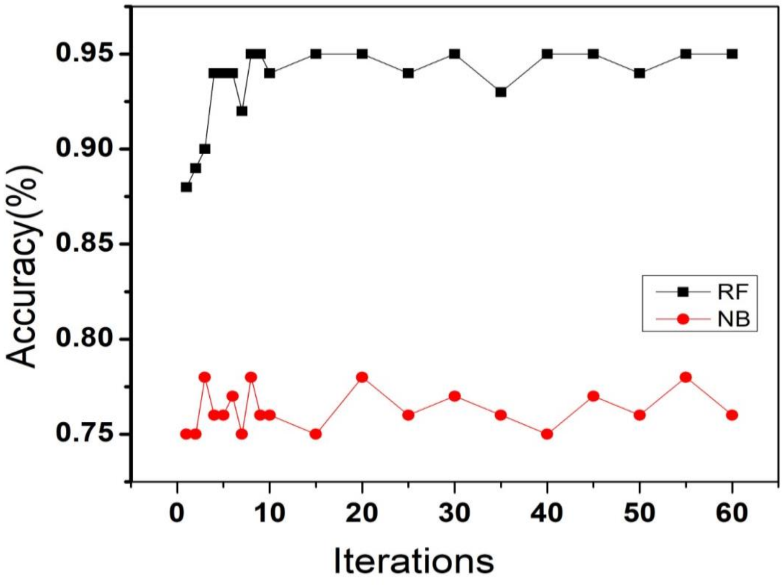


FIGURE 1. Comparison of RF and NB at different accuracy values. It has taken through different iterations on the X-axis and accuracy on Y-axis.

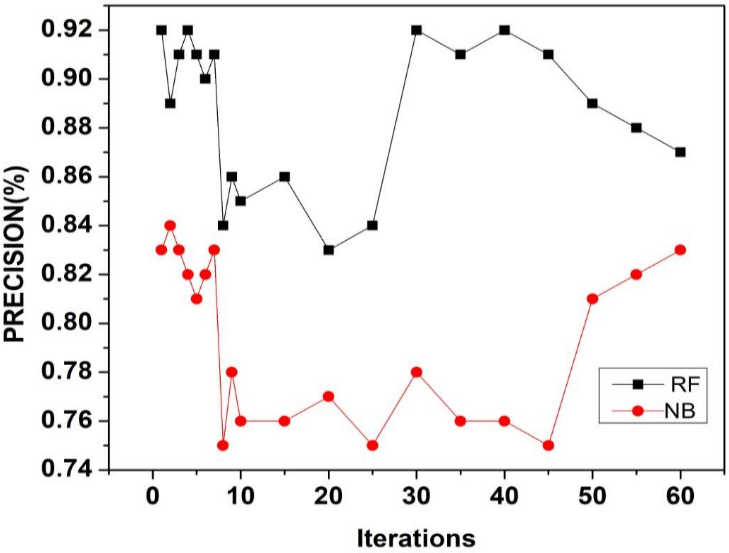


FIGURE 2. Comparison of RF and NB at different precision values. It has taken through different iterations on the X-axis and precision on Y-axis.

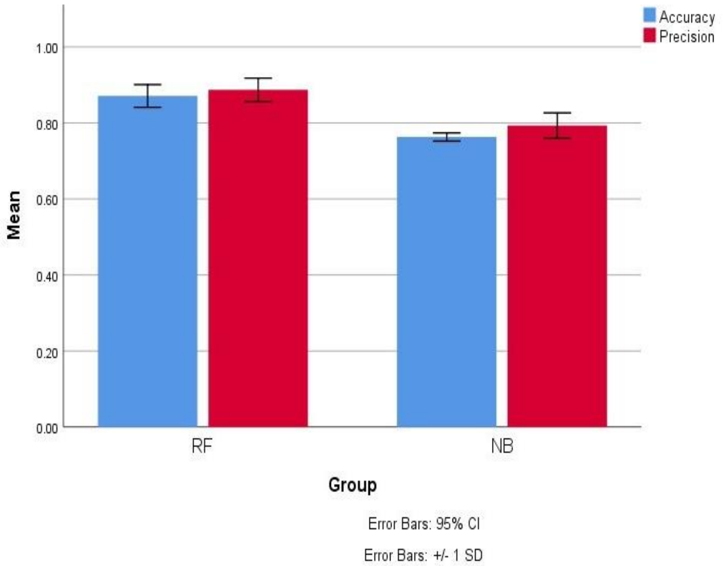


FIGURE 3. The Bar chart compares mean (+/- 1 SD) accuracy and precision of RF vs NB with different iterations.

7. CONCLUSIONS

Irregular Forest gives altogether great execution in distinguishing diabetes sickness contrasted with Naive Bayes. Arbitrary Forest created an exactness and accuracy of 91% and 92% contrasted with the Naive Bayes precision and accuracy of 78% and 84% individually. Author Contribution. Data collection, data analysis, and manuscript writing were all done by author BN. Author MJK was involved in the manuscript's conceptualization, data validation, and critical evaluation. Acknowledgement. The authors would like to thank Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences for providing the essential infrastructure for this research to be completed successfully. Funding. We thank the following organizations: Biozone Research Technologies Pvt Ltd, India. Saveetha University. 3. Saveetha Institute of Medical and Technical Sciences. Saveetha School of Engineering.

REFERENCES

[1]. Chaki, J. et al. (2020) ‘Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review’, Journal of King Saud University - Computer and Information Sciences [Preprint]. doi:10.1016/j.jksuci.2020.06.013.

[2]. Li, K. et al. (2020) ‘Convolutional Recurrent Neural Networks for Glucose Prediction’, IEEE journal of biomedical and health informatics, 24(2), pp. 603–613.

[3]. Landau, S. and Everitt, B.S. (2003) A Handbook of Statistical Analyses Using SPSS. Chapman and Hall/CRC

[4]. Goswami, A. et al. (2014) 'Bleeding disorders in dental practice: A diagnostic overview', Journal of the International Clinical Dental Research Organization, p. 143. doi:10.4103/2231-0754.143529

[5]. Manju, A. and Valarmathie, P. (2021) 'Video analytics for semantic substance extraction using OpenCV in python', Journal of ambient intelligence and humanized computing, 12(3), pp. 4057–4066.

[6]. Kovatchev, B. et al. (2016) 'The Artificial Pancreas in 2016: A Digital Treatment Ecosystem for Diabetes', Diabetes Care, pp. 1123–1126. doi:10.2337/dc16-0824.

[7]. Alfian, G. et al. (2018) 'A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing', Sensors, 18(7). doi:10.3390/s18072183.

[8]. Sai, P.M.S. et al. (2020) 'Survey on Type 2 Diabetes Prediction Using Machine Learning', 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) [Preprint]. doi:10.1109/iccmc48092.2020.iccmc-000143.

[9]. Kavakiotis, I. et al. (2017) 'Machine Learning and Data Mining Methods in Diabetes Research', Computational and Structural Biotechnology Journal, pp. 104–116. doi:10.1016/j.csbj.2016.12.005.

[10]. Gupta, S., Karanam, N. K., Konijeti, R., & Dasore, A. Thermodynamic Analysis and Effects of Replacing HFC by Fourth-Generation Refrigerants in VCR Systems. International Journal of Air-Conditioning and Refrigeration, 26(2): 1850013 (2018).

[11]. Returi, MC, Konijeti, R and Dasore, A. Heat transfer enhancement using hybrid nanofluids in spiral plate heat exchangers. Heat Transfer—Asian Res. 48: 3128- 3143 (2019).

[12]. Chaitanya, N.C. et al. (2017) 'Role of Vitamin E and Vitamin A in Oral Mucositis Induced by Cancer Chemo/Radiotherapy- A Meta-analysis', Journal of clinical and diagnostic research: JCDR, 11(5), pp. ZE06–ZE09.

[13]. Christo, M.S. et al. (2021) 'Ensuring Improved Security in Medical Data Using ECC and Blockchain Technology with Edge Devices', Security and Communication Networks, 2021. doi:10.1155/2021/6966206.

[14]. Christo, M.S., Vasanth, K. and Varatharajan, R. (2020) 'A decision based asymmetrically trimmed modified winsorized median filter for the removal of salt and pepper noise in images and videos', Multimedia tools and applications, 79(1), pp. 415–432.

[15]. Devi, T., Deepa, N. and Jaisharma, K. (2020) 'Client-Controlled HECC-as-a-Service (HaaS)', in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, pp. 312–318.

[16]. Kane, S.P., Phar and BCPS (no date) Sample Size Calculator. Available at: https://clincalc.com/stats/samplesize.aspx (Accessed: 27 March 2021).

[17]. Boateng, E.Y., Otoo, J. and Abaye, D.A. (2020) 'Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review', Journal of Data Analysis and Information Processing, pp. 341–357. doi:10.4236/jdaip.2020.84020.

[18]. Logeshwari, R. and Rama Parvathy, L. (2020) 'Generating logistic chaotic sequence using geometric pattern to decompose and recombine the pixel values', Multimedia tools and applications, 79(31-32), pp. 22375–22388.

[19]. Shabtari, M.M. et al. (2021) 'Analyzing PIMA Indian Diabetes Dataset through Data Mining Tool "RapidMiner"', 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) [Preprint]. doi:10.1109/icacite51222.2021.9404741.

[20]. Waty, H. et al. (2014) 'The Development of Web Based Expert System for Diagnosing Children Diseases Using PHP and MySQL', International Journal of Computer Trends and Technology, pp. 197–202. doi:10.14445/22312803/ijctt-v10p134.

[21]. Ji, Y., Yu, S. and Zhang, Y. (2011) 'A novel Naive Bayes model: Packaged Hidden Naive Bayes', 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference [Preprint]. doi:10.1109/itaic.2011.6030379.

[22]. Pasha, S.M. (2020) 'Diabetes and Heart Disease Prediction Using Machine Learning Algorithms', International Journal of Emerging Trends in Engineering Research, pp. 3247–3252. doi:10.30534/ijeter/2020/60872020.

[23]. Gao, Y. and Chai, F. (2021) 'Risk of non-vertebral fractures in men with type 2 diabetes: A systematic review and meta-analysis of cohort studies', Experimental gerontology, p. 111378.

[24]. Jayaraj, G. et al. (2015) 'Stromal myofibroblasts in oral squamous cell carcinoma and potentially malignant disorders', Indian journal of cancer, 52(1), pp. 87–92.

[25]. Khalid, W. et al. (2016) 'Role of endothelin-1 in periodontal diseases: A structured review', Indian journal of dental research: official publication of Indian Society for Dental Research, 27(3), pp. 323–333.

[26]. Khalid, W. et al. (2017) 'Comparison of Serum Levels of Endothelin-1 in Chronic Periodontitis Patients Before and After Treatment', Journal of clinical and diagnostic research: JCDR, 11(4), pp. ZC78–ZC81.

[27]. Marling, C. and Bunescu, R. (2020) 'The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020', CEUR workshop proceedings, 2675, pp. 71–74.

[28]. Verma, T. N., Rajak, U., Dasore, A., Afzal, A., Manokar, A. M., Aabid, A., & Baig, M. (2021). Experimental and empirical investigation of a CI engine fuelled with blends of diesel and roselle biodiesel. *Scientific Reports*, *11*(1), 18865.

[29]. Dhanalakshmi, R. et al. (2021) ‘Association rule generation and classification with fuzzy influence rule based on information mass value’, Journal of ambient intelligence and humanized computing, 12(6), pp. 6613–6620.

[30]. Palaniappan, S. et al. (2020) ‘Secure user authentication using honeywords’, in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, pp. 896–903.