IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 15, Issue 6, June 2025, ||Series -1|| PP 38-44

Machine Learning Techniques for forecasting cardiovascular or heart disease

Laveti. Rama Mohan
Associate Professor, Sri venkateswara Engineering college, Etcherla, Srukakulam AP.
Received 15 June 2025; Accepted 29 June 2025

Abstract— *Heart is the most crucial & critical organ of the human body. Life is completely dependent on the efficient working & functioning of our heart. It is one of the major causes of mortality in today's world. Heart disease remains one of the most serious health issues of our day. It is said to be the primary motive in death globally. Many times it's difficult for medical professionals to expect a heart disease on time. Nowadays, the health sector contains a lot of precious hidden facts & information which could prove to be very helpful in making predictive decisions especially in the field of medicine. Data mining is a method or technique used to analyze vast datasets and then derive significant and useful results with the use of extraordinary AI-based techniques. This article attempts to use three of these AI-based methods namely Decision Tree, Naïve Bayes, & Neural Network for forecasting cardiovascular or heart disease. All of these methods will be evaluated based on different unique & parameters with optimizations for better accuracy. The accuracy of each method will then be compared depending on accuracy based on various parameters. The best & accurate technique is then implemented for predicting whether or not a man or a woman will have coronary heart disease. This technique can be used by medical practitioners for early prediction of the disease so that timely care can be taken by the patient.*

Keywords—*Data Mining, Artificial Intelligence, Heart Disease, Prediction*

I. INTRODUCTION

Cardiovascular disease has become one of the most widespread diseases in the world at present. It is estimated to have caused around 17.9 million deaths in 2017 which constitutes about 15% of all natural deaths [13]. Cardiovascular disease is chronic heart disease and can be detected at the initial stages by measuring the levels of various health parameters like blood pressure, cholesterol level, heart rate, and glucose level [13]. The cardiovascular disease not only affects human health but also the economics and cost of the countries [14]. Nowadays, several data mining algorithms and machine learning algorithms are being developed and researched for predicting the different types of diseases [28]. Similarly, there are many research article which shows that numerous data mining, machine learning, and the hybrid algorithms are being studied, developed and investigated which can help detect the and predict the early stage of heart disease [22-26]. The heart disease diagnosis is the process of detecting or predicting heart disease from a patient's records. Doctors may not able to diagnose a patient properly in a short time, especially when the patients suffer from more than one disease [10]. The authors in [18] have surveyed numerous research papers from different years on the prediction of heart diseases and they concluded that data mining techniques are better at predicting heart diseases.

Classification techniques are used widely in healthcare because of their capabilities of processing very large data sets. The commonly used techniques in healthcare are Naïve Bayesian, support vector machine, nearest neighbor, decision tree, Fuzzy logic, Fuzzy based neural network, Artificial neural network, and genetic algorithms [1].

II. RELATED WORK

Several researchers and authors have studied, experimented with, and analyzed numerous techniques for heart disease predictions which includes the techniques for classification and feature selection.

The authors proposed the hybrid HRFLM approach by combining the characteristics of the Linear Method (LM) and Random Forest (RF). They obtained a prediction accuracy of 88.4% [1].

The authors in one of the research done in 2019, tried to mainly increase the accuracy of prediction by using the various feature selection techniques. Different data mining techniques i.e. Decision Tree, Logistic regression, Logistic regression SVM, Naïve Bayes, and Random forest are applied individually in Rapid miner on a UCI heart disease date set and compared results with the past researches and finally, the results concluded that the Logistic regression which obtained an accuracy of is 84.85% is the best feature selection technique for predicting heart disease[2]

In 2018, the researchers used the Prediction models by using the different combinations of features, and seven classification techniques: k-NN, DT, NB, LR, SVM), NN, and VOTE (a hybrid technique with Naïve Bayes and Logistic Regression). And their experiment results showed that the best-performing data mining technique, the VOTE technique with NB and LR achieved an accuracy of 87.4% in heart disease prediction [3]. The 10-fold cross-validation technique was used to validate the performance of the models[3].

The authors in 2019 developed an automated diagnostic system based on χ^2 statistical model and DNN (χ2-DNN MODEL) for the improved diagnosis of heart disease. Their proposed method targeted the two main problems i.e., the problem of underfitting and overfitting, and proposed a diagnostic system that neither under fits nor overfits the training data and their proposed model gave the testing accuracy of 93.33%[4].

The authors in [5] proposed a hybrid model or system in which the researchers used the decision tree technique, i.e., the C4.5 algorithm, and combined it with ANN and named it as hybrid DT to produce the desired result. When this model was analyzed and compared with the C4.5 algorithm and ANN on the same data set, it proved to be more accurate with an accuracy of 78.14%[5].

In 2019 the researchers implemented a hybrid approach combining various techniques that exploited the Fast Correlation-Based Feature Selection (FCBF) method to filter redundant features to improve the quality of heart disease classification. This method proved to be more than 90% accurate [6].

Few authors [7] used an ensemble of classifiers. The ensemble algorithms bagging, boosting, stacking and majority voting were employed for experiments. The proper selection techniques for feature sets helped to improve the accuracy of the ensemble algorithms. The highest accuracy was obtained with majority voting with the feature set FS2[7].

In 2020 the authors used the seven different intelligent techniques to predict coronary heart disease using the Starlog and Cleveland heart disease dataset and in their comparative study, the deep neural network performed better and obtained an accuracy of 98.15% with the Starlog dataset and in the case of Cleve- land dataset, SVM achieved an accuracy of 97.36%[8].

In 2019 in one of the research for diagnosis of heart disease, the authors used UCI machine learning repository for heart disease dataset and proposed a Multi-Layer PiSigma Neuron Model (MLPSNM) for heart disease diagnosis which was based on PI-Sigma model in which, as per the authors, the architecture and calculation are less complex as compared to other previously pro- posed models. For the learning of the network, the BP algorithm was used with bipolar sigmoid function activation function and PCA and LDA preprocessing methods are used to reduce the dimensionality of the dataset. In the SVM-LDA method, the attributes that are closer to the hyperplane are selected. For validation of the network, the k-fold validation method is used. The network converges after 50 iterations. The proposed model achieves 94.53% classification accuracy for diagnosis of heart disease by using PCA [9].

The authors in [11] compared the use of several supervised machine learning (ML) algorithms for predicting clinical events in terms of their internal validity and accuracy and the results, which were obtained using two statistical software platforms that is R-Studio and Rapid Miner were then compared and showed that the decision tree algorithm gave better results.

The authors in [15] performed the comparative study of heart disease diagnosis system using top ten data mining classification algorithms [27]. The data mining algorithms discussed were C4.5, SVM, Ada Boost, KNN, Naive Bayes, and CART, Random Forest, Bagging Algorithm, Logistic Regression, and Multilayer Perceptron (MLP). From their experimental study in terms of accuracy, the top three algorithms were Random Forest with 78.0%, kNN with 71.6%, and MLP with 63.8% and the top three based on speed were AdaBoost, kNN, and Naive Bayes.

The authors in [16] carried the implementation of prediction algorithm and reached to the conclusion that the accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose[16].

In 2011 the researchers in[19]used the classification algorithm such as RIPPER (Repeated Incremental Pruning to Produce Error Reduction) proposed by William W Cohen, Decision tree, ANN, and Support Vector Machine, and their experimental results showed that Support Vector machine achieved the highest prediction accuracy[19]. The authors, Sellappan & Palaniappan in [20]proposed an advanced and Intelligent coronary heart disorder prediction machine (IHDPS) using three data mining techniques (naïve Bayes, decision tree, neural network).

Authors K. Srinivas, B. Kavita Rani, and A. Govardhan presented the use of numerous data mining techniques to predict a heart attack. They used methods such as Decision Tree, Naive Bayes, and ANN [21]. Data mining tools, such as TANAGRA, were used in statistical learning algorithms.

III. PROPOSED METHODOLOGY

Based on the conclusion from our literature review we concluded that the three below mentioned techniques are better & efficient in classifying and predicting in terms of accuracy. Therefore we experimented with these three techniques that are;

1. Neural Network
2. Decision Tree
3. Naïve Bayes.

IV. EXPERIMENTATION AND PERFORMANCE ANALYSIS

a) DATASET

We have used the dataset from the UCI repository from this website link https://archive.ics.uci.edu/ml/datasets/Heart+Disease. We also consulted the doctor nearby who helped us to add more data to our database.

Our datasets consisted of 14 attributes with 668 records, details of which are given in Table 1, below.

Sr. no	Attribute		Description	Values
1	Age		Age in years	Continuous
2	Sex		Male or female	1 = male 0 = female
3	Cp		Chest pain type	1 = typical type 1 2 = typical type agina 3 = non-agina pain 4=asymptomatic
4	thestbps		Resting blood pressure	Continuous value in mm hg
5	chol		Serum cholesterol	Continuous value in mm/dl
6	Restecg		Resting electrographic results	0 = normal 1 = having_ ST_T wave abnormal 2 = left ventricular hypertrophy
7	FBS		Fasting blood sugar	1 ≥ 120 mg/dl 0 ≤ 120 mg/dl
8	thalach		Maximum heart rate achieved	Continuous value
9	exang		Exercise-induced agina	0= no 1 = yes
10	oldpeak		ST depression induced by exercise relative to rest	Continuous value
11	slope		The slope of the peak exercise ST segment	1 = unsloping 2 = flat 3 = downsloping
12	Ca		Number of major vessels colored by floursopy	0-3 value
13	thal		Defect type	3 = normal 6 = fixed 7 = reversible defect

Table 1. Attributes of Heart Disease Dataset

After manipulating the dataset i.e. increasing and decreasing the training and testing data, we got the following results for the three data mining techniques for the prediction of heart disease shown in various tables and graphs below.

Table 2. Decision Tree Data (Snapshot with few

b) DECISION TREE

Below table 2 and Figure 1, shows the Decision Tree tested for a different number of testing data, how the accuracy can be improved by removing some of the attributes and testing again.

configuration changes)

ATTRIBUTES	ACCURACY			Training Time		
Number of testing data	50	100	200	50	100	200
With all attributes	98.38	92.42%	84.37	0.20	0.51	0.20
Without smoke	98.38	94.36%	85.01	0.20	0.10	0.09
Without thal	97.41	93.65%	83.51	0.30	0.14	0.27
Without ca	98.1	94.18%	84.37	0.19	0.14	0.19
Without slope	98.54	93.29%	86.72	0.15	0.08	0.27
Without oldpeak	97.57	93.65%	83.73	0.11	0.33	0.19
Without exang	98.70	93.47%	84.15	0.14	0.61	0.25
Without thalach	98.38	94.53%	84.80	0.19	0.55	0.55
Without restecg	98.38	92.95%	84.15	0.30	0.50	0.16
Without fbs	98.38	94.53%	85.01	0.56	0.19	0.28
Without chol	97.56	93.36%	85.65	0.11	0.14	0.21
Without restbp	97.73	94.36%	86.72	0.19	0.31	0.16
Without cp	96.27	94.36%	83.08	0.20	0.17	0.20
Without sex	98.54	92.59%	86.29	0.14	0.88	0.33
Without age	96.76	94.18%	85.44	0.13	0.33	0.23

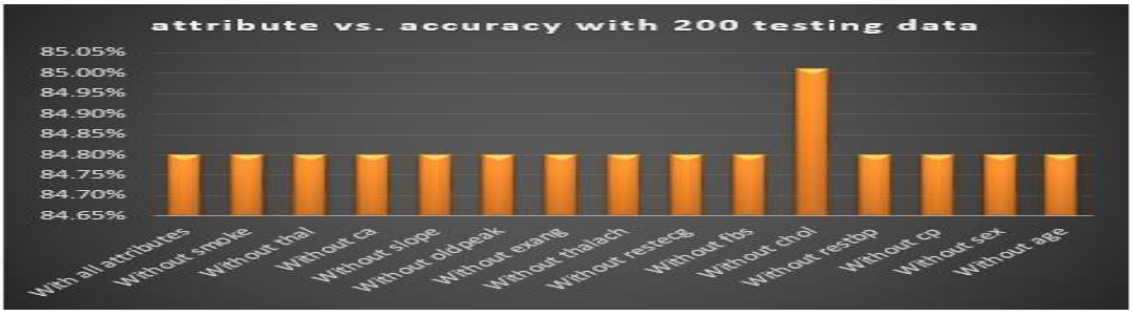
Figure 1. Attribute v/s Accuracy graph for table2

c) NAÏVE BAYES

The below table3 shows the Naïve Bayes tested for different numbers of testing data, how the accuracy can be improved by removing some of the attributes and testing again.

Table 3: Naïve Bayes Data (Snapshot with few

ATTRIBUTES	ACCURACY			Training Time		
Number of testing data	50	100	200	50	100	200
With all attributes	81.36%	82.72%	84.80%	0.17	0.57	0.33
Without smoke	81.36%	82.89%	84.80%	0.85	1.28	0.34
Without thal	81.36%	82.71%	84.80%	0.41	0.89	0.24
Without ca	81.20%	82.54%	84.80%	0.53	0.42	0.21
Without slope	81.20%	82.72%	84.80%	0.52	1.99	0.50
Without oldpeak	81.20%	82.72%	84.80%	0.41	0.38	0.70
Without exang	81.36%	82.54%	84.80%	0.74	0.25	0.47
Without thalach	81.20%	82.54%	84.80%	0.72	0.53	0.39
Without restecg	81.36%	82.54%	84.80%	0.58	0.59	0.33
Without fbs	81.20%	82.56%	84.80%	0.70	0.20	0.50
Without chol	81.36%	83.07%	85.01%	0.74	0.53	0.36
Without restbp	81.36%	82.54%	84.80%	0.67	0.48	0.44
Without cp	81.20%	82.72%	84.80%	0.61	0.38	0.55
Without sex	81.36%	82.54%	84.80%	0.53	0.58	0.45
Without age	81.20%	82.54%	84.80%	0.53	0.41	0.41



d) NEURAL NETWORK

Below table 4 and Figure 3 show the accuracy obtained for neural networks tested on different hidden layers, changing the number of epochs, increasing and decreasing learning rate and folds, and the activation functions. For improving the accuracy, we removed some of the attributes

Table 4: Neural Network Data (Snapshot with few

HIDDEN LAYERS	5		5		2		2		2		5		4	
EPOCHS	500		500		500		500		500		500		500	
LEARNING RATE	0.3		0.3		0.3		0.3		0.3		0.3		0.3	
FOLDS	10		3		8		6		5		8		8	
ACTIVATION FUNCTION	SIGMOID		RELU		SIGMOID		SIGMOID		SIGMOID		SIGMOID		SIGMOID	
	ACCURACY	TIME	ACCURACY	TIME	ACCURACY	TIME	ACCURACY	TIME	ACCURACY	TIME	ACCURACY	TIME	ACCURACY	TIME
With all attributes	74.70%	417.4	81.85%	81.7	78.61%	238.088	79.43%	55.972	78.50%	136.277	76.36%	327.01	76.39%	267.42
Without smoke	76.67%	828.534	81.85%	75.25	77.86%	274.637	79.73%	95.656	78.20%	133.671	74.40%	314	76.51%	271.4
Without thal	74.53%	450.008	81.85%	75.02	80.27%	272.604	79.13%	68.746	79.55%	47.502	75.30%	308.38	76.05%	302.04
Without ca	78.03%	778.8	81.85%	74.9	80.72%	245.195	80.03%	151.588	80.75%	146.27	76.05%	268.5	78.21%	277.31
Without slope	74.09%	400.655	18.17%	75.25	78.82%	218.972	79.13%	93.028	78.65%	90.442	77.41%	270.36	77.71%	274.05
Without oldpeak	74.85%	401.925	18.17%	74.72	79.01%	267.17	78.88%	89.761	78.95%	63.478	76.81%	271.13	75.15%	281.69
Without exang	75.91%	347.75	81.85%	76.56	79.22%	182.557	79.58%	170.945	78.80%	143.908	78.01%	277.94	77.71%	280.01
Without thalach	73.32%	345.48	81.85%	75.17	78.46%	232.427	76.88%	169.04	77.14%	137.324	74.40%	270.86	75.00%	258.2
Without restecg	76.67%	347.69	81.85%	74.78	81.02%	82.102	79.13%	96.13	79.40%	131.615	78.31%	268.02	78.16%	261.6
Without fbs	75.30%	345.31	18.17%	74.92	78.61%	74.162	78.23%	118.912	77.25%	50.4	76.81%	273.44	76.66%	226.97
Without chol	76.67%	341.59	18.17%	74.98	78.77%	103.116	77.48%	125.428	77.90%	96.166	75.60%	266.41	76.36%	235.7
Without restbp	76.97%	344.8	18.17%	81.47	79.97%	91.072	78.88%	92.525	78.65%	86.145	76.05%	264.95	77.56%	226.38
Without cp	76.36%	344.13	81.85%	76.42	79.52%	78.015	78.68%	118.454	78.95%	81.503	74.70%	271.02	77.11%	228.83
Without sex	75.15%	345.39	81.85%	76.44	79.07%	80.262	81.08%	150.983	78.65%	55.645	76.05%	267.27	76.51%	227.79
Without age	75.30%	344.13	81.85%	76.65	79.97%	77.703	79.73%	194.882	78.65%	123.84	74.40%	266.54	76.05%	227.44

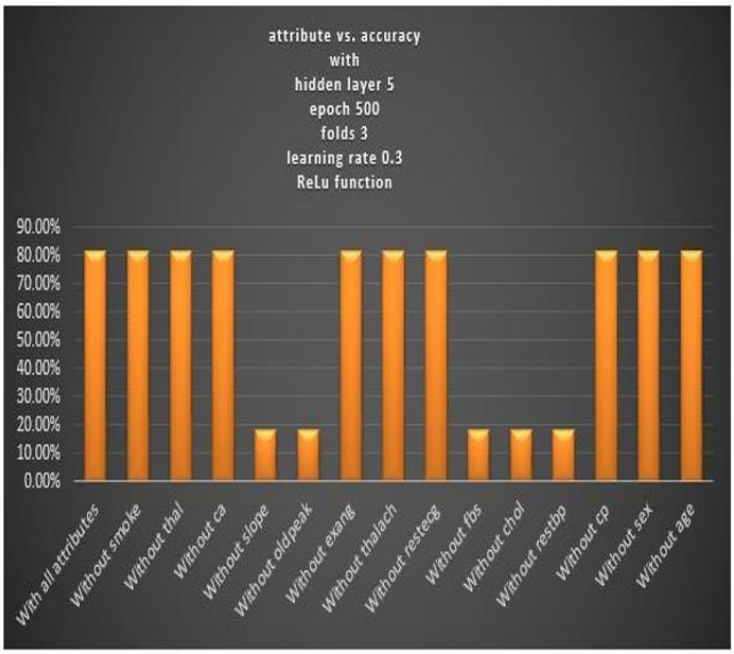


Figure 3. Attribute v/s Accuracy graph for table 4

e) MODEL COMPARISON

Figure 5 shows the graph for the accuracy of the three data mining techniques.

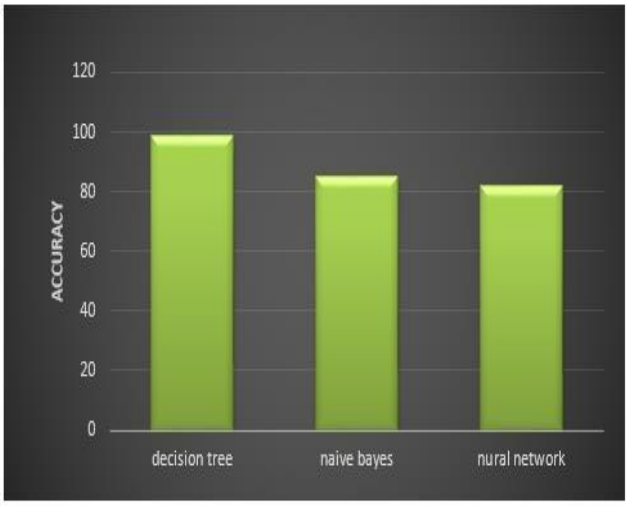


Fig 5: Accuracy levels for three data mining techniques

V. CONCLUSION AND FUTURE SCOPE

From the above graphs obtained from our implementation, we can conclude that when we increase hidden layers, the result becomes less accurate and it also consumes more time i.e. not efficient. Also After decreasing the learning rate the accuracy decreased. In the neural network, we got the highest accuracy i.e. 81.08% when we used a smaller number of hidden layers with increased learning rate and increased training dataset.

When we changed the attributes, the result was also changing. Removal of the chest pain and cholesterol attributes decreased the accuracy of a decision tree since both are important attributes for heart disease prediction. But after removing the sex attribute the accuracy remained unchanged, which led to our conclusion that this attribute doesn't play an important role in disease prediction.

We also tried to check the accuracy of Naïve Bayes by removing some attributes but the results didn't change much because the Naïve Bayes algorithm is independent of other attributes.

In a neural network, for finding better accuracy we tried with different hidden layers, learning rates, and changing Attributes. When we increased hidden layers, it gave better accuracy but its computation time increased

which was not good for prediction, but when we reduced the number of hidden layers it gave us better results with much shorter calculation time which was reliable. After analyzing the above graphs we concluded that the decision tree was giving more accurate results with 98.54% as compared to other methods which we're giving 85.01% (Naïve Bayes) and 81.83% (neural network). As we can see from the below graph.

We can make this system more efficient & reliable by using a more number of training datasets and evaluating the datasets. We can also try to increase the number of features such as Junk food, exercise, and tobacco to be more precise.

Also, there is a scope to improvise this system by integrating these approaches and forming a hybrid model that can deliver better outcomes than individual methods.

REFERENCES

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

[2] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (pp. 619-623). IEEE.

[3] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82-93.

[4] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}} $ Statistical Model and Optimally Configured Deep Neural Network. *IEEE Access*, *7*, 34938-34945.

[5] Maji, S., & Arora, S. (2019). Decision tree algorithms for prediction of heart disease. In *Information and Communication Technology for Competitive Strategies* (pp. 447-454). Springer, Singapore.

[6] Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *Int. J. Intell. Eng. Syst.*, *12*(1), 242-252.

[7] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

[8] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, 1-20.

[9] Burse, K., Kirar, V. P. S., Burse, A., & Burse, R. (2019). Various preprocessing methods for neural network-based heart disease prediction. In *Smart innovations in communication and computational sciences* (pp. 55-65). Springer, Singapore

[10] Tarawneh, M., & Embarak, O. (2019, February). Hybrid approach for heart disease prediction using data mining techniques. In *International Conference on Emerging Internetworking, Data & Web Technologies* (pp. 447-454). Springer, Cham.

[11] Beunza, J. J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., ... & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of biomedical informatics*, *97*, 103257.

[12] Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019, July). Prediction of coronary heart disease using machine learning: an experimental analysis. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies* (pp. 5156).

Authors

Laveti. Rama Mohan, MCA, MTech, Working as Associate Professor.. In Sri venkateswara Engineering college, Etcherla, Srukakulam, Guest Faculty in (college of Engineerng) Dr. BR. Ambedkar university, Etcherla, Srikakulam, having 12+ years.