S.N.BOSE SUMMER INTERNSHIP REPORT

National Institute of Technology, Silchar

Heart Disease Prediction Using Machine Learning

Under the supervision of Dr. Malaya Dutta Borah
Computer Science and Engineering Department, NIT Silchar

By

Hrishi Raj Deka(2014004)
Mrinal Kalita(2014000)
Mainur Islam Ahmed(2013009),
Manash Pratim Pathak(2013036)

ABSTRACT:

In today's fast-paced life, many people are neglecting their health while focusing on their daily work and responsibilities. This has led to a significant increase in the number of individuals falling ill. Among various diseases, heart disease has emerged as a major concern, accounting for approximately 31% of global deaths, according to data from the World Health Organization (WHO). Predicting the occurrence of heart disease has therefore become crucial for the medical field. However, the volume of data received by hospitals and medical institutions is often overwhelming, making analysis challenging. Employing machine learning techniques can greatly enhance the efficiency of healthcare professionals in predicting and managing heart disease.

This study addresses the topic of heart disease, its risk factors, and the utilization of machine learning techniques. By leveraging these techniques, the study aims to predict the occurrence of heart disease and provide a comparative analysis of various machine learning algorithms employed in the prediction experiment. The primary objective of this research is to accurately predict heart disease using machine learning techniques and analyse their effectiveness

LITERATURE REVIEW:

SL No	Author, Year	Methods Used	Datasets Used	Findings
1	Senthilkumar Mohan , Chandrasegar Thirumalai And Gautam Srivastava 2019 [6]	Language model,K-Nearest Neighbour, Support vector machine,Naïve Bayes,Decision Tree,Random forest, Neural Networks	UCI machine learning repository	From the Proposed Models SVM kernel has the highest accuracy of 72.6%
2	Rahul Katarya,Sunit Kumar Meena 2020 [1]	Logistic Regression,K-Nearest Neighbour,Support vector machine, Naïve Bayes,Decision Trees,Random Forest,Decision Trees,Artificial Neural Network,Deep Neural Network, Multi-Layer Perceptron (MLP)	UCI dataset	From the proposed Models Logistic Regression has the Highest accuracy of 93.40% and MLP having the lowest accuracy of 75.42.

3	Ashok Kumar Dwivedi 2018 [3]	Logistic Regression, K-Nearest Neighbour,Support vector machine, Naïve Bayes,Classification Trees, Artificial Neural Network	UCI heart disease dataset	From the proposed Models ANN has the Highest classification Accuracy of 0.84 withF1 score of 0.86
4	Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, Harshali Rambade 2016 [2]	Support Vector Machine (SVM),Logistic Regression,Naïve Bayes Algorithm.	UCI dataset	The algorithms were judged based on their accuracy and it is observed that the SVM is the most accurate out of the three with 64.4% efficiency.
5	R.Jane Preetha Princy,Saravanan Parthasarathy,P. Subha Hency Jose,P. Subha Hency Jose,Selvaprabu Jeganathan,2020[5]	Logistic Regression, K-Nearest Neighbour, Support vector machine, Naïve Bayes, Decision Tree, Random forest	Cardiovascular Disease datasets Kaggle	The decision tree algorithm outperformed others by producing 73% accuracy. The logistic regression and SVM delivered 72% and Random forest made it with 71%. The KNN and Naive-Bayes algorithms delivered 66% and 60% of accuracy respectively.
6	Chiradeep Gupta,Athina Saha, NV Subba Ready, U Dinesh Acharya 2013 [7]	Logistic Regression, K-Nearest Neighbour, Support vector machine, Naïve Bayes, Decision Tree, Random forest	UCI datasets	Conclusion and Future Scope The results of the proposed work depict that Logistic Regression is better than the other supervised classifiers in terms of the discussed performance metrics – accuracy, precision, sensitivity (or recall), specificity and F1 score. The model gives the results with the highest accuracy of 92.30%
7	Anupama Yadav, Levish Gediya, Adnanuddin Kazi 2021[4]	K-Nearest Neighbour, Support vector machine, Decision Tree, Random forest.	UCI datasets	This experiment showed that Nave Bayes is more accurate than ID3 when it comes to predicting heart disease, even if the input data is clean and well maintained. A combination of algorithms like Nave Bayes and K-means can be used to get accuracy. Machine learning is proving to be valuable in predicting heart disease, and new methods may soon come to make it more helpful in healthcare. The algorithms used in this experiment have performed well using the available attributes.

DESCRIPTION OF DATA-SET AND FEATURES:

REST ECG:

Rest ECG (Electrocardiogram) is a non-invasive test that measures the electrical activity of the heart at rest. It helps detect heart disease by identifying abnormal heart rhythms, signs of ischemia (lack of blood flow to the heart), heart muscle damage, and structural abnormalities, aiding in the diagnosis and assessment of various cardiac conditions.

ST abnormality refers to deviations from the normal pattern of the ST segment on an electrocardiogram (ECG), indicating potential myocardial injury or ischemia.

LV hypertrophy (left ventricular hypertrophy) is the thickening of the muscular wall of the left ventricle of the heart, often resulting from conditions such as hypertension or valve diseases, leading to an enlarged and overworked left ventricle.

CHEST PAIN:

Chest pain serves as a warning sign of possible heart disease and prompts further evaluation. Characteristics of the chest pain, such as location and triggers, provide clues about the underlying cause. Diagnostic tests are used to assess heart function and blood flow when cardiac-related chest pain is suspected.

Typical angina: Chest pain or discomfort with a predictable pattern, triggered by physical exertion or stress, relieved by rest or nitro-glycerine.

Atypical angina: Chest pain or discomfort that does not fit the typical pattern, with different symptoms, triggers, or medication response.

Angina not: Chest pain not related to reduced blood flow to the heart, caused by other factors like musculoskeletal or gastrointestinal issues.

Asymptomatic: No symptoms despite underlying heart disease or reduced blood flow, can have significant blockages without noticeable chest pain. Diagnosed through tests.

FBS:

FBS (Fasting Blood Sugar) is primarily used to assess blood glucose levels and diagnose diabetes. While FBS itself is not a direct indicator of heart disease, it can indirectly contribute to its detection. Elevated fasting blood sugar levels may indicate underlying conditions such as diabetes or prediabetes, which are known risk factors for cardiovascular disease. By identifying and managing diabetes, the risk of developing heart disease can be reduced. However, other specific tests, such as lipid profile, ECG, stress tests, and imaging studies, are typically used to directly detect and diagnose heart disease.

EXANG(EXERCISE INDUCED ANGINA):

Exercise-induced angina helps detect heart disease by causing chest pain or discomfort during physical activity, indicating reduced blood flow to the heart. It is often assessed through exercise stress tests to monitor the onset of angina

SERUM CHOL:

Serum cholesterol levels are essential in detecting heart disease. High levels of LDL cholesterol, known as "bad" cholesterol, can contribute to the development of atherosclerosis and increase the risk of heart disease. Measuring serum cholesterol through a lipid profile test helps identify individuals with elevated LDL cholesterol, allowing for appropriate interventions to reduce the risk of heart disease, such as lifestyle changes and medication. Monitoring cholesterol levels helps in assessing and managing the overall cardiovascular health of an individual.

OLD PEAK:

ST depression during exercise compared to rest is indicative of reduced blood flow to the heart muscle, suggesting underlying coronary artery disease. This finding helps in detecting heart disease by identifying individuals at risk. Evaluating ST segment changes during exercise stress testing provides valuable information for further diagnostic testing and treatment decisions.

ST SEGMENT:

The pattern of ST segment changes (upsloping, down sloping, or flat) during exercise stress testing can provide valuable information in the detection of heart disease.

1. Upsloping ST segment: This pattern is considered less specific for heart disease and may be seen in individuals without significant coronary artery disease. However, when accompanied by other symptoms or risk factors, it can still be an indication of possible heart disease and may warrant further evaluation.
2. Down sloping ST segment: This pattern is more concerning and associated with a higher likelihood of coronary artery disease. It suggests an inadequate blood supply to the heart during exercise and can be an important sign of underlying heart disease.
3. Flat ST segment: A flat ST segment during exercise stress testing is generally considered a normal finding and not suggestive of heart disease. However, it is important to consider other factors such as symptoms and risk factors to make a comprehensive assessment.

CA:

The number of major vessels (0-3) coloured by fluoroscopy, typically determined through procedures like coronary angiography, is crucial in detecting and assessing heart disease.

1. 0 vessels coloured: This indicates no visible blockages or significant narrowing in the major coronary arteries. It suggests a lower likelihood of significant heart disease.
2. 1 or 2 vessels coloured: This suggests the presence of blockages or narrowing in one or two major coronary arteries. It indicates a moderate risk of heart disease and may warrant further evaluation or intervention.

3. 3 vessels coloured: This indicates blockages or narrowing in all three major coronary arteries. It suggests a higher risk of severe heart disease and may require immediate medical attention or interventions such as angioplasty or coronary artery bypass grafting.

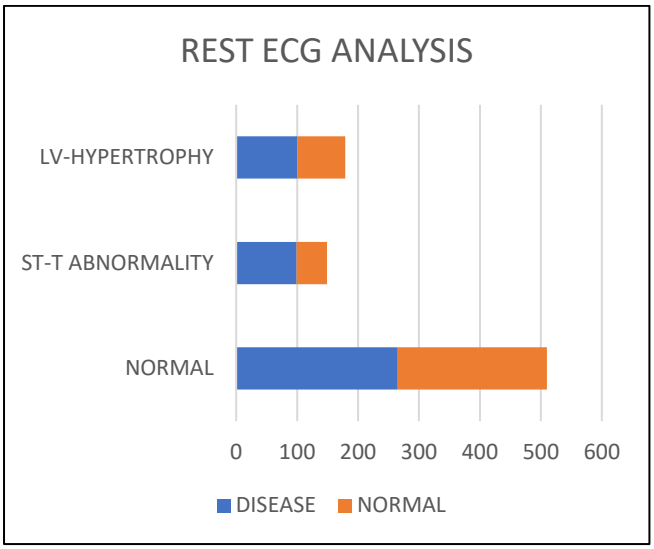
THALASEMIA:

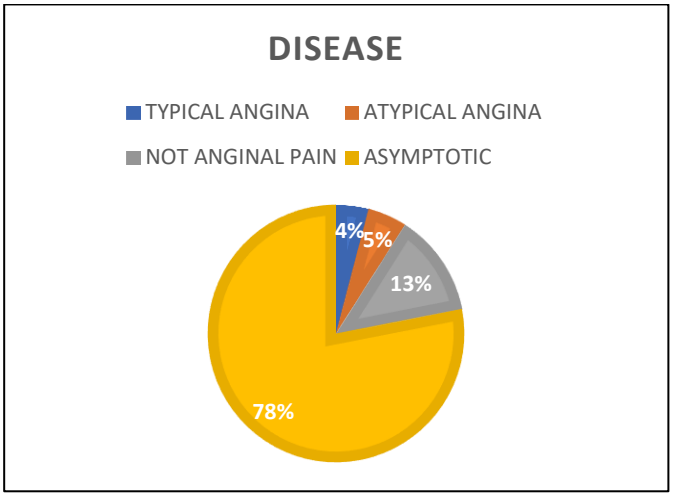
Thalassemia, a genetic blood disorder, can contribute to heart disease development. Chronic anaemia and increased cardiac workload are associated with thalassemia, leading to potential heart complications. Regular monitoring using tests like echocardiography and ECG helps detect and manage heart-related issues in individuals with thalassemia, improving long-term outcomes.

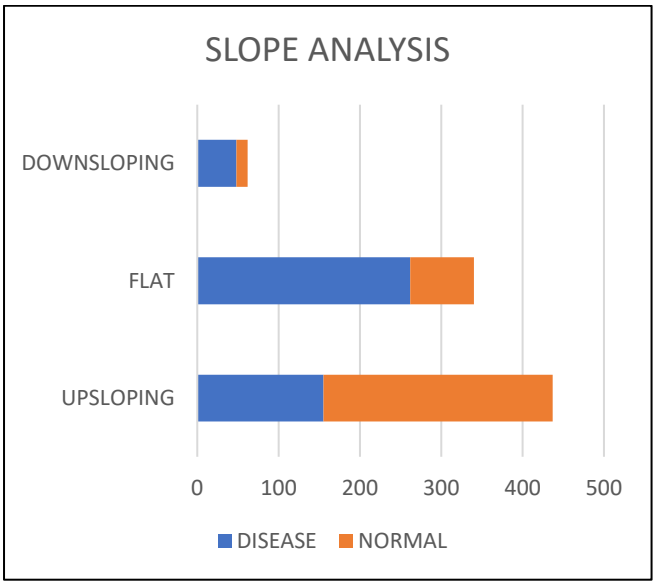
REST BPS:

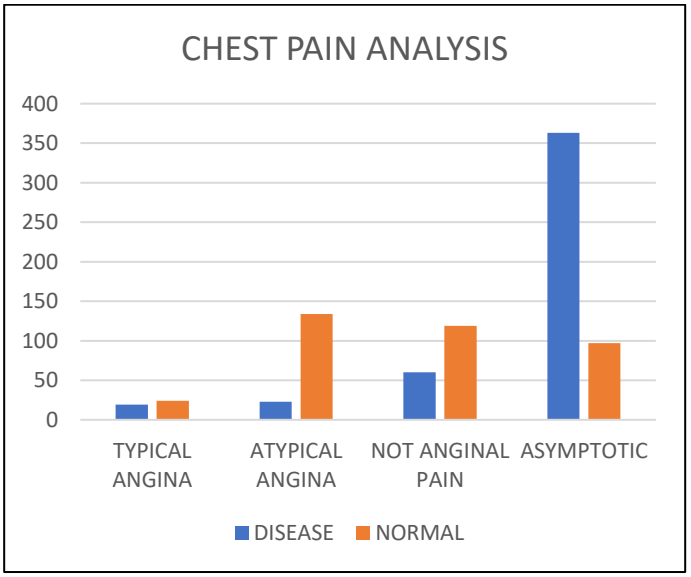
Resting blood pressure measurements are crucial for detecting heart disease. High blood pressure (hypertension) is a major risk factor for cardiovascular conditions. Elevated resting blood pressure indicates increased strain on the heart and arteries. Regular monitoring helps identify individuals with hypertension, enabling interventions like lifestyle changes and medication to manage blood pressure and reduce the risk of heart disease. Effective blood pressure management improves overall cardiovascular health.

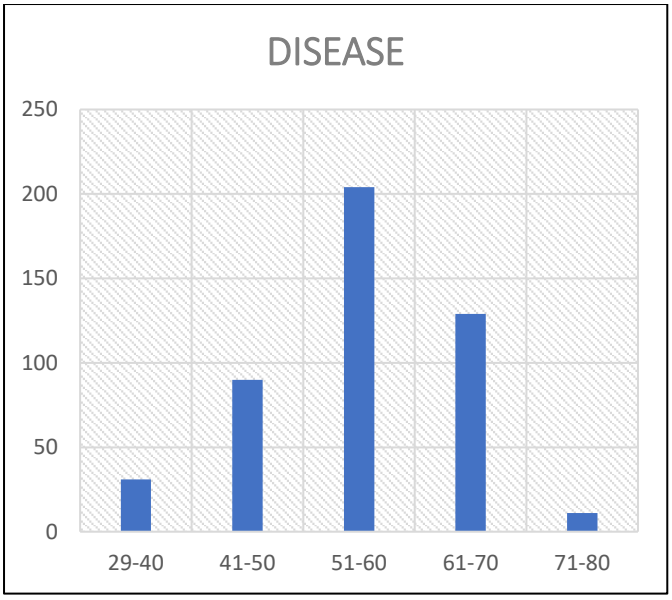
ANALYSIS OF DATASETS:

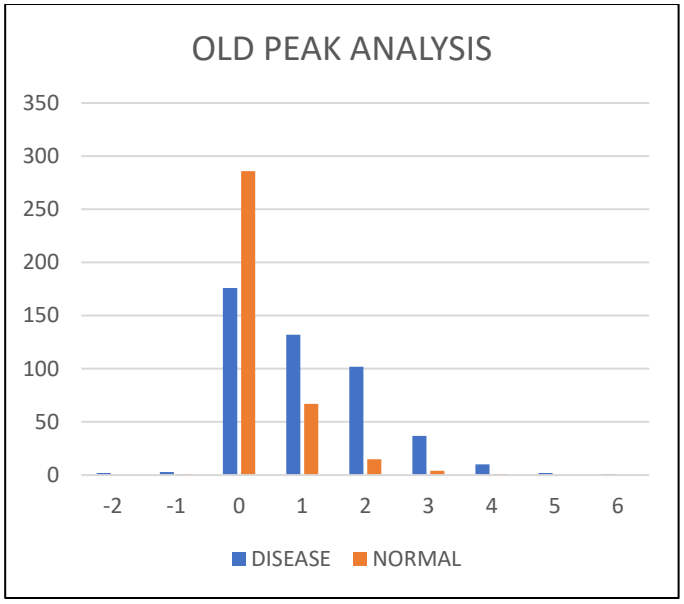


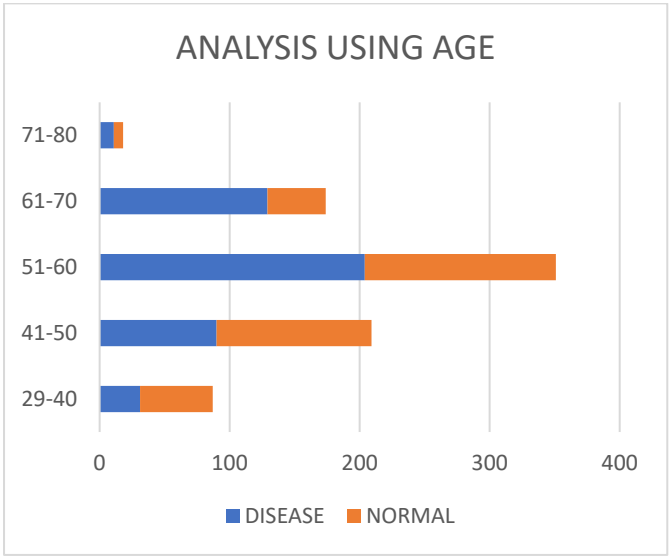


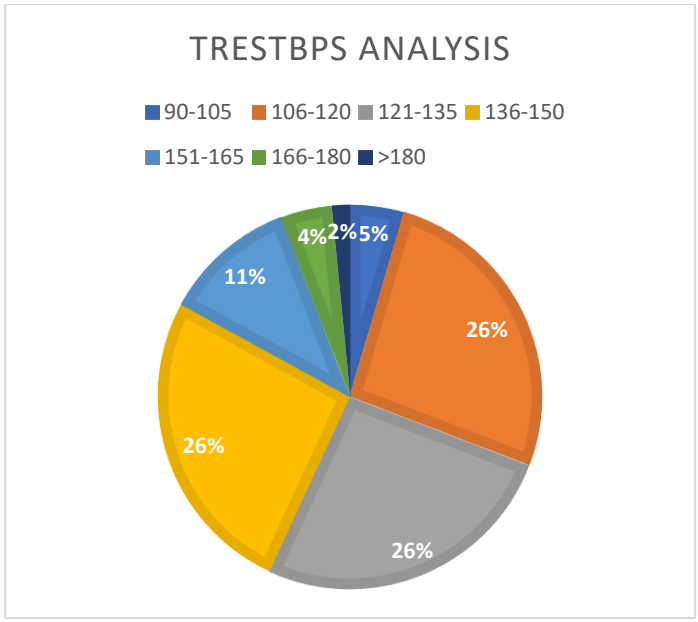


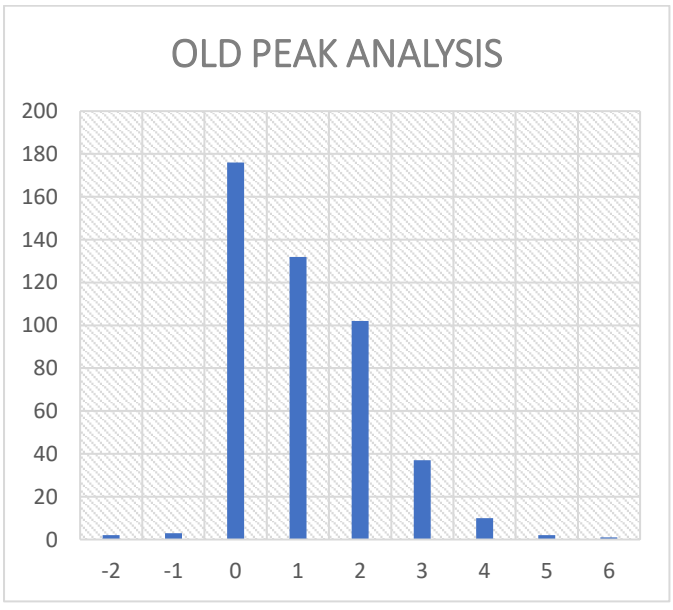


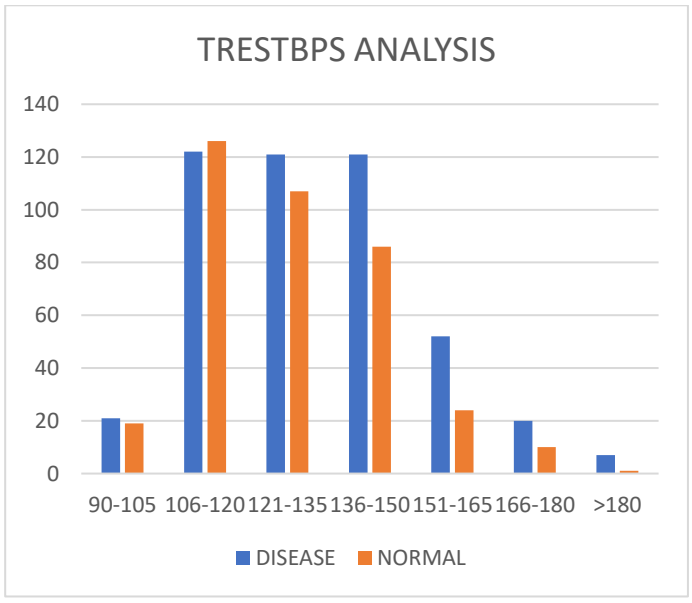

DISEASE
250
200
150
100
50
0
29-40
41-50
51-60
61-70
71-80

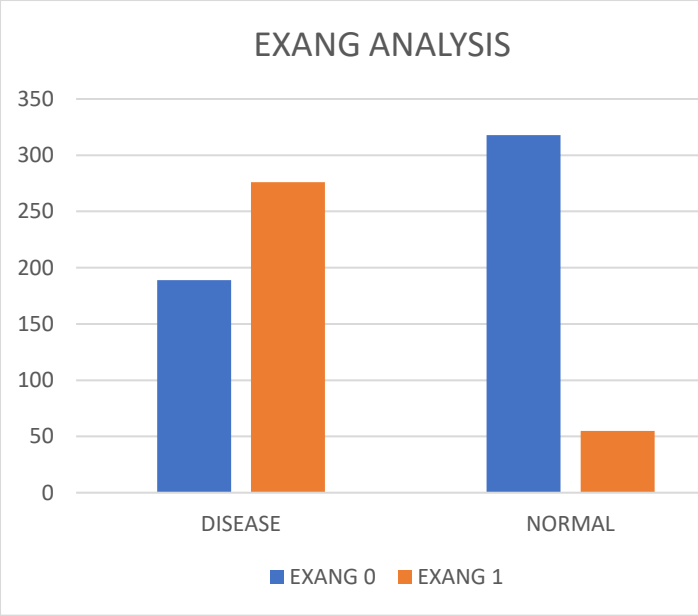

OLD PEAK ANALYSIS
350
300
250
200
150
100
50
0
-2
-1
0
1
2
3
4
5
6
DISEASE
NORMAL


ANALYSIS USING AGE
71-80
61-70
51-60
41-50
29-40
0
100
200
300
400
DISEASE
NORMAL

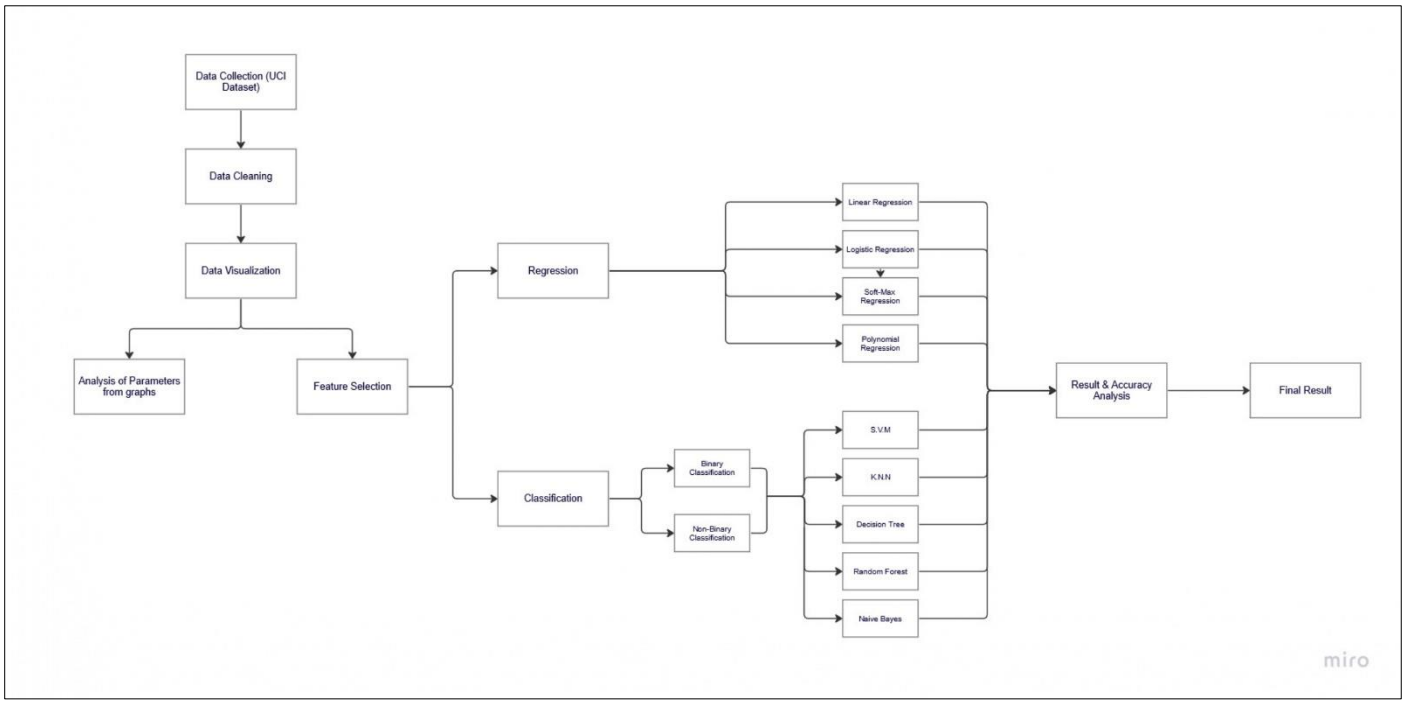

TRESTBPS ANALYSIS
90-105
106-120
121-135
136-150
151-165
166-180
>180
11%
4%
2%
5%
26%
26%
26%


OLD PEAK ANALYSIS
200
180
160
140
120
100
80
60
40
20
0
-2
-1
0
1
2
3
4
5
6


TRESTBPS ANALYSIS
140
120
100
80
60
40
20
0
90-105
106-120
121-135
136-150
151-165
166-180
>180
DISEASE
NORMAL



ARCHITECHTURE:



MACHINE LEARNING MODEL DESCRIPTION:

The approaches utilised for the experiment in this study have been covered in this section. We have described how we have taken into account the significant risk factors for the experiment and the methods we employed for heart disease prediction. We have explained the algorithms we have

used for comparison and prediction of heart disease.

1. K-Nearest Neighbours (KNN):

K-Nearest Neighbours (KNN) is a non-parametric algorithm used for classification and regression tasks. It assigns a class or predicts a value for a data point based on the classes or values of its k nearest neighbours in the feature space. KNN operates on the assumption that similar instances are likely to share the same class or have similar values. It uses distance metrics, such as Euclidean distance, to determine the proximity between data points. KNN is a lazy algorithm as it does not explicitly build a model during training but instead stores the training instances to make predictions. It is versatile and can handle multi-class classification and regression problems.

2. Linear Regression:

Linear regression is a linear model used for regression tasks. It assumes a linear relationship between the input features and the target variable. The algorithm estimates the coefficients of a linear equation by minimizing the sum of squared errors between the predicted and actual values. This is typically achieved using optimization techniques like ordinary least squares. Linear regression is parametric, as it assumes a specific functional form, and provides interpretable coefficients that represent the contribution of each feature. It can handle continuous target variables and capture linear dependencies between features and the target variable.

3. Logistic Regression:

Logistic regression is a widely used algorithm for binary classification tasks. It models the probability of the target variable belonging to a particular class using a logistic function. Logistic regression estimates the coefficients of a linear equation through maximum likelihood estimation. The algorithm applies the logistic function to the linear equation to obtain class probabilities. A threshold is then applied to assign the predicted class. Logistic regression is interpretable and can provide insights into the influence of features on the probability of a binary outcome. It can handle both categorical and continuous input features.

4. Polynomial Regression:

Polynomial regression extends linear regression by including polynomial terms of the input features. It allows capturing non-linear relationships between the features and the target variable. The algorithm fits a polynomial curve to the data by estimating the coefficients of the polynomial equation. The degree of the polynomial determines the flexibility of the model. Higher degrees can better fit complex patterns but may lead to overfitting. Polynomial regression is useful when linear models are insufficient to capture the underlying relationships, and it can handle continuous target variables.

5. SoftMax Regression:

SoftMax regression, also known as multiclass logistic regression, is an extension of logistic regression for multiclass classification problems. It estimates the probabilities of each class using the SoftMax function, which normalizes the outputs. SoftMax regression uses a linear equation for each class and applies the SoftMax function to obtain class probabilities that sum to one. The predicted class corresponds to the class with the highest probability. SoftMax regression is widely used in multiclass classification tasks and can handle multiple classes effectively.

6. Support Vector Machines (SVM):

Support Vector Machines (SVM) is a powerful algorithm used for both classification and regression tasks. SVM finds an optimal hyperplane in a high-dimensional feature space that maximally separates data points of different classes or predicts continuous values. SVM can handle linearly separable and non-linearly separable data by using kernel functions to transform the data into higher-dimensional spaces. It selects the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points. SVM is effective in handling high-dimensional data, can mitigate the risk of overfitting, and provides flexibility in choosing kernel functions to capture complex relationships.

7. Decision Trees:

Decision trees are tree-like models that make decisions based on the values of input features. They recursively split the data based on features to create a hierarchy of decisions. Decision trees can handle both classification and regression tasks. At each split, the algorithm selects the feature that best separates the data according to a chosen criterion, such as Gini impurity or information gain. Decision trees are easy to interpret and visualize, as the splits represent decision rules. However, they can be prone to overfitting, especially when the trees become deep, and they may struggle to capture complex relationships in the data.

8. Random Forest:

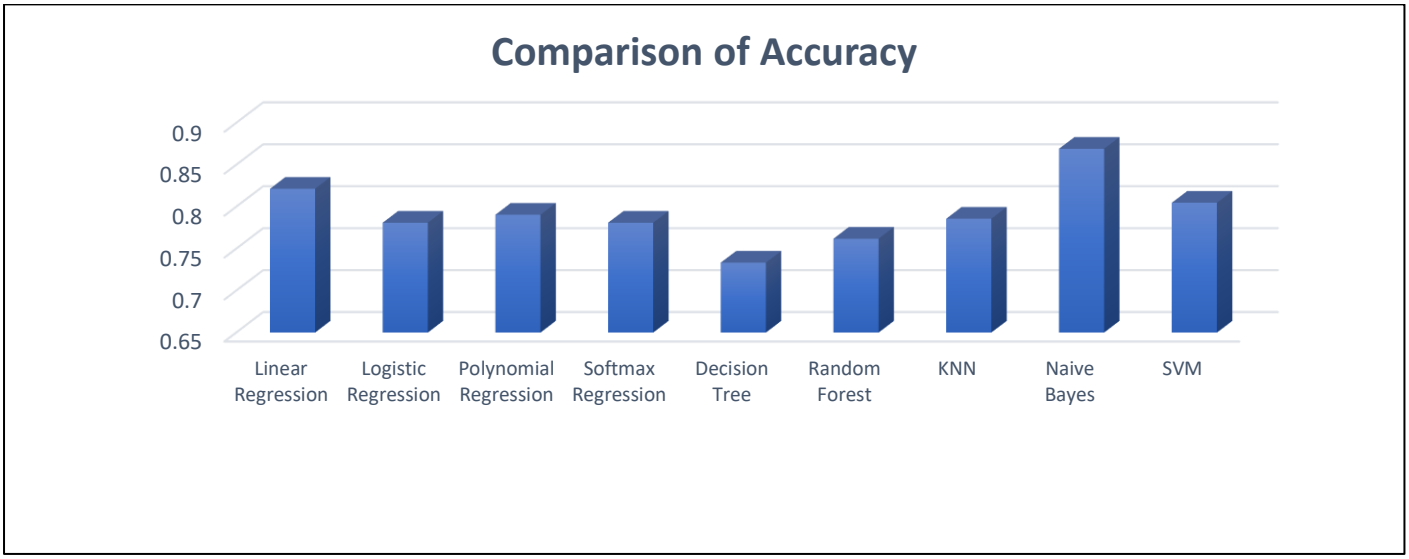
Random Forest is an ensemble algorithm that combines multiple decision trees to make predictions. It constructs a forest of decision trees, where each tree is trained on a random subset of the data and features. The algorithm aggregates the predictions of individual trees, typically through voting or averaging, to make the final prediction. Random Forest is robust to noise and outliers, handles high-dimensional data well, and can mitigate the overfitting issues of decision trees. It provides improved accuracy and generalization by reducing variance and maintaining the interpretability of decision trees.

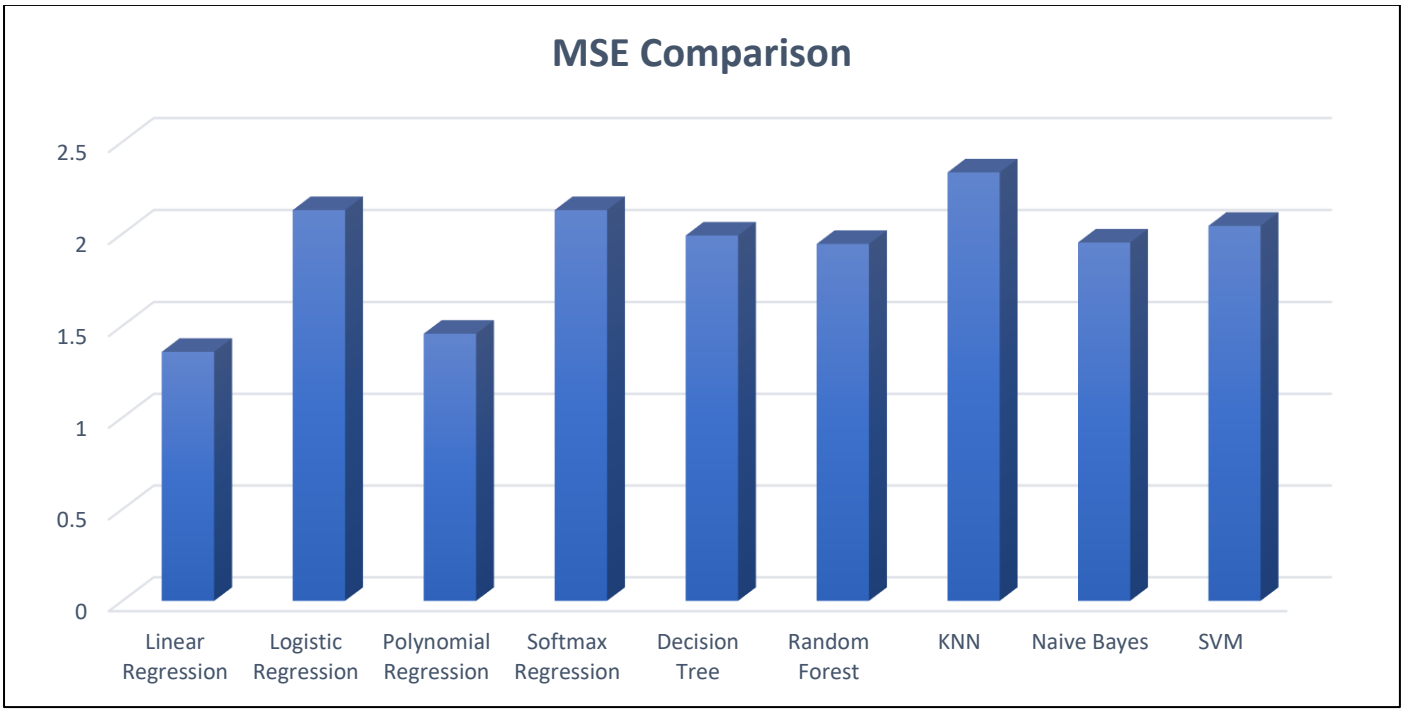
9. Naive Bayes:

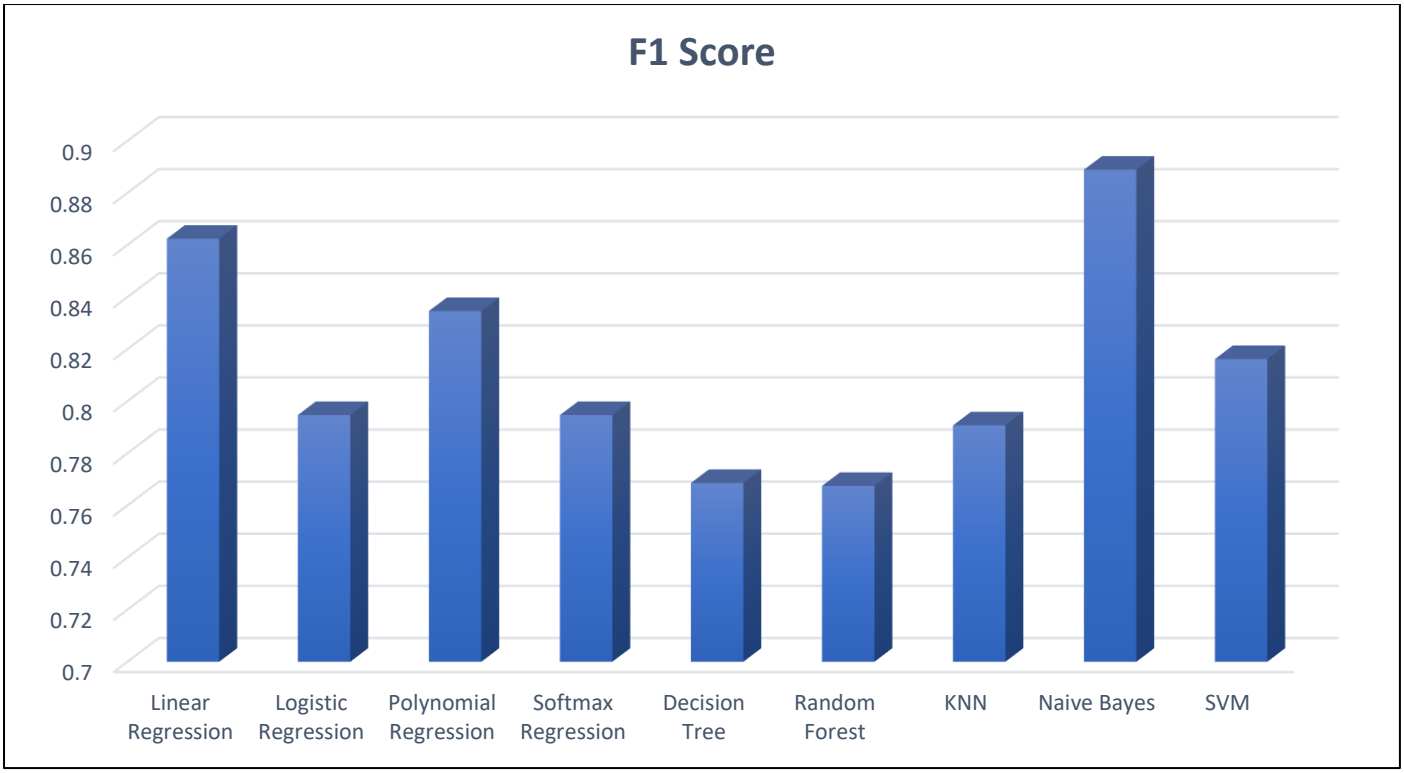
Naive Bayes is a probabilistic algorithm based on Bayes' theorem. It assumes independence between the features given the class variable. Naive Bayes is widely used for classification tasks, particularly in natural language processing and text classification. It estimates

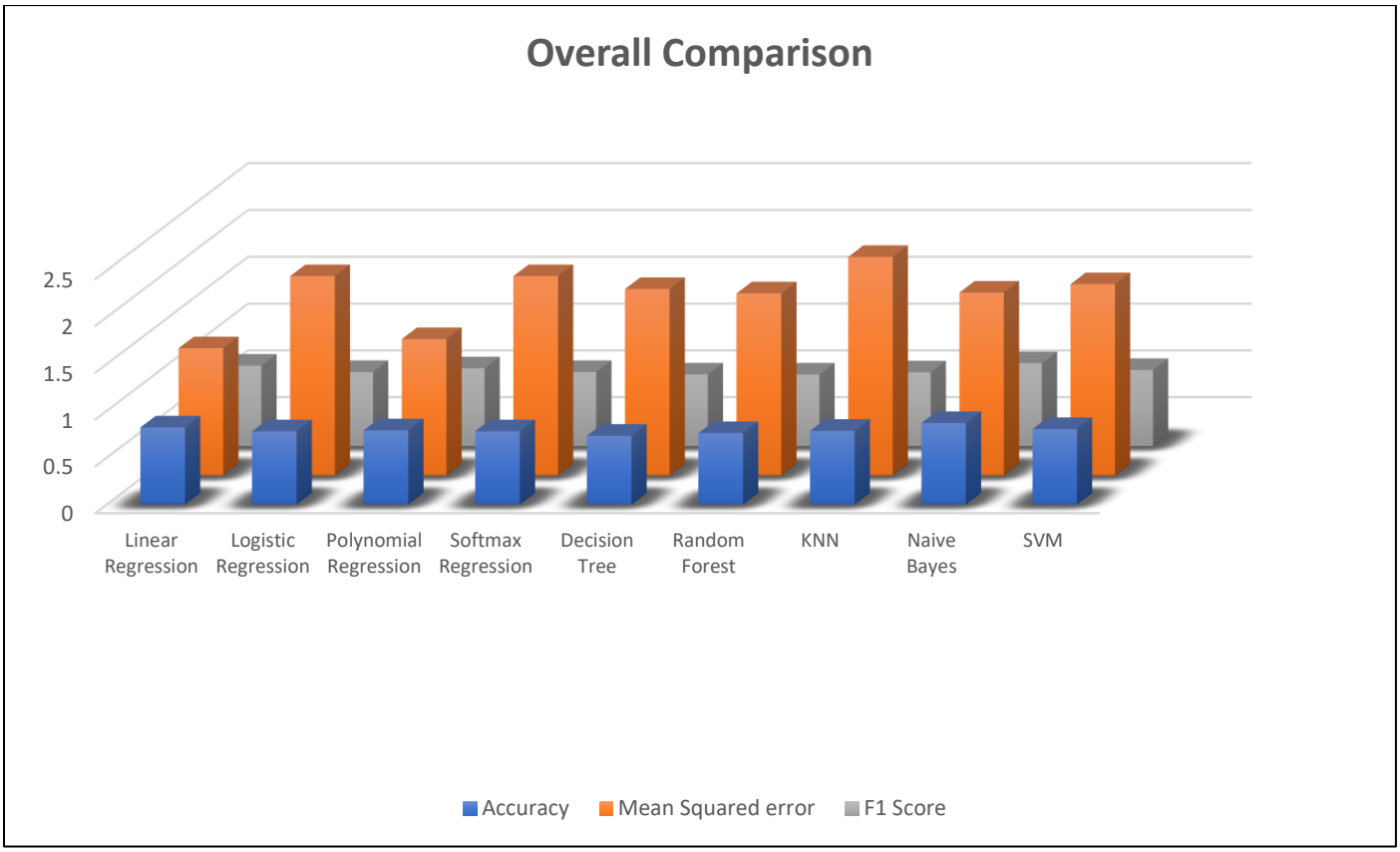
the probabilities of each class given the features and assigns the most probable class to a data point. Naive Bayes is computationally efficient, especially with high-dimensional data, as it requires estimating only class and feature probabilities. Despite its "naive" assumption, Naive Bayes can achieve competitive performance and can be trained with small datasets.

RESULT:






F1 Score
0.9
0.88
0.86
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
Linear Regression
Logistic Regression
Polynomial Regression
Softmax Regression
Decision Tree
Random Forest
KNN
Naive Bayes
SVM


Overall Comparison
2.5
2
1.5
1
0.5
0
Linear Regression
Logistic Regression
Polynomial Regression
Softmax Regression
Decision Tree
Random Forest
KNN
Naive Bayes
SVM
Accuracy
Mean Squared error
F1 Score

RESULT TABLE:

Algorithm	Mean Squared Error	F1	Accuracy	Confusion Matrix
Linear Regression	1.35813	0.86238	0.82142	[[44 25] [5 94]]
Logistic Regression	2.12857	0.79464	0.78095	[[75 14] [32 89]]
Polynomial Regression	1.45653	0.83458	0.79047	[[55 34] [10 111]]
SoftMax Regression	2.12857	0.79464	0.78095	[[75 14] [32 89]]
Decision Tree	1.99047	0.76859	0.73333	[[61 33] [23 93]]
Random Forest	1.94444	0.76744	0.76190	[[93 17] [43 99]]
KNN	2.33333	0.79069	0.78571	[[64 5] [31 68]]
Naive Bayes	1.95238	0.88888	0.86904	[[58 11] [11 88]]
SVM	2.04285	0.81614	0.80476	[[78 11] [30 91]]

CONCLUSION:

In conclusion, this research paper assessed ten machine learning algorithms for predicting heart disease, focusing on accuracy and F1 score metrics. Naive Bayes exhibited the highest accuracy of 86.9% and an impressive F1 score of 88.8%, indicating its effectiveness in correctly classifying heart disease cases. Conversely, the decision tree algorithm achieved the lowest accuracy of 73.3%, while random forest had the lowest F1 score of 76.7%. While Naive Bayes shows promise for heart disease prediction, other factors like interpretability and scalability should be considered. Future research can explore optimizing the underperforming algorithms and assembling techniques to enhance prediction accuracy and F1 score.

REFERENCE:

[1] Katarya, R., Meena, S.K. "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis." Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7

[2] Patel, Jaymin & Tejalupadhyay, Samir & Patel, Samir. (2016). Heart Disease prediction using Machine learning and Data Mining Technique. 10.090592/IJCSC.2016.018.

[3] Dwivedi, A.K. "Performance evaluation of different machine learning techniques for prediction of heart disease." Neural Comput & Applic 29, 685–693 (2018). https://doi.org/10.1007/s00521-016-2604-1

[4] C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab

Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.

[5] D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 367-372, doi: 10.1109/TENCON.2019.8929434.

[6] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[7] Chiradeep Gupta et al 2022 J. Phys.: Conf. Ser. 2161 012013DOI 10.1088/1742-6596/2161/1/012013