Advanced Generative AI for optimised disease diagnosis

Mahaswi[1], Gayathri[2], Keerthi[3], and Adetya[4]

VELLORE INSTITUTE OF TECHNOLOGY,AP vsc@vitap.ac.in
https://vitap.ac.in/

Abstract. Correct and timely diagnosis is the idea of powerful remedy and patient care. With the advent of artificial intelligence (AI) era, a big opportunity has emerged to improve diagnostic fashions and optimising techniques. This text explores the mixing of synthetic intelligence techniques such as generative adverse networks (GAN) and variable autoencoders (VAE) in ailment analysis. We purpose to increase the accuracy, robustness and performance of sickness diagnoses by leveraging generative fashions. specially, we observe using generative AI in responsibilities consisting of records processing, most beneficial problem solving, disease modeling, and drug discovery models. through evaluation and analysis, we reveal the ability of AI to trade illnesses and enhance patient results.

Keywords: Generative Artificial Intelligence · optimising techniques · GAN · VAE · disease modeling.

1 Introduction

In medical disciplines, artificial intelligence(AI) primarily focuses on developing the algorithms and ways to determine whether a system's geste is correct in complaint opinion. Medical opinion identifies the complaint or conditions that explain a person's symptoms and signs. generally, individual information is gathered from the case's history and physical examination [1]. It's constantly delicate due to the fact that numerous suggestions and symptoms are nebulous and can only be diagnosed by trained health experts. thus, countries that warrant enough health professionals for their populations, similar as developing countries like Bangladesh and India, face difficulty furnishing proper individual procedures for their maximum population of patients[2]. also, opinion procedures frequently bear medical tests, which downward - income people frequently find precious and delicate to go.

As humans are prone to error, it isn't surprising that a case may have overdiagnosis do furtheroften.However, problems similar as gratuitous treatment will arise, impacting individualities ' health and frugality [3], If overdiagnosis. According to the National Academics of Science, Engineering, and Medicine report of 2015, the maturity of people will encounter at least one individual mistake during their lifetime [4]. colorful factors may impact the misdiagnosis, which includes

* lack of proper symptoms, which frequently inconspicuous
* the condition of rare complaint
* the complaint is neglected inaptly from the consideration

Machine literacy(ML) is used virtually far and wide, from cutting- edge technology(similar as mobile phones, computers, and robotics) to health care(i.e., complaint opinion, safety). ML is gaining fashionability in colorful fields, including complaint opinion in health care. numerous experimenters and interpreters illustrate the pledge of machine- literacy- grounded complaint opinion(MLBDD), which is affordable and time-effective [5]. Traditional opinion processes are expensive, time- consuming, and frequently bear mortal intervention. While the existent's capability restricts traditional opinion ways, ML- grounded systems have no similar limitations, and machines don't get exhausted as humans do. As a result, a system to diagnose complaint with outnumbered cases ' unanticipated presence in health care may be developed. To produce MLBDD systems, health care data similar as images(i.e.,X-ray, MRI) and irregular data(i.e., cases ' conditions, age, and gender) are employed [6].

Machine literacy(ML) is a subset of AI that uses data as an input resource [7]. The use of destined fine functions yields a result(bracket or retrogression) that's constantly delicate for humans to negotiate. For illustration, using ML, locating nasty cells in a bitsy image is constantly simpler, which is generally grueling to conduct just by looking at the images. likewise, since advances in deep literacy(a form of machine literacy), the most current study shows MLBDD delicacy of above 90 [5]. Alzheimer's complaint, heart failure, bone cancer, and pneumonia are just a many of the conditions that may be linked with ML. The emergence of machine literacy(ML) algorithms in complaint opinion disciplines illustrates the technology's mileage in medical fields.

Recent improvements in ML difficulties, similar as imbalanced data, ML interpretation, and ML ethics in medical disciplines, are only a many of the numerous grueling fields to handle in a nutshell [8]. In this paper, we give a review that highlights the new uses of ML and DL in complaint opinion and gives an overview of development in this field in order to exfoliate some light on this current trend, approaches, and issues connected with ML in complaint opinion. We begin by outlining several styles to machine literacy and deep literacy ways and particular armature for detecting and grading colorful forms of complaint opinion.

Many academics and practitioners have used machine learning (ML) procedures in disorder prognosis. This segment describes many styles of machine-getting to know-based totally ailment diagnosis (MLBDD) that have acquired lots interest because of their importance and severity. for example, due to the global relevance of COVID-19, numerous research targeting COVID-19 disorder detection using ML from 2020 to the prevailing, which additionally obtained extra priority in our take a look at. severe sicknesses consisting of coronary heart sickness, kidney sickness, breast most cancers, diabetes, Parkinson's, Alzheimer's, and COVID-19 are discussed in short, while other illnesses are blanketed in brief underneath the "different disease".

Generative artificial Intelligence (AI) consists of a collection of algorithms and methodologies designed to synthesize information outputs, which include pix, texts, or sounds, which could on occasion be indistinguishable from actual-world information. fundamental to generative AI are neural networks, particularly Generative detrimental Networks (GANs) and Variational Autoencoders (VAEs) (Rosca et al. 2017; Mescheder, Nowozin, and Geiger 2017). GANs, conceptualized thru (Goodfellow and Pouget-Abadie 2014), feature via a dualistic structure concerning neural networks: the Generator and the Discriminator. The Generator's role is to provide information samples, whereas the Discriminator evaluates those samples against real statistics, functioning as a binary classifier. The mathematical goal of the Generator is to maximise the possibility of the Discriminator making an errors, formulated as a minimax exercise. this could be represented via using the cost feature

(V(G, D)), wherein **(G)** and **(D)** denote the generator and discriminator, respectively:

[min G max D V (D,G) = Ex pdata(x)[logD(x)] + Ez pz(z)[log(1-D(G(z)))]]

here, **(E)** denotes the expectancy, **(pdata)** is the distribution of the actual facts, and **(pz)** is the enter noise distribution for the generator.

Variational Autoencoders, then again, are built upon the ideas of Bayesian inference (Doersch 2016) . They encompass an encoder and a decoder. The encoder transforms facts right into a latent area representation, on the equal time because the decoder reconstructs the records from this latent space (Girin et al. 2020). The objective is to research the parameters of a risk distribution modeling the statistics. A key thing of VAEs is the usage of a variational approach for approximating the posterior distribution of latent variables. that is executed with the aid of minimizing the Kullback-Leibler (KL) divergence a number of the approximate and actual posterior, it is a degree of ways one opportunity distribution diverges from a second, reference hazard distribution (Prokhorov et al. 2019). The loss characteristic of a VAE comprises two terms: the reconstruction loss, which guarantees that the decoded samples resemble the specific inputs, and the KL divergence, selling the encoding distribution to approximate the previous distribution. The loss characteristic ($\mathcal{L}$) for a VAE may be expressed as:

$$\mathcal{F}c = -\mathcal{F}o\{Ez \sim q\vartheta(z|x)[\log p_{\Theta}(x|z)]\} + \mathrm{KL}(q_{\vartheta}(z|x)||p(z))$$

Here, $q_{\phi}(z|x)$ is the approximate posterior, $p_{\theta}(x|z)$ is the probability, $p(z)$ is the prior over latent variables, and (KL) represents the Kullback-Leibler divergence.

These methodologies arise at the intersection of statistical learning theory and computational performance to enable generative AI to provide complex, high-dimensional statistical distributions. Generative AI's utility extends to developing synthetic medical data, which is crucial for research and training without compromising patient privacy. This data is used notably to train machine learning models, especially in scenarios where actual patient data is limited or

sensitive. Moreover, AI-driven biomarker discovery is gaining traction. These AI systems facilitate early disease detection and the development of targeted treatments, significantly impacting areas like cancer research by identifying new biomarkers. Generative AI models simulate various disease outbreak scenarios, aiding public health officials in preparing for and responding to health crises. For medical trials, AI simulates patient responses, providing insights into treatment effectiveness and potential side effects. AI optimizes healthcare workflows, including patient scheduling and aid allocation.

1.1 Motivation

The reason of this paper is to give perceptivity to rearmost and fortune experimenters and interpreters regarding contrivance- literacy- grounded optimised complaint opinion(MLBDD) on the way to useful resource and enable them to choose the most applicable and advanced contrivance studying deep studying strategies, thereby growing the probability of rapid-fire and dependable complaint discovery and class in analysis. also, the estimate points to come apprehensive of implicit studies associated with the MLBDD. In wide, the compass of this take a look at is to give the proper reason for the following questions:

1.How does advanced Generative AI decorate diagnostic accuracy?: This question delves into the particular mechanisms through which advanced generative AI strategies improve the accuracy of sickness analysis by using studying complex and heterogeneous scientific statistics.

2.What are the advantages of early sickness detection facilitated by way of these models?: Exploring the implications of early ailment detection on remedy outcomes, affected person prognosis, and healthcare resource usage.

3.How do those models help customized medication?: information how advanced generative AI strategies tailor diagnostic and remedy strategies to person affected person characteristics, consisting of genetic profiles, disorder threat elements, and treatment responses.

4.What role do these models play in optimizing resource allocation inside healthcare structures?: Investigating how these models prioritize excessive-threat people for diagnostic trying out and intervention, main to extra efficient and effective useful resource utilization.

5.How do these models provide decision support to healthcare carriers?: examining the approaches wherein advanced generative AI strategies synthesize and interpret complicated scientific statistics to aid in diagnostic interpretation, treatment making plans, and affected person management.

In this research paper, we summarize the different machine learning (ML) methods used in various optimised disease diagnosis applications. The remainder of this paper is structured as given. In Section 2, we discuss the background and overview of Machine Learning, whereas in Section 3, we detail the article selection technique. Section 4 includes the bibliometric analysis of data. In Section 5, we discuss the use of ML in various optimized disease diagnosis, and in the Section 6, we identify the most frequently utilized ML methods and data types based on

linked research. In the Section 7, we discuss the anticipated trends, and problems. Lastly,In Section 9 concludes the article with a general conclusion.

2 BASICS AND BACKGROUND

Machine learning (ML) approach analyzes data samples to create main conclusions using mathematical, numerical and statistical approaches, allowing the machines to learn without programming.

2.1 MACHINE LEARNING ALGORITHMS:

RANDOM FOREST is a classifier that has many decision trees on various subsets of a given dataset and averages them to improve the prediction accuracy of the dataset.

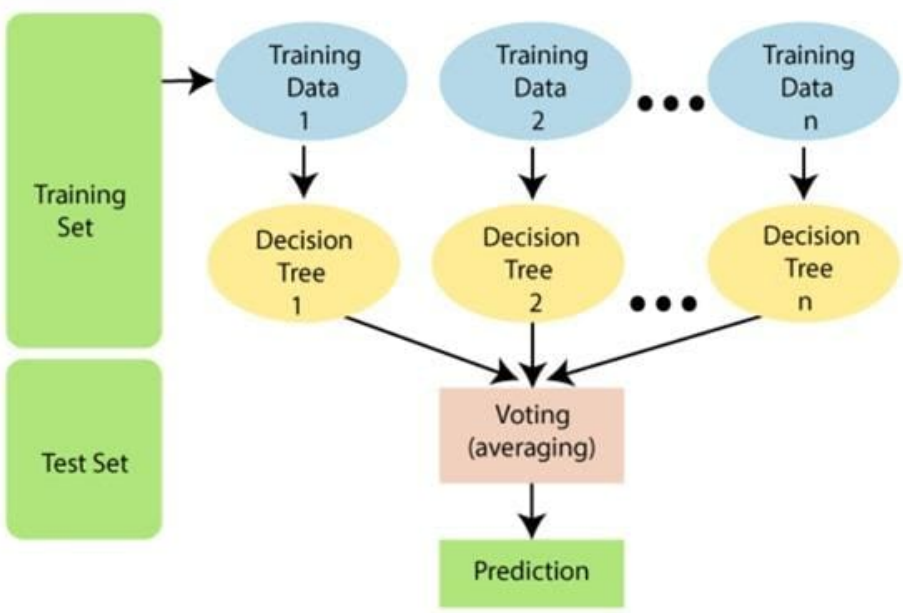


Fig. 1: Random Forest

GRADIENT BOOSTING is a powerful boosting algorithm that combines multiple weak learners into a strong learner, where each new model is trained using gradient descent to minimize the loss function, such as the mean square error or cross-entropy of the previous model. At each iteration, the algorithm

computes the gradient of the loss function based on the predictions of the current ensemble and then trains a new weak model to minimize this gradient. Predictions from the new model are then added to the ensemble and the process is repeated until the stopping criterion is met.

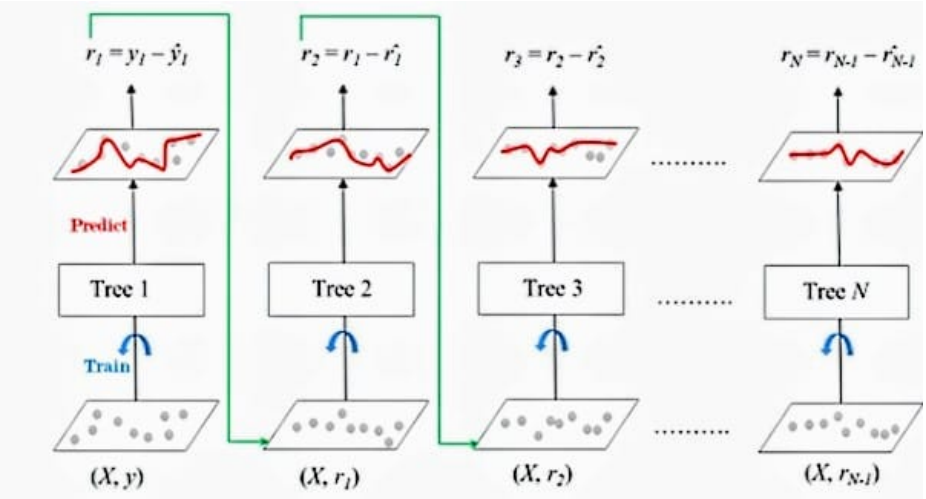


Fig. 2: Gradient boosting

LINEAR SVM is used for linearly separable data. That is, if a data set can be divided into two classes using a single straight line, such data is called linearly separable and a classifier called linear SVM classifier is used. SVM algorithm helps in finding the best line or decision boundary. This optimal boundary or region is called the hyperplane. The SVM algorithm finds the closest point between two classes of lines. These points are called support vectors. The distance between the vector and the hyperplane is called the margin. And the goal of SVM is to maximize these profits. The hyperplane with the maximum margin is called the optimal hyperplane.

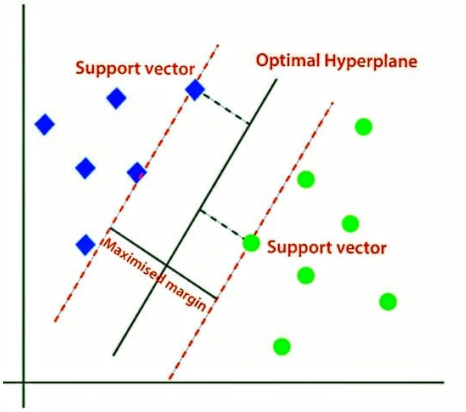


Fig. 3: Linear SVM

POLYNOMIAL SVM Kernel functions, also known as polynomial SVM, are methods used to transform data into the required format for processing. The term "kernel" refers to the mathematical functions used in support vector machines to manipulate data. The kernel function typically transforms the training data set so that the nonlinear decision surface can be expressed as linear equations in a higher dimensional space. Essentially, it computes the dot product between two points in the standard dimensions of the object.

The standard kernel function equation is: $\mathrm{K}(\bar{\mathbf{x}}) = \begin{cases} 1 & \text{if } \|\bar{\mathbf{x}}\| \geq 1 \\ 0 & \text{otherwise} \end{cases}$.

CODE:

```
from sklearn.svm import SVC
classifier = SVC(kernel ='rbf',
random_state = 0) classifier.fit(x_train, y_train)
```

RBF SVM Radial basis function support vector machine (RBF SVM) is a powerful machine learning algorithm that can be used for classification and regression. It is a non-parametric model suitable for non-linear and high-dimensional data. This algorithm uses kernel functions (such as radial basis functions) to measure the similarity between pairs of data points in the feature space. This means:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

where x and x' are the input points and gamma are the control kernel width hyperparameters. , And. is the Euclidean distance between points. The kernel function evaluates the similarity between pairs of data points based on their distance from a given location.

2.2 PERFORMANCE EVALUATIONS:

Accuracy (Acc):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision (Pn):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Rc):

$$\text{Recall} = \frac{TP}{TP + FN}$$

Sensitivity (Sn):

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity (Sp):

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F-measure:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3 MACHINE LEARNING TECHNIQUES FOR OPTIMISED DISEASE DIAGNOSIS:

Machine learning techniques for optimized sickness prognosis contain using algorithms to research medical records and make correct predictions about disorder presence, development,and remedy consequences. these techniques leverage advanced statistical methods and computational algorithms to pick out patterns, correlations, and predictive biomarkers in large-scale datasets. by using analyzing numerous information assets along with medical imaging, genomic statistics, medical statistics, and patient-reported results, machine studying models can offer valuable insights into ailment prognosis and control. Key strategies consist of supervised gaining knowledge of for class responsibilities, unsupervised mastering for clustering and anomaly detection, and deep learning for complicated statistics representations. these methods permit early sickness detection, personalized remedy techniques, and progressed affected person effects, in the end reworking healthcare delivery and biomedical research.

Diseases such as heart disease, kidney disease, breast cancer, diabetes, Parkinson's, Alzheimer's, and COVID-19 are discussed in brief, while other diseases are covered briefly under the "other disease"

3.1 Heart Disease

Researchers use machine learning (ML) approaches to identify cardiac diseases and optimize them [8,9].

Ansari et al. (2011) offered an automated coronary heart disease diagnosis system based on neurofuzzy integrated systems that could yield around 95% accuracy [8]. However, this study is inefficient as it lacks a clear explanation of how the technique would work in various scenarios like multiclass classification and big data analysis.

Rubin et al. (2017) used a deep convolutional neural network-based approach to detect non-regular cardiac sounds. The authors adjusted the loss function to improve the training dataset's sensitivity and specificity, resulting in a final prediction of 0.95 specificity and 0.73 sensitivity (Rubin)Researchers use machine learning (ML) approaches to identify cardiac diseases and optimize them [10]

In addition to ML, deep learning-based algorithms have recently garnered attention in detecting cardiac diseases. Miao and Miao (2018) offered a DL-based technique to diagnose cardiotocographic fetal health based on morphologic patterns, achieving an accuracy of 0.85 F-score [7].

The table below summarizes some of the cited publications that employed machine learning and deep learning approaches in the diagnosis of cardiac diseases. (For more information, see [5]).

REFERENCE	CONTRIBUTION	ALGORITHM	DATA SET	DATA TYPE	PERFORMANCE EVALUATION
[11]	Predict coronary heart disease	Gaussian NB, Bernoulli NB, and RF	Cleveland dataset	Tabular	TabularC Accuracy:85.00%, 85.00% and 75.00%RF (Accuracy: 80.327%, Precision: 82%, Recall:80%, F1-score:80%), CNN
[12]	Predicting heart diseases	RF,CNN	Cleveland dataset	Tabular	(Accuracy-78.688, Precision-80%, Recall-79%, F1-score-78%)
[13]	Heart disease classification	SVM	Cleveland dataset	Tabular	Accuracy—73–91%
[14]	Heart disease classification	Back-propagation NN,LR	Cleveland dataset	Tabular	Accuracy (BNN—85.074%, LR—92.58%)

In paper , uses various data from Cleveland dataset .Propose heart disease prediction using gradient boosting and random forest .Based on performance

from different factor gradient boosting, LINEAR SVM, polynomial SVM ,RBF SVM (87% Accuracy) achieves optimum performance than Random forest (84% accuracy).
As the accuracy by the use of gradient boosting , linear SVM , polynomial SVM ,RBF SVM showed the best accuracy of 87% in contrast to other models providing the better precision and transparency of the model used for predicting heart disease.

4 Methodology:

The main aim of this study is to develop a predictive model for early detection of heart disease using multiple machine learning algorithms. A combined health screening approach to accurately predict the onset of heart disease. Additionally, information from the first article has been incorporated into the method combining Naive Bayes, k-nearest neighbor (KNN), decision trees, network of neural networks (ANN) and random forest algorithms.The purpose of the proposed method is to predict the occurrence of heart disease for early detection of the disease in less time. In this approach, we are using different machine learning algorithms like gradient boosting, LINEAR SVM, polynomial SVM , RBF SVM to predict the heart disease based on some health parameters. It was done using Anaconda Navigator's Jupyter Notebook, an open source software platform that enables the implementation of various machine learning algorithms. The environment supports real-time code generation, visualization, data processing, and graphing to make the predictive modeling process more efficient and effective. These studies are designed to develop powerful predictive models for early detection of heart disease, helping to improve health outcomes and the health of patients.
The first paper has an accuracy of 86%, while the accuracy of the present paper is 87%.So the present paper is more efficient than the cited one.

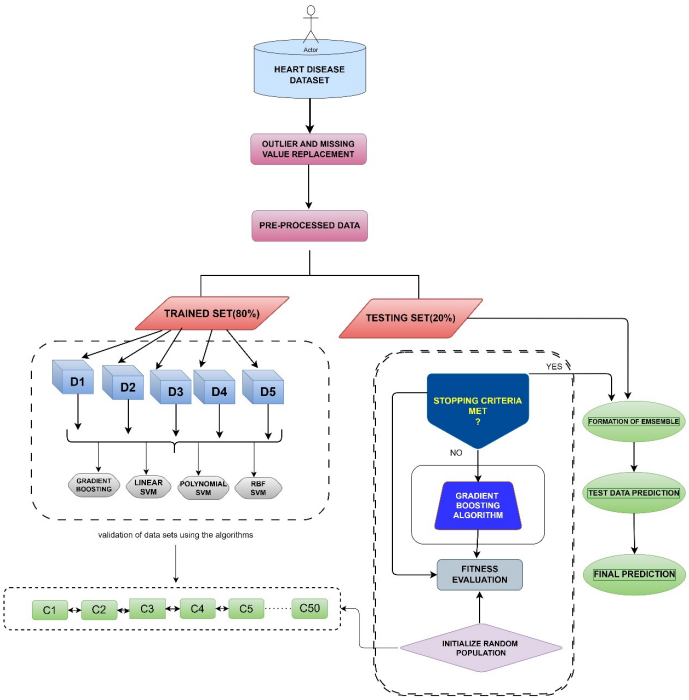


Fig. 4: PROPOSED MODEL

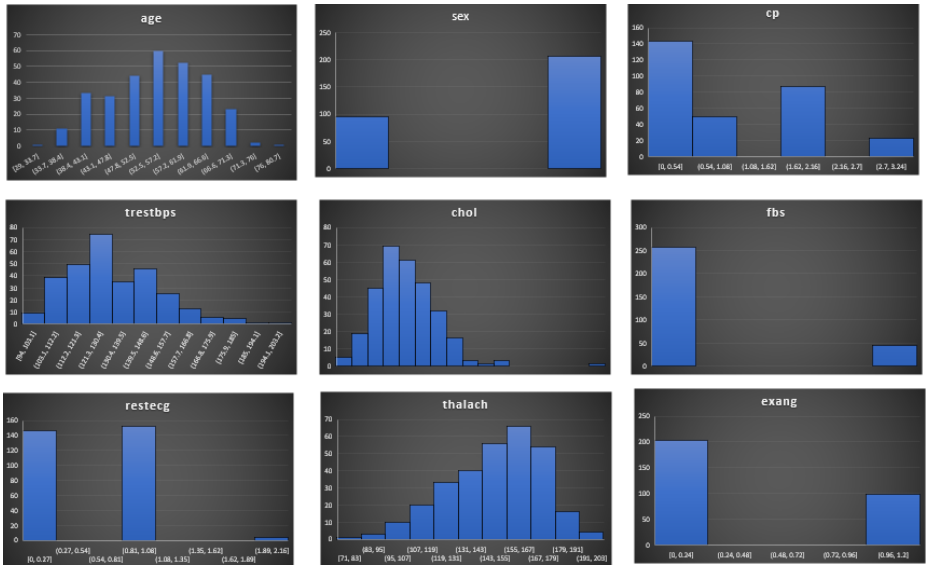
4.1 Dataset for implementation:

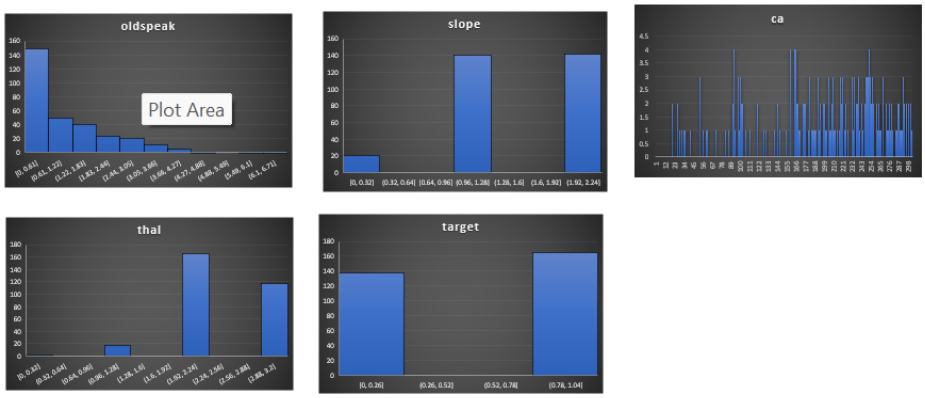
The dataset used to predict heart disease is taken from Kaggle , the dataset contains nearly 300 data elements to be trained or tested based on groups.
Age
Sex
Cerebral palsey/chest pain (CP)
Blood Pressure(bps) in mm HG
Cholesterol in mg
Fasting blood sugar test (fbs)
resting electrocardiographic results
thalach (Maximum heart rate achieved)
exang (Exercise induced Angina)
oldpeak (ST depression induced by exercise relative to rest)
slope(the slope of the peak exercise ST segment)
ca(Number of major vessels)
thal(Reversible defect)
Target(0,1)

4.2 Data splitting:

Data is divided into training and testing data. 20% data is used for testing purpose while 80% data is used for training purpose. We performed data normalization for removing Nan values.

5 Visualization of data:





The performance and the accuracy of every experiment is evaluated by standard metrics such as TP , TN , precision, recall and F-measure which are calculated by Confusion Matrix which is known as predictive classification table. All these measures will be used to compare the performance of these selected algorithms and implemented algorithms.

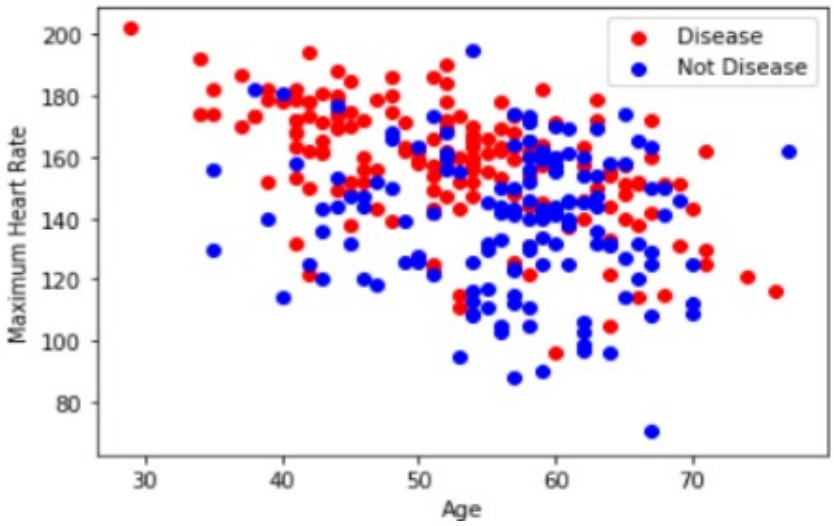


Fig. 5: heart disease rate plot

Graph above visualizes the patients who have heart disease with red dots on graph while the blue dots represent the patient who are not suffering from heart disease.

Random Forest Algorithm: The random forest algorithm includes a collection of decision trees; Each tree in the collection contains data from the training set, called the bootstrap model. One-third of these training samples are put into testing data, called out-of-bag (oob) samples, which we will discuss later. Another example of randomness is post-injection through packing, which introduces more variance into the data and reduces the correlation between decision trees. Depending on the type of problem, the decision of the forecast will be different. For regression tasks, individual decision trees will be averaged, and for classification tasks, majority voting will be done, meaning the variance of the majority distribution will produce the predicted class. Finally, the model is used in the competition to complete the prediction.

Confusion Metrics:

The precision of Random forest depends on the distance metric and the K value. It can be measured by confusion metrics.

		Predicted	
		Diseased	Healthy
Actual	Diseased	TP = 25	FN = 4
	Healthy	FP = 3	TN = 29

Random forest result:

```
Random Forest Test Accuracy: 0.84
              precision    recall  f1-score   support

           0       0.83      0.83      0.83        29
           1       0.84      0.84      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61

Random Forest ROC AUC Score: 0.84
```

Gradient Boosting Algorithm: Gradient boosting is a common machine learning technique that combines predictions from multiple weak learners (usually decision trees). The goal is to improve the overall prediction by improving the model's weights based on the errors of previous studies and gradually reducing the prediction error to improve the model's accuracy.

Confusion Metrics:

The precision of gradient boosting depends on the distance metric and the K value. It can be measured by confusion metrics.

		Predicted	
		Diseased	Healthy
Actual	Diseased	TP = 23	FN = 6
	Healthy	FP = 7	TN = 25

Gradient boosting result:

```
Gradient Boosting Test Accuracy: 0.79
              precision    recall  f1-score   support

           0       0.77      0.79      0.78        29
           1       0.81      0.78      0.79        32

    accuracy                           0.79        61
   macro avg       0.79      0.79      0.79        61
weighted avg       0.79      0.79      0.79        61

Gradient Boosting ROC AUC Score: 0.79
```

Linear SVM Algorithm: LINEAR SVM is used for linearly separable data. That is, if the data set is available Data divided into two classes using a single straight line is called linear. It is separable and uses a classifier called linear SVM classifier. SVM algorithm It helps you find the best decision line or boundary. This optimal limit or That region is called the hyperplane. The SVM algorithm finds the closest point. Between the lines of two classes. These points are called support vectors. distance The space between the vector and the hyperplane is called the boundary. and goal The goal of SVM is to maximize these benefits. The hyperplane with the maximum margin is:
It is called the optimal hyperplane.

Confusion Metrics:

The precision of Linear SVM depends on the distance metric and the K value. It can be measured by confusion metrics.

		Predicted	
		Diseased	Healthy
Actual	Diseased	TP = 25	FN = 4
	Healthy	FP = 4	TN = 28

Linear SVM result:

```
Linear SVM Test Accuracy: 0.87
              precision    recall  f1-score   support
...
   macro avg       0.87      0.87      0.87        61
weighted avg       0.87      0.87      0.87        61

RBF SVM ROC AUC Score: 0.87
```

Polynomial SVM Algorithm: Polynomial SVM (Support Vector Machine): The Polynomial SVM is a variant of the traditional SVM algorithm that uses a polynomial kernel function to map the input data into a higher-dimensional space. In this higher-dimensional space, a hyperplane is constructed to separate the classes by maximizing the margin between them. The polynomial kernel function computes the dot product between two vectors in the feature space, raised to a certain power (degree), and optionally scaled by a coefficient (gamma). By adjusting the degree of the polynomial, you can control the complexity of the decision boundary. Higher degrees allow for more complex decision boundaries, but they also increase the risk of overfitting. Polynomial SVMs are effective for non-linear classification tasks where the decision boundary cannot be adequately described by a linear function.

Radical Based Function SVM Algorithm: RBF SVM (Radial Basis Function SVM): The RBF SVM is another variant of the SVM algorithm that uses a radial basis function (RBF) kernel. Unlike the polynomial kernel, the RBF kernel does not require specifying the degree of the polynomial. Instead, it measures the similarity (or distance) between data points in the original feature space. The RBF kernel function computes the similarity between two data points based on their Euclidean distance. Data points that are closer in feature space have a higher similarity. The RBF kernel is characterized by a parameter called gamma, which controls the width of the Gaussian kernel. A smaller value of gamma makes the decision boundary smoother, while a larger value makes it more complex. RBF SVMs are highly flexible and can capture complex decision boundaries in both linear and non-linear datasets. They are widely used in various machine learning applications, especially when dealing with high-dimensional data.

Ensembling Soft Voting: Soft Voting is a method of combining the predictions of multiple base models in classification tasks. Unlike Hard Voting, which simply takes the majority vote of the base models, Soft Voting takes into account the probability estimates (or confidence scores) of each model for each class. In Soft Voting, the final prediction is made by averaging the predicted probabilities from all base models and selecting the class with the highest average probability as the final prediction. This method is particularly useful when the base models are capable of estimating class probabilities, such as in the case of probabilistic classifiers like Logistic Regression, Softmax Regression, or models with built-in probability estimates like Random Forests with predict proba() method. Soft Voting tends to provide more robust and accurate predictions compared to Hard Voting, especially when the base models have varying degrees of confidence in their predictions. Soft Voting can handle situations where there is uncertainty in the predictions made by individual models, allowing it to make more nuanced decisions. Soft Voting is commonly used in ensemble learning techniques like the Voting Classifier in scikit-learn, where multiple classifiers are combined to improve overall prediction performance. In summary, Soft Voting Ensemble is a flexible and effective method for combining the predictions of multiple clas-

sifiers by considering the probability estimates of each model, leading to more accurate and robust predictions, especially in situations where individual models may have varying degrees of confidence

5.1 Results analysis and Comparison

The goal of our research is to use machine learning algorithms to treat heart disease in clinical practice. For this purpose, we conducted experiments using different algorithms on heart patients. Using this we can know which classification algorithm is best in predicting heart disease.

Table 1: The result of the cited paper

ALGORITHMS	ACCURACY	TN	FP	FN	TP
RANDOM FOREST	0.82	42	11	5	33
KNN	0.87	30	2	8	36
NAVIE BAYES	0.88	31	4	5	36
DECISION TREE	0.78	42	15	5	29
ANN	0.87	30	2	8	36

After using different algorithms, the second step is to compare the different learning machines used in these experiments and choose the best algorithm that provides the most accuracy. Different power values were used to compare these tests; that is, true positive,true negative,false positive ,false negative and ROC curves were used. The algorithm is summarized in the table below.

Table 2: The result of present paper

ALGORITHMS	ACCURACY	TN	FP	FN	TP
RANDOM FOREST	0.84	27	5	5	24
LINEAR SVM	0.87	28	4	4	25
POLYNOMIAL SVM	0.90	29	3	3	26
RBF SVM	0.87	27	5	3	26
GRADIENT BOOSTING	0.79	25	7	6	23

6 Conclusion and Future scope

Our study mainly focused on the use of MACHINE LEARNING techniques in healthcare especially in the detection of heart disease. Heart disease is a fatal disease which may cause death. Machine learning techniques were implemented using the following algorithm, LINEAR SVN, POLYNOMIAL SVN,GRADIENT BOOSTING,RBF SVN and Random Forest. We measured performance on the

basis of Accuracy, TN, FP, FN and TP rate and in some of the algorithm.We conducted five experiments with the same data set to predict the heart disease. The result of all the implemented algorithm are shown in tabular form for better understanding and comparisons .The experiment shows that GRADIENT BOOSTING gives the highest accuracy which is 87% followed by SVN and RBF SVM with accuracy of 87% .Our findings indicate that machine learning can be used and applied in the healthcare industry to predict and diagnose the disease at early stages.

7 References

1. McPhee S.J., Papadakis M.A., Rabow M.W., editors. Current Medical Diagnosis & Treatment. McGraw-Hill Medical; New York, NY, USA: 2010.
2. Ahsan M.M., Ahad M.T., Soma F.A., Paul S., Chowdhury A., Luna S.A., Yazdan M.M.S., Rahman A., Siddique Z., Huebner P. Detecting SARS-CoV-2 From Chest X-ray Using Artificial Intelligence. IEEE Access. 2021;9:35501–35513. doi: 10.1109/ACCESS.2021.3061621.
3. Coon E.R., Quinonez R.A., Moyer V.A., Schroeder A.R. Overdiagnosis: How our compulsion for diagnosis may be harming children. Pediatrics. 2014;134:1013–1023. doi: 10.1542/peds.2014-1778.
4. Balogh E.P., Miller B.T., Ball J.R. Improving Diagnosis in Health Care. National Academic Press; Washington, DC, USA: 2015.
5. Ahsan M.M., Siddique Z. Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review. arXiv. 20212112.06459
6. Ahsan M.M., E Alam T., Trafalis T., Huebner P. Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients. Symmetry. 2020;12:1526. doi: 10.3390/sym12091526.
7. Stafford I., Kellermann M., Mossotto E., Beattie R., MacArthur B., Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ Digit. Med. 2020;3:1–11. doi: 10.1038/s41746-020-0229-3
8. Ansari A.Q., Gupta N.K. Automated diagnosis of coronary heart disease using neuro-fuzzy integrated system; Proceedings of the 2011 World Congress on Information and Communication Technologies; Mumbai, India. 11–14 December 2011; pp. 1379–1384.
9. Ahsan M.M., Mahmud M., Saha P.K., Gupta K.D., Siddique Z. Effect of data scaling methods on machine Learning algorithms and model performance. Technologies. 2021;9:52. doi: 10.3390/technologies9030052.
10. Rubin J., Abreu R., Ganguli A., Nelaturi S., Matei I., Sricharan K. Recognizing abnormal heart sounds using deep learning. arXiv. 20171707.04642
11. Bemando C., Miranda E., Aryuni M. Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms; Proceedings of the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM); Pekan, Malaysia.

24–28 August 2021; pp. 232–237.
12. Kumar R.R., Polepaka S. ICCII 2018, Proceedings of the Third International Conference on Computational Intelligence and Informatics. Springer; Singapore: 2020. Performance Comparison of Random Forest Classifier and Convolution Neural Network in Predicting Heart Diseases; pp. 683–691.
13. Singh H., Navaneeth N., Pillai G. Multisurface Proximal SVM Based Decision Trees For Heart Disease Classification; Proceedings of the TENCON 2019–2019 IEEE Region 10 Conference (TENCON); Kerala, India. 17–20 October 2019; pp. 13–18.
14. Desai S.D., Giraddi S., Narayankar P., Pudakalakatti N.R., Sulegaon S. Advanced Computing and Communication Technologies. Springer; Berlin/Heidelberg, Germany: 2019. Back-propagation neural network versus logistic regression in heart disease classification; pp. 133–144.