Heart Disease Prediction Using Deeplearning

Mr. K. Sathyaseelan[1], M. Abdulsalam[2], E. K. Akilnanda[3], T. NijeethKumar[4]

[1] Assistant Professor, Department of Information Technology, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, Email: sathyaseelan.it@srit.org.

[2] Student, Department of Information Technology, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, Email: abdulsalam.2006@srit.org.

[3] Student, Department of Information Technology, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, Email: akil.2006@srit.org.

[4] Student, Department of Information Technology, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, Email: nijeeth.2006@srit.org.

Abstract :

Cardiovascular diseases (CVDs) continue to be a major worldwide health concern that require prompt management and an efficient early diagnosis. The goal of this research is to predict heart disease using deep learning, a powerful technique for identifying complex patterns from a variety of datasets. The main goal is to use deep neural networks and patient data integration to improve prediction accuracy and offer insightful information for proactive healthcare decision-making. The importance of this research resides in its capacity to revolutionize risk assessment techniques, promoting customized treatment and building confidence via the creation of clear and understandable models. We aim to improve the prediction accuracy of heart disease and reveal new insights into the complex relationships found in heterogeneous datasets by utilizing the power of deep learning. It is expected that this study's findings will contribute to a paradigm shift in the medical field by empowering medical professionals to customize interventions, make well-informed decisions, and ultimately improve patient outcomes in the area of cardiovascular health.

Index Term :

Machine learning, decision support system, deep learning, ensemble classifiers, heart disease diagnosis, accuracy, performance, cross validation.

Introduction :

Cardiovascular diseases (CVDs) are a major worldwide health problem that must be effectively diagnosed at an early stage in order to minimize their detrimental effects. In order to predict cardiac disease, this study uses deep learning, a reliable technique for identifying intricate patterns in diverse datasets. By integrating patient data and applying deep neural networks, the main goal is to improve prediction accuracy and offer significant insights for proactive healthcare choices. The technology that is being examined has the potential to transform risk assessment and bring in a new age of personalized healthcare. Our technique explores the subtleties of many datasets to find patterns and connections that might be early markers of cardiovascular diseases. Our technique relies heavily on deep neural networks, which are capable of extracting complex insights from convoluted data structures and providing a powerful predictive analytics tool. This finding is significant because it has the potential to revolutionize risk assessment techniques and encourage customized treatment regimens based on specific patient profiles. Additionally, by emphasizing interpretability and openness, the study aims to increase patient and healthcare professional confidence in prediction models. In conclusion, our study uses deep learning to make a contribution to the worldwide conversation on cardiovascular health. The project aims to shift healthcare into a proactive paradigm and improve the effectiveness of managing cardiovascular disease globally by advancing predicting accuracy, providing actionable insights, and promoting transparent models.

Literature Survey :

1.Chae et al. used big data and deep learning to forecast the incidence of infectious illnesses, including chickenpox, malaria, and scarlet fever. The paper examined how well the conventional Autoregressive Integrated Moving Average (ARIMA) model performed in comparison to the Long Short-Term Memory (LSTM) and Deep Neural Network (DNN) models. Utilizing a dataset of 576 observations obtained from several sources, including KCDC, Naver Data Lab, KMA, and Twitter, the information included temperature, query data, and social media data. While ARIMA was used to analyze non-stationary time series data and show its application to infectious disease prediction, the Deep Neural Network used a back-propagation method to forecast time series data. The research provided in-depth insights by delving into the complex variations and oscillations in the data.

2. Raju et al. focused on utilizing a deep learning system to automatically detect diabetes in their work. Their suggested architecture used a convolutional technique using deep neural networks trained on a benchmark heart disease dataset downloaded from Kaggle in order to categorize diabetic retinopathy stages based on funduscopic pictures. High sensitivity and specificity values were found for different phases of diabetic retinopathy by the classification analysis. About 35,000 photos were used in the network training to obtain a sensitivity of 80.28% and specificity of 90.29%. An additional 8,810 images were used to contribute to a model accuracy of 93.28%. In order to increase classification accuracy, the research recommended taking into account other variables including demographics and family history. It also underlined the need of removing low-quality photos in order to lower mistake rates.

3. Reddy and Khare developed a hybrid heart disease prediction system using Rule-Based Fuzzy Logic (RBFL) and Oppositional Firefly with BAT (OFBAT). By making feature selection easier, Locality Preserving Projection (LRP) helped create a classification model using a fuzzy logic system. After that, fuzzy rules were created using example data, and the relevant rules were chosen using the OFBAT algorithm. These fuzzy rules and membership functions were used to create an efficient fuzzy system for categorization. Testing using Cleveland, Hungarian, and Switzerland, three publicly accessible UCI datasets, demonstrated RBFL's 78% accuracy for UCI datasets. The research makes recommendations for possible improvements by incorporating deep learning techniques.

4. Khateeb et al. used the 303-record UCI dataset, which is available to the public, to develop a diagnostic paradigm for heart illness. They evaluated six different machine learning algorithms in their investigation. While the second case only used seven of the fourteen relevant characteristics, the first scenario assessed classifiers without feature reduction.

In the third scenario, general data such as age, sex, and resting blood pressure were removed before accuracy estimates were made. Re-sampling was done on datasets with different properties in examples four and five. Using KNN in conjunction with the Synthetic Minority Over Sampling Technique (SMOTE), the sixth case achieved the best accuracy, almost 80%. Nevertheless, the suggested method is limited to using a single dataset.

5. Verma et al. put out a thorough framework that includes the identification of important risk factors for the diagnosis of heart disease. The study successfully identified these risk variables using K-means clustering and Particle Swarm Optimization (PSO). Then, supervised learning algorithms were used to classify data from three different medical datasets: 335 cases from Indhira Gandhi Medical College, patient data from India, and Cleveland heart disease data. Multinomial Logistic Regression performed better than other classifiers in all datasets, according to the results analysis, highlighting the diagnostic framework's usefulness.

6. Chen et al. presented a decision support system in their study that is intended to help doctors identify cardiac disease. The model used existing data to predict and assess cardiac disease. It was based on the Learning Vector Quantization (LVQ) technique. 13 important qualities were chosen for the framework, and then an artificial neural network was used to classify heart disease based on these attributes. By using Receiver Operating Characteristics (ROC) curve analysis, the user-friendly system showed an impressive 80% accuracy. The study made the case that text mining might improve the model's performance, especially when it comes to predicting unstructured data in the diagnosis of heart disease.

7. The Naïve Bayes classifier was utilized by Jabbar et al. in their work to improve the predicted accuracy of early heart disease detection. To remove unnecessary and redundant features, they used discretization and genetic search. The Genetic Algorithm (GA) was then used for feature selection, which involved removing the attributes with the lowest ranking from the dataset. When comparing Naïve Bayes' performance to alternative methods, the statlog dataset showed an astounding 86.29% accuracy in diagnosing heart disease.

8. Srinivas et al. developed a clinical decision support system for the detection of cardiac disease by utilizing machine learning techniques. The technique is based on three basic algorithms for classification, with testing conducted on benchmark datasets. In order to improve the quality of the data, data preprocessing uses an expanded form of Naïve Bayes to resolve missing values using imputation techniques like mean and mode. Compared to other current frameworks, the suggested framework has the benefit of being able to handle complicated questions in heart disease diagnosis in an effective manner.

9. Chitra and Sinivasagam introduced a model for heart disease diagnosis utilizing supervised learning algorithms and medical records. Employing a Cascaded Neural Network (CNN), the authors achieved classification based on information extracted from patients' medical records. The heart disease classification utilized 13 selected attributes as input for the CNN classifier, and the model's efficiency was evaluated on records from 270 patients, demonstrating effectiveness compared to state-of-the-art methods. The proposed CNN system stands as a valuable tool to aid physicians in the accurate diagnosis of heart disease.

10. Helmy et al. presented a machine learning-based decision support system for diagnosing cardiac disease in their study. Individual classifiers are used in the approach for training, and the Bagging algorithm is then used to aggregate the outcomes. Analysis of the results showed that, in comparison to alternative ensemble approaches and single classifiers, the fusion of heterogeneous classifiers using Bagging provided more effective results. Even if the method shows encouraging results, more performance research using a variety of benchmark and real-time datasets is advised.

Proposed Methodology :

In the context of diagnosing cardiac illness, we present a thorough approach in this study that makes use of cutting-edge methods for model selection, interpretability, and comparative analysis. Our method is based on the Automated Optimal Model Selection pipeline. This pipeline is designed to optimize a wide variety of neural network topologies by methodically combining Grid Search, Random Search, cross-validation, and hyperparameter tuning. By finding the best possible configurations, this thorough investigation seeks to improve the models' performance and guarantee a reliable and effective prediction framework. Our Interpretable Deep Learning method brings transparency to the diagnosis of heart disease. To provide comprehensible and unambiguous predictions, this combines LIME (Local Interpretable Model-agnostic Explanations) with attention processes. In order to ensure model interpretability, we use XGBoost combined with LIME in conjunction with Convolutional and Recurrent Neural Networks with attention processes. The goal of this hybrid method is to give understandable insights into the model's decision-making process in addition to precise forecasts. We do a Comparative Analysis of Ensemble Models to thoroughly evaluate the effectiveness of our suggested models. In this examination, deep learning techniques and conventional ensembles are carefully compared. We assess many datasets' worth of performance indicators, including sensitivity, specificity, accuracy, and more. The aim of this study is to clarify the advantages and disadvantages of each method, offering a comprehensive comprehension of their respective benefits within the particular framework of diagnosing heart disease. Our suggested technique aims to provide a comprehensive and perceptive approach to heart disease detection by integrating automated model selection, interpretable deep learning, and a rigorous comparative examination of ensemble models. This multimodal approach seeks to provide transparency and interpretability—two essentials for practical healthcare applications—while pushing the boundaries of predictive modeling.

Data Collection :

We carefully considered each dataset, obtaining information from the UCI Data Repository as well as Data World. After first analyzing four datasets (called "Cleveland," "Hungarian," "Switzerland," and "Long Beach"), we combined their characteristics. By means of group discussions, we were able to identify and choose two datasets that had fourteen important factors. Finalized variables include age, sex, type of chest pain (cp), maximum heart rate, exercise-induced angina, oldpeak, slope, ca (number of major vessels colored by fluoroscopy), cholesterol levels, fasting blood sugar, resting electrocardiographic results, and thal (thalassemia). The adoption of the best dataset for our continuing analysis and research projects is ensured by this cooperative team effort.

Data Preprocessing :

On the chosen datasets, our team used careful preprocessing methods for maximum accuracy.

To deal with missing values, imputation was used, guaranteeing a complete dataset for analysis. To scale numerical characteristics, standardization was applied in order to encourage consistency and make robust model training easier. Furthermore, the robust scaler was utilized to reduce the impact of outliers, improving the data's resilience against extreme values. All of these methods worked together to produce a clean, well-organized dataset that would enhance the performance of the model. The meticulous amalgamation of imputation, standardization, and resilient scaling conforms to optimal methodologies in data preparation, therefore establishing a foundation for dependable and precise analytical results in our research undertaking.

Logistic Algorithm :

When there are two classes in a categorical outcome variable, a statistical technique called logistic regression is employed to solve binary classification issues. Logistic regression predicts the likelihood that an input falls into a certain category, as opposed to linear regression, which is utilized for continuous outcomes.

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

K Nearest Algorithm :

Based on a data point's closeness to the closest group of points, the K-Nearest Neighbor (KNN) classification method classifies the data point. The Euclidean distance formula, which is the square root of the sum of squared differences between the corresponding coordinates of two points in an n-dimensional space, is used to calculate this closeness. KNN essentially uses the least Euclidean distance between a data point and any points in the nearest group to classify the data point. KNN is a simple yet efficient method for pattern recognition and classification applications as the classification is decided by taking into account the closest neighbors and their related classes.

$$\mathrm{d}\,(\mathrm{p},\mathrm{q})=\sqrt{\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{qi}-\mathrm{pi})\,2}$$

Gaussian Process :

Using properties inherited from the normal distribution, a Gaussian process is a random process defined by a set of variables indexed by time or space, which is a generalized form of the Gaussian probability distribution. Unlike the Gaussian probability distribution, which sums up random variables, a Gaussian process summarizes the properties of a function, making it particularly useful in statistical modeling. The Gaussian process is defined by a specific function that calculates its characteristics, providing a potent tool for identifying and comprehending complex dependencies in data.

$$\mathrm{f}\,(\mathrm{x})=\mathrm{a.exp}((\mathrm{x}-\mathrm{y})^2/2\mathrm{c}^2)$$

SVM :

Support Vector Machines (SVM) is the name of the classification algorithm that is discussed and is used to find classes in a given dataset. SVM creates a model that may be used to classify new data points once it has been trained using pre-classified data into two groups. SVM seeks to provide a distinct distinction by maximizing the margins between classes on a plot, going beyond simple classification. In this approach, support vectors are essential for identifying the critical spots that determine the decision limits and improve the classification model.

$$W_0^Tx+b_0=1orW_0^Tx+b_0=-1$$

Decision Tree :

When dealing with datasets that include a large number of attributes, some of which may be less important, a decision tree is an invaluable tool for pattern discovery and data cleansing. A criterion for determining whether qualities are unimportant or only marginally significant is the Gini Index. When building a decision tree, characteristics with low Gini Index are picked to produce rules, and nodes are selected based on the computation of conditional probabilities. Applications for this approach may be found in bioinformatics, particularly in the diagnosis and prognosis of illness. A formula is used to construct the Gini Index, which measures a dataset's impurity.

$$G = \sum_{I=1}^{C} p\,(i) * (1 - p\,(i))$$

Naïve Bayes :

The Naïve Bayes classifier operates under the assumption that the presence of a specific feature in a class is independent of the presence of any other feature. This classifier relies solely on the class feature, considering other class features as unrelated to each other [6]. The calculation of this classifier is based on a specific formula that leverages the independence assumption to make efficient and straightforward probabilistic predictions.

$$P(C_K|X) = \frac{P(C_K) * P(X|C_K)}{P(X)}$$

QDA :

For non-linear data separation, the Quadratic Discriminant Analysis (QDA) classifier is used. It fits data to class conditional densities and establishes a quadratic decision boundary using Bayes' rule. The main difference between QDA and Linear Discriminant Analysis (LDA) is that classes in QDA do not always have the same covariance matrix. When it comes to covariance matrix requirements, QDA is more flexible than LDA and shows better data-fitting capabilities [8]. Certain formulae designed for the quadratic decision boundary method are required to calculate QDA.

$$\mathrm{f_K(X)} = \frac{1}{2} \log \left| \sum \mathrm{K} \right| - (\mathrm{X} - \mathrm{U_K})^{\mathrm{T}} \sum\nolimits_{\mathrm{K}}^{-1} (\mathrm{X} - \mathrm{U_K}) + \log \pi_{\mathrm{K}}$$

Ada Boost :

Adaptive Boosting, or AdaBoost, is an ensemble learning method applied to regression and classification problems. By combining the outputs from weak learners—usually straightforward decision trees or stumps—it builds a robust model. AdaBoost uses a weighted process for every data point, and iteratively modifies the weights to emphasize cases that were previously misclassified. This iterative approach makes sure that the algorithm prioritizes fixing incorrect classifications in later iterations, which enables later weak learners to train on these difficult examples more successfully.

Bagging Classifier :

Bagging, also known as Bootstrap Aggregation, is a group machine learning technique that combines the outputs of many learners to improve model performance. Bagging is a common technique used to reduce variation in noisy data. It entails choosing random chunks of training data. After being separately trained on a variety of machine learning models, these subgroups' predictions are aggregated to get an overall forecast. Although Bagging has shown more efficacy in classification problems, it may also be used to regression methods. It is mostly utilized for classification tasks such as Decision Trees and Naïve Bayes. To calculate it, one must use a certain equation.

$$\mathrm{f_{bag}} = \mathrm{f_1\,(x)} + \mathrm{f_2\,(x)} + \mathrm{f_3\,(x)} + \ldots\ldots\ldots\ldots \mathrm{f_n\,(x)}$$

Boosting :

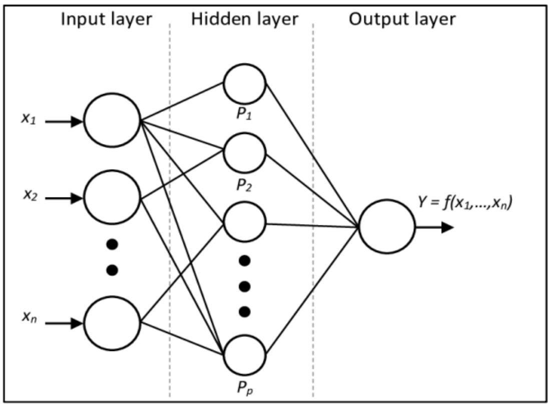
Boosting is an ensemble machine learning strategy that combines many weaker models—often constructed with decision trees—to produce a robust model. It is especially well-known for its efficient dataset classification, which uses a weighted minimization method similar to Adaboost. Weighted inputs and classifiers are regenerated when weights are minimized. As the difference between real and forecasted values, loss is the main objective of boosting [10]. An exact equation must be used in order to calculate boosting.

$$F_b\,(x_i) = F_{b-1}\,(x_i) - \gamma - \times \nabla L(y_i, F\,(x_i))$$

Dense Neural Network :

An Artificial Neural Network (ANN) having one or more hidden layers, one input layer, and one output layer is called a Dense Neural Network (DNN). DNN components, which include neurons, weights, biases, and activation functions, simulate the workings of the human brain and are used to train machine learning algorithms . DNN neurons, which are modeled like real neurons, fire in response to particular stimuli and carry out predetermined tasks. Every neuron adds a continuous bias to a dot product of random weights and various inputs. An activation function is used to the output to determine how well the network predicts particular patterns. Weights are adjusted algorithmically if they are found to be erroneous . A generalized function is applied throughout the computation.

$$F_i = f(\sum_{i=1}^{n} x_i w_i + b)$$



Architecture of Dense Neural Network

Experimental Evolutions and Measures:

Several performance indicators are used to assess the suggested framework in order to demonstrate the efficacy of the activity.

Algorithm for Proposed Methodology

Accuracy :

The ratio of a classification model's correct evaluations to all evaluations performed on the test data is known as accuracy.

$$Accuracy = \frac{Truenegatives + Truepositives}{positives + negatives}$$

Sensitivity :

the proportion of accurate evaluations that a categorization model labels as accurate. This formula is used to compute it.

$$Sensitivity = \frac{Truepositives}{Falsepositives + Truepositives}$$

Specificity :

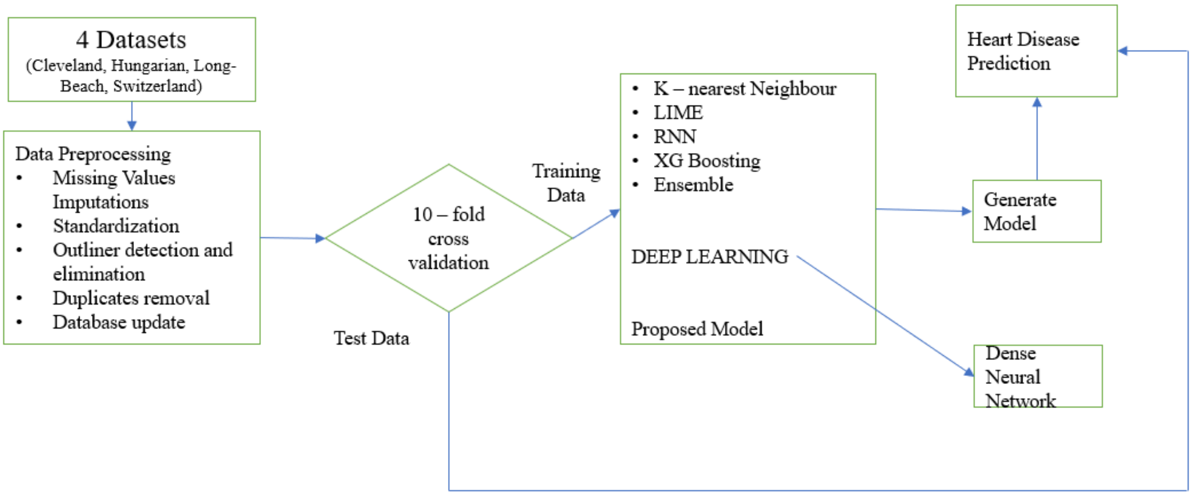
the proportion of incorrect evaluations that a categorization model flags as incorrect. This formula is used to compute it.

$$Specificity = \frac{Truenegatives}{Falsenegatives + Truenegatives}$$

F – Measure :

F-Measure is the ratio between sensitivity and specificity.

$$F - Measure = \frac{2 * Sensitivity * Specificity}{Sensitivity + Specificity}$$



Analysis and Result :

The results are obtained through the use of Jupiter Notebook, a tool that allows Python code to be run in a web browser. Users of Jupiter notebook get access to 12GB of RAM for running programming. The suggested approach starts by evaluating the output from each classifier separately. Next, the best-performing individual classifiers are selected and tested on benchmark datasets related to heart disease.

The heart disease datasets used in this work come from the reliable UCI machine learning data repository, which is highly recommended by several researchers. Cleveland and Hungarian are the two standard datasets for heart disease that are used. Every dataset is distinguished by particular characteristics that are utilized to determine whether individuals have heart disease or not. The output feature is represented by the label attribute (num), which has categorical values that indicate who is well and who is ill. In the analysis, values that are healthy are represented by 1, and those that are sick by 0.

Cleaveland Dataset :

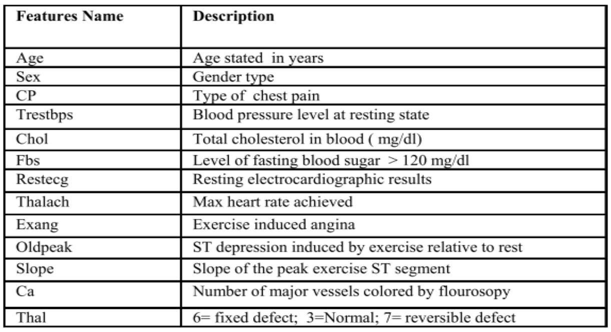
There are fourteen features in the Cleveland dataset. The 13 characteristics are independent of one another and are input features. The final column represents the output feature, which is essentially a label attribute. It is dependent upon the input characteristics; for example, if it is 0, it indicates that the patient has heart disease; if not, it does not.

There are 303 cases in the collection.

Hungarian Dataset :

There are 294 cases and 12 characteristics in the Hungarian dataset.

While the "num" output feature is reliant on input characteristics, the 13 input features are independent of one another. There are four number values in the label property "num" (0–4). The patient has cardiac disease if the readings are zero; otherwise, they are not.



Features Name	Description
Age	Age stated in years
Sex	Gender type
CP	Type of chest pain
Trestbps	Blood pressure level at resting state
Chol	Total cholesterol in blood (mg/dl)
Fbs	Level of fasting blood sugar > 120 mg/dl
Restecg	Resting electrocardiographic results
Thalach	Max heart rate achieved
Exang	Exercise induced angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
Ca	Number of major vessels colored by flourosopy
Thal	6= fixed defect; 3=Normal; 7= reversible defect

PIDD - Description

Sno	Attributes	Description	Data Type	Range
1	Age	Age stated in years	Numeric	[Correct age]
2	Sex	Gender type	Numeric	[0 & 1]
3	CP	Types of chest Pain	Numeric	[1-4]
4	Trestbps	Blood pressure level at resting state	Numeric	[94-200]
5	Chol	Total cholesterol in blood (mg/dl)	Numeric	[126-564]
6	Fbs	Level of fasting blood sugar > 120 mg/dl	Numeric	[1 & 0]
7	Restecg	Resting electrocardiographic results	Numeric	[0-2]
8	Thalach	Max heart Rate achieved	Numeric	[71-202]
9	Exang	Exercise induced angina	Numeric	[0 & 1]
10	Oldpeak	ST depression induced by exercise relative to rest	Numeric	[0-6.2]
11	Slope	Slope of the peak exercise relative to rest	Numeric	[1-3]
12	Ca	Number of major vessels colored by flourosopy	Numeric	[0-3]
13	Thal	Types of defect	Numeric	3 = normal 6 = fixed 7 = reversible

Cleaveland statistical Summary

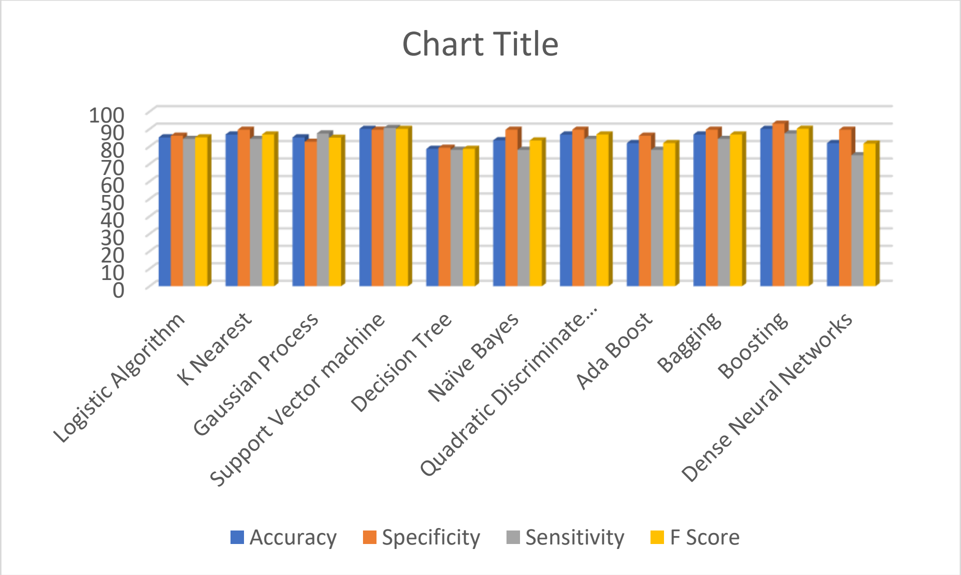
	age	sex	cp	Trestbps	cholestrol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
Count	303	303	303	303	303	303	303	303	303	303	303	303	303	303
Mean	54.4	0.67	3.15	131.68	246.69	0.14	0.99	149.6	0.32	1.03	1.6	0.69	4.72	0.45
Std	9.03	0.46	0.96	17.59	51.77	0.35	0.99	22.87	0.46	1.16	0.61	0.96	1.93	0.49
Min	29	0	1	94	126	0	0	71	0	0	1	0	3	0
25 %	48	0	3	120	211	0	0	135	0	0	1	0	3	0
50 %	56	1	3	130	241	0	1	153	0	0.80	2	0	3	0
75 %	61	1	4	140	275	0	2	166	1	1.6	2	1	7	1
Max	77	1	4	200	564	1	2	202	1	6.2	3	3	7	1

Hungarian Statistical Summary

	age	sex	cp	trestbps	cholestrol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
Count	294	294	294	294	294	294	294	294	294	294	294	294	294	294
Mean	47.82	0.72	2.98	132.13	231.22	0.06	0.21	138.65	0.30	0.58	0.67	0.001	0.53	0.36
Std	7.81	0.44	0.96	19.22	93.65	0.25	0.46	24.90	0.46	0.90	0.92	0.02	1.72	0.48
Min	28.0	0	1	0	0	0	0	0	0	0	0	0	0	0
25 %	42	0	2	120	198	0	0	122	0	0	0	0	0	0
50 %	49	1	3	130	237	0	0	140	0	0	0	0	0	0
75 %	54	1	4	140	277	0	0	155	1	.1	2	0	0	1
Max	66	1	4	200	603	1	2.1	190	1	5	3	0.4	7	1

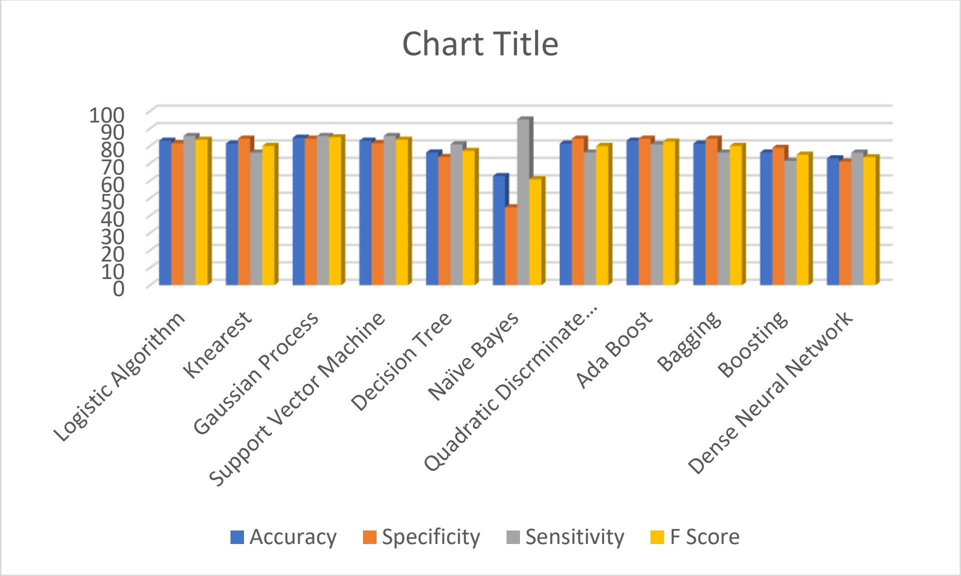
Cleaveland

	Accuracy	Specificity	Sensitivity	F Score	Exe time
Logistic Algorithm	85.25	86.21	84.38	85.28	0.0188
K Nearest	86.89	89.66	84.38	86.93	0.0220
Gaussian Process	85.25	82.76	87.50	85.06	0.7899
Support Vector machine	90.16	89.66	90.62	90.14	0.0274
Decision Tree	78.69	79.31	78.12	78.71	0.0078
Naïve Bayes	83.61	89.66	78.12	83.49	0.0106
Quadratic Discriminate Analysis	86.89	89.66	84.38	86.93	0.0302
Ada Boost	81.97	86.21	78.12	81.97	0.0625
Bagging	86.89	89.66	84.38	86.93	0.1115
Boosting	90.16	93.10	87.50	90.21	0.0560
Dense Neural Networks	81.97	89.66	75	81.68	1.1631



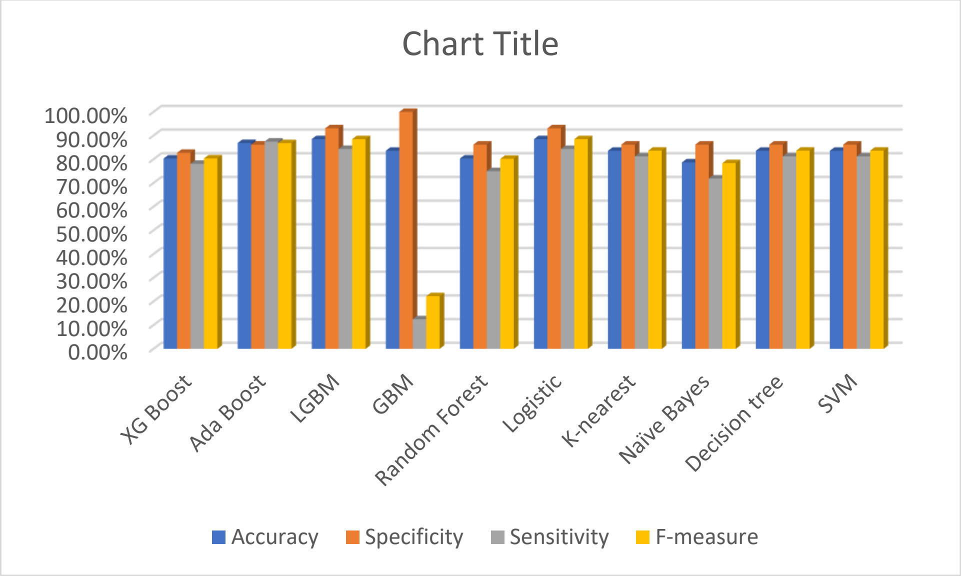
Hungarian

	Accuracy	Specificity	Sensitivity	F Score	Exe time
Logistic Algorithm	83.05	81.58	85.71	83.6	0.0198
Knearest	81.36	84.21	76.19	80	0.0309
Gaussian Process	84.75	84.21	85.71	84.96	0.4587
Support Vector Machine	83.05	81.58	85.71	83.60	0.0238
Decision Tree	76.27	73.68	80.95	77.15	0.0061
Naïve Bayes	62.71	44.74	95.24	60.88	0.0869
Quadratic Discrminate Analysis	81.36	84.21	76.19	80	0.1135
Ada Boost	83.05	84.21	80.95	82.55	0.0807
Bagging	81.36	84.21	76.19	80	0.1216
Boosting	76.27	78.95	71.43	75	0.0489
Dense Neural Network	72.88	71.05	76.19	73.53	1.1340

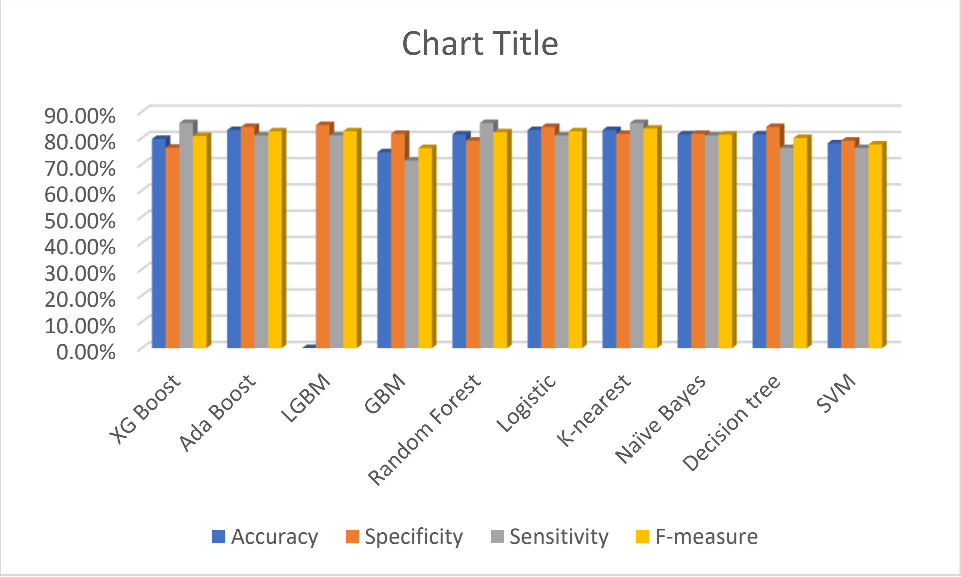


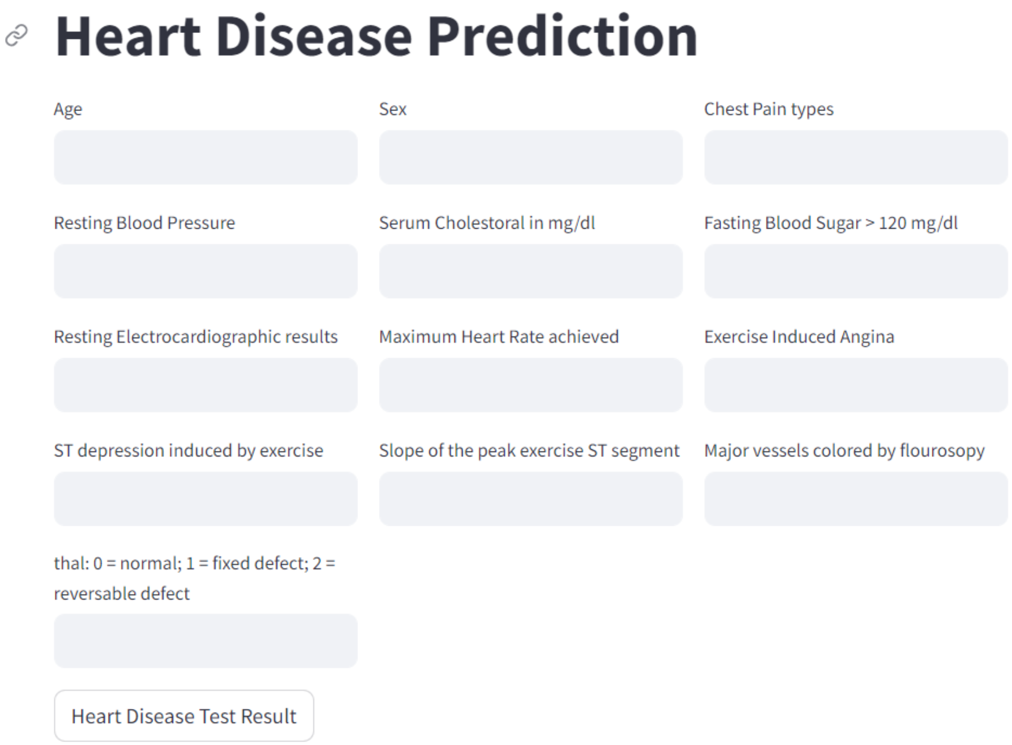
Cleveland	Accuracy	Specificity	Sensitivity	F-measure	Execution time
XG Boost	80.33%	82.76%	78.12%	80.38%	28.7102
Ada Boost	86.89%	86.21%	87.50%	86.85%	27.3439
LGBM	88.52%	93.10%	84.38%	88.52%	0.0578
GBM	83.61%	100%	12.50%	22.22%	29.0419
Random Forest	80.33%	86.21%	75%	80.21%	28.5142
Logistic	88.52%	93.10%	84.38%	88.52%	29.3103
K-nearest	83.61%	86.21%	81.25%	83.66%	28.1663

Naïve Bayes	78.69%	86.21%	71.88%	78.39%	27.6944
Decision tree	83.61%	86.21%	81.25%	83.66%	27.7524
SVM	83.61%	86.21%	81.25%	83.66%	30.2198



Hungarian	Accuracy	Specificity	Sensitivity	F-measure	Execution time
XG Boost	79.66%	76.32%	85.71%	80.74%	27.4996
Ada Boost	83.05%	84.21%	80.95%	82.55%	28.7462
LGBM	83,05%	84.95%	80.95%	82.55%	0.0439
GBM	74.58%	81.58%	71.43%	76.17%	27.1836
Random Forest	81.36%	78.95%	85.71%	82.19%	27.4080
Logistic	83.05%	84.21%	80.95%	82.55%	28.0320
K-nearest	83.05%	81.58%	85.71%	83.60%	28.4060
Naïve Bayes	81.36%	81.58%	80.95%	81.26%	28.6789
Decision tree	81.36%	84.21%	76.19%	80%	26.6422
SVM	77.97%	78.97%	76.19%	77.54%	26.6306





Heart Disease Prediction

Age

Sex

Chest Pain types

Resting Blood Pressure

Serum Cholestoral in mg/dl

Fasting Blood Sugar > 120 mg/dl

Resting Electrocardiographic results

Maximum Heart Rate achieved

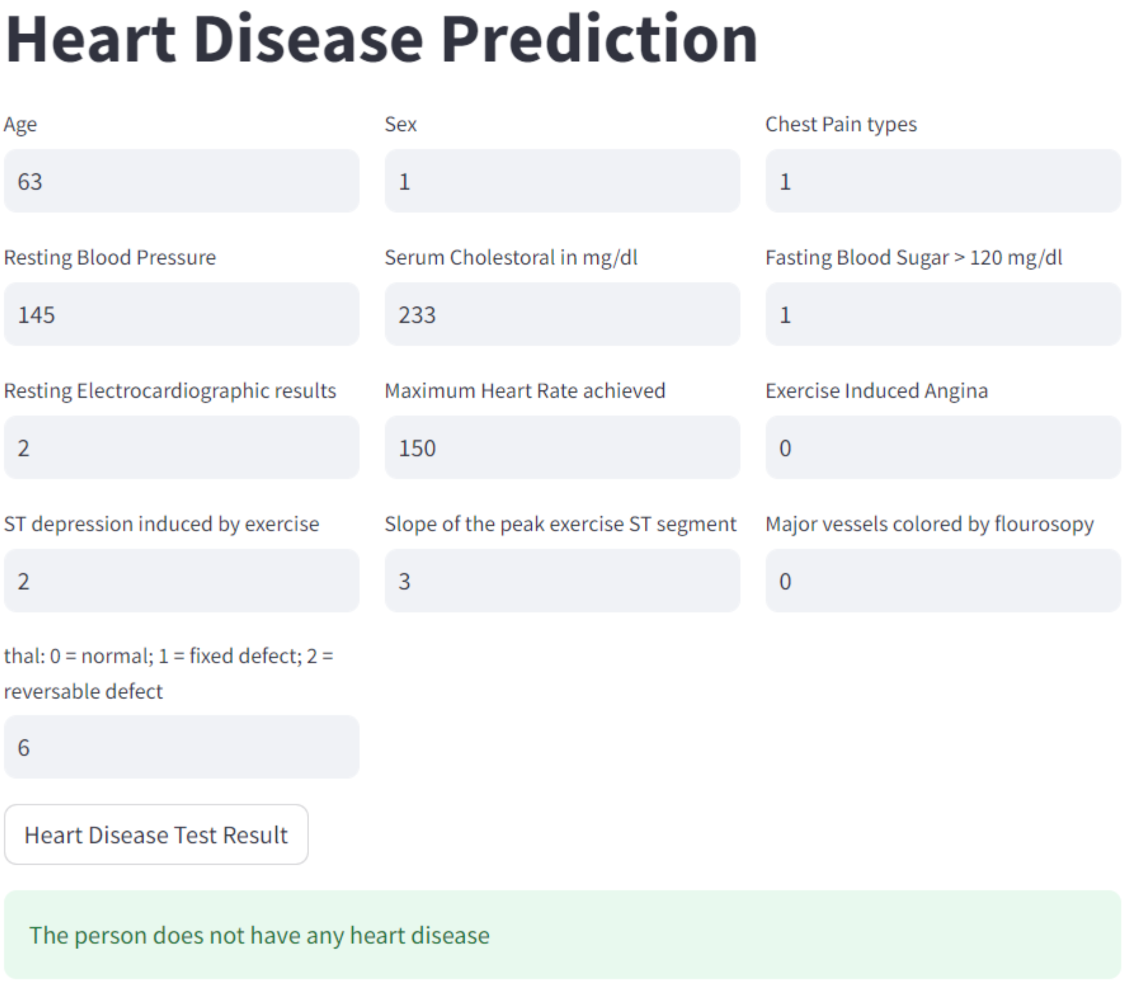
Exercise Induced Angina

ST depression induced by exercise

Slope of the peak exercise ST segment

Major vessels colored by flourosopy

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect



Heart Disease Prediction

Age: 63

Sex: 1

Chest Pain types: 1

Resting Blood Pressure: 145

Serum Cholestoral in mg/dl: 233

Fasting Blood Sugar > 120 mg/dl: 1

Resting Electrocardiographic results: 2

Maximum Heart Rate achieved: 150

Exercise Induced Angina: 0

ST depression induced by exercise: 2

Slope of the peak exercise ST segment: 3

Major vessels colored by flourosopy: 0

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect: 6

The person does not have any heart disease

Conclution :

In conclusion, our study leveraged machine learning and 10-fold cross-validation to significantly enhance heart disease prediction. Successful data scaling and null value handling during preparation set a strong foundation. The diverse deep neural network models explored, such as KNN, Decision Trees, and SVM, yielded distinct results. Our findings hold promise for revolutionizing healthcare analytics, enabling early diagnosis, personalized treatment recommendations, and resource optimization. Furthermore, our research contributes to medical advancements, evaluating new heart disease medicines and predicting their efficacy in advance. The significance of sophisticated preprocessing, ensemble learning, and machine learning methodologies underscore their pivotal role in improving patient care, fostering medical research, and advancing healthcare analytics for the benefit of researchers, data scientists, and healthcare professionals alike.

Reference :

[1]. N.-S. Tomov and S. Tomov, ‘‘On deep neural networks for detecting heart disease,’’ 2018, arXiv:1808.07168.

[2]. A. Kumar, P. Kumar, A. Srivastava, V. D. A. Kumar, K. Vengatesan, and A. Singhal, ‘‘Comparative analysis of data mining techniques to predict heart disease for diabetic patients,’’ in Proc. Int. Conf. Adv. Comput. Data Sci. Singapore: Springer, 2020.

[3]. C. Sowmiya and P. Sumitra, ‘‘Analytical study of heart disease diagnosis using classification techniques,’’ in Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS), Mar. 2017.

[4]. Y -J. Huang, M. Parry, Y. Zeng, Y. Luo, J. Yang, and G.-P. He, ‘‘Examination of a nurse-led community-based education and coaching intervention for coronary heart disease high-risk individuals in China,’’ Asian Nursing Res., vol. 11, no. 3, pp. 187–193, Sep. 2017.

[5]. R. Hasan, ‘‘Comparative analysis of machine learning algorithms for heart disease prediction,’’ in Proc. ITM Web Conf., vol. 40, 2021.

[6]. S. Khan and S. T. Rasool, ‘‘Current use of cardiac biomarkers in various heart conditions,’’ Endocrine, Metabolic Immune Disorders-Drug Targets, vol. 21, no. 6, pp. 980–993, Jun. 2021.

[7]. S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, ‘‘A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction,’’ IEEE Access, vol. 9, pp. 130805–130822, 2021.

[8]. S. Chae, S. Kwon, and D. Lee, ‘‘Predicting infectious disease using deep learning and big data,’’ Int. J. Environ. Res. Public Health, vol. 15, no. 8, p. 1596, Jul. 2018.

[9]. M. Raju, V. Pagidimarri, R. Barreto, A. Kadam, V. Kasivajjala, and A. Aswath, ‘‘Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy,’’ in Proc. MEDINFO, 2017, pp. 559–563.

[10]. G. T. Reddy and N. Khare, ‘‘An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model,’’ J. Circuits, Syst. Comput., vol. 26, no. 4, Apr. 2017, Art. no. 1750061.

[11]. N. Khateeb and M. Usman, ''Efficient heart disease prediction system using K-nearest neighbor classification technique,'' in Proc. Int. Conf. Big Data Internet Thing, Dec. 2017, pp. 21–26.

[12]. L. Verma, S. Srivastava, and P. C. Negi, ''A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data,'' J. Med. Syst., vol. 40, no. 7, p. 178, Jul. 2016.

[13]. A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, ''HDPS: Heart disease prediction system,'' in Proc. Comput. Cardiol., Sep. 2011, pp. 557–560.

[14]. M. A. Jabbar and S. Samreen, ''Heart disease prediction system based on hidden naïve Bayes classifier,'' in Proc. Int. Conf. Circuits, Controls, Commun. Comput. (I4C), Oct. 2016, pp. 1–5.

[15]. K. Srinivas, B. K. Rani, and A. Govrdhan, ''Applications of data mining techniques in healthcare and prediction of heart attacks,'' Int. J. Data Mining Techn. Appl., vol. 7, no. 1, pp. 172–176, Mar. 2018.

[16]. R. Chitra and V. Seenivasagam, ''Heart attack prediction system using cascaded neural network,'' in Proc. Int. Conf. Appl. Math. Theor. Comput. Sci., 2013, p. 223.

[17]. T. Helmy, S. M. Rahman, M. I. Hossain, and A. Abdelraheem, ''Nonlinear heterogeneous ensemble model for permeability prediction of oil reservoirs,'' Arabian J. Sci. Eng., vol. 38, no. 6, pp. 1379–1395, Jun. 2013.

[18]. G. Alfian, M. Syafrudin, and J. Rhee, ''Real-time monitoring system using smartphone-based sensors and NoSQL database for perishable supply chain,'' Sustainability, vol. 9, no. 11, p. 2073, Nov. 2017, doi: 10.3390/su9112073.

[19]. M. Syafrudin, G. Alfian, N. Fitriyani, and J. Rhee, ''Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing,'' Sensors, vol. 18, no. 9, p. 2946, Sep. 2018, doi: 10.3390/s18092946.

[20]. M. Syafrudin, N. Fitriyani, G. Alfian, and J. Rhee, ''An affordable fast early warning system for edge computing in assembly line,'' Appl. Sci., vol. 9, no. 1, p. 84, Dec. 2018, doi: 10.3390/app9010084.