**Sustainable IoHT-based Machine-Learning-Modelled System for Prediction of Cardiovascular Disease Risk**

Ashish K. Biswal1,\*, Akash Nair2, Sushruta K. Mishra3

School of Computer Engineering, KIIT Deemed to be University, Bhubaneshwar, India

E-mail: 1ashishbiswal02@gmail.com, 2mailtoakashnair@gmail.com,

3sushrutamishrafcs@kiit.ac.in

**Abstract: This paper presents development of a sustainable cardiovascular-disease-risk-detection system / model by integrating Internet of Health Things (IoHT) and various machine learning models / algorithms. Logistic Regression, Decision Tree and Random Forest models are employed and evaluated using accuracy, confusion matrix, and classification report with receiver operating characteristic curves to assess the performance of the models. The results indicate that the Random Forest model achieves the highest accuracy of 88.36%, outperforming the other models. In addition, data visualization has been conducted to gain insights into the relationships between different variables and their impact on heart disease risk. Further, the education level of individuals seems to have a little correlation with their heart disease risk. The developed model can aid in early intervention and prevention strategies i. e. assisting healthcare professionals in identifying individuals at higher risk of heart disease and implementing targeted interventions**.

**Keywords: Machine learning, Internet of Health Things, Random Forest, Cardiovascular Disease, Sustainable Healthcare**

1. Introduction

Machine learning models, powered by computational intelligence, have shown great potential in predicting and assessing the risk of several chronic diseases based on blood profiles and other health metrics. These models can analyze large volumes of data and identify patterns, correlations, and risk factors associated with the diseases. By leveraging machine learning algorithms, healthcare professionals can make informed decisions and implement targeted interventions to prevent and manage cardiovascular diseases effectively. This can be done collecting data either online or offline. In this context, Internet of Health Things (IoHT) plays a crucial role in sustainable healthcare by connecting various medical devices, wearable, and sensors to collect real-time health data. By amalgamating smart devices, wearable and data analytics, it establishes an intricate healthcare network capable of real-time monitoring of vital signs, health parameters, and environmental factors. Within this ecosystem, sustainability emerges as a central tenet, ensuring not only the effectiveness of healthcare but also its environmental viability. This interconnected network enables continuous monitoring and analysis of vital health metrics, providing valuable insights into an individual's health status and potential disease risks. In this way, the synergy of sustainability, the IoHT and the prowess of machine learning have emerged as a transformative paradigm in the ever-evolving landscape of healthcare (Fig. 1) [1].

|  |  |
| --- | --- |
| **Fig. 1** A Venn diagram illustrating the intersection of healthcare, technology and sustainability infrastructure | **Fig. 2** Schematic diagram showcasing the interconnectivity of IoHT devices, wearable sensors, and healthcare |

Cardiovascular diseases (CVDs) remain a global health challenge of unparalleled magnitude, accounting for an alarming number of deaths and healthcare costs. Therefore, we must not only better understand the multifaceted risk factors that contribute to CVDs, but also develop innovative strategies to predict, prevent and manage these diseases with the utmost precision. However, at present, there are several issues such as inadequate predictive decision, resource-intensive healthcare systems, socio-economic disparities coexist making things more complicated and challenging. Traditional risk factors such as age, gender, and cholesterol levels, while important, do not provide the granularity required to offer precise, patient-specific predictions. The existing models lack the ability to incorporate real-time data streams from the IoHT, which can offer a dynamic, moment-by-moment assessment of a patient's health. Further, the global healthcare system's environmental impact is often overlooked. From energy-hungry hospital facilities to the growing pile of electronic waste generated by obsolete medical devices, the healthcare industry contributes significantly to carbon emissions. Healthcare disparities persist, with underserved communities bearing a disproportionate burden of cardiovascular diseases. Therefore, there is an urgent need to address this healthcare disparity, sustainability issue and design healthcare systems that not only improve patient outcomes but also minimize ecological footprints and socioeconomic division.

In view of the above, the present work proposes a sustainable IoHT-based machine-learning-modelled system that aims to integrate the principles of sustainability, the capabilities of IoHT, and the predictive power of machine learning algorithms and develop an efficient and effective system for predicting CVDs risk. By utilizing real-time health data collected through IoHT devices, the system can continuously monitor and analyze an individual's health parameters, identify potential risk factors, and provide personalized risk assessments. This approach not only enhances the accuracy and timeliness of cardiovascular disease risk prediction at an unprecedented level but also minimizes resource consumption and environmental impact to a great extent.

**2. Literature Survey**

Heart disease is the leading cause of death worldwide, accounting for an estimated 17.9 million deaths in 2019. It is a complex disease with multiple risk factors, including age, gender, family history, lifestyle choices, and medical conditions such as high blood pressure, high cholesterol, and diabetes. Machine learning is a field of artificial intelligence that allows computers to learn without being explicitly programmed. Machine learning algorithms can be used to identify patterns in data and make predictions.

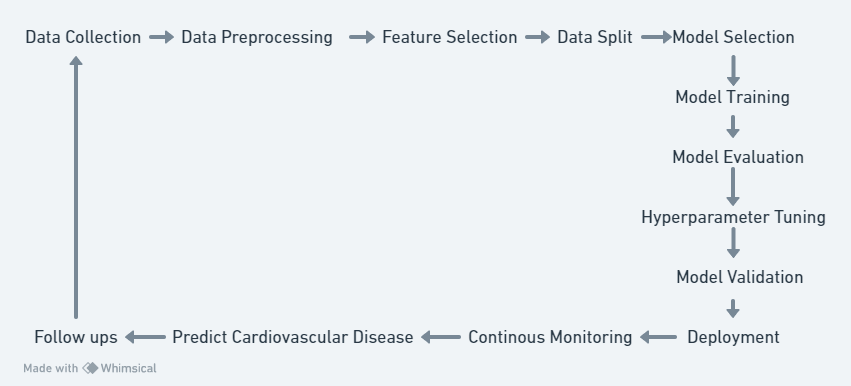
In recent years, machine learning has been increasingly used to predict heart disease. A number of studies have been conducted to evaluate the effectiveness of machine learning algorithms for heart disease prediction. These studies have shown that machine learning algorithms can achieve high accuracy in predicting heart disease and are summarised in table 1. One study found that random forest gave an accuracy of 85.51% in predicting heart disease [2].Another paper used MLP and Cross validation to achieve an accuracy of 87.28% [3].One study concludes that Support Vector Machine (SVM) algorithm achieved accuracy of 84.9% [4] while other study found that the SVM algorithm achieved an accuracy of 85.7% in predicting heart disease [6]. Another study reported that neural networks achieved 85.3% accuracy in one experiment [7,8].This suggests that machine learning algorithms are continuing to improve in their ability to predict heart disease.

**Table 1** Existing works on cardiovascular disease risk prediction using machine learning.

|  |  |  |
| --- | --- | --- |
| **Ref. / Authors** | **Model used** | **Accuracy achieved** |
| Ahire N, Patel R. (2021) | Random forest | 85.51% |
| Bhatt M. C., Patel P. (2023) | MLP and Cross Validation | 87.28% |
| Pal M., Parija S. (2022) | SVM | 85.7% |
| Nissa N., Jamwal S. (2020) | Random forest | 85.5% |
| Parmar M. (2020) | Deep Neural Network | 85.3% |
| Ali J., Das C. B. (2022) | SVM | 84.9% |
| Sarkar K. B. (2022) | Logistic regression and KNN | 84.7% |

**3. Data Handling and Processing**

Data handling and data processing are important part of any machine learning model. Fig. 3 shows a flowchart illustrating how machine learning processes healthcare data to predict cardiovascular disease risk. It starts with data collection and data processing which are followed by feature selection, data split, model selection, model training, model evaluation, hyper parameter tuning, model validation, deployment, continuous monitoring, CVD risk prediction and follow up.

****

**Fig. 3** Flowchart illustrating how machine learning processes healthcare data to predict cardiovascular disease risk.

**3.1 Data collection**

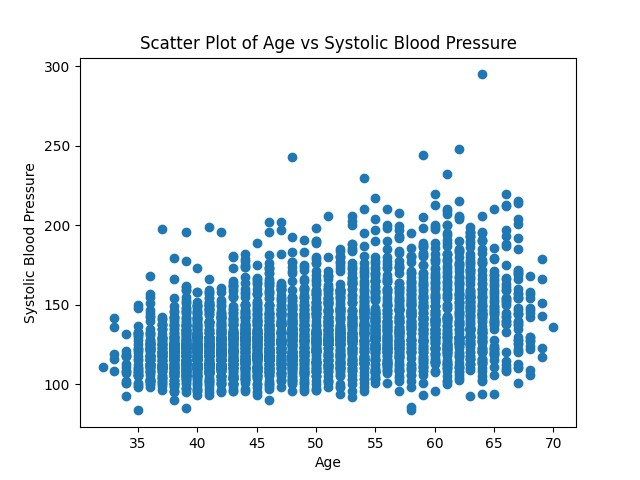
Data collection is the heart of any machine learning model. In the present study, Framingham dataset, which contains information related to cardiovascular disease risk in a .csv comma separated value, have been used. The dataset consists of 4240 instances and 16 attributes and is summarised in table 2.

**Table 2**: Cardiovascular risk predictor dataset

|  |  |  |
| --- | --- | --- |
| **Features** | **Details** | **Range** |
| Male | Indicates the gender of the individual | 0 for female,1 for male |
| Age | Represents the age of the individual. | 32 years-70 years |
| Education | Indicates the level of education of the individual. | 1-4 |
| currentSmoker | Indicates whether the individual is a current smoker | 0 for non smoker,1 for smoker |
| CigsPerDay | Represents the number of cigarettes smoked per day by the individual. | 0-70 |
| BPMeds | Indicates whether the individual is on blood pressure medication | 0 for not on medication, 1 for on medication |
| prevalentStroke | Indicates whether the individual is on blood pressure medication | 0 for no history, 1 for history |
| PrevalentHyp | Indicates whether the individual has prevalent hypertension | 0 for no hypertension, 1 for hypertension |
| diabetes | Indicates whether the individual has diabetes | 0 for no diabetes, 1 for diabetes |
| totChol | Represents the total cholesterol level of the individual. | 107-696 |
| sysBP | Represents the systolic blood pressure of the individual. | 83.5-295 |
| diaBP | Represents the diastolic blood pressure of the individual. | 48-142.5 |
| BMI | Represents the body mass index of the individual. | 15.54-56.8 |
| heartRate | Represents the heart rate of the individual. | 44-143 |
| glucose | Represents the glucose level of the individual. | 40-394 |
| TenYearCHD | Indicates whether the individual developed coronary heart disease within ten years | 0 for no,1 for yes |

**3.2 Data set**

The dataset provides valuable insights into the relationships between these attributes and the risk of cardiovascular disease. For example, age, smoking status, blood pressure, cholesterol levels, and diabetes are known risk factors for cardiovascular disease. Exploring the relationships between these attributes can help in understanding the risk factors and developing predictive models. One significant relationship is between age and systolic blood pressure. A scatter plot of systolic blood pressure versus age (Fig. 4) reveals a positive correlation, indicating that as individual’s age, their blood pressure tends to increase. This relationship is crucial as high blood pressure is a known risk factor for cardiovascular disease.



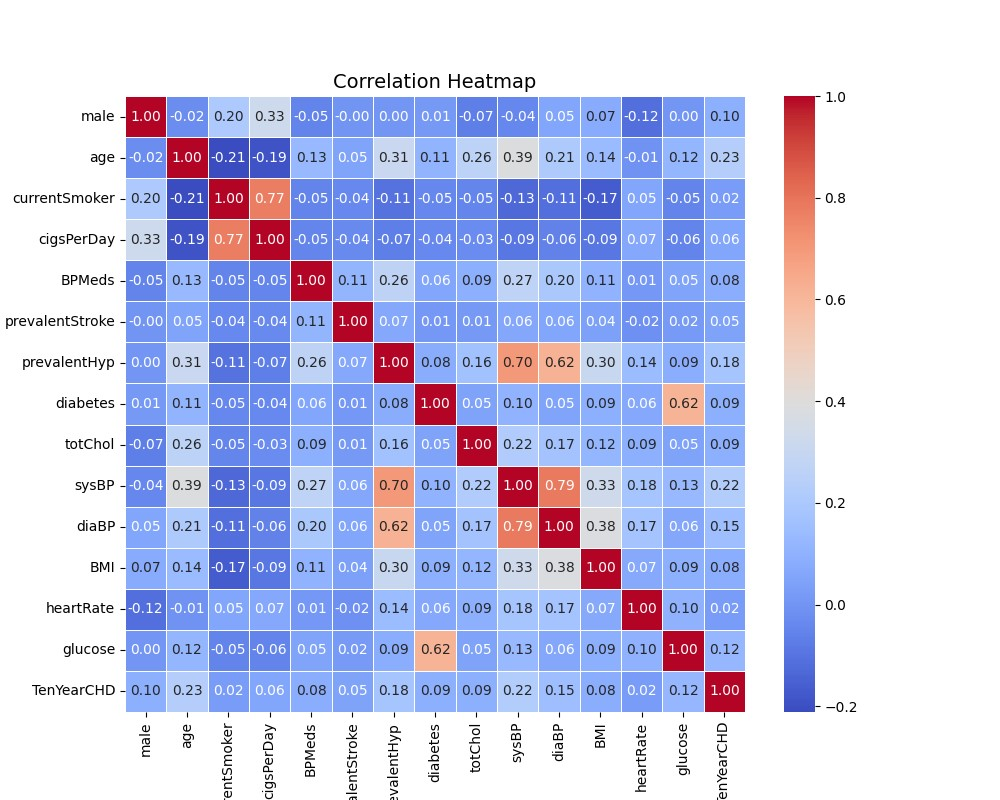
**Fig. 4** Scatter plot of systolic pressure vs. age

Fig. 5 shows the correlation heatmap that provides insights into the relationships between different variables in the Framingham dataset. From the heatmap, it can be observed that,

(i) There is a positive correlation between current smoker status and the number of cigarettes smoked per day (cigsPerDay). This suggests that individuals who are current smokers tend to smoke more cigarettes per day.This indicates that smoking is a well-established risk factor for cardiovascular disease, and this relationship highlights the importance of smoking cessation interventions.

(ii) Total cholesterol levels (totChol) and systolic blood pressure (sysBP) show a positive correlation. This correlation is expected, as high cholesterol levels can contribute to the development of hypertension, which is a significant risk factor for cardiovascular disease.

(iii) There is a positive correlation between diabetes and glucose levels. This relationship indicates that individuals with diabetes tend to have higher glucose

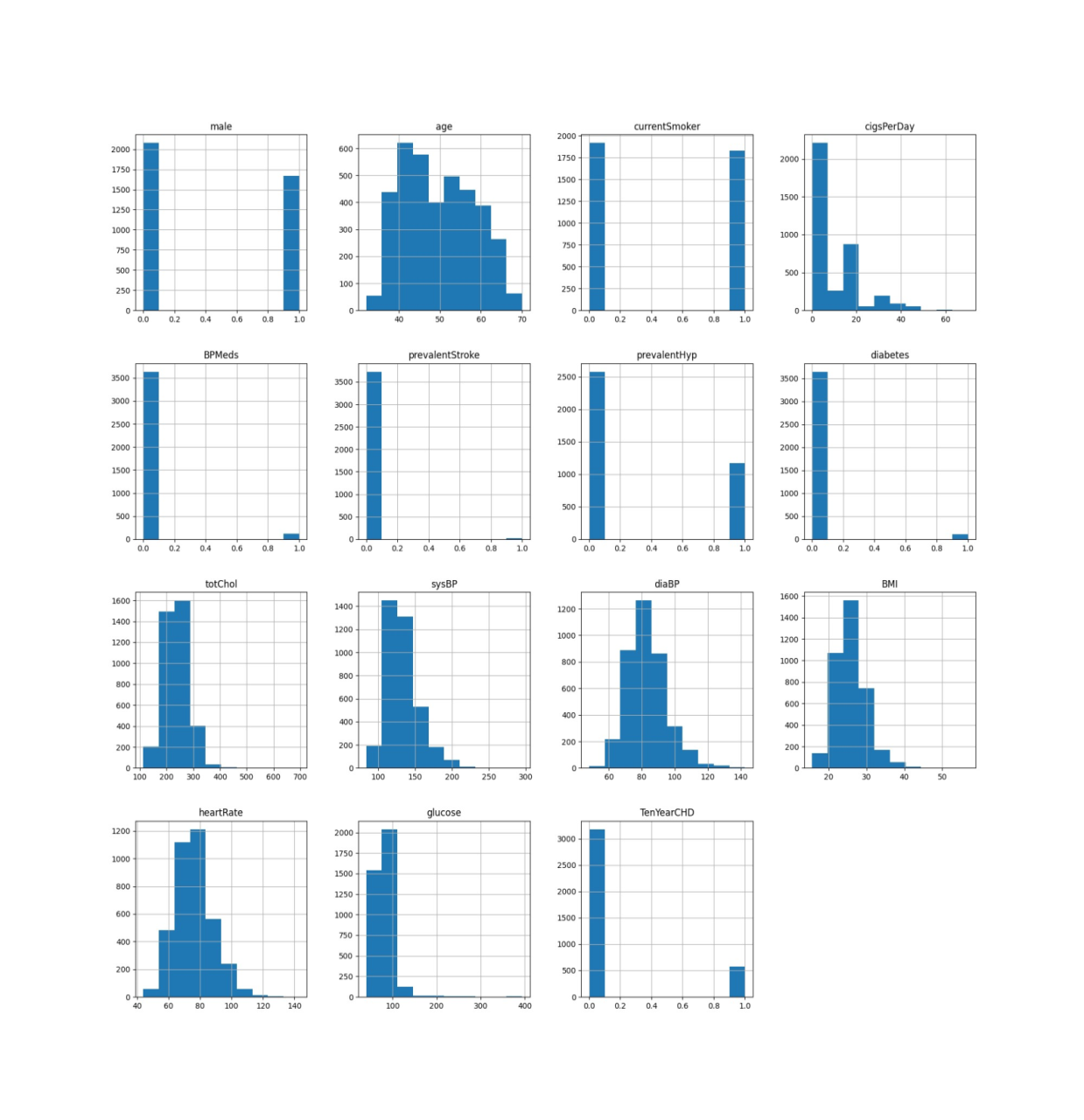


**Fig. 5** Correlation heatmap showing interlinking of different variables

levels. Diabetes is a known risk factor for cardiovascular disease, and this correlation emphasizes the importance of managing blood glucose levels in individuals with diabetes.

Fig. 6 shows the histograms providing insights into the distribution of various attributes in the dataset. From the histogram, it can be observed that,

1. The age histogram shows a relatively normal distribution, with a peak around middle age.



**Fig. 6** Histogram for variables

1. The BMI (Body Mass Index) histogram indicates that the distribution is skewed towards higher BMI values, suggesting a higher prevalence of overweight or obese individuals in the dataset.
2. The glucose histogram shows a right-skewed distribution, indicating that most individuals have glucose levels within a normal range, but there is a tail towards higher glucose levels.
3. The histogram for systolic blood pressure (sysBP) shows a distribution with multiple peaks, suggesting the presence of different blood pressure groups within the dataset.

The target value taken is near to 1 for a person has CVD and close to 0 for a person does not have CVD.

**3.3 Working with the data**

**3.3.1 Data pre-processing**

Data processing is a crucial step in the journey from raw health data to insightful predictions. The integrity of the data and the effectiveness of predictive models depend significantly on the methodologies applied during this processing phase. Missing values are addressed which is a common challenge in health-related datasets, and proceed to discuss feature engineering, scaling, and the crucial task of managing class imbalance. The aim is to elucidate the strategies used to harness the potential of the dataset and facilitate robust predictive modeling [9],[10].

**3.3.1a Handling missing values**

Prior to model development, a systematic approach was applied to handle missing values. Specifically, the following steps were taken:

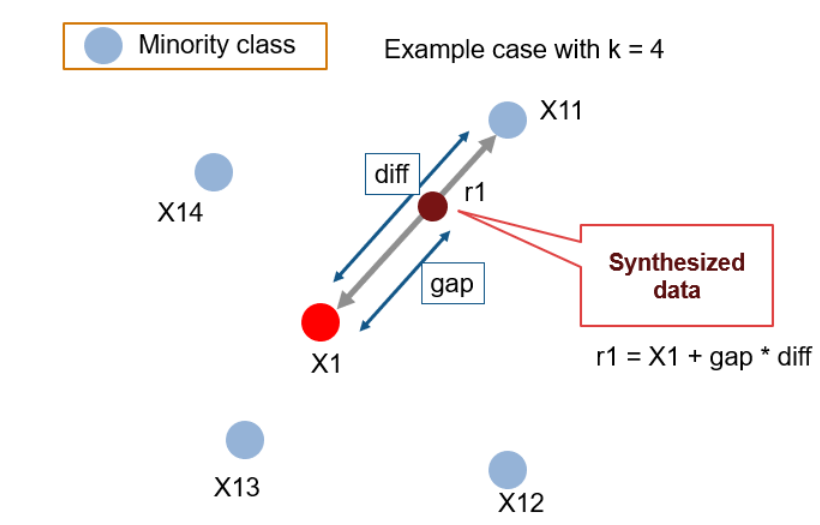
* Missing values in the 'bmi' feature were addressed by calculating the Body Mass Index (BMI) for each individual and filling missing values with the mean BMI.
* Records with missing data in other features were excluded from the analysis to maintain data integrity.

**3.3.1b Handling class imbalance**

Addressing class imbalance is critical for accurate prediction of cardiovascular disease risk. To overcome class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was employed. The 'smote' module from the 'imbalanced-learn' library was used to oversample the minority class, ensuring a balanced dataset. Its primary purpose is to address class imbalance, a situation in which one class (the minority class) is significantly underrepresented in the dataset compared to the other class (the majority class). It deals with imbalanced datasets because it helps to mitigate the issues that can arise due to the unequal distribution of classes.

This technique balances the class distribution by generating synthetic samples for the minority class and is illustrated in Fig. 7. The method works as follows:

* For each minority class instance, SMOTE selects k nearest neighbours from the same class.
* A random neighbour is chosen, and synthetic instances are created by interpolating between the original instance and the chosen neighbour.
* This interpolation is performed along each feature dimension.
* The result is a set of synthetic instances that are added to the minority class, balancing the class distribution.



**Fig.7** SMOTE demonstration

**3.3.2 Data Splitting**

In the quest to develop a robust and reliable predictive model, a critical step involved partitioning the pre-processed dataset into separate training and testing sets. This division is essential to assess the model's performance accurately and to ensure that it can generalize effectively to unseen data [11], [12]. The dataset was split using a widely accepted partitioning strategy known as an 80-20 split, where 80% of the data was allocated for training, and the remaining 20% was earmarked for testing. This strategic allocation offers several advantages:

* Training Set (80%): The training set serves as the foundational cornerstone for model development. It is used to train the Random Forest Classifier by exposing it to a substantial portion of the data, allowing the model to learn the underlying patterns, relationships, and intricacies within the dataset. The model fine-tunes its parameters and decision boundaries based on this training data.
* Testing Set (20%): The testing set, on the other hand, remains untouched during the model development phase. It serves as an isolated evaluation set that is reserved for assessing the model's performance and generalization capabilities. By withholding a portion of the data, we can simulate the real-world scenario where the model encounters new, unseen instances. This testing phase helps identify potential overfitting issues and provides a reliable measure of the model's predictive accuracy.

This division into training and testing sets is a fundamental practice in machine learning and statistics, ensuring that the model's performance is rigorously evaluated while maintaining the integrity of the assessment process. The 80-20 split represents a balanced compromise that allocates a substantial portion of the data for training to allow the model to learn effectively, while still reserving a sizable portion for testing to provide a stringent and reliable measure of performance.

**3.3.3 Feature standardization**

To prepare the dataset for modelling, a critical step involved the standardization of its features. Standardization, also known as feature scaling, is a fundamental pre-processing technique that ensures uniformity in the scale of features. This process is vital, especially for machine learning models like the Random Forest Classifier, which rely on distance-based calculations and may be sensitive to variations in feature scales. The goal of feature standardization is to transform the data in such a way that each feature exhibits a mean (average) value of zero and a standard deviation of one. This transformation centers the data around a common reference point (mean) and scales it consistently (standard deviation), ensuring that no single feature dominates the modelling process simply due to its scale. By standardizing the features, a level playing field has been created for the model to consider all features with equal importance. This process of standardization is mathematically achieved by subtracting the mean of the feature from each data point and then dividing by the standard deviation. The resulting standardized feature has a mean of zero and a standard deviation of one. It helps prevent features with larger numerical values from having a disproportionately large influence on the model's decision-making process. By ensuring that all features are on the same scale, a more stable and effective learning process has been facilitated for the Random Forest Classifier, which combines multiple decision trees to make predictions. Furthermore, it helps the model generalize well to unseen data and aids in avoiding potential issues that can arise when features are measured in different units or have different ranges.

**4. Prediction**

In the present study, the primary machine learning model employed for predicting cardiovascular disease risk is the Random Forest Classifier. The Random Forest algorithm falls under the category of ensemble learning techniques, renowned for their ability to enhance model accuracy and generalization. Random Forest is a robust and versatile ensemble method that harnesses the collective wisdom of multiple decision trees. Each decision tree in the ensemble independently generates predictions, and the final prediction is made by aggregating the individual tree outputs. Some key features of Random forest are:

* Decision Tree Ensemble: A Random Forest comprises a multitude of decision trees. Each tree is constructed based on a random subset of the training data (bootstrapping) and a random subset of features (feature bagging).
* Strength in Diversity: The ensemble approach capitalizes on the strength of diverse decision trees. By considering a range of perspectives, the model is more resilient to over fitting and capable of capturing complex patterns in the data.
* Predictive Power: When it's time to make predictions for a new, unseen data point, each tree in the forest provides its prediction. For classification tasks, the majority vote of the trees determines the final prediction, while for regression tasks, the mean or median of the individual tree predictions is taken.
* Optimal Parameters: Random Forests allow for the fine-tuning of hyper parameters to optimize performance. Parameters such as the number of trees, maximum tree depth, and minimum samples per leaf can be optimized to achieve the best results.

**4.1 Attribute importance**

This is a crucial step to gain insights into which features (attributes) are most influential in making predictions. Random Forest provides a built-in mechanism for assessing attribute importance. Random Forest calculates attribute importance based on the Gini impurity or Mean Decrease in Accuracy. These metrics quantify the reduction in impurity or accuracy that each attribute contributes when used for splitting nodes in the decision trees. In general, higher values of importance indicate more influential attributes. There are two common methods of calculating attribute importance in Random Forest: Gini impurity-based importance and Mean Decrease in Accuracy (MDA).

**4.1.1 Gini impurity-based importance**

In this method, attribute importance is calculated based on the reduction in Gini impurity. Gini impurity measures the disorder or impurity of a node in a decision tree. A lower Gini impurity indicates a more pure node.

Mathematically, the Gini impurity for a node N is given as,

Gini(N) = 1−∑((pi)2) (1)

where Gini(N) is the Gini impurity for node N and “pi” is the proportion of training instances in class i among the samples in node N.

The attribute importance is given as,

Attribute\_Importance(attr) = ∑(Gini(N) − Gini(Nattr)) × (|N|/|Nattr ​|) ​ (2)

where Attribute\_Importance(attr) is the importance of attribute 'attr', Gini(N) is the Gini impurity of the current node, Gini(N attr) is the Gini impurity of the node after the split on attribute 'attr', |Nattr| is the number of samples in the node after the split on attribute 'attr' and |N| is the total number of samples in the current node.

**4.1.2 Mean Decrease in Accuracy**

MDA measures the decrease in the accuracy of the model when an attribute is removed. It's a measure of the attribute's contribution to the model's accuracy. Mathematically, the MDA for an attribute 'attr' is calculated as,

MDA(attr) = (1/(Ntrees)). ∑(Accuracy\_Without\_attr - Accuracy\_With\_attr) (3)

where MDA(attr) is the Mean Decrease in Accuracy for attribute 'attr', Ntrees is the number of decision trees in the Random Forest, Accuracy\_Without\_attr is the accuracy of the model without attribute 'attr' and Accuracy\_With\_attr is the accuracy of the model with attribute 'attr'. The pseudocode used for the attribute importance is presented in table 3.

**Table 3** Pseudocode for attribute importance

Step 1: Initialize an empty dictionary 'attribute\_importance' to store importance

scores for each attribute.

Step 2: For each decision tree 'tree' in the Random Forest:

- Initialize an empty dictionary 'tree\_attribute\_importance' to calculate

importance scores for attributes in this tree.

For each split node in 'tree':

- Calculate the reduction in Gini impurity or accuracy when using the

attribute for the split.

- Update 'tree\_attribute\_importance' for the attribute based on this

calculation.

Step 3: Aggregate attribute importance from all trees:

For each attribute 'attr':

- Calculate the mean importance score across all decision trees for

'attr'.

- Store this mean score in 'attribute\_importance[attr]'.

Step 4: Sort the 'attribute\_importance' dictionary in descending order based on

importance scores.

Step 5: Output:

The 'attribute\_importance' dictionary, which provides a ranked list of

attributes and their corresponding importance scores.

Step 6: End

Above algorithm calculates importance scores for each attribute and ranks them based on their influence in making predictions. These importance scores can be used for feature selection, dimensionality reduction, and gaining insights into the dataset's predictive factors.

**4.2 Working of the model**

After feature selection it is crucial to implement the random forest model to predict the result. It operates by assembling an ensemble of decision trees. Each decision tree is developed independently, using subsets of the training data and features. During prediction, the model aggregates the outputs of individual trees, either through majority voting for classification tasks or mean/median calculations for regression tasks. The resulting combined prediction offers robust and accurate results by harnessing the wisdom of diverse decision trees, thereby mitigating over fitting and enhancing generalization. Table 4 gives an overview of working of the algorithm whereas table 5 presents hyper-parameters of Random forest used in the study.

**Table 4** Pseudocode for Random Forest algorithm

Step 1: Initialize an empty list 'forest' to store decision trees.

Initialize a variable 'n\_trees' to define the number of trees in the forest.

Step 2: For i from 1 to n\_trees:

- Create a random subset of the training data with replacement

(bootstrapping).

- Create a random subset of features (feature bagging).

- Build a decision tree using the selected data and features.

Step 3: Make predictions on the testing data using each decision tree

**Table 5** Hyper-parameters of Random forest used in the study

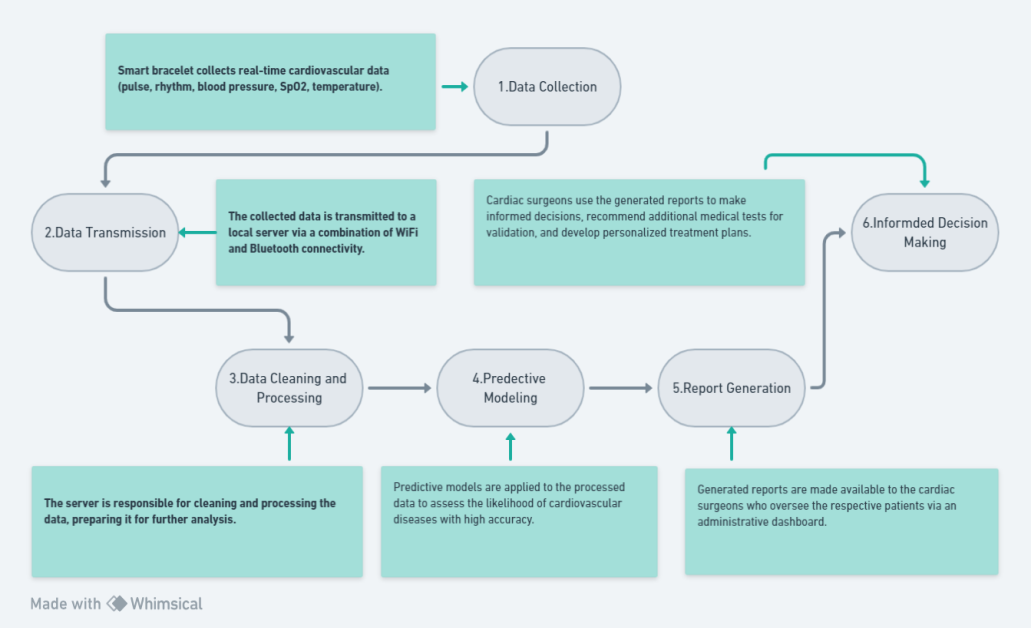
|  |  |  |
| --- | --- | --- |
| **Parameter** | **Detail** | **Range** |
| n\_estimators | The number of decision trees in the Random Forest | 10-100 |
| max\_depth | The maximum depth of each decision tree in the Random Forest | 5-30 |
| min\_samples\_split | The minimum number of samples required to split an internal node. | 2-20 |
| min\_samples\_leaf | The minimum number of samples required to be at a leaf node | 1-10 |
| max\_features | The number of features to consider when looking for the best split | 1-no of features in a dataset |
| bootstrap | A Boolean value indicating whether bootstrap samples should be used when building trees. | True/False |

**5. Methodology**

In the present work, an IoT-enabled smart wearable bracelet designed for comprehensive cardiovascular health monitoring is being introduced. The device is strategically placed on the patient's left arm to optimize data collection from the subclavian artery, ensuring the utmost accuracy. Real-time data pertaining to essential cardiovascular parameters, including blood pulse, heart rhythm, blood pressure, blood oxygen (SpO2) levels, and body temperature, is gathered for a duration of 1-2 minutes. Data collected by the smart bracelet is transmitted to a local server via a combination of WiFi and Bluetooth connectivity to minimize the risk of network-related issues. The server is responsible for data cleaning and processing, preparing it for further analysis. Subsequently, predictive models are applied to assess the likelihood of cardiovascular diseases with high accuracy. Generated reports are made available to the cardiac surgeons who oversee the respective patients via an administrative dashboard. These reports empower cardiac surgeons to make informed decisions, recommend additional medical tests for validation, and develop personalized treatment plans. This approach not only optimizes patient care but also contributes to cost efficiency and increased chances of patient survival [13], [14]. Components of the system are as follows:

* Sensor Technology: The smart bracelet is equipped with medical industry-standard sensors to facilitate accurate data collection:
* Photoplethysmography (PPG) Sensors: PPG sensors utilize emitted light to assess light absorption or reflection by blood vessels, enabling precise measurement of pulse rate and heart rhythm.
* SpO2 Sensors: These sensors gauge oxygen saturation levels in the blood, assisting in the detection of hypoxemia, a condition linked to elevated heart rates.[15]
* Temperature Sensors: Temperature sensors detect and monitor abrupt changes in body temperature, a valuable metric for identifying cardiac irregularities.
* Miniature ECG Sensors: These sensors monitor the heart's electrical activity, making them instrumental in detecting conditions such as atrial fibrillation. They ensure accurate results by establishing direct contact between electrodes on the bracelet and the patient's skin. [16,17]

The workflow of the IoT-Enabled Smart Wearable Bracelet for Cardiovascular Health Monitoring is presented in Fig. 8. This innovative wearable technology represents a significant advancement in cardiovascular health monitoring, offering both accuracy and convenience for patients and healthcare providers alike. It holds the potential to revolutionize early disease detection and improve patient outcomes.



**Fig. 8** IoT-Enabled Smart Wearable Bracelet for Cardiovascular Health Monitoring Workflow

Benefits of the IoT-Enabled Smart Wearable Bracelet for Cardiovascular Health Monitoring are:

* Accurate and comprehensive cardiovascular health monitoring
* Convenient and non-invasive data collection
* Real-time data transmission and analysis
* Predictive modelling for early disease detection
* Informed decision making and personalized treatment planning
* Improved patient care and outcomes

Overall, the IoT-enabled smart wearable bracelet for comprehensive cardiovascular health monitoring is a promising technology with the potential to significantly improve the lives of patients and healthcare providers alike.

**6. Proposed model and result analysis**

The work comprises the development of a smart and intelligent cardiovascular disease detection model for a sustainable environment. Leveraging a heuristic-based approach, a meticulous examination of cardiovascular health datasets to curate an optimized feature set has been conducted in this study. The heart of our model relied on the Random Forest algorithm to effectively classify individuals exhibiting risk factors associated with cardiovascular diseases. Further, the present work unfolds the versatile landscape of the Python programming language, allowing for a seamless and efficient implementation of the proposed predictive model. The outcomes of this investigation were not only insightful but also visually compelling. The results were meticulously documented, culminating in a collection of graphs and tables for comprehensive comparison with various evaluation parameters. To ensure the model's optimal performance, different hyper-parameters were fine-tuned meticulously with the Random Forest algorithm. In addition, various performance metrics were used for the implementation analysis, and these were derived from the confusion matrix shown in table 6.

**Table 6** Confusion matrix obtained from Random forest model

|  | **Actual Disease** | **Actual No Disease** |
| --- | --- | --- |
| Predicted Disease | 508 (TP) | 70 (FP) |
| Predicted No Disease | 76 (FN) | 564 (TN) |

The performance parameters of the model such as accuracy, precision, recall and F1 score are also calculated and analysed. “Accuracy” is defined as the ratio of the accurate disease predictions to the total number of predictions and can be calculated using equation (4) as,

(4)

“Precision” is a measure of how many of the positive predictions made are correct (true positives) and can be calculated using equation (5) as,

(5)

“Recall” is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data and can be calculated using equation (6) as,

(6)

“F1 Score” is a measure combining both precision and recall. It is generally described as the harmonic mean of the two and can be calculated using equation (7) given as,

(7)

where, TP = True positives, TN = True negatives, FP = False positives, FN = False negatives. TP is an outcome where the model correctly predicts the positive class, TN is an outcome where the model correctly predicts the negative class, FP is an outcome where the model incorrectly predicts the positive class and FN is an outcome where the model incorrectly predicts the negative class.

The efficiency of any machine learning model is determined by it’s ability to outperform the other existing models. The proposed Random forest model was tested against all other models such as KNN, CNN, RNN, Naive Byes, SVM and logistic regression model. The results of the comparison are presented in table 7.

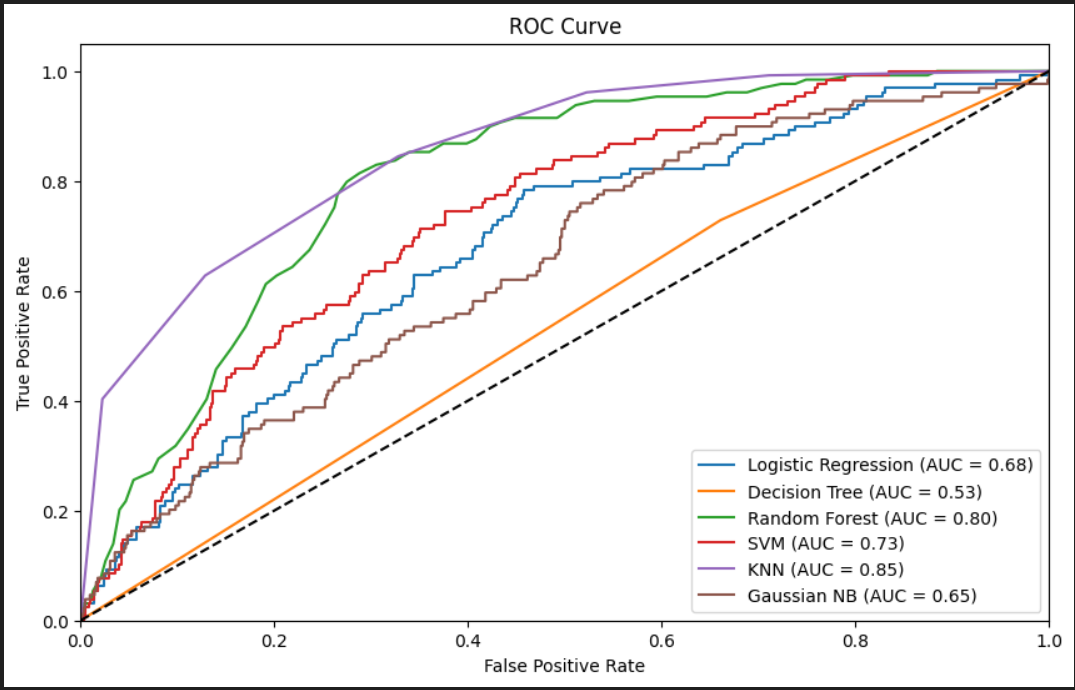
**Table 7** Accuracy comparison for different models

|  |  |
| --- | --- |
| Model | Accuracy (%) |
| Random Forest (proposed) | 88.364 |
| CNN | 80.69 |
| ANN | 80.29 |
| KNN | 76.1 |
| SVM | 73.899 |
| Logistic regression | 68.082 |

The enhanced classification accuracy is attributed due to the fact that over fitting and variance have been reduced through random forest with embedded attribute optimization feature in the present study. Also, it has inbuilt attribute selection functionality during model training phase.

Further, an essential facet of evaluating the performance of our predictive models is the Receiver Operating Characteristic (ROC) analysis. ROC curves provide a comprehensive understanding of a model's ability to discriminate between positive and negative instances across varying thresholds. ROC curves visually represent a model's performance by illustrating the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity). A perfect model would have an ROC curve that reaches the top-left corner (100% sensitivity and 0% false positive rate), while a random classifier's ROC curve would be a diagonal line from the bottom-left to the top-right.

To provide a visual representation of these results, the ROC curves for each model have been thoughtfully prepared and are shown in Fig. 9. It is observed that while KNN exhibits a higher AUC, indicating its strong ability to discriminate between positive and negative instances, the decision to opt for Random Forest over KNN is guided by the consideration of overall model performance and practical utility. Random Forest's superior accuracy, coupled with its robustness and generalization capability, makes it an apt choice for the task of cardiovascular disease risk prediction. Accuracy is often the most critical metric when the objective is to minimize the misclassification of individuals at risk. Additionally, Random Forest's ensemble nature helps mitigate over-fitting and enhances its practicality in real-world healthcare applications. Therefore, the decision to prioritize Random Forest is driven by the holistic view of model performance and its alignment with the study's specific objectives.



**Fig. 9** ROC curves for different predictive models compared

A new predictive healthcare model is said to be efficient and reliable if it generates consistent outcome with heterogeneous disease datasets. Evaluation of the proposed IoHT enabled predictive model is effective only when the accuracy is good enough. Various other performance indicators like specificity, sensitivity, and f-score metrics were used to determine the proposed hybrid heuristic classifier model’s effectiveness.

The application of the Internet of Health Things (IoHT) in smart cities offers substantial benefits for promoting sustainability in the proposed model. This innovative approach allows healthcare providers to continually monitor patients in real-time, enabling early detection of disease risks and preventing their escalation [19]. Key advantages of this IoHT-based model include:

* Technological Advancement: The system eliminates the need for unnecessary medical visits, streamlining the processing and analysis of patient health data in a timely manner.
* Enhanced Accessibility: Physicians can access and monitor patient data from anywhere, at any time, without the need to leave their location.
* Reduced Errors: The IoHT model ensures accurate data collection and automates workflows, minimizing resource wastage and reducing the likelihood of errors.
* Cost-Effectiveness: This system is straightforward to manage and deploy, leading to a reduction in patient visits to medical facilities and expediting the diagnosis and treatment of conditions like lung cancer. The interconnected nature of the model further reduces hospital stays and readmissions, making it a cost-effective solution.
* Remote Care: The IoHT model enables effective remote patient monitoring, reducing the necessity for face-to-face interactions with medical professionals. This approach not only minimizes unnecessary administrative burdens but also operates swiftly and cost-effectively. It conserves resources, thereby contributing to a more reliable and sustainable smart environment.

In essence, the IoHT model streamlines healthcare practices, making them more efficient and accessible while promoting sustainability and reducing the environmental footprint within the smart city framework.

**7. Conclusions**

In conclusion, a comprehensive approach to sustainable healthcare through the integration of the IoHT and machine learning for cardiovascular disease risk prediction is presented. The Random Forest model, with its high accuracy, stands out as a promising tool in this context. The introduction of IoT-enabled wearable technology further advances the field of healthcare by providing real-time data for accurate monitoring. The IoT technology not only offers convenience and precision but also contributes to a sustainable healthcare system by reducing the need for unnecessary medical visits, enhancing accessibility for physicians, minimizing errors, and lowering healthcare costs. This approach aligns with the principles of a sustainable smart city, where efficient healthcare practices can promote a healthier environment and improve patient care. The research work showcased in this paper has the potential to make a significant impact on the field of cardiovascular health monitoring and sustainability. It provides a foundation for further research and development in the intersection of healthcare, IoT, and machine learning, ultimately benefiting both patients and healthcare providers. Therefore, this work marks a significant step in advancing sustainable healthcare practices through the integration of IoT and machine learning, offering a glimpse into the potential future of healthcare in smart cities.

**References**

[1] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P., Heart disease prediction using machine learning algorithms, IOP Conference Series: Materials Science and Engineering, 1022(1), 012072 (2021).

[2] Ahire, N., Patel, R., & Gagare, S., Heart disease prediction using machine learning, International Journal of Engineering and Advanced Technology (IJEAT), 10(5), 351545128 (2021).

[3] Bhatt, C. M., Patel, P., Ghetia, T., Mazzeo, P. L., Effective heart disease prediction using machine learning techniques, MDPI, 16(2), 88 (2023).

[4] Chaimaa Boukhatem; Heba Yahia Youssef; Ali Bou Nassif, Heart Disease Prediction Using Machine Learning ,ASET 21668689 (2022).

[5] Pal, M., Parija, S., Panda, G., Dhama, K., Mohapatra, R. K., Risk prediction of cardiovascular disease using machine learning classifiers, Frontiers in Cardiovascular Medicine, 9, 857662 (2022).

[6] Hussain, A., Pawan, K., Chaurasia, P., Sabri, M. S., A review on heart disease prediction using machine learning techniques, IEEE Access, 7, 72162-72176 (2019).

[7] Parmar M., Heart Diseases Prediction using Deep Learning Neural Network Model, International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-9 Issue-3, (2020).

[8] Nissa, N., Jamwal, S., Ganie, S. M., Early detection of cardiovascular disease using machine learning techniques: An experimental study, Journal of Physics: Conference Series, 237552771(2020).

[9] Farley A, McLafferty E, Hendry C., Cardiovascular Disease,In book: StatPearls [Internet], Nurs. Stand. 27(9):35-9 (2012).

[10] Edgardo Olvera Lopez, Brian D. Ballard, Arif Jan, Cardiovascular Disease, National Library of Medicine-/NBK535419(2023)

[11] Ye Ruan, Yanfei Guo, Yang Zheng, Zhezhou Huang, Shuangyuan Sun, Paul Kowal, Yan Shi, Fan Wu, Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1 (2018).

[12] Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et. al., Global, regional and national burden of cardiovascular diseases for 10 causes, 1990 to 2015, J. Am. College Cardiol. 70(1):1–25 (2017).

[13] Suliman A,1,2 Abdul N, Waheb A. Jabbar, Mukarram A. M. Almuhaya, Anupam Kumar Bairagi, Md. Al-Masrur Khan, Seong-HK, IoT-Based Healthcare-Monitoring System towards Improving Quality of Life: A Review,36292441(2022).

[14] Alekya R., Boddeti N.D., Monica K.S., Prabha R., Venkatesh V., IoT based smart healthcare monitoring systems: A literature review, Eur. J. Mol. Clin. Med. 7:2020 (2021).

[15] Gera S., Mridul M., Sharma S., IoT based Automated Health Care Monitoring System for Smart City, Proceedings of the 5th International Conference on Computing Methodologies and Communication (ICCMC); Erode, India. 8–10 April 2021; pp. 364–368 (2021).

[16] Corrine Y. Jurgens, Christopher S. Lee, Dawn M. Aycock, Ruth Masterson Creber, Quin E. Denfeld, Holli A. DeVon, Linda R. Evers, Miyeon Jung, Gianluca Pucciarelli, Megan M. Streur, Marvin A. Konstam, The Relevance of Symptoms in Cardiovascular Disease and Research: A Scientific Statement From the American Heart Association, State of the Science:Vol 146,Issue 12, 2022

[17] Harshal S Yeotkar,Dr. V.T. Gaikwad, IoT based human body parameters monitoring by using wearable wireless sensor network, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056

Volume: 06 Issue: 07 , 2019

[18] Mohammad Shahidul Islam, Mohammad Tariqul Islam, Ali F. Almutairi, Gan Kok Beng, Norbahiah Misran, Nowshad Amin, Monitoring of the Human Body Signal through the Internet of Things (IoT) Based LoRa Wireless Network System, 9091884 applied sciences, 9(9), 1884, 2019.

[19] Ju-Yu Wu,Yuhling Wang, Congo Tak Shing Ching, Hui-Min David Wang, Lun-De Liao, IoT-based wearable health monitoring device and its validation for potential critical and emergency applications, 1188304 Front. Public Health, Sec. Digital Public Health Volume 11 - 2023

[20] Nafisa Shamim Rafa, Basma Binte Azmal, Abdur Rab Dhruba, Mohammad Monirujjaman Khan, Turki M. Alanazi, Faris A. Almalki, Othman AlOmeir, IoT-Based Remote Health Monitoring System Employing Smart Sensors for Asthma Patients during COVID-19 Pandemic,Theories, Technologies, and Applications of Artificial Intelligence in Cloud-Based Internet of Things Volume 2022 , 6870358 2022