engineering proceedings

MDPI

Proceeding Paper

Predicting Heart Disease Using Sensor Networks, IoT, and Machine Learning: A Study on Physiological Sensor Data and Predictive Models †

Neelamadhab Padhy

School of Engineering and Technology, Department of computer science and Engineering, GIET University, Gunupur, Odisha, India; dr.neelamadhab@giet.edu

† Presented at the 10th International Electronic Conference on Sensors and Applications (ECSA-10), 15–30 November 2023; Available online: https://ecsa-10.sciforum.net/.

Abstract: The Internet of Things (IoT) and sensor networks are used for structural health monitoring (SHM). This research aims to create a model for predicting cardiac disease using sensor networks, IoT, and machine learning. Through the wearable sensors the physiological data such as heart rate, blood pressure, and oxygen saturation levels collected from patients. The data is subsequently processed and translated into an analysis-ready format. The most important predictors of heart disease are identified using feature selection techniques. Accuracy, precision, recall, F1-score etc. are used to evaluate the performance of the proposed model. SVM obtained the highest accuracy with 93.87%.

Keywords: sensory data; CVD; Machine learning; performance analysis

Citation: Padhy, N. Predicting Heart Disease Using Sensor Networks, IoT, and Machine Learning: A Study on Physiological Sensor Data and Predictive Models. *Chem. Proc.* **2023**, *56*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2023

1. Introduction

In the 21st century, heart disease is one of the important causes of death in the world. There are several researchers discussed the HDP using machine learning. However early prediction and detection are important components in the healthcare sector. Wearable sensors plays a crucial role in tracking physiological variables like heart rate, blood pressure, pulse oximeters, activity sensors, temperature sensors, respiration sensors, and electrocardiogram (ECG) signals. With the help of these sensors, the heart disease system be monitored properly, if any kind of irregularities can be identified easily. Machine learning algorithms can be suitable in this situation to create predictive models that can identify those who have a high risk of acquiring heart disease and allow for early therapies. Structural health monitoring is one of the leading technologies for the health industry which focuses on the sensory that monitor the health of structures. The SHM allows to monitoring of the health of the cardiovascular system and detects any structural abnormalities or defects. ML algorithm effectively finds the unseen data from the database as well and it gives early warning signal to patients and healthcare providers. This can enable timely interventions such as lifestyle changes, medication, or surgery, and improve patient outcomes. The objective of this project is to develop a predictive model for heart disease using sensor-based monitoring and machine-learning algorithms. The model will be trained on a dataset of physiological parameters collected from patients using wearable sensors. The model's effectiveness will be assessed using established benchmarks like accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve. The ultimate objective is to create a sturdy and dependable model that can anticipate the probability of heart disease in patients, allowing for timely interventions that can enhance patient results and decrease healthcare expenses. The main motivation of this article is to enhance the HDP model's accuracy using sensory data. The sensory data are electrocardiography (ECG) and blood tests, which are often limited by their reliance on intermittent measurements and subjective interpretation. These data are provided continuously and analyzed using novel classification approaches. These algorithms provide the patterns and predict the outcomes. The goal of establish a heart disease prediction model employing sensor-based monitoring of the model's accuracy.

This paper discusses 4 sections. Section 1 discussed the introductory part of the structural health monitoring system along with sensory devices. Section 2 provides insight into heart disease prediction through machine learning classifiers. Section 3 provides the proposed model and is followed by Section 5 result discussion.

2. Literature Review

Many wearable sensor-based systems have been proposed recently to enhance the process of predicting heart disease. Tang, C., et al., [1] has used mechanical sensors that allow the monitoring the heart diseases. Their objective was to check pulse waves, heart rhythms, and BP. They also focus the real-time sensory data for CVD prediction. Lin, J., et al., [2] conducted an SLR (Systematic Literature Review) about various physiological signals for cardio problems. They also provide future direction. et al., [3] have developed one androidApp which monitors the patient record and provides information to the doctors taking care of the patients. Finally, they integrated into the cloud where the machine learning predictions were done. Salvi, S., et al. [4]. The research uses machine learning to present a combined hybrid feature selection and classification strategy for heart disease prediction in a cloud-based IoT healthcare system. Kumar, R., et al., [5] he study examine the application of linear regression as a machine learning approach for air quality time series prediction. This study employed sensor data from three separate locations in Delhi and the National Capital Region to estimate air quality for the following day using linear regression as a machine learning approach, and the findings were significant.. Kumar, P. M., et al., [6] various machine learning methods have been applied to predict heart disease, and logistic regression with majority voting achieved 88.59% accuracy, which is superior to previous techniques. A framework for dependable at-home assistance in healthcare was proposed by Hongxu and Niraj [7]. The primary focus of this architecture was data transmission between installed servers located in patient homes and distant hospitals, closer to the edge. Daniele et al.'s study [8] concentrated on classifying sensor data using an advanced computing technique. They provided sensor data analytics from mobile or wearable devices using the Deep Learning Approach concept. In their line of work, these gadgets serve as data collection points.

3. Proposed Model

The model is divided in to two phases.

- **Phase#1:** We have developed the proposed model for heart disease prediction using machine-learning classification algorithms. These are some of the steps followed:- Data collection done from the different sources through the sensors. We have deployed the sensors (heart rate, blood pressure, pulse oximeters, activity sensors, temperature sensors, respiration sensors, and electrocardiogram (ECG) signals) into the patient's body for testing.
- **Phase#2:** The major steps for this model are data cleansing, normalization, and feature engineering called DPA (data preparation and analysis). The objective of the step is to train the model. During DPA, we also identified the relevant features from the dataset and then used standard statistical tests and correlation analysis. Once feature engineering was over, we used the novel machine learning classification algorithms (RF,DTC,K-NN, SVM, GNB, AdaBoost, Bagging, KNN, LR) for CVD prediction. The below-mentioned Figure 1 indicates the Structural Healthcare monitoring model for cardiovascular prediction. In our proposed model we have used the machine learning classification algorithm like Random forest, decision tree, K-NN, Gussian Naive

Bayes, AdaBoost, Bagging and Logistic Regression. It is one kind of ensemble learning algorithm that generates a multiple numbers of decision trees during training and produces the output of the predicted classes of the individual trees. Normally this classifier is good when we are dealing with the high dimensional data as well as when missing values are available in the dataset. Decision Tree (DT): It is one of the classifiers that allows to splitting of the data into different homogeneous sets based on good features. As it is simple because this classifier can handle both the numerical and categorical data. K-NN is used for the same purpose that classifies the unseen heart disease instances of a patient. Support vector machine is one of the powerfull classifier which finds the best hyperplane and that sepatates the data into the different classes. It performs well on high-dimensional data and is capable of handling both linear and non-linear data. Gussian Naive Bayes (GNB): GNB is a probabilistic method that computes the likelihood of each class given the input data and chooses the class with the highest likelihood. It is simple and quick, and it can handle data with multiple dimensions.. AdaBoost: It is another kind of ensemble learing-techinques which combines the different weak classifiers to make another strong classifier. As our dataset is large so we have used for this algorithm for handling the imbalance issue which can enhance the performance of weak classifiers. Bagging is also an ensemble learning the generates the different subsets of the training data and trains a classifier on each subset. We have used for handling the overfitting and improve the performance of model purpose. Logistic Regression is also used for classification purpose which estimates the probability of binary outcome. Feature selection: Mutual Information Feature Selection (MIFS) is a technique that chooses relevant features by evaluating their mutual information with the target variable. MIFS can be particularly beneficial when working with sensor data because it can identify features that have a strong correlation with the target variable. This method evaluates how much information a feature provides about the target variable and selects the ones that have a high degree of mutual information. By using MIFS, it is possible to select the most relevant features and reduce the dimensionality of the dataset. This can improve the accuracy and efficiency of heart disease prediction models that use sensor data. The Figure 1 shows the proposed Structural Healthcare monitoring model for cardiovascular prediction.

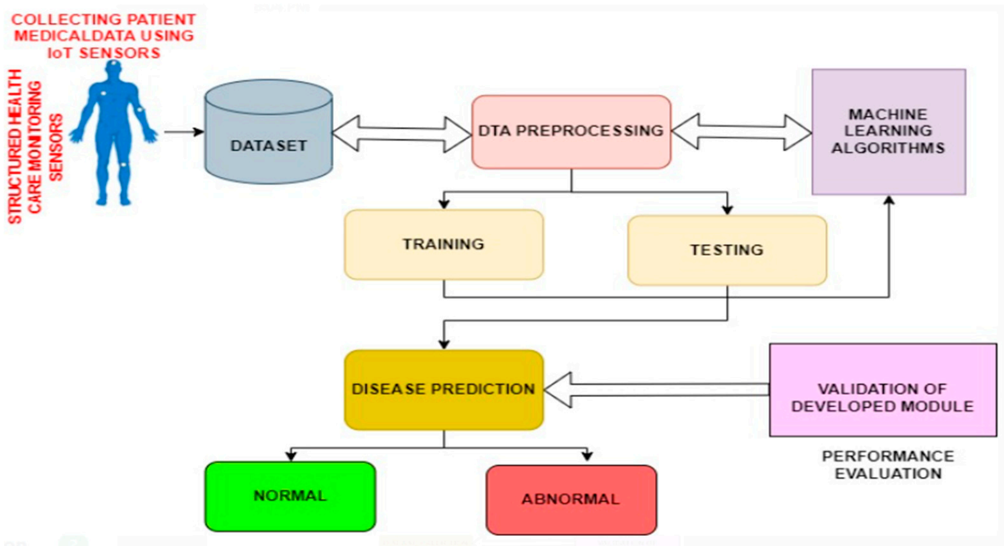

Figure 1. Structural Healthcare monitoring model for cardiovascular prediction.

- **Phase#3:** This phase mainly trains the model and tests its effectiveness. The best model is identified during the training phase and the trained model is tested on unseen data will be one of the challenging tasks. The performance metrics are evaluated such as accuracy, precision, recall, and F1 score area under the receiver operating characteristic (ROC) curve. The required hyper parameters are used to enhance CVD prediction accuracy.

- **Phase#4:** The model will be validated using the performance evaluation parameters on the test dataset which we have originally collected from the patients through sensory data. To ensure that it is not over fitting to the training data, validation required.

4. Result and Discussion

The above-mentioned Table 1 is meant for analysis of some classification models. Here we compare the performance metrics of the ensemble classification model like accuracy, precision, recall, F1-Score, true-negative rate (TNR), and true-positive rate (TPR). There are 9 classification model has been developed to achieve and these are Random Forest (RF), Decision Tree Classifier (DTC), K-Nearest Neighbour (K-NN), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), AdaBoost, Bagging, KNN, and Logistic Regression (LR). Our observation stated that SVM achieved the best accuracy with (93.87%). Again if we consider the TPR, LR obtained 90.20%. The Table 2 shows the confusion matrix for the various classifiers.

Table 1. Comparative analyses of the ensemble classification models.

Algorithms	Accuracy	Precision	Recall	F1-Score	TNR	TPR
RF	89.90	88.47	87.07	85.11	78.4%	84%
DTC	88.87	85.62	86.82	82.30	82.5%	54%
K-NN	87.16	86.34	85.99	88.40	89.2%	82%
SVM	93.87	93.33	93.67	93.35	77.6%	82%
GNB	87.25	89.20	87.54	91.20	89.2%	86%
AdaBoost	85.20	88.45	85.56	87.20	78.6%	84%
Bagging	89.54	89.99	90.20	81.25	87.7%	86%
KNN	87.55	89.25	88.89	90.20	89.7%	87%
LR	92.25	90.20	89.99	90.09	90.4%	88%

Table 2. Confusion matrix and ROC curve values for the various classifiers.

Name of the model	Accuracy	AUC	TNR	TPR	F-score
RF	89.90	0.88	78.4%	84%	85.11
DTC	88.87	0.87	82.5%	54%	82.30
K-NN	87.16	0.86	89.2%	82%	88.40
SVM	93.87	0.91	77.6%	82%	93.35
GNB	87.25	0.88	89.2%	86%	91.20
AdaBoost	85.20	0.85	78.6%	84%	87.20
Bagging	89.54	0.90	87.7%	86%	81.25
KNN	87.55	0.87	89.7%	87%	90.20
LR	92.25	0.93	90.4%	88%	90.09

The Critical Observation

In the above-mentioned, the highest accuracy obtained is SVM while AdaBoost has the lowest, with a range of 85.20% to 93.87%. Similarly, when we considered AUC (Area Under the Curve) it ranges from 0.85 to 0.93. The TPR and TNR are important components of any kind of classifier. The highest TPR and TNR are found in SVM and LR, whereas the lowest TPR is seen in DTC. Similarly, when we consider the F-score as well as Precision, SVM obtained the highest classifier that can find the positive cases.

The Table 3 is meant for error analysis of several machine learning algorithms. We have used 9 models to exhibit the performance in Tables 1 and 2 as well for error in Table 3. The MSE an RMSE are considered the best error measurement for the classification model. From Table 3 it is concluded that SVM has the lowest MSE and RMSE which indicates the most accurate model for heart disease prediction. The model with the greatest MSE and second-highest RMSE is LR, indicating that it is the least accurate.

Table 3. Details of the implementation environment used in the current study.

Model	MSE	RMSE
RF	0.37322	0.47185
DTC	0.33547	0.57805
K-NN	0.45281	0.59233
SVM	0.18752	0.32589
GNB	0.32558	0.50785
AdaBoost	0.45287	0.60890
Bagging	0.32897	0.54780

As the dataset is one kind of binary class classification nature and performs AUC = 0.94 which perfectly classifies the positive and negative classification. The ranges of AUC value start from 0 to 1. Here zero indicates that random classifier and 1 indicates the perfect classifier. Figure 2 shows an AUC value of 0.94 and the model suggested a good level of discrimination between positive and negative classes. This graph is meant for the performance of the binary classifier.

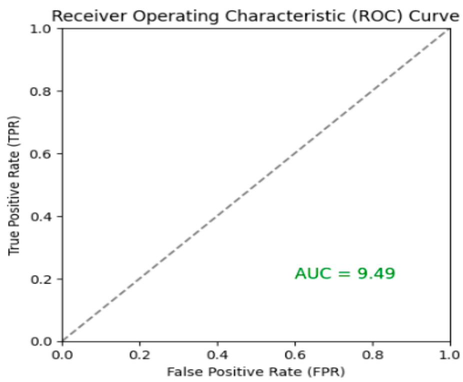

Figure 2. True positive vs. false positive rate.

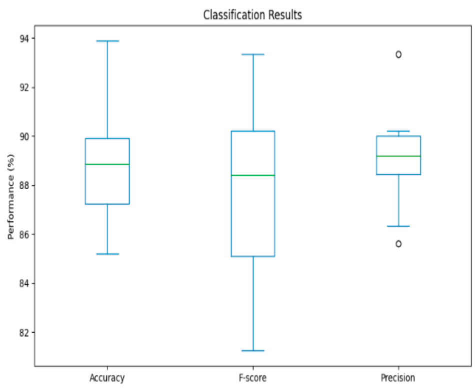

Figure 3. Boxplot representation for CVD prediction.

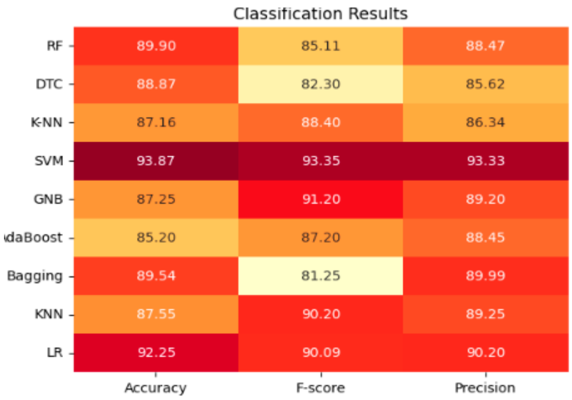

Figure 4. Classification Result for heart disease data.

The above figure demonstrates the classification results for heart disease prediction. As concerned of precision metrics, our result demonstrated the highest one i.e., 93.33% which indicates a low false positive rate. The above figure represents the normalized confusion matrix for the heart disease classification task. This graph has a square matrix where the column represents the predicted classes whereas the rows represent the actual classes. These classes are divided into the total number of samples in each true class that's why this graph is called a normalized matrix. The reason behind this is to compare the performance of the different classes even though they have a number of independent samples. In the above figure, the diagonal matrix of each graph indicates the correctly identified classes and the off diagonal indicates the misclassified observation. A classifier is said to be good where the diagonal matrix value represents 1 which means all the observations that were collected correctly classified the heart disease.

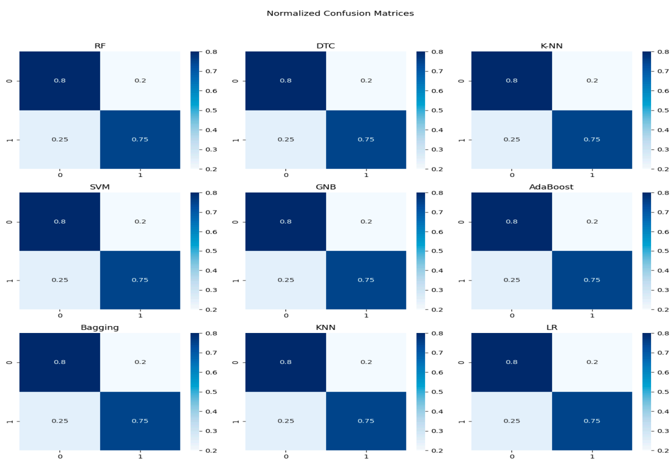

Figure 5. Normalized Confusion Matrix.

5. Conclusions

The heart disease prediction model developed in this work has the potential for improving patient outcomes while also lowering healthcare costs by identifying patients at risk of developing heart disease and offering appropriate interventions and treatments. Future research can expand on the possibilities of sensor networks, IoT, and machine learning approaches in healthcare, allowing for the development of more accurate and effective predictive models for heart disease and other medical diseases. SVM has the highest accuracy of 93.87%, followed by LR at 92.25%. Overall, SVM and LR seem to be the most effective models for this dataset, while AdaBoost may need further optimization. SVM has the best overall performance, with the lowest Mean-squared Error (MSE) and Root-mean-squared Error (RMSE) of 0.18752 and 0.32589, respectively.

References

1. Tang, C.; Liu, Z.; Li, L. Mechanical sensors for cardiovascular monitoring: From battery-powered to self-powered. *Biosensors* **2022**, *12*, 651.
2. Lin, J.; Fu, R.; Zhong, X.; Yu, P.; Tan, G.; Li, W.; Zhang, H.; Li, Y.; Zhou, L.; Ning, C. Wearable sensors and devices for real-time cardiovascular disease monitoring. *Cell Rep. Phys. Sci.* **2021**, *2*, 100541.
3. Esther, G. M.; Ahila, S. S.; Kumar, P. H. Coronary Heart Disease (Cad) Monitoring System Based On Wireless Sensors. In *Journal of Physics: Conference Series*; IOP Publishing: 2019; Volume 1362, p. 012045.
4. Salvi, S.; Dhar, R.; Karamchandani, S. IoT-Based Framework for Real-Time Heart Disease Prediction Using Machine Learning Techniques. In *Innovations in Cyber Physical Systems: Select Proceedings of ICICPS 2020*; Springer: Singapore, 2021; pp. 485–496.
5. Kumar, R.; Kumar, P.; Kumar, Y. Time series data prediction using IoT and machine learning technique. *Procedia Comput. Sci.* **2020**, *167*, 373–381.
6. Kumar, P. M.; Gandhi, U. D. A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Comput. Electr. Eng.* **2018**, *65*, 222–235.
7. Yin, H.; Jha, N. K. A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles. *IEEE Trans. Multi-Scale Comput. Syst.* **2017**, *3*, 228–241.
8. Ravi, D.; Wong, C.; Lo, B.; Yang, G. Z. A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE J. Biomed. Health Inform.* **2016**, *21*, 56–64.