Hybrid Convolutional neural network for detecting heart disease in patients

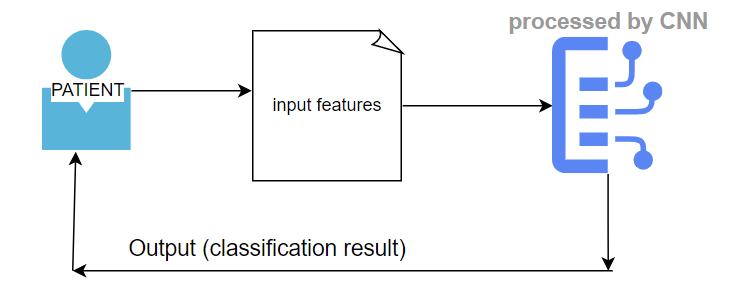
*Abstract*- If anyone wants to live a good life, having a healthy lifestyle and maintaining good physical and mental health is very important. The health industry includes all the people that contribute their lives to make others’ lives better. Health industry is one of the most important industries and has recently been getting huge technological advancements due to automation. This means that the level of healthcare that is being administered to patients has been improving as time passes. It is essential to note that artificial intelligence is widely being applied in order to ease the prediction and improve accuracy of detecting fatal illnesses/defects at an early stage. Heart disease is a critically overlooked sector in the health industry. There are several symptoms that play a major role in predicting the presence of a heart disease in a patient. Machine learning and deep learning models are known to have a high accuracy rate in data analysis of medical patients. Hence, we use this already existing knowledge to predict fatal heart conditions based on already-available features. In this paper, we devise a hybrid-convolutional neural network (hybrid CNN model) in order to achieve high accuracy rate by using selective-features after working with the Cleveland heart-disease dataset that is available in UCI repository.

# Introduction (*Heading 1*)

We have all heard the famous quotation ‘health is wealth’. Especially in this generation, people are becoming more and more aware of the benefits on maintaining a healthy lifestyle and regular exercise. This is supplemented by the health industry staff, that includes doctors, nurses, paramedics and so on… it is important to note that the health industry is also having a steep increase in the productivity and is becoming more and more advanced both in terms of medical knowledge as well as technical components used in the industry such as x-rays. One fairly new sector is the use of automation inside this sector so that the doctors will get more assistance from the robots used to perform these tests, that is., the robots can themselves make certain decisions that will assist the doctor in coming to an analysis and then diagnosis faster. This means that the robots will themselves be able to analyze the data that is taken by them and provide an outcome. Here, artificial intelligence can come in handy to take in the data samples and come to a conclusion. We propose such an AI model for the prediction of heart disease based on certain features that are recorded during the patient’s checkup.

Heart disease is a sensitive topic and must be handled with care because it has a high fatality rate. Incorrect predictions can result in permanent heart damage or even premature death. We should be aware of both false-positive predictions as well as false-negative predictions and both the above cases will prove to be fatal for the patient. Hence, this is a very sensitive topic and must be handled with care. Heart disease prediction has always been a tough job due to the high number of factors that may affect the outcome drastically. Usually, the main causes of people having heart diseases is due to an unhealthy lifestyle/sedentary lifestyle. Genetic heart problems are also a major issue as even though the person may be athletic, he/she could have a defective heart since birth. We need to be incredibly careful if we are implementing robots to do the analysis and prediction of heart disease and the model must be extremely precise and not have any false negative predictions, that is., predicting someone who actually does have a heart problem and medically fit. We must note that it is better to not predict at all than to appear at a false conclusion, hence putting the patient’s health at significant risk.

Artificial intelligence has recently gained a lot of popularity. Automation of robots has taken place in a lot of industries, including the medical industry. One integral sub-part of artificial intelligence is machine learning, where a model is made and the model learns to process data and come to a conclusion on its own. Machine-learning algorithms have widely been in use for the past-few years and are continually being used in the medical sector. These hybrid machine-learning models work on pre-existing datasets that have data collected from credible sources. One such dataset is the Cleveland heart disease dataset from the uci repository. In many datasets, we can see there are at least 10-20 important attributes that majorly affect the outcome of the machine-learning model used for prediction. Cleveland dataset is also a similar dataset with 70+ features, but mainly 13 selected attributes and 1 target attribute which is used for prediction. Many machine-learning models have been proposed till date in various papers using well-known algorithms such as Decision-trees, Random-forest classifier, KNN(k-nearest-neighbors) classifier, and several boosting algorithms. These models have developed a decently accuracy by themselves but even small variations in the dataset lead to drastic fall in accuracy and hence fails the model.Deep learning is a part of machine learning but is way more complex and has many hidden layers that have modifiable weights and biases that provide more importance and priority to certain features over other features. These weights and biases also have the ability to self-learn and change their values in each iteration (also called epoch) using back-tracking. Hence, we dive into deep-learning and construct a neural-network using convolution in order to get a better prediction with lower false-predictions. Deep learning is a part of machine learning and is a type of supervised learning that uses multiple learning parameters and many hidden layers to give a better output. A hybrid convolutional neural network model (CNN) has been implemented in this paper. Chest pain, bad cholesterol levels, thalassemia, blood sugar, rest ECG are some important features used for prediction in this model. The hybrid CNN is proposed and implemented with a high accuracy of above 99% which has never been attained before. Many proposed models till date have limited feature-selection for algorithmic use. Our proposed hybrid model does not provide any such restrictions and hence has a stronger capability to predict the heart diseases as compared to pre-existing methods even in case of outliers/variations in the newly-fed dataset.



Identify applicable funding agency here. If none, delete this text box.

# related work

Heart-disease is a topic on which many surveys and research has been conducted. Several machine-learning models-both supervised and unsupervised have been used in recent years to predict presence or absence of disease.

Two frequently used datasets are the Statlog and Cleveland datasets, both openly available datasets. One of the papers used a hybrid random forest with a linear model (HRFLM) and got an accuracy of 88.7% [1]. HRFLM model in the paper was compared with other pre-existing models and proved to have the highest accuracy score as compared to the other models. Another model used got a 98.4% accuracy using heart disease clinical decision support system (hdcdss) using multiple models that were stacked on top of each other [2]. This model was implemented by integrating DBSCAN, SMOTE-ENN, and XGBoost-based MLA to improve prediction accuracy. Paper [3] first designed a Fuzzy-GBDT algorithm combining fuzzy logic and gradient boosting decision tree (GBDT) to reduce data complexity and increase the generalization of binary classification prediction. Then, they integrated Fuzzy-GBDT with bagging to avoid overfitting. As a lot of features are present in the dataset and using association rules will only further increase the number of features and rules, an algorithm[4] was proposed to reduce the number of rules. The algorithm performs several train and test cycles to achieve basic cross-validation and reduce the number of rules with poor generalization potential. One other paper also used a cnn model[5] but just implemented the model as it is and got 95% accuracy.

The problem of not having a robust model for real world was acknowledged in [6]. The proposed OCI-DBN for heart disease prediction resolves three main problems: overfitting, underfitting, and finding an optimal network configuration. One paper to breakthrough in accuracy used a fast conditional mutual info model(FCMIM) and got an accuracy of 92.7%. Another paper did use deep learning: they used both cnn and rnn models and got an accuracy of 87.9% using cnn,89.4% using rnn respectively. Another MSSO-ANFIS method was proposed that had a superior accuracy of 97.89%, which was higher than all other approaches. (modified salp swarm optimization (MSSO) and an adaptive neuro-fuzzy inference system (ANFIS)) [8]. Two other papers addressed issues on heart disease in IoT platform[9][11][12][13][14] and then also on renal dysfunction in heart-disease patients[10]. An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection[15] was one of the first papers to use a hybrid of 2 algorithms and got a 3.3% increase in accuracy as well. Another paper proposed the CHD risk prediction method based on two DNN models and applied it to the KNHANES dataset[16] but the main limitation of this implementation is that it does not allow for missing values, hence work needs to be done before it can be put into the real-world environment.

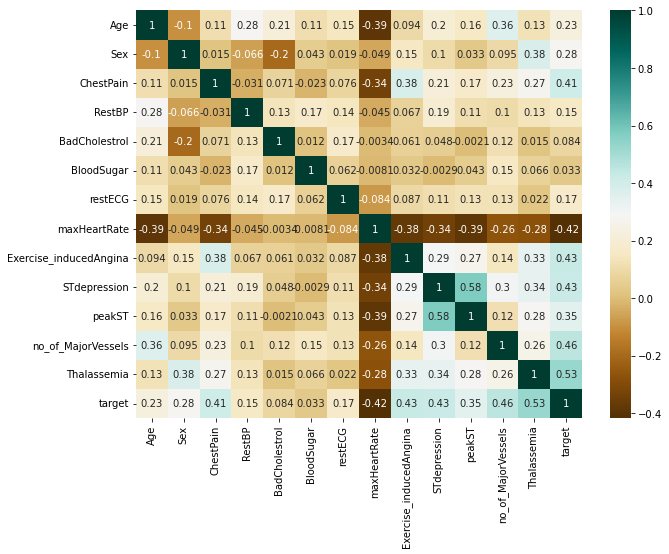
It is important to note that all these above papers also implemented their models on the same set of important features present in the Cleveland uci dataset and used classification models to predict presence or absence of heart disease- 1 or 0. Based on these works, a more flexible hybrid convolutional neural network model is proposed by us in this paper.

# Dataset

The dataset used is the Cleveland heart disease dataset that is available in the UCI repository. (<https://archive.ics.uci.edu/ml/datasets/heart+disease)>. It consists of 75 input attributes and 1 output target class label. But, all these 75 attributes were not considered during the implementation as some private attributes such as names and social security numbers have been removed (these attributes are called PII-personally identifiable information and do not hold any weightage to predict the presence or absence of heart disease. For ex: whether someone’s name is Adam or Dominic, or their bank balance does not help predict presence or absence of heart disease in that patient). The 13 main usable attributes that have a big impact on the final target result are listed below. The 14th attribute, which is target is just a dummy attribute used to map the output classification value, i.e., the presence ( with output 1) or absence (with output 0) of heart disease.

1. Age in years
2. Sex: 1->male, 0->female
3. Chest pain:
   1. Value 1: typical angina
   2. Value 2: atypical angina
   3. Value 3: non-anginal pain
   4. Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to hospital)
5. Cholesterol: serum cholesterol level
6. FBS: blood sugar on fasting
7. RestECG: electro-cardiogram results while resting
8. thalach: maximum heart rate achieved
9. exang: exercise-induced angina (1 -> yes, 0 -> no)
10. Oldpeak: ST depression induced by exercise relative to rest
11. Slope: the slope of the peak exercise ST segment:
    1. Value 1: upsloping
    2. Value 2: flat
    3. Value 3: downsloping
12. Ca: number of major vessels (0-3) coloured by flouroscopy
13. Thal: 3=normal, 6=fixed defect, 7= reversable defect
14. Target : the prediction attribute for presence(1) or absence(0) of heart-disease in the patient

The heatmap corresponding to the above attributes:



# Proposed methodology:

A total of 75 input attributes are present for the Cleveland heart-disease dataset. We are using only 13 key attributes in order to arrive at a conclusion.

The model that we are using is a convolutional neural-network model. Below is an image from theclickreader.com.

A CNN model is a deep-learning model that takes in an input(usually in the form of an image), assigns importance(weights and bias) to the various attributes and be able to differentiate one from the other. We implement a hybrid CNN model having 3 convolution layers, 1 max-pooling layer to reduce the features, dropout and flattening, followed by 3 dense layers using leaky relu activation function, and finally 1 dense layer using softmax activation function.

There are a total of 10245 rows are present, hence we divide it into 70% for training and 30% for testing. As the python inbuilt features for the CNN model(such as convolution) are made to act on images, we first reshape the train and test arrays into 1-d arrays. Also, we use conv1d for the convolution layers as the input is in a single row of attributes(i.e., in the form of an array), and not 2-dimensional like an image.

Below is the summary of the hybrid CNN model showing the number of learnable parameters at each layer. Reshaping was also done in the pre-processing stage. A visual representation is also pasted below, created using draw.io.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Model: "sequential"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

conv1d (Conv1D) (None, 13, 32) 128

conv1d\_1 (Conv1D) (None, 13, 64) 6208

conv1d\_2 (Conv1D) (None, 13, 128) 24704

max\_pooling1d (MaxPooling1D (None, 7, 128) 0

)

dropout (Dropout) (None, 7, 128) 0

flatten (Flatten) (None, 896) 0

dense (Dense) (None, 256) 229632

dense\_1 (Dense) (None, 512) 131584

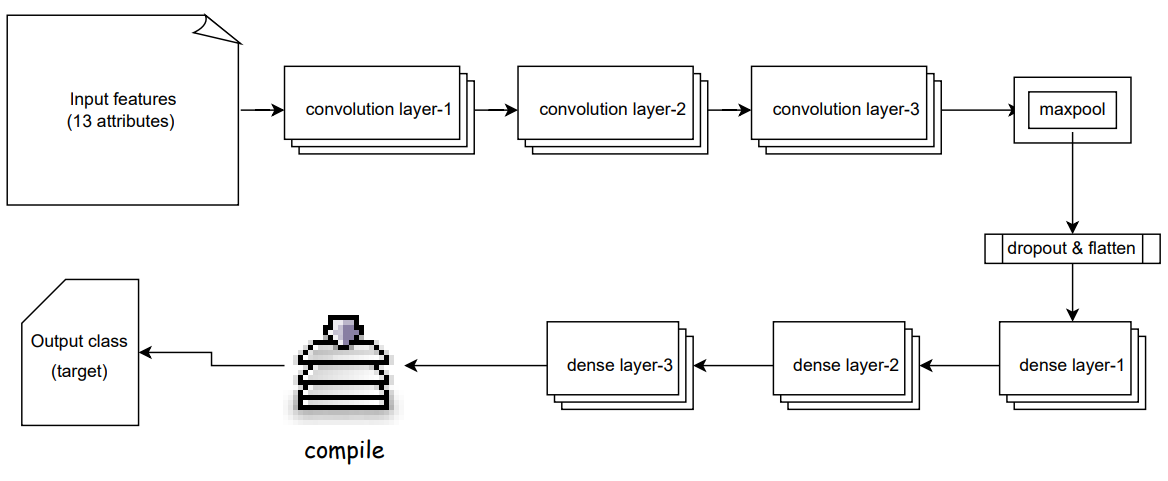
dense\_2 (Dense) (None, 10) 5130

=================================================================

Total params: 397,386

Trainable params: 397,386

Non-trainable params: 0



1-d convolution means that convolution occurs over a single temporal dimension.

LeakyReLU activation function was used for all 3 layers. Here, f(x)=max(0.01x,x).

This function returns value x as it is if input x is positive. If x is negative, it returns a very small value close to 0 (x\*alpha). leaky relu hence has a small slope for negative values instead of a flat slope(flat slope is seen in normal relu function).

In our case, alpha for leaky relu is set to 0.001 in the convolution layers, and 0.05 in the dense fully connected layers. Alpha value is the value that leaky relu substitutes in place of 0(for negative inputs).

The activation function used is sigmoid. Sigmoid takes any real value as input and outputs values in range of 0 and 1.

S(x) =

Binary cross-entropy is the loss function used as we had to squash the outputs between 0 and 1- depicting absence or presence of heart disease respectively. Binary cross entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value. That means how close or far from the actual value. It is also called log loss.

Given y as the label and p(y) being the probability of y, we have

Logloss=

Finally, metrics- accuracy and precision were used to record the performance of the proposed cnn model. Below is a custom made diagram of the proposed model.

A picture containing chart

Description automatically generated

As we can see above, a hybrid CNN model has been created using unorthodox 1-dimensional convolution. Three such convolution layers were used with increasing feature sizes, and then we cut down the sizes using pooling, dropout and flattening. Finally, we properly finish training the model and getting the features to map required output by passing through 3 dense fully- connected layers that gradually decrease in size respectively. We are able to map the output to a binary classification of 2 labels: 1 or 0 using the ‘sparse\_categorical\_crossentropy’ option as loss function during the compilation process. Finally, after training the model, we test it with the remaining values from the dataset that were reserved for testing and got a very high output accuracy of 99.4%, which is the highest ever accuracy recorded in this dataset.

# Future works:

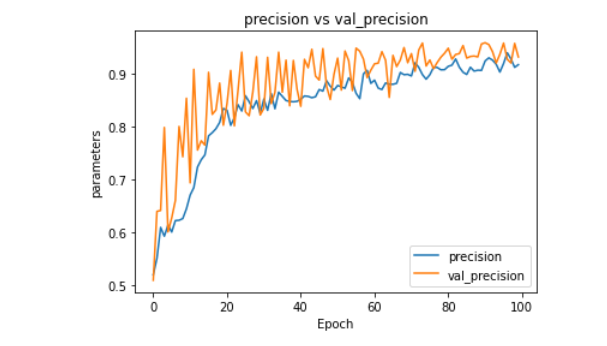
In the future, we can put the model into real-world environment, i.e., by collecting data from near-by hospitals and pre-processing to only include certain important features like in Cleveland dataset and then using the CNN model to predict the presence/absence of heart disease. We can also use this model to create better models in the future. This model can be used as a baseline to create ML and DL applications for other problems present in the field of medicine.

Furthermore, the model can be made more robust and provide higher accuracy by tweaking certain hyperparameters to better suit the characteristics of patients in a particular region. This can help increase the consistency of results based on identified outliers that occurred in the past.

# results and discussions:

As we can see above, our hybrid convolutional neural network model was proposed and implemented on the Cleveland dataset. Our proposed model has the highest accuracy of above 99% and high precision rate also. High precision along with high accuracy indicates that there is very small rate of misclassification, i.e., false positives (fp) and false negatives(fn). This means there are lesser number of cases where a heart disease in patient was classified as normal. Below is a plot of accuracy on training as well as test data with increasing epochs.

Chart, line chart

Description automatically generated 

Similarly, precision on both training and testing data has also been plotted. Clearly, precisino on the testing data is really high aswell starting from epoch(8-10) onwards.

Comparing the results obtained with the previously proposed models,

|  |  |
| --- | --- |
| **Model** | **Recorded accuracy** |
| Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets(cnn,rnn) | 87.9%,89.4% respectively |
| hybrid random forest with a linear model (HRFLM) | 88.7% |
| heart disease clinical decision support system (hdcdss) | 98.4% |
| fast conditional mutual info model(FCMIM) | 92.7% |
| modified salp swarm optimization (MSSO) and an adaptive neuro-fuzzy inference system (ANFIS) | 97.89% |
| An optimally configured and improved deep belief network(OCI-DBN) approach | 94.61% |
| Our hybrid CNN model | 99.4% |

Hence, our hybrid cnn model yields the best results on the test dataset. No proposed paper has come this close in terms of precision while also having a high accuracy. Further work can be done on the model by trying to further optimize its working by training and testing on other datasets aswell, and finally in the real world as mentioned in future works section above.

# References:

1. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
2. N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
3. X. Yuan, J. Chen, K. Zhang, Y. Wu and T. Yang, "A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT," in IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2032-2040, March 2022, doi: 10.1109/TII.2021.3098306.
4. C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," in IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 2, pp. 334-343, April 2006, doi: 10.1109/TITB.2006.864475.
5. Y. Pan, M. Fu, B. Cheng, X. Tao and J. Guo, "Enhanced Deep Learning Assisted Convolutional Neural Network for Heart Disease Prediction on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 189503-189512, 2020, doi: 10.1109/ACCESS.2020.3026214.
6. S. A. Ali et al., "An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo–Tompa and Stacked Genetic Algorithm," in IEEE Access, vol. 8, pp. 65947-65958, 2020, doi: 10.1109/ACCESS.2020.2985646.
7. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
8. M. A. Khan and F. Algarni, "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," in IEEE Access, vol. 8, pp. 122259-122269, 2020, doi: 10.1109/ACCESS.2020.3006424.
9. C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020, doi: 10.1109/ACCESS.2020.2981159.
10. B. Wang et al., "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients With Electronic Health Records," in IEEE Access, vol. 7, pp. 178392-178400, 2019, doi: 10.1109/ACCESS.2019.2956859.
11. L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on Statistical Model and Optimally Configured Deep Neural Network," in IEEE Access, vol. 7, pp. 34938-34945, 2019, doi: 10.1109/ACCESS.2019.2904800.
12. S. S. Sarmah, "An Efficient IoT-Based Patient Monitoring and Heart Disease Prediction System Using Deep Learning Modified Neural Network," in IEEE Access, vol. 8, pp. 135784-135797, 2020, doi: 10.1109/ACCESS.2020.3007561.
13. D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.
14. M. A. Khan, "An IoT Framework for Heart Disease Prediction Based on MDCNN Classifier," in IEEE Access, vol. 8, pp. 34717-34727, 2020, doi: 10.1109/ACCESS.2020.2974687.
15. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in IEEE Access, vol. 7, pp. 180235-180243, 2019, doi: 10.1109/ACCESS.2019.2952107.
16. T. Amarbayasgalan, V. -H. Pham, N. Theera-Umpon, Y. Piao and K. H. Ryu, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," in IEEE Access, vol. 9, pp. 135210-135223, 2021, doi: 10.1109/ACCESS.2021.3116974.
17. S. E. A. Ashri, M. M. El-Gayar and E. M. El-Daydamony, " ork Based on Hybrid Classifiers and Genetic Algorithm," in IEEE Access, vol. 9, pp. 146797-146809, 2021, doi: 10.1109/ACCESS.2021.3122789.
18. G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection," in IEEE Access, vol. 10, pp. 23808-23828, 2022, doi: 10.1109/ACCESS.2022.3153047.
19. A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.
20. P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in IEEE Access, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
21. G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in IEEE Access, vol. 10, pp. 80151-80173, 2022, doi: 10.1109/ACCESS.2022.3165792.
22. W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou and W. Cao, "A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease," in IEEE Access, vol. 7, pp. 175248-175258, 2019, doi: 10.1109/ACCESS.2019.2957367.
23. J. Chen, A. Valehi and A. Razi, "Smart Heart Monitoring: Early Prediction of Heart Problems Through Predictive Analysis of ECG Signals," in IEEE Access, vol. 7, pp. 120831-120839, 2019, doi: 10.1109/ACCESS.2019.2937875.