International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 02, 2021

# Heart Disease Classification And Risk Prediction By Using Convolutional Neural Network

V.Archana Reddy[1], K Venkatesh Sharma[2]

*[1]Assistant Professor, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, Telangana, India.*
*[2]Professor, Department of Computer Science and Engineering, CVR College of Engineering,Rangareddy, Telangana, India.*

*Email:[1]vareddy.cse@gmail.com,[2]venkateshsharma.cse@gmail.com*

***Abstract: one of the biggest causes of death in today's globe is heart disease. Heart disease is the most common cause of death for both men and women.This has a significant impact on human life. The majority of the time, heart disease diagnosis is based on a complex mixture of clinical and pathological evidence. Machine learning effectively aids in making decisions and predictions from the massive amount of data generated by the healthcare industry, which pays to more noise, such as missing data, duplicate data, etc. The efficiency of classification and the accuracy of disease prediction are also affected by duplicate and null values. Various traditional machine learning algorithms have been implemented to increase heart disease prediction performance in the current method. Our proposed work includes Machine Learning-based classifiers to assess classification accuracy and the use of Deep Learning techniques such as Convolutional Neural Network Unidirectional Risk Prediction (CNN-UDRP) to enhance the accuracy of heart disease prediction. We also compare the classification accuracy of the KNN, SVM, and Naive Bayes Classifiers on many healthcare datasets.***

***Keywords: Machine Learning,Deep Learning,Coronary Heart Disease,CNN-UDRP***

## 1. INTRODUCTION

Heart disease is the greatest threat to our lives today. About 13% of all deaths in the United States are attributed to CHD. In order to minimize cardiac arrest and mortality, timely detection of heart disease is essential. According to research from the American Heart Association, heart disease care is predicted to rise by almost 100% by 2030. The two main risk factors for hypertension are the Body Mass Index (BMI) and systolic blood pressure. Higher levels of blood pressure are related to age, sex, and BMI. A large amount of the evidence that we have seen suggests that the higher levels of creatinine in the blood raise the risk of CHD, including chronic, meaning present for a considerable period, and severe increases in both Cholesterol and glycohemoglobin in CHD patients. Cardiovascular velocity and axis deviations have been previously demonstrated to be risk factors for future coronary heart disease using echocardiography and electrocardiography study models. Early diagnosis and early treatment are just as critical to the financial health of an organization as saving lives and improving the quality of life of its customers.

Non_communicable diseases, such as heart disease, cancer, and diabetes, account for nearly 61% of all disease-related deaths in India. A significant cause of the disease is the environmental and lifestyle conditions that people face. An earlier diagnosis of the disease and decreased disease and diagnosis risk are made possible through machine learning. Today, big data is a better method for disease prediction, detection, and diagnosis than in the past. They process any data from a massive set of data to gather the needed information. The historical data that precedes future predictions informs them [5]. Data in the form of numbers (such as statistics, artificial intelligence, database techniques, and machine learning) can be found in many fields of big data. When making decisions in the medical field, hidden information is contained in the medical data, making it hard to conclude.

Machine learning has significant implications in analyzing medical data and discovering hidden patterns in medical data. Many different types of machine learning can be applied when analyzing different data types, including healthcare, finance, government, transportation, and marketing. In the medical field, machine learning is used to detect, diagnose, and predict disease. These techniques aim to discover heart disease earlier to be treated with proper heart disease care[6]. Heart disease can be successfully treated with a mixture of dietary improvements, medication, and sometimes procedures. The appropriate therapy will reduce the effects of cardiac failure and increase the functionality of the heart. The projected outcomes will be used to postpone surgery and reduce the expense of more costly therapies.

Convolutional Neural Network Unidirectional Risk Prediction (CNN-UDRP) is developed using Naive Bayes, SVM, and KNN algorithms to predict heart disease risk. The machine uses age, sex, blood pressure, cholesterol, and obesity to determine. The CNN-UDRP predicts the probability of heart disease in patients. It makes for considerable knowledge. Cf. e.g., Relationships should be formed between medical factors contributing to heart disease and habits.

The primary aim is to compare the classification performance of the proposed classifiers and reliably predict the probability of heart disease prediction (CNN-UDRP) of the Convolutional Neural Network, which we are using the UCI repository data collection.Machine learning plays a crucial role in the identification and study of discretely hidden patterns. After data processing, computer training methods aid in the identification and early detection of heart disease. This paper analyses the efficiency of various ML strategies such as Naive Bayes, SVM, Decision Tree, Logistic Regression, and Random Forest for the early prevision of heart disease. That is the key justification for this study.

The paper's main contribution is to perform the data pre-processing techniques upon the Heart diseases dataset to improve the disease classification performance and measure the classification and prediction accuracy among Machine learning and Convolutional Neural Network Techniques.

The remainder is arranged as follows. Section 2 discusses the background work;related work has been presented in Section 3. Some widely used Methods and Datasets are presented in Section 4. Results and discussions arepresented in section 5. Summarizes our conclusions in section 6.

## 2. BACKGROUND

### *A. Machine Learning (ML) Techniques for Classification*

Here, various methods for analyzingMachine learning-based classifiers. This is shown in

Figure 1

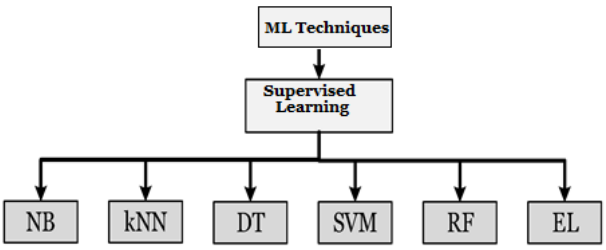



Figure 1. ML techniques for Classification

*Naive Bayes (NB) Classifier:*

Bayes' theorem is applied to estimate the probability of an event occurring from observations of previously observed events. This is useful for classifying normal and abnormal behaviors in supervised learning scenarios. The formula for Bayes' theorem is given as:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad \ldots\ldots\ldots \quad (1)$$

Where, $P(A/B)$ is Posterior probability: Probability of hypothesis A on the observed event B.

$P(B/A)$ Is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

$P(A)$ is Prior Probability: Probability of hypothesis before observing the evidence.

$P(B)$ is Marginal Probability: Probability of Evidence.

*K-Nearest Neighbor (KNN):*

No parameters are required for KNN to function. When it comes to measuring the distance between neighbors, Euclidean distance is used. Some previous studies [7] have successfully used KNN-based classification to process the medical data analysis like heart disease data analysis for classification. The only challenge when using KNN is finding the optimal value of k and identifying missing nodes.

*Support Vector Machine:*

SVM is a supervised learning algorithm. SVM learns from labeled data. Classifying new data is determined by what SVM learned during training. While SVM has numerous benefits, including the capacity to handle classification and regression tasks, it is most notable for being suitable for these two specific areas.

### *B. Convolutional Neural Networks*

CNN's are neural networks with spectral layers to learn features at multiple levels. These examples have demonstrated how CNNs are effective models for predicting future statistics, modeling, and the like. Three additional concepts such as local filters, max-pooling, and weight sharing result in an abundance of power compared to DNNs [8]. Figure 2 visually illustrates the architecture of CNN used for predicting heart disease. CNN is built from a small number of pairs of convolution and max-pooling layers. Whenever a pooling layer is applied, it is always preceded by a convolutional layer. In general, spectrum pooling is used in the frequency domain. Using the max pooling, the results for the variability problem are favorable. A max-pooling layer computes the maximum filter activation across a defined window of positions. Convolutional neural network generation of lower resolution features at this step. Minimizing positioning discrepancies and increasing the speed of convergence are both benefits of using max-pooling. The fully connected layers perform 2-D feature extraction, combining inputs from all positions into a 1-D feature vector. Finally, the softmax activation function is applied to the overall inputs for classification.

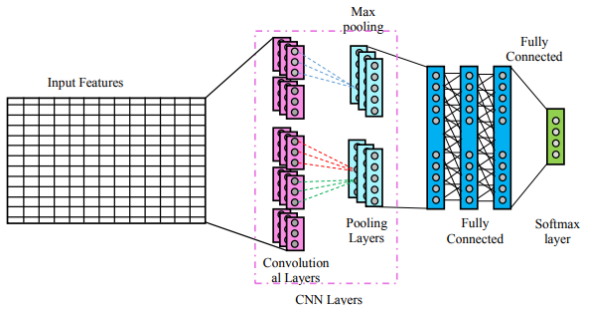



Figure 2. The architecture of CNN for prediction.

## 3. RELATED WORK

Many researchers are actively engaged in researching heart disease prediction.

AwaisMehmood et al. [9] the proposed method, CardioHelp, employs a deep learning algorithm called convolutional neural networks to predict the probability of cardiovascular disease in a patient (CNN). The proposed approach is dedicated to modeling temporal data by utilizing CNN in the earlier stages of HF prediction. The dataset was prepared, and results were compared with state-of-the-art methods, and the results were good. In terms of performance evaluation metrics, the proposed method outperforms the existing methods. The accuracy of the proposed method has been calculated and results in 97 percent.

In [10], ElhamNikookar et al. discussed implementing data science concepts for heart disease detection using hybrid techniques. The author proposed a hybrid model by combining three algorithms. Firstly, they used ANN (Artificial Neural Network) classifier. The results are given to the Support Vector Machine (SVM) classifier, and finally, the results are given to the Naive Bayes classifier to get the prediction results. The proposed model proved high accuracy on the hybrid algorithm. Around 88% has been achieved.

Poornima Singh et al. [11] proposed a heart disease prediction system that utilizes a neural network. The algorithm considered 15 different characteristics for prediction. Using a multilayer perception neural network with backpropagation, the training method employed a

multilayer perception neural network. The dataset was all in a tidy, ordered fashion, and the model delivered 100% accuracy.

Naive Bayes and decision tree data mining techniques were utilized by Gomathi et al. [12] in predicting different types of diseases. They focused on forecasting various types of cardiovascular disease, diabetes, and breast cancer. The confusion metrics provided the results.

Naive Bayes classifier approach for the prediction of cardiovascular diseases was suggested by Miranda et al. [13]. The authors have put forth only a few important risk factors in making the final decision on whether someone has cardiovascular disease. Seventy-five percent of the proposed concept has been accurate, sensitive, and specific to 85% accuracy, sensitivity, and specificity.

## 4. METHODOLOGY

The paper's main idea is to determine whether the patient has heart disease. Not only does predicting the risk of heart disease put the patient at risk, but it also predicts the risk of heart disease, which lays the patient at high risk or low risk.

An accurate prediction becomes difficult due to the missing values found in the dataset, mainly comprised of medical data. So, to accomplish the imputation and data cleaning task, it is necessary to complete this step. After this data imputation, we need to use a data cleaning and data imputation process to transform the missing data into structured data. Once the naïve Bayes SVM and KNN algorithm has been implemented on the input values, the program next classifies all patients based on whether or not they have had a heart attack. The three algorithms mentioned above look at Bayesian naive, SVM, and KNN to aid in classification. However, the classifier that provided the most incredible accuracy in the test dataset yielded the most accurate classification concerning risk assessment using the CNN-UDRP [14] algorithm. Classifier performance is better in the case of the naïve Bayes classifier, and CNN-UDRP thus chooses to use that classifier as input. This system can predict whether the patient suffers from high or low risk by using the CNN-UDRP algorithm. Feature extraction is performed with the Convolutional Neural Network algorithm. The softmax classifier is utilized to determine the likelihood of heart disease.

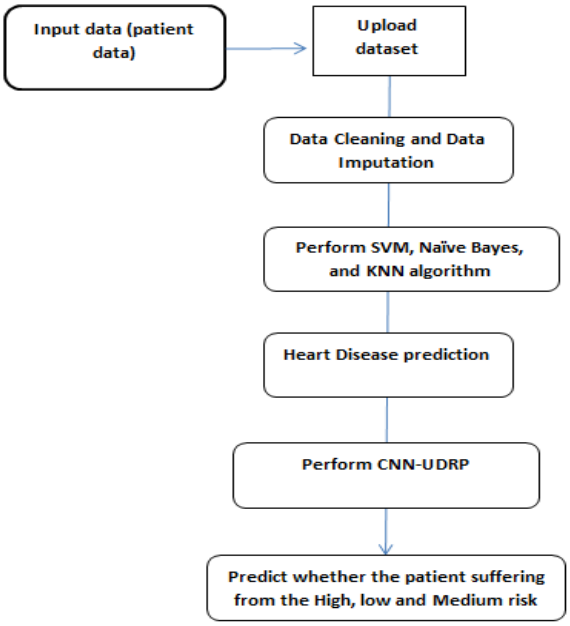



Figure 3. Proposed Workflow model

*Dataset:*we use the dataset from the UCI Repository. This dataset has over 19,392 records and 14 different attributes.

Table 1.Attributes from the heart risk prediction dataset.

| S.No | Attribute Name | Description |
|---|---|---|
| 1 | Age | Age of the patientin a year |
| 2 | Sex | Gender of thepatient |
| 3 | Cp | Chest Pain Type |
| 4 | Trestbps | Blood Pressure |
| 5 | Chol | Cholesterol |
| 6 | Restecg | Resting results of Electrocardiography |
| 7 | Ca | Number of major vessels that colored by fluoroscopy |
| 8 | Fbs | Fast Blood Sugar |
| 9 | Oldpeak | ST segments that induced by exercise relative to rest |
| 10 | Slope | Peak exercise ST segments |
| 11 | Thal | Defect values |
| 12 | Exang | Exercise-induced angina |

*A. Data cleaning and data imputation*

Dataset consists of unstructured data, which means well-formed data is not a part of it. The majority of medical data is in unusable formats. Data cleaning and data imputation are necessary when the missing data are present. Both unwanted and noisy data must be removed from a dataset to obtain a well-structured dataset.

*B.Class Distribution:*
(Class value 1 is interpreted as "tested positive for diabetes")

Table 2 Class distribution table

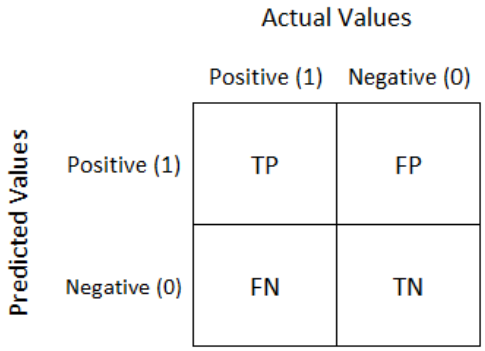
| **Class Value** | **Number of Instances** |
|---|---|
| 0 | 8832 |
| 1 | 10560 |

"Outcome" is the feature we are going to predict. 0 means No Heart diseases, 1 means Heart diseases. Of these 19392 data points, 8832 are labeled as 0 and 10560 as 1:

Finally, let us split the training and test data. The data has been divided into an 80:20 ratio in this paper. 80% of the training data is used in training, and 20% of the testing data is used for validation.

*C.Training:*

The metrics used for th




The confusion matrix shows correctly predicted and incorrectly predicted values and values that are correctly predicted and values that are incorrectly predicted by a classifier. This calculation, known as the total number of correctly classified entries by the classifier, is the sum of TP and TN.

*D.Evaluation Method:*

To verify the classifier's accuracy and efficiency, the essential task is completed after the classifier is finished. Classifiers can be assessed in many ways. We will review these options below.

*Holdout Method*

In many cases, classifiers can be evaluated this way. 20% and 80% respectively. In training the data set, the train set is used, and in testing the predictive power of the train set, the unseen test set is used [15].

*Cross-Validation*

Many machine learning models suffer from over-fitting. To verify if the model is over-fitted, you can conduct cross-validation using K-fold splitting illustrated in figure 4.

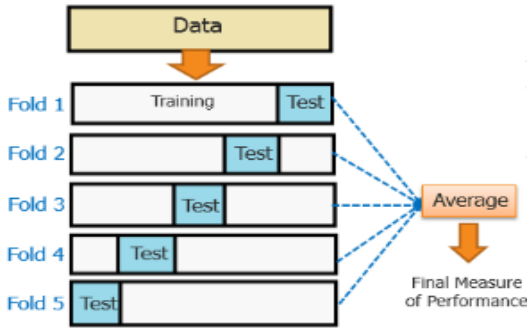



Figure 4. K-fold cross-validation

Each subset is the same size. a portion of these is used to test the model, while the rest are used to train itFor the experimental result, we have notations as follows:

TP: True Positive (the number of instances that the prediction was correct),
FP: False positive (the number of instances where the prediction was inaccurate),
TN: True negative (As expected, the number of cases was underestimated.),
FN: false negative (The number of instances was mistakenly predicted as not being required.).

*E.Classification Report*

Accuracy: Accuracy is a ratio of correctly predicted observation to the total observations
True Positive (TP): The number of correct predictions that the occurrence is positive.
True Negative (TN): Number of correct predictions that the occurrence is negative.
F1- Score: It is the weighted average of precisionand recall

*Precision (P)and Recall(R)*

Precision refers to the proportion of relevant examples found among the retrieved instances, whereas recall refers to the proportion of relevant instances found among the overall occurrences. They are mainly used as a metric for determining significance.

Based on this parameter, we can calculate four measurements.
1. Accuracy (A):A=(TP+TN) / (Total no of samples)
2. Precision (P): P = TP / (TP+ FP)
3. Recall (R): R =TP / (TP+FN)
4. F-Measure (F): F=2*(P*R) / (P+R)

***F.Algorithms used***

*Naive Bayes (NB) Classifier:*The Naive Bayes method is a supervised learning algorithm for addressing classification issues based on the Bayes theorem. It is primarily utilized in text classification tasks that require a large training dataset. The Naive Bayes Classifier is a primary and effective classification method that aids in developing fast machine learning models capable of making quick predictions.

*Support vector machine:* The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. A hyperplane Support vector machine is the optimal decision boundary. Its decision function uses a subset of training points, making memory efficient and very effective in high-dimensional spaces.
*K-Nearest Neighbor (KNN):*This technique is straightforward to develop and is resistant to noisy training data. It is pretty efficient, even if the training data is enormous.

***CNN-UDRP Algorithm****:* We must conduct five phases of the algorithm in order to forecast the risk of heart disease. The dataset is transformed into vector form in the first stage. The data was then filled with zero values using word embedding. The convolutional layer is the result of word embedding. We use this convolutional layer to input the pooling layer and execute a maximum pooling operation on it. The entire connected neural network is coupled to the pooling layer. Last but not least, the softmax classifier is coupled to the complete connection layer.

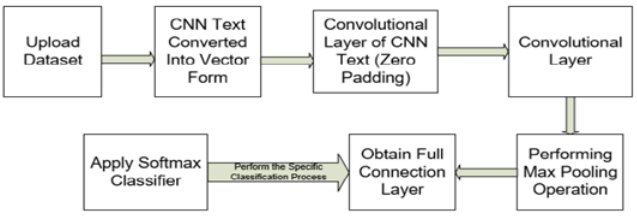



Figure 5:CNN-UDRP Algorithm Steps

## 5. RESULT AND ANALYSIS

On the dataset, we performed heart disease classification and heart disease risk prediction. Training and test datasets are included in the dataset. The dataset is divided into two sections: training, approximately 19392 entries, and test, which contains 3878 entries. The accuracy of the naive Bayes, SVM and KNN algorithms in predicting heart disease is compared.

*A. SVM Accuracy model:*
*Confusion Matrix data*

| | True Positive | True Negative |
|---|---|---|
| Predicted Positive | 1682 | 59 |
| Predicted Negative | 22 | 2116 |

Table 3. Performance of the SVM Classifier Model

| **Measure** | **Value** |
|---|---|

| | |
|---|---|
| Sensitivity | 0.9871 |
| Specificity | 0.9729 |
| Precision | 0.9661 |
| Negative Predictive Value | 0.9897 |
| False Positive Rate | 0.0271 |
| False Discovery Rate | 0.0339 |
| False Negative Rate | 0.0129 |
| Accuracy | 0.9791 |
| F1 Score | 0.9765 |
| Matthews Correlation Coefficient | 0.9579 |

*B.KNN Classifier*

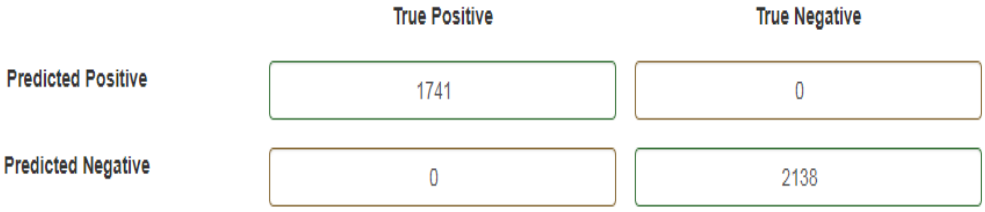



Table 4. Performance of the KNN Classifier Model

| **Measure** | **Value** |
|---|---|
| Sensitivity | 1.0000 |
| Specificity | 1.0000 |
| Precision | 1.0000 |
| Negative Predictive Value | 1.0000 |
| False Positive Rate | 0.0000 |
| False Discovery Rate | 0.0000 |
| False Negative Rate | 0.0000 |
| Accuracy | 1.0000 |
| F1 Score | 1.0000 |
| Matthews Correlation Coefficient | 1.0000 |

*C.NB Classifier*

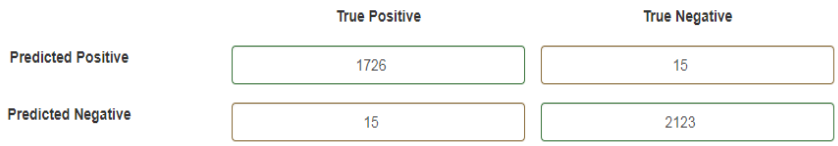



Table 5.Performance of the NB Classifier Model

| **Measure** | **Value** |
|---|---|
| Sensitivity | 0.9914 |
| Specificity | 0.9930 |

| | |
|---|---|
| Precision | 0.9914 |
| Negative Predictive Value | 0.9930 |
| False Positive Rate | 0.0070 |
| False Discovery Rate | 0.0086 |
| False Negative Rate | 0.0086 |
| Accuracy | 0.9923 |
| F1 Score | 0.9914 |
| Matthews Correlation Coefficient | 0.9844 |

Performance of The Classifiers among Heart diseases Dataset

Table 6. Performance of The Classifiers among Heart diseases Dataset

| **Classifiers** | **Accuracy (%)** |
|---|---|
| SVM | 0.97 |
| KNN | 100 |
| Naïve Bayes | 0.99 |

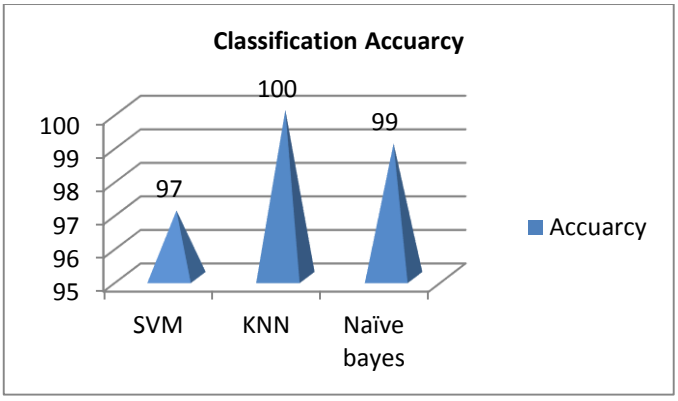



Figure 6. Overall Classification accuracy

Compared to the Support vector machine and the Naive Bayes classifiers, KNN provides high Classification Accuracy, i.e., 100 percent.

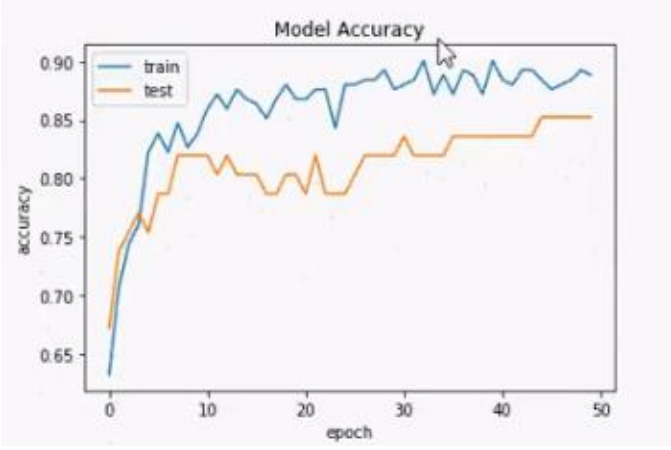



Figure 7.CNN-UDP model accuracy plot diagram

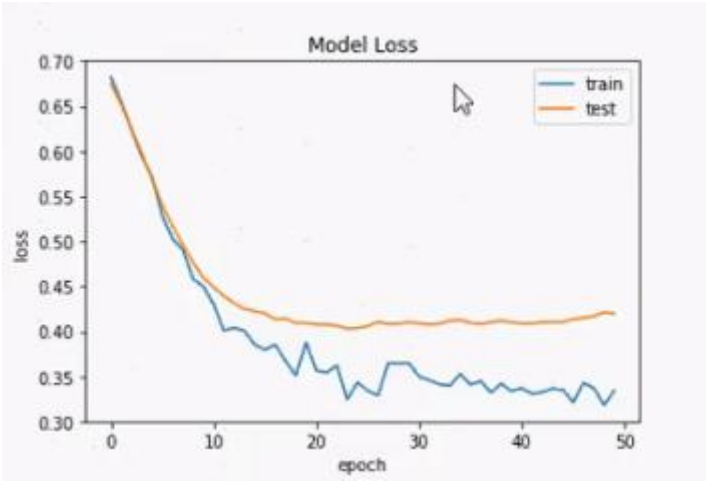



Figure 8. CNN-UDP model loss plot diagram

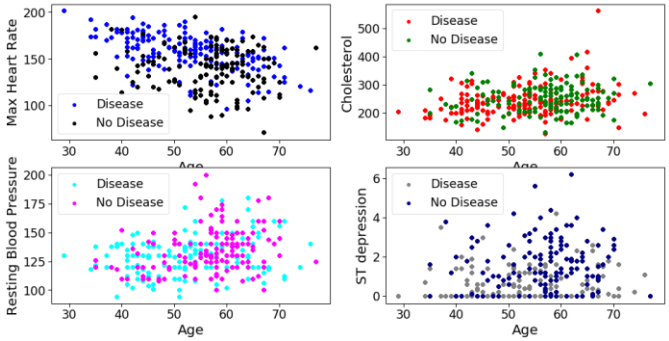



Figure 9. Frequency of Heart diseases Prediction with age vs max heart rate, age vs Cholesterol, age Vs BP, age vs. depression

*Overall Prediction Accuracy*

Table 7. Overall Prediction Accuracy

| **Prediction Model** | **Train Score** | **Test Score** | **Precision** | **Recall** | **f1** | **Accuracy (%)** |
|---|---|---|---|---|---|---|
| KNNeighbors Algorithm | 0.88 | 0.54 | 0.95 | 0.47 | 0.54 | 63.49 |
| SVM Algorithm | 0.82 | 0.64 | 0.87 | 0.66 | 0.64 | 75.67 |
| Naive Bayes Algorithm | 0.72 | 0.6 | 0.86 | 0.61 | 0.6 | 72.22 |
| CNN-UDRP | 1.0 | 0.72 | 0.91 | 0.73 | 0.72 | 81.57 |

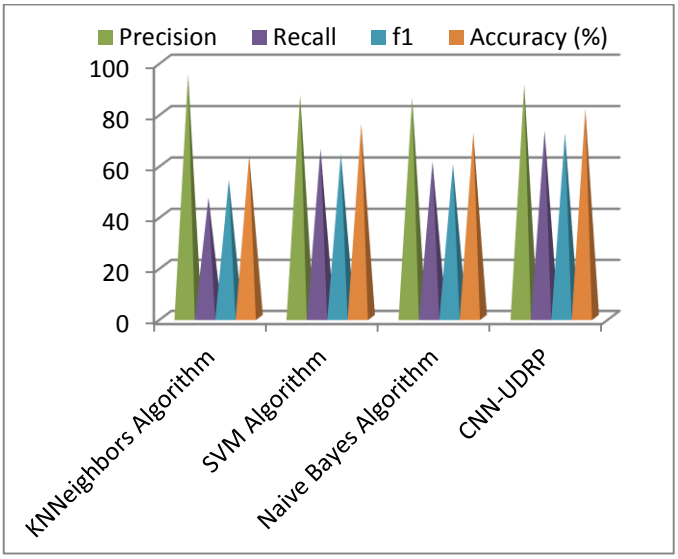



Figure 10. Overall prediction accuracy

In Figure 10, several characteristics such as precision, recall, F1-measure, and accuracy are shown. Heart disease risk prediction accuracy is close to 81.57 percent. The recall and precision scores are quite close to 73 and 91 percent, respectively. As a result, CNN-UDRP outperforms traditional machine learning-based classifiers in terms of heart disease risk prediction. The result is presented to the patient in the form of a percentage with high, low, and medium risk levels.

## 6. CONCLUSION

To conduct this experiment, we have used structured data for disease risk prediction with CNN-UDRP. We used the naïve Bayes algorithm, the Support Vector Machine (SVM), and KNN to perform heart disease prediction. We find that KNN provides a classifier with a Classification Accuracy of 100%. Concerning heart disease risk prediction, the CNN-UDRP is a better method for using machine learning techniques. We were provided with a correct disease risk prediction, which was produced as an output, thanks to our input, which consisted of patients' reporting, which allowed us to comprehend the level of disease risk prediction. The heart disease risk is predicted to be low, medium, and high. Low time consumption and minimal cost make this system especially effective for disease risk prediction. In the future, we will include even more diseases and calculate the likelihood that a specific patient will contract that disease.

## 7. REFERENCES

[1] Benjamin, E. J. et al. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association. American Heart Association, 139, 56–528. https://doi.org/10.1161/CIR.0000000000000659.

[2] Fava, A., Plastino, M., Cristiano, D., Spanò, A., Cristofaro, S., Opipari, C., Chillà, A., Casalinuovo, F., Colica, C., Bartolo, M. D., Pirritano, D. & Bosco, D. (2013). Insulin

resistance possible risk factor for cognitive impairment in fibromialgic patients. Metabolic Brain Disease, 28(4), 619–627. https://doi.org/10.1007/s11011-013-9421-3.

[3] Nakanishi, R., Dey, D., Commandeur, F., Slomka, P., Betancur, J., Gransar, H., Dailing, C., Osawa, K., Berman, D. & Budoff, M. (2018) Machine learning in predicting coronary heart disease and cardiovascular disease events: Results from the multi-ethnic study of atherosclerosis (MESA). Journal of the American College of Cardiology, 71(11), Supplement. https://doi.org/10.1016/S0735-1097(18)32024-2.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[5] Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," inIEEE big data analytics and computational intelligence, Oct 2017 pp.23-25.

[6] Disease and symptoms dataset –www.github.com.

[7] Zhang, Shichao. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software. 85. 2012,2541–2552. 10.1016/j.jss.2012.05.073.

[8] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in PatternRecognition (ICPR), 2012 21st International Conference on, 2012, pp. 3288-3291.

[9] Mehmood, A., Iqbal, M., Mehmood, Z. et al. Prediction of Heart Disease Using Deep Convolutional Neural Networks. Arab J Sci Eng 46, 3409–3422 (2021). https://doi.org/10.1007/s13369-020-05105-1

[10] Nikookar, Elham, and Ebrahim Naderi. "Hybrid Ensemble Framework for Heart Disease Detection and Prediction." International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, 2018.

[11] Singh, Poornima, et al. "Effective heart disease prediction system using data mining techniques." International Journal of Nanomedicine, vol. 13, 2018, pp. 121-124.

[12] Gomathi K, ShanmugaPriyaa D. Multi disease prediction using data mining techniques. Int J Syst Softw Eng. 2016;4(2):12–4.

[13] Miranda E, Irwansyah E, Amelga AY, Kom S, Maribondang MM, Kom S, Salim M, Kom S. Detection of cardiovascular disease risk's level for adults using Naive Bayes classifier. Healthc Inf Res. 2016;22(3):196–205.

[14] S. Ambekar and R. Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697423.

[15] Ajitesh Kumar. "Holdout Method for Training Machine Learning Models." https://vitalflux.com.20. Accessed 22 Dec. 2020.