Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python

Saiyed Faiayaz Waris, S. Koteeswaran

Dept. of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India

ARTICLE INFO

Article history:
Received 8 January 2021
Accepted 19 January 2021
Available online xxxx

Keywords:
K-means neighbour
Improved k-means neighbour
Heart disease
Prediction
Accuracy and python

ABSTRACT

The various existed machine learning classifiers are defined for prediction of early heart disease than the schedule of it, the improved version of K-means neighbour classifier is used that guarantees the more accuracy than actual K-means neighbour classifier and other related classifiers. The terms that influence the heart diseases significantly smoking, food habits, diabetes, blood pressure and other related. In such scenarios, a specific approach or hybrid combination of predicting approaches are required to predict the heart disease very early. In the stages of predicting the heart disease, the classifier is the fourth stage which is significant stage for achieving accuracy, sensitivity, and specificity. When k-means neighbour classification is compared, the ideology titled heart diseases prediction very early using improved k-means neighbour classifier is more efficient in the processing and computing the accuracy. The improved k-means neighbour in the python environment is illustrated with few information sets and produces the output with more accuracy when compared with actual k-means neighbour classifier. The working of the proposed system with architecture and ER Diagrams are defined in the methodology along with pseudo procedure.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

The factors that influence the heart are in more number. The factors include bad habits and irregular food habits may cause other diseases such as blood pressure and diabetes. These diseases affect blood pumping and may influence the heart. The heart may be attacked in many ways which are assumed as heart diseases Fig. 1. Fig. 2 Fig. 3 Fig. 4 Table 1.

The kind of heart diseases are identified as shown as below:

1) Coronary (or) Valvular describes the blood pumping vessels are damaged and stops blood to go to the heart.

2) Hypertension indicates that condition of blood pumping is high towards artery walls.

3) Cardiac Arrest indicates that sudden error about the heart functioning and consciousness.

4) Heart Failure represents heart doesn't pump blood.

5) Arrhythmia represents heart functions as irregular like too fast or too low.

6) Peripheral artery indicates that condition of pumping the blood from narrowed vessels to the limbs.

7) Stroke represents damage to the brain by stopping the pumping of the blood.

8) Congenital represents abnormality raised in the heart before birth itself.

9) Any combination of above few factors also cause heart to not function properly or leads to death.

The kinds of attacks to the human heart are specified in the above, there may be chance of sudden failure or malefic of the heart because of blood pressure and diabetes. Not only by these two, many other factors may influence the heart functioning. Whenever the heart won't function well, it leads to financial crisis and loss of the death of that person. Not only above list of factors which impact the heart attack or disease, may be combination of above specified factors also may influence the heart.

There are many methodologies that are used in heart disease prediction where such methods are applied over a lot of samples and few required attributes. With reduced set of attributes, the analyzation is to be done and yields the reports with respect to the accuracy importantly. Any method is considered as best if its performance is better and is providing best accuracy. Here, the

https://doi.org/10.1016/j.matpr.2021.01.570
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

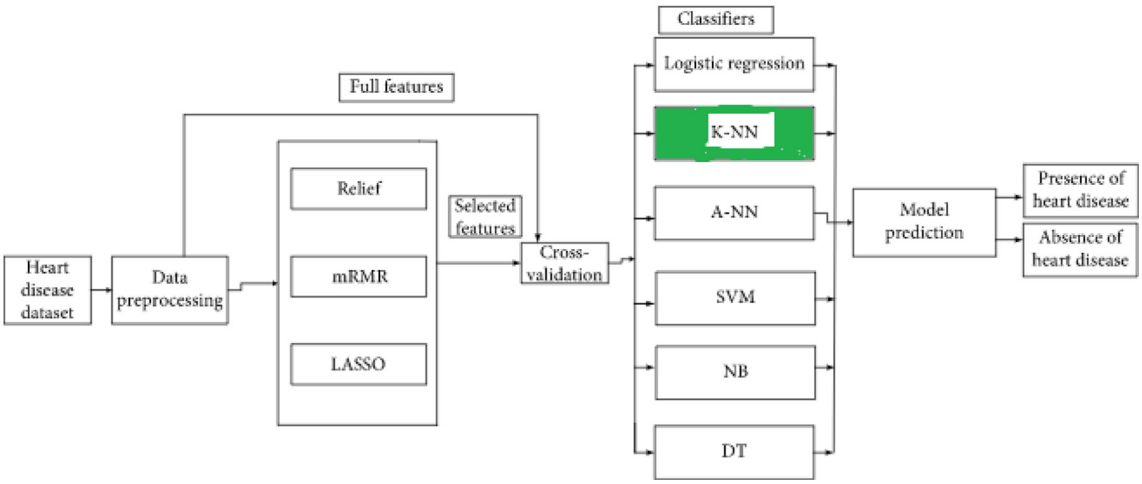

Fig. 1. The stages in the prediction of heart disease.

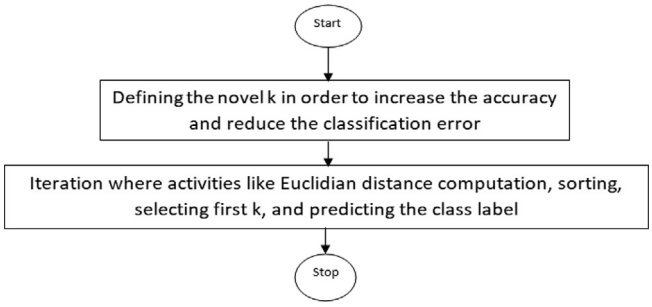

Fig. 2. Flow chart of novel KNN approach.

review is made on predicting the heart problem in advance than naturally occurred accidentally so that the patient or user would be under care for handling the situation and save the life of that person. The importance of this task is to save the life by predicting the heart disease over a few significant parameters and aims to obtain more accuracy than the existing methodologies or approaches. There were many studies taken as an importance in which the more ML methodologies are used and analyzed and their descriptions are discussed in the literature review chapter. Any study that is demonstrated and listed in literature review is focused on the kind of classifier used in their study and how much accuracy they are producing. Among the mentioned ML and together DM approaches, k-means neighbor classification is proved the best in its kind in predicting the heart disease very early with more accuracy. Hence, we took the k-means neighbor classifier as seed and modified it in the reliable manner such that it would produce some more accuracy than the existing approach accuracy. For running the scripts of the proposed algorithm, the environment suitable is python with few predefined libraries. The classification is the fourth stage in the process of five stages where the first stage is data preprocessing where representation of dataset is possible in efficient manner, second stage is feature selection where RELIEF algorithm, mRMR and Least absolute shrinkage and selection operator(LASSO) are used, third stage is cross validation where k-fold cross validation is applying k times and takes the average of all, fourth stage is classification where logistic regression, naïve bayes, Artificial neural network, decision tree classifier, and K-nearest classifier are used and the fifth stage is prediction of class where the class label is computed such as negative means 0, otherwise 1 and there were predefined formulas for True accuracy, False accuracy by classification error, specificity, sensitivity, precision and prediction ability of the classifier using MCC, its graph curves are monitored by ROC and AUC measures of the classifier.

The below five stage approach where our field of study is specifically on classification in which significantly k-means neighbor classifier.

The following is the pseudo procedure of k-means nearest neighbor classification:

Step1: Feed the dataset

Step2: Fix the value for k

Step3: Iterate till the last sample in the training set for the correct prediction

Step 3.1: Compute Euclidian distance between test and training sample

Step 3.2: Arrange the distances computed in ascending order

Step 3.3: Choose the first k rows from the sorted dataset

Step 3.4: Consider the rows that have most frequent class as label

Step 3.5: Output the most frequent class

The benefits of this approach is easy to implement, flexible for no need of building the model to tune with additional parameters, and is mutable for classification, regression, and searching. The disadvantage of this it becomes slower when number of samples would be increased.

2. Literature review

There were many studies happened and undergone on the heart disease prediction. In which certain studies are analyzed for bringing the new style of determining the heart diseases in the early stages.

With respect to the resource mentioned in [1], Heart is very significant for any living mammal. The diseases of it are to be many ways. The reasons are more in factors such as smoking, unbalanced diets, daily habits of food and drinking, Cholesterol level, Blood pressure, diabetes and others. This approach used Data mining method called k-nearest neighbor algorithm because of its proven accuracy when compared with few other contemporary methods. Later, the proposed algorithms with certain arguments are given as input to the web app that was developed in python using flask and piggle packages. In future, may be any advanced or improved algorithms are used in the aspect of enhancing the accuracy of prediction. According to the source specified in [2], there are machine learning algorithms used to predict the heart disease especially

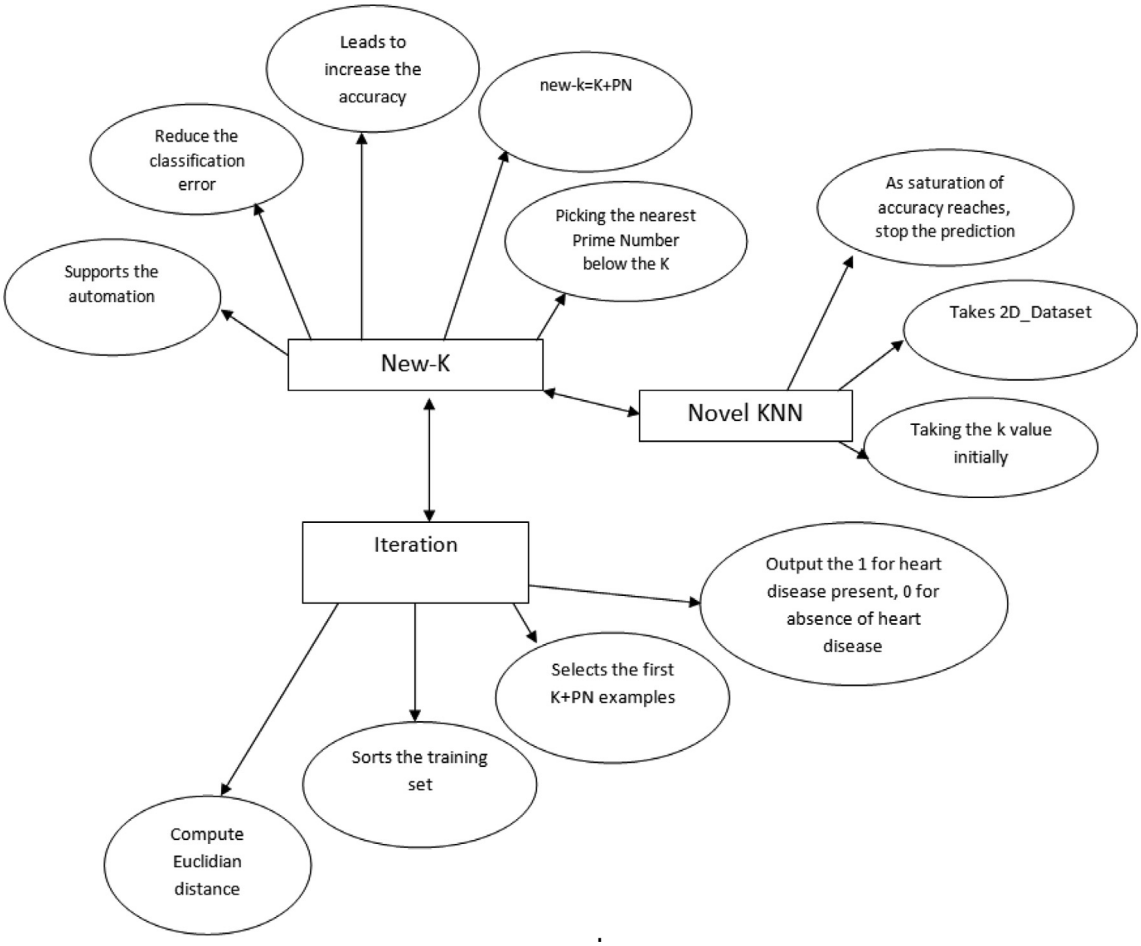

Fig. 3. Architecture of Novel KNN method.

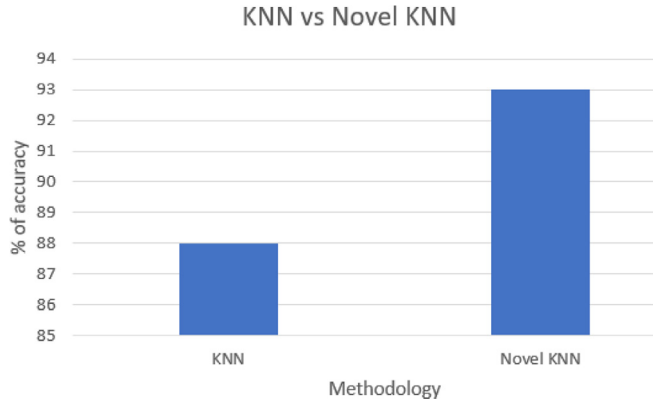

Fig. 4. Accuracy of KNN and Novel KNN.

Table 1
Statistics of KNN and Novel KNN over 14 attributes of Cleveland dataset.

Methodology	Examples	Accuracy
KNN	303	88
Novel KNN	303 + 301 = 604	93

coronary illness. The approaches used in this review, Support Vector Machines are proved with best accuracy. It would be demanding based on the people living in the villages where hospitals and facilities available. The parameters notified are stored in a google sheet and are observed for preparing and testing the information. With the data provided in [3], the bulletins which cause heart disease. To predict the disease, the data mining and machine learning techniques such as k-means neighbor, random forest, Naive bayes, and decision trees. It is applied over certain samples and against certain attributes in which 14 attributes are found necessary to analyze. This demonstrate the which factors represent the high probability of causing the heart disease. Among the reviewed methods, the k-means neighbor gives highest accuracy in the prediction of heart diseases early. As per the information given in [4], there are data mining techniques such as naive bayes and DT methodologies that produce best accuracy in determining the heart disease. The various medical parameters are used in evaluating the heart disease. With respect to [5], there are machine learning methods such as k-nearest neighbor, decision tree, linear regression and support vector machine (SVM) over a UCI repository dataset for a training and testing. The best environment found for analyzation and programming is python anaconda (Jupytor). As per [6], there are various classifiers defined over a heart disease are SVM , Adaboost , J48 Decision tree, K-NN, Naive Bayes, JRip, Stochastic Gradient Decent (SGD) and Decision Table (DT). The comparative analyses of such methods are defined in order to predict the heart disease (HD) cases with minimal attributes. This study has taken samples over the countries such as Cleveland, Hungary, Switzerland, and Long Beach. As per description men-

tioned in [7], there is clinical patient parameters are given, predicting the patient has heart disease or not to be determined using various machine learning classifiers in which k-means nearest neighbor classification is proved with best accuracy 86% in both training and testing aspects. As the dataset mentioned in [8], there are certain tools, equipment to determine and analyze the heart diseases using various machine learning and data mining classifiers. The environment is also specified in terms of python because it has extensive and supportive libraries for heart disease prediction. With the data provided in [9], the sign of the sample is calculated in terms of distance to the nearest sample, K-NN extends this idea to k nearest samples. It gets the sign of the majority of computed k samples. Larger the k value, that consist of reduce less noisy points within the training set. If 10-fold cross validation is used for given dataset, 90% is the training set and 10% only is the test set. As per the resource given in [10], k-NN is described in terms of pseudo procedure. It is also demonstrated with two examples in which one is over a cluster where output is to be the pineapple is on the pizza or not, and other is finding the 5 recommendations that are similar to the post movie and sends to the MovieDB website. With the view of source described in [11], the description of k-means neighbor classification is provided and their pseudo codes are defined in few programming environments. In the view of [12], the implementation of k-means neighbor would be done in python as well as R programming environments. The value of k should be optimal based on validation error and training error. As per the data provided in [13], the various distance functions are defined over a sample of points. The optimal k value is to be picked based on training data set and validation data set. The credit assessment is an example and is taken as demonstration on age and loan. As per the information noticed in [14], the description and implementation of k-nearest neighbors are discussed. With reference to [15], the algorithms used in this are used to predict the heart disease in which logistic regression is found more accuracy over random forests, ANN and KNN. Along with more accuracy in the prediction, minimum error rate it guarantees compared with other classifiers used. As per [16], the cardiovascular heart disease to be predicted with 92% accuracy using a combination of methods such as random forests and linear regression. Without any equipment, the accuracy is computed by applying the combination of existing machine learning techniques. With the data provided in [17], the various machine learning algorithms are used in determining the heart disease. The purpose of predicting in advance is medical practitioners would provide better treatment and avoid mishaps. As per the dataset given in [18], the various machine learning algorithms are used in order to predict the heart disease. The review is done over such algorithms and these methods are compared based on accuracy. With regard to [19], the data generated from clinical approach and EHR results huge amount of records. To analyze such samples, the machine learning and the deep learning techniques are used for extracting valuable information for decision making. In this also, the heart disease prediction is made by various machine learning and deep learning approaches where 60% by former and 30% by latter are involved. As per the resource mentioned in [20], as the k value increases, the accuracy would be increased that could be explained in step by step in this approach. As the saturation of accuracy is reached, the further process is to be stopped. As per the dataset given in [21], the machine learning methods such as KNN and genetic algorithm is used in order to get more accuracy than other existing methods. The method defined here is optimal in diagnosing the heart disease. As per the source demonstrated in [22], there are many classification methods are defined for getting accuracy. One of the factors assumed in the stages of class prediction is feature selection. Specifically PSO is used that removes the noise so that accuracy is increased in which classifier assumed is KNN process. With regard of [23], this study aims at choosing best k value based on error rate or accuracy of the model. The detailed steps of the KNN are demonstrated in python environment. In regard of [24], there are certain restrictions to install the python with certain libraries on platforms such as Windows, Mac OS and linux. Here, the detailed steps and descriptions are provided how to install suitable libraries after installing the python on the specified platforms.

The advantages as well as disadvantages are mentioned in which one of the disadvantage mentioned is k should be known to proceed further. All these approaches are somehow used for heart disease prediction early and accuracy is matter and plays a significant role. There were varieties in k-NN such as globally adaptive NN, Locally adaptive NN, FML NN which is scalable machine learning on big data, fuzzy KNN and few others are used in different applications but not in the case of heart disease prediction.

The description of proposed system is provided in the next chapter called proposed approach where the architecture, pseudo procedure is provided. This proposed system aims to get more accuracy than existing k-means neighbor approach.

3. Proposed approach

The improved k-means neighbor is discussed and is compared with actual k-means neighbor classifier. The architecture of proposed method and its pseudo procedure is also to be described in this session. The pseudo code of actual k-means nearest neighbor is defined in introduction chapter. Based on that, the few steps are added in the novel k-means nearest neighbor approach. The enhanced k-means NN approach is refined as follows:

Pseudo_procedure Novel_KNN(2D_dataset[][],k):

Input: k and 2D_dataset

Output: 1 means heart disease present or 0 means absence of heart disease

Step1: Fix the value for k

Step2: new k is defined which is actual k + prime_number below k value i.e. new_k = K + PN

/* This would increase the accuracy because of examples are increased and classification error rate also is reduced. For example, if k is 25, PN is 23, so new_k is 25 + 23 = 48. The reason behind taking prime number is it is powerful in set-righting any situations */

Step3: Iterate to the last sample in the training set in order to predict the class

Step 3.1: Compute Euclidian distance between test and training sample

Step 3.2: Arrange the distances computed in ascending order

Step 3.3: Choose the first K + PN rows from the sorted dataset

Step 3.4: Consider the rows that have most frequent class as label

Step 3.5: Output the most frequent class

The modules identified in the novel KNN approach is as follows:

1. **Redefining new_k:** It is the first module where new_k = K + PN where k is number of times the training examples are mixed with test set. The prime number is powerful number that set-righted many such difficult scenarios in the many real time applications. In this process, the computing is stopped when saturation in accuracy is reached beyond k and within the K + PN value.
2. **Iteration:** For each step in the repetition, the activities such as computing Euclidian distance between the examples, arrange the examples in sorted order, selecting first k rows from the table, prediction of class label based on most frequent labeled examples.

The flow chart of a novel k-means NN in terms of modules is demonstrated as below:

The activities involved for first module new-k include picking the nearest prime number below the k value, finding new_k value, increases the accuracy, decreases the classification error, and supports tuning to additional parameters and automation. The activities involved for second module include distance computing the training example and test set example using euclidian, sorting the examples, selecting the top K + PN examples, and output the prediction of the example.

4. Results

The screen shots of the k-means neighbor and improved k-means neighbor are to be demonstrated in this session. The factor that is used to show the differentiation is accuracy.

In python environment, the steps are simplified and rewritten as follows:

1) Data Handling: In this, reader function is used over the patient dataset that would be opened using open().

The following code that performs the assigned task:

```
with open(r'C:clevelandpatient.data.txt') as csvfile:
    lines = csv.reader(csvfile)
    for row in lines:
        print (', '.join(row))
```

For splitting, the following code is provided:

```
import csv
import random
def handleDataset(filename, split, trainingSet=[] , testSet=[]):
    with open(filename, 'r') as csvfile:
        lines = csv.reader(csvfile)
        dataset = list(lines)
        for x in range(len(dataset)-1):
            for y in range(4):
                dataset[x][y] = float(dataset[x][y])
            if random.random() < split:
                trainingSet.append(dataset[x])
            else:
                testSet.append(dataset[x])
```

For handling the dataset, the following code is used:

```
trainingSet=[]
testSet=[]
handleDataset(r'iris.data.', 0.66, trainingSet, testSet)
print ('Train: ' + repr(len(trainingSet)))
print ('Test: ' + repr(len(testSet)))
```

2) Calculation of distance: For summing up the examples, the following code is provided.

```
import math
def euclideanDistance(instance1, instance2, length):
    distance = 0
    for x in range(length):
        distance += pow((instance1[x] - instance2[x]), 2)
    return math.sqrt(distance)
```

3) K nearest neighbors finding: For a given test instance, finding the k nearest neighbors are done through the following code:

```
def getKNeighbors(trainingSet, testInstance, k+PN):
    distances = []
    length = len(testInstance)-1
    for x in range(len(trainingSet)):
        dist = euclideanDistance(testInstance, trainingSet[x], length)
        distances.append((trainingSet[x], dist))
    distances.sort(key=operator.itemgetter(1))
    neighbors = []
    for x in range(k):
        neighbors.append(distances[x][0])
    return neighbors
```

4) Class prediction: To get the majority of voting response form a more number of neighbors, the following code is used.

```
def getResponse(neighbors):
    classVotes = {}
    for x in range(len(neighbors)):
        response = neighbors[x][-1]
        if response in classVotes:
            classVotes[response] += 1
        else:
            classVotes[response] = 1
    sortedVotes = sorted(classVotes.items(), key=operator.itemgetter(1), reverse=True)
    return sortedVotes[0][0]
```

5) Validating the accuracy: To provide accuracy over a correct number of examples, the following code is used.

```
def getAccuracy(testSet, predictions):
    correct = 0
    for x in range(len(testSet)):
        if testSet[x][-1] is predictions[x]:
            correct += 1
    return (correct/float(len(testSet))) * 100.0
```

The demonstration of the accuracy difference between the KNN and Novel KNN in terms of accuracy obtained.

The accuracy is to be increased with the increase in value of k although it causes the approach becomes slower. There is no limitation for storing in the memory in the today world.

The following graph shows the accuracy of prediction of heart disease when charted KNN and Novel KNN is:

5. Conclusion

When reviewed few classification approaches in which k-means neighbor classifier is the best one with proved accuracy. There is a novel method defined such as a new k-means neighbor classifier which gets more accuracy and is better in performance compared with actual k-means neighbor classifier. The accuracy when gets saturated, the process of computing for correct examples is stopped that itself would enhance performance. By taking nearest prime number below k value, new k is computed and applied as input to the Novel KNN approach. This derived approach might be used in predicting the heart diseases but also many other simple to complex applications as well. Irrespective of intensity of application, the accuracy is improved as well as classification and validation errors are to be reduced. In future, hybrid frameworks are possible that might increase still accuracy in some percentage.

CRediT authorship contribution statement

Saiyed Faiayaz Waris: Conceptualization, Methodology, Software, Visualization, Writing - original draft. **S. Koteeswaran:** Data curation, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Keshav Srivastava, Dilip Kumar Choubey, Heart Disease Prediction using Machine Learning and Data Mining, May,2020, DOI: 10.35940/ijrte.F9199.059120.

[2] Honey Pnadey,S.Prabha, Smart Health Monitoring System using IOT and Machine Learning Techniques, January 02,2021, IEEE Explore

[3] Devansh Shah, Samir Patel & Santosh Kumar Bharti, Heart Disease Prediction using Machine Learning Techniques, October, 2020, SN Computer Science volume 1, Article number: 345(2020),https://link.springer.com/article / 10.1007/s42979-020-00365-y

[4] K.J. Santhana, S. Geetha, Prediction of Heart Disease Using Machine, Learning Algorithms ICIICT (2019).

[5] Archana Singh, Rakesh Kumar, Heart disease prediction using machine learning algorithms, ICE3 (2020).

[6] Khaled Mohamad Almustafa, Prediction of heart disease and classifiers' sensitivity analysis, 02 July 2020, BMC Bioinformatics volume 21, Article number: 278 (2020), https://bmcbioinformatics.biomedcentral.com /articles/ 10.1186/s12859-020-03626-y.

[7] Predicting heart disease using machine learning, Python notebook using data from Heart Disease UCI, https://www. kaggle. com /faressayah/predicting-heart-disease-using-machine-learning

[8] Amit Chauhan, Heart Disease Prediction using Machine Learning with Python, October 2020, https://towardsai.net/p/machine-learning/heart-disease-prediction-using-machine-learning-with-python

[9] Nearest Neighbour Classifier, https://www.robots.ox.ac.uk/~dclaus/digits/neighbour.htm

[10] Onel Harrison, Machine Learning Basics with the K-Nearest Neighbors Algorithm, September, 2018, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[11] K-Nearest Neighbours, https://www.geeksforgeeks.org/k-nearest-neighbours/

[12] TAVISH SRIVASTAVA, Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R), https:// www. analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[13] K Nearest Neighbors - Classification, https://www.saedsayad.com/ k_nearest_neighbors .htm

[14] M.W. Kenyhercz, N.V. Passalacqua, Missing data imputation methods and their performance with biodistance analyses, biological distance, Analysis (2016).

[15] S.Raguvaran, R.Anandhi , A.Anbarasi, T.Megala, Heart Disease Prediction Using Hybrid Machine Learning Algorithms, Vol. 13 No. 01 (2020): Vol 13 No 1 (2020), http://sersc.org/ journals/index.php/IJGDC/article/view/26283

[16] Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy, Rajalakshmi, Heart Disease Prediction Using Machine Learning Techniques, April, 2020, https://www.irjet.net/archives/V7/i4/IRJET-V7I4993.pdf.

[17] Dr Dilbag Singh, Jasjit Singh Samagh, A COMPREHENSIVE REVIEW OF HEART DISEASE PREDICTION USING MACHINE LEARNING, ISSN- 2394-5125, Vol 7, Issue 12, 2020, http://www.jcreview.com/fulltext/197-1592483859.pdf.

[18] M. Kamboj, Heart disease prediction with machine learning approaches, Int. J. Sci. Res. (IJSR), ISSN 2319–7064 (2018).

[19] Kusuma.S, Divya Udayan.J, Machine Learning and Deep Learning Methods in Heart Disease (HD) Research, Volume 119 No. 18 2018, 1483-1496, ISSN: 1314-3395 (on-line version), https://acadpubl.eu/hub/2018-119-18/2/116.pdf

[20] Nagesh Singh Chauhan, Building Heart disease classifier using K-NN algorithm, https://www.kdnuggets.com/2019/07/classifying-heart-disease-using-k-nearest-neighbors. html/2

[21] M.Akhil jabbar, B.L Deekshatulu, Priti Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, Procedia Technology 10 (2013) 85 – 94, https://arxiv.org/ftp/arxiv/papers/1508/1508.02061.pdf.

[22] Jabbar MA, Prediction of heart disease using k-nearest neighbor and particle swarm optimization, Biomedical Research (2017) Volume 28, Issue 9, https://www.alliedacademies.org/articles/prediction-of-heart-disease-using-knearest-neighbor-and-particle-swarm-optimization.html
[23] Tharuka Sewwandi, Predicting Cardiovascular Disease Using K Nearest Neighbors Algorithm, September,2020, https://towardsdatascience.com/predicting-cardiovascular-disease-using-k-nearest-neighbors-algorithm-614b0ecbf122
[24] Jason Brownlee, How to Setup Your Python Environment for Machine Learning with Anaconda, September, 2020, https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/.