



## Predictive Model to Diagnose Heart Disease using a novel approach to handle heterogeneous data

Sujata Joshi<sup>1</sup>, Mydhili.K.Nair<sup>2</sup>,

<sup>1</sup>Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India

<sup>2</sup>School of Computer Science, RV University, Bangalore, India

---

**Article History:** Received: 12.07.2023      Revised:02.08.2023      Accepted: 15.08.2023

---

### ABSTRACT

In the recent times, healthcare sector is using data mining for a variety of tasks out of which, one of them is predictive modelling. The role of data is of utmost important in any data mining task. In order to perform predictive modelling, data is collected, pre-processed, and then predictive model is developed using appropriate learning algorithms. It is observed that medical datasets are inherently heterogeneous having a mix of categorical, nominal, binary, numeric and non-numeric types. Dealing with such heterogeneity is a challenge for mining of heterogeneous datasets. This work aims to design a predictive model for diagnosis of heart disease taking into consideration, the data heterogeneity. In this work, the CVD dataset from Cleveland database of UCI repositories and Echocardiogram data collected from hospital is utilized. The predictive model is developed using a novel approach which employs K Nearest Neighbour learning algorithm which takes into consideration the type of data and then constructs distance measures accordingly. The results are compared with the two learning algorithms namely Decision Tree and Naive Bayes which can handle data heterogeneity inherently. It is found that the novel approach has given promising results. The accuracy of the model developed using the novel approach is found to be 88% as compared to 81% given by baseline k-NN learning algorithm.

**KEYWORD:**Heart disease,Prediction,K-Nearest neighbour,Heterogeneous data

---

**DOI: 10.48047/ecb/2023.12.8.653**

### Corresponding Author:

Dr. Sujata Joshi

Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India

Email: sujata.joshi@nmit.ac.in

### INTRODUCTION (10 PT)

Healthcare data is collected from electronic medical records, various images, patient interviews, reports, laboratory tests, physician observations and interpretations. This data may be heterogeneous in nature as it may contain different types and forms of data. Also the data may be structured, unstructured or semi-structured form. This requires efficient mining on complex data types, storage devices of high capacity, novel tools for analysis of such data and visualization techniques and computer translation for processing physician's interpretation [1,2]. All these pose an additional challenge to the data mining task.

Predominantly classification is a vital data mining task. Typically a medical dataset is heterogeneous in nature as it is made up of values having different attribute types such as numeric, non numeric, binary, categorical or ordinal types. The major challenge in mining medical datasets containing heterogeneity is how to develop predictive models which can handle heterogeneity of data[3]. Some existing learning algorithms like Decision trees and Naïve Bayes techniques can handle heterogeneity, but still there is a scope to develop specialized techniques taking into consideration the inherent type of data [4].

In case of Coronary heart disease(CHD), the blood vessels are narrowed down or clogged because of the deposit of plaque on the interior walls of the blood vessels. This condition is called Atherosclerosis. As the build-up progresses, the blood vessels are further narrowed down, stopping the flow of the required amount of oxygenated blood to the heart. In extreme conditions, this may cause heart attack or stroke[5].

CHD has emerged as the major cause of deaths worldwide. It is noted that 1970 onwards, the death rate as a consequence of heart diseases has reduced in developed countries. Nevertheless, mortality rate due to heart diseases has escalated at a much faster pace in developing countries. However, the number of people having heart disease has been always on the rise globally which is a matter of concern[6].

Though heart disease is found to affect older adults generally, young people also are equally vulnerable to be affected. The symptoms for the underlying heart disease may begin at an early age, which makes initial efforts of prevention necessary from a young age. In this regard, a predictive model to predict and diagnose heart disease, would be useful which could be used to assess the health of the heart and thereby the individual[7,8].

Owing to the advances in information technology, data mining has found crucial importance in the early diagnosis of diseases. Though it is applied in varied areas, its role is becoming more prominent in healthcare industry. Techniques are developed to diagnose and predict of diseases based on available information. In this regard, various models are developed for diagnosis of heart diseases. Though many other researchers have also developed models related to heart diseases, the aspect of data heterogeneity is not taken into account in their models explicitly[9,10].

Although Decision trees and Naïve Bayes techniques can handle heterogeneity, they have their own drawbacks. Since K Nearest Neighbour learning algorithm uses distance measures to find similar neighbours, there is a scope to introduce the concept of computing distance measures based on the type of attribute[11,12].

Several measures are proposed to compare data objects of homogeneous type. For numerical data, the most popular distances used are Minkowski distance which is a generalized form of Euclidean, Manhattan, and L-supremum distance. For categorical data types, the distance measures commonly use are Simple matching, Jaccard, Cosine and Extended Jaccard. Ordinal data types can also be treated in a similar way as numeric or categorical, but the order or the rank has to be established[13,14]. But when the data objects are of mixed types, they have to be dealt with in a different way.

The researchers have used the techniques described above in different ways to deal with mixed data. In one of the studies, the mixed data is preprocessed through transformation before applying the learning algorithm[15]. The data is converted into homogeneous type before classification.

In another work, the researchers have worked on improving the existing classifiers to handle heterogeneous data. They have used hybrid ensemble classifiers to same data type objects and the result is taken as the weighted average of all classifiers[16].

In [17], the researcher studied classification problem for mixed data and has proposed a method called Extended Naïve Bayes (ENB) for data with numerical and categorical features. Here the native Naive Bayes algorithm is used compute the posteriori probabilities of categorical features, but for numeric features, the statistical features such as mean, standard deviation and variance are used to compute the probabilities.

In a recent study the researchers have proposed an algorithm to divide the initial data records based on pairwise similarity. Here the objective is to increase the quality of the data subsets by reducing noise and apply specialized classifier models on them. The performance of the method is evaluated on heterogeneous datasets and is found to achieve better performance[18,19].

## **1. RESEARCH METHOD (10 PT)**

### **1.1.1 Dataset**

In this work, two datasets are used. The dataset for Coronary heart diseases is obtained from UCI repositories, which has 13 predictors, 1 target attribute and 303 records and 2 classes. The other dataset used is the echocardiogram reports collected and curated from M.S.Ramaiah Narayana Heart Center, Bangalore, India[.It has 345 records, 34 predictors attributes, 1 target attribute and 4 classes. The data in both the repositories is heterogeneous in nature with a mix of nominal, categorical, ordinal and numeric types[20,21].

### **1.1.2 Method**

To obtain an efficient predictive model, a new approach based on k-Nearest Neighbour learning algorithm is proposed. In this approach, the HET-DATA-kNN employs a new distance measure to compute distance between the objects in which the distance is computed taking into consideration, the type of data and its respective metric for the purpose of classification as shown in Eq 2.1. In the previous works, the models developed using kNN have used homogeneous distance measures though the data was mixed type. In this work, the distance between objects measure used in k-NN is computed by combining distance measure obtained separately for categorical attributes, continuous attributes, ordinal attributes based on relevant measures. This

combined distance measure is then used to compute distances between objects and it is found to improve the accuracy of the model

The proposed approach considers the type of data and its respective metric for the purpose of classification. The dissimilarity between two objects  $i$  and  $j$  of mixed data types for  $m$  attributes is computed as

$$d(i, j) = \sum_{f=1}^m d_{ij}^{(f)} \quad (2.1)$$

Where,  $f$  is the feature which ranges from 1 to  $p$ , and

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}} \quad \text{if } f \text{ is numeric}$$

$$d_{ij}^{(f)} = 0 \quad \text{if } x_{if} = x_{jf}, 1 \text{ otherwise, if } f \text{ is categorical or binary}$$

compute the ranks  $r_{if}$  and  $z_{if} = \frac{r_{if}-1}{M_f-1}$  and treat  $z_{if}$  as numeric if  $f$  is ordinal.

### 1.1.3 Algorithm: HET-DATA-kNN

Computes the  $k$  nearest neighbours for the query objects and returns their class label using a novel heterogeneous distance measure.

**Input:** Dataset  $X$  with predictor attributes  $F = \langle f_1 \dots f_m \rangle$ , and target attribute  $Y$  and  $k$  classes, Query objects  $x_q$

**Output:** Target class of query objects.

**Method:**

Given a new query object  $x_q$  to be classified

- Find  $k$  neighbours of  $x_q$  as
  - For all categorical attributes  $p \in F$  in dataset  
Compute distance between query object  $i$  and object  $j$  from  $X$  as

$$d_1(i, j) = \frac{p - m}{p}$$

where  $m$  is the number of matches.

- For all numeric attributes  $r \in F$  in dataset  
Compute distance between query object  $i$  and object  $j$  as

$$d_2(i, j) = \sqrt{\sum (x_{ir} - x_{jr})^2}$$

where  $x_{ir}$  and  $x_{jr}$  are the values of the objects.

- Compute new distance

$$d(i, j) = d_1(i, j) + d_2(i, j)$$

- Sort the distance in increasing order.
- Choose top  $k$  data objects as the neighbours of  $x_q$
- Let  $x_1 \dots x_k$  denote the  $k$  data objects from training data objects that are nearest to  $x_q$
- Determine the class label of  $x_q$  by taking majority vote of the class labels of  $x_1, \dots, x_k$

## 2. RESULTS AND DISCUSSION

The algorithm HET-DATA-kNN is based on the novel distance measures as described in Section 2.1.2. The challenge here is to derive the parameter  $k$  for which the classification is accurate and error is minimum. The graph in Figure 1 shows the error at different values of  $k$ . It is found that at  $k=12$  the mean error is minimum which is 12% and hence the value of  $k=12$  is selected as the number of neighbours.

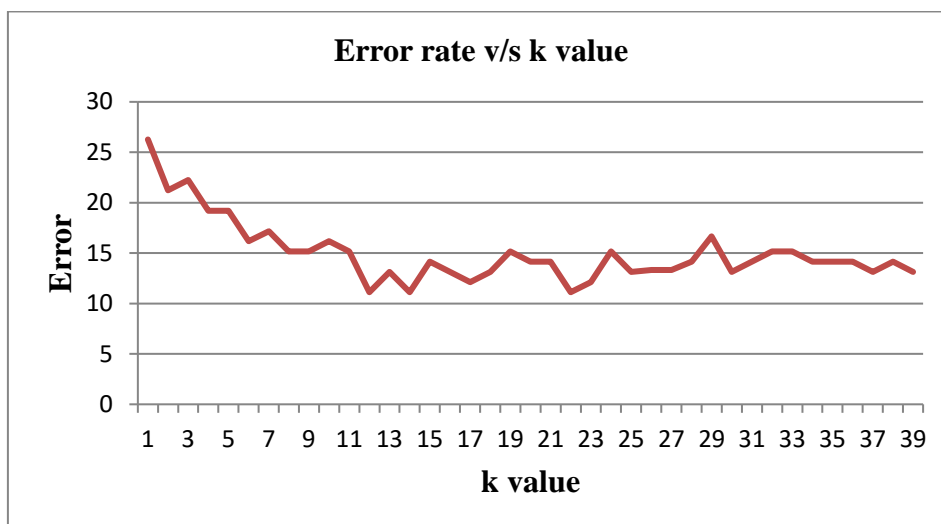


Figure 1. Error at different values of k using HET-DATA-KNN

This section discusses the performance of the proposed technique based on k-NN with the novel distance measure meant to handle mixed data types on both the datasets.

### 2.1 Comparison of Baseline kNN and proposed HET-DATA-kNN on heart.csv

Table 1 . Baseline k-NN v/s HET-DATA-kNN

Prediction Model	Accuracy %	Error rate %	Sensitivity %	Specificity %
Baseline k-NN	81.0	19.0	78.18	84.44
HET-DATA-kNN (k=12)	88.0	12.0	85.7	89.6

It is found that HET-DATA-kNN has better performance for the evaluation measures Sensitivity, Specificity, Accuracy and Error rate as compared to Baseline kNN. As shown in Table 1 and Figure 2, the proposed method has shown accuracy of 88% and error rate of 12% as compared to 81% and 19% in Baseline kNN model. Also the sensitivity and specificity of the proposed method is better than the baseline kNN.

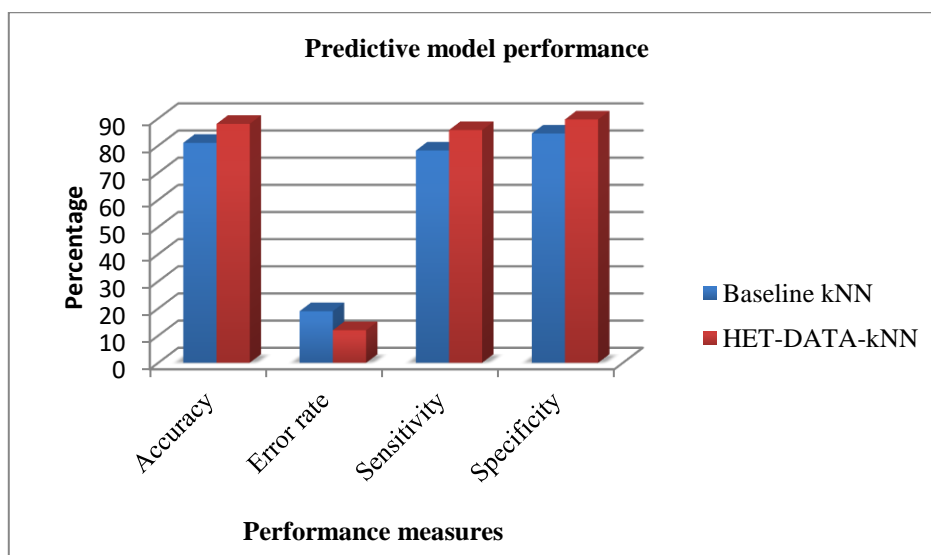


Figure 2. Performance of HET-DATA-kNN

### 2.2 Comparison of Baseline kNN and proposed HET-DATA-kNN on Echocardiogram dataset

It is found that HET-DATA-kNN has better performance for the evaluation measures Accuracy and Error rate as compared to Baseline kNN on Echocardiogram dataset. As shown in Figure 3, there is improvement in

accuracy and reduction in error rate in HET-DATA-kNN model as compared to Baseline-k-NN on echocardiogram dataset.

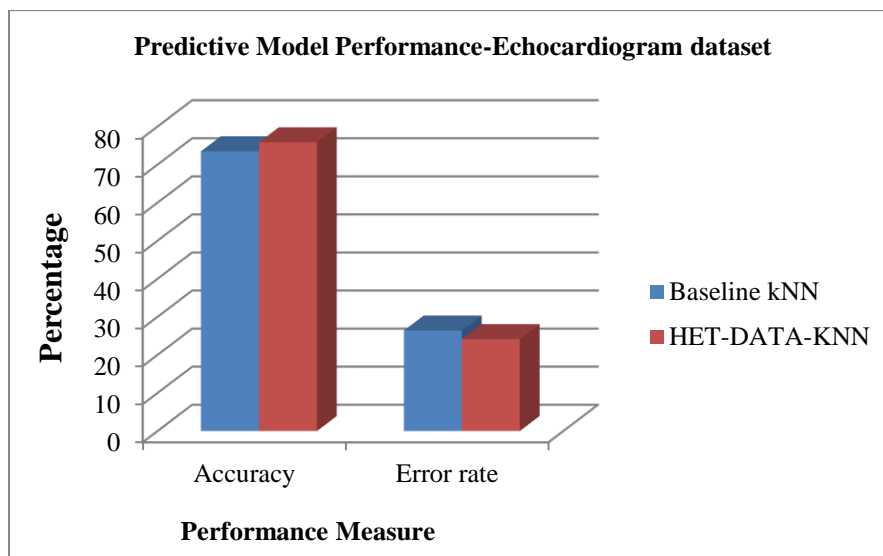


Figure 3. Performance of HET-DATA-kNN on Echocardiogram data

### 3. CONCLUSION

In this work we have developed predictive models for heart disease diagnosis using CVD dataset from Cleveland database of UCI repositories and Echocardiogram dataset. The dataset is analyzed and predictive models are developed accordingly using the learning algorithms namely Basic k Nearest neighbour and proposed HET-DATA-kNN. The accuracy of the HET-DATA-kNN model is found to be 88% when compared with the baseline kNN model which has accuracy of 81%. The comparison shows that the proposed method has improved accuracy as compared to Basic k NN.



The models developed are also applied on the echocardiogram dataset for prediction. The presented results show that the HET-DATA-kNN model is better than Baseline kNN model for predicting the class attribute from the echocardiogram dataset

### REFERENCES

- [1] Cios, K. J., & Moore, G. W., "Uniqueness of medical data mining". *Artificial intelligence in medicine*, vol. 26, no.1-2, pp.1-24, 2002
- [2] Yang, Q., & Wu, X., "10 challenging problems in data mining research", *International Journal of Information Technology & Decision Making*, vol. 5 no. 4 pp. 597-604, 2006.
- [3] Koh, H. C. and Tan, G. "Data mining applications in healthcare". *Journal of healthcare information management*, vol. 19 no. 2 pp 65, 2011.
- [4] Canlas, R. D., "Data mining in healthcare: Current applications and issues", *School of Information Systems & Management, Carnegie Mellon University, Australia*, 2009.
- [5] Heart Disease – General Info and Peer reviewed studies : [Online] Available <http://www.aristoloft.com>
- [6] Prabhakaran, D., Jeemon, P., & Roy, A. (2016). Cardiovascular diseases in India: current epidemiology and future directions. *Circulation*, 133(16), 1605-1620.
- [7] Types of heart disease- <https://www.heartandstroke.ca/heart/what-is-heart-disease/types-of-heart-disease>
- [8] Cardiovascular diseases - [https://www.who.int/cardiovascular\\_diseases/en/cvd\\_atlas\\_01\\_types.pdf](https://www.who.int/cardiovascular_diseases/en/cvd_atlas_01_types.pdf).
- [9] Causes and symptoms of heart diseases - <https://www.mayoclinic.org/symptoms-causes>.
- [10] Milovic, B., & Milovic, M., " Prediction and decision making in health care using data mining", *Kuwait Chapter of the Arabian Journal of Business and Management Review*, vol 1, no.12, pp.126, 2012.
- [11] Bellazzi, R., and Zupan, B., "Predictive data mining in clinical medicine: current issues and guidelines", *International journal of medical informatics*, vol. 77, no. 2, pp. 81-97, 2008.
- [12] Cataloluk, H., & Kesler, M., "A diagnostic software tool for skin diseases with basic and weighted K-NN", In *2012 international symposium on innovations in intelligent systems and applications* pp. 1-4. IEEE, 2012.
- [13] Han, J., Pei, J., & Kamber, M. " *Data mining: concepts and techniques* ". Elsevier, 2012.
- [14] Tan, P. N., Steinbach, M., & Kumar, V. " *Introduction to data mining*." Pearson Education India, 2016.
- [15] Chandrasekar, P., Qian, K., Shahriar, H., & Bhattacharya, P. " Improving the prediction accuracy of decision tree mining with data preprocessing". In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* , Vol.. 2, pp. 481-484, IEEE, 2017.
- [16] Verma, A., & Mehta, S., "A comparative study of ensemble learning methods for classification in bioinformatics", In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, pp. 155-158, IEEE,2017
- [17] Hsu, C. C., Huang, Y. P., & Chang, K. W. (2008). Extended Naive Bayes classifier for mixed data. *Expert Systems with Applications*, vol. 35, no. 3, pp.1080-1083.

- [18] Ali, N., Neagu, D., & Trundle, P, "Classification of heterogeneous data based on data type impact on similarity", In *UK Workshop on Computational Intelligence*, (pp. 252-263, Springer, Cham, 2018.
- [19] Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F, "The distance function effect on k-nearest neighbour classification for medical datasets". *SpringerPlus*, vol. 5, no. 1, pp. 1-9, 2016.
- [20] UCI Machine Learning Repository. [Online] Available: <http://archive.ics.uci.edu/ml/datasets.html>.
- [21] Sujata Joshi, Mydhili K. Nair, "Detection of Myocardial Ischemic Events from Echocardiogram using Linear Discriminant Analysis and Multilayer Perceptron", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Volume 9, Issue 2, 12/2019, 2875-2880.
- [22] Medical tests for heart disease- <https://www.heartfoundation.org.au/your-heart/living-with-heart-disease/medical-tests>.
- [23] Diagnostic test for heart disease- Echocardiogram - <https://my.clevelandclinic.org/health/diagnostics/16947-echocardiogram>.
- [24] Echocardiogram - <https://contenidos.bupasalud.com/en/health-and-wellness/bupa-life/echocardiogram>

## BIOGRAPHIES OF AUTHORS

	<p>Dr. Sujata Joshi is currently working as Associate Professor in the department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India. She has completed PhD from Visvesvaraya Technological University Bangalore. She has authored and co-authored several journal and conference papers and book chapters. She has guided many students for their project work and student research publications. Her research interests include prediction modelling, machine learning and artificial intelligence particularly in the medical domain. She has been a member of the organizing committee for many conferences and also in the review committee of conferences and journals.</p>
	<p>Dr. Mydhili K Nair is currently working as a Professor &amp; Head, School of Computer Science, RV University, Bangalore. Earlier she has worked for 18 years in M.S.Ramaiah Institute of Technology, Bangalore. She has a mixed bag of both academic as well as Industrial experience both spanning close to a decade each. In the IT Industry she has adorned various roles ranging from Technical Lead to Project Manager in different companies such as Wipro, IBM, Integra Microsystems, and Comviva &amp; Hexaware. In Academia she is working in MSRIT since 2004. She has done her PhD from Anna University &amp; her research interest includes Distributed Computing and Data Science. She has won the prestigious IBM Faculty Award in the year 2011 for her collaborative research association with IBM. She &amp; her students have won many project competitions &amp; best paper awards at State &amp; National level. She is also in the editorial panel of referred journals &amp; has been a member of the organizing committee as well as session chair for many symposiums &amp; conferences.</p>