# Heart Disease Prediction-based on Conventional KNN and Tuned-Hyper Parameters of KNN: An Insight

Sanjeeva Polepaka, R. P. Ram Kumar,
Department of AI & ML Engineering,
Gokaraju Rangaraju Institute of Engineering and Technology (GRIET), Hyderabad, India

Srujan Katukam, Sai Vinay Potluri, Sunell Dutt Abburi, Mohith Peddineni, Nagaraj Islavath, Muralidhar Reddy Anumandla,
B. Tech., Department of AI & MLE, Gokaraju Rangaraju Institute of Engineering and Technology (GRIET), Hyderabad, India

2023 International Conference on Computer Communication and Informatics (ICCCI) | 979-8-3503-4821-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCI56745.2023.10128508

*Abstract*— **The proposed research work aims to predict the heart disease-based on conventional KNN and tuned-hyper parameters of KNN for enhanced performance. Initially, the conventional KNN Model is adopted to determine and predict the heart diseases with respect to the dataset collected from Kaggle Domain. Later, the hyper-parameters are tuned for determining the prediction performance on the same dataset. The various parameters identified for tuning the performance includes, the parameter grid, score and cross validation. Experimental observations depict that the hyper-tuned KNNs algorithm has enhanced performance than the conventional KNN model in terms of prediction accuracy.**

***Keywords-KNN, Supervised Machine Learning, Hyper-parameters, Precision, Recall, F1 – Score, Accuracy***

## I. INTRODUCTION

The K-Nearest Neighbors algorithm also termed as k-NN or KNN algorithm is a supervised machine learning algorithm is used to solve classification and regression problem statements. A supervised machine learning (SML) algorithm completely depends on the labelled input data to know the function that produces appropriate output for the given new unlabeled data. For instance, when an infant is trained by parents to identify a specific fruit, the parents show pictures of different fruits to the infant and train the infant to identify the specific fruit. The term "non-parametric" refers to the property of no prior assumptions about the data being processed. Since there is no training phase in this SML, KNN is also termed as "Lazy Learner" algorithm. In classification problem, the output data hold discrete data with which numerical operations are impossible. The standard method of data representation in classification is -1, 0, and 1. In regression problem, the data is real that holds decimal points. The data consists of both dependent and independent variable. Often in the data, every row – data point, example, observation while every column (excluding label and dependent variable) – predictor, dimension, feature or independent variable[15].

KNN algorithm works and relies completely on the fact that similar things are close to each other. Quantifying the closeness determines the output of KNN. To enunciate, data points are close to each other. The closeness, proximity or similarity of the query point with other data points is calculated by determining the distance between the query point and other data points (individually). Some of the distance measurement techniques are: Euclidean distance, Manhattan distance, Mincowski distance, Hamming distance. The number of neighbors/distances to be checked can be determined manually. It is denoted by 'k'. In general, the odd value of k yields good results (act as a tiebreaker). Low value of 'k' might cause low bias and high variance; on the contrary high value of 'k' brings low variance [13] and [14].

The KNN algorithm is one of the most preferred algorithms as it is simple and easy to understand and implement. Moreover, the algorithm doesn't involve any model building and value of k can be determined manually depending upon the input data. KNN can be used for classification, regression and search problems; hence a versatile algorithm that match the need accordingly. When the volume of data increases, KNN significantly slows down thereby reducing rapid functionality. This major defect is overthrown when suitable and sufficient computing resources are utilized.

Figure 1 depicts the procedure regarding the traditional KNN Model. Initially, the (input) data is loaded and initialized to a variable. Later the iteration process is initiated from the first data point to the nth data point (the last point); during the iteration, the distance between each point is determined. Possibly, the Euclidean distance (ED) is adopted to determine the distance between the points. The determined distances are stored for further processing. Later, these values are sorted in a descending order. By extricating the top "k" value(s), the concerned neighbors and their classes are determined.

## II. EXISTING APPROACHES

The following section depicts few of the existing approaches that adopts KNN model for prediction process.

Alkhatib, Hassan Najadat, Ismail Hmeidi, Mohammed K. Ali Shatnawi [1] designed a stock price prediction model for Jordanian based 6 listed companies in Jordanian stock exchange with sample dataset of 200 in each company. The model utilized data mining technique – KNN algorithm with the 'k' value '5'. The stock price predicted by the model assisted the users of several field; in addition, KNN algorithm

**979-8-3503-4821-7/23/$31.00 ©2023 IEEE**

was rapid and dynamic with minimum error value. Thus, the real stock values were almost parallel to the predicted values.

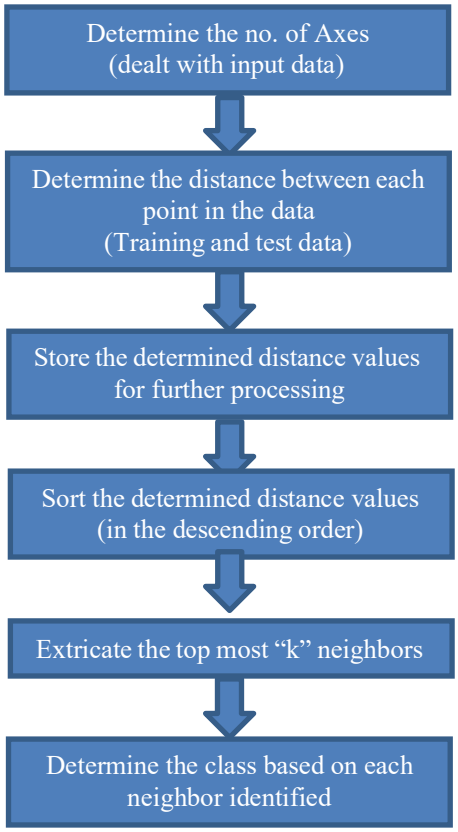


Figure 1. Block diagram of traditional KNN Model

Bijalwan, Kumar, Kumari and Pascual [2] applied KNN algorithm on Reuters – 21578 datasets for Document Classification; documents were categorized under five labels – people, exchange, places, topics, and organization respectively. The phases in document mining involve the following steps, namely, (i) pre-processing and (ii) classification by KNN, Term Graph algorithm and Naïve Bayes algorithm. Compared to other techniques in the study, KNN outperformed in terms of accuracy with respect to afore mentioned labels 99.7, 98, 99.27, 99.29 and 98.5 respectively.

Biau, Cadre and Rouviere [3] analyzed the statistics provided by KNN in Collaborative Recommendation. The recommendation model suggests the websites, places, or things that a user would be interested based on the history and comparative analysis of similar interests among several users.

Dramé, Mougin and Diallo [4] classified large scale of biomedical documents facilitating automatic processing using partial information (keywords) by means of KNN and ESA algorithms. The datasets used were BioASQ organizers and MeSH Thesaurus consisting of 133770 and 2,268,724 documents. The performances of KNN and ESA in both datasets were measured in terms of example-based Precision, Recall and F- Score (EBP, EBR, EBF). The results obtained from KNN with respect to F-Score was comparatively good than other classification methods.

Guo, Wang, Bell, Bi and Greer [5] categorized documents based on content as a part of automatic document classification. The prototypes were experimented on Reuters-21578 (ModApte version) and 20-newsgroup document collection utilizing two similarity-based learning algorithms such as k-nearest neighbor (KNN) classifier and Rocchio classifier. The result proved KNN to be superior to Rocchio algorithm.

Justiawan, Sigit, and Arief [6] used KNN Classifier to detect accurate tooth color during tooth reconstruction. Sixteen type of teeth images from UNAIR, RSGM and SURABAYA were used for the study. The accuracy of KNN Classifier improved from 97.19% to 97.8% upon using Principal Component Analysis technique for feature selection which reduced features from 12 to 8. Zhang [7] discussed the challenges in using KNN algorithm such as Computation of 'k' value, selecting and searching the nearest neighbor and finally the classification rules. Zhang suggested few resolution techniques regarding the aforesaid challenges by examining the recent improvements. The author used 15 UCI datasets for the purpose of evaluating the resolution.

Khan and Ahmed [8] designed a reasonable in-vehicle video camera that present drivers real-time roadway weather especially snow. The model used texture-based image features such as GLCM and LBP; features were further classified based on clear, light and heavy snow by applying three algorithms namely SVM, KNN, RF. Upon comparing the accuracies of SHRP2 NDS dataset on the model, LBP with SVM exhibited 95.9% while LBP with KNN showed 93%. Avand, Janizadeh, Naghibi, Pourghasemi, Bozchaloei and Blaschke [9] carried out Gully Erosion Susceptibility in Iran using data mining techniques – RF and KNN respectively. The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were determined. The validation test of AUC showed 87.4% in Random Forest. Moreover, rainfall, altitude and distance from the river were identified as pivotal factors for Gully Erosion.

## III. PROPSED METHOD

The proposed method aims to predict the Heart Diseases (HD) based on KNN algorithm followed by effectively tuning the hyperparameters for improved prediction accuracy. Initially the HD dataset is collected from Kaggle domain [10] which has 14 fields including the decision field. The various attributes in the HD dataset includes, (i) age represented in years, (ii) sex – male/female represented as 1/0, respectively, (iii) cp denotes the chest pain types- typical angima, atpical angima, non-anginal pain and asymptomatic represented by 1, 2, 3 and 4, respectively, (iv) trestbps denotes the normal blood pressure during the resting period, (v) chol symbolizes the serum cholesterol, (vi) fbs denotes the blood sugar during the fasting period, (vii) rest ECG represents the Electrocardiographic results during the resting period (varies from 0, 1, and 2), (viii) thalach symbolizes the maximum value of the determined heart rate, (ix) exang denotes the angima induced due to exercise, (x) oldpeak illustrates the depression caused by exercise rather than rest period, (xi) slope denotes the slope generated during the peak exercise period, (xii) ca represents the major vessels count, (xiii) thal denotes the thalassemia values ranging from 0

to 3 corresponding to NULL, fixed defect, normal flow of the blood and reversible defect respectively, and finally (xiv) the target indicating the presence of HD as '1' and non-presence as '0', respectively. The dimensions of the dataset are 303 X 14, comprising 14 fields and 303 rows; the rows denote the number of patient records. With the help of Exploratory Data Analysis (EDA), the various parameters in the HD Dataset are analyzed which includes (i) target (ii) Age (iii) Sex (iv) trest bps (v) chol (vi) thalach (vii) slope (viii) exang (ix) rest ECG followed by (x) thal and (xi) old peak values for better understanding of correlation of features (that is the parameters) that causes the HD. After analyzing the dataset features and with prior knowledge on dependent attributes, KNN classifier is applied to predict the HD. The various training phase ranges from 70%, 80% and 90% such that the testing phase incurs 30%, 20% and 10% respectively. The value of 'K' is chosen randomly from 1 to 20. Later, the tuning of KNN parameters is done to determine whether there is any input on the predicted HD using the conventional method [11] and [12]. The optimal values for the KNN algorithm are determined from the pre- determined set of values regarding the hyper parameters including (a) 'p' represents parameter grid object to hold the parameters (b) 's' represents the score of the evaluation metric (c) 'cv' represents the cross validation for the concerned selected hyper parameters (d) 'v' represents the verbose regarding the detailed view of concerned hyper parameters (e) 'nj' represents the number of processes that can be made to execute in parallel respectively.

## IV. Experimental Results and Discussions

The proposed method of HD prediction-based on KNN is evaluated on the dataset collected from Kaggle domain. Initially the predictions are determined using the traditional KNN approach followed by hyper-parameter tuning method. The various parameters considered to determine the HD prediction includes precision (p), recall (r), f1-score (f1) and accuracy (A). Table 1 depicts the determined predictions regarding the performance measures. Further, the determined performances regarding the hyper parameters are represented in Table 1. The various values and their ranges are as follows: (a) parameter grid 'p' ranging between 1 and 2 (b) neighborhood values for 'K' ranges between 1 to 35 and (c) controlling the leaf size parameter between 1 and 40 respectively.

TABLE I. Performance Measures

| Description | p (%) | r (%) | F1 (%) | A (%) |
|---|---|---|---|---|
| Traditional KNN model | 70 | 70 | 76 | 71 |
| Tuned Hyper-parameters | 79 | 80 | 84 | 80 |

## V. Conclusion

The proposed HD prediction was built using KNN Model and evaluated on the dataset collected from Kaggle domain. There were two phases of implementation, namely, (i) traditional KNN model, and (ii) tuning of hyper parameters for improved performance. The various parameters used to evaluate the proposed HD prediction model based on KNN model were p, r, f1 and A respectively. Observing Table 1, the conclusion is that the performance of KNN based HD prediction model has significantly improved performance while tuning their hyper-parameters than the traditional algorithm. As a future enhancement, KNN's traditional and hyper-tuned approaches are to be evaluated on various dataset for detailed exploration.

## References


[1] K. Alkhatib, H. Najadat, I. Hmeidi, & M. K. A. Shatnawi, "Stock Price Prediction Using K-Nearest Neighbor Algorithm. International Journal of Business", Humanities and Technology, vol. 3, no. 3, pp. 32-44, 2013.

[2] V. Bijalwan, V. Kumar, P. Kumari, & J. Pascual, "KNN based machine learning approach for text and document mining", International Journal of Database Theory and Application, vol. 7, no. 1, pp. 61-70, 2014, https://doi.org/10.14257/ijdta.2014.7.1.06

[3] G. Biau, B. Cadre, & L. Rouviére, "Statistical analysis of k-nearest neighbor collaborative recommendation", Annals of Statistics, vol. 38, no. 3, pp. 1568-1592, 2010, https://doi.org/10.1214/09-AOS759

[4] K. Dramé, F. Mougin, & G. Diallo, "Large scale biomedical texts classification: A kNN and an ESA-based approaches", Journal of Biomedical Semantics, vol. 7, no. 1, pp. 1-34, 2016, https://doi.org /10.1186/s13326-016-0073-1

[5] G. Guo, H. Wang, D. Bell, Y. Bi, & K. Greer, "An kNN model-based approach and its application in text categorization", Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 2945, pp. 559-570, 2004, https://doi.org/10.1007/978-3-540-24630-5_69

[6] Justiawan, R. Sigit, & Z. Arief, "Tooth Color Detection Using PCA and KNN Classifier Algorithm Based on Color Moment", EMITTER International Journal of Engineering Technology, vol. 5, no. 1, pp. 139-153, 2017, https://doi.org/10.24003/emitter.v5i1.171

[7] S. Zhang, "Challenges in KNN Classification", IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, pp. 4663-4675, 2021. https://doi.org/10.1109/TKDE.2021.3049250

[8] M. N. Khan, & M. M. Ahmed, "Snow Detection using In-Vehicle Video Camera with Texture-Based Image Features Utilizing K-Nearest Neighbor", Support Vector Machine, and Random Forest. Transportation Research Record, vol. 2673, no. 8, pp. 221-232, 2019, https://doi.org/10.1177/0361198119842105

[9] M. Avand, S. Janizadeh, S. A. Naghibi, H. R. Pourghasemi, S. K. Bozchaloei, & T. Blaschke, "A comparative assessment of Random Forest and k-Nearest Neighbor classifiers for gully erosion susceptibility mapping", Water (Switzerland), vol. 11, no. 10, 2019, https://doi.org /10.3390/w11102076

[10] "UCI Heart Disease Data", https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

[11] "Tune Your Hyperparameters," OpenClassrooms. [Online]. Available: https://openclassrooms.com/en/courses/6401081-improve-the-performance-of-a-machine-learning-model/6559796-tune-your-hyperparameters

[12] "Grid Search For Hyperparameter Tuning," Medium, 21-Mar-2020, https://mathanrajsharma.medium.com, [Online]. Available: https:// towardsdatascience.com/grid-search-for-hyperparameter-tuning-9f63945e8fec

[13] "K-Nearest Neighbors Algorithm." https://www.ibm.com/in-en/topics /knn

[14] "Machine Learning Basics with The K-Nearest Neighbors Algorithm," Medium, 14-Jul-2019, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761, https://onelharrison.medium.com

[15] Paul Stone Brown Macheso, Angel G Meela, "IoT Based Patient HealthMonitoring using ESP8266 and Arduino", International Journal Of Computer Communication And Informatics, volume-3, issue- 2, 75-83, 2021