Performance Comparison of K-NN, Minimum Distance and SVM Classifiers Used to Predict Hypertension Based on Manuscript

Seema V. Kedar
JSPM's RSCOE, Dept. of Electronics Engg., S.P.P.U., Pune, India
Email: seemaahkeddar@gmail.com

D. S. Bormane and S. C. Patil
AISSMSCOE, JSPM's RSCOE, S.P.P.U., Pune, India
Email: bdattatraya@yahoo.com, shailaja.patil11@gmail.com

***Abstract*—Cardiovascular disease is the main cause of fatality for women and men in developing and developed countries. Hypertension also known as high blood pressure is one of the major type of cardiovascular disease which causes more deaths than other cardiovascular diseases such as heart failure, coronary artery disease, arrhythmia, congenital heart disease etc. It is the time to take necessary steps to control death rate due to hypertension. In this paper, an approach to predict the hypertension based on handwritten manuscript using K-NN, Minimum distance and SVM classifiers has been presented. The proposed approach extracts geometric features such as number of right and left slants, number of horizontal and vertical lines, total length of horizontal and vertical lines, total length of left and right diagonal lines, and regional features such as Euler, major axis length and extend from a manuscript. Based on this information it predicts the existence of high blood pressure. The dataset contains total 140 manuscripts. The dataset is prepared using manuscripts of high blood pressure and control group people. The proposed system provides accuracy of 76.47% using K-NN classifier, 70.59% using minimum distance classifier, and 85.29% using SVM classifier.**

***Index Terms*—high blood pressure, writing features, handwriting analysis, manuscript**

I. Introduction

The mortality rate due to cardiovascular diseases is increasing at an alarming rate across the globe. This count is greater than deaths due to AIDS and all cancers combined [1]-[3]. Cardiovascular disease also known as heart disease has become a bigger health concern in the developed as well developing countries. Cardiovascular disease is number one cause of mortality globally [4]. Every year around 30% of all global deaths occur due to cardiovascular disease. Around 80% deaths take place due to heart disease in economically backward and developing countries. If appropriate actions are not taken in time, the deaths due to cardiovascular disease will increase to 23.6 million by 2030 [5]. Cardiovascular diseases kill over three million people every year in India [6]-[9]. Globally 8.5 million deaths occur in women due to cardiovascular disease which is one-third of all deaths in women. High Blood Pressure (HBP) which is type of heart disease is leading cause of deaths as it leads to heart, kidney and brain damage. In United States, everyday around 1000 people die because of HBP and the death rate of HBP has increased 23% since 2000 [10]. About 33% of adults in Scotland and England are suffering from HBP [11]. Around half of the Singaporeans aged 60 to 69 suffer from HBP [12]. The governments across the world are working on reducing the mortality count due to HBP.

We can detect any type of heart disease after physical symptoms are apparent using different tests such as blood pressure test, stress test, blood sample analysis, ECG scanning, chest X-Ray, CT Heart scan and heart MRI [11], [13]. Some of the researchers have proposed a soft computing approach to identify cardiovascular disease using the parameters like resting electrographic results, cholesterol, blood pressure, sex, age, exercise induced angina, defect type, maximum heart rate, fasting blood sugar, and chest pain. The researchers have used classifiers such as neural network, fuzzy logic, SVM to perform the work [14]-[18] and they have obtained the accuracy between 90% to 100%. In paper [19], authors have presented a system for monitoring and diagnosis of patients suffering from coronary artery disease. The diagnosis module finds heart problems using algorithms such as SVM, BayesNet, and Functional Trees. The monitoring module uses wearable sensor for detecting heart rate. It sends this data to a mobile device. If there is any change in heart rates, the monitoring algorithm sends a signal to the patients' relatives and closest hospital. The proposed approach has got 88.3% accuracy using SVM classifier for diagnosis component and 100% accuracy for monitoring component. In paper [20], authors have reviewed the research carried on heart disease prediction and diagnosis using the data mining algorithms. As per literature, SVM and neural networks are good for predicting the presence of coronary heart diseases.

Manuscript received January 16, 2016; revised March 27, 2017.

doi: 10.18178/ijsps.5.1.39-43

Decision tree is good for diagnosing cardiovascular disease. The performance of data mining techniques to detect coronary arteries diseases varies between 60%-75% which is not encouraging and further improvements are required. A model proposed in paper [21] uses Bagging algorithm and J4.8 Decision Tree for finding of heart disease. The proposed approach works in different steps like data discretization, partitioning of data, selection of decision Tree type, and preparation of pruned decision trees by the application of reduced error pruning. The proposed approach used 13 attributes such as CP, Sex, Age, Chol, Thal, CA, Exang, Restecg, Fbs, Thalach, Old peak ST, Trestbps, Slope, and Diagnosis class. The system provided sensitivity of 77.9%, Specificity of 85.2%, and Accuracy of 84.1%. In paper [22], authors have proposed a solution to detect heart attack using smart phone. Using proposed approach one can detect heart condition either by using peak of blood or by using heartbeat. Authors claim that the proposed system is also useful for detecting abnormal blood, valve circulation, and heart blockage, The paper [23], presents an approach to diagnose heart disease using techniques such as Decision Tree, Naive Bayes, and Classification by Clustering based on six attributes such as Chest pain type, Number of major vessels colored, Old peak, Exercise induced angina, Resting Blood pressure, and Maximum heart rate achieved. The proposed approach has got 99.2 % accuracy using decision tree, 96.5% accuracy using Naïve Bayes, and 88.3% accuracy using clustering algorithm.

Thus, numerous techniques have been proposed by researchers for detecting heart diseases, but they require costly hardware or software. In this paper, we have presented an approach based on handwritten text also known as manuscript to predict existence of HBP at early stage. It is an instant and cost effective method.

Usually handwriting analysis is used to understand the personality, emotional state, behavior etc. of the writer [24]-[26]. Handwriting of a person shows pre-illness warnings about the disease before physical symptoms are apparent [27]. According to research study carried out by University of Plymouth, England, Graphology can be used detect health conditions [28]. When a person writes, what he/she writes is controlled by conscious mind while how he/she writes is controlled by subconscious mind. Instructions about writing style are given to hand by subconscious mind. Subconscious mind and central nervous system are related. Thus, if there are any health issues, they are reflected through handwriting. According to Graphology, people suffering from same disease have same writing features. These writing features are studied to predict the disease [29]-[32].

A naïve approach to predict healthy person (also known as control group) and person suffering from hypertension has been proposed in this paper. The proposed approach extracts geometric features such as number of right and left slants, number of horizontal and vertical lines, total length of horizontal and vertical lines, total length of left and right diagonal lines, and regional features like Euler, major axis length and extend from a manuscript. These features are used to foresee the existence of hypertension using K-NN, Minimum Distance (MD) and SVM classifiers. The performance of K-NN, MD, and SVM classifiers used to implement the proposed system has been discussed in this paper.

The remainder of the paper is organized as follows: Section II briefs the proposed system, Section III explains the proposed algorithm, Section IV discusses experimental results, and Section V concludes the paper.

II. Proposed System

Proposed research work requires the dataset of manuscripts of both high blood pressure and Control Group (CG) people. As such standard dataset is not available; we have prepared our own dataset by collecting the manuscripts of both CG people and people suffering from HBP. A plain white paper has been used to collect known manuscripts of CG and HBP people. These manuscripts are then converted into JPEG images using scanner. Fig. 1 demonstrates the working of the proposed system.

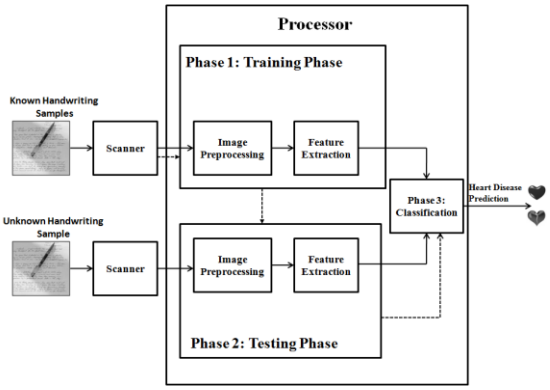

Figure 1. Working of the proposed system.

The system works in three phases namely training, testing, and classification. The training phase acquires JPEG images of known manuscripts of CG and HBP people. The testing phase acquires JPEG image of unknown writing sample which is of either person suffering from high blood pressure or control group person [33].

Initially the system operates in training phase. In this phase, it passes through preprocessing and feature extraction steps respectively. During preprocessing, it converts the input image to binary image. It removes noise from the image by applying median filter. It then resizes the image to [512, 512] size. The binary image is inverted and then after it is thinned.

The feature extraction step uses the preprocessed image as input and extracts geometric features such as number of right and left slants, number of horizontal and vertical lines, total length of horizontal and vertical lines, total length of left and right diagonal lines, and regional features such as Euler, major axis length and extend from it. The reference feature vector is prepared by combining geometric and regional features of all preprocessed images from training dataset [34].

After preparing reference feature vector, the system enters into testing phase. It acquires JPEG image of

unknown writing sample. It then again passes through preprocessing and feature extraction steps and prepares testing feature vector by combing geometric and regional writing features of test image.

Finally system enters into classification phase. The reference feature vector and testing feature vector are given as input to classifier. The classifier compares reference and testing feature vectors, and predicts whether the test image belongs to CG or HBP class. The performance of the system is tested using three different classifiers namely K-NN, MD and SVM [35].

III. Proposed Algorithm

Proposed algorithm works on a dataset which has been partitioned into training and testing dataset. A training dataset is prepared using JPEG images of manuscripts of CG and HBP people and testing dataset is created using JPEG images of an unknown manuscript. Different steps of the proposed algorithm are given below.

- Repeat the steps A1 to step A3 for all the images in training dataset so as to prepare reference feature vector.

1. Select an image from training dataset.
2. Preprocess the selected image using following steps:
   a) Convert the image to grayscale image.
   b) Convert the grayscale image into binary image.
   c) Remove noise from an image by applying median filter.
   d) Resize the image to [512,512] size.
   e) Invert the binary image and thin it.
3. Extract geometric and regional features from the preprocessed image by using following steps:
   a) Skeletonize the preprocessed image.
   b) Select the universe of discourse.
   c) Divide the preprocessed image into equal size nine parts.
   d) Compute geometric writing features like number of right and left slants, number of horizontal and vertical lines, total length of horizontal and vertical lines, total length of left and right diagonal lines, for each part of partitioned image.
   e) Add together the geometric features of each part.
   f) Compute the regional features such as Euler, major axis length and extend for the entire image.
   g) Combine geometric and regional features and store in reference vector with appropriate class label.

- Select a JEPG image for testing from testing dataset.
- Prepare testing feature vector by applying steps A2 and step A3 on testing image.
- Find appropriate class label i.e. CG or HBP for test image using either K-NN, MD or SVM classifier.
- Display the predicted class label i.e. CG or HBP.

IV. Experimental Results

The data set with 140 manuscripts of HBP and CG people is used for performing experimentation. The training dataset is created using 106 manuscripts which are categorized into two classes such as HBP, and CG (53 manuscripts of each class). The testing dataset contains 34 manuscripts (17 manuscripts of each class). Performance comparison of K-NN, MD, and SVM classifiers is discussed below.

A. Results Obtained Using K-NN Classifier

Overall accuracy of 76.47% is obtained using K-NN Classifier. It is computed using (1).

$$Accuracy = \frac{Correctly\ predicted\ manuscripts}{Total\ number\ of\ manuscripts} * 100$$

$$= \frac{26}{34} * 100 \quad (1)$$

$$= 76.47\%$$

Table I represents the confusion matrix with actual and predicted class labels for K-NN classifier.

B. Results Obtained Using Minimum Distance Classifier

Overall accuracy of 70.59% is obtained using MD Classifier. It is computed using (1).

$$Accuracy = \frac{Correctly\ predicted\ manuscripts}{Total\ number\ of\ manuscripts} * 100$$

$$= \frac{24}{34} * 100$$

$$= 70.59\%$$

Table II represents the confusion matrix with actual and predicted class labels for MD classifier.

C. Results Obtained Using SVM Classifier

Overall accuracy of 85.29% is obtained using SVM classifier. It is computed using (1):

$$Accuracy = \frac{Correctly\ predicted\ manuscripts}{Total\ number\ of\ manuscripts} * 100$$

$$= \frac{29}{34} * 100$$

$$= 85.29\%$$

The confusion matrix for SVM classifier is given in following Table III.

The graph in Fig. 2 depicts the recognition rate for HBP and CG classes. We obtained 70.59% and 82.35% accuracy for CG and HBP class respectively using K-NN

classifier. The accuracy of 64.70% and 76.48% has been obtained for CG and HBP classes respectively using MD classifier. While using SVM classifier we obtained 82.35% and 88.23% accuracy for CG and HBP class respectively.

TABLE I. CONFUSION MATRIX FOR K-NN CLASSIFIER

		Predicted Class		Marginal Sum of Actual Values
		CG	HBP	
Actual Class Values	CG	12	5	17
	HBP	14	3	17
Marginal Sum of Predicated Values		26	8	T = 34

TABLE II. CONFUSION MATRIX FOR MD CLASSIFIER

		Predicted Class		Marginal Sum of Actual Values
		CG	HBP	
Actual Class Values	CG	11	6	17
	HBP	13	4	17
Marginal Sum of Predicated Values		24	10	T = 34

TABLE III. CONFUSION MATRIX FOR SVM CLASSIFIER

		Predicted Class		Marginal Sum of Actual Values
		CG	HBP	
Actual Class Values	CG	14	3	17
	HBP	15	2	17
Marginal Sum of Predicated Values		29	5	T = 34

TABLE IV. RECOGNITION RATE OF CG AND HBP CLASSES

Sr. No.	Classifier	Accuracy for CG in %	Accuracy for HBP in %
1.	K-NN	70.59	82.33
2.	MD	64.70	76.48
3.	SVM	82.35	88.23

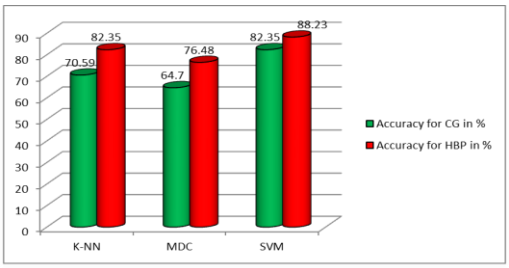

Figure 2. Recognition Rate for CG and HBP Classes.

V. CONCLUSION

An approach to predict the hypertension based on manuscript using three classifiers such as k-Nearest Neighbor, Minimum Distance and Support Vector Machine is presented here. The performance comparison of all these classifiers is discussed. A training dataset consists of 106 manuscripts of CG and HBP classes, whereas testing dataset includes 34 manuscripts. The proposed system provides an overall accuracy of 76.47% using the K-NN classifier, 70.59% using MD classifier, and 85.29% using SVM classifier. The SVM classifier predicts CG and HBP manuscripts more accurately than K-NN and MD classifiers. Predicting hypertension based on manuscript is a cost effective and easy solution. Our next objective is to improve the accuracy by extracting Graphological features such as zone, pressure, baseline, size, page margin and slant etc.

ACKNOWLEDGMENT

The authors are thankful to Savitribai Phule Pune University, Pune, for sponsoring the proposed research work.

REFERENCES

[1] *Cardiovascular Diseases in India Challenges and Way ahead*, International Heart Protection Summit, 2011.
[2] BBC News. [Online]. Available: http://news.bbc.co.uk
[3] World Heart Federation. [Online]. Available: http://www.world-heart-federation.org
[4] R. Gupta, S. Guptha, K. K. Sharma, A. Gupta, and P. Deedwania, "Regional variations in cardiovascular risk factors in India: India heart watch," *World J Cardiol*., pp. 112-120, 2012.
[5] World Health Organization0. [Online]. Available: http://www.who.int
[6] S. Chauhan and B. T. Aeri, "Prevalence of cardiovascular disease in India and it is economic impact - A review," *International Journal of Scientific and Research Publications*, vol. 3, no. 10, 2013.
[7] The Indian Express Archive. [Online]. Available: http://archive.indianexpress.com
[8] India Today. [Online]. Available: http://indiatoday.intoday.in
[9] Times of India. [Online]. Available: http://timesofindia.indiatimes.com
[10] American Heart Association. [Online]. Available: http://www.heart.org/HEARTORG
[11] British Heart Foundation. [Online]. Available: http://www.bhf.org.uk
[12] Beat Helath Exchange. [Online]. Available: https://www.healthxchange.sg/high-blood-pressure
[13] C. S. Dangare and S. S. Apte, "A data mining approach for prediction of heart disease using neural networks," *International Journal of Computer Engineering & Technology*, vol. 3, no. 3, pp. 30-40, 2012.
[14] K. S. Kavitha, K. V. Ramakrishnan, and M. K. Singh, "Modeling and design of evolutionary neural network for heart disease detection," *IJCSI International Journal of Computer Science Issues*, vol. 7, no. 5, pp. 272-283, 2010.
[15] M. N. Elbedwehy, H. M. Zawbaa, N. Ghali, and A. E. Hassanien, "Detection of heart disease using binary particle swarm optimization," *Computer Science & Information Systems*, vol. 11, pp. 177-182, 2012.
[16] S. Kumar and G. Kaur, "Detection of heart diseases using fuzzy logic," *International Journal of Engineering Trends and Technology*, vol. 4, no. 6, 2013.
[17] C. S. Dangare and S. S. Apte, "A data mining approach for prediction of heart disease using neural networks," *International Journal of Computer Engineering & Technology*, vol. 3, no. 3, pp. 30-40, 2012.
[18] K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," *International Journal of Computer Applications*, vol. 19, no. 6, 2011.
[19] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease,"

International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 143-156, 2015.
[20] S. M. Alzahani, A. Althopity, A. Alghamdi, B. Alshehri, and S. Aljuaid, "An overview of data mining techniques applied for heart disease diagnosis and prediction," *Lecture Notes on Information Theory*, vol. 2, no. 4, December 2014.
[21] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in *Proc. 9th Australasian Data Mining Conference*, Ballarat, Australia, 2011.
[22] M. Ashrafuzzaman, M. Huq, C. Chakraborty, M. R. M. Khan, T. Tabassum, and R. Hasan., "Heart attack detection using smart phone," *International Journal of Technology Enhancements and Emerging Engineering Research*, vol. 1, no. 3, 2013.
[23] S. B. Patel, P. K. Yadav, and D. P.Shukla, "Predict the diagnosis of heart disease patients using classification mining techniques," *IOSR Journal of Agriculture and Veterinary Science*, vol. 4, no. 2, Jul.-Aug. 2013.
[24] S. Kedar, V. Nair, and S. Kulkarni, "Personality identification through handwriting analysis: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 1, 2015.
[25] S. Kedar and D. S. Bormane, "Automatic emotion recognition through handwriting analysis: A review," in *Proc. IEEE International Conference on Computing Communication Control and Automation*, 2015.
[26] S. Kedar and D. S. Bormane, "Automatic personality assessment: A systematic review," in *Proc. IEEE International Conference on Information Processing*, Dec. 2015.
[27] GROW. [Online]. Available: http://www.handwritingexplore.com/grow-paid-services.html
[28] Zoom Health. [Online]. Available: http://www.zoomhealth.net/WhatYourHandwritingSaysAboutYourHealth.html
[29] Christina Strang, "Handwriting in the early detection of disease – A research study," *The Graphologist: The Journal of the British Institute of Graphologists*, vol. 31, no. 1, 2013.
[30] A. Rahiman, D. Varghese, and K. G. Manoj, "HABIT-Handwriting analysis based individualistic traits prediction," *International Journal of Image Processing*, vol. 7, no. 2, 2013.
[31] D. J. Antony and O. F. M. Cap, *A Textbook of Handwriting Analysis, Personality Profile through Handwriting Analysis*, Anugraha Publications, India, 2008.
[32] Natural News. [Online]. Available: http://www.naturalnews.com
[33] S. Kedar and D. S. Bormane, "Heart disease prediction using k-nearest neighbor classifier based on handwritten text," *Advances in Intelligent Systems and Computing*, vol. 410, December 2015.
[34] D. Dileep, "A feature extraction technique based on character geometry for character recognition," Department of Electronics and Communication Engineering, Amrita School of Engineering, Kollam, India, 2012.
[35] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second ed., Elsevier Inc., 2006.

Mrs. Seema Kedar is pursuing Ph.D. from S.P.P.U, Pune, India. She has completed M.E (Computer) in 2006 from Bharati Vidyapeeth, Pune, India.
Image Processing, Database Management Systems, Object Oriented Modeling and Design are her areas of interest. She author more than 25 research publications. She is a member of professional bodies such as ISTE, and CSI.

Dr. D. S. Bormane is working as Professor at AISSMSCOE, affiliated to Savitribai Phule Pune University, Pune, India. He received PhD in Engineering (EC & CSE) from S.R.T.M.U., Nanded in 2003. He has completed M.E (EC) from Shivaji University, Kolhapur in year 1996.
He is having more than 30 years' experience in teaching field. He is working as Chairman BOS in Electronics, Member of Research & Recognition Committee at S.P.P.U, Pune, India. He is recognized Ph.D. guide of seven universities in India. He has guided number of research scholars. Digital signal processing, Image and speech processing, VLSI design, Digital systems are his areas of interest. He author more than 125 research publications. He has worked as session chair/ reviewer/ chief guest for various national/ international conferences.

Dr. Shailaja Patil is Professor at the Department of Electronics and Telecommunication of RSCOE, affiliated to Savitribai Phule Pune University, Pune. She received Ph.D. in Computer Engineering from SVNIT, SURAT.
She has total 25 years of research and teaching experience in the field of Wireless Technologies, Internet of Things and Pervasive computing. She has authored many research articles in International conferences and Journals, Book chapters, Patent, and Book. She has worked as session chair/ reviewer/ for various national/ international conferences. She is a member of professional bodies such as ISTE, IEEE and ACM.