Intelligent Syncope Disease Prediction Framework using DM-Ensemble Techniques

Ammar Asjad Raja
Fakultät für Informatik,
Hoschschule Heilbronn,
Heilbronn, Germany
araja@stud.hs-heilbronn.de

Madiha Guftar
Department of Computer Engineering
College of Electrical and Mechanical Engineering,
National University of Sciences & Technology,
Islamabad, Pakistan
madiha12@ce.ceme.nust.edu.pk

Irfan-ul-Haq
Department of Software Engineering,
Bahria University Islamabad Campus
Islamabad, Pakistan
Irfan.ul.haq@univie.ac.at

Tamim Ahmed Khan
Department of Software Engineering,
Bahria University Islamabad Campus
Islamabad, Pakistan
tamim@bui.edu.pk

Dominik Greibl
Fakultät für Informatik,
Hoschschule Heilbronn,
Heilbronn, Germany
dominik@greibl.de

***Abstract*—Data mining can be used in various fields' i.e. mobile computing, web mining, expert predictions, crime analysis, engineering, management and medicine. In medical field, data mining techniques can be used by the researchers for the diagnosis and prediction of various diseases. A framework is proposed to predict Syncope Disease using Ensemble technique that contains Naïve Bayes, Gini Index and Support Vector Machine classifiers. Patient's data set for this research work is obtained from Armed Forces Institute of Cardiology (AFIC & NIHD) in Pakistan. Thirty one attributes have been used to predict Syncope using Ensemble techniques but each technique uses its own way to predict Syncope based on specific rules. In the end results are compared and accuracy is measured on majority voting from applied data mining ensemble techniques. Results prove that proposed research framework is accurate and can be used for future development.**

Keywords—Ensemble techniques; Syncope Prediction; Data mining; Naïve Bayes; Support Vector Machine; Gini Index

I. INTRODUCTION

There are many Intelligent System models used for prediction and speculation in many areas as health, weather forecasting and energy etc. The process of Knowledge discovery from data contains many decisions and subtasks. Data transformation for decision making is the core procedure of knowledge discovery [1]. Valuable knowledge can be extracted from health care data by applying different data mining techniques [2].

Syncope is immediate and temporary loss of awareness which is caused by a fall in blood pressure. Commutative Incidence of Vasovagal Syncope is 35-39% in the lifetime of subjects up to age of 65 years [3, 4].Quality of life of a Vasovagal Syncope patient significantly reduces, particularly in patients that go through recurrent episodes of Syncope [5, 6]. Furthermore symptoms of syncope are directly related to 3% of the total visits to emergency rooms and 6% admissions of patients in hospital. Currently the diagnosis of syncope is based on HUTT (Head up Tilt Test) during which the symptoms are reproduced [7].

A large group of people among our population experience Syncope. The quality of life of an individual is not only reduced by experiencing Syncope it can also lead to death in some cases. Falls during an episode of Syncope can cause severe injuries and if a person undergoes an episode of Syncope while driving, this not only risks his life but also puts lives of other people in danger. Hence Syncope should be taken seriously and proper treatment should be adapted. Researchers in area of Machine learning give different data mining techniques to analyze and predict diseases, Syncope is hardly discussed in their area of interest for research practices. Therefore, this paper presents a model to predict Syncope disease in a patient using vote based ensemble technique.

During 55 minutes of HUTT initially a patient is positioned in a supine position for 5 minutes on the tilt table, after that the patient was tilted at an angle of 70 degrees, if any symptom of syncope did not occur during 30 minutes in tilt position the patient is administered sublingually a Glyceryl Trinitrate (GTN) tablet of 0.5mg and is maintained in the same upright position for another 20 minutes. B.P and Pulse of patient is monitored during whole test and the patient is

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

immediately returned to supine position any symptom of pre-syncope or syncope occurs during any stage of HUTT. After HUTT, report of Patient's is considered positive if patient experienced syncope as defined by the European Society of Cardiology [8] and the report of the test is considered negative if patient did not experience syncope.

In this paper, a framework is proposed for prediction of syncope disease in patients. Syncope prediction is performed on the basis of an ensemble technique using three classifiers namely Naïve Bayes, Support Vector Machine (SVM) and Decision Tree Induction based on Gini index. Finally the label is assigned using majority voting on the three classifiers result. A discussion on the experimental results based on case study is performed in order to achieve better understanding of results. The following sections in this paper present literature review of work carried out by other researchers. Section 3 describes the dataset in detail. In section 4 proposed ensemble technique is explained. Section 5 is based on discussion on experimental results. In the end the section 6 presents conclusion and future work.

II. Related Work

Syncope is a short term, self-limited loss of consciousness, usually leading to falling [9, 10]. Abu Khousa et al. [11] presents the heart disease prediction system and predictive models of data mining are used for making decision support systems for heart disease. Based on the Framingham study [12], the occurrence of syncope was 6.2 per 1000 person-year, which means a 42% generality of syncope during the life of a person living 70 years (assuming stable occurrence rate over time). The incidence increases with age starting at the age of 70 (23% prevalence during a 10-year period in the population older than 70).1n another research Jens et al. [13] performed a study on patients who had a history of undefined faints. They monitored their Blood Pressure, Heart Rate and Pulse arrival time during HUTT. For this purpose initially data of 51 patients was considered but after detailed evaluation of dataset only data of 39 patients was considered for further processing. Only relevant features are considered which reduces the complexity of proposed model. An expert system based on ensemble techniques is proposed by Saba et al. [14]. Basic goal of this research was the prediction of heart disease, that whether a person has heart disease or not. Several data mining techniques are used for identification of heart disease. Dataset that are used in the research for training and test purpose is obtained from UCI online repository. They used three data mining algorithms classifiers mainly SVM, decision tree using Gini- Index based induction and naïve Bayes are trained using obtained dataset and the classifiers are then used to classify patients as healthy or sick. This system predicted heart disease up to 81.82% accurately. A framework for Classification of Syncope using k-Means Clustering Algorithm is proposed by Madiha et al. [15]. Basic goal of this research was the prediction of syncope disease, that whether a person has syncope disease or not. K-means data mining techniques used to identify the presence of syncope disease. Dataset used in this research for training and test purpose is obtained from local hospital of Cardiology in Rawalpindi, Pakistan which is run by Armed Forces of Pakistan. A framework was proposed to classify patients using K-Means clustering algorithm

Although researchers give various data mining techniques regarding (diseases), syncope is rarely discussed in contemporary research practices. Therefore, this paper presents a detailed framework for prediction of syncope using Ensemble techniques of datamining including Naïve Bayes, Support vector machine and Gini index.

III. Case Study for the Proof of Concept

The data set used in this research was obtained from a local Cardiology hospital (AFIC & NIHD) in Pakistan running by Armed Forces of Pakistan. The list of attributes was finalized with the help of domain experts at AFIC & NIHD so that no important attribute related to syncope prediction slips of the list. For attributes we decided to follow '*Lazy Tactics*' — i.e. drop attributes on analysis-steam when not required. If at any later stage set of attributes don't appear to resonate with our objectives complete attribute-set can be truncated to a subset for clarity. The dataset comprised of unstructured text form reports of 72 patients that contain information related to results of the test. Test reports did not contain detailed information related to the symptoms that patient had at the time when he/she experienced syncope, for that reason there was also a closed end questionnaire with 31 attributes designed with the help of Cardiologists for taking history manually from patient prior to test. The obtained dataset included patients of various ages. All 31 attributes were included in the questionnaire filled prior to each patients test while last three attributes; blood pressure, pulse and test result were extracted from the report collected after the test. List of attributes is given in the table 1 below.

TABLE I. List of Attributes

Gender	Age
Heart rate	Blood pressure
Blurred vision	Lightheadedness
Drowsiness	Dizziness
Vomiting	Nausea
Heart sinking	Palpitation
Shortness of breath	Chest pain
Loss of consciousness	Blackout
Diaphoresis	Feeling of warmth
Myoclonic jerks	While driving
Fainting after meal	Traumatic injury
Fainting after exercise	Family history of syncope
Patients history of other diseases	Prolonged standing before syncope
Feeling unsteady or weak while standing	Volume depletion
History of cardiac arrhythmia	

Data preparation and variable selection are the two most important steps that need to be taken care of in order to perform rest of the research. Data understanding is proposed because it is very important to develop an understanding of domain knowledge before performing any data mining activity.

- Data preparation

All this Information was stored in a database so that it could be used for further processing. After the database was created textual data was transformed into numeric values as they are compatible with machine learning algorithms. Data is

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

moved towards the next phase once it is converted into structured format.

- Data preprocessing

All the data preprocessing is performed using Rapid miner tool. During this phase missing values of attributes are replaced with the help of "replace missing vales" operator in Rapid miner. After that data type is converted by using "numeric to binomial" operator. After that data normalization is performed using "Normalize" operator to improve the performance of ensemble techniques of data mining.

- Data Mining

We start with the Syncope disease dataset which is then partitioned into training and test datasets to avoid over-fitting of the classifiers. Next, the variables are selected for further classification. 31 attributes were selected for this purpose. Selection of attributes is very important because all results are based on section of relevant attributes. Three data mining algorithms namely Naïve Bayes, Support Vector Machine (SVM) and Decision Tree Induction based on Gini index were applied. Finally the label was assigned using majority voting on the three classifiers result.

Naïve Bayes Classifier

The Naïve Bayes classifier is based on the rule that presence or absence of a particular disease is not dependent on the other attributes. The features of a class are independent of others. The probability model of Naïve Bayes can be efficiently trained using supervised learning [16]. The Naïve Bayes classifiers work very well, for many complex situations, than any other algorithm. The recent analysis of research shows that Naïve Bayes algorithm has theoretically unreasonable efficacy of classification [17].The main advantage of Naïve Bayes classifier is that it requires a small amount of dataset for training and estimation such as central tendency and spread of the parameters that are required for classification. As the attributes as independent of each other, only the attribute of given class are need to be estimated instead of entire covariance matrix [28].

Following is the formula to calculate probability of given dataset.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

After we apply this formula we get outcome class that has more information depending on given input from dataset values. The data is then added into classifier to get accurate results and to get better probability. This classifier assumes that all features in given dataset are independent to each other so it can easily done classification. For example if we talk about classification of a fruit such that an apple and if we list of features; round, red and 3 in diameter. If these are dependent on other features set then Bayesian classifier consider all of them as independent attributes to the probability of an apple.

1) Decision tree using Gini Index

In a excel file we mostly save large number of values and datasets which contain large number of attributes and some time it happens that we need some values which are actually not attributes for some other classifiers. So the purpose to use Gini index is to reduce that number of items to produce informative attribute procedures which are more compact and those attributes which have lowest Gini Index. Here is formula for this.

$$Gini\ Index(t) = 1 - \sum_{i=0}^{c-1} \left[p\left(\frac{i}{t}\right) \right]^2$$

2) Support Vector Machine (SVM)

Support vector machine classifier is used to perform binary classification in form of 1 or 0 and to determine high dimensional feature space using support vectors. We have used two attributes for this classifier and they are selected on the basis of maximum value which is gain value. Let p_i be the probability that an arbitrary tuple in dataset D belongs to class C_i, estimated by $|C_{i,D}|/|D|$. The expected results are needed to classify a tuple in D is given as

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Information gained by attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

In the dataset of Syncope 'blood pressure' and 'heart attack' are two main attributes have the highest information gain and thus are selected for Support Vector Machine classifier. SVM classifies data into binary format like 1=Yes and 0=No.

IV. PROPOSED FRAMEWORK

The proposed framework based on ensemble technique is given in fig. 1. We start with the Syncope disease dataset which is then partitioned into training and test datasets to avoid over-fitting of the classifiers. Next, the variables are selected for further classification. We choose 31 attributes for this purpose. We applied three data mining classifiers namely Naïve Bayes, Support Vector Machine (SVM) and Decision Tree Induction based on Gini index. Finally the label is assigned using majority voting on the three classifiers result.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

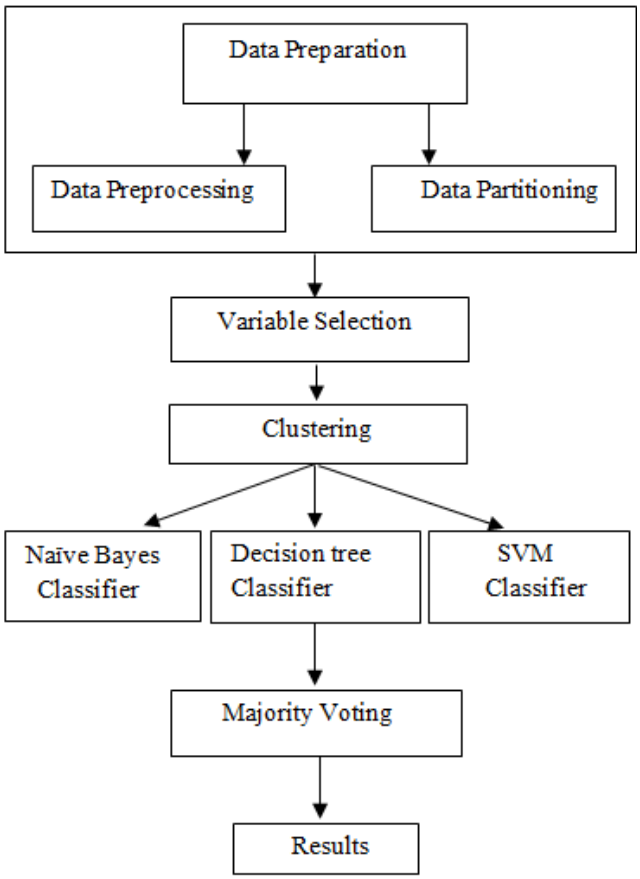

Fig. 1. Proposed Framework

The tasks in proposed framework were carried out in seven major steps which are as under:

1) First step is to prepare data in appropriate way. Data with missing values is removed from the dataset. Final prepared data set is fed to the Classifiers to perform further operations.

2) In The Second step, Naïve Bayes classifier (NBC) is applied on preprocessed dataset using Rapid minor.

3) 3rd step is to prepare decision tree using Gini Index to construct decision tree classifier (DTC). It is model that is based on Gini index and generated set of rules as a result of tree composition.

4) In 4th step Support Vector Machine (SVM) is applied on Syncope disease data set. Only two attributes were used in this step and these two attributes were Heart rate and Blood Pressure.

5) In the 5th step we used ensemble architecture to classify data. Data is classified in two classes which are; patients having/not having Syncope disease (Yes or No) with values 1 and 0.

6) Each classifier is fed with testing data for classification and each of three classifiers shown results in class 0 means healthy person or class 1 means a person with Syncope disease. Final label is assigned on the basis of majority voting.

7) Output of three classifiers is used as an input to the voting system and class with high voting result is assigned the tuple.

8) Effectiveness of proposed method is evaluated with the help of sensitivity, specificity and accuracy of proposed technique.

V. Discussion on Experimental Results

We have applied our algorithms on dataset of 72 patient's containing 31 attributes. These all attributes have continuous values. Three classifiers which are Naïve Bayesian classifier, SVM Classifier, and Decision tree based on Gini index classifier were used to classify and identify the attribute values for identification of Syncope disease in a patient. 31 attribute values were fed in Naïve Bayes and decision tree whereas for SVM classifier only 2 attributes (Blood pressure and heart rate) were used. Results for three classifiers are given in the confusion matrices shown in Table 2. Table 3 shows the result of sensitivity and specificity of the three classifiers and the accuracy comparison is shown in Table 4. The results show the superiority of the proposed approach.

TABLE II. Performance Evaluation of Classifiers Results

Predicted Class			
Classifier		**Healthy Patient**	**Sick Patient**
Naïve Bayes	Healthy Patient	33	3
	Sick Patient	9	27
Decision tree	Healthy Patient	31	5
	Sick Patient	11	25
SVM	Healthy Patient	29	7
	Sick Patient	13	23
Proposed	Healthy Patient	33	3
	Sick Patient	9	27

TABLE III. Sensivity and Specificity of Classifiers

Classifier	**Sensitivity**	**Specificity**
Naïve Bayes	97.42%	97.86%
Decision Tree	63.16%	87.71%
SVM	73.68%	88.57%
Proposed	97.42%	97.86%

TABLE IV. Accuracy Comparison

Classifiers	**Correctly Classified**	**Incorrectly Classified**	**Accuracy**
Naïve Bayes	70	2	98%
Decision Tree	63	9	87.5%
SVM	64	8	88.8%
Proposed	70	2	98%

VI. Conclusion and Future Work

Proposed technique predicts the presence of syncope disease more accurately. Three classifiers naïve Bayes, decision tree and support vector machine are used. Classification of attributes and diagnosis of syncope disease in a patient is performed using these classifiers. The prediction of syncope disease is also computed for the proposed ensemble

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

technique. Dataset contained continuous values before the model construction. Results exhibit that the proposed technique has higher accuracy as compared to the accuracy of all three techniques if individually applied.

Database of syncope patients is designed that could be used in future for research and for practical implementation of prediction systems. Our proposed technique can be extended in order to calculate intensity of syncope. For this purpose fuzzy learning models can be applied. This technique can also be tested on datasets other than syncope.

REFERENCES

[1] Abu Khousa, E.; Campbell, P., "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems," International Journal of Computer Applications, vol. 17, pp. 267-272, 2012.

[2] Walid Moudani, Dynamic Features Selection for Heart Disease Classification, World Academy of Science, Engineering and Technology Issue 0074 February.

[3] Sheldon RS, Sheldon AG, Connolly SJ, Morillo CA, Klingenheben T, Krahn AD et al. Age of first faint in patients with vasovagal syncope. J Cardiovasc Electrophysiol 2006;17:49–54.

[4] Ganzeboom KS, Mairuhu G, Reitsma JB, Linzer M, WielingW, van Dijk N. Lifetime cumulative incidence of syncope in the general population: a study of 549 Dutch subjects aged 35–60 years. J Cardiovasc Electrophysiol 2006;17:1172–6.

[5] Rose MS, Koshman ML, Spreng S, Sheldon R. The relationship between health-related quality of life and frequency of spells in patients with syncope. J Clin Epidemiol 2000;53:1209–16.

[6] van Dijk N, Sprangers MA, Boer KR, Colman N, Wieling W, Linzer M. Quality of life within one year following presentation after transient loss of consciousness. Am J Cardiol 2007;100:672–6.

[7] D. G. Benditt, D. W. Ferguson, B. P. Grubb, W. N. Kapoor, J. Kugler, B. B. Lerman, J. D. Maloney, A. Raviele, B. Ross, R. Sutton, M. J. Wolk, and D. L. Wood, "Tilt table testing for assessing syncope," Am. Coll. Cardiol, vol. 28, pp. 263–275, 1996.

[8] N. Colman, K. Nahm, K. S. Ganzeboom, W. K. Shen, J. Reitsma, M. Linzer, W. Wieling, and H. Kaufmann, "Epidemiology of reflex syncope," *Clin Auton Res,* vol. 14 Suppl 1, pp. 9-17, Oct 2004.

[9] Guidelines on management (diagnosis and treatment) of syncope. Eur Heart J 2001; 22:1256-1306.

[10] Benditt D.G., Brignole M., Raviele A. and Wieling W., Syncope and transient loss of consciousness: multidisciplinary management, Blackwell Publishing 2007, ISBN: 978-1 4051-7625-5.

[11] Walid Moudani, Dynamic Features Selection for Heart Disease Classification, World Academy of Science, Engineering and Technology Issue 0074 February 2013

[12] Soteriades ES et al, "Incidence and prognosis of syncope", N Engl J Med 2002; 347 (12): 878-885.

[13] Jens Muehlsteff, Anita Ritz, Thomas Drexel, Christian Eickholt, Paulo Carvalho, Ricardo Couceiro, MalteKelm, Christian Meyer, Pulse Arrival Time as Surrogate for Systolic Blood Pressure Changes during Impending Neurally Mediated Syncope, 34th Annual International Conference of the IEEE EMBS San Diego, California USA, 28 August - 1 September, 2012.

[14] Saba Bashir, Farhan Hassan Khan, UsmanQamar, Intelligent Heart Disease Prediction System using Ensemble Techniques.

[15] Madiha Guftar, Syed Hasnain Ali, Ammar Asjad Raja, Usman Qamar, A Novel Framework for Classification of Syncope Disease using k-means Clustering Algorithm, SAI Intelligent Systems Conference 2015(IntelliSys), London, England, 10 - 11 November, 2015.

[16] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.

[17] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications for scientificdatabases", Proc. of the 9th Int. Conf. on Scientific andStatistical Database Management, Olympia, Washington,USA, 2-11, 1997.

[18] Intelligent Heart Disease Prediction System Using Data Mining Techniques-SellappanPalaniappan, RafiahAwang 978-1-4244-1968-5/08/ ©2008 IEEE.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE