Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Introduction to a New Emphatic Model for Predicting Patients and Understanding Mining Techniques

Karandeep Kaur
M.Tech.(scholar)
Computer Engg.& Technology
Department GNDU Amritsar, Punjab, India

Kiranbir Kaur
Assistant professor,
Computer Engg.& Technology
Department GNDU Amritsar, Punjab, India

Abstract: Data mining is the act of inspecting substantial prior databases to create new data, choosing and exploring large volumes of information. This procedure has turned into an inexorably forceful activity in every aspect of health care and research. A lot of scattered data about health is accessible. This paper actualizes an imaginative thought to recognize ailments influenced in humans and gives the results in terms of accuracy. Hybrid algorithm approach of simple logistic and IB1 is used. The image set of the sick humans are caught with the assistance of high pixel cameras or cell phones like android, iPhone or remote PDA. These images are then nourished for application for distinguishing human ailments and recommend solutions. Result in terms of accuracy shows improvement by15% than existing algorithms.

Keywords: Data Mining, Diseases, Simple logistic, IB1, accuracy.

1. INTRODUCTION

Image processing is the investigation and control of a digitized picture particularly keeping in mind the ultimate goal to enhance its quality. It is type of flag handling for which the information is a picture, for example, a photo; the yield of picture preparing might be either a picture or an arrangement of attributes or parameters identified with that picture. Most picture preparing strategies regard the picture as a two-dimensional flag and apply standard flag handling systems to it. Computerized picture preparing is the utilization of PC calculations to perform picture handling on pictures.

Data mining is the act of looking at vast prior databases to produce new data [1]. It is a computational procedure of discovering patterns in extensive informational collections including techniques like machine learning, insights, and database frameworks. The objective of data mining is not just to extract the information from database is to learn new patterns and facts from this collected extensive measures of information. The general objective of the data mining process is to remove the data from an informational index and change it into a justifiable structure [2] [3]. There are two types of information examination steps that can be utilized for separating models.

These two structures are –

- Prediction
- Classification

The prediction model performs information examination, while classification model performs information investigation.

Basically in image processing, the process begins with the digitized shading picture of human sickness [4]. At that point a technique for arithmetic morphology is utilized to section these pictures. Here in this paper, Disintegration technique has been utilized.

Following are some of the basic steps that are followed in this paper, for acquiring the results.

- Collecting dataset
- Performing correction of dataset.
- Apply preprocessing strategy in order to achieve rectification of missing values by eliminating them
- Achieving Normalization
- Applying K-means Clustering to accomplish Clustering.
- Generating Decision Tree for classification and Results.

2. EXISTING SCHEMES USED FOR CLASSIFICATION PURPOSE

There exists number of techniques which are available and analyzed for prediction and achieving accuracy. These methods are discussed in this section.

2.1 NAÏVE BAYES

The Bayesian Classification is a learning technique and a measurable strategy for performing arrangement of data [5] [6]. It makes use of different classifiers that are based on bayes theorem. It basically accepts a fundamental probabilistic model and then enables us to know the vulnerability about the model by deciding probabilities of the results. It can take care of analytic and prescient issues.

To describe the utilization of Naïve Bayes Approach consider a supervised learning algorithm for problem of approximation a target f:X->Y or equivalently P(X/Y). It basically use the conditional probability for performing experiments on different sets of data.

Probabilistic values generated from P(X/Y) along with “if - then” rules are used for classifiction.

2.2 J48

J48 classifier is a basic C4.5 Decision tree for grouping[7]. It makes a paired tree. The choice tree approach is also used in J48. With this procedure of choice tree, a tree is developed to show the characterization procedure. Once the tree is assembled, it is connected to each tuple in the database and brings about arrangement for that tuple. While building a tree, J48 overlooks the missing qualities i.e. the incentive for that thing can be anticipated in view of what is

thought about the characteristic qualities for the other records[8]. The fundamental thought is to separate the information into range in view of the characteristic qualities for that thing that are found in the preparation test. J48 permits arrangement by means of either choice trees or principles created from them.

2.3 SIMPLE LOGISTIC

Logistic regression is used to fit the probablistic function into the curve[9]. The probablistic function which is used is of the following form.

$$\pi(x) = \exp(\alpha + \beta x)\exp(1 + \exp(\alpha + \beta x))$$

A calculated relapse tree display offers one approach to at the same time hold the graphical interpretability of straightforward models and the prescient exactness of wealthier ones[11]. Its fundamental inspiration is to make an unpredictable arrangement of information and then divide this information into adequately numerous subsets ,so as to achieve the final goal. Information subsetting is performed recursively, with the specimen part on one variable at any given moment. This outcomes in the parcels being representable as a double choice tree. A Simple Logistic relapse is a Logistic relapse with just a single parameters. For the speculation (ie with more than one parameter) is reffered to as Multi-variation strategic relapse .

A linear regression will only work with values outside the acceptable range i.e. they will work with the probabilities having values outside the range of 0 and 1. But in the logistic regression the values of probability of an outcome can have only two values i.e. 0 and 1. This logistic regression is similar to linear regression but in case of plotting a graph they both vary. The logistic regression uses the concept of natural logarithms rather than probability.

2.4 SMO

SMO is actually abbreviated for sequential minimal optimization. It is a sort disintegration strategy under a general and adaptable method for picking the two-component working set.It is an algorithm that is used to solve the problems of quadratic programming (QP). These quadratic programming problems arise during the training of SVM (support vector machines).SMO takes care of the SVM QP issue by breaking down it into QP sub-issues, then further handles these small sub parts. Disintegration strategies are at present one of the significant techniques for preparing bolster vector machines[12] [13]. Changes occur in them in large extent as per distinctive working set determinations.

2.5 IB1

Lazy classifiers in weka tool has this classifier IB1. Lazy classifiers are actually those classifiers that normally delay the construction of classifiers until the classification time is achieved. This classifier is comprised of many algorithms in which IB1,IBK, KSTAR algorithms are present. IB1 is a abbreviation used for instance based learning. It is a nearest neighbor algorithm which classifies any of the instance according to nearest neighbor distance , which is found by using euclidean distance. This algorithm works same as the nearest neighbor algorithm but along with that normalization is also done on the attribute's range and also it processes the instances incrementanlly which the NN is not able to do. IB1 is an algorithm which does not make any assumptions regarding convexity of the target, the number of components attached to it, nor the relevant positions of those components. Only the informative instances are stored in memory due to which we can reduce the memory requirement.

Occasion Based learner stores all the illustrations, when performing task it considers the most suitable instances and make experminents on the premise of those occurrences[14], [15]. The established case of IBL is K nearest neighbor. It discovers K most comparable occasions and gives most class. Similarity based function is used for the purpose of classifiation. The equation for the same is shown as under

$$x_i, y_i = (x_i - y_i)^2$$

$x_i and\ y_i$ gives the closest points that can be tackled through the system being examined.

Hybridization can be used to enhance the performance further[16]. The enhancment in terms of accuracy is obtained. The existing literature consider the same but the proposed work involving hybridization of Simple logistic and IB1 is better in terms of accuracy. IB1 ia closest neighbor calculation that arranges an occurrence as indicated by the closest neighbor distinguished by eculedian distance. IB1 algorithms have capability of using classes i.e. Nominal class, Missing class values, Binary class and attributes like Nominal attributes, Binary attributes, Date attributes, Missing values, Empty nominal attributes, Numeric attributes, Unary attributes accident[16]–[18]. Instance based Learning (IBL) is characterized into sluggish classifier for putting away all of preparing cases and does not truly work until the order procedure runs. IB1 generally exact in ordering yet require an expansive storage room for putting away all the preparation cases. IBL calculations order case by contrasting it and preparing information (predefined class) so that a comparable case will have a comparable grouping utilizing the cosine comparability work. Class choice is finished by taking a number n of the most elevated closeness esteem class that turned into the most broadly arrangement comes about. Each test archive that effectively characterized will be included the preparation report[19], [20]. To decide the level of closeness, this review utilizes Cosine Similarity to ascertain the separation between two purposes of the record and produce a rundown of coordinating preparing occurrence and test example sorted by level of 79 comparability. Cosine Similarity acquired from the condition.

$$similarity = \cos(\theta) = A.\frac{B}{|A||\ \ ||B|}$$

$$= \sum_{i=1}^{x} Ai\ X\ Bi / \sqrt{\sum_{i=1}^{n} (Ai)2\ X\ \sqrt{\sum_{i=1}^{n} (Bi)2}}$$

Where: A = test document
B = training document
n = amount of terms in a document
i = index of a term

3. PROPOSED WORK

The proposed work is based on estimating disease in humans by hybridizing Simple Logistic with IB1.
The steps for this model is as under

I. Understand the data set of Cancer in perspective of machine learning

II. Convert the data set in machine understandable format
III. Investigate missing values and cleaning of data using appropriate algorithm.
IV. Visualizing dataset patterns
V. Understanding various data mining algorithms for prediction.
VI. Propose an approach for predicting Cancer
VII. Model Evaluation
VIII. Comparison of model with exiting approaches
IX. Applying test dataset in the model to get results
X. Clustering

The model for the same is described as under

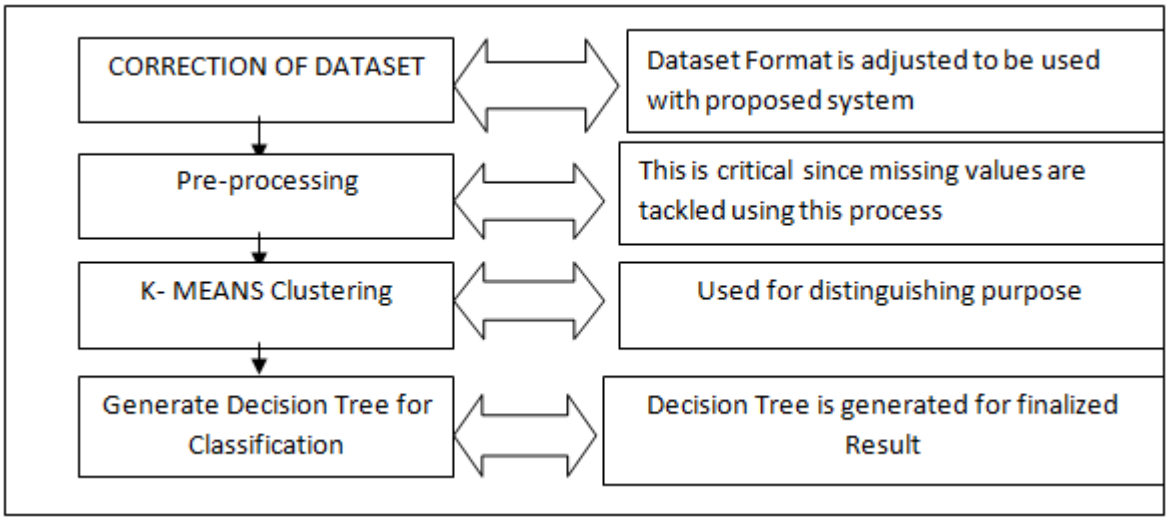

Figure 1: Model for proposed work

3.1 Correction of Dataset

Dataset is obtained and feed into proposed model for result. The problem of format within the dataset is mostly present like some of the values are not clearly present, some data is not according to the format, all values are not present etc. . .

In order to tackle this problem format adjustment is performed. This is accomplished with the help of correction of dataset. Simulation is conducted in Weka tool which accept corrected dataset only hence correction becomes critical to classification

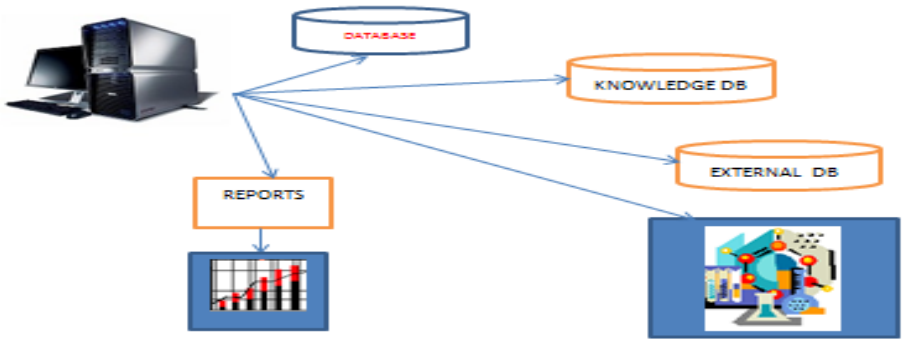

Figure 2: Model Attachment with dataset

3.2 PRE-PROCESSING

Again a critical phase, describing artficats removal from the dataset used for classfication. Artificats in terms of missing values is present. The pre-processing mechanism provides best possible dataset free from distortion. Distortion elimination enhances accuracy greatly. Earlier approaches such as Naïve bayes does not consider this phase and degradation in performance is substantial. Pre-processing performed at appropriately can increase performance almost by 5%. So, this task should be performed at the initial phases to improve the efficiency of work.

3.3 K-MEANS CLUSTERING

Clustering mechanism is used in order to distinguish values from each other[4] . Most of the researchers uses a K-means technique for detecting disease and performing prediction accurately by simplifying parameters[2][21]. The elements that have homogenous properties are grouped together by using grouping functions and these elements have been identified by nearest neighborhood algorithm. For determining the problem the comparison of threshold values against the values generated by grouping function are to be done. Problems are reflected in the form of deviation. The process is described by considering two points ‘A’ and ‘B’. Let distance(A,B) is the distance between points A and B then

a. distance(A,B)=0 and distance(A,B) >=0 iff A=B
b. distance(A,B)=distance(B,A)
c. distance(A,C)<=distance(A,C)+distance(C,B)

Property c is also known as transitive dependency. Distance if close to zero then prediction is accurate otherwise error is said to be present. Error calculating metrics is applied to determine accuracy of the approach. Accuracy is given as

Accuracy=1-Error_rate

where Error rate is given as

$$\text{Error rate} = \frac{|X - X_a|}{X_a}$$

K-mean algorithm is used in many different areas such as classification, interpolation, problem solving, teaching and learning etc. [22]. Major limitation of K means is that its performance depends upon value of k, accuracy is low, and further work is required to be done to improve accuracy.

Clusters of similar features are collected and analyzed. Proposed work uses K-means for clustering for faster detection rate.

.

3.4 DECSION TREE FOR RESULTS

Decsion Tree consist of nodes[23], [24]. These nodes performs test for the particular attributes. Edges indicates outcome for the test and leaves of the tree gives the final outcome. Classification process through decision tree corrsponds to following steps

3.4.1 Start at the root
3.4.2 Perform the test
3.4.3 Move towards the nodes following the edges
3.4.4 Label the outcome
3.4.5 Goto step 3.4.2 until leaf node is not reached
3.4.6 Predict whether outcome corresponds to the leaf

Accuracy is high if outcome belongs to the leaf node otherwise misclassification is obtained

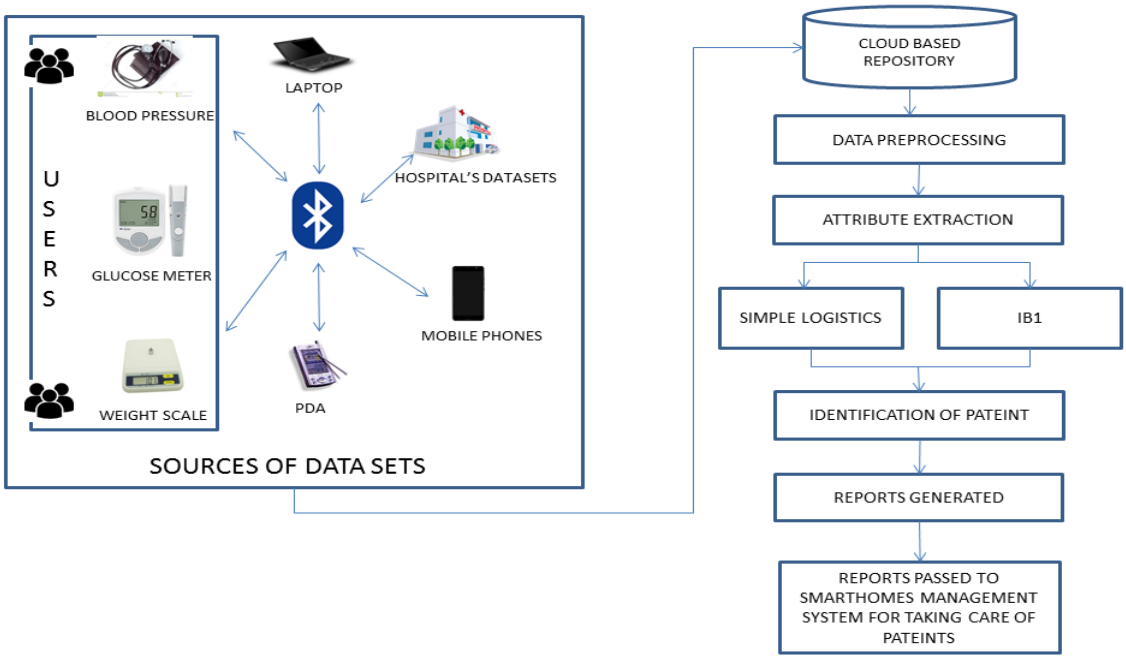

Figure 3: proper working of Proposed system

Dataset required for performing operations on, and for identification of the pateint is basically collected form different sources. The sources here are mobile phones, computers, laptops, different pda machines,sensors which with use IOT technology or even from hosptial datacenters who have a complete knowlegde of the pateints. Data present in the datacenters of hospital is collected by performing various tests on the pateints using different tools like blood pressure meter, glucose meters, and many more. All these information are gathered and stored. This information about various people can be stored online on cloud also or on personal harddrives . but to store this data in cloud is a good option as this data is present in huge amounts. One more advantage is also provided, that when the data is stored on cloud various inbuilt functions and tools required to process data.

Now data form these cloud repositories are collected and various pre processing techniques on this data is performed. The preprocessing of data involves proper checking of data for missing values, checking whether the data is in correct format or not , and data cleaning is performed etc.. After that attritube extraction takes place which basically performs feature extraction.Here, attribute is selected for which we will further work on. Various different patterns are visualized for data present in these data centers. After this step, different data mining algorithms are analyzed and then selecting the best from all and applying that algorithm on data. Here, in our proposed model we have used the hybrid of two agorithms i.e. simple logistics and IB1 algorithms. After applying we will compare our values with some existing approaches and the performance of our algorithm is analysed . We can also perform clustering of data which will group homogeneous groups of data together and then check for performance. So by this method we will be able to identify correctly that which pateint is suffering from disease. After this identification of pateint we can pass this information about the pateint to hospitals and their family members so that proper care and diagonsis can be performed

for the pateint, to recover patient from the chronic disease. These reports can be passed to smart home system managements so that pateint can be examined continously at their homes and if any problem occurs , i.e. if it is observed that patient is not feeling well immediately doctors,or nurse, or any of relative of pateint can be informed.

4. PERFORMANCE ANALYSIS

Performance analysis is accomplished be comparing perposed system(Simple Logistic with IB1) with other classifiers. The obtained results in terms of accuracy is listed as under Parameter-Accuracy

Table 1: Accuracy obtained from various classifier and its comparison with proposed approach

Algorithm	Accuracy
Naive Bayes	75.1748
J48	75.8741
Naive Bayes+J48	74.4755
SMO+J48	76.2238
Simple Logistics	76.2238
Simple Logistics+Naive Bayes	75.1748
Simple Logistics+SMO	76.2238
SMO+IB1	82.1678
Simple Logistics+IB1	97.5524

The degree of missclassification is also obtained as listed below

Table 2: Degree of missclassification

Algorithm	Misclassification
Naive Bayes	24.8252
J48	24.1259
Naive Bayes+J48	25.5245
Naive Bayes Updatable+J48	25.5245
SMO+J48	23.7762
Simple Logistics	23.7762
Simple Logistics+Naive Bayes	24.8252
Simple Logistics+SMO	23.7762
SMO+IB1	17.8322
Simple Logistics+IB1	2.4476

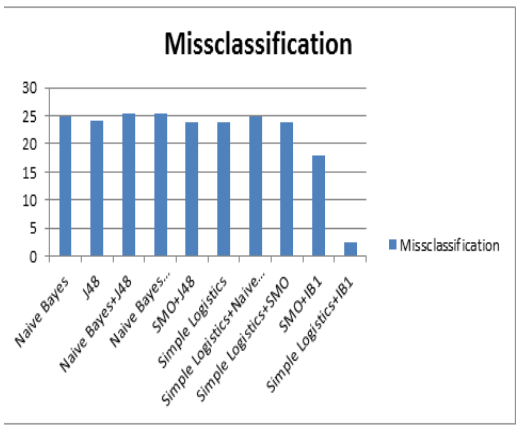

Figure 4: Plots for missclassification describibg existing and proposed technqiues

Image mining is the way toward looking and finding significant data and learning in huge volumes of information. Data mining draws fundamental standards from ideas in databases, machine learning, measurements, design acknowledgment and "delicate" figuring. Utilizing information mining procedures empowers a more proficient utilization of information. It is in this manner turning into a rising examination field in geosciences in view of the expanding measure of information which prompt new encouraging applications. For instance, the utilization of high determination satellite pictures now empowers the perception of little questions, while the utilization of high transient determination pictures empowers checking of changes at high recurrence. To be sure, earth's perception information (obtained from optical, radar and hyperspectral sensors introduced on earthbound, airborne or spaceborne stages) is regularly heterogeneous, multi-scale, inadequate, and made out of complex articles. Division calculations, unsupervised and regulated order techniques, clear and prescient spatial models and calculations for expansive time arrangement investigation will be introduced to help specialists in their insight revelation.

Performance analysis indicates that proposed work (Simple Logistics and IB1) performs better in terms of accuracy as well as missclassfication.

5. CONCLUSION

This paper actualizes a creative thought to distinguish sicknesses influenced in humans and gives the cure/answer for concerns in the form of accuracy in identifying pateint. The pictures are then bolstered to our application for distinguishing proof of human illnesses and the cures are recommended to the concerns. This execution gives better decision to each class of disease detection group especially in terms of accuracy.

Performance analysis also indicates algorithms like Naïve bayes and J48 lack performance or degrades since preprocessing phase is absent. By merging multiple algorithms(Simple Logistics+IB1) error rate(Missclassification) decreases and accuracy enhances.

In future, LBP, PCA , DWT and their hybridization can be used to enhance accuracy and performance of the system being utilized.

6. REFERENCES

[1] D. Tyagi, "using Local Ternary Pattern with GA and SVM classifier," pp. 421–426, 2016.

[2] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, 2013.

[3] K. Thenmozhi and P. Deepika, "Heart Disease Prediction Using Classification with Different Decision Tree Techniques," vol. 2, no. 6, pp. 6–11, 2014.

[4] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, "Spectral Ensemble Clustering via Weighted K-means: Theoretical and Practical Evidence," *IEEE Trans. Knowl. Data Eng.*, vol. XXX, no. XXX, pp. 1–1, 2017.

[5] K. Chai, H. T. Hn, and H. L. Cheiu, "Naive-Bayes Classification Algorithm," *Bayesian Online Classif. Text Classif. Filter.*, pp. 97–104, 2002.

[6] T. M. Mitchell, "CHAPTER 1 GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND

LOGISTIC REGRESSION Learning Classifiers based on Bayes Rule," *Mach. Learn.*, vol. 1, no. Pt 1-2, pp. 1–17, 2010.
[7] T. R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *Int. J. Comput. Sci. Appl. ISSN 0974-1011*, vol. 6, no. 2, pp. 256–261, 2013.
[8] Y. Chauhan and J. Vania, "J48 Classifier Approach to Detect Characteristic of Bt Cotton base on Soil Micro Nutrient," vol. 5, no. 6, pp. 305–309, 2013.
[9] W.-Y. Loh, "Logistic Regression Tree Analysis," *Handb. Eng. Stat.*, pp. 537–549, 2006.
[10] S. J. Sheather, "Logistic Regression," *A Mod. Approach to Regres. with R*, pp. 263–303, 2009.
[11] I. Schuster and P. Jähnichen, "Classification using Logistic Regression," 2012.
[12] S. Minimal, "Part One - Background What ' s a QP problem ?," *Optimization*.
[13] P.-H. Chen, R.-E. Fan, and C.-J. Lin, "A study on SMO-type decomposition methods for support vector machines.," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 893–908, 2006.
[14] M. Sahu, N. K. Nagwani, S. Verma, and S. Shirke, "Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal," *Int. J. Knowl. Eng.*, vol. 1, no. 2, pp. 141–145, 2015.
[15] T. Mitchell, "Instance Based Learning," *Mach. Learn.*, pp. 199–214, 1997.
[16] D. Adriaans, Pieter and Zantinge, "Data mining," 2001.
[17] S. Das, V. J. Nandeshwar, and G. S. Phadke, "Discrimination of adulteration orange juice by Linear Discriminant Analysis (LDA)," *2015 IEEE Int. WIE Conf. Electr. Comput. Eng. WIECON-ECE 2015*, pp. 39–42, 2016.
[18] S. Vijayarani and M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 8, pp. 3118–3124, 2013.
[19] T. S. Xu, H. D. Chiang, G. Y. Liu, and C. W. Tan, "Hierarchical K-means method for clustering large-scale advanced metering infrastructure data," *IEEE Trans. Power Deliv.*, vol. PP, no. 99, 2015.
[20] R. Kumar and R. Dwivedi, "Quaternion Domain k-Means Clustering for Improved Real Time Classification of E-Nose Data," *IEEE Sens. J.*, vol. 16, no. 1, pp. 177–184, 2016.
[21] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient â€™ s Health Parameters," vol. 8, no. 12, 1843.
[22] Y. Ning, X. Zhu, S. Zhu, and Y. Zhang, "Surface EMG decomposition based on K-means clustering and convolution kernel compensation," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 2, pp. 471–477, 2015.
[23] M. Trees, "Decision-Tree Learning."
[24] T. M. Mitchell, "Decision Tree Learning," *Machine Learning*. pp. 52–80, 1997.

