A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms

Md. Razu Ahmed[1], S M Hasan Mahmud[1, 2,*], Md Altab Hossin[3], Hosney Jahan[4], Sheak Rashed Haider Noori[1]
[1]Faculty of Science and Information Technology, Daffodil International University, Bangladesh
[2]Department of Computer Science and Engineering, University of Electronic Science and Technology of China, China
[3]Department of Management Science & Engineering, University of Electronic Science and Technology of China, China
[4]Department of Computer Science, Sichuan University, China
e-mail: razu35-1072@diu.edu.bd[1],{hasan.swe, drnoori}@daffodilvarsity.edu.bd[1], altabbd@163.com[3], hosney.hstu08@gmail.com[4]

***Abstract*—Heart disease prediction and detection has long been considered as a critical issue. Early detection of heart disease is an important issue in health care services (HCS). In growing amount of health care systems, patients are offered expensive therapies and operation that is quiet expensive for developing countries. Recently, heart disease is a prominent public chronic disease, ex. it's a growing concern in the US. The main reason of these diseases are tobacco consumption, bad life style, lack of physical activity and the intake of alcohol. Therefore, there is a need for the cloud based architecture that can efficiently predict and track health information. Recently, machine learning techniques have already been established to solve clinical problem and medical diagnosis. In this study, we proposed a cloud-based 4- tier architecture that can significantly improve the prediction and monitoring of patient's health information. Hence, we used five popular supervised learning based machine learning technique for early detection of heart disease. The major purpose of this study is to examine the performance of the selected classification techniques. In addition, we use prominent evaluation criteria to observe the best performance of these machine learning techniques. Moreover, we used the ten-fold cross-validation technique to evaluate the performance of the five classifiers. The analysis results indicate that the Artificial Neural Network (ANN) achieved the highest performance of all. However, health care researchers and practitioners can obtain independent understanding from this work while selecting machine learning techniques to apply in their area.**

Keywords-machine learning; cloud computing; heart disease; supervised learning; disease prediction

I. INTRODUCTION

Chronic heart disease (CHD) defined as a common term for the rise of plaque inside the coronary arteries that could lead to heart attack [1]. Heart diseases are severe events which are caused by blockage inside the heart arteries. CHD is a disease of long period and progresses slowly. Actually, chronic heart disease is getting widespread proportion day by day [2]. As a result, CHD is the leading cause of death globally [3]. According to the report of World Health Organization (WHO), heart disease causes death of millions of peoples annually than any other disease. About 17.7 million deaths around the world have been caused from CHD (indicating 31% of all global deaths) in 2015 [4]. Moreover, over three-quarters of CHD deaths (75% are caused by CVDs) took place in many low-income countries in the world [4]. Considering, the low-income countries, chronic diseases are a severe problem, such as Bangladesh where a shocking amount of male and female are at risk of Heart Disease at young stage (99.6% male and 97.9% of females are in risks of CVDs) [5]. World Health Organization projected that more than 23.6 million individuals will be dead by 2030, because of heart disease [6]. There are some major heart disease risk factors such as peculiar glucose metabolism, extreme blood pressure, dyslipidemia, smoking, lack of physical exercise and growing age are well recognized [7]. Most of the time heart disease diagnosis is very costly and complicated. This takes obsessive time, as a result, incorrect or delayed decisions are likely to cause death. But, early detection and diagnosis can prevent it. Consequently, a computational intelligent system is suggested that uses the real-time decision from the cloud. This method can be applied to generate a better-quality decision with less exertion.

In the last decades, data has been generated in a volume of scale from diverse fields including health care services (HCS) and medical fields [8]. State of the art machine learning techniques have been applied to obtain knowledge from the health care data for research, and make effective prediction and decision for heart disease diagnosis. Machine learning (ML) methods have drawn that aim to solve different medical and clinical problem [9]. Many of the studies show that machine learning techniques have gained significantly high accuracies in classification-based problems. In recent years, machine learning-based classification methods are one of the most operational appraisement method for the research community and real-world applications [10]. Therefore, the use of these techniques for early detection, diagnosis, and cure of heart disease can significantly reduce clinical error and treatment cost. Moreover, the prediction accuracy of the machine learning techniques may vary on different conditions. Hence, most of the studies, applying ML classification techniques have been focused on the prediction and their accuracies [10] [11] [12]. To the best of our knowledge, there is no study which has referred to as the real-time cloud-based architecture for early detection of heart diseases. In this study, we present the first attempt to propose a cloud-based four-tier architecture for

978-1-5386-8339-2/18/$31.00 ©2018 IEEE

early heart disease detection. The aim of this paper is to evaluate the performance of different classification algorithms and obtain more accurate results by eliminating the high cost of heart disease diagnosis. The algorithms used in our study include: Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Naïve Bayes (NB). Moreover, the performance is compared using the confusion matrix and Receiver operating characteristic (ROC) curve. Hence, the most significant machine learning technique has been denoted for the early detection of heart disease with proposed architecture.

The rest of the paper is organized as follows, the dataset, proposed architecture, and workflow of the proposed system are described in Section 2. Section 3 describes the different classification techniques. The evaluation results are illustrated in Section 4. Finally, conclusions and ideas for future work are discussed in Section 5.

II. Materials and Methodology

A. Data Collection

In this experiment, the prediction performance of different classification algorithms has been evaluated using the Stat Log Heart Disease dataset provided by the UCI Machine Learning Repository [13] [14]. We analyzed data from 270 instances of which 120 (44.4 % true cases) samples are the presence and 150 samples (55.60% false cases) are the absence of heart disease. In the following, we provide the details of the final set of attributes [15] we choose for the data prepossessing such as,

1) Age
2) Sex (This is the binary attribute that can assume value 1 for female and 0 for male)
3) Chest pain type (categorical with 4 values)
4) Resting blood pressure
5) Serum cholesterol in mg/dl (continuous)
6) Fasting blood sugar > 120 mg/dl (binary)
7) Resting electrocardiographic results (categorical with 3 levels)
8) Maximum heart rate achieved
9) Angina provoked by exercise (binary)
10) The slant of the peak exercise ST segment (0-3 levels)
11) Number of major vessels (categorical with 4 levels) colored by fluoroscopy
12) Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
13) Old peak = ST depression provoked by workout qualified to rest

B. Proposed Architecture

In this present study, the proposed cloud-based four-tier architecture including machine learning techniques has been presented. The proposed cloud-based heart disease prediction and monitoring system consists of a four-tier architecture to store and process a huge volume of wireless sensors and device data. Tier 1 focuses on collecting and combining data from different health tracking sensors and devices. Tier 2 uses Kafka pipeline and Cassandra to store huge amount of real-time data. Afterwards, tier 3 uses machine learning classification algorithm for training and feature extraction in order to develop a real-time based architecture for early detection of heart disease. In addition, tier 4 represents the results of the whole system for the users. The proposed real-time cloud-based architecture for early heart disease detection is shown in Fig. 1. Moreover, Fig.2 represents the flow diagram for the proposed architecture.

In the proposed architecture, the real time health information is collected using different tracking devices and sensors. The tracking request is accepted by the cloud application in tier 3. Here, the racking context will check the request, if the request is equal to pre trained data, then this observed goes to cloud server and store the value. Moreover, if the observed value is higher or lower than the pre trained data, then users can get a notification.

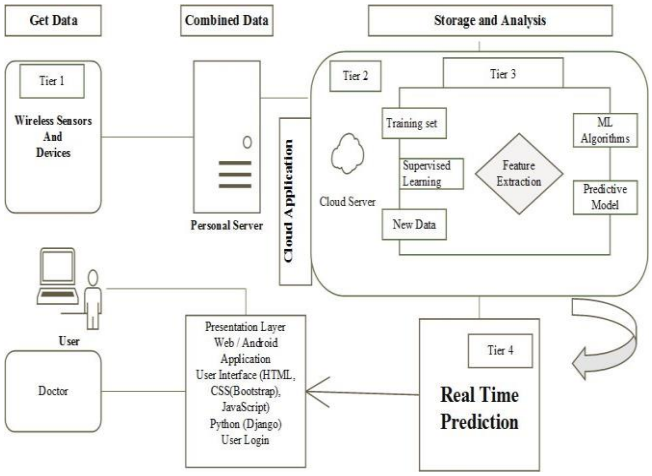

Figure 1. Proposed architecture for early detection and monitoring of heart disease

Inspired by the expressive performance of machine learning based disease predictions, this paper considers appropriate classification algorithms as well as Kafka pipeline, live stream datasets, NoSQL database (for handling the huge amount of data), cloud server and real-time data prediction service to develop a powerful solution for heart disease patients. The significant contributions of the paper are summarized as follows,

- A cloud-based architecture using machine learning for early detection and monitoring of heart disease is proposed. This architecture helps the heart disease patients to take effective suggestions and decisions for their daily life activities.
- Considering the vast amount of healthcare services data and real-time data from different health tracking devices, our proposed architecture is able to handle this large amount of data. Therefore, if we do not process this large amount of data effectively, the main aspects of those data could be missed.
- Most of the study does not consider real-time prediction. Besides, few of the studies consider the f-score, precision and recall. However, our study

provides the real-time prediction by considering the f-score, precision, and recall values.

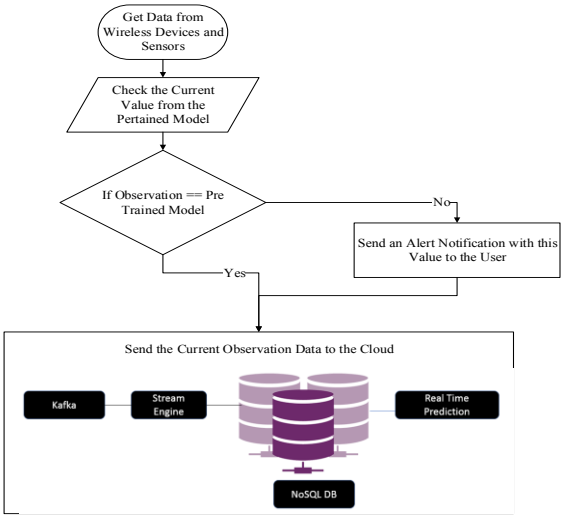

Figure 2. Flow diagram for proposed architecture

The proposed cloud-based architecture consists of four modules, specifically, data collection module, data storage module, analysis module, and application presentation module. The data collection module is used for extracting the particular patient's or person's data using health tracking sensors and devices. Health tracking devices are integrated with the human body to collect the particular person's health data in a continuous manner. In addition, tracking sensors and devices send health data uninterruptedly. For this vast amount of data, their storing and analysis becomes difficult by using the traditional database tools and techniques. The proposed architecture uses cloud computing and NoSQL database technologies to store the continuous healthcare data. Moreover, in the application module users can see their health report through the mobile application.

III. Description of the Classification Technique

A. *Artificial Neural Network (ANN))*

The AI expert Maureen Caudill defines artificial neural network as "a computing technique made up of a significant number of simple, vastly interconnected processing elements, which process information by their dynamic state response to external inputs" [16]. In machine learning, artificial neurons are the basic concept of ANN which works similar to biological neural network. However, ANN consist of three layers (an input layer, multiple hidden layers, and an output layer). Here, every node in one layer is connected to every node in the other layers. For creating the deeper neural network it expands the number of hidden layers. In addition, the outcome of the output layer is defined as its node value or activation [17].

B. *Support Vector Machine (SVM)*

Support vector machine has been first introduced by Vladimir Vapnik and Alexey Chervonenkis [18] [19]. SVM is a method of machine learning that can solve both linear and nonlinear problems. It provides good performance to solve both regression and classification problem. The SVM classification technique inspects for the optimal separable hyperplane in order to classify the dataset between two classes [20]. Finally, the model can estimate noisy data problems for new cases.

C. *Decision Tree (DT)*

Decision tree is one of the well-known supervised learning algorithm of machine learning. DT is a classification technique that divides the dataset into smaller subsets. Here, each branch represents a decision and each leaf represents an outcome. In decision tree, root of the tree is used to estimate the entropy, i.e., the information gain.

D. *Random Forest (RF)*

Leo Breiman first introduced random forest in his study [21]. Random Forest algorithm is a popular algorithm in machine learning. RF work well for many clinical and biological problems. It is able to solve both classification and regression problems in health care services. It creates a forest by different ways and make it random. In the forest of trees has been the direct relationship between the combine trees and the result it can get. To acquire more efficient and accurate prediction, random forest inserts an extra layer of randomness to bagging.

E. *Naïve Bayes (NB)*

Naive Bayes is one of the simple, most effective and commonly-used, machine learning technique [22]. It is a probabilistic classifier that classifies using the hypothesis of a conditional independence with the pre-trained datasets [23]. Henceforth, Naive Bayes classifiers are techniques for finding the traditional solution of classification problems, such as spam detection, and also well fit for medical problems.

IV. Results and Discussion

A. *Classification Performance Measurement*

In this work, five machine learning techniques were applied for the early detection of heart disease. We employed the 10-fold cross-validation approach to evaluate each classification results. Performance of the classifiers are assessed by different statistical techniques. Such as confusion matrix (TP, FP, TN, FN), Recall, Precision, f-measure etc. Hence, the validation matrix is defined by,

True Positive (TP): Prediction results are true and the patient has heart disease

True Negative (TN): Prediction results are false and the patient does not have heart disease.

False Positive (FP): Prediction results are true but the patient does not have heart disease.

False Negative (FN): Prediction results are false and the patient has heart disease.

The computation method of the measurement factors are as follows,

$$\text{Accuracy} = (TP+TN) / (TP+FP+TN+FN) \quad (1)$$

$$\text{TPR} = \text{Sensitivity} = \text{Recall} = TP / (TP+FN) \quad (2)$$

$$\text{Specificity} = \text{TNR} = TN / (TN+FP) \quad (3)$$

$$\text{Precision} = TP / (TP+FP) \quad (4)$$

$$F1 = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (5)$$

The F1 measure is defined by the average of the recall and precision. For the best performance of the classifiers, the value must be one and for the worst performance, it must be zero.

B. *Analysis of the Results*

In this experiment, we consider different analysis to examine the five machine learning classification techniques for the classification of Stat log heart disease dataset. From this datasets, total 270 samples with 120 true samples and 150 negative samples of heart disease were taken into analysis. Therefore, all of the data samples were segregated in ten folds for ten-fold cross-validation. Figure 3 shows the confusion matrix of prediction results for Naive Bayes (NB), random forest (RF), artificial neural network (ANN), support vector machine (SVM) and decision tree (DT) algorithms. Figure 4 shows the prediction accuracy of these machine learning classifiers for heart disease detection.

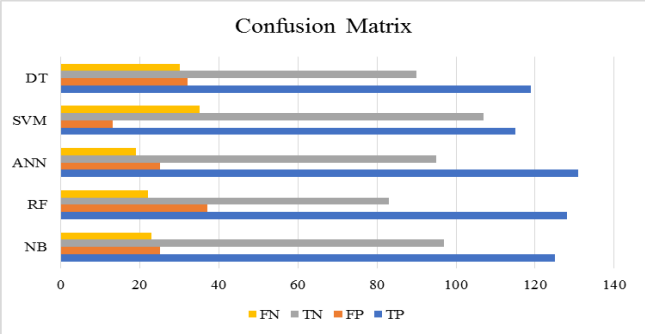

Figure 3. Classification results using confusion matrix

Here, ANN achieved better performance than the other classification techniques by obtaining 84% accuracy, whereas DT shows the worst performance by attaining 77% accuracy.

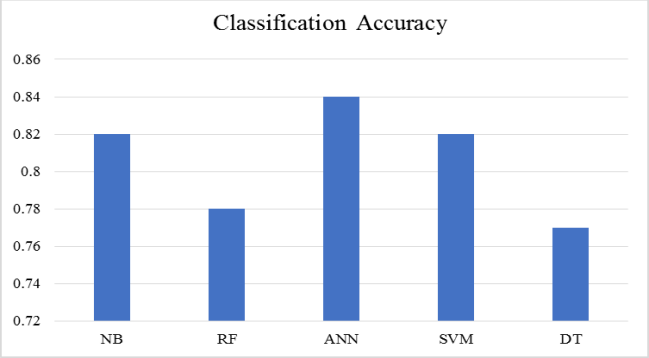

Figure 4. Prediction accuracy of classification techniques

C. *Performance Evaluation*

Predictions of all the classifiers are presented in figure 5 and Table 1, according to their accuracy, specificity, sensitivity, precision, and f- measure. The results clearly shows that the ANN achieved the highest accuracy and DT achieved the lowest accuracy than the remaining algorithms. On the other hand, SVM reached to the highest specificity and precision of 89% and 90%, respectively. Considering f1 measure and accuracy, ANN has the best performance than other classifiers, which is 84% and 86%, respectively. However, ANN indicates that this classification technique is more suitable than the other classifiers.

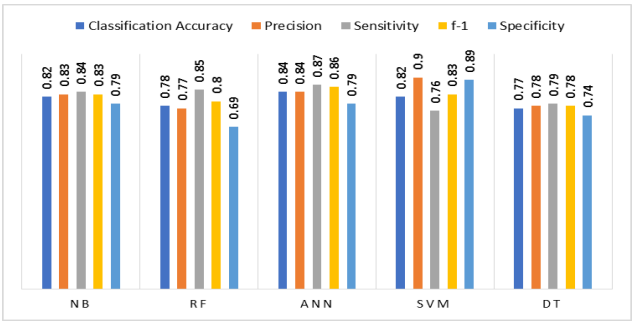

Figure 5. Classification performance of machine learning techniques.

TABLE I. CLASSIFICATION PERFORMANCE MEASUREMENT

Parameter/Algorithms	NB	RF	ANN	SVM	DT
Classification Accuracy	0.82	0.78	0.84	0.82	0.77
Precision	0.83	0.77	0.84	0.90	0.78
Sensitivity	0.84	0.85	0.87	0.76	0.79
F-1	0.83	0.80	0.86	0.83	0.78
Specificity	0.79	0.69	0.79	0.89	0.74

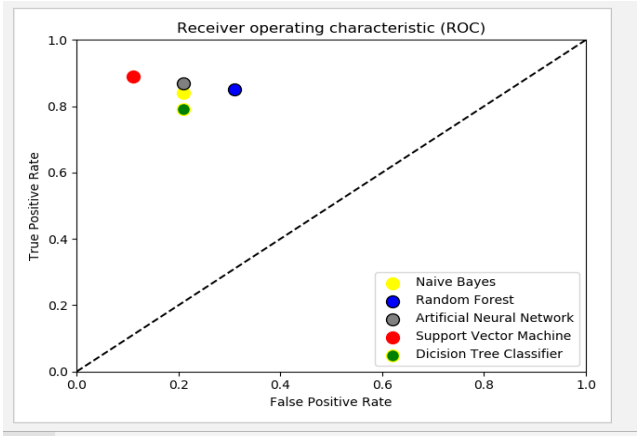

Figure 6. ROC for five machine learning classifier for prediction of heart disease

Most of the machine learning classification techniques exhibit the accuracy level above 80% which indicates that the performance of these algorithms are pretty good. Moreover, another important measure for classification is

receiver operating characteristics curve which is based on the true positive rate and false positive rate of the selected classification results. Figure 6 presents the ROC curve for the five machine learning classification algorithms, where SVM hits all other classification techniques (absence of heart disease). In addition, ANN is outperformed than the other classification algorithms in forecasting of presence of heart disease.

V. CONCLUSION

In this paper, we have described five supervised learning-based machine learning techniques. Afterwards, we compared the performance of the five classifiers which are used in the prediction of heart disease and evaluated their performance using confusion matrix and ten-fold cross-validation method. This study provides a workflow on machine learning based cloud application for the early detection and monitoring of heart disease. Therefore, we proposed a real-time-based four-tier cloud architecture for heart disease prediction and tracking record. This examination has used five machine learning techniques for the early detection of heart disease based on several parameters. In addition, this cloud-based application is able for early detection of heart disease by collecting real-time data from patient and health care center. We are currently developing a cloud-based real-time application for detecting heart disease and send the alert message from any dangerous situation to patients. Hence, this application can be used for operational heart disease prediction which will be helpful for the heart disease patient for tracking their health information.

REFERENCES

[1] "Coronary Heart Disease | National Heart, Lung, and Blood Institute (NHLBI)." [Online]. Available: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease. [Accessed: 18-Sep-2018].

[2] A. M. Islam, A. Mohibullah, and T. Paul, "Cardiovascular Disease in Bangladesh: A Review," Bangladesh Hear. J., vol. 31, no. 2, pp. 80, Apr. 2017.

[3] V. Hopwood, C. Donnellan, V. Hopwood, and C. Donnellan, "Current context: neurological rehabilitation and neurological physiotherapy," Acupunct. Neurol. Cond., pp. 39–51, Jan. 2010.

[4] "WHO | Cardiovascular diseases (CVDs)," *WHO*, 2018.

[5] K. Fatema, N. A. Zwar, A. H. Milton, L. Ali, and B. Rahman, "Prevalence of Risk Factors for Cardiovascular Diseases in Bangladesh: A Systematic Review and Meta-Analysis," *PLoS One*, vol. 11, no. 8, pp. e0160180, Aug. 2016.

[6] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," Procedia Comput. Sci., vol. 85, pp. 962–969, Jan. 2016.

[7] W. B. Kannel and D. L. McGee, "Diabetes and glucose tolerance as risk factors for cardiovascular disease: the Framingham study.," Diabetes Care, vol. 2, no. 2, pp. 120–6, Mar. 1979.

[8] M. R. Ahmed, M. Arifa Khatun, A. Ali, and K. Sundaraj, "A literature review on NoSQL database for big data processing," Int. J. Eng. Technol., vol. 7, no. 2, pp. 902–906, 2018.

[9] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," in *2017* International Conference on Intelligent Computing and Control (I2C2), 2017, pp. 1–8.

[10] A.K.Dwivedi,"Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Comput & Applic. Vol.29, Issue10, May 2018, pp. 685–693, doi: 10.1007/s00521-016-2604-1

[11] M. Kukar, I. Kononenko, C. Grošelj, K. Kralj, J. Fetich, "Analysing and improving the diagnosis of ischaemic heart disease with machine learning," intelligence in, Vol. 16, Issue 1, May 1999, pp 25-50, doi: 10.1016/S0933-3657(98)00063-3

[12] R. Das, I. Turkoglu, A Sengur, "Effective diagnosis of heart disease through neural networks ensembles," E. systems with applications, Vol. 36, Issue 4, May 2009, pp. 7675-7680, doi: 10.1016/j.eswa.2008.09.013

[13] X. Liu et al., "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method.," Comput. Math. Methods Med., vol. 2017, pp. 8272091, 2017, doi: 10.1155/2017/8272091

[14] "UCI Machine Learning Repository: Statlog (Heart) Data Set." [Online].Available:http://archive.ics.uci.edu/ml/datasets/statlog+(heart). [Accessed: 27-Sep-2018].

[15] "heart: Statlog (Heart) Data Set in VarSelLCM: Variable Selection for Model-Based Clustering of Mixed-Type Data Set with Missing Values."[Online].Available:https://rdrr.io/cran/VarSelLCM/man/heart.html. [Accessed: 23-Sep-2018].

[16] M. Caudill, "Neural networks primer, part I," AI expert., Volume 2 Issue 12, Dec. 1987, pp. 46 - 52

[17] R. H. Nielsen, "Theory of the backpropagation neural network," Neural networks for perception, 1992, Pages 65-93, doi: 10.1016/B978-0-12-741252-8.50010-8

[18] A. Y. Chervonenkis, "Early History of Support Vector Machines," in Empirical Inference, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–20, doi: 10.1007/978-3-642-41136-6_3

[19] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[20] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Stat. Comput., vol. 14, no. 3, pp. 199–222, Aug. 2004.

[21] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[22] F. Jensen, An introduction to Bayesian networks. 1996.

[23] K. M. Leung, "Naive bayesian classifier," Polytech. Univ. Dep. Comput. Sci. Risk Eng., 2007.