UACEE International Journal of Computer Science and its Applications - Volume 2: Issue 3 [ISSN 2250 - 3765]

Data Mining for Medical Systems: A Review

Muhamad Hariz Muhamad Adnan,Wahidah Husain, Nur'Aini Abdul Rashid
School of Computer Sciences
Universiti Sains Malaysia
11800 USM, Penang, Malaysia
mhma.com08@student.usm.my, wahidah@cs.usm.my, nuraini@cs.usm.my

***Abstract*—Data mining is a growing area of research that intersects with many disciplines such as Artificial Intelligence (AI), databases, statistics, visualization, and high-performance and parallel computing. The goal of data mining is to turn data that are facts, numbers, or text which can be processed by a computer into knowledge. Nowadays, the reliance of health care on data is increasing. Therefore, this paper aims to allow the readers to understand about data mining and its importance in medical systems.**

Keywords-data mining; medical system; decision tree; neural network; Bayesian classifiers, Support Vector Machine

I. INTRODUCTION

Data mining is the process of analyzing and summarizing data from different perspectives and converting it into useful information [1]. Data are facts, numbers, or text which can be processed by a computer [1]. In a large database, data mining is employed to detect patterns to extract hidden pieces of information [2].

The data mining tasks can be descriptive or predictive [3]. The data mining tasks are related to machine learning process such as data pre-processing, classification, regression, clustering, association rules, and visualization [4]. Data mining tasks can be classified as summarization, classification, clustering, association and trend analysis [1]. Summarization is the generalization or abstraction of a set of data resulting in a smaller set that gives a general overview of the data [1]. Classification derives the class of an object based on its attributes. It will help in better understanding of the object class in databases. Clustering refers to identifying groups or clusters of objects with unknown classes. By looking at the objects common features, they can be summarized to form class descriptions.

The Association is the discovery of togetherness or the connection of objects [5]. This connection of objects is termed the association rule – for example if an appearance of a set of objects in a database is strongly related to the appearance of another set of objects, the two sets are said to be associated. Trend analysis is finding the patterns and common attributes in data that change over time [1]. Data mining can uncover existing patterns and associations in the available data or derived prediction model from the data [3].

The data mining processes include formulating a hypothesis, collecting data, performing preprocessing, estimating the model, and interpreting the model and draw the conclusions (Figure 1).

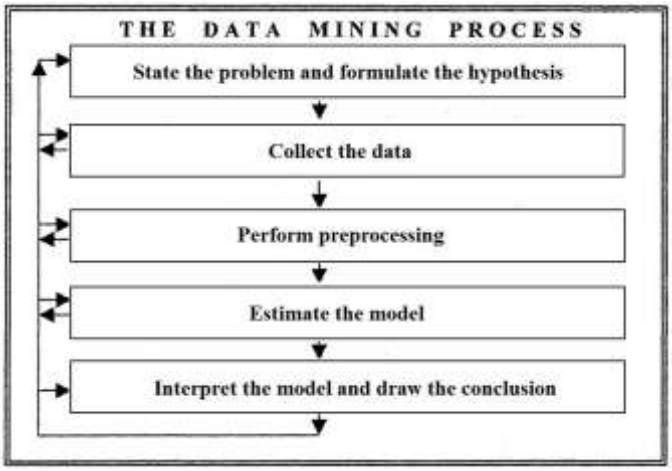

Figure 1. Data Mining Process [6].

The paper is organized as follows. In Section 2, several well known and widely used data mining techniques are presented. Section 3 briefly discusses the importance of data mining in medical systems. In section 4, several data mining classifications and predictions for problems related to medical domain are presented. Section 5 presents the summary on the study of the data mining techniques for medical systems.

II. DATA MINING TECHNIQUES

Well-known data mining techniques include the Artificial Neural Network (ANN), decision tree, Bayesian classifiers, Support Vector Machine (SVM) and many others. In this section, we introduce these four widely used data mining techniques.

A. Artificial Neural Network

ANN is based on the biological neural networks in the human brain (Figure 2) and described as a connectionist model [7].

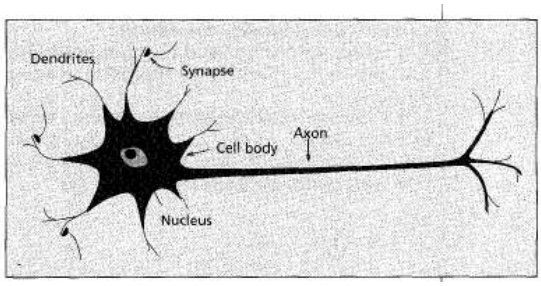

Figure 2. A Sketch of a Neuron in the Human Brain [8].

It is based on the neuron, a cell that processes information in the human brain [8]. The neuron cell body contains the nucleus, and has two types of branches, the axon and the dendrites. The axon transmits signals or impulses to other neurons while the dendrites receive incoming signals or impulses from other neurons. Every neuron is connected and communicates through the short trains of pulses (Figure 3) [8]. The nodes are the artificial neuron and the directed edges represented the connection between output neurons and the input neurons.

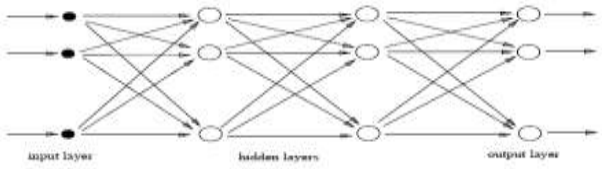

Figure 3. A Neural Network Model. [8].

B. Decision Tree

The decision tree is a powerful classification algorithm that is popular in the information systems [9]. Decision tree algorithms include Iterative Dichotomiser 3 (ID3), assistant algorithm, C4.5, C5, and CART [9, 10]. The decision tree is performed with separate recursive observation in branches to construct a tree for prediction. The splitting algorithms – i.e. Information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID) – are used to identify a variable and the corresponding threshold, and then split the input observation into two or more subgroups [9]. The steps are repeated until a complete tree is built as shown in Figure 4. The reason for the splitting algorithm is to find the variable-threshold pair which maximizes the order of the samples [9].

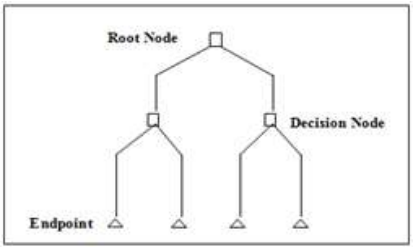

Figure 4. A Structure of a Decision Tree.

C. Bayesian Classifiers

The Bayesian classifiers have a structural model and a set of conditional probabilities [11]. The structural model (Figure 5) is represented as a directed graph where the nodes represent attributes and arcs represent attribute dependency.

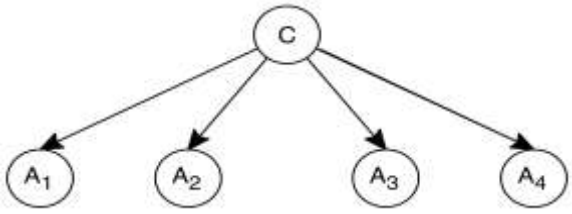

Figure 5. A Representation of a Bayesian Classifier Structure [11].

For classifications, the Bayesian networks are used to construct classifiers from a given set of training examples with class labels. The classifier of a general Bayesian network:

$$c(E) = \arg\max_{c \in C} P(c)P(a_1, a_2, \ldots, a_n|c). \quad (1)$$

The a_1 to a_n refers to the attributes or nodes in the Bayesian network [12]. The variable C represents the class variable that refers to the class node in a Bayesian network. The variable c represents the value of C and c(E) denotes the class of E. For Naïve Bayes (NB), all the attributes have to be assumed as independent. Therefore, the definition of Naïve Bayes is as follows:

$$c(E) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(a_i|c). \quad (2)$$

The attributes a are independent attributes [12]. For example, X is a child whose class is to be determined. Then, H is the class that child X going to be predicted. The formula to calculate P (H|X) using Bayesian networks is as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}. \quad (3)$$

P(H) is the probability of class H, P(X|H) is the posterior probability that X is conditioned on H, and P(X) is the probability of X [12]. These values can be obtained from the data used for training. For Naïve Bayes classifiers, the steps to predict probability for data X with the assumption that all the attributes are independent of each other is as follows:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k, C_i). \quad (4)$$

P(X1|Ci), P(X2|Ci),…,P(Xn|Ci) can be calculated from the training samples [12].

D. Support Vector Machine

The Support Vector Machine (SVM) is a classification algorithm in statistical learning theory [13]. It can provide accurate models because it can capture nonlinearity in the data. The classification tasks are performed by maximizing the margin separating both classes and minimizing the classification errors [13]. The training of SVM involves the optimization of a convex cost function where the learning process is not complicated by local minima [14]. The testing used the support vectors to classify a test dataset and the performance is based on error rate determination [14]. For a training set of l samples, the learning procedure are as the followings [12]:

$$\min_{\alpha} : \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j . \quad (5)$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, i = 1, \cdots, l . \quad (6)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 . \quad (7)$$

The y_i is the label of the i^{th} sample x_i [12]. The a_i is the Langrangian multiplier of x_i. The C is the upper bound of a_i and $K(x_i,x_j)$ is the kernel. The samples with a > 0 are called support vectors [12]. The decision function is as follow, where n_s is the number of support vectors [12] :

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n_s} y_i \alpha_i^* K(x_i, x) + b^*\right). \quad (8)$$

In this section, we have introduced four famous data mining techniques. In the subsequent section, the importance of data mining in medical systems are briefly discussed.

III. Application of Data Mining in medical systems

The reliance of health care on data is increasing [6]. Medical researchers, physicians, and health care providers face the problem to use stored data efficiently when more medical information systems with large database are used [6]. The medical information system databases contain many data such as patient records, physician diagnosis, and monitoring information where the data has been useful in many medical decision support systems to save lives [15].

A medical decision support systems are systems that help in the decision making process in the medical domains such Clinical Decision Support Systems (CDSS), medical imaging, and Bioinformatics [15]. The contributions of these systems are to reduce medical errors and costs, earlier disease detection, and to achieve preventive medicine [15]. The advantages of using computerized CDSS are the decision support systems can help to manage overloaded data and turn them into knowledge, reduce the complexity of the work such as automatic complex workflows, and help to identify obese children while reducing the errors, time, and variety of practices [15].

Continuous usage of the information systems result to the size of the database increasing. Therefore the usage of knowledge discovery and data mining in the database (KDD) for the growing databases is important. KDD attempts to gather knowledge by identifying relations from the data sets to help predictions [15]. KDD utilization is increasing in medical informatics and researchers have used it in many areas such as statistics, machine learning, intelligent databases, data visualization, pattern recognition, and high performance computing [16, 17].

The data mining has been used in the medical domain for other purpose like to improve the decision making such as diagnostic and prognostic problems in oncology, liver pathology, Neuropsychology, and Gynaecology [18]. For better data analysis and decision support, data mining and decision support can be integrated [19]. The task of detecting associations between risk factors and outcomes in the medical area is a difficult work even for experienced biomedical researcher or health care manager [6]. Data mining usage has helped clinicians to improve their health service by assisting in detecting regularities, trends, and unexpected events from the data [6].

The usage of data mining tools with advanced algorithms are popular for pattern discovery in biological data [3]. The biological problems include protein interactions, sequence and gene expression data analysis, drug discovery, discovering homologous sequences or structure, construction of phylogenetic trees, gene finding, gene mapping, and sequence alignment [3].

Machine learning was not fully accepted in the medical community because medical practitioners feel that their work is more complicated using such tools [20]. For an example, different models used in healthcare applications have a different explanation especially for model-specific methods [20]. Therefore, important things that must be considered when developing an application for medical practitioners are simplicity and the way of explaining the decisions. Simple techniques used for medical predictions have shown reasonable results [21]. Another challenge is that the systems must able to present discovered knowledge in an easy and fast manner [6].

The data that can be captured by a patient record are classified in three groups: a structured data, semi-structured data, and unstructured data (Figure 6) [22]. Data mining and knowledge discovery techniques and tools based on rule induction are important to analyze the growing size of clinical data.

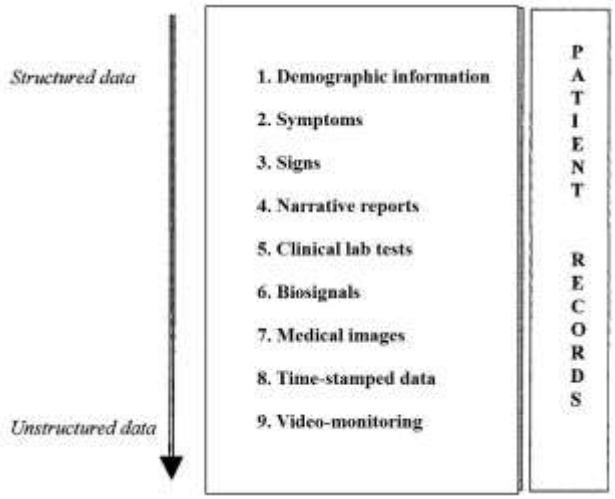

Figure 6. Data That Can be Captured From a Patient Record [6].

IV. Utilization of Data Mining for Predictions in Medical Domain

One task of data mining in medical domain is for classification and predictions. The classification derives the class of an object based on its attributes while prediction means an indication in advance based on observations, experiences, or scientific reasons [23]. This section discusses several classifications and predictions using data mining that was applied for medical related problems.

A. Coronary Heart Diseases

Coronary heart disease (CHD) is a serious disease that causes many deaths especially in China [13]. The study on CHD patients was aimed to identify the syndrome of CHD using data mining techniques [13]. The study used 1069 CHD cases that were collected using surveys on 5 clinical centres located in two provinces. 80 symptoms that are closely related to CHD and always appear in the literatures of CHD were selected.

Four well-known data mining techniques were tested for classifications: Bayesian network (BN), decision tree (C4.5), neural network mode (MLP), and SVM. The performances of these techniques were determined by their sensitivity, specificity, and accuracy. The sensitivity = TP / (TP + FN); the specificity = TN / (TN + FP); and the accuracy = (TP + TN) /(TP + FP + TN + FN) [12, 13, 18, 24]. TP is the number of samples classified as true while they were actually true; FN is the number of samples classified as false while they were actually false; TN is the number of samples classified as false while they were actually false; and FP is the number of samples classified as true while they were actually false.

The results showed that MLP accuracy is the highest (88.6%). The second highest accuracy is SVM (82.5%) and the third highest is Bayesian model (82.0%). Decision tree (C4.5) has the lowest accuracy among them (80.4%). From these results, data mining techniques have showed good accuracy on predicting the coronary heart diseases.

B. Breast Cancer

The study of breast cancer is interesting because the risk factors are difficult to identify similar to childhood obesity. The study on breast cancer recurrences involve 1035 breast cancer patient [20]. 22 medical patient features were recorded at the time of surgery while 10 more features were recorded through follow-up. The study took more than 10 years.

The data mining classifiers that were used are Naïve Bayes (NB), decision tree, a tuned SVM, Random Forests, multilayer ANN, and bagging with NB (bag). The recurrence and no-recurrence classes are almost fairly distributed. The results of these data mining classifiers were also compared with the expert oncologist predictions. The mean accuracy of oncologists prediction tested on 100 instances is 0.65 and 063. Meanwhile among the classifiers, NB and bag have shown the highest accuracy (0.7). The decision tree and Random Forests showed slightly lower accuracy (0.67 and 0.68 respectively).

The models were also tested on binary datasets. Bag and NB have the highest mean accuracy (0.68 and 0.678 respectively) follow by Random Forests (0.676), decision tree (0.674) and ANN (0.608). In this study, the negative and the positive groups are fairly distributed. The Naïve Bayes technique has shown good and consistent accuracy in this study.

C. Diabetes

Diabetes is a metabolic disorder where the body cannot make proper use of carbohydrate and greatly affected by the patient lifestyle [18, 24]. The study on diabetes prediction used 2017 diabetic patient clinical information [18]. There are 425 features in the database. The first step in the study was data pre-processing: data integration and reduction. The following step was feature selection using Relief to reduce the number of parameters.

Three classification algorithms were used to analyze the data. The algorithms are: decision tree C4.5, IB1, and Naïve Bayes. The features were discretized to improve the results of C4.5. The performance was determined based on the specificity and sensitivity. The result indicated that Naïve Bayes and IB1 accuracy improved when the input parameters were reduced.

Discretized C4.5 was the best in classifying bad blood glucose control patients (sensitivity) while Naïve Bayes was the best in classifying good blood glucose control patients (specificity). In terms of differences in both sensitivity and specificity, Naïve Bayes has the least differences. This study has shown that decision tree may produce high sensitivity. But in terms of fair distribution between sensitivity and specificity, the Naïve Bayes is better.

D. In-vitro Fertilization

The prediction was to identify the number of embryos to be transferred to the woman's womb and the selection of embryos with the highest reproductive viabilities [25]. The prediction was to determine whether the embryos are suitable for implant. The similarity of IVF implantation prediction with childhood obesity prediction is the imbalances distribution of positive and negative samples, which is common in medical datasets [25].

The study used 2453 records with 89% of them are positive samples [25]. There are 18 features that are related to the embryo and the patients. Six classification algorithms were used that are the Naïve Bayes, k-Nearest Neighbor, Decision Tree, SVM, Multilayer Perceptron (MLP), and Radial Basis Function (RBF). Receiver Operating Characteristics (ROC) analysis was used to evaluate the classifier performance. The results showed that the Naïve Bayes and RBF performance are significantly better than other classifiers. The Naïve Bayes sensitivity is 67%.

This study has used many features for prediction. The predictions are difficult because the positive group is very large compared to the negative group. The Naïve Bayes and RBF significantly higher accuracy than other classifiers has shown that these techniques can perform well on difficult predictions.

E. Childhood Obesity

Six data mining techniques and logistic regression were used for childhood obesity predictions [12]. The techniques are decision tree (C4.5), association rules, Neural Network, Naïve Bayes, Bayesian networks, linear SVM and RBF SVM. The prediction aims to identify obese and overweight children at 3 years old using the data recorded at birth, 6 weeks, 8 months, and 2 years. The prediction used 16653 instances, where only 20% of the samples are obese or overweight cases. The accuracy was measured using the sensitivity and specificity.

The first step for the prediction was data pre-processing that consists of data cleaning to discard abnormal instances and data discretization to change the continuous values into nominal values. The following step of the prediction was featured selection using the Cfs- Subset Evaluator and BestFirst Search method. The features were selected in terms of individual predictive ability, high correlation with overweight/obesity and low inter-correlation between features.

The selected attributes for training are sex, adjusted SDS birth weight, adjusted SDS length, time of gestation, adjusted SDS weight gain from birth to 6 weeks old, adjusted SDS weight gain from between 6 weeks to 8 months old, BMI at 8 months old, adjusted SDS height at 2 years old, and adjusted SDS BMI at 2 years old.

The results showed that for overweight predictions at 3 years old, the linear SVM and RBF SVM have the highest sensitivity (59.6% and 60% respectively). However, the specificity of both SVMs is the lowest compared to other techniques and logistic regression. Other good accuracy belongs to the Naïve Bayes and Bayesian Network (54.7%). The specificity of both Bayesian classifiers is very high (93.1%) hence they have the highest overall accuracy (91.9%).

For obesity prediction at 3 years old, the sensitivity of the decision tree is 0%; logistic regression is 11.2%; association rules is 21.9%; Neural Network is 24.6%; RBF SVM is 38%; and both Bayesian classifiers are 62%. In this study, the Bayesian classifiers and the SVMs have shown good accuracy for overweight prediction.

V. Summary

This paper presents a review of data mining importance in medical systems. Data mining is the process of analyzing and summarizing data from different perspectives and converting it into useful information. Well-known data mining techniques include the Artificial Neural Network (ANN), decision tree, Bayesian classifiers, Support Vector Machine (SVM). Data mining utilization is increasing in medical informatics and for improving the decision making such as diagnostic and prognostic problems in oncology, liver pathology, Neuropsychology, and Gynaecology. The challenge of data mining utilizations for medical practitioners are complexity and knowledge representation.

Acknowledgment

The authors would like to thank the Ministry of Higher Education (MOHE), Malaysia (Grant No. 203/PKOMP/6730002) and University Sains Malaysia for supporting this study.

References

[1] Q. Luo, "Advancing knowledge discovery and data mining," in *WKDD '08 Proceedings of the First International Workshop on Knowledge Discovery and Data Mining*, Washington, DC, USA, 2008.

[2] E. Papagergiou, *et al*., "Data mining: a new technique in medical research," *International Journal of Endocrinology and Metabolism*, pp. 189-191, 2005.

[3] H. Cheng, *et al.* (2010). *Data mining for protein secondary structure prediction*. *134*.

[4] I. H. Witten, *et al*., *Data mining: practical machine learning tools and techniques*, 3rd ed.: Morgan Kaufmann, 2011.

[5] Y. Fu. Data mining.potentials [Online].

[6] A. S. Elmaghraby, *et al.* (2006). *Data Mining from multimedia patient records*. *6*.

[7] B. Novak and M. Bigec, "Application of artificial neural networks for childhood obesity predicition," in *ANNES '95 Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems* 1995.

[8] A. K. Jain, *et al.* Artificial neural network : a tutorial [Online].

[9] Y. Xing, *et al.*, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," presented at the Proceedings of the 2007 International Conference on Convergence Information Technology, 2007.

[10] W. Peng, *et al.* An Implementation of ID3 --- decision tree learning algorithm [Online].

[11] L. Jiang, *et al*., "A novel bayes model: hidden naive bayes," *IEEE Trans. on Knowl. and Data Eng*., vol. 21, pp. 1361-1371, 2009.

[12] S. Zhang, *et al.*, "Comparing data mining methods with logistic regression in childhood obesity prediction," *Information Systems Frontiers*, vol. 11, p. 51, 2009.

[13] J. Chen, *et al.* (2007). *A comparison of four data mining models: bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease*. *4491/2007*.

[14] I. Maglogiannis, *et al.*, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers," *Applied intelligence*, vol. 30, 2007.

[15] W. Lord and D. Wiggins, "Medical Decision Support Systems Advances in Health care Technology Care Shaping the Future of Medical." vol. 6, G. Spekowius and T. Wendler, Eds., ed: Springer Netherlands, 2006, pp. 403-419.

[16] J. C. Prather, *et al.*, "Medical data mining: knowledge discovery in a clinical data warehouse," in *AMIA Annual Fall Symposium* 1997, pp. 101-105.

[17] J. Han and M. Kamber, *Data Mining, concepts and techniques*, 1st ed.: Academic Press, 2001.

[18] Yue Huang, *et al.*, "Evaluation of outcome prediction for a clinical diabetes database ", ed, 2004.

[19] A. Pur, *et al.*, "Data mining for decision support: an application in public health care," 2005.

[20] E. Strumbelj, *et al.*, "Explanation and reliability of prediction models: the case of breast cancer recurrence," *Knowl. Inf. Syst.*, vol. 24, pp. 305-324, 2010.

[21] D. Gregori, *et al.*, "Using Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better?," *Journal of Medical Systems*, 2011.

[22] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*: Wiley-IEEE Press, 2003.

[23] C. Park, "A dictionary of environment and conversation," in *A Dictionary of the Internet*, ed. Oxford University Press, 2007.

[24] Y. Huang, *et al.*, "Evaluation of Outcome Prediction for a Clinical Diabetes Database, Knowledge Exploration in Life Science Informatics." vol. 3303, J. López, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2004, pp. 181-190.

[25] A. Uyar, *et al.*, "ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction, Electronic Healthcare." vol. 27, P. Kostkova, Ed., ed: Springer Berlin Heidelberg, 2010, pp. 108-111.