

Performance Evaluation of the Tree Augmented Naïve Bayes Classifier for Knowledge Discovery in Healthcare Databases

Mário W. L. Moreira^{1,2}, Joel J. P. C. Rodrigues^{1,3,4}, Antonio M. B. Oliveira²,
Kashif Saleem⁵, Augusto J. Venâncio Neto⁶

¹Instituto de Telecomunicações – Universidade da Beira Interior (UBI)
R. Marquês D'Ávila e Bolama – 6201-001 – Covilhã – Portugal

²Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Aracati, CE – Brazil

³Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, MG – Brazil

⁴Universidade de Fortaleza (UNIFOR)
Fortaleza, CE – Brazil

⁵Center of Excellence in Information Assurance (CoEIA)
King Saud University (KSU) – Riyadh, Kingdom of Saudi Arabia

⁶Departamento de Informática e Matemática Aplicada
Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brazil

{mario.moreira, joeljrc}@ieee.org, amauroboliveira@gmail.com

ksaleem@ksu.edu.sa, augusto@dimap.ufrn.br

Abstract. *Smart decision support systems (DSSs) have been successfully employed in several areas. In healthcare, these systems offer solutions for uncertain reliably acts and moments. Systems based on Bayesian networks (BNs) can generate predictions even in information lack situations. This paper proposes the modeling and presents a performance evaluation study of the Bayesian classifier named Tree Augmented Naïve Bayes (TAN). Results show that the proposed algorithm obtained good performance for a pregnancy database, presenting F-measure 0.92, Kappa statistic 0.8932, and ROC area 0.993. The proposed method allows representing more complex connections between variables. Nevertheless, it requires major computational effort and time that are not needed in other Bayesian algorithms.*

1. Introduction

Smart decision support systems (DSSs) are major tools that help health experts in the prevention, diagnosis, diseases treatment, and patient accompaniment [Raffaeli et al. 2016, Manirabona et al. 2017]. It uses clinical information (sometimes, referred to as knowledge) that is incorporated into the system, helping health experts to analyze patient data as well as the decision-making process [Mushcab et al. 2017, Al-Khasawneh 2016]. These systems considers a knowledge base and an inference mechanism that uses the clinical data collected generating specific recommendations for each particular case. In

[Arisha and Rashwan 2016], the authors discuss different development phases of modeling approaches, presenting evidence that these methods can significantly improve decisions related to health care management at various levels. [Arshad et al. 2016] argue challenges to the existing healthcare industry. This study shows that despite the growing of extensive databases and valuable information no effective knowledge management gives experts substantial resources to an efficient decision-making process. The data classification provides decision makers with useful recommendations and guidelines based on the integrated knowledge and experience. In [Anand et al. 2016], the authors show how data mining (DM) can produce knowledge from a health data set. In this study, the performance evaluation and comparison between several DM techniques is presented. Results show that the use of these approaches in health systems provide useful identification and relevant information on patient health care. According to [Wang et al. 2008] problems involving DM are categorized into four groups, namely, statistical, data accuracy and standardization, technical, and organizational. Among them, there are the approaches based on the Bayes theorem. The main classifiers based on the well-known Bayes theorem in health are Naïve Bayes (NB) [Orphanou et al. 2016], TAN [Moore et al. 2014], and Averaged One-Dependence Estimators (AODE) [Kovács and Hajdu 2011]. The NB classifier assumes the nodes are conditionally independent, *i.e.*, the one event information is not informative about any other. The AODE classifier minimizes nodes dependency considering that, in real health problems, the attributes independence rarely occurs. Finally, the TAN classifier also relaxes the independence between nodes. However, the first presents a computational performance advantage in comparison to the second approach given that the AODE method lends itself directly to incremental learning.

For enhancing the biometrics systems performance, [Chiroma et al. 2014] compare the TAN classifier with the Markov Blanket (MB) approach in an iris dataset, using statistical indicators. Results show that the TAN classifier had a better performance in relation to MB. Regarding analyzing the time and computational complexity, both classifiers was found to be equal. This result is significant to evaluate the TAN model in other datasets. In [Wan and Freitas 2016], the authors propose a TAN algorithm variant, which considers removing the hierarchical redundancy during the classifier learning process. Results show that this approach obtained a better predictive performance than the conventional classifier. This study enhanced the robustness concerning imbalanced class distributions in aging-related gene datasets, using ontology terms as features. This research shows that TAN classifier has potential to be adjusted, thus improving its reliability. [Ren 2015] studies the breast cancer prediction problem in the aging population, using the TAN classifier in a clinical dataset. It is based on mammography examination features and demographic information to predict the malignancy probability, for the biopsy threshold setting and decision making. The author uses the 10-fold cross-validation technique and the Receiver Operating Characteristic (ROC) curves to evaluate the proposal, showing an interesting result. Considering the above mentioned details, the TAN classifier presented in this paper contains the potential for diseases identification inherent to risk pregnancy and can become a useful inference mechanism in health, helping clinicians in the decision-making process at uncertainty times in an efficient way for knowledge discovery.

The rest of the paper is organized as follows. Section II discusses the TAN classifier modeling and Section III shows the nodes relation proposal used to identify hyper-

tensive disorders as from symptoms presented by the patient. Section IV presents the performance evaluation, considering the comparison of Bayes-based methods, and the results analysis. Finally, Section V provides the conclusion and suggestions for further works.

2. TAN Classifier Modeling Proposal

The TAN classifier is a search algorithm that decides which node attributes depend on the class node. It uses a greedy search to add connections to the BN, initially assuming a set of blank child nodes for each node. At each step, it adds a new child node to another node, if that child node manages a better result of the used punctuation metric. The tree created by the TAN classifier is a structure that represents the dependencies between the attributes. This algorithm is similar to the NB, allowing, however, dependencies between the attributes. The constraints are that each attribute depends conditionally on the class (as it does on the NB) and that there are $n - 1$ attributes that conditionally depend on another attribute. This last condition implies that if there is a connection from X_i to X_j , these two attributes are not independent given the class. Instead, the influence of X_j on the probability of class C depends on the value of X_i . The dependencies structure is defined from the NB classifier. For TAN, it is also defined that all attributes depend on the class, but it is necessary to find the dependencies between the attributes following the described rules. To find the dependencies between the attributes this work uses the algorithm reported by [Huang et al. 2002]. Since each attribute can have, at most, one parent node, it is necessary to find the one with the greatest conditional probability given the value of C . In a simpler way, it is necessary to find the two attributes that have the highest correlation. When choosing for each attribute the one that has greater conditional dependence, it theoretically obtains the best dependencies representation. Equation (1) describes the TAN calculation, where the parents of the attribute X_i are represented by Π_{X_i} and the attribute has as parents the class and at most another attribute.

$$P_A(X_1, \dots, X_n) = P_A(X_i | \Pi_{X_i}) \quad (1)$$

Dependency tree construction differentiates the TAN from the NB. This tree determines the TAN classifier performance. Theoretically, it is due to the dependence between the attributes that TAN improves the performance about the NB. To construct the tree that maximizes the mutual information between the attributes, follow the steps considered by the Algorithm 1.

Algorithm 1 General pseudo-code for TAN classifier

- Step 1:** Obtain the mutual information $I_P(X_i, X_j | C)$ between each pair of attributes i, j and construct a vector with this mutual information;
 - Step 2:** Construct a complete non-oriented graph, having as nodes the attributes and as cost of the connections the mutual information between the attributes;
 - Step 3:** Calculate the tree that maximizes mutual information, without creating cycles;
 - Step 4:** Transform the non-oriented tree into orientated, choosing as root the higher mutual information;
 - Step 5:** Add the class as the parent of all attributes.
-

Equation (2) shows mutual information I_P calculation necessary to complete the first step above considering that all attributes are class dependent.

$$I_P(X; Y|C) = \sum_{x,y,c} P(x, y, c) \frac{P(x, y|c)}{P(x|c)P(y|c)} \quad (2)$$

Equation (2) assumes that both attributes are class-dependent, thus computing the information X provides on Y given the class value.

According to these steps, next topic presents the relationship construction between nodes, whose attributes represent the symptoms presented by pregnant women who suffered hypertensive disorders during pregnancy.

3. Classifier Graph Structure for Hypertensive Disorders Identification in Pregnancy

During pregnancy, hypertension is the elevation of blood pressure that occurs in women who have never presented the problem before and who, if untreated, can have serious consequences for both mother and fetus. Hypertension during pregnancy occurs in 5 to 8% of pregnant women, being the main cause of maternal death during pregnancy [Khan et al. 2006]. The most dangerous of these disorders are preeclampsia and eclampsia. Preeclampsia is the increase in blood pressure accompanied by the protein elimination in the urine. Eclampsia is one of the most severe state of preeclampsia presenting, in addition to high blood pressure, various complications and repeated seizures, which may end in a coma and, eventually, death. Women who were already hypertensive before gestation and who continue to be hypertensive do not are the same diagnostic group as those who only present hypertension specifically during pregnancy. Hypertensive women require special care when they become pregnant, some of them similar to those due to preeclampsia or eclampsia, but their hypertension usually does not assume the proportions and potential preeclampsia and eclampsia risks. In the other way, pregnant women who have had high blood pressure or hyperglycemia during pregnancy are more likely to have these disorders later [Magee et al. 2014]. Figure 1 shows the graph constructed from symptoms information collected from 25 pregnant women who presented with a clinical condition of hypertensive disorder during pregnancy. Obstetrician and gynecologist physicians have contributed to this research in the construction of this database by their experience. For the modeling of this, the steps described above in algorithm 1 were used.

The four main hypertensive disorders in gestation that this article seeks to identify are the pre-existing hypertension complicating pregnancy, childbirth and puerperium (CH), the pre-existing hypertensive disorder with superimposed proteinuria (PS), gestational hypertension (pregnancy-induced) without significant proteinuria (GH), and preeclampsia or eclampsia (PE). A risk factor that needs to be identified is the pregnant woman gestational age. This identification is pertinent because some hypertensive disorders often occur in a certain gestation period. This study separates this period in two categories, to know, before the 20th week and after the 20th gestational week. Regarding the symptoms presented by the pregnant woman during pregnancy the main ones are arterial hypertension and urine protein loss (proteinuria). Hypertension is considered normal

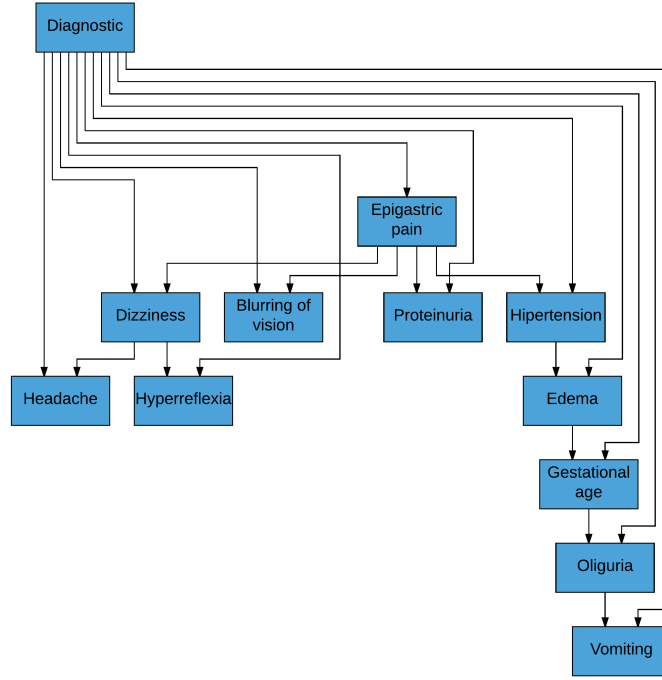


Figure 1. Graphical construction of the TAN classifier from a database of hypertensive disorders during risk pregnancy.

when both systolic and diastolic pressure is below $140mmHg$ and $90mmHg$, respectively. It is high when both the systolic and diastolic pressure is between $140 - 159mmHg$ and $90 - 119mmHg$, respectively. It is very high when systolic pressure is greater than or equal to $160mmHg$ and/or diastolic is greater than or equal to $110mmHg$. Regarding to proteinuria, the pregnant woman cannot present protein loss in the urine, present traces (more or less $300mg/24hrs$) and have severe loss (greater than $5g/24hrs$). This indicator is fundamental to differentiate PE from CH and GH.

It is important to detect early hypertensive states that constitute maternal and perinatal risk. Every pregnant woman with a hypertensive condition should be referred for high-risk prenatal care at the specialized service, since CH can progress to PE, when the hypertension is associated with proteinuria. PE is characterized by the elevated blood pressure presence levels during pregnancy. It usually occurs after the 20th week of gestation, by the gradual hypertension development, significant proteinuria and/or edema in hands or face. Differentiating a CH from PE is an important task of the prenatal care staff by the level of protein concentration in the urine.

4. Performance Evaluation and Result Analysis

For the proposed method performance evaluation, this paper uses the cross validation approach [Duda et al. 2012]. Under this method, data are divided into one part for training and other for testing. The k -fold cross validation method consists in dividing the total patterns set into k groups of approximately equal sizes. With this, training is performed k times, each time leaving one of the subgroups for testing. For each training, a classification error is calculated, and in the end, these errors average are calculated to achieve a final result. Then, from these results, a classification matrix is constructed, which com-

pares current values with predicted ones for each specified predicted state. The rows in the matrix represent the predicted values for the model and the columns representing the real values. The categories used in the analysis are false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN). For the FP Rate, the best values for this indicator are closer to zero. Using the same database as in [Moreira et al. 2016], the table 1 below shows the performance evaluation of the TAN classifier and compares it with other Bayesian methods. The used indicators are the precision, which is the classifier ability to recognize a determining given class, and reject the others. The recall is defined by the classifier's ability to recognize all instances of an interest class. The F-measure is a harmonic mean between precision and recall. The equations for calculating precision (equation (3)), recall (equation (4)), and F-measure (equation (5)) are shown below.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

Table 1. Standard metrics values of Bayesian predictive classifiers computed on confusion matrix predictive parameters using the 5-fold cross validation method.

Approach	TP Rate	FP Rate	Prec.	Rec.	F-measure	Class
AODE	0.667	0.000	1.000	0.667	0.800	CH
NB	0.667	0.000	1.000	0.667	0.800	
TAN	0.500	0.158	0.500	0.500	0.500	
AODE	0.167	0.263	0.167	0.167	0.167	PS
NB	0.333	0.211	0.333	0.333	0.333	
TAN	0.167	0.211	0.200	0.167	0.182	
AODE	0.500	0.211	0.429	0.500	0.462	GH
NB	0.667	0.211	0.500	0.667	0.571	
TAN	0.167	0.211	0.200	0.167	0.182	
AODE	0.714	0.167	0.625	0.714	0.667	PE
NB	0.714	0.111	0.714	0.714	0.714	
TAN	0.571	0.278	0.444	0.571	0.500	

From all the presented results, the TAN classifier obtained a good performance only in the false positive rate (FPR) for the PS and GH classes. This rate is also known as false alarm rate, usually referring to the probability that the model result indicates a disease presence when, in fact, it does not exist. A value closest to zero for this indicator indicates a better predictive model. This occurs because the TAN classifier is not efficient at predicting unknown cases. This was due to the fact that the data are all different from each other. Table 2 shows the results obtained for the classifier when the test set is the same training set, that is, it checks the predictive capacity of the classifiers when the case already exists in the database.

Table 2. Standard metrics values of Bayesian predictive classifiers computed on confusion matrix predictive parameters when the test set is the same that the training set.

Approach	TP Rate	FP Rate	Prec.	Rec.	F-measure	Class
AODE	0.833	0.000	1.000	0.833	0.909	CH
NB	0.833	0.000	1.000	0.833	0.909	
TAN	1.000	0.000	1.000	1.000	1.000	
AODE	0.833	0.053	0.833	0.833	0.833	PS
NB	0.333	0.053	0.667	0.333	0.444	
TAN	0.833	0.053	0.833	0.833	0.833	
AODE	1.000	0.053	0.857	1.000	0.923	GH
NB	1.000	0.158	0.667	1.000	0.800	
TAN	1.000	0.000	1.000	1.000	1.000	
AODE	0.857	0.056	0.857	0.857	0.857	PE
NB	0.857	0.111	0.750	0.857	0.800	
TAN	0.857	0.056	0.857	0.857	0.857	

Results presented in Table 2 show that TAN classifier presented excellent performance for all classes, having optimal performance for the CH and GH classes. Even for classes that have a difficult prediction, such as the PE and PS classes, the proposed classifier performed above the NB classifier performance and was equal to the AODE classifier. The NB classifier, which proved to be an excellent predictor for unknown cases, showed only a reasonable performance to predict cases already registered in the database. Table 3 shows two other important indicators that are the receiver operator characteristic (ROC) area and the Kappa statistic. The ROC curve represents the relationship between specificity (FPR) and recall (TPR) of a quantitative diagnostic test. Area under ROC curve is ideal when its value is closest to 1. Kappa statistics is an agreement measure used on nominal scales that gives an idea of how far observations depart from those expected, thus indicating how reliable the interpretations are. The values presented in Table 3 are an arithmetic mean of the values of Table 2 for each class.

Table 3. Average standard metrics values of Bayes-based classifiers.

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Kappa statistic
AODE	0.880	0.041	0.886	0.88	0.880	0.979	0.8397
NB	0.760	0.082	0.770	0.76	0.741	0.942	0.6788
TAN	0.920	0.028	0.920	0.920	0.920	0.993	0.8932

Kappa statistic results for the AODE and TAN classifiers are considered excellent, whereas for the NB classifier it is considered substantial. Figure 2 shows the ROC curve for the Bayesian classifiers using Table 2 results for PE class. Models that present curves closest to the point (0, 1) are considered optimal models. The TAN classifier shows the best results for this indicator. The AODE and NB classifiers show similar results, thus indicating that the three models have a good relation between FPR (less possible value) and TPR (highest possible value).

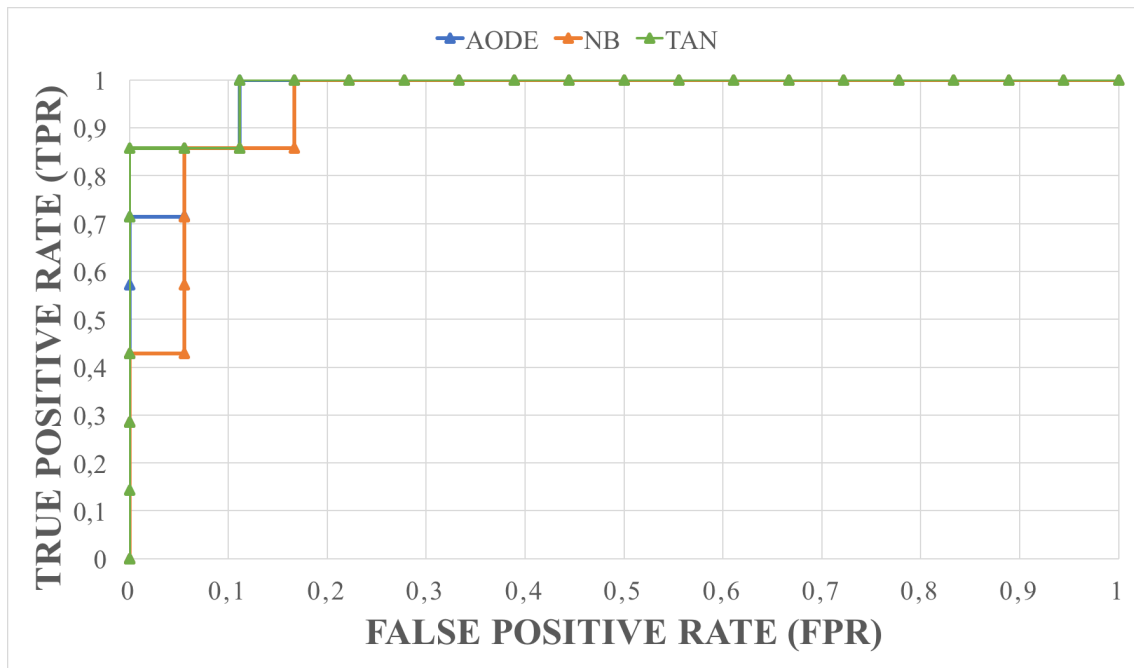


Figure 2. ROC curve to the all the Bayesian-based classifiers to the PE class.

5. Conclusion

This paper presented the modeling and performance comparison between AODE, NB, and TAN Bayesian classifiers. Using a database acquired from the medical experts' experience in risk pregnancy, the proposed models were evaluated using two methods, namely, the 5-fold cross-validation and a training set, *i.e.*, the classifier was tested on the same set in which it is trained. Although it did not perform well in the first validation method that identified the prediction accuracy for unknown data, the TAN classifier presented an excellent performance in the second validation method, which determined the prediction accuracy for data already registered in the database. An important result of this work was to show that the NB classifier that until then was a good classifier for recognizing cases not yet registered did not achieve the same performance for recognizing cases already known.

Further works suggest the inclusion of other variants for the classifiers presented, namely, NBTree, lazy Bayesian rules (LBR), super parent TAN (SP-TAN), backward sequential elimination and joining (BSEJ), backward sequential elimination (BSE), Bayesian classifier based wound characterization (BWC) [Chakraborty et al. 2016], and forward sequential selection (FSS). This work also proposes a study with a larger database to prove the TAN classifier efficiency in know data recognition. Other models based on artificial intelligence should also be experimented for the recognition of hypertensive disorders in pregnancy. Among them, tree-based models and those based on neural networks stand out. For the proposed model it is advisable to insert other risk factors and a set of nodes for the delivery outcome.

This study is an effort to develop a smart DSS that uses data-mining techniques to derive valuable insights from health databases. This system could contribute to the reduction of maternal mortality caused by the various types of hypertensive disorders

in the risk gestation. This task is part of the UN's sustainable development objectives. Achieving this goal will require an even greater effort on the part of all actors who make up those who care for high-risk pregnancies.

Acknowledgments

This work has been partially supported by the Instituto de Telecomunicações, Next Generation Networks and Applications Group (NetGNA), Portugal, by National Funding from the FCT - Fundação para a Ciência e a Tecnologia through the UID/EEA/50008/2013 Project, by Finep, with resources from Funttel, Grant No. 01.14.0231.00, under the Centro de Referência em Radiocomunicações - CRR project of the Instituto Nacional de Telecomunicações (Inatel), Brazil, and by Ciência sem Fronteiras of CNPq, Brazil, through the process number 207706/2014 – 0.

References

- Al-Khasawneh, A. (2016). Diagnosis of breast cancer using intelligent information systems techniques. *International Journal of E-Health and Medical Communications (IJEHMC)*, 7(1):65–75.
- Anand, T., Pal, R., and Dubey, S. K. (2016). Data mining in healthcare informatics: Techniques and applications. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, March 16-18, New Delhi, India, pages 4023–4029. IEEE.
- Arisha, A. and Rashwan, W. (2016). Modeling of healthcare systems: past, current and future trends. In *Proceedings of the 2016 Winter Simulation Conference, December 11-14, Washington, DC, USA*, pages 1523–1534. IEEE Press.
- Arshad, A., Noordin, M. F. B., and Othman, R. B. (2016). A comprehensive knowledge management process framework for healthcare information systems in healthcare industry of pakistan. In *2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, November 22-24, Jakarta, Indonesia, pages 30–35. IEEE.
- Chakraborty, C., Gupta, B., and Ghosh, S. K. (2016). Chronic wound characterization using bayesian classifier under telemedicine framework. *International Journal of E-Health and Medical Communications (IJEHMC)*, 7(1):76–93.
- Chiroma, H., Gital, A., Abubakar, A., and Zeki, A. (2014). Comparing performances of markov blanket and tree augmented naive-bayes on the iris dataset. In *International MultiConference of Engineers and Computer Scientists (IMECS)*, March 12-14, Hong Kong, China, pages 1–4.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
- Huang, K., King, I., and Lyu, M. R. (2002). Constructing a large node chow-liu tree based on frequent itemsets. In *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02)*, November 18-22, Singapore, volume 1, pages 498–502. IEEE.

- Khan, K. S., Wojdyla, D., Say, L., Gülmezoglu, A. M., and Van Look, P. F. (2006). Who analysis of causes of maternal death: a systematic review. *The lancet*, 367(9516):1066–1074.
- Kovács, G. and Hajdu, A. (2011). Extraction of vascular system in retina images using averaged one-dependence estimators and orientation estimation in hidden markov random fields. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, March 30 - April 2, Chicago, IL, USA*, pages 693–696. IEEE.
- Magee, L. A., Pels, A., Helewa, M., Rey, E., von Dadelszen, P., Committee, S. H. G., et al. (2014). Diagnosis, evaluation, and management of the hypertensive disorders of pregnancy: executive summary. *Journal of Obstetrics and Gynaecology Canada*, 36(7):575–576.
- Manirabona, A., Fourati, L. C., and Boudjit, S. (2017). Investigation on healthcare monitoring systems: Innovative services and applications. *International Journal of E-Health and Medical Communications (IJEHMC)*, 8(1):1–18.
- Moore, L., Kambhampati, C., and Cleland, J. G. (2014). Classification of a real live heart failure clinical dataset-is tan bayes better than other bayes? In *2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), October 5-8, San Diego, CA, USA*, pages 882–887. IEEE.
- Moreira, M. W., Rodrigues, J. J., Oliveira, A. M., Saleem, K., and Neto, A. V. (2016). An inference mechanism using bayes-based classifiers in pregnancy care. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), September 14-17, Munich, Germany*, pages 1–5. IEEE.
- Mushcab, H., Kernohan, W. G., Wallace, J., Harper, R., and Martin, S. (2017). Self-management of diabetes mellitus with remote monitoring: A retrospective review of 214 cases. *International Journal of E-Health and Medical Communications (IJEHMC)*, 8(1):52–61.
- Orphanou, K., Dagliati, A., Sacchi, L., Stassopoulou, A., Keravnou, E., and Bellazzi, R. (2016). Combining naive bayes classifiers with temporal association rules for coronary heart disease diagnosis. In *2016 IEEE International Conference on Healthcare Informatics (ICHI), October 4-7, Chicago, IL, USA*, pages 81–92. IEEE.
- Raffaeli, L., Spinsante, S., and Gambi, E. (2016). Integrated smart tv-based personal e-health system. *International Journal of E-Health and Medical Communications (IJEHMC)*, 7(1):48–64.
- Ren, P. (2015). A tree augmented naive bayesian network experiment for breast cancer prediction. *ArXiv e-prints*, pages 1–24.
- Wan, C. and Freitas, A. A. (2016). A new hierarchical redundancy eliminated tree augmented naive bayes classifier for coping with gene ontology-based features. In *International Conference on Machine Learning (ICML 2016), June 19-24, New York, NY, USA*, pages 1–5.
- Wang, J., Hu, X., and Zhu, D. (2008). Minimizing the minus sides of mining data. In *Data Mining and Knowledge Discovery Technologies*, pages 254–279. IGI Global.