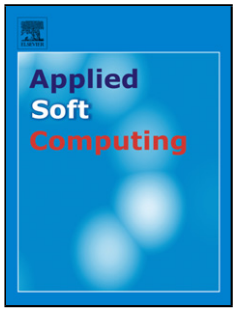


Title: An Application of Belief Merging for the Diagnosis of Oral Cancer

Author: Sameem Abdul Kareem Pilar Pozos-Parra Nic Wilson

PII: S1568-4946(17)30087-X
DOI: http://dx.doi.org/doi:10.1016/j.asoc.2017.01.055
Reference: ASOC 4058

To appear in: *Applied Soft Computing*

Received date: 21-11-2013
Revised date: 18-8-2016
Accepted date: 30-1-2017

Please cite this article as: Sameem Abdul Kareem, Pilar Pozos-Parra, Nic Wilson, An Application of Belief Merging for the Diagnosis of Oral Cancer, *<![CDATA[Applied Soft Computing Journal]]>* (2017), http://dx.doi.org/10.1016/j.asoc.2017.01.055

An Application of Belief Merging for the Diagnosis of Oral Cancer

Sameem Abdul Kareem[a,*], Pilar Pozos-Parra[b,**], Nic Wilson[c]

[a] *Dept. of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia*
[b] *IRIT - Université de Toulouse, France*
[c] *Insight Centre for Data Analytics, School of Computer Science and IT, University College Cork, Ireland*

Abstract

Machine learning employs a variety of statistical, probabilistic, fuzzy and optimization techniques that allow computers to "learn" from examples and to detect hard-to-discern patterns from large, noisy or complex datasets. This capability is well-suited to medical applications, and machine learning techniques have been frequently used in cancer diagnosis and prognosis. In general, machine learning techniques usually work in two phases: training and testing. Some parameters, with regards to the underlying machine learning technique, must be tuned in the training phase in order to best "learn" from the dataset. On the other hand, belief merging operators integrate inconsistent information, which may come from different sources, into a unique consistent belief set (base). Implementations of merging operators do not require tuning any parameters apart from the number of sources and the number of topics to be merged. This research introduces a new manner to "learn" from past examples using a non parametrised technique: belief merging. The proposed method has been used for oral cancer diagnosis using a real-world medical dataset. The results allow us to affirm the possibility of training (merging) a dataset without having to tune the parameters. The best results give an accuracy of greater than 75%.

*Corresponding author
**Principal corresponding author
Email addresses: sameem@um.edu.my (Sameem Abdul Kareem), maripozos@gmail.com (Pilar Pozos-Parra), nic.wilson@insight-centre.org (Nic Wilson)

Preprint submitted to Applied Soft Computing *March 13, 2017*

Keywords: Artificial intelligence, knowledge modelling, decision support systems, belief merging

1. Introduction

Oral or mouth cancer is a malignancy that occurs in any part of the mouth, namely, the lips, on the tongue's surface, inside the cheek, the gums, in the roof and floor of the mouth, in the tonsils, and also the salivary glands. Oral Squamous Cell Carcinoma (OSCC) results from a combination of risk habit factors such as tobacco use, betel-quid chewing, alcohol consumption and genetic damage that leads to DNA alterations in key cellular genes. Genetic instability can result in an uncontrolled cellular growth and researchers have shown that genetic polymorphisms are able to affect the risk of a wide range of cancers [1, 2, 3, 4, 5]. Most oral cancer cases occur when the patient is at least 40 years old. It affects more men than women [6].

In Peninsular Malaysia, oral cancer incidence as reported in 2006, was divided into Tongue cancer, Salivary Gland cancer and Mouth cancer [7]. Both Tongue and Salivary Gland cancers are higher among men as compared to women. However, mouth cancer is higher among women (56.5%) than men (43.5%) with women of Indian ethnicity forming 71% of the total incidence among Malaysian women. As in other parts of the world, the incidence of oral cancer in Malaysia increases with age, with more oral squamous cell carcinoma occurring in individuals over 40 years of age [8].

A coordinated and standardized data collection of oral cancer cases from multiple centers involving eight hospitals in Malaysia was initiated by the Malaysian Oral Cancer Research and Coordinating Center (OCRCC), University of Malaya, Malaysia [9]. The data collected includes parameters on sociodemographic, clinical, pathological, quality of life measures, details of treatment methods, vital status and dietary intake. The establishment of this database, named as, the Malaysian Oral Cancer Database and Tumour Bank System, aims in encouraging and supporting researches on oral cancer.

Machine learning, a sub-branch of artificial intelligence, has been in existence for more than 50 years with a wide variety of learning tasks and successful applications [10]. Machine learning techniques have been applied to medicine for diagnosis, prognosis, bio-medical analysis, image analysis and drug development [11, 12]. Machine learning techniques invariably involve parameter tuning with regards to the underlying technique, such as, the number of neurons, layers or epochs in a neural network technique; membership

function selection in fuzzy logic; population size, selection strategy, mutation rate, crossover rate in genetic algorithms as well as in the hybrid techniques that use fuzzy logic or neural network or both.

On the other hand, belief merging looks at strategies for combining symbolic information, expressed in propositional logic, coming from different sources. Every source is coded as a set of propositional formulae and known as a belief base, where the group of belief bases in conjunction may be inconsistent; the strategies aim at obtaining a consistent belief base representing the group. Logic-based belief merging has been studied extensively in the literature [13, 14, 15, 16, 17]. In particular [17] explains in detail the works on belief merging of propositional bases; the authors have presented an overview of logic based merging with relevant strategies known as merging operators. They have also mentioned the scarcity of belief merging applications and the expectation on the applications of these merging techniques.

As discussed in [17], a well known strategy involves the use of an operator Δ which takes as input the belief bases (profile) E and outputs a new consistent merged belief base $\Delta(E)$. In each case, the belief bases are described using a finite number of propositional symbols; there is no hierarchy, priority, or any difference in reliability of the sources assumed. While the belief merging framework has theoretical strength, most of the approaches lack implementation. Experimental assessments of algorithms for merging operators use random belief bases as a way of evaluating performance [18, 19]. Another operator that has been implemented is $PS\text{-}Merge$ [20, 21], which uses toy examples from the literature as an experimental evaluation. There is, generally, no accepted method for evaluating belief merging algorithms, nor a library of standard belief merging problems. Thus, a good strategy for testing the merging operator is to evaluate their results when they are used to implement algorithms for some real world problems, such as, the diagnosis of a particular medical condition using real world medical datasets.

In this paper, we discuss an application of belief merging on a set of data for which a prior research involving oral cancer diagnosis has been carried out using machine learning techniques [22]. We implement a new method for such diagnosis based on the $PS\text{-}Merge$ operator implementation. Our aim is to give a real world application to belief merging, for which the technique is sadly lacking. Any reference made to machine learning is merely to put the research in perspective and not to act as a comparison on the classification superiority of either technique. As far as we know, this is the first attempt to use real world data in order to test the implementation of belief merging

operators.

The rest of the paper is organized as follows. After providing some technical preliminaries and discussing the motivating works done using the same data in Section 2, the research methodology is given in Section 3 while the results and conclusion are provided in Sections 4 and 5, respectively.

2. Preliminaries and Previous Work

2.1. Preliminaries

We consider a language $\mathcal{L}$ of propositional logic using a finite ordered set of symbols $P := \{p_1, p_2 \ldots, p_n\}$ where the formulae are in Disjunctive Normal Form (DNF) or Conjunctive Normal Form (CNF). A formula ϕ is in DNF iff ϕ is a disjunction of terms $\phi = D_1 \vee \cdots \vee D_m$, where each term D_i is a conjunction of literals $D_i = l_1 \wedge \cdots \wedge l_{m'}$, with $l_k = p_j$ or $l_k = \neg p_j$. A formula ϕ is in CNF iff $\phi = C_1 \wedge \cdots \wedge C_m$, where each clause C_i is a disjunction of literals $C_i = l_1 \vee \cdots \vee l_{m'}$, with $l_k = p_j$ or $l_k = \neg p_j$.

A *belief base* K is a finite set of propositional formulae of $\mathcal{L}$ representing the beliefs from a source (we identify K with the conjunction of its elements).

The set of models of the language is denoted by $\mathcal{W}$; its elements will be denoted by vectors of the form $(w(p_1), \ldots, w(p_n))$ and the set of models of a formula ϕ is denoted by $mod(\phi)$. K is consistent iff there exists a model of K.

A belief *profile* $E = \{K_1, \ldots, K_m\}$ is a multiset (bag) of m belief bases.

2.2. Previous work

Commonly used machine learning techniques are artificial neural networks, fuzzy logic, support vector machines and genetic algorithms. In the works of medical diagnosis, machine learning has been used in the diagnosis of breast cancer, sexually transmitted diseases, sepsis, oral cancer, leukaemia, etc. [23, 24, 22, 25]. In medical prognosis, machine learning has been used in naso-pharyngeal carcinoma, colo-rectal cancer, cardiac diseases, gestational trophoblastic tumours, septicaemia, etc. [12, 26, 27].

A retrospective cohort study was conducted using a sample of 171 data obtained from the Malaysian Oral Cancer Database and Tumour Bank System (MOCDTBS) courtesy of the OCRCC. The full dataset were split randomly into a modeling dataset (67% of the total) and testing dataset (the remaining 33%). Note that this is the same data that is used in the current research on belief merging and will be discussed in detail in Section 3. The

four machine learning classifier methods used were, namely, fuzzy neural networks, fuzzy logic, logistic regression and fuzzy regression. The demographic profiles, risk habits and clinical variables as well as the oral cancer gene marker expression of GSTM1 and GSTT1 of cancer patients were reported to be associated risk factors to oral cancer. Peripheral blood was obtained from consenting individuals and the GSTM1 and GSTT1 genotypes were determined using Polymerase Chain Reaction (PCR) and restriction enzyme digestion at the Cancer Research Initiatives Foundation, Malaysia (http://www.hati.my/health/cancer-research-initiatives-foundation-carif/).

The risk factors of the oral cancer patients and the demographic profile and other associated parameters of the control group were used as input variables in developing the machine learning prediction models, with the outcome being the health condition of "cancer" or "healthy" [22]. Bootstrapping and cross-validation techniques were applied to the sample used in order to minimize the effects of model over-fitting and overcome the problems associated with small sample size. The data was prepared by converting them to binary values; each possible combination of the input variables was used in the research and the variables were subsequently reduced from the initial total of 8 based on the wrapper method of variable selection. In the wrapper method, a subset of variables was selected and a classifier was run on the training data; the model accuracy on a test set was used to evaluate the performance of the variable subset. As mentioned earlier, the demographic and clinical variables of patients were used as the predictor variables in developing the machine learning prediction models. A complete description about wrapper feature selection procedures can be found in [28].

The results obtained were in the range of 0.456–0.828 for the fuzzy neural network, 0.472–0.766 for fuzzy logic, 0.452–0.833 for the logistic regression and 0.455–0.824 for fuzzy regression for various numbers of input variables as shown in Table 1 [29, 30, 31, 22]. The best results were obtained using 4 input variables, namely, age, alcohol drinking, betel quid chewing and smoking, with the exception of the fuzzy logic model which obtained the best results with only 3 input variables of age, alcohol drinking and betel quid chewing.

Table 1: Summary of Results Obtained from Machine Learning (AUC) by Dom et al

Fuzzy Neural Network	Fuzzy Logic	Logistic Regression	Fuzzy Regression
0.456 - 0.828	0.472 - 0.766	0.452 - 0.833	0.455 - 0.824

3. Materials and Methods

3.1. Materials

Table 2 lists all the binary variables used as input for the prediction models together with a brief description of the variables.

Table 2: Input Variables and their Descriptions

Factors	Descriptions
Age	Less than 40 or Greater
Gender	Male or Female
Ethnicity	Aborigines or Non-Aborigines
Cigarette Smoking	Smoker or Non-Smoker
Alcohol Drinking	Drinker or Non-Drinker
Betel Quid Chewing	Chew or Do Not Chew
GSTM1	Positive or Negative Gene Mapped to Chromosome 1p13.3
GSTT1	Positive or Negative Gene Mapped to Chromosome 22q11.2

In order to investigate the possibility of applying belief merging in the diagnosis of oral cancer, we use the data set discussed in the last section.

The data was earlier obtained from the Oral Cancer Database and Tumour Bank System (MOCDTBS) provided by the Oral Cancer Research and Coordinating Center (OCRCC), University of Malaya, Malaysia. Demographic profiles (age, gender) and oral cancer risk habits (cigarette smoking, alcohol drinking, tobacco and betel-quid chewing) of the cancer patients and control group were used as input variables as listed in Table 2. The dichotomous output or outcome refers to the health state of either cancer (1) or healthy (0).

3.2. Diagnosis based on PS-Merge

The diagnosis proposal, as in classical machine learning, is divided into the training and testing phases; the training (merging) phase is based on the following operator.

Definition 1. *Let $E = \{K_1, \ldots, K_m\}$ be a belief profile and let PS-Merge be a function which maps a belief profile to a belief base, PS-Merge: $\mathcal{L}^n \to \mathcal{L}$ with the following property: the set of models of the resulting base, (i.e., the Partial Satisfiability Merge PS-Merge(E) of E) is:*

$$\left\{ w \in \mathcal{W} \middle| \sum_{i=1}^{m} w_{ps}(K_i) \geq \sum_{i=1}^{m} w'_{ps}(K_i) \text{ for all } w' \in \mathcal{W} \right\},$$

where w_{ps} is defined formally as follows:

Definition 2 (Normal Partial Satisfiability). *For $K \in \mathcal{L}$, $w \in \mathcal{W}$, the Normal Partial Satisfiability of K for w, denoted as $w_{ps}(K)$, is defined as follows:*

- *If $K \in P$, then $w_{ps}(K) = w(K)$;*
- *if $K = \neg p$, then $w_{ps}(K) = 1 - w_{ps}(p)$;*
- *if $K = D_1 \vee \cdots \vee D_n$, then $w_{ps}(K) = max\{w_{ps}(D_1), \ldots, w_{ps}(D_n)\}$ and*
- *if $K = C_1 \wedge \cdots \wedge C_n$, then $w_{ps}(K) = \sum_{i=1}^{n} \frac{w_{ps}(C_i)}{n}$.*

The testing phase is also divided into two subphases. The first subphase uses the merged base provided by the training phase; however, given that there is a large proportion of cases with imprecise results, i.e., unknown or ambiguous diagnoses, then we introduce a second subphase of testing in order to obtain a diagnosis when a non-answer is given (unknown diagnosis) and to reduce the number of ambiguous diagnoses when the given answer is cancer or healthy. For the second subphase we also need an extended version of the operator, $PS\text{-}Merge_{\mu}$, considering a set of belief constraints μ, where the result must satisfy the constraints as follows:

$$\left\{ w \in mod(\mu) \middle| \sum_{i=1}^{m} w_{ps}(K_i) \geq \sum_{i=1}^{m} w'_{ps}(K_i) \text{ for all } w' \in mod(\mu) \right\}.$$

The method considers every case in the medical database as a source of information; thus, every case (belief base) K is of the form $l_1 \wedge \cdots \wedge l_8 \to l_9$ with the first variables or factors represented as follows: l_1 represents age, l_2 represents gender, l_3 represents ethnicity, l_4 represents cigarette smoking,

l_5 represents alcohol drinking, l_6 represents betel quid chewing, l_7 represents GSTM1 and l_8 represents GSTT1, while, the last variable, l_9, represents the current diagnosis. For example, suppose that the medical database contains case 13 which is a diagnosis of oral cancer for a smoker, non-aboriginal male over forty years of age, who does not drink and does not chew betel quid and is positive for chromosome GSTM1 but negative for chromosome GSTT1. Then, we have $K_{13} = \neg p_1 \wedge p_2 \wedge \neg p_3 \wedge p_4 \wedge \neg p_5 \wedge \neg p_6 \wedge p_7 \wedge \neg p_8 \rightarrow p_9$; given that the input and output data to the proposed algorithm is in a binary format, we use an equivalent notation with 1's and 0's, thus writing $K_{13} = (0, 1, 0, 1, 0, 0, 1, 0, 1)$.

Based on the following equivalence: $p \rightarrow q \equiv \neg p \vee q$, a DNF or CNF equivalent to K can be easily found negating every factor and connecting all the elements with disjunctions, $\neg l_1 \vee \cdots \vee \neg l_8 \vee l_9$. For the previous example, we have $K'_{13} = p_1 \vee \neg p_2 \vee p_3 \vee \neg p_4 \vee p_5 \vee p_6 \vee \neg p_7 \vee p_8 \vee p_9$. After the transformation of the 171 cases into their normal forms we obtain the medical dataset in normal form E, then we apply Algorithm 1 to E, where the profile holds two third ($\frac{2}{3}$) of the 171 belief bases in normal forms. Then the algorithm obtains the diagnosis D for one third ($\frac{1}{3}$) of the cases. The diagnosis has the format: $(target, outcome)$ where $target$ is the current diagnosis stored in the medical dataset and $outcome$ is the diagnosis obtained by the algorithm.

The description of Algorithm 1 is as follows: we split the belief bases into two sets, the "training" set and the testing set, using the following method: the first two cases were assigned to the training set then the third was assigned to the testing set, the next two cases were then assigned to the training and the following one sent to the testing set; the process was repeated until the complete set of 171 cases has been split into two sets, E_{train} with 114 cases and the E_{test} with 57 cases. Then we merged the 114 cases using the operator $K_{Merged} = PS\text{-}Merge(E_{train}) = PS\text{-}Merge(K'_1, \ldots, K'_{114})$.

Initially the method was run in a single testing phase, using the training phase (result of the merging) as follows: for every one of the 57 testing cases we had to verify if there existed models in K_{Merged} such that the first variables (factors) had the same value as the case in question; if yes, we had to verify the last variable value, it could be: 0, 1 or both; the diagnosis 0, 1 or B respectively would then be assigned to the outcome result.

Let us show in a simplified example the outputs of Algorithm 1. Consider a scenario where there are four cases with only two diagnosis factors, namely Gender and Cigarette Smoking.

Algorithm 1: Diagnosis based on $PS\text{-}Merge$

Data:
E : Medical dataset in normal form
Result:
D : diagnosis for one third of the cases in (target, outcome) format

begin
  $[E_{train}, E_{test}] \leftarrow split_{[\frac{2}{3},\frac{1}{3}]}(E)$;
  $K_{Merged} \leftarrow PS\text{-}Merge(E_{train})$;
  % *Testing using the merged belief base*;
  for $s \leftarrow 1 \ldots |E_{test}|$ **do**
    $D(1,s) \leftarrow Last\_Variable(E_{test}(s))$;
    $D(2,s) \leftarrow B$;
    for $b \leftarrow 1 \ldots |K_{Merged}|$ **do**
      if $factors(E_{test}(s)) = factors(K_{Merged}(b))$ **then**
        if $D(2,s) = B$ **then**
          $D(2,s) \leftarrow Last\_Variable(K_{Merged}(b))$
        end
        else
          $D(2,s) \leftarrow B$;
        end
      end
    end
  end
end

- First case: male smoker is diagnosed positive for oral cancer.
- Second case: non-smoker male diagnosed positive for oral cancer.
- Third case: female non-smoker diagnosed positive for oral cancer.
- Last case: female smoker diagnosed negative for oral cancer.

Then if we represent male by the Boolean value of 1, female by 0, both in the first position; smoking by 1, non-smoking by 0, both in the second position; and oral cancer by 1, non-oral cancer by 0 in the last position. Then, we have, $K_1 = (1, 1, 1)$, $K_2 = (1, 0, 1)$, $K_3 = (0, 0, 1)$ and $K_4 = (0, 1, 0)$. Which in DNF is as follows: the first case is represented by $K'_1 = \neg p_1 \vee \neg p_2 \vee p_3$, the second case is represented by $K'_2 = \neg p_1 \vee p_2 \vee p_3$, the third case by $K'_3 = p_1 \vee p_2 \vee p_3$ and the last case by $K'_4 = p_1 \vee \neg p_2 \vee \neg p_3$. Applying Algorithm 1, we have $E_{train} = \{K'_1, K'_2, K'_4\}$ and $E_{test} = \{K'_3\}$. $K_{Merged} = \{(0 \wedge 0 \wedge 0) \vee (0 \wedge 0 \wedge 1) \vee (0 \wedge 1 \wedge 0) \vee (1 \wedge 0 \wedge 1) \vee (1 \wedge 1 \wedge 1)\}$ and $D = (1, B)$ which means that Algorithm 1 could not find an answer in the testing part. It is worth noticing that there are only four possible combinations between the two diagnosis factors, and in any other case Algorithm 1 can provide a diagnosis. This testing case could be solved by Algorithm 2 which provides D = (1,1), which means that our approach found a true positive in the testing part.

Coming back to our data set using a single testing phase based on the merge phase of the algorithm, we obtained 20 ambiguous results, for 8 factors, i.e., cases in which the diagnosis was both, 0, and 1, see line 14 in Algorithm 1; also the algorithm could not obtain a diagnosis for 11 cases, for 8 factors, i.e., cases in which the algorithm obtained a non-answer, see line 7 in Algorithm 1. In both cases the outcome was 'B' (imprecise). This means, the algorithm was not able to determine whether the health condition was cancer or healthy. Since this defeats the whole purpose of the exercise we subsequently "forced" the algorithm to form an opinion and tried to turn the imprecise diagnosis to precise ones.

This was done by carrying out the testing into a second phase in order to reduce the ambiguous cases, marked with B, i.e., the cases where the diagnosis was both cancer or healthy, this second phase was used in order to resolve the cases that resulted in non-diagnosis (non-answer) marked with B too.

Algorithm 2: Diagnosis of imprecise and non-answers outcomes using $PS\text{-}Merge_{\mu}$

```
% Testing the undefined outcomes B using a merging operator under
  constraints;
for s ← 1 ... |E_test| do
    if D(2, s) = B then
        Constraints ← Factors(E_test(s));
        K_const ← PS-Merge_μ(K_Merged, Constraints);
        if |K_const| = 1 then
            D(2, s) ← Last_Variable(K_const);
        end
    end
end
```

In the second phase the algorithm reduced the number of B appearing in the result of the first phase, see Algorithm 2. Therefore, for every case with a diagnosis of B, the factors of the case in question needed to conform to the set of constraints imposed for the second merge, i.e., the results must satisfy the factors of the case in question and then forced to have at least one answer. Also, the cases resulting in non-diagnosis were not allowed in this second phase.

As mentioned earlier, this exercise was an attempt to investigate a real world application for belief merging using data as described in Section 3.1. The model's accuracy was measured in terms of Sensitivity and Specificity and also the Area Under the receiver operator Characteristics (AUC) as is commonly done in medical statistics and machine learning exercises. The Receiver-Operating Characteristic (ROC) is a plot of sensitivity versus specificity for different test results. A person with the disease who has a *positive* test result is termed a True Positive (TP), whereas a person with the disease but a *negative* test is termed a False Negative (FN). On the other hand, a person without the disease who has a *positive* result is termed a False Positive (FP), while person without the disease but a *negative* test is termed a True Negative (TN) [32]. This is summarized in Table 3.

There are some derivations from a confusion matrix such as Sensitivity and Specificity, which are defined as in Table 4.

Table 3: Confusion Matrix

Outcome \ Target	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Table 4: Some derivations from a confusion matrix

Sensitivity (Sen) or True Positive Rate (TPR) $Sen = (TP)/(TP + FN)$
Specificity (Spe) or True Negative Rate (TNR) $Spe = TN/(FP + TN)$
Positive Prediction Value (PPV) $PPV = TP/(TP + FP)$
Negative Prediction Value (NPV) $NPV = TN/(TN + FN)$
Accuracy (Acc) $Acc = (TP + TN)/(TP + FN + FP + TN)$
False Positive Rate (FPR) $FPR = FP/(FP + TN) = 1 - Spe$
Harmonic mean of precision and sensitivity or F_1 Score (F_1) $F_1 = 2TP/(2TP + FP + FN)$

4. Results and Discussion

In this section we discuss the results of Algorithm 1. In Section 4.1, we present the results using a second method in which the imprecise cases were assigned randomly. In Section 4.2 we discuss the results of the approach presented in the previous section where the imprecise cases were handled by $PS\text{-}Merge_{\mu}$.

4.1. Handling Imprecise Results

In this section we discuss the imprecise diagnosis approach without the second phase. As previously mentioned, for each case, the algorithm either returns a prediction of 1, or of 0, or B, where B represents "ambiguous", or "unknown" diagnosis. Table 5 lists these results, where CBD stands for Cannot Be Determined. For example, the True Positives in Table 5 are defined to be cases where the method predicted 1 and this turned out to be correct. FP is the number of cases where the method predicted 1, but the

correct result was 0. The results indicate that, among cases where there is a prediction, the method can be very accurate, at least for the 8, 7, and 6 factor methods. For example, for the 8 factors case, a prediction is given in a little less than half the cases (26 out of the 57), with 92% accuracy. Also, the sensitivity (i.e., the true positive rate) is 91%, and with a very low (7%) false positive rate.

It is not immediately obvious how one should compare this imprecise method (which does not give a prediction for all cases) with a precise method that always gives a prediction. A simple approach is to consider a random algorithm, parameterised by a probability value $p \in [0, 1]$: for each unknown case, where the imprecise method gives value B indicating no prediction, the value of B is changed to a prediction of 1 with probability p, and otherwise to a prediction of 0. We can then compute the expected prediction of such a random algorithm for each case, leading to the expected number of true positives, true negatives etc., and the associated statistics. For example, for 8 factors, and with $p = 0.5$, there were 17 of the 57 cases with no prediction (represented by B) which turned out to actually be 1, and 10 cases which were correctly predicted as 1; the TP score will then be $10 + 0.5 \times 17 = 18.5$. Results for the $p = 0.5$ case (where there is an equal chance that unknown value B is changed to 1 or 0), are shown in Table 6, where the figures for TP, FP, FN and TN are rounded to the nearest whole number, and the other figures calculated from these. It turns out the results of the revision based merging, shown in Table 8, correspond to the case where $p = 1$, where all unknown predictions B are changed to a prediction of 1.

An alternative choice for p is to use the proportion of *number of predictions of 1* to *number of predictions of either 1 or 0*. For example, with 8 factors the method predicted 11 ones, and 15 zeros, giving a value of p of $\frac{11}{26} \approx 0.42$. We call this method the *Proportional p* method; the results are shown in Table 7, again rounding to the nearest whole number as in Table 6.

The accuracy scores are fairly similar in Tables 8, 6 and 7. The $p = 1$ method (Table 8) has very high values of Sensitivity (True Positive Rate) but the $p = 0.5$ and *Proportional p* methods have better values of Specificity (True Negative Rate). On this dataset, the slightly more sophisticated *Proportional p* method does not do better than the simpler $p = 0.5$ method.

4.2. Results using PS-Merge$_\mu$

As can be seen from Table 8, the preliminary results obtained using Belief Merging under various input factors were comparable to those obtained in

Table 5: Belief Merging Results for Imprecise Algorithm

Factors	TP	FP	FN	TN	Sen	Acc	Spe	PPV	NPV	F1
8	10	1	1	14	91	92	93	91	93	0.91
7	15	2	0	17	100	94	89	88	100	0.94
6	18	2	0	11	100	94	85	90	100	0.95
5	8	2	0	0	100	80	0	80	CBD	0.89
4	7	1	0	0	100	88	0	88	CBD	0.93

Table 6: Belief Merging Results for Imprecise Algorithm: $p = 0.5$

Factors	TP	FP	FN	TN	Sen	Acc	Spe	PPV	NPV	F1
8	19	8	9	21	68	70	72	70	70	0.69
7	21	7	7	22	75	75	76	75	76	0.75
6	23	10	5	19	82	74	66	70	79	0.75
5	18	15	10	14	64	56	48	55	58	0.59
4	17	15	11	14	61	54	48	53	56	0.57

the machine learning environment. The best accuracy of (78.95%) and the greatest AUC of (0.793) were obtained when 7 input factors were used. These factors were, namely; age, alcohol drinking, betel quid chewing, smoking, GSTT, gender and ethnicity. Recall that, the AUC is the area under the Receiver Operating Characteristics (ROC), while the ROC is a plot of the Sensitivity versus Specificity as shown in Figure 1. In our research we use several evaluation criteria such as the TP, TN, FP, FN, accuracy as discussed above. We have included the AUC in order to compare it with Dom et al's work as this was an evaluation criterion used by Dom [22].

With this work being an initial exploration of using belief merging in a medical diagnosis scenario, we have not used all permutations of the input variables nor reduced the variables based on the variable selection methods carried out by [22]. We have merely repeated a subset of the set of all permutations of the input factors that produced the best results in the works by [29, 30, 31, 22].

Notice that only with 8 factors (see Table 8), we obtained a value of FN different from zero, i.e., FN=1. A shallow analysis would lead to the conclusion that the lack of preprocessing introduced noise to the merge (training) phase. When the input of 8 factors was used, the medical dataset included two instances of the form $p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4 \wedge \neg p_5 \wedge \neg p_6 \wedge \neg p_7 \wedge \neg p_8 \rightarrow \neg p_9$

Table 7: Belief Merging Results for Imprecise Algorithm: *Proportional p*

Factors	TP	FP	FN	TN	Sen	Acc	Spe	PPV	NPV	F1
8	17	7	11	22	61	68	76	71	67	0.65
7	21	7	7	22	75	75	76	75	76	0.75
6	24	12	4	17	86	72	59	67	81	0.75
5	28	29	0	0	100	49	0	49	CBD	0.66
4	28	29	0	0	100	49	0	49	CBD	0.66

Table 8: Belief Merging Results for Different Factors

Factors	TP	FP	FN	TN	Sen	Acc	Spe	PPV	NPV	F1	AUC
8	27	15	1	14	96.43	71.93	48.28	64.29	93.33	0.771	0.723
7	28	12	0	17	100	78.95	58.62	70	100	0.824	0.793
6	28	18	0	11	100	68.42	37.93	60.87	100	0.757	0.69
5	28	29	0	0	100	49.12	0	49.12	CBD	0.659	0.5
4	28	29	0	0	100	49.12	0	49.12	CBD	0.659	0.5

and one instance of the form $p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4 \wedge \neg p_5 \wedge \neg p_6 \wedge \neg p_7 \wedge \neg p_8 \rightarrow p_9$, i.e., the dataset held a contradiction when the factors were of the form $p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4 \wedge \neg p_5 \wedge \neg p_6 \wedge \neg p_7 \wedge \neg p_8$, having two instances with the diagnosis of non-cancer and one instance with a cancer diagnosis.

The two instances holding non-cancer diagnosis were in the training base and the instance holding cancer diagnosis was in the testing base; the system thus learned from the training data that when the input factors were $p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4 \wedge \neg p_5 \wedge \neg p_6 \wedge \neg p_7 \wedge \neg p_8$, the output should be 0 (because of the two instances in the training base), and so, in the testing phase, the system responded with 0 for these factors.

We cannot blame the system for the incorrect answer, given that the system learned correctly from the training instances. Analysing the TN case, we found that all adequate diagnoses of non-cancer were made after learning, from the training phase, instances (rules) that did not hold contradictions, i.e., the system learned correctly rules that solely held non-cancer diagnoses. It is worth noticing that for 4 and 5 factors, all the instances concerning the diagnosis of non-cancer were contradictory, i.e., there existed rules with the same factors holding cancer diagnosis; for this reason we obtained in both cases TN=0. Another characteristic of the system is that when it faces contradictory instances, the outcome of the diagnosis is cancer; this

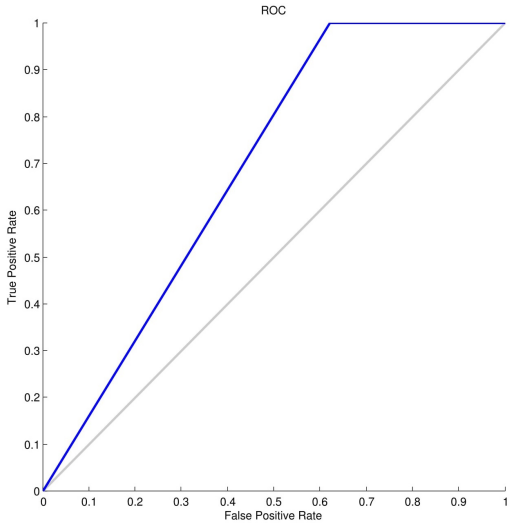


Figure 1: ROC for seven factors

explains the high number of FP and TP. Bearing this in mind, it would appear that a satisfactory explanation of the behaviour of belief merging can been attempted, thus overcoming the black box effects associated with certain machine learning techniques such as the artificial neural networks.

Table 9: Best Belief Merging Results

Factors	TP	FP	FN	TN	Sen	Acc	Spe	PPV	NPV	F1	AUC
7	28	12	0	17	100	78.95	58.62	70	100	0.824	0.793

5. Conclusion and future work

Belief merging is a deterministic method and our implementation for oral cancer diagnosis preserves the deterministic property of belief merging. In this research we have been successful in showing that belief merging can be used for classification or prediction (in this case for oral cancer) and is not merely an academic exercise to be considered with hypothetical examples only. Belief merging has the capability to diagnose oral cancer with an accuracy of (49% – 79%) and an AUC (0.5 – 0.79), which is comparable to what was achieved using machine learning techniques based on the same set of data, as discussed in Section 2. The result from the machine learning

exercise, as conducted by Dom et al, is given in Table 1, and that the AUC obtained using for logistic regression for the diagnosis for oral cancer in this exercise was in the range of 0.45 - 0.83. Hence, the performance of belief merging may be slightly below this, at 0.79 but it is only a difference of 0.04 in terms of AUC.

As we mentioned earlier, our aim is not to compete with machine learning as this is a relatively early experiment for belief merging using real world data. Machine learning techniques, on the other hand, have been using real world data for numerous medical diagnosis and prognosis instances together with various techniques of data pre-processing and variable selection well in place. In this research, we have not considered data pre-processing at all and have only carried out a rudimentary form of variable selection, roughly based on what was carried out by the machine learning researchers. Bearing this in mind, the scope for research in the application of belief merging (learning) in medicine is promising and there is yet much work that can be explored in this field. The proposal uses $PS\text{-}Merge$ twice; first the merge was used in order to learn from the cases and then a second merge subject to constraints was carried out in order to eliminate the unknown cases and reduce the ambiguous cases without having to randomly assign the results. In Section 4 we present the results using a random algorithm that assigns diagnosis values to unknown and ambiguous cases in a proportional way and the comparable results given by $PS\text{-}Merge_{\mu}$. Finally, given that the FN result is rather dangerous to the life of patients as they may have the potential to delay the detection of the cancer and the subsequent treatment, we consider our proposed method may be acceptable as a medical tool, given that the method resulted in a small number of FN, actually the only case where the FN was not 0, obtained using 8 factors, i.e., $FN = 1$, was the case in which the training set holds two contradictory diagnoses to that given by the system. So we can conclude that the method did not make mistakes concerning FN.

In addition, the best results in the belief merging experiments are obtained with 7 input factors with an accuracy of 79%, AUC of 0.72, no FNs as shown in Table 9. The only drawback with this result is the FP is rather high, meaning that if this algorithm is used in a clinical scenario, it would cause a lot of distress to patients, as they would be deemed to have cancer when they do not actually have it. It is important, therefore, to improve the performance of the algorithm further.

Given that, in theory, belief merging and revision are methods which are not restricted by the number of variables, this proposal can be used by any set

of data which is represented by binary vectors. However, the implementation may be limited by the available resources (hardware and software). The belief merging method for oral cancer diagnosis presented was easily implemented with a response time of 65 seconds, using a PC operating a Windows 7 and Matlab 9.

In future, we would like to try to create an open source implementation that could be used by researchers for "training" and "testing" their data sources. In order to attain this goal, we would need to determine the maximum size of the input binary data set allowed. That is, we would need to analyse how many factors and how many cases would be allowed using similar resources to mentioned earlier. We would also like to adapt this implementation of oral cancer, in order to carry out the diagnosis for other types of cancer, such as breast cancer, depending on the availability of datasets.

Acknowledgements

We would like to thank Weiru Liu and Queen's University Belfast for hosting this work. We would also like to extend our appreciation to Rosma Mohd Dom for the extensive use of her previous work in this research. This research was supported by the University of Malaya Sabbatical Leave in the case of the first author and the Program Postdoctoral and Sabbatical Stays Abroad for the Consolidation of Research Groups - CONACyT and UJAT in the case of the second author.

[1] S. N. Drummond, R. S. Gomez, M. J. C. Noronha, I. A. Pordeus, A. A. Barbosa, L. De Marco, Association between gstt1 gene deletion and the susceptibility to oral squamous cell carcinoma in cigarette-smoking subjects, Oral Oncology 41 (2004) 515–519.

[2] K. Tanimoto, S. Hayashi, K. Yoshiga, T. Ichikawa, Polymorphisms of cyp1a1 and gstm1 gene involved in oral squamous cell carcinoma in association with cigarette dose, Oral Oncology 35 (1999) 191–196.

[3] M. Sato, T. Sato, T. Izumo, T. Amagasa, Genetically high susceptibility to oral squamous cell carcinoma in terms of combined genotyping of cyp1a1 and gstm1 genes, Oral Oncology 36 (2000) 267–271.

[4] T. Sreelekha, K. Ramadas, M. Pandey, G. Thomas, K. Nalinakumari, M. Pillai, Genetic polymorphism of cyp1a1, gstm1 and gstt1 genes in indian oral cancer, Oral Oncology 37 (2001) 593–598.

[5] A. Reichart, Identification of risk groups for oral precancer and cancer and preventive measures, Clin Oral Invest 5 (2001) 207–213.

[6] C. Nordqvist, What is mouth cancer? what causes mouth cancer? what is oral cancer?, Medical News Today (September 2014).
URL http://www.medicalnewstoday.com/articles/165331.php

[7] K. Kesihatan, Malaysian cancer statistics - data and figure - peninsular malaysia 2006.
URL http://www.moh.gov.my/images/gallery/Report/Cancer/MalaysiaCancerStatistics_2006.pdf

[8] NCPR, National cancer patient registry (ncpr), clinical research centre (April 2013).
URL http://www.acrm.org.my/ncpr/

[9] R. Zain, R. Latifah, I. Razak, S. Ismail, A. Samsuddin, S. Atiya, Oral cancer and precancer research in malaysia - the database and tissue resource bank, Oral Oncology 1 (2005) 123–164.

[10] E. Coiera, Artificial intelligence in medicine: an introduction, guide to medical informatics, the internet and telemedicine (July 2013).
URL http://www.openclinical.org/aiinmedicine.html

[11] J. G. Carbonell, R. S. Michalski, T. M. Mitchell, Machine learning: A historical and methodological analysis, The AI Magazine 4 (3) (1983) 69–79.

[12] P. J. F. Lucas, A. Abu-Hanna, Prognostic methods in medicine, Artificial Intelligence in Medicine 15 (1999) 105–119.

[13] P. Revesz, On the Semantics of Arbitration, Journal of Algebra and Computation 7 (2) (1997) 133–160.

[14] P. Liberatore, M. Schaerf, Arbitration (or how to merge knowledge bases), IEEE Transactions on Knowledge and Data Engineering 10 (1) (1998) 76–90.

[15] J. Lin, A. Mendelzon, Knowledge base merging by majority, in: R. Pareschi, B. Fronhoefer (Eds.), Dynamic Worlds: From the Frame Problem to Knowledge Management, Kluwer Academic, 1999, pp. 195–218.

[16] S. Konieczny, R. Pino-Pérez, Merging information under constraints: a logical framework, Journal of Logic and Computation 12 (5) (2002) 773–808.

[17] S. Konieczny, R. Pino-Pérez, Logic based merging, Journal of Philosophical Logic 40 (2) (2011) 239–270.

[18] J. Hue, O. Papini, E. Wurbel, Syntactic propositional belief bases fusion with removed sets, in: Symbolic and Quantitative Approaches to Reasoning and Uncertainty, 2007, pp. 66–77.

[19] N. Gorogiannis, A. Hunter, Implementing semantic merging operators using binary decision diagrams, Int. J. Approx. Reasoning 49 (1) (2008) 234–251.

[20] V. Borja-Macías, P. Pozos-Parra, Implementing ps-merge operator, in: A. H. Aguirre, R. M. Borja, C. A. R. Garca (Eds.), MICAI, Vol. 5845 of Lecture Notes in Computer Science, Springer, 2009, pp. 39–50.

[21] P. Pozos-Parra, L. Perrussel, J. M. Thévenin, Belief merging using normal forms, in: I. Z. Batyrshin, G. Sidorov (Eds.), MICAI (1), Vol. 7094 of Lecture Notes in Computer Science, Springer, 2011, pp. 40–51.

[22] R. M. Dom, A fuzzy regression model for the prediction of oral cancer susceptibility:phd thesis, Ph.D. thesis, Faculty of Computer Science, University of Malaya (2009).

[23] P. Szolovits, R. S. Patil, W. B. Schwartz, Artificial intelligence in medical diagnosis, Annals Of Internal Medicine 108 (1998) 80–87.

[24] J. Nagi, S. Abdul-Kareem, F. Nagi, S. Ahmed, Automated breast profile segmentation for roi detection using digital mammograms, in: Proceeding of Biomedical Engineering and Sciences, IEEE Conference Publications, 2010, pp. 87–92.

[25] H. Madhloom, S. Abdul-Kareem, H. Ariffin, An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images, Journal of Medical Systems 36 (2012) 2149 –2158.

[26] H. Hamdan, S. Abdul-Kareem, N. Mohd-Taib, C. Yip, Back propagation neural network for the prognosis of breast cancer: Comparison on different training algorithms, in: Proceedings of the Second International Conference on Artificial Intelligence in Engineering and Technology, 2004, pp. 445–449.

[27] O. F. Baker, S. Abdul-Kareem, Assessment of the use of soft computing models into survival analysis data (nasopharyngeal carcinoma cancer), The International Medical Journal 15 (2005) 105–119.

[28] G. Papakostas, D. Koulouriotis, A. Polydoros, V. Tourassis, Evolutionary feature subset selection for pattern recognition applications, in: E. Kita (Ed.), Evolutionary Algorithms, InTech, 2011, pp. 443–458.

[29] R. Dom, S. Abdul-Kareem, Variable selection and data preprocessing for neural network models, in: Proceedings of The Seventh International Conference on Information Integration and Web Based Applications and Services (iiWAS2005), Austrian Computer Society, Kuala Lumpur, Malaysia, 2005, pp. 915–922.

[30] R. Dom, S. Abdul-Kareem, I. Razak, B. Abidin, A learning system prediction method using fuzzy regression, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, Newswood Limited, Hong Kong, 2006, pp. 904–912.

[31] R. Dom, S. Abdul-Kareem, B. Abidin, R. Raja Jallaludin, C. Sok, R. Zain, Oral cancer prediction model for malaysian sample, Austral-Asian Journal of Cancer 7 (2008) 545–559.

[32] ROCOnline, Introduction to roc curves (May 2013).
URL `http://gim.unmc.edu/dxtests/ROC1.htm`

Brief Biography of Sameem Abdul Kareem

Sameem Abdul Kareem received the B.Sc. in Mathematics (Hons) from the University of Malaya, in 1986, the M.Sc. in Computing from the University of Wales, Cardiff, UK, in 1992, and the Ph.D. in computer science from University of Malaya, Malaysia, in 2002. She is currently a Professor of the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University Of Malaya. Her research interests include medical informatics, machine learning, data mining and intelligent techniques, ontologies and image processing. She has published over 100 peer reviewed articles in journals and conference proceedings.

Brief Biography of Pilar Pozos-Parra

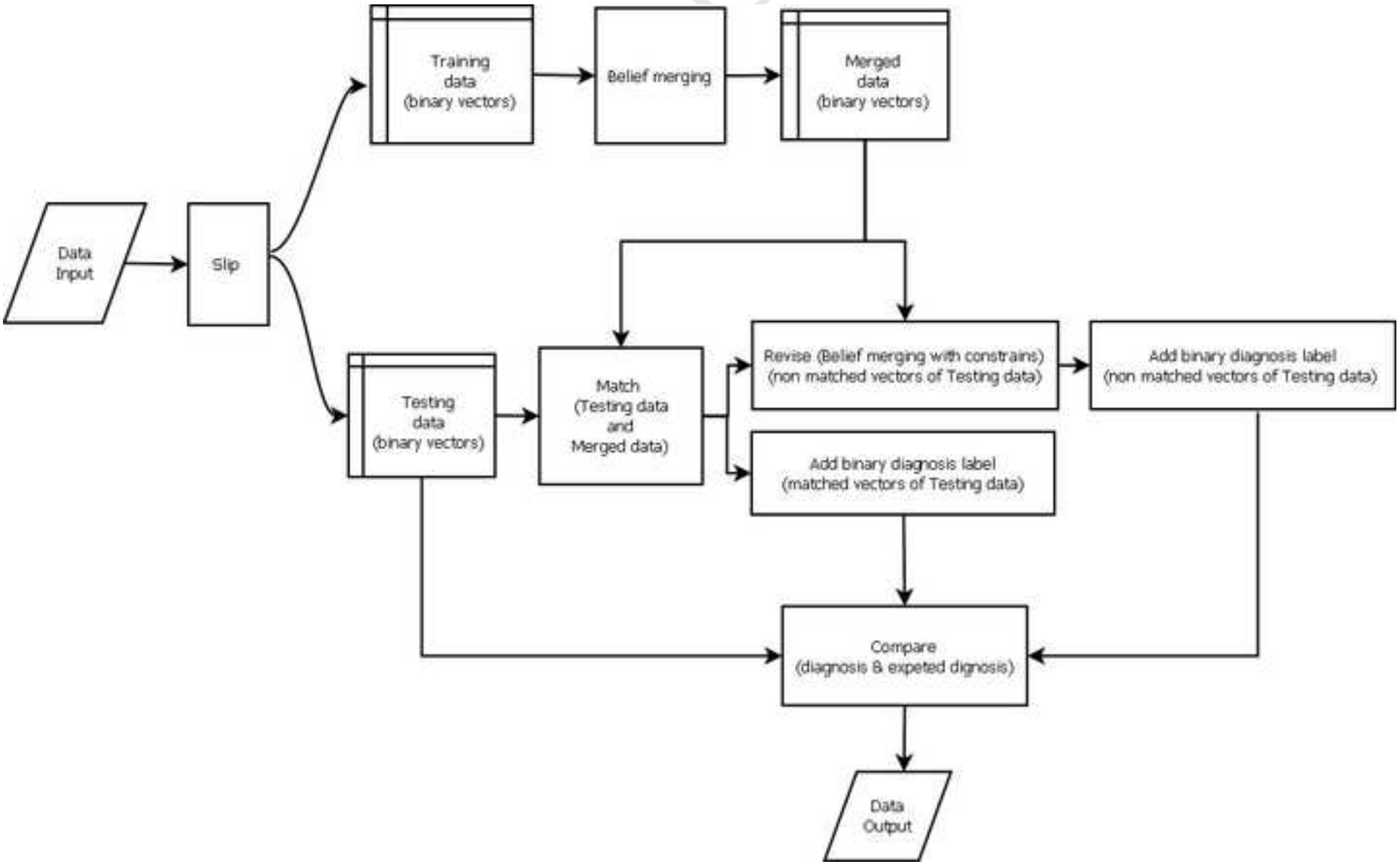
Pilar Pozos-Parra received her B.S. (1996) in Computer Sciences from the University of Puebla, Mexico, her M.S. (1998) and Ph.D. (2002) both in Computer Sciences (Knowledge Representation and Formalization of Reasoning) from SUPAERO College, France. She is currently a Post-Doctoral Research Fellow at the School of Information Technology (Faculté d'informatique),

Université Toulouse 1 Capitole, France. Her research interests include classical and modal logics, situation calculus, automated deduction strategies, belief merging and revision and Artificial Intelligence approaches to nutritional meal planning.

Brief Biography of Nic Wilson

Nic Wilson is a senior research fellow at Insight Centre for Data Analytics, University College Cork, Ireland, and during the preparation of this paper was also a professor at Queen's University Belfast. Before joining Cork in 2003 he was a senior research fellow at Keele University and Oxford Brookes University, UK. His research interests are in areas related to reasoning with preferences, uncertainty and constraints. He has published widely, with around eighty peer-reviewed conference and journal publications.

***Graphical abstract (for review)**



Highlights:

- Using a non parameterized technique, namely, belief merging in order to learn from past examples.
- Using belief merging operators to integrate inconsistent information, which may come from divergent sources, into a unique consistent belief set (base).
- Using belief merging and revision algorithms in order to diagnose oral cancer.
- Testing the implementation of belief merging operators using real world data.