Towards Integrating the Gene Ontology and the Hierarchical Bayesian Network Classification Model: An Empirical Case Study

Hasanein Alharbi, Al-Mustaqbal University College, Iraq

Abstract— Data Mining (DM) is knowledge-intensive process that can be significantly enhanced by integrating the domain knowledge. Recent research claimed that ontology can play various roles in the DM process. Additionally, ontology can facilitate different steps in the Bayesian Network (BN) construction task. To this end, this paper investigates the advantages of consolidating the Gene Ontology (GO) and the Hierarchical Bayesian Network (HBN) classifier in a flexible framework which preserves the advantages of both ontology and Bayesian theory. The proposed Semantically Aware Hierarchical Bayesian Network (SAHBN) classification model introduces a flexible framework that systematically consolidates domain knowledge in the form of ontology and the DM process. Furthermore, it establishes a solid foundation to explore the possibility of integrating more comprehensive ontological knowledge in the DM process. SAHBN is tested using three datasets in the biomedical domain to predict the effect of the DNA repair gene on the human ageing process. DNA repair genes are classified as either ageing-related or non-ageing related based on their GO biological process terms. Overall, SAHBN classifier shows a very competitive performance compared with the existing Bayesian-based classification algorithms. SAHBN has outperformed existing algorithms in more than 50% of the implemented experiments. Six performance criteria were used to evaluate the performance of the proposed SAHBN model.

Keywords— DNA Repair Gene, Hierarchical Bayesian Network, Human Ageing Process, Ontology, Semantic Data Mining.

1. INTRODUCTION

THE term Data Mining (DM) has been used to refer to methods that aim to extract useful information and knowledge from data. Fayyad et al. have defined these methods as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in databases [1], [2].

Although the ultimate goal of DM algorithms is to identify useful, understandable, and previously unknown knowledge from data, the majority of the existing mining algorithms are confined to the generation of frequent patterns and do not illustrate how to act upon them[3]–[5]. Some evidences suggest that the drawbacks in the existing mining algorithms are partially caused as a result of viewing the mining process as data-driven trial and error practices and ignoring the domain knowledge[4], [6]. Hence, the data mining philosophy has faced a paradigm shift from being a data-centered process to a knowledge-centered process that aims to cater to domain knowledge and its integration in the mining process [5], [7].

Domain knowledge can be represented using various techniques. However, recent research indicated that ontology playing significant role in the process of knowledge acquisition and representation [8], [9]. In fact, the formal structure of ontology makes it a strong candidate for knowledge integration in the DM algorithms. Ontology could be intertwined with the DM algorithms to bridge the semantic gap, to provide prior knowledge and constraints, to formally represent the mining results [10], [11].

Hence, the process of developing a framework which systematically consolidates ontology and the mining algorithms in an intelligent mining environment is investigated in this paper. The aim of this paper to explore the potential advantage obtained from coupling the domain knowledge in form of Gene Ontology (GO) and the hierarchical Bayesian Network classifier and then utilizing the developed model to predict the DNA repair gene effect in the human ageing process.

The contribution of this paper can be summarized in the following points:

- Propose an automatic, systematic and flexible framework to integrate ontology and the HBN.
- Exploit the implicit semantic knowledge of ontology to enrich the data classification process.
- Merge the deterministic nature of ontology and the uncertainty aspect in form of BN in such a way that preserves the advantages of both.
- Lay out a solid foundation to explore the advantages of integrating more ontological knowledge in the data classification process.

The structure of this paper is organized as follows. Section 2 discusses the underpinning techniques. Section 3 illustrates the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

proposed model in details. Section 4 presents the experimental results and evaluation. Finally, section 5 presents the discussion and conclusion.

II. Background techniques

A. Semantic Data Mining

Many researchers hold the view that the integration of domain knowledge in the DM process enriches the mining task and improves the quality of the generated patterns/mining model [12], [13]. Although domain knowledge can be expressed in various formats, recent research in the data mining filed suggested that ontology is the natural way to encode the domain knowledge for DM use [11]. The process of integrating domain knowledge in the DM task is known as Semantic Data Mining. At this stage, it is necessary to clarify exactly what is meant by Semantic Data Mining. It refers to data mining tasks which systematically incorporated domain knowledge, especially formal semantic, into the mining process. the branch of the semantic data mining which use ontology to represent the domain knowledge is referred to as the ontology-based semantic data mining which is the core of this paper [10], [14], [15]. Fig. 1 depicts the schema of ontology-based semantic data mining as proposed by Novak et al. [15].

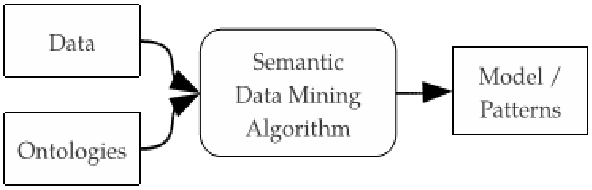

Fig. 1 Semantic Data Mining Schema [15]

According to D. Dou et al [11], the roles ontology can plays in the data mining task can be summarized in the following points:

1. Overcome the existing semantic gap between data, applications, data mining algorithms and data mining results.
2. Empower the mining algorithm with prior knowledge, which helps to guide the mining process or reduce the search space.
3. Formally represent the mining flow.

Recent researches have suggested that the challenge of developing fully automatic and systematic approach to integrate ontology and data mining process has not been realized. Furthermore, it has been reported that semantic data mining still in its early stages and has a promising future [11], [14]. Hence, this paper studied the advantages of integrating the hierarchical Bayesian network and the Gene Ontology (GO) then proposes Semantically Aware Hierarchical Bayesian Network (SAHBN) classifier.

B. Gene Ontology (GO)

Gene Ontology (GO) is a collaborative effort to construct controlled vocabularies that describes in consistent manner the roles the gene can play in the life of various organisms. Its main objectives are [16]–[18]:

1. Assemble a set of structured vocabularies to describe the domain of molecular biology.
2. Use the assembled vocabularies to annotate the gene and gene products.
3. Make the gene annotation dataset available to other researchers.

It has been reported that the GO structure consists of three hierarchies which cover the following biological aspects:

1. Molecular Functions (MF): MF terms are used to describe the abilities that gene products have or the jobs they may implement. This include activities such as transporting things around, binding to things, holding things together and changing one thing into another[16], [19], [20].
2. Biological Process (BP): BP terms are used to define a biological goal or objective implemented by an ordered series of molecular functions. The begging and the end of the MF activities which contribute to the BP are precisely defined [16], [19], [20].
3. Cellular Component (CC): the locations where the activities of the gene products took place are defined by cellular component terms. The locations may include a structural component of a cell such as the nucleus or it may refer to a location as part of a molecular complex such as ribosome [16], [19], [20].

As indicated previously, the GO structure consists of three hierarchies which organize the GO terms as a Directed Acyclic Graph (DAG) where each GO term represented as a node and the relationships between nodes defined as arcs. Parent-Child relation is the backbone of the GO structure which indicates that the parent nodes are more general than the child nodes [16], [20], [21].

Another significant aspect of the GO structure consistency constraints is the True-Path-Rule (TPR). The “is-a” (parent-child) GO relation obligated to follow the TPR which states that if an instance of GO node is proved to be true, so it’s ancestors all the way to the root must be true. Otherwise, if an instance founded to be false, so all its descendants to the leaf nodes must be false [16], [21], [22].

Additionally, gene ontology provides a comprehensive resource for annotating gene products [23]. According to previous studies [19], the gene ontology consortium (GOC) published 126 million annotations which cover more than 347,000 species. These annotations are created either automatically by computerized or manual means by experts who studied the relevant literature or examined biological data. Hence, this research not only used the GO to annotate the mined data but also exploited the semantic information integrated in the GO to guide the construction process of the mining model.

C. Hierarchical Bayesian Network (HBN)n

Hierarchical Bayesian Network (HBN) is defined as an extension or generalization of the standard Bayesian network (BN). In particular, the structure of the HBN provides more knowledge about the organization of the variables involved in the network and builds a more realistic probabilistic model [24]–[26].

In contrast to standard BN, which cannot represent non-propositional domains, HBN variables represent an aggregation of simple variables. Hence, HBN is an effective model that decomposes the investigated problem into smaller sub-tasks and provides more control over the data flow and better modeling techniques [24], [26].

Standard BN consists of a set of nodes and arcs that form the structure of the BN. Likewise, HBN structure compose of a set of nodes and arcs that represent the variables and relations, respectively. Additionally, the strength of the arcs between nodes is quantified by a set of conditional probability tables (CPTs). However, unlike standard BNs, the HBN arcs not only represent the probabilistic dependency between nods but also the “part-of” relationship, which can be either represented as nested nodes or tree-like hierarchical structures [25], [27]. Fig. 2 depicts both interpretation of the “part-of” relationship.

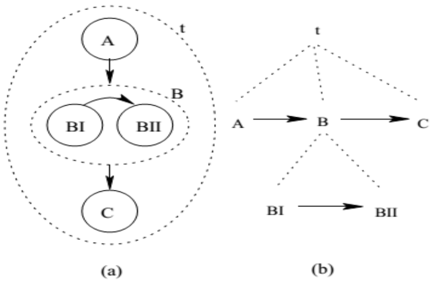

Fig. 2 Hierarchical Bayesian Network Structure. a) A nested representation of the HBN structure. b) A tree like representation of the HBN structurer [25]

Additionally, another significant aspect of HBN is the basic dependency rule that underpins the HBN structure. The HBN dependency rule states that a node is conditionally independent of its non-descendant nodes, given the value of its direct parent in the graph [24], [25].

The techniques underpin the proposed model have been explained, the next section discusses the proposed model in details.

III. PROPOSED MODEL

The GO structure represents and reflects high quality knowledge of the biomedical domain [23]. Likewise, the structure of the HBN implicitly provides more knowledge about the targeted domain [24]. As a results, the integration of these two concepts, GO and HBN, produce a classification model which seamlessly reflects the domain knowledge.

The proposed SAHBN classification model shares some initial steps with the standard classification algorithms such as data pre-processing and feature selections. However, the essential steps related to BN structure construction and variables probability estimation are designed in such a way that exploits the semantic nature of the GO. Fig. 3 compares the process sequence of the standard classification algorithms and the proposed SAHBN model.

Fig. 3 shows that the selected prediction attributes have been further processed based on the semantic knowledge extracted from the GO. This can be seen in the steps surrounded by the sold line in the same figure. The new steps introduced by the SAHBN model are summarized in the following subsections.

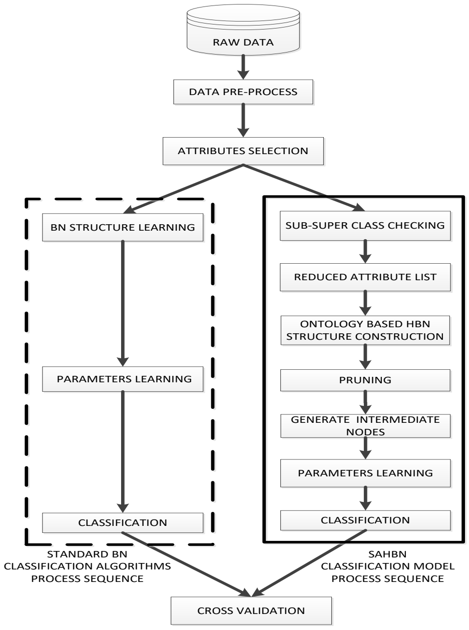

Fig. 3 SAHBN Model Vs BN Algorithms Process Sequence

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

A.Sub-Super class checking

The first step which follows the attributes selection task is to check whether there is a semantic relation between the selected attributes. This is done by matching the selected attributes to the GO concepts. The data sets covered in this paper used the GO biological process (GOBP) terms as prediction attributes. Hence, one-to-one matching between the selected attributes and the GO concepts was implemented. Consequently, the GO structure was exploited to extract the semantic relation between these attributes.

The relation that was targeted in this research is the parent-child ("is-a") class relations. GO used the "is-a" relation to represents the subtype relation between concepts. For example, "Replicative Cell Aging" is a subtype of and less general that the "Cell Aging" process. Likewise, the intermediate nodes in the HBN structure represent an aggregation of simpler nodes. Hence, the "is-a" relation was selected to identify the structure of the HBN.

The "is-a" relation not only facilitates the construction of the HBN structure, but also achieves the following objectives:

1. Maintain data consistency: the GO "is-a" relation follows the True Path Rule (TPR) which states that if an instance of GO node is proved to be true, so it's ancestors all the way to the root must be true. Otherwise, if an instance found to be false, so all its descendants to the leaf nodes must be false [16], [21], [22]. Thus, any two GO terms connected via the "is-a" relation and used as prediction attributes must follow the TPR. Otherwise, an inconsistent data set can be used to train the classification model, which may lead to inaccurate results.

Table 1 shows some records from the DNA repair gene-PPI data set (discussed in the 4th section), which highlights the inconsistency in the training data set.

Table 1 Sample of inconsistent training dataset

ROW No.	GO:0007568	GO:0001302	Label Class
1	TRUE	FALSE	TRUE
2	FALSE	FALSE	FALSE
3	FALSE	TRUE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	TRUE	TRUE
7	TRUE	FALSE	TRUE
8	FALSE	TRUE	TRUE
9	FALSE	TRUE	TRUE
10	FALSE	FALSE	FALSE
11	FALSE	FALSE	TRUE
12	FALSE	TRUE	TRUE
13	FALSE	FALSE	FALSE

According to the GO structure, the GO:0001302 attribute is a child class of the GO:0007568. While the former refers to the replicative cell ageing, the later refers to the ageing biological process, and there is an indirect "is-a" relationship between them. Hence, it can be seen that records 2,4,6,7 and 9 (bold and italic in table 1) are inconsistent because the value of the parent class is false; while the value of its child class is true and this violates the TPR.

Thus, this research proposed the use of the Chi-squared to break the conflict between the contradicted prediction terms and eliminate data inconsistency. This is done in four steps, as follows:

a. Indentify the contradictory GO prediction terms, which connected via the "is-a" relationship.
b. Calculate the Chi-squared value between each term and the label class.
c. Delete the GO term, which has the lowest dependency with the label class.
d. Repeat steps 1 throw 3 until all contradictions are removed.

2. Reduce prediction attributes list dimension: removing the contradicted attributes using GO "is-a" relation not only eliminates the inconsistency in the training data set but also reduces the dimension of the prediction attribute list. High dimensional data poses a serious challenge for data mining techniques, especially in medical domain.

B. Ontology-based HBN structure construction

The second step, which follows the parent-child class checking, is the HBN structure construction. The structure construction task is implemented based on the reduce attributes list and the structure of the GO. The steps involved in this process are summarized in the following points:

a. Match each attribute in the reduced list generated after the parent-child class checking step to node in the GO.

b. Extract the path for each matched node (i.e. attribute node) using the "is-a" relation and the GO structure. The path is extracted from the matched node all the way to root node. We began by extracting the parent class of the attribute node, and then the extracted parent class was considered as an attribute node and its parent class extracted. This process was repeated until the root node was reached.

c. Combine the extracted paths to form a tree-like hierarchical structure.

Fig. 4 depicts a sample of ontology-based HBN structure for attributes list consists of five GO terms {GO1, GO2, GO3, GO4, and GO5}. The predication attributes form the terminal nodes in the HBN structure and their parent classes shape the rest of the structure.

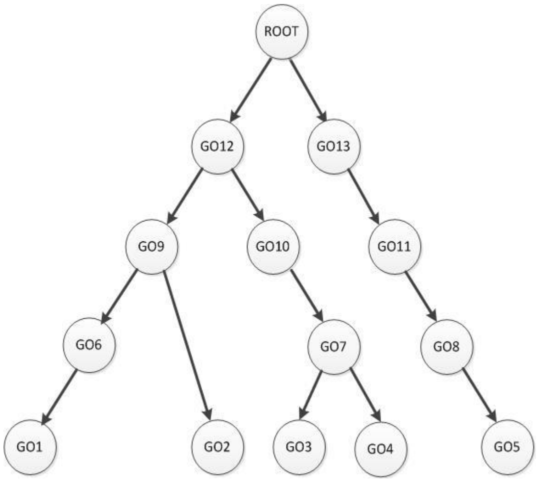

Fig. 4 Ontology-based HBN structure

C. Structure pruning

The structure pruning step exploits the transitive nature of the “is-a” relationship in the GO. The “is-a” relation is transitive which meant that if “A is-a B”, and “B is-a C”, we can infer that “A is-a C”. Hence, is save to aggregate terms connected by the “is-a” relationship [28]. Fig. 5 illustrates the structure pruning process.

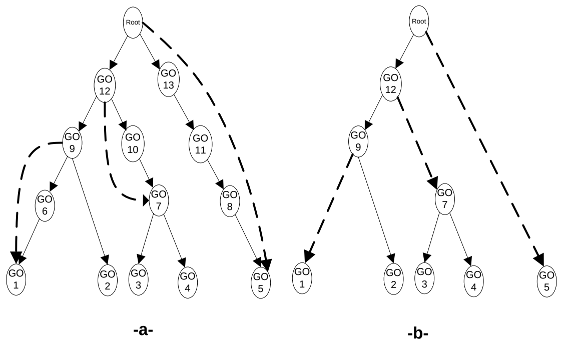

Fig. 5 HBN structure pruning process

The aim of this step is to remove redundant nodes that do not affect the principles of the HBN structure and maintain the semantic consistency of the targeted domain. There are two main basic principles underpinning the structure of the HBN. These principles can be summarized in the following points.

a. Aggregation: each node is the HBN structure represents an aggregation of simple nodes.
b. Independency: each node in the HBN structure is conditionally independent of its non-descendant node given the value of its direct parent.

Consequently, and in order to prune the created HBN structure without violating the above principles, the following steps were followed:

a. Delete all intermediate nodes that have only one child class.
b. The child class of the deleted node will be a child class of the deleted node parent class.

To demonstrate the pruning process, the above steps were applied to the structure of the HBN depicted in Fig. 5 (a), which was constructed in the previous step. As a results GO6, GO8, GO10, GO11 and GO13 terms, and the associated arcs, were deleted. The steps of the pruning process are summarized in Fig. 5 (b).

D. Generate intermediate nodes

Fig. 5 (b) shows that three intermediate nodes have been added to the structure of the HBN, namely, GO7, GO9, and GO12. Unlike the observed prediction attributes (i.e. terminal nodes), the value of the added intermediate nodes are unknown. However, as previously explained, the GO “is-a” relation is subject to the TPR. Consequently, the TPR principle was exploited to define the values of the intermediate nodes. This is done by implementing the following rule: “the value of any intermediate node is equal to true if and only if the value of any of its child classes is equal to true. Otherwise, its value is equal to false”. Consequently, semantically consistent and complete training data set was generated.

E. Parameters learning

Having filled the intermediate nodes with values, the next step is to learn the SAHBN variables probability. There are two main approaches for estimating the probability values in the BN model for complete dataset, namely, Maximum Likelihood Estimation (MLE) and Bayesian Estimation [29]–[32].

Despite its various advantages, the MLE method has the following limitation [30].

a. The size of the observed data set has no effect on the estimation process.
b. MLE does not consider the prior knowledge. Therefore, it entirely relies on the observed data set.

Hence, this research has used a Bayesian-based approach, namely, Maximum a Posterior Estimation (MAP) method to estimate the probability values of the SAHBN variables.

Thus far, this paper has argued that the dependency rule, which forms the base of the HBN structure, states that each node is conditionally independent of its non-descendant node given the value of its direct parent in the graph. Additionally, each node in the HBN structure represents an aggregation of simpler nodes. Likewise, the GO structure organizes its terms in hierarchical structure using the "is-a" relation, in which each GO parent term is more general that the child term. Hence, the assumption made in this research claimed the following observations:

"Each prediction attribute is represented as terminal node, which is independent of other prediction attributes given the value of its parent. Furthermore, each intermediate node is independent of its non-descendant nodes given its parent. Finally, the label class is placed as the root node". These assumptions meet the principles of the HBN and the GO structure.

Having discussed how to construct the SAHBN model and the associated techniques, the next Section discusses the experimental results in detail.

IV.EXPERIMENTAL IMPLEMENTATION AND EVALUATION

This section discusses the experimental implementation and the obtained results. A brief introduction to the human ageing process was given in the first subsection. Then, the second subsection discusses in details the process of DNA repair gens dataset creation. Finally, in the third subsection obtained results are thoroughly analyzed.

A. Human Ageing Process

Human aging is defined as the gradual failure of the physiological functions in various cells, tissues, and organs in the human body, which ultimately leads to the fragility of body functionalities with the time growth and increases the probability of death [33]–[35].

Recent researches have suggested that the advancements in the healthcare sector in developed countries have led to substantial increase in human lifespan. In fact, some findings indicated that almost 20% of the world's population will be aged 60 or older by 2050. Accordingly, the increase of the centenarians' percentage in different countries has led to many challenges, such as boost in age-related disease, increase in healthcare cost, and shortage of caregivers. Hence, the task of understanding human ageing process was the focus of many researchers to develop new techniques for preventing or delaying the diseases associated with the aging process or even treat them in more successful and rational ways [33]–[36].

Human ageing is an extremely complex, mysterious, controversial, and puzzling process that requires more investigation. Furthermore, studying the ageing process has led to challenges, such as ethical factors associated with doing experiments on human data, time for implementing the experiments on human data, and the comprehensive elements that must be considered when analyzing the ageing process. Thus, researchers have alternatively used the gene/protein databases of short living organism models to implement their experiments. Consequently, data mining techniques have been recently applied to analyze the large amount of open access gene/protein databases to gain some insights into the human ageing process [37]–[40].

Human genome preserves its integrity by protecting the cellular DNA from both internal and external attacks. Thus, the cellular DNA is steadily monitored by the repair enzymes to correct damages resulting from these attacks. Accordingly, DNA damage is an essential element in the human ageing process, and the modification of DAN repair process will result in advance understanding of the cellular ageing phenomena [39], [41]. Hence, the proposed SAHBN classification model applied to the DNA repair genes database to classify their effect as either an ageing-related or non-ageing related gene. The following subsection illustrates in details the data set creation process.

B. DNA repair gene data set creation

The datasets used in this research have been created in two different approaches. In the first approach, the protein-protein interactions (PPI) database is used to represent each DNA repair gene in the form of proteins. Then, the proteins are complied in form of Gene Ontology Biological Process (GO BP) terms. While, the second approach directly converts each DNA repair gene into its GO BP terms using the Gene2GO database. Fig. 6 clarifies the data sets creation process.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

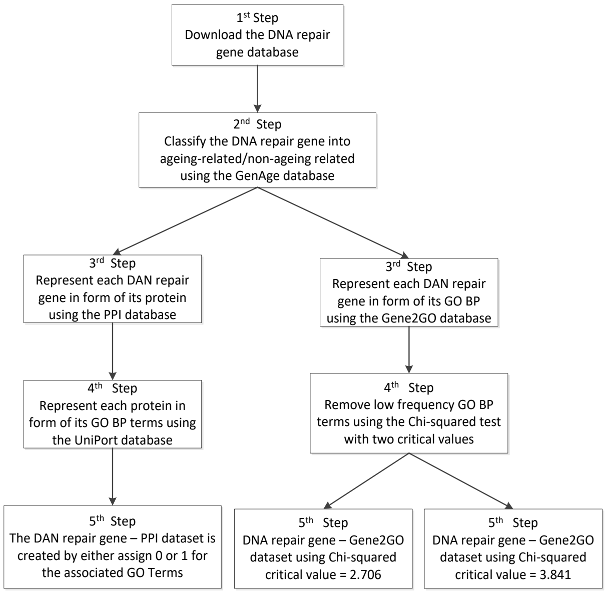
1st Step
Download the DNA repair gene database
2nd Step
Classify the DNA repair gene into ageing-related/non-ageing related using the GenAge database
3rd Step
Represent each DAN repair gene in form of its protein using the PPI database
3rd Step
Represent each DAN repair gene in form of its GO BP using the Gene2GO database
4th Step
Represent each protein in form of its GO BP terms using the UniPort database
4th Step
Remove low frequency GO BP terms using the Chi-squared test with two critical values
5th Step
The DAN repair gene – PPI dataset is created by either assign 0 or 1 for the associated GO Terms
5th Step
DNA repair gene – Gene2GO dataset using Chi-squared critical value = 2.706
5th Step
DNA repair gene – Gene2GO dataset using Chi-squared critical value = 3.841

Fig. 6 DNA repair gene data set creation process

The process depicted in Fig. 6 is further explained in the following steps:

- DNA repair gene-PPI data set.

1) Download the DNA repair gene database from the human DNA repair gene website [42].
2) Classify the downloaded DNA repair genes into two categories, namely, ageing-related and non-ageing-related. Those DNA repair genes appearing in the GenAge [43] database are classified as ageing related, while the DNA repair genes that do not appear in the GenAge database classified as non-ageing related.
3) Extract the protein-protein interactions from the human reference database [44]. The extracted protein interactions meet the following criteria.
    a) At least one of the interacted proteins is located in the DNA repair gene.
    b) The type of the evidence for the interaction is obtained from either in-vitro or in-vivo experiments.
4) Represent each protein in form of GO BP using the UniPort [45] database.
5) Finally, each DNA repair gene was represented in the form of GO PB terms, which are associated with the proteins that represent the gene. The value of the GO PB term was equal to 1 if it appears in the protein associated with the gene, otherwise, it was equal to 0.

- DNA repair gene-Gene2GO data set.

The first and second step of the DNA repair gene-Gene2GO approach is similar to the corresponding steps in the DNA repair gene-PPI approach. Hence, this approach is explained starting from the third step.

6) Represent each DNA repair gene in the form of GO BP terms using the National Center for Biotechnology Information (NCBI) Gene2Go [46]
7) BP terms with low frequency and possess no or very low prediction power were removed. A previous research [39] has used a predefined frequency threshold to remover low frequency terms. In a slightly different manner, this research used the Chi-squared technique to measure the dependency between each attribute and the label class. Then, attributes that appeared to be independent from the label class were removed. Two critical values were used in the Chi-squared test, precisely, 2.706 and 3.841.
8) Finally, each DNA repair gene was represented by its GO BP terms. The value of the GO term is equal to 1 if it is associated with the given gene. Otherwise, it assigned to 0.

C. Experimental Results

This subsection gives a detailed description of the experimental implementation and the obtained results. It discusses the implementation of the proposed SAHBN model, comparison with existing classification algorithms, and analysis the results obtained. Accordingly, Weka [47] and Netica-J API [48] software tools are used to accomplish these tasks.

For each data set, 11 attributes selection methods are used and 6 performance criteria are measured. The calculate performance criteria are: precision, recall, F1 measure, accuracy, average accuracy and harmonic accuracy. Additionally, the performance of SAHBN model was compared with existing Bayesian-based classification algorithm, such as: ICC, K2, TAN, Hill-Climbing and Tabu. In total, 297 experiments are implemented.

The following subsections summarize the results for each data set.

- DNA repair gene-PPI data ser results summary:

DNA repair gene-PPI data set is used to compare the performance of the proposed SAHBN classification model against the performance of three Bayesian-based classification algorithms, namely, ICSS, K2, and TAN. In this experiment 11 attributes selection methods are used and 6 performance

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

criteria are measured. Table 2 summarize the results in terms of the number when SAHBN model outperformed, equal to, or less than the performance of existing algorithms with respect to all 6 performance criteria.

Table 2 DNA repair gene-PPI dataset results summary

SAHBN Vs Existing Algo.	Outperform	Equal to	Less Than
Precision	13	6	14
Recall	23	9	1
F1 Measure	18	5	10
Accuracy	15	10	8
Average Accuracy	24	5	4
Harmonic Accuracy	27	2	4
Total	120	37	41

Table 2 indicates that the total number of tests when SAHBN model outperformed existing algorithms is almost triple time the number of tests when SAHBN model is exceeded by existing algorithms. Additionally, SAHBN harmonic accuracy surpassed existing algorithms in 27 out of 33 experiments. Furthermore, SAHBN exceeded existing algorithms with respect to average accuracy and recall. All in all, SAHBN model outperformed existing algorithms in all performance criteria, except precision. The obtained results are further visualized in figure below

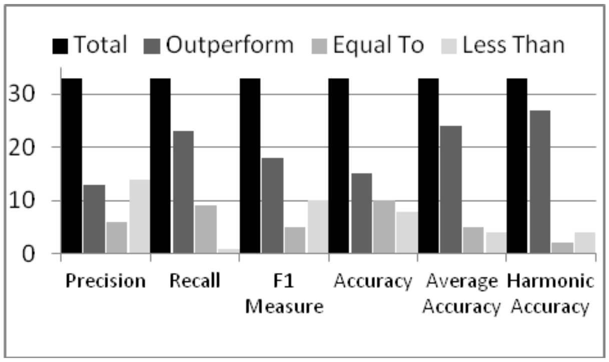

Fig. 7 DNA repair gene-PPI dataset results summary

- DNA repair gene-Gene2GO (CV=2.706 and CV= 3.841) data sets results summary:

The DNA repair gene-Gene2GO (CV 2.706 and CV = 3.841) data sets are utilized to compare the performance of the proposed SAHBN model with Hill-climbing and Tabu Bayesian-based classification algorithms. SAHBN are tested using 6 scoring measures and 11 attributes selection approaches. Consequently, 132 experiments are implemented for each data set, and the obtained results are summarized in Table 3 and Table 4 respectively.

Table 3 DNA repair gene-Gene2go (CV= 2.706) results summary

SAHBN Vs Existing Algo.	Outperform	Equal to	Less Than
Precision	62	21	49
Recall	74	10	48
F1 Measure	72	11	49
Accuracy	67	22	43
Average Accuracy	70	10	52
Harmonic Accuracy	74	10	48
Total	419	84	289

Table 3 clearly shows that SAHBN model has outperformed existing algorithms in all performance criteria. For instance, SAHBN exceeded existing algorithms in 74 of out 132 experiments with respect to recall and harmonic accuracy. Additionally, the numbers of experiments when SAHBN exceeded existing algorithms in terms of F1 measures and average accuracy were 72 and 70, respectively. Overall, SAHBN surpassed the existing algorithms in more that 50% of the total number of implemented tests for all performance criteria. This can be clearly seen in Fig. 8.

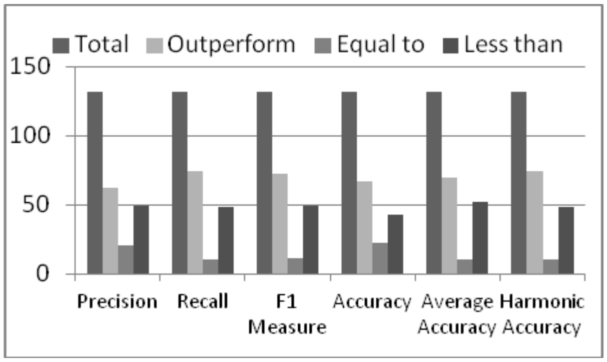

Fig. 8 DNA repair gene-Gene2GO (CV=2.706) results summary

While Table 3 presents the obtained results for DNA repair gene-Gene 2GO data set with CV = 2.706, Table 4 summarizes the results for the same data set for Chi-squared critical value = 3.841.

Table 4 DNA repair gene-Gene2GO (CV=3.841) RESULTS summary

SAHBN Vs. Existing Algo.	Outperform	Equal to	Less Than
Precision	62	14	56
Recall	75	7	50
F1 Measure	81	5	46
Accuracy	71	24	37
Average Accuracy	78	11	43
Harmonic Accuracy	73	6	53
Total	440	67	285

The DNA repair gene-Gene2GO (CV=3.841) data set results summary presented in Table 4 confirmed the findings of the previous data set. SAHBN outperformed the existing algorithms in all performance criteria. For example, the number of experiments when SAHBN model outperformed existing algorithms in term of F1 measure was 81 experiments out of 132 experiments. Additionally, SAHBN exceeded the existing algorithms with respect to accuracy, harmonic, and average accuracies in 71, 73, and 78 experiments, respectively. Overall, SAHBN has outperformed existing algorithms in 440 out of 792 tests.

Compared to the results presented in table X, the total number of tests when SAHBN model outperformed the existing algorithms is slightly increased. Hence, it can be concluded using higher critical value in the Chi-squared test has led to better results. Fig. 9 depicts the result summary of the DNA repair gene-Gene2GO dataset with a critical value equal to 3.841.

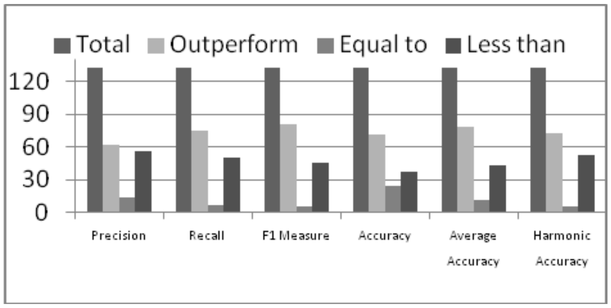

Fig. 9 DNA repair gene-Gene2GO (CV=3.841) results summary

V. DISCUSSION AND CONCLUSION

This paper investigated the potential advantages of integrating the domain knowledge in the form of ontology and HBN classifier. Accordingly, the proposed Semantically Aware Hierarchical Bayesian Network (SAHBN) classification model was tested using three data sets in the biomedical domain. Consequently, the findings extracted from analyzing the obtained results indicated that SAHBN model demonstrated a very competitive performance in comparison with existing Bayesian-based classification algorithms. SAHBN model outperformed the existing algorithms in more than 50% (149+ experiments out of 297 experiments) of the experiments with respect to five performance criteria and slightly less than 50% with respect to precision. Table 5 and Fig. 10 below summarize the overall results for all three data sets in terms of the number when SAHBN has either outperformed, equal to or less than existing algorithms for all performance criteria.

Table 5 Three data sets results summary

SAHBN Vs. Existing Algo.	Outperform	Equal to	Less Than
Precision	137	41	119
Recall	172	26	99
F1 Measure	171	21	105
Accuracy	153	56	88
Average Accuracy	172	26	99
Harmonic Accuracy	174	18	105
Total	979	188	615

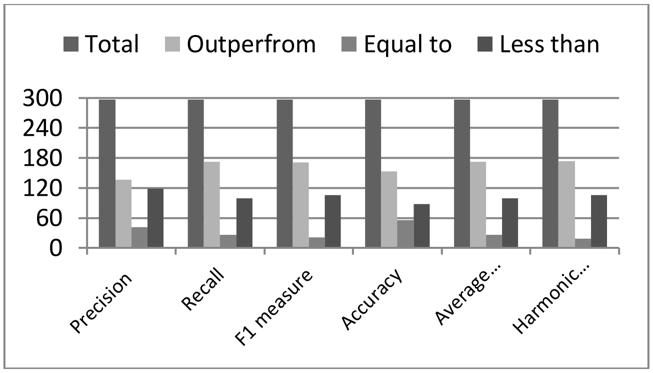

Fig. 10 Overall results summary

The proposed SAHBN model exploited the ontological knowledge to construct a consistent training dataset and eliminate contradictions between prediction attributes. Additionally, SAHBN structure implicitly reflected the background knowledge of the targeted domain. Hence, it was a self-explanatory structure that can readily be maintained.

SAHBN model not only highlighted the advantages of integrating ontology with the HBN classifier but also laid out the foundations to consider a more semantic relation between prediction attributes, such as equivalent, disjoint, union, and intersection. Future works will investigate the advantages of integrating more ontological knowledge with the SAHBN classification model.

In summary, the SAHBN model consolidated the gene ontology and the HBN classification algorithms in a flexible framework that preserves the advantages of ontology and Bayesian theory. Initial results revealed very promising findings that establish a solid foundation for future research.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI magazine, vol. 17, no. 3, pp. 37–54, 1996.

[2] C. Zhang and S. Zhang, Association Rule Mining: Models and Algorithms. Springer-Verlag Berlin Heidelberg. XII, 244., 2002.

[3] M. Sexton and S. Lu, “The challenges of creating actionable knowledge: an action research perspective,” Construction Management and Economics, vol. 2, pp. 683–694, 2009.

[4] L. Cao, P. S. Yu, C. Zhang, and Y. Zhao, Domain driven data mining. New York: Springer, 2010.

[5] L. Cao, “Domain-driven data mining: Challenges and prospects,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 6, pp. 755–769, 2010.

[6] H. Dahan, S. Cohen, L. Rokach, and O. Maimon, Proactive Data Mining with Decision Trees. Springer Science & Business Media., 2014.

[7] C. Antunes and A. Silva, “New Trends in Knowledge Driven Data Mining a position paper,” Proceedings of the 16th International Conference on Enterprise Information Systems, pp. 346–351, 2014.

[8] S. Staab and R. Studer, Hand Book on Ontologies. Springer Science & Business Media, 2013.

[9] G. Mansingh and L. Rao, “The Role of Ontologies in Developing Knowledge Technologies,” In Knowledge Management for Development. Springer US, pp. 145–156, 2014.

[10] H. Liu, “Towards semantic data mining,” In Proc. of the 9th International Semantic Web Conference (ISWC2010). 2010.

[11] D. Dou, H. Wang, and H. Liu, “Semantic data mining: A survey of ontology-based approaches,” in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), 2015, pp. 244–251.

[12] J. Han and M. Kamber, Data Mining: Concepts and Techniques, vol. 12. 2011.

[13] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, “Data mining techniques and applications – A decade review from 2000 to 2011,” Expert Systems with Applications, vol. 39, no. 12, pp. 11303–11311, 2012.

[14] P. Ristoski and H. Paulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey,” Web Semantics: Science, Services and Agents on, 2016.

[15] P. K. Novak, A. Vavpetic, I. Trajkovski, and N. Lavrac, “Towards semantic data mining with g-segs,” in Proceedings of the 11th International Multiconference Information Society, IS, 2009.

[16] J. A. Blake and M. A. Harris, “The Gene Ontology (GO) Project: Structured vocabularies for molecular biology and their application to genome and expression analysis,” Current Protocols in Bioinformatics, no. SUPPL. 23. 2008.

[17] M. Harris, J. Deegan, and J. Lomax, “The Gene Ontology project in 2008,” Nucleic Acids Research, vol. 36, no. Database issue, pp. D440–D444, 2008.

[18] P. Gaudet, N. Škunca, J. C. Hu, and C. Dessimoz, “Primer on the Gene Ontology,” arXiv preprint arXiv:1602.01876, 2016.

[19] R. Balakrishnan, M. A. Harris, R. Huntley, K. Van Auken, and J. Michael Cherry, “A guide to best practices for gene ontology (GO) manual annotation,” Database, vol. 2013, 2013.

[20] The Gene Ontology Consortium, “Gene Ontology Consortium: going forward,” Nucleic Acids Research, vol. 43, no. D1, pp. D1049–D1056, 2015.

[21] S. Götz and A. Conesa, Visual Gene Ontology Based Knowledge Discovery in Functional Genomics. INTECH Open Access Publisher, 2011.

[22] R. P. Huntley, T. Sawford, M. J. Martin, and C. O’Donovan, “Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt.,” GigaScience, vol. 3, no. 1, p. 4, 2014.

[23] J. A. Blake, “Ten quick tips for using the gene ontology,” PLoS Comput Biol, vol. 9, no. 11, p. e1003343, 2013.

[24] E. Gyftodimos and P. a Flach, “Hierarchical Bayesian Networks : An Approach to Classification and Learning for Structured Data,” Proceedings of the ECML/PKDD - 2003 Workshop on Probablistic Graphical Models for Classification, vol. 3025. pp. 291–300, 2004.

[25] E. Gyftodimos and P. A. Flach, “Hierarchical bayesian networks: A probabilistic reasoning model for structured domains,” in Proceedings of the ICML-2002 Workshop on Development of Representations, 2002, pp. 23–30.

[26] M. M., L. D., F. N., and S. K., “A hierarchical, ontology-driven Bayesian concept for ubiquitous medical environments--a case study for pulmonary diseases.,” Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, pp. 3807–3810, 2008.

[27] E. Gyftodimos and P. Flach, “Learning hierarchical bayesian networks for human skill modelling,” in Proceedings of the 2003 UK workshop on Computational Intelligence (UKCI-2003). University of Bristol, 2003.

[28] “Gene Ontology Consortium | Gene Ontology Consortium.” [Online]. Available: http://www.geneontology.org/. [Accessed: 21-Dec-2016].

[29] T. D. Nielsen and F. V. Jensen, Bayesian Network and Decision Graph. Springer Science & Business Media, 2009.

[30] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques. MIT press, 2009.

[31] R. G. Almond, R. J. Mislevy, L. S. Steinberg, D. Yan, and D. M. Williamson, “Learning in Models with Fixed Structure,” Bayesian Networks in Educational Assessment. Springer New York, pp. 279–330, 2015.

[32] Z. Ji, Q. Xia, and G. Meng, “A Review of Parameter Learning Methods in Bayesian Network,” in Advanced Intelligent Computing Theories and Applications: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part III, D.-S. Huang and K. Han, Eds. Cham: Springer International Publishing, 2015, pp. 3–12.

[33] H. E. Wheeler and S. K. Kim, “Genetics and genomics of human ageing.,” Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 366, no. 1561, pp. 43–50, 2011.

[34] H. Lees, H. Walters, and L. S. Cox, “Animal and human models to understand ageing,” Maturitas, 2016.

[35] T. B. Kirkwood, “The origins of human ageing.,” Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 352, no. 1363, pp. 1765–72, 1997.

[36] P. Rashidi and A. Mihailidis, “A survey on ambient-assisted living tools for older adults,” IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 3, pp. 579–590, 2013.

[37] C. Wan, A. A. Freitas, and J. P. De Magalhaes, “Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 2, pp. 262–275, 2015.

[38] J. P. de Magalhães et al., “The Human Ageing Genomic Resources: Online databases and tools for biogerontologists,” Aging Cell, vol. 8, no. 1. pp. 65–72, 2009.

[39] A. a Freitas, O. Vasieva, and J. P. de Magalhães, “A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related.,” BMC genomics, vol. 12, no. 1, p. 27, 2011.

[40] C. Wan and A. Freitas, “Prediction of the pro-longevity or anti-longevity effect of Caenorhabditis Elegans genes based on Bayesian classification methods,” in Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, 2013, pp. 373–380.

[41] R. D. Wood, M. Mitchell, J. Sgouros, and T. Lindahl, “Human DNA repair genes.,” Science (New York, N.Y.), vol. 291, no. 5507, pp. 1284–9, 2001.

[42] “Human DNA repair genes.” [Online]. Available: http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html. [Accessed: 08-Dec-2016].

[43] “GenAge: The Ageing Gene Database.” [Online]. Available: http://genomics.senescence.info/genes/. [Accessed: 08-Dec-2016].

[44] “Human Protein Reference Database.” [Online]. Available: http://www.hprd.org/index_html. [Accessed: 08-Dec-2016].

[45] “UniProt.” [Online]. Available: http://www.uniprot.org/. [Accessed: 08-Dec-2016].

[46] “National Center for Biotechnology Information.” [Online]. Available: https://www.ncbi.nlm.nih.gov/. [Accessed: 08-Dec-2016].

[47] I. H. W. Eibe Frank, Mark A. Hall, “The WEKA Workbench. Online Appendix for ‘Data Mining: Practical Machine Learning Tools and Techniques.’” Morgan Kaufmann, 2016.

[48] “Norsys Software Corp. - Bayes Net Software.” [Online]. Available: https://www.norsys.com/. [Accessed: 22-Dec-2016].