Semantically Aware Hierarchical Bayesian Network Model for Knowledge Discovery in Data: An Ontology-based Framework

Hasanein Alharbi

School of Computing, Science and Engineering

University of Salford

Manchester, UK

Submitted in Partial Fulfilment of the Requirements of the Degree of Doctor of Philosophy, 2017

Table of Content

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank ALLAH ALMIGHTY for giving me the patience and strength to conduct and complete this thesis.

My most profound gratitude goes to my sponsor, The Iraqi government represented by the Ministry of Higher Education and Scientific Research and the Iraqi cultural attaché in London, for their continuing help and support whilst undertaking this research.

It is difficult for me to express adequately my gratitude and appreciation to my parents, brothers, sisters, nephews and nieces, who never ceased praying for me and wishing me every success.

I wold also like to use this opportunity to express my gratitude to my supervisor, Dr. Mohamad Saraee, for his support throughout this research.

The following papers have been published as part of this research. The published papers include some of the content of this thesis.

1. Alharbi, Hasanein, and Mohamad Saraee. "Semantic Aware Bayesian Network Model for Actionable Knowledge Discovery in Linked Data". Machine Learning and Data Mining in Pattern Recognition. Springer International Publishing, 2016. 143-154.
2. Alharbi, Hasanein, and Mohamad Saraee. "Towards Integrating Ontology and Hierarchical Bayesian Network: A Flexible Framework", CSE 2017 Annual PGR Symposium (CSE-PGSym 17), 17th March 2017, University of Salford, UK.

Declaration

As the authors of this thesis, we hereby confirm that no portion of the work presented in this thesis is submitted in support of an application for another degree or qualification at Salford or any other university.

List of abbreviations

ALS	Airborne Laser Scanner
BN	Bayesian Network
BP	Biological Process
CC	Cellular Component
CDSS	Clinical Decision Support System
CPG	Clinical Practice Guideline
CPTs	Conditional Probability Tables
CRISP-DM	Cross-Industry Standard Process for DM
CVS	Concept Vector Space
DAG	Directed Acyclic Graph
df	Degree of Freedom
DFS	Depth First Search
DM	Data Mining
FMEA	Failure Model and Effect Analysis
FN	False Negative
FP	False Positive
GO	Gene Ontology
GOC	Gene Ontology Consortium
HBN	Hierarchical Bayesian Network
HTTP	Hypertext Transfer Protocol
ICVS	Improved Concept Vector Space
IPFP	Iterative Proportional Fitting Procedure
JPD	Joint Probability Distribution
KDD	Knowledge Discovery in Database
KNN	K Nearest-Neighbour
LD	Linked Data
MAP	Maximum a Posterior Estimation
MEBN	Multi-Entity Bayesian Networks
MeSH	Medical Section Head
MF	Molecular Function
MLE	Maximum Likelihood Estimation
NCBI	national Centre for Biotechnology Information
OOBN	Object Oriented Bayesian Network

OWL	Web Ontology Language
PIN	Protein Interaction Networks
PPI	Protein-Protein Interactions
RDF	Resource Description Framework
RDFS	RDF Schema
RF	Random Forest
SAHBN	Semantically Aware Hierarchical Bayesian Network
SPARQL	Simple Protocol & RDF Query Language
SVM	Support Vector Machine
SW	The Semantic Web
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TNM-O	Tumour Node Metastasis Ontology
TP	True Positive
TPR	True Path Rule
UMLS	Unified Medical Language System
URIs	Uniform Resource Identifiers
WBC	White Blood Cell

Abstract

Several mining algorithms have been invented over the course of recent decades. However, many of the invented algorithms are confined to generating frequent patterns and do not illustrate how to act upon them. Hence, many researchers have argued that existing mining algorithms have some limitations with respect to performance and workability.

Quantity and quality are the main limitations of the existing mining algorithms. While quantity states that the generated patterns are abundant, quality indicates that they cannot be integrated into the business domain seamlessly. Consequently, recent research has suggested that the limitations of the existing mining algorithms are the result of treating the mining process as an isolated and autonomous data-driven trial-and-error process and ignoring the domain knowledge. Accordingly, the integration of domain knowledge into the mining process has become the goal of recent data mining algorithms.

Domain knowledge can be represented using various techniques. However, recent research has stated that ontology is the natural way to represent knowledge for data mining use. The structural nature of ontology makes it a very strong candidate for integrating domain knowledge with data mining algorithms. It has been claimed that ontology can play the following roles in the data mining process:

- Bridging the semantic gap.
- Providing prior knowledge and constraints.
- Formally representing the DM results.

Despite the fact that a variety of research has used ontology to enrich different tasks in the data mining process, recent research has revealed that the process of developing a framework that systematically consolidates ontology and the mining algorithms in an intelligent mining environment has not been realised. Hence, this thesis proposes an automatic, systematic and flexible framework that integrates the Hierarchical Bayesian Network (HBN) and domain ontology.

The ultimate aim of this thesis is to propose a data mining framework that implicitly caters for the underpinning domain knowledge and eventually leads to a more intelligent and accurate mining process. To a certain extent the proposed mining model will simulate the cognitive system in the human being.

The similarity between ontology, the Bayesian Network (BN) and bioinformatics applications establishes a strong connection between these research disciplines. This similarity can be summarised in the following points:

- Both ontology and BN have a graphical-based structure.
- Biomedical applications are known for their uncertainty. Likewise, BN is a powerful tool for reasoning under uncertainty.
- The medical data involved in biomedical applications is comprehensive and ontology is the right model for representing comprehensive data.

Hence, the proposed ontology-based Semantically Aware Hierarchical Bayesian Network (SAHBN) is applied to eight biomedical data sets in the field of predicting the effect of the DNA repair gene in the human ageing process and the identification of hub protein. Consequently, the performance of SAHBN was compared with existing Bayesian-based classification algorithms. Overall, SAHBN demonstrated a very competitive performance.

The contribution of this thesis can be summarised in the following points.

- Proposed an automatic, systematic and flexible framework to integrate ontology and the HBN. Based on the literature review, and to the best of our knowledge, no such framework has been proposed previously.
- The complexity of learning HBN structure from observed data is significant. Hence, the proposed SAHBN model utilized the domain knowledge in the form of ontology to overcome this challenge.
- The proposed SAHBN model preserves the advantages of both ontology and Bayesian theory. It integrates the concept of Bayesian uncertainty with the deterministic nature of ontology without extending ontology structure and adding probability-specific properties that violate the ontology standard structure.
- The proposed SAHBN utilized the domain knowledge in the form of ontology to define the semantic relationships between the attributes involved in the mining process, guides the HBN structure construction procedure, checks the consistency of the training data set and facilitates the calculation of the associated conditional probability tables (CPTs).
- The proposed SAHBN model lay out a solid foundation to integrate other semantic relations such as equivalent, disjoint, intersection and union.

Chapter 1 Introduction

The term data mining (DM) is used to refer to methods that aim to extract useful information and knowledge from data. Fayyad et al. defined these methods as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in a database [1], [2].

Despite the fact that the ultimate goal of DM algorithms is to identify useful and understandable knowledge from data, many of the existing mining algorithms are confined to generating frequent patterns and do not illustrate how to act upon them. Accordingly, some researchers have argued that existing mining algorithms demonstrate some serious drawbacks with respect to performance and workability [3]–[5].

These drawbacks have affected the extracted knowledge in terms of both quantity and quality. While the former states that the generated patterns are abundant, the latter indicates that they cannot be integrated into the business domain seamlessly [4], [6]. There is some evidence to suggest that these drawbacks have been partially caused as a result of viewing the mining process as data-driven, trial-and-error practices that ignore the domain knowledge [4], [7]. Consequently, DM philosophy has faced a paradigm shift from being a data-centred to a knowledge-centred process that aims to accommodate the domain knowledge in the mining process [5], [8].

Although domain knowledge can be represented through various techniques, recently ontology has played a significant role in the process of knowledge acquisition and representation. The term ontology has different interpretations in different communities. Nevertheless, this thesis has used the interpretation provided by the knowledge engineering community, which defines ontology as “explicit specification of conceptualizations”. In other words, ontology is a simplified, abstract and formal representation of an area of interest, which describes the concepts, entities and relationships that hold among them. Hence, ontology has bundled data and its semantics into one package [9], [10].

The formal structure of ontology makes it a strong candidate for knowledge incorporation on DM algorithms. It is believed that ontology can be intertwined with DM algorithms to perform the following tasks:

- Bridging the semantic gap.
- Providing prior knowledge and constraints.

- Formally representing the DM results.

However, the process of developing a framework that systematically consolidates ontology and mining algorithms in an intelligent mining environment remains an open research question [11], [12].

Hence, this thesis studied the potential advantages of consolidating the surrounding knowledge in the form of ontology, and the extant mining algorithms then proposed an automatic, systematic and flexible framework that integrates the Hierarchical Bayesian Network (HBN) and domain ontology. The proposed Semantically Aware Hierarchical Bayesian Network (SAHBN) framework uses gene ontology (GO) to define the semantic relations between the attributes involved in the mining process, guides the BN structure construction process, checks the consistency of the training data set and facilitates calculation of the associated conditional probability tables (CPTs).

Consequently, a data classification model was developed based on the proposed SAHBN framework, and tested using eight real-life data sets in the biomedical domain. Finally, the obtained results were compared with the results of the existing Bayesian-based mining algorithms. In conclusion, the proposed SAHBN framework demonstrated a very competitive performance in comparison to the existing algorithms.

1.1 Problem Statement

It is widely agreed that integration of the domain knowledge in DM algorithms enriches various stages of the knowledge-extraction process. Furthermore, it is believed that ontology is the most appropriate approach to represent the domain knowledge for data mining use because of its structural format [1], [11]–[13]. However, to the best of our knowledge, the challenges associated with ontology and DM algorithm integration have not been solved. For example:

- There is no standard framework that systematically integrates ontology and DM algorithms.
- Ontology is developed based on description logic and does not accommodate the uncertainty factor that is characteristic of many real-life situations.
- The majority of existing DM algorithms follow an interactive data-driven process, which does not cater for any external source of knowledge.

Hence, this thesis proposes a SAHBN frame to address the above challenges.

1.2 Research Motivations

Many researchers hold the view that domain knowledge can play an important role in the DM process and bridge the gap between business requirements and DM algorithms. Thus, this research is motivated by the following facts:

1. The gap between the output of the current DM algorithms and business requirements. While end-users expect a consistent set of understandable, actionable and usable patterns, the output of the existing mining algorithms is abundant and cannot be integrated into the business domain transparently [8], [14].
2. The vast proliferation of ontology forms an attractive environment for data mining. Recently, many domains have constructed an ontology knowledge base that embodies the domain knowledge. Health care is one of the leading sectors to have embraced this concept and used it voraciously [15]–[17].
3. The new philosophy in knowledge representation, coined in the form of linked data (LD), ontology and the Semantic Web (SW), has presented new techniques to couple data with its semantics in one package. Hence, these techniques can be a vital source of external knowledge in various steps of the data mining process [13].
4. Existing DM algorithms are designed to analyse a domain-specific silo of data. Hence, they cannot cope with the multidisciplinary nature of LD and ontology. As a result, multidisciplinary knowledge cannot be obtained using the existing DM algorithms [18].
5. The open source nature of ontology releases research in the DM field from the burdens of the strict data access approval process, especially medical data, when data sets are safeguarded by various rules and the police [19].

1.3 Research Objectives

The ultimate aim of this research is to propose an automatic, systematic and flexible framework that analyses and exploits the semantic nature of ontology in the mining process. In fact, there is an absolute necessity and high demand for an ontology-based DM framework in various domains. Accordingly, in order to materialise the aim of this research, the following objectives must be fulfilled.

1 Develop an ontology-based DM framework that analyses, exploits and preserves the semantic nature of the targeted domain.

2 Convert the conceptual model of ontology structure into a graphical probabilistic model, namely, the HBN, which compactly represents the logical relations between concepts in the targeted domain.

3 Use the maximum a posteriori estimation (MAP) parameter estimation techniques to calculate the probabilistic values associated with the targeted domain concepts in such a way that reflects the semantic relationship between them.

4 Evaluate the proposed framework by applying it to various data sets in the biomedical domain and measure its performance using different performance criteria.

The primary research hypothesis is that integration of the domain knowledge in the form of ontology can help to improve the DM algorithms' performance. Furthermore, exploiting the semantic relation attached to the domain ontology will lead to better DM and KDD processes.

1.4 Research Methodology

Research is defined as the art of scientific investigation, which aims to know the unknown and to gain new knowledge. However, in the academic environment the term "research" is used in a technical sense to refer to the following activities [20].

1. Defining and redefining problems.
2. Formulating hypothesis or suggest solution.
3. Collecting, organising and evaluating data.
4. Finally, testing the original hypothesis.

Kothari et al. summarised the steps involved in the research process in a flow chart, which is depicted in Figure 1-1 (below).

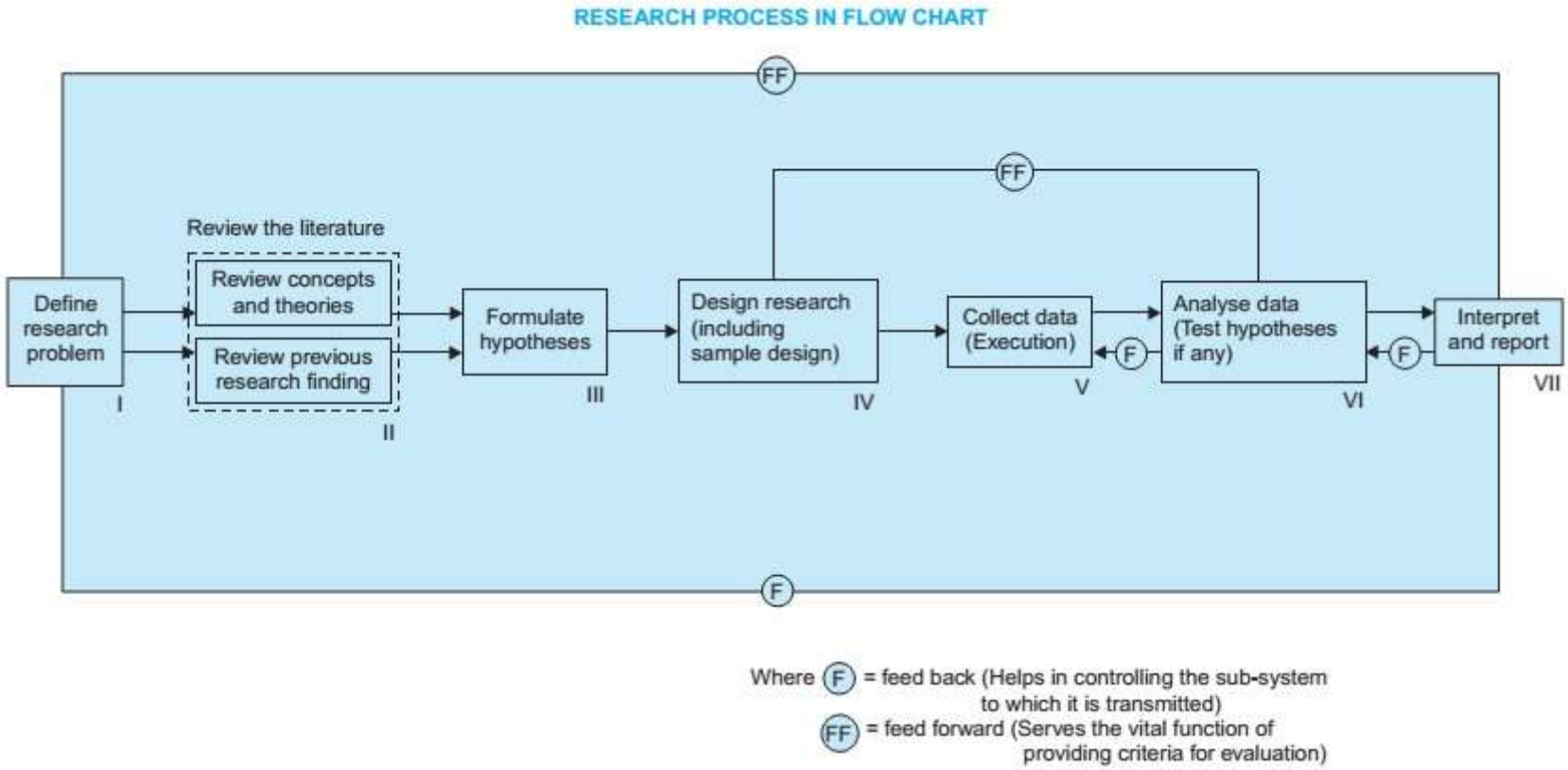


Figure 1-1 Research process in flow chart [20]

Having defined what is meant by research and the main steps involved in the research process, the next subsections will discuss the basic types of research.

1. **Descriptive versus analytical research**

   Social science and business research mainly use the descriptive research approach, which includes various types of survey and fact-finding enquiry. The term *ex post facto* has been used to refer to descriptive research studies when researchers report what has happened or what is happening and have no control over the input variables. On the other hand, the analytical research approach is used to analyse and study existing information and to make a critical evaluation of the available information [20]–[22].

2. **Applied versus fundamental research**

   Research that aims to find the solution to a current problem in society or an industrial/business organisation is called applied research. Meanwhile, research deals with generalisations, and the formulation of a theory is fundamental. For example, research related to natural phenomena or pure mathematics is an example of fundamental research. Likewise, studies concerned with human behaviour are also an example of fundamental research. In contrast, research that investigates whether particular social, economic or political trends may affect the problem facing a certain organisation is an example of applied research [20].

3. **Quantitative versus qualitative research**

   Quantitative research is concerned with an experimental process, which could be measured using certain statistical and mathematical criteria. In contrast, qualitative research deals with phenomena in which quality or kind are involved, for example, when researchers investigate a human behaviour such as why people think or do certain things. How do they feel or what do they think about a particular subject?[20].

4. **Conceptual versus empirical research**

   Research studying an abstract idea(s) or theory is known as conceptual research. Philosophers and thinkers use this approach to define new concepts or to redefine an existing one. On the other hand, an experience or data-based driven research is known as empirical research. In this type of research, researchers must propose a working hypothesis and gather enough data to prove or disapprove it. Hence, the researcher has complete control over the variables under study and can purposely

modify the values to study their effects. It has been argued that empirical research is the most powerful approach to justify a hypothesis under investigation [20].

It is common practice for researchers to use various types of the above research approaches to conduct their research. Likewise, the research presented in this thesis has used a combination of quantitative and empirical research approaches. In fact, the proposed framework was applied to eight real-life data sets, and statistical criteria such as accuracy were used to measure the performance and to compare it with the existing algorithms. Hence, the approach used in this research is in line with quantitative and empirical research approaches. The following steps were followed to identify the research gap and define a hypothesis.

1. Identifying the research gap and highlighting the key concepts that lay out the foundations of the identified gap. This step is derived by the fact that there is wide acceptance that DM is a knowledge-intensive process, which could be further improved by integrating the domain knowledge in the form of ontology. However, no standard framework to integrate ontology and DM algorithms has been proposed yet.
2. Carrying out an in-depth literature review of the state-of-the-art literature, which covers the following topics:
    - Data mining and knowledge discovery process.
    - Linked data (LD), ontology and the Semantic Web (SW).
    - Semantic data mining.
    - Ontology-based data classification methods.
    - Integration ontology and Bayesian Network (BN).
3. Designing and implementing a solution that aims to propose an automatic, systematic and flexible framework to consolidate ontology and HBN in such a way that preserves the advantages of both.
4. Empirical evaluation, including implementation of the proposed model using different real-life data sets. Additionally, the performance of the proposed model was measured using various quality criteria and compared against existing Bayesian-based mining algorithms using the 10-fold cross-validation.
5. Results analysis, including studying and contrasting the results of the empirical implementation and then drawing a conclusion and outlining future work from the established findings.

1.5 Thesis Overview

This section summarises the organisation of this thesis. The structure of this thesis is divided into six chapters, as follows:

- **Chapter 1: Introduction**
  Chapter 1 gives an introduction to the work presented in this thesis and covers the research motivation, objectives, methodologies and hypothesis.
- **Chapter 2: Background and Literature Review**
  Chapter 2 presents the background techniques and reviews the state-of-the-art literature, which covers the following topics:
  - Data mining and knowledge discovery process;
  - Linked data (LD), ontology and the Semantic Web (SW);
  - Semantic data mining;
  - Ontology-based data classification methods;
  - Integrating ontology and the Bayesian Network.
- **Chapter 3: Fundamental Techniques and Proposed Framework**
  Chapter 3 covers the technical aspects that form the scientific foundations of the proposed framework. Additionally, it presents the steps involved in the development of the proposed Semantically Aware Hierarchical Bayesian Network (SAHBN) framework.
- **Chapter 4: Empirical Implementation and Experimental Results**
  Chapter 4 explains the structure of the empirical implementation in terms of the tested data sets and the measured performance criteria. It covers performance quality criteria, data pre-processing, attribute selection methods, implementation of the SAHBN framework and the existing algorithms.
- **Chapter 5: Conclusions and Future Work**
  Finally, Chapter 5 concludes the thesis by revisiting the initial research objectives and discussing potential future work.

1.6 Research Limitation

The model proposed in this thesis addressed an important topic of integrating the SW and HBN classifier. However, it yet has the following limitations:

1. It has been assumed that there is an ontology which describes the hierarchical relationship between attributes in the targeted domain.

2. Only attributes and their parent's classes are utilized to construct the structure of the HBN. Hence, synonyms of attributes are ignored.
3. The calculation of the conditional probability tables relied heavily on the observed data (i.e. A-Box) and does not reflect all semantic relationships between attributes.
4. The description logic sentences associated with the underpinning ontology were not exploited to elect more semantic knowledge.

Chapter 2 Background and Literature Review

As mentioned in the previous chapter, this thesis investigates the advantages of integrating ontology with an existing data mining algorithm, namely, the Hierarchical Bayesian Network (HBN) classifier. Hence, in order to develop a DM framework that exploits the semantic nature of ontology in the data classification process, the following areas must be studied.

a. Data mining and knowledge discovery process;
b. Linked data, ontology and the Semantic Web;
c. Semantic data mining;
d. Ontology-based data classification;
e. Integrating ontology and the Bayesian Network.

Consequently, the purpose of this chapter is to review the state-of-the-art literature on the above-mentioned topics in such a way that rationalises their connections to the proposed research hypothesis. Figure 2-1 (below) depicts the studied areas and explains the relations between them.

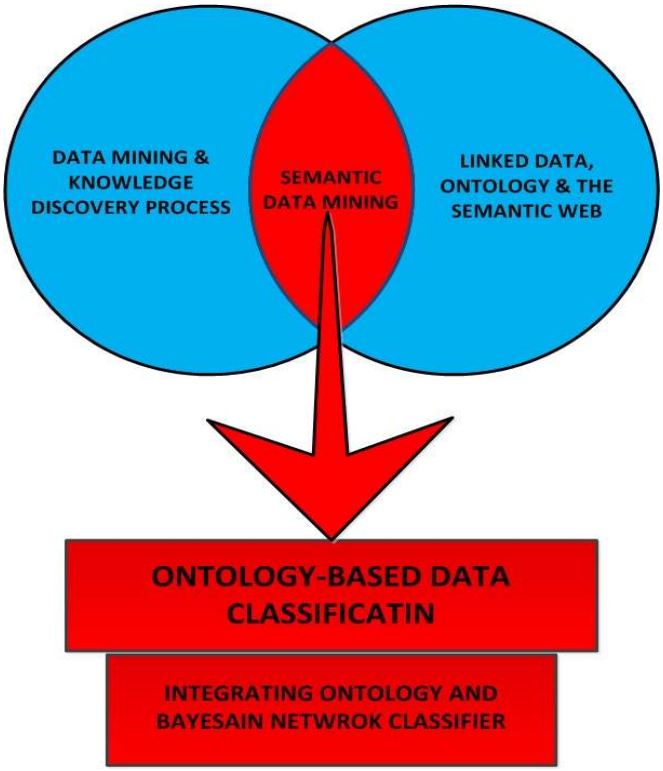


Figure 2-1 Literature Review Structure

2.1 Data Mining and Knowledge Discovery Process

The steady development of modern computers has accumulated enormous amounts of data at a rapid pace. Furthermore, the affordable cost of electronic storage devices has encouraged various organisations to collect data on a daily basis. Consequently, it has been reported that the volume of data is doubled every 20 months. This situation has rightly been described as "data rich but knowledge poor" [18], [23], [24]. Useful examples of sources for data generation may include, but are not limited to, the following:

a. NASA Earth observation satellites generate a terabyte of data every day.
b. The human genome project generates thousands of bytes for several billions of genetic bases.
c. A hundred million customer transactions have been maintained by many companies.
d. There are a vast number of automatic recording devices, such as cash machines, sensors, CCTV recording, Web logs, and many others.
e. Hundreds of millions of websites generate a vast volume of data every day.
f. Hundreds of millions of Facebook and Twitter users are sending and sharing a huge volume of data every day.

It is commonly believed that the availability of a huge amount of data has resulted in an imminent need for tools that can analyse the existing data and generate some useful knowledge. The extracted knowledge could be used in market analysis, fraud detection, customer retention, production control, science exploration, Internet agents, telecommunication and manufacturing, among other things. Hence, in order to respond to this need, the concept of knowledge discovery in database (KDD) was formally introduced at the first KDD workshop in 1989.

KDD is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable, patterns in data. The notion of extracting knowledge from data has been investigated under various titles such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing. However, it is widely agreed that the term KDD is used to refer to a practical, interactive and iterative process, which consists of many steps, coupled with the data mining (DM) process. Although DM is only one step in the KDD process, it is an essential one [1], [23]–[26].

The steps involved in the KDD process vary from one DM model to other. For example, the academic-oriented model proposed by Fayyad consists of nine steps, while the cross-

industry standard process for DM (CRISP-DM) models consists of six steps. Figure 2-2 (below) summarises the steps used in each model [27].

Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specification			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

Figure 2-2 Side-by-side comparison of the major existing KDDM models [27]

It is evident that some steps are common to all models. The following points give a brief description of the actions taken in each step [23].

a. **Developing an understanding of the application domain:** this is an opening step in the mining process, which includes analysis of the business domain and understanding the goal of the mining task from the domain experts. Furthermore, an initial decision about the forthcoming activities such as transformation, mining algorithms, representation, and so on, must be taken. However, the mining process is an iterative and interactive process; therefore, when the process is advancing, decisions can always be reviewed.

b. **Selecting and creating a data set on which discovery will be performed:** the formation of the data that will be used in the KDD process is a crucial step. The quality of the data created in this step will affect the quality of the generated patterns. This step includes the following activities: 1) Identifying what data is available. 2) Collecting extra data if required. 3) Merging the data into one set and identifying the attributes that will be utilised in the KDD process.

c. **Pre-processing and cleansing:** the main goal of this step is to improve the quality of the targeted data. This step includes the removal of noise or outliers and handling missing data.

d. **Data transformation:** this step varies from one project to other. It is described as a project-specific activity. However, the essential actions involved in this step include two major tasks. The first task is dimension reduction, such as feature selection, and the second task is attribute transformation, such as discrimination of numerical attributes.

e. **Choosing the appropriate Data Mining task:** as explained earlier, the goal of the KDD process identified in the initial step affects the later decisions. Consequently, the type of mining algorithm must meet the goals identified in the first step. Generally speaking, there are two types of mining algorithm, namely, prediction and description. The prediction algorithms are also known as supervised algorithms such as classification and regression. Meanwhile, the description algorithms are referred to as unsupervised algorithms such as clustering.

f. **Choosing Data Mining algorithm:** having defined the goal and type of the KDD process in the previous step, the next step is to define the tool needed to accomplish this goal. This stage includes identification of a specific mining algorithm to generate the required knowledge.

g. **Employing the Data Mining algorithm:** applying the chosen mining algorithm to the created data set is at the core of the KDD process. This step may include execution of the mining algorithm many times until the required knowledge is generated.
h. **Evaluation:** taking into consideration the initial goal identified in the first step, the generated knowledge is evaluated and interpreted. Additionally, it is documented for future use.
i. **Using the discovered knowledge:** finally, in this step the generated knowledge is integrated into other systems for further actions.

It is a widely held view that the KDD process is multidisciplinary, which requires a combination of tools from various domains such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, machine learning, artificial intelligence, knowledge acquisition for expert systems, information retrieval, image and signal processing, visualisation and social science methodologies [25], [28], [29].

As mentioned earlier, various types of mining algorithm are used to achieve different tasks. Rokach et al.'s book classifies the mining algorithms into two main categories: verification and discovery. While verification algorithms are concerned with authenticating user hypotheses, discovery methods aim to generate new rules and patterns. Discovery methods have been further divided into two types: predictive and descriptive. Alternatively, the machine learning community has used the term supervised learning to refer to the prediction methods, and unsupervised learning to refer to the descriptive methods. Figure 2-3 (below) gives detailed information about the taxonomy of the data mining algorithms [23], [24], [30].

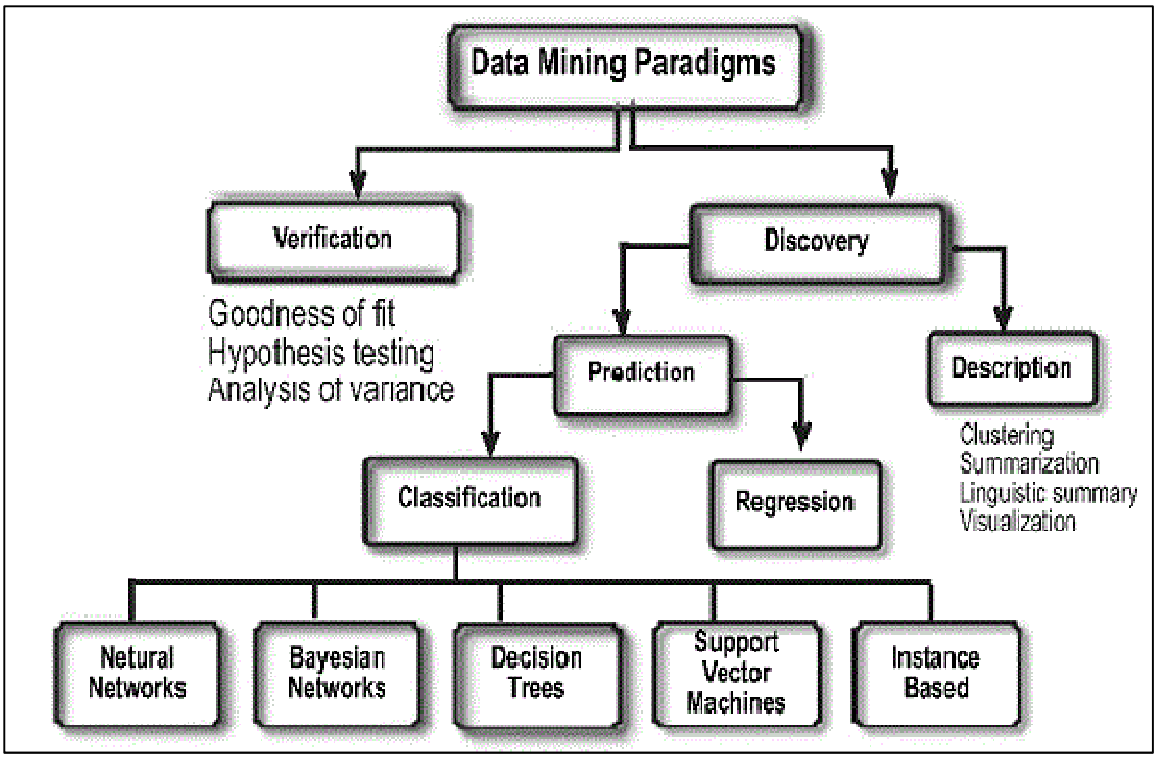


Figure 2-3 Taxonomy of data mining methods [30]

Although there is large number of published researches in data mining field, only few of them are applicable in real life applications. In fact, much research has argued that there is a significant imbalance in the data mining field with respect to workability and performance. This situation can be summarised in the following points [4], [5], [7].

a. **Algorithm imbalance:** many published algorithms versus several really workable ones in the business environment.
b. **Pattern imbalance:** many patterns mined versus only a small portion of, or no, satisfying business expectations.
c. **Decision imbalance:** many patterns identified versus effectively very few able to be taken over for business use.

Consequently, it is believed that there is a gap between the outputs of existing DM algorithms and business requirements. This is evident in the differences between the output of the existing data mining algorithms and end-user expectations. Therefore, initial observation suggests that the extracted patterns and knowledge do not offer the desired workable, actionable and operable capabilities [4], [31]. It seems possible that the limitations of the extracted patterns and knowledge is due to considering the mining process as a data-driven, trial-and-error process that overlooks the domain knowledge. It is a widely held view that the mining algorithms extract patterns that meet the technical interestingness measures and ignore business interest. As a result, the data mining process has faced a paradigm shift from being a data-centred process to domain-driven knowledge discovery. The ultimate aim of domain-driven knowledge discovery is integrating the

background knowledge into the mining process. In fact, it has been proven that integrating the background knowledge can improve the quality of the generated patterns [32], [33].

In recent years, the emergence of linked data (LD), ontology and the Semantic Web (SW) has played a significant role in the process of knowledge acquisition and representation. Hence, the aim of this thesis is to integrate background knowledge in the form of ontology with the DM algorithm. The proposed framework not only caters for the domain knowledge integration but also facilitates the integration of different silos of data. Therefore, more comprehensive knowledge could be generated in contrast to an area-based data mining approach [4], [7], [16], [18].

2.2 Linked Data (LD), Ontology and the Semantic Web (SW)

The conventional Web has met its objectives as a global document repository that can be easily accessed and consumed by human beings. However, the goal of constructing a document repository that is processed by not only humans but also machines has not been realised. Hence, the SW was introduced as an extension of the current Web, providing the technical capabilities to publish data in a machine-understandable format [34]–[37].

The ultimate vision of the SW is to provide a common framework that intertwines data and its semantics in one package. Furthermore, it facilitates data sharing and reuse across different applications. Thus, the essential step that must be materialised to meet the SW vision is the development of a technique for publishing and connecting data over the Web. This technique is the backbone of the SW and lies at the heart of its architecture. Accordingly, the concept of linked data (LD) was introduced in response to this need [38]–[40]. Figure 2-4 (below) depicts the architecture of the SW.

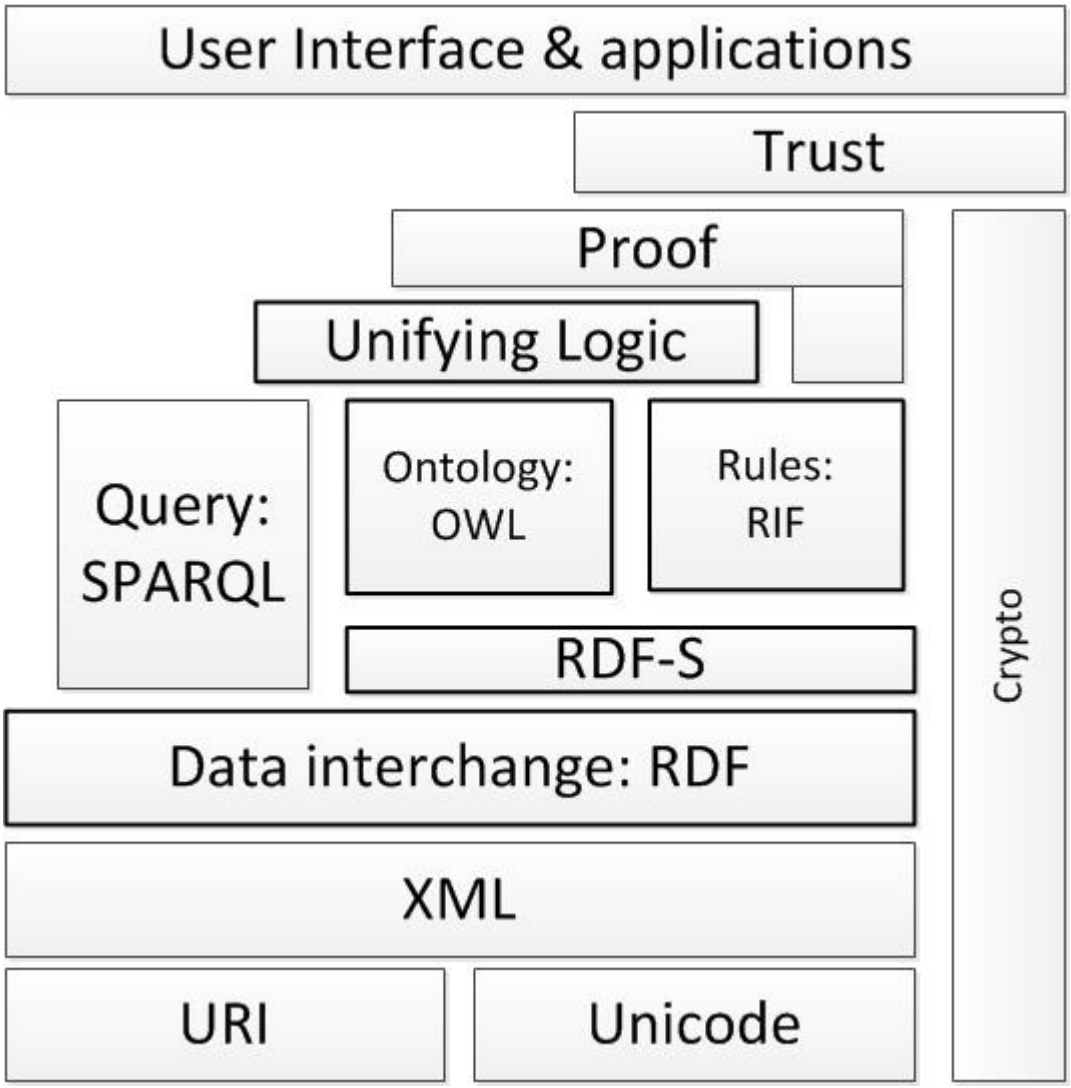


Figure 2-4 The SW reference architecture [41]

Linked data (LD) was introduced by Sir Tim Berners-Lee in 2006 and defined as the set of best practices for publishing and connecting structured data on the Web. These practices are currently known as LD principles. They can be summarised in the following points [42].

a. Uniform resource identifiers (URIs), used to name things.
b. They use Hypertext Transfer Protocol (HTTP) and URIs, so that people can look up names.
c. When someone looks up a URI, they provide useful information, using the standard resource description framework (RDF) and the simple protocol & RDF query language (SPARQL).
d. They include links to other URIs, so that they can discover more things.

The technical foundation of the LD concept is the use of HTTP and URI, not only to access Web documents but also to define and access real-world entities. Additionally,

RDF is used to represent the defined entities [43].]. In fact, RDF is the basic building-block for information representation in the SW.

The structure of the RDF statement is built based on a triple format, that is, subject–predicate–object, which forms a graph-based data model. Furthermore, RDF is enriched by the RDF schema (RDFS for short). RDFS is utilised to define the vocabularies used in a specific RDF data model. It uses a sub-class, sub-property relationship, in addition to domain and range restrictions, to describe the relationship between objects and to identify which property applies to which object. In contrast to RDF and RDFS, the Web Ontology Language (OWL) has added more vocabularies to describe properties and classes, leading to better machine readability of the Web content [42], [44], [45].

The word "ontology" is rooted in philosophy. Philosophers used ontology to refer to the study of the nature of existence. Nevertheless, the concept of ontology has been hijacked by the computer science field and used in a different sense. In computer science literature, ontology is used to provide a shared understanding of domain knowledge and treated as a special kind of information object or computational artefact. It has been defined as "formal, explicit specification of shared conceptualization". The terms used in the ontology definition could be further explained as follows [9], [36], [37], [46]:

a. *Formal* indicates that ontology must be machine-readable.
b. *Explicit* refers to the fact that the types of concept, and the constraints on their use, are explicitly defined.
c. *Shared* means that the notation captured by ontology must reflect a consensual knowledge that is accepted by a group, and not a private vision of certain individuals.
d. *Conceptualisation* refers to an abstract model of some phenomenon in the world, having identified the relevant concepts of the phenomenon.

A number of researchers hold the view that computational ontology has been used to formally represent the conceptual structure of a domain. Different entities in the targeted domain and their relations can be encoded in the term ontology [9], [47], [48]. Figure 2-5 (below) shows an example of how ontology is used to represent a publication domain.

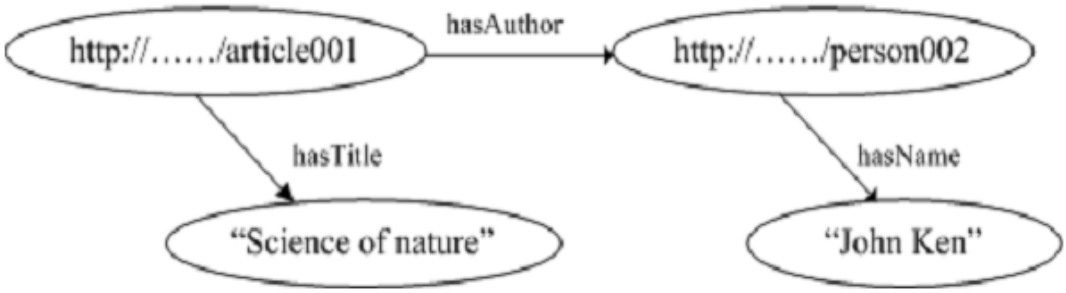


Figure 2-5 Ontology RDF triple relation [44]

According to [44], the structure of ontology is defined as tuple consisting of the following elements: S = (C, R, H, rel, A) [44], where:

- C is the set of entities describing the conceptual structure for the targeted system;
- R is the set of relation types;
- H is the set of taxonomy relationship of C;
- rel is the set of relationship of C with relation of type R, where rel ⊆ C X C;
- A is the set of description logic sentences.

Furthermore, rel has been further defined as a set of three tuple relations, that is, rel = (s, r, o), which represents the subject–relation–object relationship, where:

- s is the subject, which is an element from C;
- r is the relation, which is an element from R;
- o is the object, which is an element of C.

To further illustrate the structure of ontology and to gain a better understanding of its basic components, the following example converts the ontology model given in Figure 2-5 (above) into its equivalent RDF statements.

Example: the RDF triples for the ontology model of the publication system (Figure 2-5).

- hasAuthor(article001,person002).
- hasTitle(articale001,"Science of nature).
- hasName(person002,"John Ken).

The above RDF statements can be serialised in RDF/XML syntax, as shown in Figure 2-6 (below).

```
<rdf:Description about="http://domain/article001">
     <hasAuthor rdf:resource="http://domain/person002"/>
            <hasName rdf:resource="John Ken"/>
     </hashAuthor>
     <hasTitle rdf:resoource="Science of nature"/>
<rdf:Descrtiption>
```

Figure 2-6 RDF/XML Syntax [44]

It has been reported that ontology can be broadly classified into two categories, namely, lightweight and heavyweight ontologies. The lightweight ontology is mainly concerned with taxonomies and includes concepts, concept taxonomies, and the relationship between concepts and properties that describe concepts. On the other hand, heavyweight ontology addresses the domain in a deeper way and provides more restrictions by adding axioms and constraints to the lightweight ontology. Figure 2-7 (below) categorises ontology based on its structural complexity. It follows a line where ontologies move from lightweight to heavyweight [46], [49].

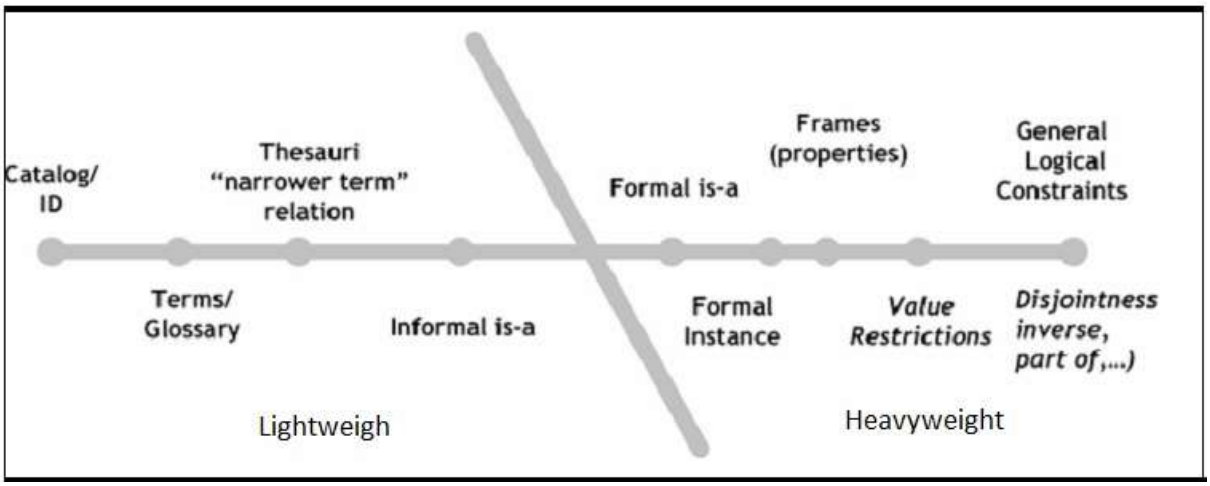


Figure 2-7 McGuiness Ontology Spectrum Classification [49]

The process of ontology development and management is not a trivial process; nor can it be implemented in a simple way. In fact, the objectives of the constructed ontologies and the development methodology must be critically assessed and closely pursued by the organisation or the individual pushing for their creation. The field that studies the methods and tools for ontology development and maintenance is known as ontology engineering. Ontology engineering refers to the set of activities that deal with the ontology development process, which includes the ontology life cycle, methodologies, tools and languages for building ontologies. It has been advised to implement the following activities during the ontology development process [9], [46].

a. Ontology management activities include scheduling, control and quality assurance.
b. Ontology-development-oriented activities are grouped into pre-development, development and post-development.
c. Ontology support activities include knowledge acquisition, evaluation, integration, merging, alignment, documentation and configuration management.

The ontology development activities have been summarised in Figure 2-8 (below).

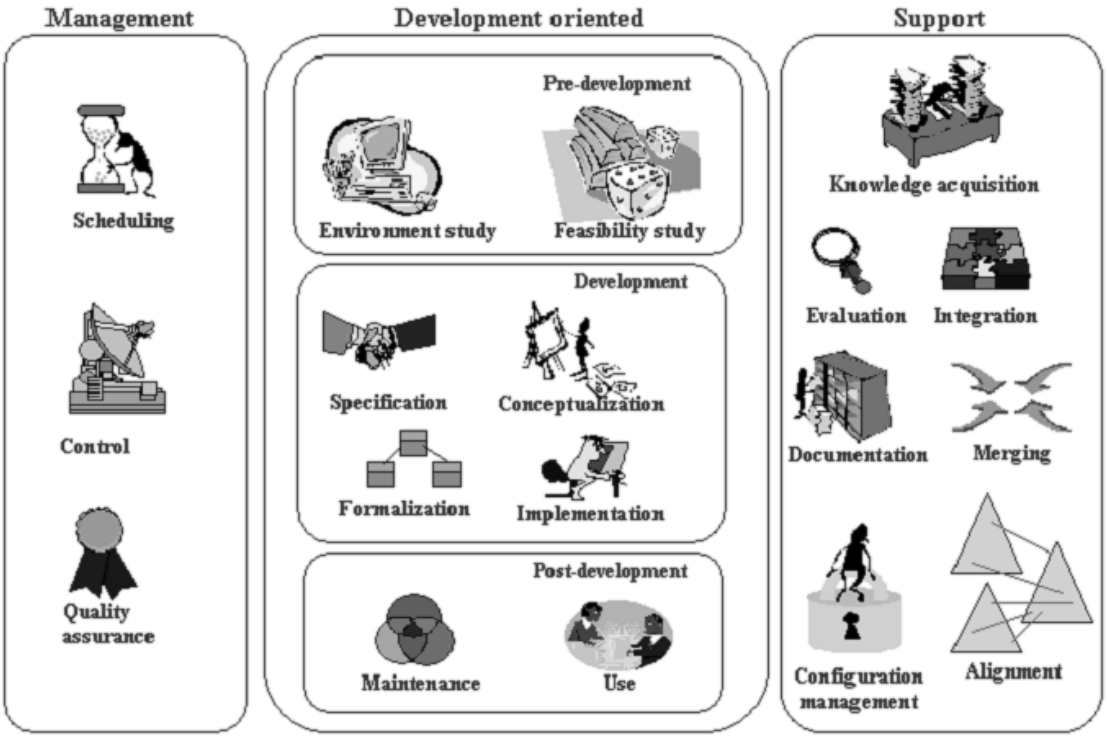


Figure 2-8 Ontology development process [46]

Having defined what ontology is and explained its structure and how it can be used to represent domain knowledge, the next section discusses the integration of ontology in the DM techniques.

2.3 Traditional Data Mining Versus Semantic Data Mining

As mentioned in the previous section, the KDD and DM processes have been defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data. Additionally, it has been believed that KDD is a multidisciplinary process, which requires the integration of various techniques spanning different disciplines such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, machine learning, artificial intelligence, knowledge acquisition for expert systems, information retrieval, image and signal processing, visualisation and social science methodologies. Furthermore, many researchers hold the view that the incorporation of domain knowledge with the KDD process enriches the mining process and improves the quality of the generated knowledge [1], [25], [29].

Although domain knowledge can be expressed in various formats, recent research in the data mining field suggests that ontology is the natural way to encode domain knowledge for data mining use. The process of integrating domain knowledge with the data mining task is known as semantic data mining. At this stage, it is necessary to clearly define what

is meant by semantic data mining. It refers to data mining tasks that systematically incorporate domain knowledge, especially formal semantic, into the mining process. The branch of semantic data mining that uses ontology to represent the domain knowledge is called ontology-based semantic data mining, which forms the core of this thesis [11], [13], [50], [51]. Figure 2-9 (below) depicts the schema of the ontology-based semantic data mining methodology, as proposed by Novak et al. [51].

Figure 2-9 The proposed semantic data mining methodology schema [51]

Recent research has suggested that ontology can play various roles in the DM task. For example, [51] exploited the "is-a" hierarchical structure of ontology to prune the pattern search space and obtain efficient search techniques. Meanwhile, [11] reported that using ontology for semi-structured data semantic annotation is an essential step in the semantic data mining process. He claimed that semantic annotation intertwines the data with its semantics. However, the survey conducted by [50] concluded that ontology can contribute to the mining task in three ways:

a. Bridging the semantic gap between the data, application, data mining algorithms and data mining results.
b. Providing data mining algorithms with a priori knowledge, which either guides the mining process or reduces/constrains the search space.
c. Providing a formal way of representing the data mining flow, from data pre-processing to the mining results.

The findings of recent research in the semantic data mining area have argued that the challenge of developing a fully automatic and systematic approach to integrating ontology and the data mining process has not been realised. Furthermore, it has been reported that semantic data mining is still in its early stages and has a promising future [13], [50]. Hence, this thesis studied the advantages of integrating the Bayesian Network classifier, and ontology then proposed a Semantically Aware Hierarchical Bayesian Network classifier, which is explained in detail in the next chapter.

2.4 Ontology-based Data Classification

Ontology has played various roles in the data classification process at different domains. Several studies have reported that text data classification methods have used ontology to represent the semantic aspects of text data and to inject it into the mining process.

For instance, the authors of [52] developed lexical variations and a synonyms ontology to define the logical relation between different words used in customers' reviews. Hence, even though the online product reviews are written using different words that share the same meaning, they can still be classified using the proposed support vector machine (SVM) augmented by the developed lexical variation ontology. The basic idea is that the weight of a word can be measured based on the weight of its synonym or lexical variant. Thus, if word-1 is a synonym of word-2 and the weight of word-2 has been calculated, then the weight of word-1 can be measured based on the weight of word-2. Similarly, the SAHBN model proposed in this thesis used the "is-a" relation enriched by the true path rule (TPR) to generate the values of the semantically related attributes. Although the authors of [52] claim that their model was tested using 20,000 product reviews for training and 6,000 reviews for testing, and that good classification accuracy was obtained, the key limitation of this approach is that the training data set must be tokenised based on the developed lexical variation ontology, and no evidence has been given to prove that the developed ontology is comprehensive enough to cover all the identified words. Additionally, it has not been discussed what alternative actions can be taken if the identified word does not match any concept class in the developed ontology. In contrast to this approach, our method guarantees that the identified prediction attributes always match a concept class in the used ontology.

Likewise, [53] used the WordNet ontology to enrich the selected attributes. For each singular and plural substantive term, the lexemes, synonyms, hyponyms and meronyms were introduced. The introduction of term variation was done based on semantic similarity measure techniques and the "is-a", "is-part-of", "is-member-of" and "is-substance-of" relations. The semantic enrichment process was started using the Wu-Palmer similarity measure to extract synonyms from a sub-tree of the WordNet ontology. The "is-a" synonyms were studied first and then the search was extended to cover other relations such as "is part-of", "is-member-of", and "is-substance-of". This process was applied for each attribute or term of each phrase sequentially. Consequently, the authors evaluated the performance of the proposed model using six classification algorithms and compared their

discriminatory power. Eventually, they concluded that the support vector machines (SVM) obtained the highest performance.

Despite the various advantages of the work presented by [53], it would have been more interesting if it had utilised the extracted ontological knowledge, not only in the attributes enhancement stage but also in the subsequent stages of the mining process. Hence, the SAHBN model proposed in this thesis was designed to use the ontological knowledge, not only to identify the semantic relations between attributes but also to guide the structure construction of the BN classifier, to check training data consistency and to assist the CPT estimation process.

In a similar way, [54] exploited the "is-a" relation in the domain ontology to extract the ancestors' concepts. Accordingly, the authors claim that extending the number of the word vector to include semantically extended words generated better results. In fact the authors used ontology not only to identify the semantically related concepts, but also to calculate the weight for each concept and its parents. Likewise, the SAHBN model proposed in this thesis used the "is-a" relation to anticipate the values of the semantically related variables, eventually using these values to calculate the conditional probability tables to quantify the strength of the semantic relation.

The work proposed by [54] was validated using a data set consisting of 952 CT abdomen and 1,128 CT neuron reports, which were classified using the k-nearest-neighbour (KNN) algorithm, with and without integrating the semantic aspect. Consequently, the authors concluded that the semantically extended vector achieved a better performance compared to the non-semantic case.

In slightly different way, [55] used ontology to calculate the weight for the concept vector space (CVS). The authors used ontology and the page-ranking algorithm to calculate the importance of each concept and then form the result as an improved concept vector space (ICVS). In fact, concept weight is an aggregation of the concept's importance and relevance. The proposed model consists of three steps:

1. Convert the domain ontology into a graph.
2. Calculate the weight of each concept using the page-ranking algorithm.
3. Construct the improved concept vector space (ICVS).

A total of 348 grant documents were randomly divided into two unequal parts (70%, 30%) and used for training and testing the classifier respectively. Consequently, various

classification algorithms were used to evaluate the performance of the proposed model. Eventually, the authors reported that the semantically improved concept vector space leads to better classification performance in terms of F1 measure compared to the standard concept vector space.

Although this research utilised the structure of ontology to identify the importance of concepts, it main weakness is that only the number of links (i.e. relations) was considered. In other words, all links were given the same importance regardless of their semantic meaning. In contrast to this approach the SAHBN model proposed in this thesis was designed to exploit the implicit semantics of the "is-a" ontological relation.

In the same way, [56] used the Unified Medical Language System (UMLS) ontology to construct a semantically aware classifier. UMLS ontology was used in two different phases, before and after the classifier training stage. In the first phase, before the training stage, the semantic kernel method was used to transfer the traditional bag of words into a bag of concepts. Consequently, the enriching vectors method was applied in the second phase after the training stage and before the prediction stage to overcome the shortage of the lexical matching.

The ultimate aim of Albitar et al.'s [56] research was to investigate the advantages of two semantic enrichment strategies, semantic kernel and enriching vectors, on conceptualised text classification. The authors concluded that the classifier's performance depends on the semantic similarity measure used in the enrichment and prediction phases. Perhaps the most important advantage of this method is the integration of ontology at two different stages of the mining process, which highlighted the increasing role of ontology. Hence, the SAHBN model proposed in this thesis used ontology, not only to define the relationship between attributes but also to guide the classifier structure construction process and assist in the conditional probability calculation stage.

In a slightly different way, the UMLS terminology was used by [57] as a basis to develop ontology that defines an entity for each word and finds the relations and rules between words. The authors argued that adding extra knowledge in the form of ontology makes it possible to find expressions that facilitate the epilepsy classification process. The model proposed by [57] used a java-based annotation engine, which makes use of ontology expressions such as anatomy or events that are necessary for classification and then use

this extra knowledge to select the most relevant features. A real data set consisting of 18 anonymous records was used to test the proposed model.
Although the testing data set was small, the authors asserted that the proposed approach demonstrated a good performance. However, the authors did not offer a clear explanation about the exact rules and relations extracted from ontology and how they empowered the entity recognition stage.

It is reported that ontology can be used not only in the classifier training and prediction stages, but also in the pre-processing steps, such as dimension reduction, and post-processing steps, such as result interpretations. For example, [58] used the Medical Section Head (MeSH) ontology during the attribute-selection stage to transform a traditional bag of words into a bag of concepts. The produced bag of concepts covers not only the extracted words but also their hyponyms. The authors' ultimate goal was to reduce the dimension of the vector and enrich its contents. This goal was met in the following ways:

1. Disambiguation mapping strategies were used to map each term onto ontology concepts.
2. Hyponyms were added to enrich the generated vector.

Elberrichi et al. [58] claimed that the proposed ontology-based bag of concepts approach led to a 30% better performance compared to the standard bag of words' method. Even though natural language disambiguation has a complex nature, since there is no definite interpretation of the meaning of words, the findings of this research reported that ontological semantic knowledge has improved the classification process by using more comprehensive ontology concepts as prediction attributes instead of using simple words.

Likewise, the SAHBN model presented in this thesis represented the prediction attributes in the form of ontology concepts. However, unlike the work of Elberrichi et al. [58], our research used ontology not only to represent the prediction attributes but also ontological semantic knowledge to construct the classification model structure and assist the process of quantifying the strength of relations between various concepts.

Similarly, [59] argued that mapping the original text features into a node in the MeSH ontology using the Least-Max-Cover approach generated a better performance. The proposed concept-based, dimension-reduction method was compared with the state-of-the-art feature selection, such as IG, CHI, One-R and LARS, and the obtained results

suggested a significant improvement in the F-measure value (3.5%) and recall (5.25%), on average.

The model proposed by [59] was tested using the most recent 1,000 abstracts of 12 journals in the MEDLINE database. The journals were selected and organised into different groups in such a way as to ensure classification accuracy. Eventually, the C4.5 classification algorithm and cross-validation technique were applied to evaluate the overall performance. However, the research of J. Wang et al. [59] would be more robust if it had included other classification algorithms such as decision trees and support vector machines, among other things. Additionally, one question that needs to be asked is why the authors did not use accuracy to measure the quality of their classification model, relying instead on the precision, recall and F1-measure.

Likewise, [60] reported that using ontology to direct and filter the results obtained by the term frequency-inverse document frequency (tf-idf) classification algorithm led to 3.125% more subjects being correctly classified. The authors used ontology in the second and third stages of the proposed process. A program has been developed to direct the broad classification of results into ontology classes. Consequently, a set of candidate classes for each input document was generated in the second stage. Finally, in the third stage candidate ontology classes were further purified to obtain one ontology class for each input document.

Perhaps the most serious weakness of the work proposed by [60] is that the developed ontology is very limited in terms of size and applicability. Unlike this study, the SAHBN model proposed in this thesis utilised the GO, which is known to be a comprehensive and high-quality source of ontological knowledge. Additionally, the GO is subject to regular maintenance and upgrading by various research groups.

In a different way, [61] used ontology to predefine the classification semantic concepts. Hence, it could be considered a step in the pre-processing stage. Ontology was developed based on the information extracted from the reviews and requirements of travellers. The proposed model consists of two main components, subject classifier and sentiment classifier, with the former used to determine the ontology class of the subject that appeared in the review text. The latter, meanwhile, was utilised to assess the sentiment value of the text with regard to the identified subject.

Consequently, various machine learning algorithms were used for each classifier. J48, Naïve Bayes and LibSVM were applied by the subject classifier, and the Linear Regression, LibSVM and SMOreg models were used by the sentiment classifier. Eventually, the proposed model was evaluated using a manually created training data set, and the authors concluded that the developed model accurately classified the text subject into predefined classes and evaluated the emotional expression presented in the text.

One major drawback of this approach is that ontology was used to predefine the possible classification classes. Hence, the quality of the classifier is totally dependent upon the quality of the developed ontology. Unlike this approach, our model did not restrict the possible classification classes based on the created ontology concepts.

In contrast to the work presented in [58]–[61], this thesis proposed a model that not only used ontology in the pre-processing or post-processing stages of the classification process but also contributed to the core classification stages, as will be discussed in the third chapter.

It is a commonly held view that the quality of the text classification algorithm relies heavily on a well-classified training corpus. However, in a real-life scenario very often the training corpus is scarce or even not available. Hence, many scholars have proposed the use of background knowledge in the form of ontology as a classifier to eliminate the need for a training data set. This is exemplified in the work undertaken by [62].

The authors of [62] converted the English language Wikipedia to ontology that divides news documents into topics of interest. The main argument of the proposed work was that the semantic similarity between the news text document and a fraction of the developed news ontology could be used for news document classification.

In the first step a group of words extracted from the news text was matched to a set of concepts in the ontology. A disambiguation technique was applied in the second step to identify more accurate entities during the matching process. Furthermore, a sub-graph of the news ontology was extracted after eliminating the less important connections. Finally, the semantic similarity measure between the news document and the news ontology was calculated to identify the class of the news.

Consequently, the model proposed by [62] was evaluated using three data sets obtained from Reuters RSS feed and the micro averaged precision (MAP) calculated. Overall, the

authors claimed that the proposed ontology-based news classification methods produced very good results.

Similar to the work presented by [61], the authors of [62] used ontology to predefine the classification classes. Hence, they restricted the application domain of the proposed models into specific cases. Unlike these studies, our approach used semantic knowledge in the form of ontology to enrich the classification process and did not restrict the domain of application.

A more recent example of a similar approach was proposed by [63] when the authors developed ontology in the domain of occupational health and security application in the oil and gas industry. The developed ontology was used as a classifier to detect accidents from unstructured text and to eliminate the need for a training data set.

The model proposed by [63] started with the standard text pre-processing tasks such as lemmatisation, stemming and stop-word removal; then a customised version of the OpenOffice Brazilian Portuguese thesaurus was used to locate words that appeared in the text and to match them to the developed ontology. The degree of matching was calculated by the crawling algorithm, which counts the number of jumps needed to get from the word to the term using the thesaurus. A smaller number of jumps indicates a higher similarity.

Consequently, the experimental results showed that the proposed algorithm outperformed the state-of-the-art machine learning approach. Although the authors of [63] exploited the structure of the developed ontology to calculate similarity between words and ontology concepts, the main drawback of this approach is that different types of relation between ontology were considered to have the same weight regardless of their semantic meaning. Hence, the notion of integrating the explicit semantics of ontology has not been fully met.

Another significant aspect of ontology-based data classification is the use of ontology, not only to integrate domain knowledge in the mining process, but also to perform reasoning and to predict the target class. An example of this is the study carried out by [64], in which the authors attempted to merge the output of various types of decision tree and ontology with the expectation that the proposed model would improve classification accuracy.

The spambase database and Weka package were used to create three different types of decision tree, namely j48, AD and LAD. The created decision trees were considered to be a kind of ontology and represented in RDF format. Finally, the Jena reasoner was used to

read the RDF file to create ontology and to check whether or not the test email was spam. In total, 500 test emails were used to evaluate the performance of the proposed model and to compare it with the output of the SVM and Naïve Bayesian classifier.
Consequently, the authors of [64] claimed that their model was more successful than other methods used to detect spam and valid emails. However, a key problem with this approach is that converting a domain-specific database into an ontology using Weka decision trees conflicts with the basic principle of ontology development rules, which state that ontology must reflect a common knowledge that could be shared unambiguously between various domains [9].

In a slightly different way, [65] developed an ontology-based classifier for tumour staging using the tumour-node-metastasis ontology (TNM-O) classification system. The first step in the proposed model is to convert the pathological information into RDF individuals, which matches the concepts presented in the TNM-O ontology. Then, the HermiT description logic reasoner was used to classify the created individuals. A total of 382 entries in the pathological data set was used to test the proposed model and to compare the results with the experts' decisions. Accordingly, the authors concluded that their model accurately classified all the 382 entries and helped to detect and explain the inconsistencies between the expert and automatic classifications.

Perhaps the most serious drawback of the work presented by [64] and [65] is that the classification was made based on an ontology reasoner that was developed based on description logic. Description logic has been defined as a decidable fragment of first-order logic. In other words, an instance either belongs or doesn't belong to a concept [66]. Hence, the uncertainty factor has not been taken into consideration, which goes against the nature of many real-life situations, especially in the health-care sector. In contrast to these methods, the model implemented in our thesis overcomes this drawback by integrating ontology and uncertainty in the form of the Hierarchal Bayesian Network (HBN).

Having discussed how ontology can be used in text data classification, the final part of this section explains examples of using ontology in image and relational data classification. This is exemplified in the work undertaken by [67], in which the authors represented the experts' knowledge on how to identify the family of a protein in the form of ontology. Then they used the developed ontology and description logic reasoner to assign a new protein instance into the protein family.

The work presented by [67] did not introduce any new bioinformatics techniques or algorithms to discover sequence features. Rather, it enhanced the existing tools used by ontology and proposed an approach to automate the protein classification process. In other words, the structure and reasoning power of ontology were exploited to formally represent the phosphate protein family classification and to automate the assignment of a new protein instance into the phosphatase family.

Accordingly, the proposed model was tested on the human and Aspergillus fumigatus genomes, and the authors claimed that their ontology-based automatic protein classification approach matched and sometimes outperformed the experts' annotation process. However, this study would have been more interesting if it had included the notion of uncertainty in the reasoning stage and not relied solely on the description logic reasoner.

Another possible example would be the research carried by [68], in which the existing medical ontology was enriched by domain ontologies to represent specific hospital requirements. Ontology played two different roles:

1. Ontology was used to explain the semantic meaning of the time point in the medical data. For example, the time associated with the white blood cell (WBC) count process did not explicitly indicate whether it was the time when the blood sample was taken or when it was analysed in the laboratory. Hence, ontology was used to overcome this challenge.

2. The uncertainty of the prediction was represented in the form of ontology. In the medical domain, the clinical decision support system (CDSS) is designed to help junior or non-expert clinicians to make more informed decisions. Hence, it is necessary to convey the accuracy of the prediction to the medical staff. In this work ontology was used to represent this information. For instance, the probability of a patient with low WBC having sepsis is 17%.

As we have seen, this study used ontology not only to enrich the semantic meaning of the data but also to convey the uncertainty of the information to the end-users. However, it seems that the study focused on the time dimension of the data and overlooked other semantic aspects in the data.

In a different way, [69] developed a seizure ontology to represent human knowledge in the field of detection of types of epilepsy. The core of the constructed ontology is the

interactions between epileptic symptoms. Thus, the structure of the seizure ontology and the Ck matrix were used to quantify the degree of association between each pair of symptoms. The number of arcs in seizure ontology was used to measure the distance between symptoms. Consequently, the mean for all the matrices was computed and seizure assigned to epileptic type with the minimum distance from the mean.

Consequently, the authors tested the proposed ontology-based classification model using a data set consisting of 129 patients, and compared the results with 7 clinicians with different degrees of expertise; they concluded that their model represented an essential step in performing automatic epilepsy classification and that it helps in the identification of symptoms that are necessary for the classification process. However, the findings of this research would be more useful if a larger data set had been used. Additionally, it seems that the authors only relied on the structure of the developed ontology by utilising the number of arcs to classify the epileptic type. Hence, all arcs were given the same weight, regardless of their semantic meaning.

As indicated previously, ontology has been used in image data classification. Hence, the following part of this section gives examples of ontology-based image data classification. This can be seen in the work presented by [70], in which the authors developed an ontology model to classify the objects in the images collected by the Airborne Laser Scanner (ALS) into various building types, such as “residential/small buildings”, “apartment buildings” and “industrial and factory buildings”.

The first step of the model proposed by [70] sought to extract the features of different building types from text data using the ensemble random forest (RF) algorithm. Then ontology was constructed based on the information generated by the RF algorithm. Finally, the FaCT++ description logic reasoner was used to classify an object in a given image.

Consequently, the proposed model was tested using a data set recorded by the Trimble Harrier 68i system, which covers 1.1 square kilometres in the area of Biberach and der Riss town in Germany. The authors claimed that a 97.7% F-Measure was obtained for the residential/small building type and a lower result was achieved for other building types, such as industrial and factory buildings, with an F-Measrue of 60% and 51% respectively. However, the main weakness of the study was its failure to integrate the uncertainty

concept into the reasoning stage. Since the authors used the FaCT++ reasoner, which is built based on description logic, it did not accommodate uncertainty.

Likewise, [71] proposed an ontology-based classification model that was designed to help radiologists in staging cancer. The proposed model utilised Semantic Web reasoning capabilities, ontologies and semantic image annotation. It includes, but is not limited to, the following steps:

a. Annotating the medical image using the ePAD tool.
b. Converting the annotation result into OWL instance format.
c. Developing or extending an existing ontology to represent the tumour staging criteria.
d. Classifying new instances using the OWL reasoner. This is done by integrating the information represented by the semantic image annotation and cancer staging ontology.

Similar to the work presented by [70], the authors of [71] used Semantic Web reasoning capabilities. Hence, the uncertainty dimension was not included in the reasoning phase.

To summarise, to the best of our knowledge, and based on the findings extracted from reviewing state-of-the-art literature in the ontology-based data classification field, it can be concluded that existing studies have the following drawbacks:

a. The majority of the existing work has investigated the advantages of integrating ontology and text data classification. Meanwhile, the relational data set has received less attention.

b. Some studies have relied heavily on ontology and the description logic reasoner to classify the test data. Hence, the notion of uncertainty, which is characteristic of many real-life situations, has been ignored and not included in the proposed model.

c. Some studies have only considered the graphical structure of ontology and not the explicit semantic meaning of ontology. Thus, it has been assumed that all arcs in the ontology structure have the same semantic weight.

d. Most of the available approaches are case-specific, which can only be applied to a specific data set in a certain domain. Hence, it could be concluded that there was no attempt to propose a general framework that systematically integrates ontology in the classification process.

e. Some of the proposed models used ontology to predefine the relation between the classification classes. Thus, they further restricted the application domains to those classes included in the ontology instead of defining the semantic relation between attributes included in the classification problem and widening the range of the application domains.

f. Some of the existing studies only investigated the advantages of ontology in the pre-processing/post-processing stages of the classification process and did not investigate the possibility of integrating ontology in the core phases of the classification process.
g. Some of the studies misunderstood the conceptualising nature of ontology, which aims to provide shared knowledge about a specific domain, which can be shared unambiguously.

Hence, the model proposed in this thesis aims to overcome the above drawbacks and to suggest a classification model that meets the following criteria:

a. Merges ontology and the concept of uncertainty by developing an ontology-based hierarchical Bayesian Network.

b. Introduces a more general framework that can be applied to different domains and systematically integrates ontology to the classification process.

c. Exploits the semantic meaning of ontology relations and not just the structure of ontology.

d. Increases the roles of ontology to contribute to different stages of the classification process.

e. Applies the proposed model to a relational data set in order to highlight the advantages of ontology in relational database classification in the biomedical domain.

2.5 Integrating Ontology and Bayesian Network

The concept of the Semantic Web (SW) introduced ontology as a means to represent and share knowledge. Although ontology has played an important role in SW technology, it has serious limitations, which restrict its use in real-life applications. Unlike many real-life problems, which are known for their uncertainty and vagueness, ontology is constructed based on description logic, which is known for its deterministic nature. Hence, much effort has been spent integrating ontology and uncertainty [72], [73].

The aim of this section is to shed some light on the notion of uncertain Bayesian ontology; a good starting point would be the overview conducted by [74], in which the integration of ontology and the Bayesian Network, and their applications in different domains, were summarised. For instance, the BayesOWL and OMEN approaches were introduced as ontology mapping techniques. These approaches are explained in the following subsections.

1. BayesOWL: The ultimate aim of the BayesOWL framework is to enrich traditional ontology by adding the capability of uncertainty. It consists of a set of construction

rules that convert ontology into a Bayesian Network directed a cyclic graph (DAG), which preserves the semantics of the original ontology. Furthermore, the iterative proportional fitting procedure (IPFP) is used to modify the values of the conditional probability tables (CPTs) attached to the variables in the constructed BayesOWL so that they can meet certain constraints [74]. The construction of BayesOWL has two main phases. In the first phase the BayesOLW graph structure is constructed and the attached conditional probability tables (CPTs) are initialised with default values. Then, in the second phase the given constraints are integrated [75], [76]. It was reported by [77] that the framework of BayesOWL has been developed and published as an open source java API, which contains the following components:

a. **Taxonomy parser (T-Parser):** this component reads an ontology file, which represents the conceptual model of the targeted domain and generates three arrays, namely, 1) nodes name array, 2) nodes type array and 3) parent–child relationships array.

b. **Bayesian Network structure constructor:** this component converts the three arrays generated by the first component into a BN graph and initialises the CPTs with default values.

c. **Probability parser (P-Parser):** the functionality of this component is to extract the probability constraints given in the form of a probability file and then generate a constraints array, which is used by the next component.

d. **CPT's constructor:** this component uses the IPFP to modify the values in the CPTs so they become consistent with the given constraints.

In BayesOWL, the process of constructing a BN DAG from the given ontology file is governed by a set of rules. The conventions underpinning these rules are summarised as follows [74], [78].

i. Every primitive or defined class is mapped onto a binary variable.

ii. Each parent superclass is connected with its child subclass by an arc.

iii. For each concept class, C, defined as the intersection of a set of classes, C_i = (C_1, ….. , Cn), a subnet is created in such a way that there is a link from each class in the set, C_i, towards the class, C. Furthermore, there is a link

from C and each class in the set Ci towards a logical node classed LNodeIntersection. Figure 2-10 (below) depicts the creation of the intersection subnet.

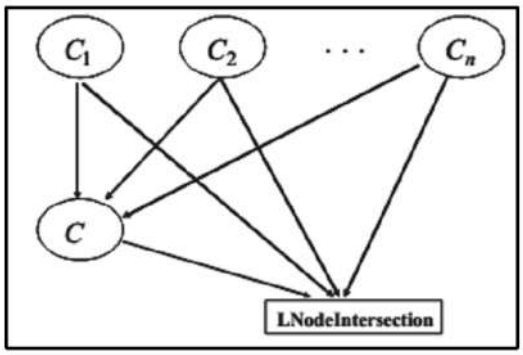


Figure 2-10 OWL:IntersectionOf [74]

iv. For each concept class, C, defined as the union of classes, Ci = (C1, ……, Cn), a subnet is created in such a way that there is a link from C to each class in the set Ci. Furthermore, there is a link from C and each class in the set Ci towards a logical node called LNodeUnion. Figure 2-11 (below) depicts the creation of the union subnet.

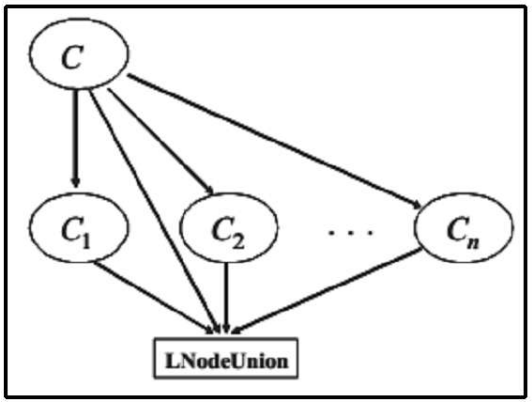


Figure 2-11 OWL:UnionOf [74]

v. For each two concept classes, C1 and C2, defined as a complement of, equivalent to, disjoint with, each other, a logical node, LNodeComplement, LNodeEquivalent, LNodeDisjoint, is created, which takes two input links from C1 and C2. Figure 2-12 (below) depicts the creation process for LNodeComplement, LNodeEquivalent and LNodeDisjoint respectively.

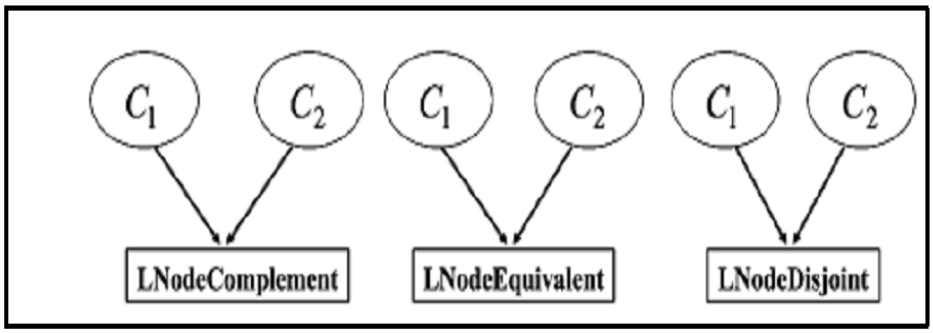


Figure 2-12 OWL:complementOf, OWL:equivalentClass, OWL:disjoitnWith [74]

It could be concluded that the generated BayesOWL contains two types of node, namely, regular nodes, which represent classes, and logical nodes, which show the logical relation among classes. Hence, the combination of these two types of node forms the structure of the BayesOWL network.

Having discussed how to construct the structure of BayesOWL, the following paragraph addresses ways to calculate probability. As indicated previously, BayesOWL contains two types of node, logical and regular; thus, the CPT for each type is calculated as follows [74], [77], [78].

a. **CPTs calculation for logical nodes:** the CPT for each logical node is determined by its logical relation. BayesOWL accommodates five different logical relations, and their CPTs are calculated accordingly. The following subsection explains the CPT calculation process for each logical node.

   i. **Complement of:** the complement relation between two concept classes, C1 and C2, is true if $c1\overline{c2} \vee \overline{c1}c2$, which generates the CPT in Table 2-1 (below).

Table 2-1 CPT of Lnode Complement [74]

C1	C2	True	False
True	True	0	1
True	False	1	0
False	True	1	0
False	False	0	1

   ii. **Disjoint with:** the disjoint with relation between two concept classes, C1 and C2, is true, if $c1\overline{c2} \vee \overline{c1}c2 \vee \overline{c1c2}$, which results in the CPT in Table 2-2.

Table 2-2 CPT of LNode Disjoint [74]

C1	C2	True	False
True	True	0	1
True	False	1	0
False	True	1	0
False	False	1	0

iii. **Equivalent to:** the equivalent to relation between two concept classes, C1 and C2, is true if $c1c2 \vee \overline{c1c2}$, which leads to the CPT in Table 2-3.

Table 2-3 CPT of LNode Equivalent [74]

C1	C2	True	False
True	True	1	0
True	False	0	1
False	True	0	1
False	False	1	0

iv. **Intersection of:** the relation that C is the intersection of C1 and C2 is true if $cc1c2 \vee \overline{cc1}c2 \vee \overline{c}c1\overline{c2} \vee \overline{cc1c2}$, which is expressed by the CPT in Table 2-4.

Table 2-4 CPT of LNode Intersection [74]

C	C1	C2	True	False
True	True	True	1	0
True	True	False	0	1
True	False	True	0	1
True	False	False	1	0
False	True	True	0	1
False	True	False	1	0
False	False	True	0	1
False	False	False	1	0

v. **Union of:** the relation that C is the union of C1 and C2 is true if $cc1c2 \vee c\overline{c1}c2 \vee cc1\overline{c2} \vee \overline{cc1c2}$, which is represented by the CPT in Table 2-5.

Table 2-5 CPT of LNode Union [74]

C	C1	C2	True	False
True	True	True	1	0

True	True	False	0	1
True	False	True	1	0
True	False	False	0	1
False	True	True	1	0
False	True	False	0	1
False	False	True	0	1
False	False	False	1	0

b. **CPTs calculation for regular nodes:** the CPTs for regular nodes are computed by applying the Bayesian theorem, as follows. $P(C|\pi_c)$, where C is the concept for the regular node and π_c is the set of its parents. The $P(C|\pi_c) = 0$ if any of its parents is false. Hence, the probability for any child concept, C, is calculated only when all of its parents are in true status. This scenario is denoted as $P(C|\pi_C^+)$, where π_C^+ represents the set of parent classes in true status. This method is used to calculate probability when data is available. Otherwise, a default value (0.5) is assigned based on equation 2.1 (below) [74], [75], [79].

$$P(C = \text{True}|\ \pi\ c+) = P(C = \text{False}|\ \pi\ c+) = 0.5 \quad \textbf{2.1}$$

Similar to the BayesOWL approach, this thesis proposes a model that represents each ontology concept as a binary variable in the BN structure and connects it to its superclass using the parent–child relation. However, BayesOWL has further augmented the BN structure by creating logical nodes corresponding to five logical relations. Hence, the future work of this thesis will investigate the advantages of integrating the logical nodes in the ontology-based hierarchical Bayesian Network. Additionally, BayesOWL used the IPFP to integrate the probabilistic information in the form of a constraint. Alternatively, our approach utilised the observed data and the maximum a posteriori estimation (MAP) parameter estimation method to compute the probabilistic information.

2. **OMEN:** an ontology mapping enhancer tool that aims to improve the ontology mapping process by amalgamating the ontology structure and Bayesian probability. The structure of ontology is exploited to construct a BN, which represents the influences and inter-relationships between ontology. Hence, nodes that are neighbours with the matched node are included in the matching process [80].

The creation of the OMEN BN structure is started by constructing m X n nodes, which represent all the possible matches between nodes across the source ontologies. Then, the semantic relation described in the given ontology is used to create links between nodes, and the down-flow edges generation process is implemented as follows: for each concept, C_1 and C_2, which are the children of P_1 and P_2, a directed edge between two nodes is added to the graph, which represents the match between (P_1 & P_2) and (C_1 & C_2) respectively. Finally, nodes that have no parents or children are removed to minimise the number of nodes.

With respect to the conditional probability tables (CPTs), the OMEN framework provides external functions to compute and produce the required CPTs. These functions are invoked based on per of nodes and their neighbour nodes, as explained in the source ontologies. Additionally, OMEN designed internal functions based on meta-rules to produce the CPTs.

In summary, OMEN utilised source ontologies, the output of other mapping tools and a set of meta-rules probability distribution to generate more accurate ontology mapping results [80]–[82]. Perhaps the most serious weakness of this model is that it does not work independently. In other words, it requires the output of other mapping tools and the initial probability distribution as input to perform the mapping process.

The following paragraphs explain the applications of BN reasoning on ontology; a good example is the work undertaken by [83], in which an extended form of BN is constructed based on the semantic relation of TBox and ontology, and CPTs are computed from the instances on the associated ABox. The authors claimed that the conditional dependencies for the BN structure can be extracted from the ontology semantic structure and the instance frequencies provide the required probability distribution. Hence, ontology can be used to perform probabilistic reasoning without any modification. The proposed method consists of three steps:

1) Convert the ontology TBox into a layered BN,
2) Compute prior and conditional probability based on the instance frequencies in the ontology ABox,
3) Use the structure of the generated layered BN to implement probabilistic reasoning.

The authors of [83] concluded that the developed layered BN model could cater for the "is-a" and object properties in the reasoning process. However, it lacks other relations

such as disjoint and cardinality, among other things [81]. Similar to the study presented by [83], the framework proposed in this thesis exploited the "is-a" relation in the TBox to construct the BN structure. However, each node in the BN structure created by our model represents an aggregation of its subclasses, while each node in the structure created by [83] is a high-level node, which in turn represents a BN. Figure 2-13 (below) depicts the structure of the two-level BN model proposed by [83].

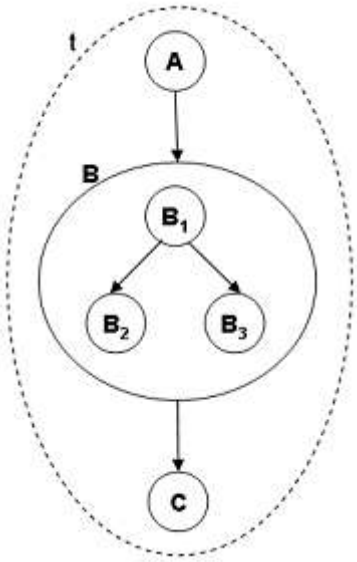


Figure 2-13 2LBN Structure [83]

According to [81], another significant domain of BN and ontology integration is the semi-automated construction of BN. Many scholars hold the view that the construction of complex BN is naturally difficult. Hence, an external knowledge base in the form of ontology could be used to help identify the variables of interest and their effect on one another. Additionally, ontology can assist in the process of probability calculation. An example of this is the study carried out by [84], in which the authors exploited the knowledge and inference capabilities of ontology to automate the construction of a BN. Ontology was used in four different steps, as follows:

a. **BN variable identification:** the aim of this step is to identify the set of concepts that exhibit a causal relationship, which can contribute to the construction of the BN structure. The concept selection process was built based on the assumption that there is an existing or developed ontology to represent the domain of interest. So, in order to identify the concepts that contribute to the construction of the BN, the authors proposed BN ontology and linked it to the original domain ontology. The root concept of the BN ontology is the BN node, and all BN variables were

initiated from the BN node root node. Finally, the combined domain ontology and BN ontology were utilised to generate the structure of the BN.

b. **BN node properties extractions:** the BN node attributes were extracted from the combined ontology created in the previous step. The concept of attribute constraints was used to restrict the values, which can be assigned to the newly created BN nodes. For example, constraints such as "hasValue" were used to specify the values a property can assume.

c. **Finding parent nodes:** the ontology reasoner was applied over the combined ontology to extract the BN nodes and assign their properties. An instance was created for each subclass of the BehaviourModelNode. The BehaviourModelNode was used to define the specifications of BN nodes in a certain application domain. Additionally, the properties were represented by nodes and assigned the appropriate values. Finally, links between nodes were created based on the domain rules, which specify the relations between concepts in BN.

d. **Calculation of the Conditional Probability Table (CPTs):** the proposed approach used existing parameter estimation methods. However, ontology semantic knowledge was used to estimate the initial probability distribution for some nodes. For instance, the CPT of the child concept can be encoded to reflect the effect of the deterministic relation with its parent.

A serious weakness of this approach, however, is that it relies heavily on the experts' knowledge to identify the concepts and relations to construct the BN structure. Additionally, it requires the extension of the domain ontology to include specific ontology concepts and properties to define the structure of the BN. Unlike this method, our approach sought to automate the creation process of the BN without any extension to the existing domain ontology or modification to its structure.

Another example of the semi-automated construction of BN is the work undertaken by [85], in which uncertainty was merged with the clinical practice guidelines (CPG) to generate an ontology-based BN, which helps doctors to estimate the risk of uncertain actions in the medical process. The proposed approach converted the CPG graph into ontology-based BN, which contains uncertainty features and provides an algorithm to construct the CPTs. Hence, medical staff can upload their observation as evidence to the developed Bayesian model and reason out the probabilities of the targeted actions.

Another attempt at an ontology-based BN construction was made by [86], in which ontology was used to:

1) Create the BN nodes based on existing ontology concepts.
2) The semantic relations between ontology were used to define the links between nodes in the BN.
3) Ontological semantic knowledge was utilized to facilitate the conditional probability calculation.

The proposed approach was tested in the area of threat detection using security ontology as the underpinning knowledge base. Accordingly, the authors claimed that their approach achieved the following:

1) Enabled the semi-automatic creation of a BN based on existing ontology.

2) Reduced the complexity of the created BN using the high-level concepts and sub-concept relation in ontology.

3) An uncomplicated maintenance process to the BN's underlying knowledge in the form of ontology.

However, the main weakness of this study is that expert intervention is required not only to select the concepts that best describe the problem but also to interpret the semantic meaning of the selected relation.

The final class of ontology and BN integration, classified by is the PR-OWL, which is a probabilistic extension of the existing OWL language. The Multi-Entity Bayesian Networks (MEBN) logic forms the logical base of the PR-OWL approach, which aims to merge first-order logic with the Bayesian Network. The ultimate aim of MEBN is to represent first-order logic as probabilistic fragments (MFrags) of BN, which can be combined to form a graphical probability model. MFrags are used to encode the uncertainty relations among a group of logically related hypotheses. In other words, MFrags represent a group of nodes in the BN that are logically connected [87], [88]. Despite its many advantages, PR-OWL is not included in the W3C standards. Hence, the existing ontology did not comply with the PR-OWL syntax, which restricted its applicability.

Having discussed the approaches presented in [81]'s survey, the following paragraphs discuss other methods that have been either overlooked by Larik et. al.'s survey or

introduced recently. This can be seen in the work implemented by [89], in which the domain of interest was represented in the form of ontology, which was then converted into BN following a predefined sequence of steps. Consequently, the associated probability distribution was calculated using the existing parameter estimation method.
The construction of the BN structure is implemented in two phases. In the first phase, the concepts and relations related to the knowledge of interest are selected, while the rest of the concepts in the ontology serve as background knowledge. It is notable that representing the domain of interest using concepts and the associated attributes provides a solid base from which to convert these attributes into statistical variables, which are represented as nodes in the BN structure. Different possible combinations of the selected concepts and attributes can form different network structures. Hence, careful selection must be followed to avoid redundancy.

In the second phase, the structure produced in the first phase is further optimised and improved. The authors tested their approach in the domain of oesophageal cancer and claimed that the proposed approach helped to address complex factors involved in the modelling process. Additionally, they compared the newly generated structure with the original network, which explains some ambiguous decisions made in the original network construction process. However, the main weakness of the study was its failure to eliminate the need for manual intervention to select the required concepts and relations related to the problem under investigation.

In similar way, [90] proposed a model to incorporate the BN and ontology. The proposed approach consists of the following steps:

1) Extend the existing OWL to encode probabilistic information for concepts and properties.
2) The probabilistically annotated OWL is converted into a BN using a set of translation rules.
3) The CPT for each node in the BN is constructed.

For the first step the authors considered the probability information as a kind of resource and proposed three different types of "WOL:Class", namely, "PriorProbObj", "CondProbObjT" and "CondProbObjF". The prior probability in the form of P(A) is represented as an instance of "PriorProbObj" and has two properties, "hasVAriable" and "hasProbValue". Additionally, conditional probability in the form of P(A|B) and P(A|$\overline{B}$) is

defined as instances of “CondprobObjT” and “CondProbObjF” respectively. All conditional probability has the following properties: “hasCondition”, “hasVariable” and “hasProbValue”. While the type of “hasCondition” and “hasVariable” are object properties, the type of “hasProbValues” is a datatype property.

Moving on to the second step, which is the structural translation rules, the core of these rules is that each subject or object class and object properties are converted into binary nodes in the BN. Additionally, an arc is created between nodes if there is a predicted link to the corresponding classes.

In the third and final step the associated CPTs are determined. In many cases the semantic and logical relations are utilised to specify the values of the CPTs. For example, the P(Man|Human,Male) = 1, which means that a man is an intersection of human and male, so the probability that man equals true, given that human and male are true, is one.

Although the authors did not test the proposed approach in a real-life scenario, they argued that the proposed OWL extension and BN translation rules are consistent with OWL semantics and the BN inference procedure. Hence, standard probability inference algorithms such as belief propagation or a junction tree can be implemented in a real-life situation. However, the key problem with this approach is that it requires the extension of the existing ontology to include the probability and BN structure information in the form of resources. Although the authors explained how to translate the BN-related resources included in the ontology, they did not comment on the sources of this information. It seems that the authors relied on human experts to provide this information; hence, manual intervention is essential to this approach.

Another example of BN and ontology integration is the “OntoBayes” model presented by [91], in which an ontology-driven uncertainty approach was proposed. The proposed approach used probabilistic information and dependency relationships of the Bayesian Network to annotate the underpinning ontology model in such a way that preserved the advantages of both. The first step in the ontology and BN integration process was ontology probabilistic extension. This was done by defining three classes, namely, “PriorProb”,”condProb” and “FullProbDist”. A “ProbValue” datatype property was defined for the first two classes and used to represent the prior and conditional probability, respectively. Additionally, “hasPrior” and “hasCond”, two disjoint object properties, were used to describe the full disjoint probability distribution class.

Turning now to the second step, which is defining the dependency relations between nodes, the authors proposed a property element called <rdfs:dependsOn> to encode the dependency information in OWL ontology. Dependency has been defined as pair X → Y, where X and Y are either a data property or an object property and read as X depends on Y. Thus, the variables in the OntoBayes model consist only of datatype and object property.

Finally, in the third step the structure of the OntoBayes was constructed by extracting all dependency triples in the form of object, predicate and subject, where the predicate is the primitive <rdfs:dependesOn>. Then, when dependency triples were merged nodes with the same identifier were aggregated to a single node. The OntoBayes approach was applied to the insurance and natural disaster management field, and the authors claimed that their approach conserves the powerful expression capability of ontology and the uncertain reasoning capability of BN.

Similar to the work presented by [90], the authors of [91] extended the existing syntax of ontology to include the information related to the BN probabilistic dependency. However, the authors of [91] introduced the <rdfs:dependsOn> property, which explicitly defines the dependency relation between variables in the BN. Hence, both approaches have the same drawbacks, in that they require the extension of the existing ontology and rely on human experts to define the dependency relation instead of deducing these relations from the semantic knowledge of ontology.

Another attempt to amalgamate ontology and BN is exemplified in the work undertaken by [92], in which the authors investigated the advantages of exploiting ontology semantic knowledge in the construction process of the Object-oriented Bayesian Network (OOBN). The proposed approach suggested the use of a set of mapping rules, which aims to compile an ontology graphical representation into OOBN. The ontology into OOBN compilation process consists of three main steps, namely, initialisation, discovery and closing.

In the initialisation step all concepts are undiscovered. Then, for each concept the OOBN class is generated. Additionally, the actions that must be taken at each detected concept are determined in the discovery step. The actions taken define input, internal and output sets for each class. Finally, for each root vertex that has no predecessor, an instance of this vertex class is added to the OOBN.

The authors argued that the proposed ontology to the OOBN mapping approach established a new connection between the two disciplines and made use of the semantic enrichment of ontology. Furthermore, it overcomes the limitations of using standard BN. Although the authors of [92] argued that the problem of constructing a BN is an NP hard, and they intended to tackle this problem by exploiting ontology semantic knowledge, they suggested the use of the depth first search (DFS) to traverse the entire ontology. Hence, it seems that this study does not take into account the size of the ontology. In contrast to this approach, the SAHBN model only extracts concepts related to the problem under investigation.

Another possible example would be the work carried out by [93], in which ontology was utilised to predefine the structure of the BN. In fact, the authors proposed the use of a two-layer network where the pathologies are represented as nodes in one layer and the symptoms forming the nodes in the other layer. The symptoms layer not only includes observations or symptoms but also represents other elements contributing to the diagnosis process, such as disease proneness, results of medical tests, effects of treatments, multi-pathology conditions and phenomena that evolve over time. Accordingly, the probability of positive diagnosis for a certain disease is conditioned by the probabilities of the presence of symptoms and other related factors. Figure 2-14 (below) depicts the structure of the proposed two-layer BN.

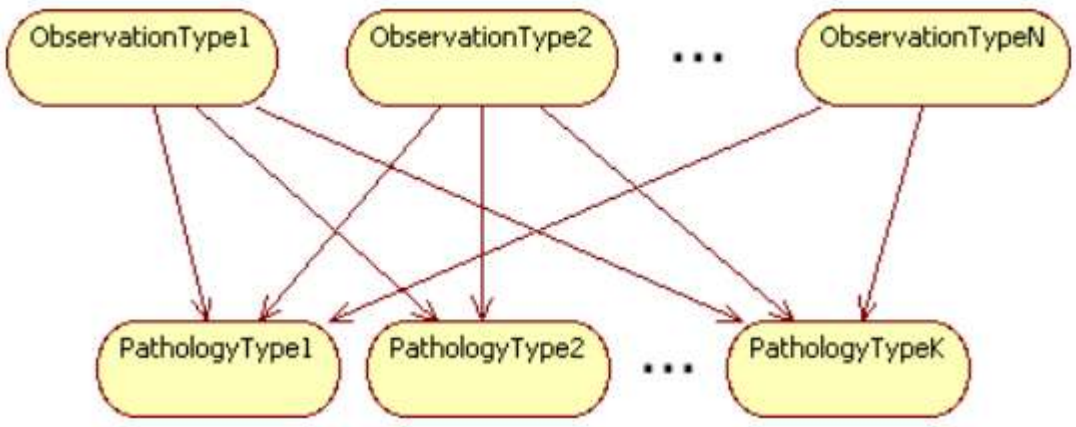


Figure 2-14 The topology of a Bayesian network for medical diagnosis [93]

The authors argued that the main advantage of their approach is that the structure of the BN is predefined; hence, ontology did not encode the information required for BN construction. Furthermore, the backbone knowledge can be easily enriched by specialist users. Such an approach, however, failed to address the point of dynamically exploiting

the explicit knowledge of ontology to create the structure of BN. Instead, they used ontology to predefine the BN structure based on existing knowledge.

Another attempt in the area of ontology-based BN construction is the work undertaken by [94], in which a semi-automatic approach to construct a BN based on existing ontology was proposed. The author claimed that ontology was used to facilitate the following BN construction steps:

1) The identification of BN nodes using ontology classes/individuals.

2) Connecting BN nodes using ontology properties.

3) Calculation of the conditional probability table based on an ontology knowledge base.

4) Findings extracted from an ontology knowledge base are used to augment the constructed BN.

The proposed model consists of the following phases:

a. **Selection of relevant classes, individuals and properties:** an expert in the targeted domain is required to select the classes, individuals and properties that contribute to the task under investigation. Care must be taken when experts are selecting the relevant entities to avoid any redundant edges. For example, transitive relation must not be considered. Hence, if A is affected by B and B is affected by C, only links from B to A and from C to B are added, and links from C to A must not be included. Although intervention of the domain expert is required in the first step, the other steps are conducted automatically based on the output of the first step.

b. **Creation of Bayesian network structure:** the classes and properties selected in the previous step are utilised to automatically construct the BN structure. The construction process adheres to the following rules: 1) the selected classes/individuals are used to create the BN nodes; 2) only the sub-classes of specific selected class are added to the BN graph; 3) for all elected individuals, their asserted individuals are included in the BN graph; 4) the selected value properties are used to assign a numerical value for each class/individual in the BN

graph. Furthermore, value properties are used to compute the CPT of the BN nodes; and, finally, 5) classes connect to their parent class using the link property.

c. **Construction of CPTs:** in order to calculate the CPTs, the entire constricted BN is scanned and nodes that have at least one parent are identified. Then, the weight property is used to define the weight of each node parent. The CPT calculation process for each node is affected by the following factors: 1) the parent state space; 2) the parent's context-specific weight; and 3) the parent distribution function, which determines the state of the targeted node. The complexity of the CPT calculation is reduced by limiting the number of states for nodes with parents to two, while nodes with no parents can have more than two states.

d. **Incorporation of existing knowledge facts:** in order to integrate ontology knowledge into the constructed model, the finding properties such as OWL object property, property assertion, some restrictions, all restrictions, minimum restriction, maximum restriction and value restriction must be considered. However, finding properties can be ignored and the BN can be kept general. If a finding property is selected, a new node must be added to the network and linked to the associated node. The state of the created node has to comply with the parent state of the associated node.

To sum up, the proposed model was used to estimate the threat probability, which uses security ontology as background knowledge. Consequently, the authors argued that their approach outperforms the existing method in the following perspectives:

1) Using existing ontology without any required extension.
2) Integrating facts in the construction process.
3) The constructed CPTs accommodate the semantics ontology aspects.

Perhaps the most notable drawback of this approach is that its quality relies heavily on the domain experts to select the most appropriate classes, individuals and properties. Hence, human expert intervention is essential and must be carefully implemented; otherwise, the consequent steps will be affected.

Another recent example is the work implemented by [95], in which the failure mode and effect analysis (FMEA) information was integrated into industrial design ontology and then the extended ontology was used to construct the structure of the Bayesian diagnostic

networks. The principle of the proposed model is to represent each ontology class as discrete random variables, which are represented as nodes in the BN, and the casual relationship is represented as arcs linking these nodes. The authors proposed an algorithm that searches the ontology model for objects and creates a node in the BN for each encountered object. The proposed algorithm consists of the following steps:

a. A node is created for each *"FailureEffect"* instance.

b. The "*effectOf"* relation is used to represent the conditional dependency between nodes in the BN. Hence, a node is created for each instance of the "*QualityCharacterisitc"* class, which is linked via the "*effectOf"* relation.

c. The "*requires"* relationship is used to identify the "*ProcessSteps", which* contributes to each "*QualityCharacteristic"*. Then, a node is created for each "*ProcessSteps"* and a link to the "*QualityCharacteristic"* node.

d. For every *"EquimentModule(s)"* instance that is linked to the *"ProcessStep"* instance via *"ImplementedBy"* relations, a node is created in the BN and linked to the *"ProcessStep"*.

e. For every instance of *"EquimpmentCharacteristics"* that is linked to *"EquimentModule(s)"* via the *"hasCharacteristics"* relationship, a node is created in the BN and linked to the associated equipment node.

f. Finally, detect and remove any cyclicity in the BN graph.

The proposed approach was tested using a typical industrial assembly system and the authors argued that their approach follows a systematic way to construct the diagnostic models and that it is more accurate and consistent than the model constructed based on human experts. Furthermore, the time and effort required to construct the model is extremely small and integrated into the system design process. Although this approach managed to exploit the implicit knowledge in ontology to identify the dependency relation between nodes in the BN structure, its main limitations are that it is focused on a very domain-specific relation such as "effectOf". Additionally, this approach requires manual intervention to remove cycles in the created BN structure.

To sum up, all the works reviewed so far suffer from the fact that they either require a domain expert's intervention or they are domain-specific, which means they cannot be

generalised to other domains. Additionally, some of them had to extend the standard syntax of existing ontologies, which restrains their applicability. Furthermore, some of the reviewed work used ontology in a static manner to predefine the BN structure. Hence, the model proposed in this thesis sought to automate the integration of ontology and uncertainty in such a way that overcomes some of the previously stated challenges. Our approach aimed to achieve the following objectives:

a. Automate the process of BN and ontology integration and eliminate human intervention.

b. Utilize the implicit knowledge of ontology in BN structure creation.

c. Propose a general integration framework that can be applied to different domains.

Chapter 3 Fundamental Techniques and the Proposed Model

3.1 Introduction

There is some evidence to suggest that the process of BN structure construction and parameter learning can either be done manually by domain experts or learned for data. Although various expert-based and data-based approaches have been studied and applied in different domains, it is believed that both approaches have their disadvantages. The main weaknesses of the expert-based approach are that it is time-consuming, error-prone and costly and experts are not always available. On the other hand, data-based approaches suffer from bias and a lack of training data [94], [96], [97]. Hence, this research investigates the advantages of integrating an external source of knowledge in the form of ontology to improve the overall quality of the Bayesian classification model. Our ultimate aim is to construct a generic BN classifier that integrates the targeted domain knowledge in the form of ontology and seamlessly exploits this knowledge in the mining process. The proposed model is generic in the sense that using different ontology will lead to the integration of different domain knowledge to the proposed BN classification model. Hence, the proposed model can be applied into different domain.

This chapter discusses in detail the proposed SAHBN model and the associated techniques. The first five sections cover the underpinning techniques, which include the following:

1) the gene ontology (GO);
2) the chi-squared test for independency;
3) the Bayesian Network;
4) the Hierarchical Bayesian Network;
5) the parameter estimation methods for the Bayesian Network.

Finally, the last section discusses the proposed SAHBN model in detail.

3.2 Gene Ontology

Gene ontology (GO) is a collaborative effort to construct controlled vocabularies that describe in a consistent manner the roles that genes play in the life of various organisms. Its main objectives are:

1) To assemble a set of structured vocabularies to describe the domain of molecular biology.
2) To use the assembled vocabularies to annotate the gene and gene products.
3) To make the gene annotation data sets available to other researchers via the open-access gene data-set repositories.

The structure of the GO consists of three hierarchies, which cover the following biological aspects [98]–[100]:

a. **Molecular Function (MF):** MF terms are used to describe the abilities that gene products have or the jobs they may implement. This includes activities such as transporting things around, binding to things, holding things together and changing one thing into another. Examples of these include, but are not limited to, the following: 1) nuclease: enzymatic activities; 2) structural constituents of chromatin: structural activities [98], [101], [102].

b. **Biological Process (BP):** BP terms are used to define a biological goal or objective implemented by an ordered series of molecular functions. The beginning and end of the MF activities that contribute to the BP are precisely defined. For example, "mitosis" is the biological process that divides the eukaryotic cell nucleus into two daughter nuclei. A more comprehensive example is the "calcium-dependent cell-matrix adhesion" process, which binds a cell to the extracellular matrix via adhesion molecules with the presence of the calcium for the interaction [98], [101], [102].

c. **Cellular Component (CC):** the locations where the activities of the gene products take 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 a ribosome [98], [101], [102].

As indicated previously, the GO structure consists of three hierarchies, which organise the GO terms as a directed acyclic graph (DAG), where each GO term is represented as a node and the relationships between nodes are defined as arcs. The parent–child relation is the backbone of the GO structure, which indicates that the parent nodes are more general than the child nodes. Table 3-1 (below) lists some of the widely applied relations in the GO DAG. However, it is not a comprehensive list covering all relations in GO [98], [102], [103].

Table 3-1 Widely Applied Relations in the GO DAG

Relation Name	Description
is a (a subtype of)	The "*is a*" relation shapes the fundamental structure of the GO. Whenever GO states that there is an "*is a*" relation between node A and node B (A is a B), this means that A is a subtype of B. It is worth mentioning that the "*is a*" relation does not mean an "instance of". An example of the "*is a*" relation in GO is "mitotic cell cycle *is a* cell cycle".
part of	The "*part of*" connection is used to express the part–whole relation between nodes. If node B is "*part of*" node A, it means that whenever B exists then A must exist as a whole part of B. However, the presence of A does not indicate the presence of B.
has part	The "*has part*" relation is the logical complement of "*part of*". It is used to express the whole–part relation from the parent's perspective. If node A has node B as part of it, it means that B is a necessary part of A, and whenever A is present B is also present. However, the presence of B does not indicate the presence of A.
regulates	The "*regulates*" relation is used to represent the influence of one process on the appearance of other processes, or its quality, such as the enzymatic reaction or cell size. It refers to certain effects of one node on the other. For example, if nodes A and B exist then B always regulates A. However, A may not be solely regulated by B; a more specific example is: the cell cycle checkpoint always regulates the cell cycle. However, the cell cycle is not only regulated by the cell cycle checkpoint.
negatively regulates and *positively regulates*	More specific types of the "*regulates*" relation are the "*negatively regulates*" and "*positively regulates*". These two relations are sub-relations of the "*regulates*" relation. Hence, if A "*negatively regulates*" B, it is valid to infer that "A regulates B". Likewise, if A "*positively regulates*" B, it is valid to conclude that "A regulates B".

Another significant aspect of the GO structure consistency constraints is the true path rule (TPR). The "*is a*" (parent–child) GO relation is obligated to follow the TPR, which states that if an instance of the GO node is proved to be true, its ancestors, all the way to the root, must also be true. Otherwise, if an instance is found to be false, all its descendants to

the leaf nodes must also be false [98], [103], [104]. Figure 3-1 (below) depicts part of the GO structure to illustrate some of its components.

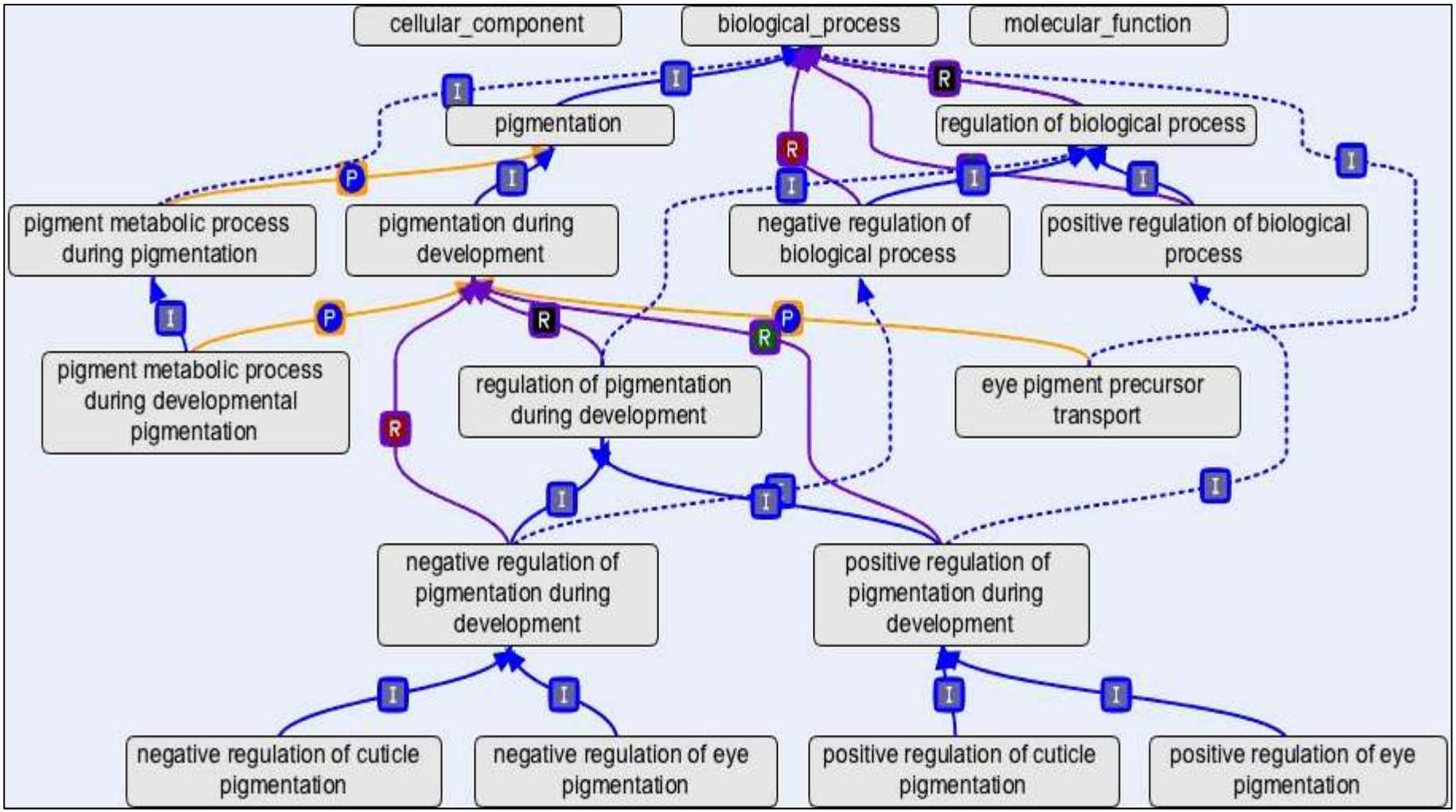


Figure 3-1 GO as Graph [105]

Many studies have reported that gene ontology provides a comprehensive resource for annotating gene products [106]. According to [101], the gene ontology consortium (GOC) published 126 million annotations, which cover more than 347,000 species. These annotations are created either automatically using a computerised method or manually by experts who have studied the relevant literature or examined the biological data. Hence, this research not only uses the gene ontology to annotate the mined data but also exploits the semantic information integrated in the GO to enrich the proposed mining model.

3.3 Chi-squared Test of Independence

The chi-squared test is a non-parametric statistical approach used to test the independency or association between two categorical variables. It checks whether the frequencies of one variable category have an effect on the frequencies of another variable category. If the occurrence frequencies of two categorical variables are related, it can be claimed that the tested variables are dependent; otherwise, they are independent [107].

An initial step in the chi-squared calculation process is to represent the occurrence of the observed data in a two-way table, which is also known as the contingency table [107]. The contingency table is an excellent tool to test the dependency relation between categorical variables, where each cell in the table represents the frequency counts of the associated row and column variables. In fact, the chi-squared test uses the contingency table not only to represent the observed data, but also to calculate the degree of freedom (denoted as *df*), which is essential to specifying the level of significance for the statistical test. The degree of freedom is calculated according to equation (3.1) (below) [108].

$$df = (number\ of\ rows - 1) * (number\ of columns - 1) \qquad (3.1)$$

So, the degree of freedom for 3 rows X 3 columns table is (3-1)*(3-1) = 4. According to [109], the chi-squared distribution forms different families and each family is associated with a specific degree of freedom. Figure 3-2 (below) depicts various distribution families with the associated degree of freedom.

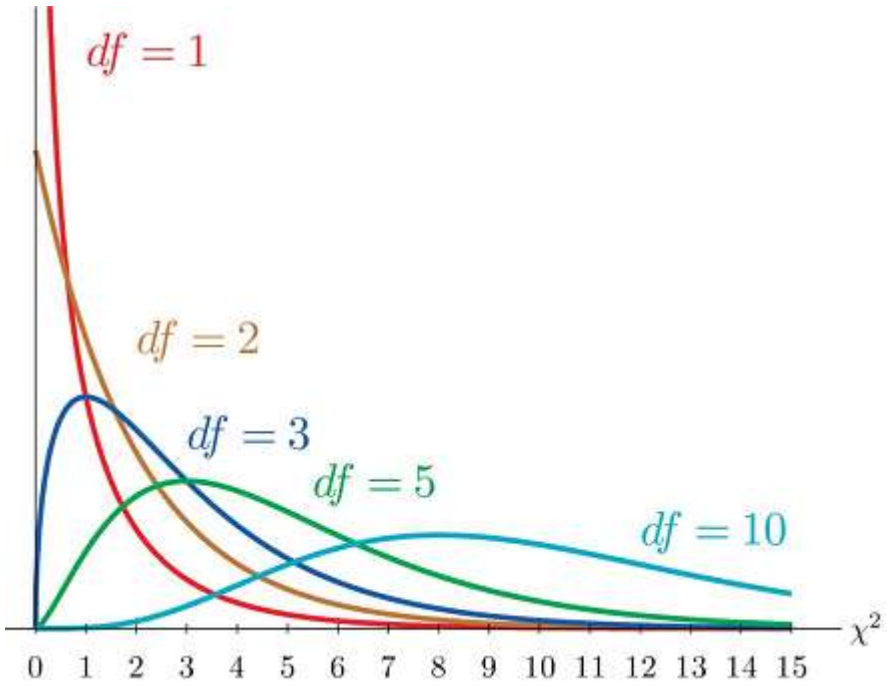


Figure 3-2 Chi-square Distributions for Different Degrees of Freedom [109]

The following example gives a full explanation of the chi-squared calculation process and discusses how to analyse and understand the obtained result. It is assumed that an owner of a company wanted to keep the health problems among his employees as low as possible. Hence, he decided to vaccinate half of the company employees with the pneumonia vaccine. Consequently, the chi-squared test was used to answer the question of whether using the pneumonia vaccine helped to reduce the number of problems related to health issues [108].

As indicated previously, the initial step in the chi-squared calculation process is to represent the observed data as a contingency table. Thus, Table 3-2 (below) represents the contingency table for a population consisting of 184 employees, of which half received the vaccine, and three possible health outcomes, namely, sick with pneumococcal pneumonia, sick with non-pneumococcal pneumonia and no pneumonia.

Table 3-2 Contingency Table Sample [108]

Health Outcome	Unvaccinated	Vaccinated
Sick with pneumococcal pneumonia	23	5
Sick with non-pneumococcal pneumonic	8	10
No pneumonia	61	77

In this example the hypothesis that needs to be validated is whether or not the action of taking the vaccine and the health outcome are independent. The no relation or null

hypothesis indicates that vaccination and health outcome are independent; the alternative hypothesis is that they are related or dependent. The null and alternative hypothesis can be formulated as follows.

H0: No relationship between vaccination and health outcome.

H1: Some relationship between vaccination and health outcome.

Or:

H0: Vaccination and health outcome are independent.

H1: Vaccination and health outcome are dependent.

The step that follows the representation of the observed data as a contingency table and states the null and alternative hypothesis is to calculate the marginal for each row and column in the contingency table. The row marginal is the sum of that row. Likewise, the column marginal is the sum of that column. Table 3-3 (below) presents the updated contingency table along with the marginal values.

Table 3-3 Contingency Table with Marginal Values [108]

Health Outcome	Unvaccinated	Vaccinated	Row Marginal (Row Sum)
Sick with pneumococcal pneumonia	23	5	28
Sick with non-pneumococcal pneumonic	8	10	18
No pneumonia	61	77	138
Column marginal (sum of the column)	92	92	N = 184

The next step is to calculate the expected values, which represent the estimated distribution of the data if there is no relation between the vaccination and the employees' health condition. The expected value for each cell in the contingency table is calculated according to equation number (3.2) (below).

$$\mathrm{E} = M_R X \frac{M_C}{n} \tag{3.2}$$

Where:

- E = the expected value for a specific cell
- M_R= the sum of the row where the cell is located.
- M_C= the sum of the column where the cell is located.
- n = the size of the population sample.

Once the expected values have been calculated the chi-squared for each cell is computed according to equation number (3.3).

$$\text{Chi} - \text{Squared}\ (\text{x}^2) = \frac{(\text{O} - \text{E})^2}{E} \quad (3.3)$$

Where:

- x^2 = the Chi-Squared value for specific cell.
- O = the observed value of the cell.
- E = the estimated value of the cell.

Table 3-4 (below) presents the results of applying equations (3.2) and (3.3) for each cell in the vaccination and health outcome contingency table to calculate the expected and chi-squared values respectively.

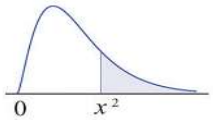
Table 3-4 Expected and Chi-squared Values [108]

Health Outcome	Unvaccinated	Vaccinated
Sick with pneumococcal pneumonia	13.92 (5.92)	12.57 (4.56)
Sick with non-pneumococcal pneumonic	8.95 (0.10)	9.05 (0.10)
No pneumonia	69.12 (0.95)	69.88 (0.73)

The final step is to sum the chi-squared value of each cell to obtain the table total chi-squared value. For this specific example, the result is 12.35 (rounded) [108].

Having discussed how to calculate the chi-squared value from a contingency table, the next step is to explain how to act upon the obtained result. The basic principle of the chi-squared statistic test is to use the calculated value either to reject or accept the null hypothesis. This is done based on two important factors, precisely, the chi-squared critical value at different degrees of freedom and the level of probability. These values are organised in a table, known as the standard chi-squared distribution table, depicted in Figure 3-3 [107][109].

The values in the left column of the chi-squared standard distribution table represent the *df* of the contingency table. Additionally, the associated chi-squared critical values for each *df* are presented at the intersected row. For example, if the *df* is equal to 1, the critical value at the level of probability of 0.10 is 2.706. Hence, if the calculated chi-squared result is greater than or equal to 2.706, then the null hypothesis is rejected; otherwise, it is accepted [107].



Critical Values of Chi-Square Distributions

	x^2 Right-Tail Area									
df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 3-3 Critical Values of Chi-Squared Distributions [109]

3.4 Bayesian Network (BN)

The Bayesian Network (BN), or belief network, has been categorised as a breed of the probabilistic graphical model family. It is a concise representation of the joint probability distribution (JPD) for a set of random variables that represent the domain of interest. The random variables depicted as vertices in the BN structure, and the conditional probability independency between these variables, are represented in the form of arcs. Thus, the BN structure is a combination of vertices (random variables) and arcs (independency relation), which form a directed acyclic graph (DAG). The conditional probability table (CPT) has been used to quantify the strength of connections between variables in the BN. For this, a CPT table is calculated and attached to each vertex in the BN DAG [110]–[112]. Figure 3-4 (below) depicts an example of a BN structure along with the associated probabilistic information and the variable status interpretation.

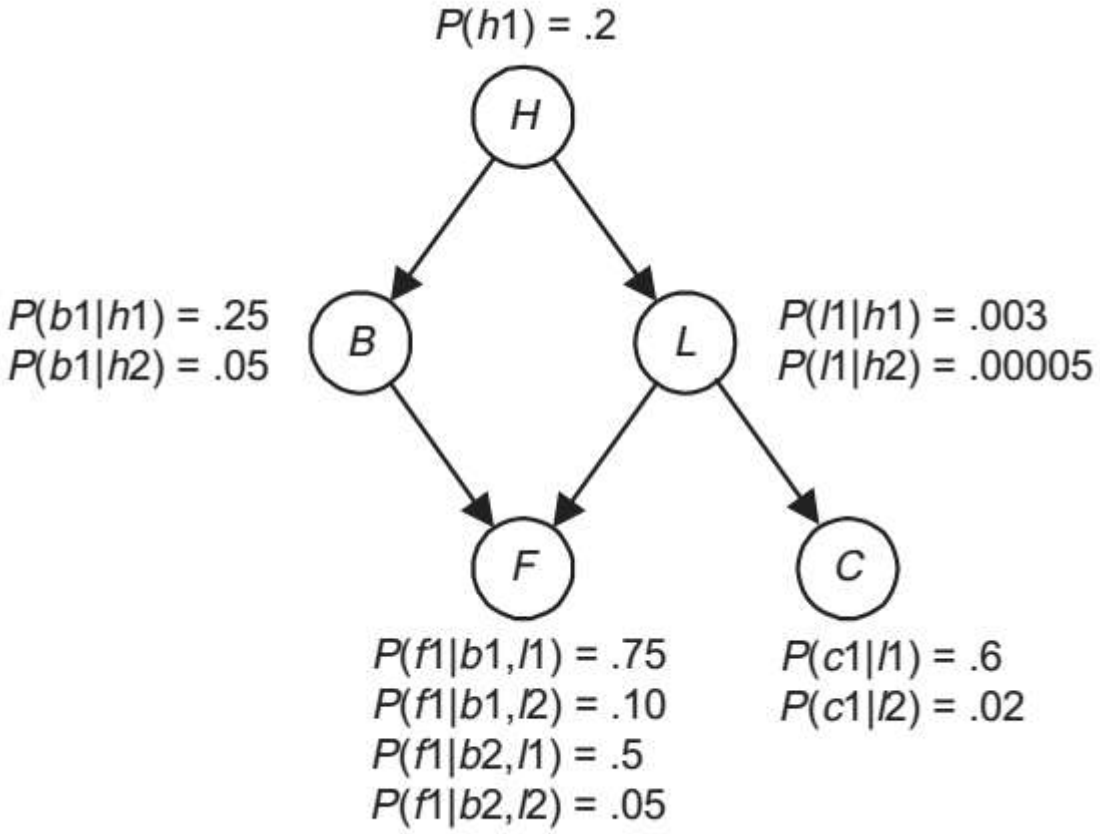


Figure 3-4 A Bayesian network [113]

The variables presented in the above BN represent a medical condition, which could be used as symptoms or signs of bronchitis or lung cancer. The possible values these variables may take are explained in Table 3-5 (below).

Table 3-5 Bayesian Network Variables Possible Values [113]

Variable	Value	When the Feature Takes this Value
H	h1	There is a history of smoking
	h2	There is no history of smoking
B	b1	Bronchitis is present
	b2	Bronchitis is absent
L	l1	Lung cancer is present
	l2	Lung cancer is absent
F	f1	Fatigue is present
	f2	Fatigue is absent
C	c1	Chest X-ray is positive
	c2	Chest X-ray is negative

According to [114], a BN consists of the following elements:

a. A set of random variables and the associated set of directed edges between variables.
b. The status of the variables is mutually exclusive.
c. The acyclic directed graph is generated by combining the variables with the attached edges.
d. A conditional probability table in the form of P(Variable|Parents) is attached to each variable in the graph.

As discussed above, a BN consists of two main components, namely, structure and probabilistic information. Hence, the process of BN learning includes structure construction and CPT calculation.

The BN structure can be manually created by experts, learned from observed data or a combination of the two. Likewise, the probabilistic information can be assigned by experts, deduced from the observed data or a combination of both techniques [113], [115].

Although the process of BN structure learning from data has been heavily investigated and several approaches have been proposed, there is some evidence to suggest that these approaches have been categorised into two main categories [116]–[119] as follows:

a. **Constraint-based methods:** the basic principle of methods under this category is to find the BN structure that best fits the constraints in the form of conditional independency between a subset of variables representing the targeted domain.

b. **Search-based methods**: the ultimate aim of the approaches belonging to this category is to find the structure with the highest scoring function. They start by using a search algorithm to find all the possible structures, and then a scoring function is used to measure the quality of the candidate network with respect to the given data set.

As discussed above, there are two main approaches to constructing the BN structure, namely, expert-based and data-based. However, there is some evidence to suggest that both approaches have their drawbacks. While the expert-based method is time-consuming, error-prone and costly, and experts are not always available, the data-based approaches are biased for the sample of the data used to construct the BN model, and in many real-life scenarios training data is scarce [94], [96], [97]. Hence, alternative approaches have been investigated to overcome the current shortages of the BN structure learning process and to construct a more accurate BN model [84], [94]. Likewise, one of the objectives of the work implemented in this thesis is to investigate the advantages of constructing a hierarchical Bayesian Network based on GO. However, before the SAHBN model is presented in detail, the concept of the Hierarchical Bayesian Network is covered in the next section.

3.5 Hierarchical Bayesian Network (HBN)

The Hierarchical Bayesian Network (HBN) is defined as an extension or generalisation of the standard Bayesian Network (BN), where the structure of the HBN provides more knowledge about the organisation of the variables involved in the network and builds a more realistic probabilistic model. In contrast to standard BN, which cannot represent non-propositional domains, each variable in the HBN structure represents an aggregation of simpler variables. Hence, it has been argued that HBN is an effective model, which decomposes the investigated problem into smaller sub-tasks and provides more control over the data flow and better modelling techniques [120]–[122].

Similar to standard BNs, HBNs consist of sets of nodes and arcs, which form the structure of the HBN. Additionally, the strength of the arcs between the nodes is quantified by a set of CPTs. However, unlike the standard BNs the arcs between the nodes in the HBN represent not only the probabilistic dependency between nodes but also the "*part-of*" relationship. The "*part-of*" relationship can be represented as either nested nodes or a tree-like hierarchical structure. Figure 3-5(below) depicts both interpretations of the "*part-of*" relationship [121], [123].

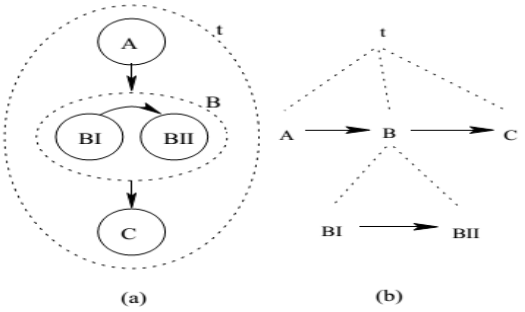


Figure 3-5 a. Nested Representation of the HBN, and b. Tree Representation of the HBN [121]

The basic dependency rule that underpins the HBN structure is that a node is conditionally independent of its non-descendant nodes given the value of its direct parent in the graph [120], [121].

To further illustrate the notion of HBN, the PlayGolf example is discussed in this section. This example investigates whether a particular day is suitable for a person to practise his/her hobby of playing golf. It is assumed that there are two independent factors affecting the action of playing golf, namely, weather and business. Furthermore, it is assumed that the weather factor is an aggregation of three variables (outlook, temperature, wind). Additionally, the business factor consists of two variables, precisely, meeting and market, where the market comprises currency and shares.

Figure 3-6 (below) depicts the structure of the HBN, which represents the hypothetical scenario explained in the PlayGolf example. It can be seen that the proposed HBN structure is very informative. For example, it explicitly shows that weather and business are independent. Furthermore, it can be readily extended or refined [121].

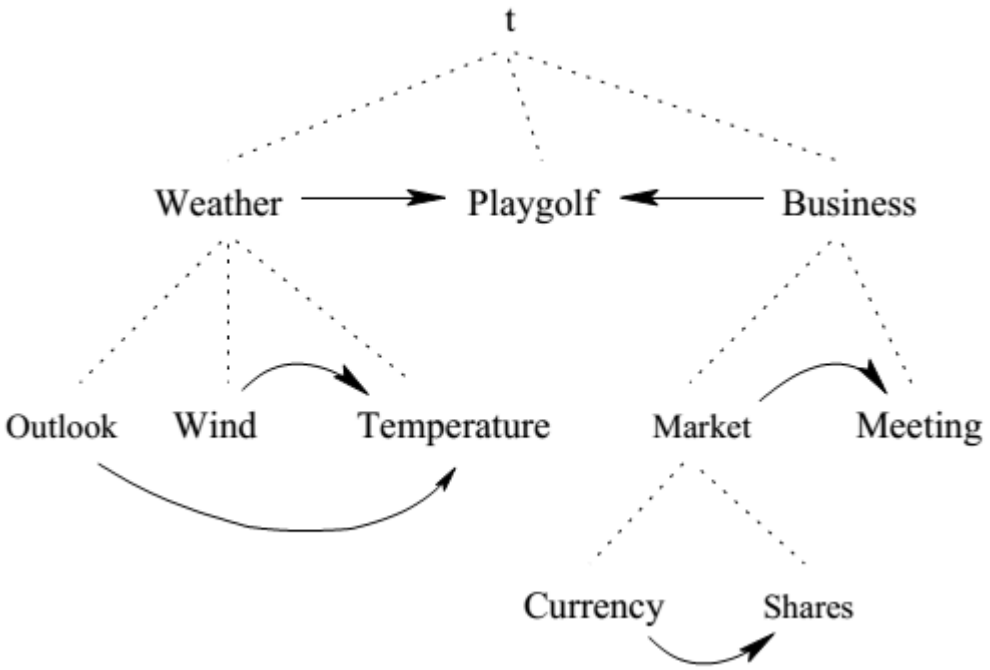


Figure 3-6 The HBN Structure of the PlayGolf Example [121]

Having discussed the structural construction of the BN and HBN networks, the next section explains the second part of the BN, namely, the parameter estimation methods.

3.6 Parameters Estimation Methods for Bayesian Network

In the preceding sections the concepts of BN structure construction were covered. What follows is a brief description of the BN parameter estimation methods. It is argued that there are two main models for estimating the parameters' probabilistic values for complete data in BN, that is, maximum likelihood estimation (MLE) and Bayesian estimation [114], [124], [125], [126]. The following subsections cover these models in detail.

a. **Maximum Likelihood Estimation (MLE):** the aim of the MLE method is to find the value of Θ that quantifies the maximum probability of the incoming event. In a given data set, D, which consists of n instants, represents the outcome of the binominal random variable X. MLE aims to estimate the maximum likelihood of the occurrence of $n+1$ incoming event [124], [127], [128].

   Let X represent the event of flipping a thumbtack, which has two possible outcomes, heads and tails, and D is a set of the observed data. For the sake of simplicity, let us assume that the size of D is five observations, such that D = {X1 = H, X2 = T, X3 = T, X4 = H, X5= H}, where H stands for heads and T stands for tails. Furthermore, the probability of X = H is equal to Θ and the probability of X =

T is 1-Θ. Since the probability of X1, X2, X3, X4 and X5 are independent and identically distributed (IID), the probability of D is defined as follows.

$$P(\langle H, T, T, H, H\rangle: \Theta) = \Theta(1-\Theta)(1-\Theta)\Theta\Theta = \Theta^3(1-\Theta)^2 \quad (3.4)$$

It can be seen that the probability depends on the value of Θ. A different value of Θ results in a different probability. Hence, the likelihood function is defined as follows.

$$L(\Theta: \langle H, T, T, H, H\rangle) = P(\langle H, T, T, H, H\rangle: \Theta) = \Theta^3(1-\Theta)^2 \quad (3.5)$$

To generalise the likelihood function, let us assume that the number of heads is *M[1]* and the number of tails is *M[0]*. Then the general form of the likelihood function would be.

$$L(\Theta: d) = \Theta^{XH}(1-\Theta)^{XT} \quad (3.6)$$

It has been realised that it is easier to maximise the likelihood function by applying the log-likelihood. Consequently, the likelihood function would be.

$$L(\Theta: d) = X^H \log\Theta + X^T \log(1-\Theta) \quad (3.7)$$

Finally, the maximum likelihood parameter $\hat{\theta}$ would be obtained by setting the derivative to 0, differentiating the log-likelihood and solving for Θ. Accordingly, the final form of the likelihood function would be.

$$\hat{\theta} = \frac{X^H}{X^H + X^T} \quad (3.8)$$

It can clearly be seen that the likelihood function is maximised by dividing the number of correct trials over the total number of trials. Although the MLE approach has various advantages, it also has some limitations. For example, the size of the observed data set has no effect on the estimation process. Furthermore, MLE does not take prior knowledge into consideration, relying entirely on the observed data set. Therefore, the Bayesian method, which integrated the prior knowledge into the estimation process, is introduced [114], [124], [127]. In the next subsection the Bayesian estimation method will be explained.

b. Maximum a Posterior Estimation (MAP): an alternative approach to parameter estimation, which injects prior knowledge in the form of prior distribution into the estimation process, is MAP. MLE aims to maximise the likelihood function.

Likewise, MAP aims to maximise the posterior of Θ given the observed data. This hypothesis is formalised in equation (3.9) [124], [127].

$$\widehat{\theta}_{MAP} = \underset{\theta}{argmax}\ p(\theta|d) \tag{3.9}$$

Equation number (3.9) can be rewritten using Bayes rule.

$$\widehat{\ }_{\mathrm{MAP}} = \underset{\theta}{\mathrm{argmax}}\ \frac{\mathrm{p(d|\theta)p(\theta)}}{\mathrm{p(d)}} \text{ where } \mathrm{p(d)} \neq \int(\theta) \tag{3.10}$$

$$\widehat{\theta}_{MAP} = \underset{\theta}{argmax}\ (log\, p(d|\theta) + log\, p(\theta)) \tag{3.11}$$

Hence, the $\widehat{\theta}_{MAP}$ function is calculated by summing the likelihood ($p(d|\theta)$) and the prior knowledge ($p(\theta)$).

It is reported that a binominal random variable, X, which has two possible outcomes, true and false, with probability Θ and 1-Θ for true and false status, respectively, follows the Bernoulli distribution. Furthermore, the Bernoulli distribution is conjugated with the Beta distribution. Hence, it turns out that the Beta distribution is the best option for prior knowledge ($p(\theta)$) representation. Consequently, calculating $\widehat{\ }_{\mathrm{MAP}}$ by summing the likelihood ($\mathrm{p(d|\theta)}$), which is a Bernoulli distribution, with the prior knowledge ($\mathrm{p(\theta)}$) in the form of Beta distribution, will result in a posterior that complies with the Beta distribution law [124], [127].

The probability distribution function (p.d.f) for Beta distribution is defined in equation (3.12) (below).

$$p(\theta) = \gamma\, \theta^{\alpha-1}\, (1 - \theta)^{\beta-1} \tag{3.12}$$

where α and β are two real positive hyper-parameters and γ is a normalising constant, which is defined in equation (3.13) (below) [129], [130].

$$\gamma = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{3.13}$$

The probability distribution function for binomial random variable, X, which follows the Bernoulli distribution, is illustrated in equation (3.14) (below).

$$p(r|n, \theta)\ \alpha\ \theta^{r}\, (1 - \theta)^{n-r} \tag{3.14}$$

where n is the total outcomes of which r are in true status [129].

The preceding paragraphs have argued that the posterior probability *($P(\theta|d)$)* is a Beta distribution, which is obtained by summing the Bernoulli distribution for likelihood and the Beta distribution of the prior knowledge. Hence, the posterior probability could be summarised as Beta distribution with (α+r) success trials out

of (α+β+n) total number of trials. Accordingly, the prior and posterior statistics for Beta distribution could be summarised in Table 3-6 (below) [125], [130].

Table 3-6 Prior and Posterior Statistics for Beta Distribution with R Success in N Trials [125]

Statistic	*Prior*	*Posterior*
Law	*Beta(a,b)*	*Beta(a+r,b+(n-r))*
Mean	$\frac{a}{a+b}$	$\frac{a+r}{a+b+n}$
Mode	$\frac{a-1}{a+b-2}$	$\frac{a+r-1}{a+b+n-2}$
Variance	$\frac{ab}{(a+b)^2+(a+b+1)}$	$\frac{(a+r)(b+n-r)}{(a+b+n)^2+(a+b+n+1)}$

Having discussed the techniques used to lay out the foundations for the proposed model, the next section explains in detail the proposed SAHBN model.

3.7 Semantically Aware Hierarchical Bayesian Network (SAHBN)

As mentioned in the previous chapter, ontology can play various roles in the DM process [50] and facilitate different tasks in the BN construction processes [94]. Hence, this thesis investigated the advantages of utilising the GO to construct the Semantically Aware Hierarchical Bayesian Network (SAHBN) data classifier.

GO was used in this research because of its high quality and comprehensive nature in biomedical domain [106]. Meanwhile, the structure of the HBN implicitly provides more knowledge about the targeted domain [120]. As a result, the integration of these two concepts, GO and HBN, generate a classification model which seamlessly reflects the domain knowledge.

The proposed SAHBN model was tested using different case studies and its performance quality was compared against existing Bayesian-based classification methods. Although SAHBN shares some initial steps with the standard data classification algorithms, the essential steps related to structure learning and parameter probability estimation are designed in such a way that exploits the semantic nature of the GO. Figure 3-7 (below) compares the process sequence of the standard BN classification algorithm and SAHBN model.

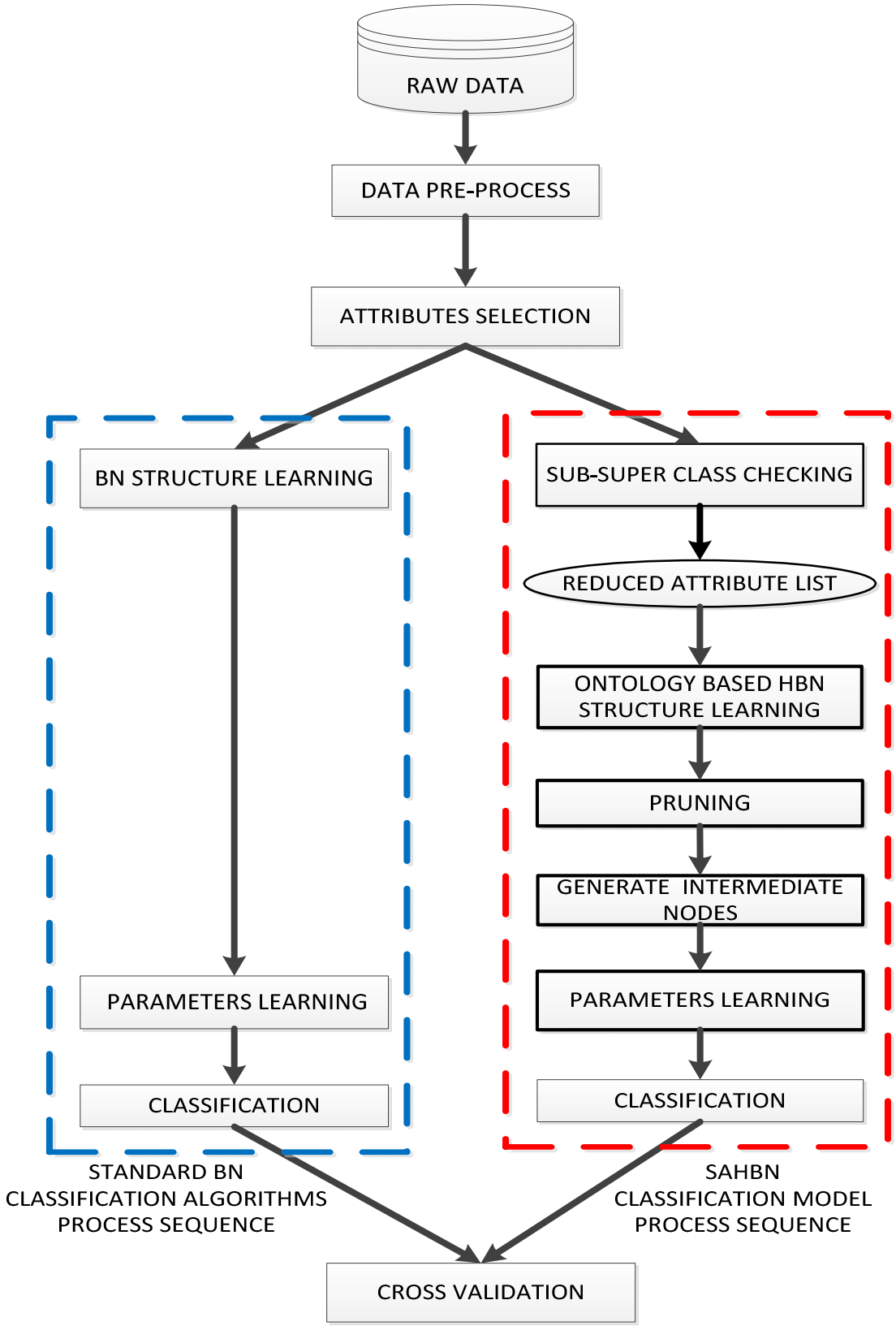


Figure 3-7 SAHBN Process Sequence Versus Standard BN Classification Algorithm Process Sequence.

It can be seen that the process sequence of the proposed model follows some standard preliminary steps, such as data pre-processing and attribute selection. However, the selected attributes were further processed based on the semantic knowledge extracted from the associated ontology (GO in this research). This can be noticed in the steps surrounded by the red dotted line in Figure 3-7. Furthermore, the new steps can be further highlighted by comparing them with the standard process sequence, which is surrounded by the blue dotted line in the same figure.

The new steps introduced by SAHBN can be summarized in the following points.

1. **Sub-Super class checking:** The first step that follows the attribute 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 thesis used the GO terms as a 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 class relation ("is-a"). 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 than the "Cell Aging" process as depicted in Figure 3-8 (below). 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. Additionally, it also achieves the following objectives.

   a. **Maintain data consistency**: as explained earlier, the "is-a" relation in the GO is subject to the TPR. The TPR states that if a GO term is observed to be true, all its super-classes, all the way to the root node, must be true. Otherwise, if it is observed to be false, then all its sub-classes, all the way to the leaf nodes, must be false. Hence, for any two GO terms connected via the "is-a" relation and used as prediction attributes, they must follow the TPR. Otherwise, an inconsistent data set can be used to train the classification model, which may lead to an inaccurate result.

For example, let us assume that there is an "is-a" relation between the first and fifth terms in a particular observed record, R1, which consists of n GO prediction terms. This example can be symbolised as follows:

$$R_1 = \{GO_1, GO_2, GO_3, GO_4, GO_5, \ldots\ldots, GO_n, \text{label class}\}$$

where, GO_5 "is-a" GO_1 and the observed data consists of a sequence of true and false values, as follows:

$$R_1 = \{\text{False, True, True, False, True, } \ldots\ldots \text{, False, True}\}$$

It can be seen that the value of GO_5 = True, while the value of its parent class GO_1 = False, which is contradictory to the TPR stated earlier. This is an example of inconsistent data, which could mislead the training phase of the classifier model. Hence, SAHBN proposed a method to delete one of the contradictory terms and eliminate the inconsistency.

Table 3-7 (below) shows some records from the DNA repair gene–PPI data set (discussed in Chapter 4), which highlights the inconsistency in the training data.

Table 3-7 Sample of Inconsistent Training Data Set

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 latter refers to the ageing biological process, and there is an indirect "is-a" relation between them, as explained in Figure 3-8 (below). Hence, it can be seen that records 3, 6, 8, 9 and 12 (highlighted in red) 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, SAHBN has utilised the concept of chi-squared, which was discussed in the third section of this chapter, to break the conflict between the contradicted terms. This is done in three steps, as follows:

1. Identify the GO prediction terms, which are connected via the "is-a" relation.
2. Calculate the chi-squared value between each term and the label class.
3. Delete the GO term that has the lowest dependency with the label class.

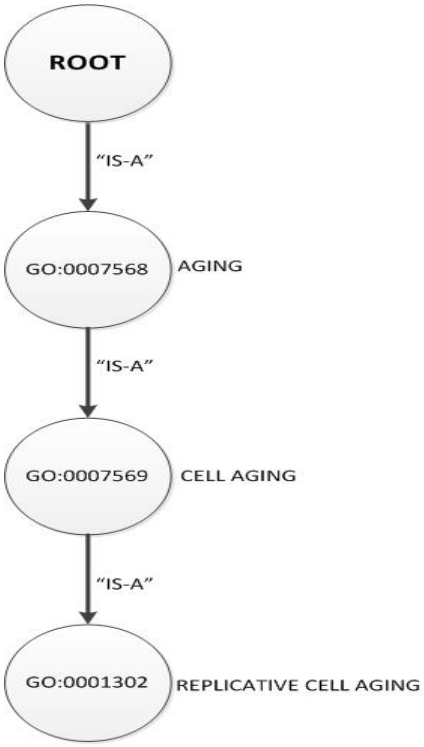


Figure 3-8 GO Attributes "is-a" Relation

The ultimate aim of the above steps is to delete one of the contradicted GO prediction terms that has the lowest prediction capability to the label class. Accordingly, the java program was developed to implement this task. The following pseudo-code describes the main steps in this process.

Read the prediction attributes list
For each GO term in the prediction attributes list
    FIRST_GO_TERM get the current GO term
    For each GO term in the prediction attributes list except the FIRST_GO_TERM
        SECOND_GO_TERM get the current GO term
        If FIRST_GO_TERM "is-a" SECOND_GO_TERM then
            GET the CHI_SQUARED value for the FIRST_GO_TERM
            GET the CHI_SQUARED value for the SECOND_GO_TERM
            If CHI_SQUARED_1st >= CHI_SQUARED_2nd

```
            Delete SECOND_GO_TERM
        Else
            Delete FIRST_GO_TERM
        End If
    End IF
End For
End For
```

Figure 3-9 Pseudo Code to Delete Contradicted GO Terms

b. **Reduce the attributes dimension:** checking the "is-a" relation between the selected attributes not only eliminates the inconsistency in the training data but also reduces the dimension of the prediction attributes. High-dimensional data poses a serious challenge for data mining techniques, especially in the medical domain. A clear example of attribute dimension reduction in this research can be seen in the second experiment of the *Homo sapiens* protein hub data set presented in Table 6-8 in Appendix A (discussed in Chapter 4) when 94 attributes were deleted.

   It is believed that reducing the attributes dimension may affect the prediction quality of the classification model. However, in this research we argued that implementing dimension reduction based on the semantic knowledge extracted from the underpinning ontology does not degrade the quality of the classification model. In fact, some results suggested that the reduced attribute list obtained better classification accuracy in various cases. A detail discussion of the experimental results, along with the associated data set, is presented in Chapter 4.

2. **Ontology-based HBN structure learning:** the second step, which follows the parent–child class checking, is HBN structure learning. The structure learning task is implemented based on the reduced attributes list and the structure of the GO. The steps involved in this process are summarised in the following points:

   a. Match each attribute in the reduced list generated after 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 the 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.

To further illustrate the process of HBN structure learning, let us assume that the reduced attributes list consists of five GO terms, as follows:

Attributes List = {GO1, GO2, GO3, GO4, GO5}. And the parent classes for each GO term in the attributes list are listed in Table 3-8 (below).

Table 3-8 Attribute List with Super-classes

Attribute Term	Super-Classes list
GO1	{GO6, GO9, GO12, Root}
GO2	{ GO9, GO12, Root}
GO3	{ GO7, GO10, GO12, Root}
GO4	{ GO7, GO10, GO12, Root}
GO5	{ GO8, GO11, GO13, Root}

Accordingly, the structure of the HBN, which is constructed as a result of combining the terms in Table 3-8 (above), is depicted in Figure 3-10 (below).

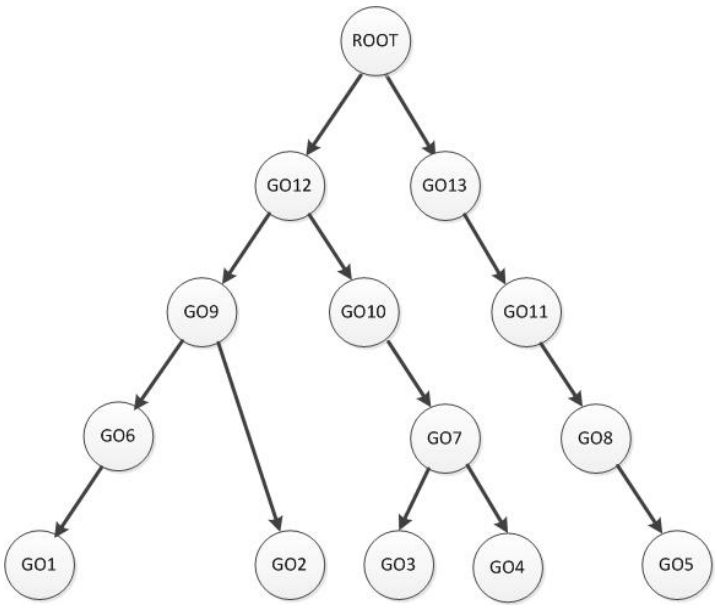


Figure 3-10 HBN Structure Example

In order to implement the HBN construction task, the java program was developed. The pseudo code presented in Figure 3-11describes the functionality of this program.

```
Read the prediction attributes list
For each GO term in the reduced attribute list
    CURRENT_TERM Get the current attribute term.
    Match the CURRENT_TERM to GO node
    CURRENT_TERM_PATH Get the path started fro the matched node to the root node
    ADD the extracted path to the paths list
End For
```

SAHBN_STRUCTURE Combine all the extracted paths

Figure 3-11 Pseudo Code for SAHBN Structure Construction Process

3. **Pruning:** the step that follows the HBN structure learning process is 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 mean that if A is-a B, and B is-a C, we can infer that A is-a C. Hence, it is safe to aggregate terms connected by the "is-a" relationship [105]. **Error! Reference source not found.** (below) explains the transitive nature of the "is-a" relation.

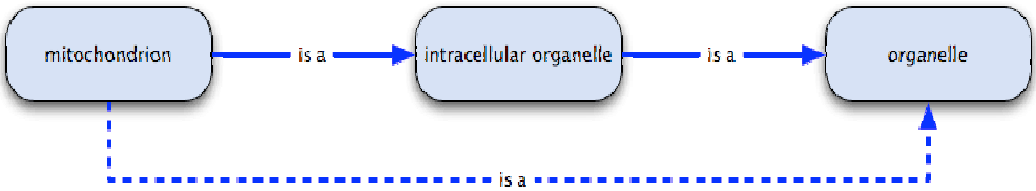


Figure 3-12 The Transitive Nature of the "is-a" Relation [105]

The aim of this step is to remove redundant nodes that do not affect the principles of the HBN structure. As described in the previous sections, there are two main basic principles underpinning the structure of the HBN. These principles can be summarised in the following points:

a. **Aggregation:** each node in the HBN structure represents an aggregation of simpler 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 Figure 3-10, which was constructed in the previous step. As a result GO6, GO8, GO10, GO11 and GO13 terms, and the associated arcs, were deleted. The steps of the pruning process are summarised in Figure 3-13 (below).

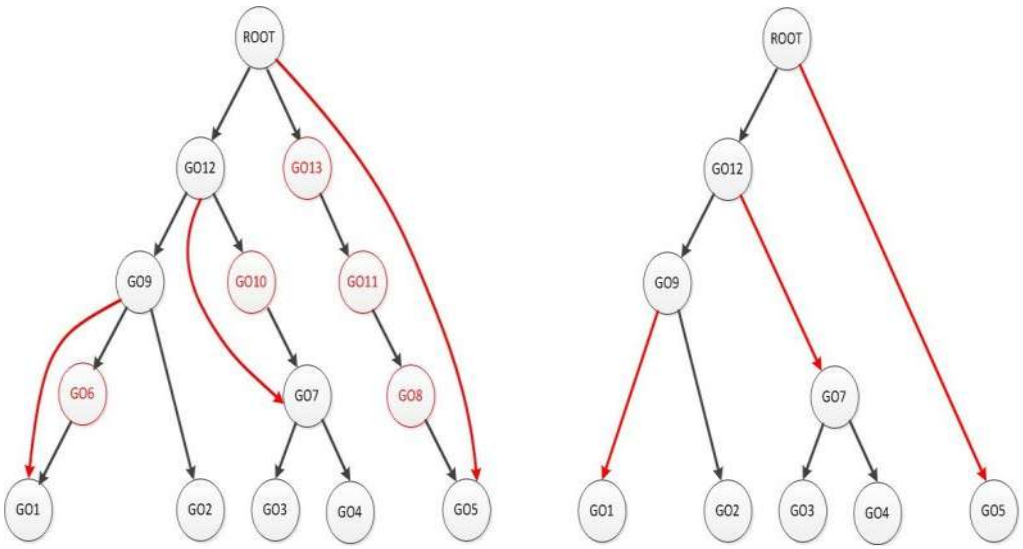


Figure 3-13 Pruning Process

As mentioned on the previous page, the java program was developed to construct the HBN structure. Likewise, the java program was used to implement the pruning task. The pseudo code presented in Figure 3-14 explains the functionality of the pruning java program.

```
READ the SAHBN structure
FOR each node in SAHBN structure
    CURRENT_NODE Get the current node
    If CURRENT_NODE is not root and not terminal node
        IF CURRENT_NODE has one sub-class
            DELETE the CURRENT_NODE and the associated arcs
            CONNECT the super-class of the deleted node to the deleted node sub-class
        END IF
    END IF
END FOR
```

Figure 3-14 Pseudo Code for SAHBN Structure Pruning Process

4. **Generate intermediate nodes:** by examining Figure 3-13, it can be seen that three intermediate nodes were added to the structure of the HBN, namely, GO7, GO9 and GO12. Unlike the terminal nodes (i.e. prediction attributes), the values of the intermediate nodes are unknown. Thus, this section is devoted to explaining the technique followed to assign values to the unobserved intermediate nodes.

   As was pointed out in the previous sections, the subtype relation between GO terms was built based on the TPR. Hence, the concept of the TPR was exploited to define the values of the intermediate nodes. This was 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*".

Applying the above rules to generate the values of the intermediate nodes leads to the creation of a more consistent and accurate data set, which is used to train the classification model and can lead to more solid results.

The task of generating the values of the intermediate nodes was implemented with the help of java programming language. The pseudo code presented in Figure 3-15 explains the actions taken in this task.

```
REPEAT_UNTIL the values of all intermediate nodes are generated
    FOR_EACH intermediate node in the SAHBN structure
        CURRENT_INTERMEDIATE_NODE Get the current intermediate node
        GET the current intermediate node sub classes list
            FOR_EACH node in the sub-classes list
                IF the sub-class value is equal to true
                    CURRENT_INTERMEDIATE_NODE_VALUE is equal to true
                ELSE
                    CURRENT_INTERMEDIATE_NODE_VALUE is equal to false
                END IF
            END FOR
    END FOR
END REPTEAT
```

Figure 3-15 Pseudo Code for SAHBN Intermediate Nodes Value Generation Process

5. **Parameters learning:** having filled the intermediate nodes with values, we have a completed training data set, which can be used to learn the variable probability. The MAP method, which was explained in Section 3.6 of this chapter, was used to calculate the probability values for each variable in the SAHBN model.

3.8 Chapter Summary

In summary, this chapter has argued that the dependency rule that forms the basis 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 organises its terms in a hierarchical structure using the "is-a" relation, where each GO parent term is more general than the child term. Hence, the assumption made in this thesis claims the following:

"Each prediction attribute is represented as a terminal node, which is independent of other prediction attributes given the value of its parent class. Furthermore, each intermediate node represents an aggregation of its child classes (i.e. subtype), which 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 model and GO structure".

Having discussed how to construct the SAHBN model and the associated techniques, the experimental implementation is discussed in detail in the next chapter.

Chapter 4 Empirical Implementations and Experimental Results

4.1 Introduction

This chapter explains in detail the experimental implementation and the obtained results. It discusses the evaluation process, data sets creation, the SAHBN model implementation, comparisons with existing algorithms and analysis of the results. Accordingly, to achieve these tasks, the following software tools were used.

a. Weka: data mining software written in Java [131].
b. RStudio: open source edition [132].
c. Netic-J API from Norsys Software Corp [133].

It is worth mentioning that for each tested data set 11 attribute selection methods were used. Table 4-1 (below) summarises the combinations of the attribute selection methods used.

Table 4-1 Attribute Selection Methods

No.	Attributes Selection Method	
	Evaluation Method	Search Method
1	CfsSubSetEval	BestFirst
2	CfsSubSetEval	GreedyStepwise
3	ConsistencySubsetEval	BestFirst
4	ConsistencySubsetEval	GreedyStepwise
5	FilteredSubsetEval	BestFirst
6	FilteredSubsetEval	GreedyStepwise
7	InfoGainAttributeEval	Ranker
8	GainRatioAttributeEval	Ranker
9	CorrelationAttributeEval	Ranker
10	ReliefFAttAttributeEval	Ranker
11	OneRAttributeEval	Ranker

All in all, the proposed model was tested using eight data sets organised in three case studies. In total, 1,093 experiments were implemented. Additionally, six performance criteria were calculated to evaluate the performance of the SAHBN model.

Finally, the results produced by SAHBN were compared with the results generated by the standard Bayesian-based classification algorithm, such as ICSS, K2, TAN, Hill-Climbing and Tabu Bayesian classifier (hereafter, we refer to them as existing algorithms).

Having discussed the software tools and structure of the experimental implementation, the next section explains the evaluation process and metrics.

4.2 Cross-validation

The ultimate aim of classification models in a real-life application is to predict the class of some unknown instances based on their observed attributes. However, in an experimental environment, and as stated by the CRISP-DM model, the performance of the developed classifier must be measured before it can be deployed [27], [134].

It has been reported that the performance of a classification model is measured in order to address the following points [135]:

a. To identify the most suitable model for a given task.
b. To anticipate the model performance when deployed.
c. To prove that the developed model meets the objectives for which it has been developed.

The basic principle of the classifier cross-validation process is to test the developed classification model using a testing data set that has not been used during the classifier training phase. Since the classes of the instances in the testing data set are known in advance, the performance of the classifier is determined by counting the frequencies when the developed classifier predicts the correct/incorrect instance class. The output of the cross-validation process is a two-dimensional matrix known as the confusion matrix [134], [135].

		predicted	
		positive	negative
truth	positive	tp	fn
	negative	fp	tn

Figure 4-1 Confusion Matrix Structure [134]

Figure 4-1shows that the confusion matrix contains four values. Each one has captured a certain performance aspect. The confusion matrix values were interpreted as follows [135]:

a. **True positive (TP):** this is the number of instances that have a positive value in the test data set and predicted to have a positive value by the classifier.
b. **True negative (TN):** this is the number of instances that have a negative value in the test data set and predicted to have a negative value by the classifier.
c. **False positive (FP):** this is the number of instances that have a negative value in the test data set and predicted to have a positive value by the classifier.

d. **False negative (FN):** this is the number of instances that have a positive value in the test data set and predicted to have a negative value by the classifier.

The confusion matrix not only provides detailed information about the predicted results, but also forms the basis for calculating other performance measures [135]. The following points cover the measures used in this thesis and calculated based on the confusion matrix:

a. **Classification accuracy:** this can take values in the range between 0 and 1; higher accuracy indicates a better performance. Equation (4.1) explains the calculation process for classification accuracy [135].

$$\text{Classification accuracy } = \frac{(\text{TP} + \text{TN})}{(TP + TN + FP + FN)} \tag{4.1}$$

a. **Precision:** this measures the certainty that a positive instance in the testing data has been correctly classified as positive by the developed classifier. It takes values in the range between 0 and 1. Higher precision indicates a better performance. Equation (4.2) shows the calculation of classifier precision [135].

$$\text{Precision} = \frac{\text{TP}}{(\,TP + FP\,)} \tag{4.2}$$

b. **Recall:** this measures the certainty that all positive instances in the testing data set have been found by the proposed model. It takes values in the range between 0 and 1. A higher recall value indicates a better classification performance. Equation**Error! Reference source not found.** shows the calculation of recall [135].

$$Recall \ = \frac{TP}{(\,TP + FN)} \tag{4.3}$$

c. **F1 Measure:** this represents the combination of precision and recall into one measure, which is a simpler alternative to the misclassification rate. Equation**Error! Reference source not found.** defines the F1 measure calculation [135].

$$\text{F1 measure} = 2 \times \frac{(precision \times recall)}{(\,precision \ + recall)} \tag{4.4}$$

d. **Average class accuracy:** the classification accuracy defined in equation**Error! Reference source not found.** (above) can misjudge the classifier performance if the tested data set is imbalanced. Hence, the average accuracy was used to overcome this issue. Average accuracy is defined in equation**Error! Reference source not found.** (below) [135].

$$Average\ class\ accuracy = \frac{1}{|levels(t)|} \sum_{l \in levels(t)} \text{recall}_l \quad (4.5)$$

Where $levels(t)$ refers to the set of levels the targeted feature t can take; $|levels(t)|$ is the set of levels size and recall_l is the recall value obtained by the model for level l.

e. **Average class accuracy (harmonic mean):** the average accuracy defined in equation (4.5) (above) used the arithmetic mean. However, other research prefers to use the harmonic mean, which highlights the effect of smaller values and produces a more realistic measure of how a model is performing. Equation**Error! Reference source not found.** (below) defines how the harmonic mean accuracy is measured.

$$\text{Average class accuracy}_{HM} = \frac{1}{\frac{1}{|levels(t)|} \sum_{l \in levels(t)} \frac{1}{\text{recall}_l}} \quad (4.6)$$

Although various approaches are available to create the test data set, recent research indicates that the 10-fold cross-validation approach has been widely used [134]. In this approach the available data was divided into 10 equal-sized partitions, and then in each run 1 partition was used as test data, while the other 9 partitions were used as training data. This process was repeated 10 times until all partitions had been used as testing data. The overall prediction model performance represents the aggregation of the model performance in each run. Figure 4-2 (below) explains the k-folds process in the form of pseudo code [134], [135].

Input: Training set S, integer constant k
Procedure:
    partition S into k disjoint equal-sized subsets $S_1, \ldots, S_k$
    for $i = 1$ to $i = k$
        let $T = S \setminus S_i$
        run learning algorithm with T as training set
        test the resulting classifier on S_i obtaining tp_i, fp_i, tn_i, fn_i
    compute $tp = \sum_i tp_i$, $fp = \sum_i fp_i$, $tn = \sum_i tn_i$, $fn = \sum_i fn_i$

Figure 4-2 K-fold Cross-validation Process [134]

Having discussed the performance measures taken to compare the proposed SAHBN model with the existing Bayesian-based classification algorithms, the next section explains the data sets created in the human ageing case studies.

4.3 Human Ageing Case Studies

Human ageing has been defined as the gradual failure of the physiological functions of various cells, tissues and organs in the human body, which ultimately leads to the fragility of body functionalities within time and increases the probability of death [136]–[138].

Recent research suggested that advances in the health-care sector in developed countries have led to a substantial increase in the human lifespan. In fact, some research has reported that almost 20% of the world's population will be aged 60 or older by 2050. Furthermore, some statistics show that the human lifespan has increased by almost three months per year since 1840. For example, the expected lifetime of Japanese women is 85 years, which is the highest in developed countries. Additionally, the number of elderly people in the US population rose from 3,700 in 1940 to approximately 61,000 in 2006. Accordingly, the boost in the percentage of centenarians in different country's populations has resulted in many challenges. These challenges can be summarised in the following points:

a. An increase in diseases such as heart failure, cancer, diabetes and Alzheimer's.

b. An increase in health-care costs.

c. A shortage of caregivers.

d. Dependency.

e. A large impact on society.

Hence, the task of understanding the human ageing process has been the focus of many researchers in various developed countries. The ultimate aim of those researchers has been to develop new techniques to prevent or delay diseases associated with the ageing process, or even to treat them in more successful and rational ways[136]–[139].

Recent research has revealed some interesting findings about the human ageing process. However, it is a widely held view that human ageing is an extremely complex, mysterious, controversial and puzzling process, which requires further investigation. Studying the human ageing process has presented some challenges, such as ethical factors associated with doing experiments on human data. Additionally, a long timescale is required to implement experiments on humans. Finally, there are comprehensive elements that must be taken into account when analysing the ageing process. Thus, research has alternatively used the gene/protein data of short living organism models to implement experiments. Consequently, data mining techniques were recently applied to analysing the large amount

of openly available gene/protein databases and to gain some insights into the human ageing process[140]–[143]. Likewise, the SAHBN model proposed in this thesis was applied to two case studies, which predicted the gene effect in the human ageing process. The data sets used in these case studies and the obtained results are discussed in detail in the next subsections.

4.3.1 DNA Repair Genes Case Study

The human genome preserves its integrity by protecting the cellular DNA from both internal and external attacks. While external attacks can be caused by exposure to ultraviolet (UV) light from the sun, inhaled cigarette smoke or incompletely defined dietary factors, internal attacks caused by mutagens elements implicitly appeared in the cells such as water, reactive oxygen species and metabolites that can act as alkylating agents. Thus, cellular DNA is steadily monitored by the repair enzymes to correct the damage resulting from these attacks. Accordingly, it has been reported that modification of the DNA repair process will result in an advanced understanding of the cellular ageing process [144]. Additionally, it has been agreed that DNA damage is an essential element in the ageing process [142]. Hence, the SAHBN model proposed in this thesis was applied to the DNA repair gene databases to classify their effects as either ageing-related or non-ageing-related genes.

4.3.1.1 Data Set Characteristics and Pre-processing

The data sets used in this case study were created using two different approaches. In the first one the protein–protein interactions (PPI) database was used to represent each DNA repair gene in the form of its gene ontology biological process (GO BP) terms. Meanwhile, the second approach directly converted each DNA repair gene into its GO BP terms using the Gene2GO database, as clarified in Figure 4-3 (below).

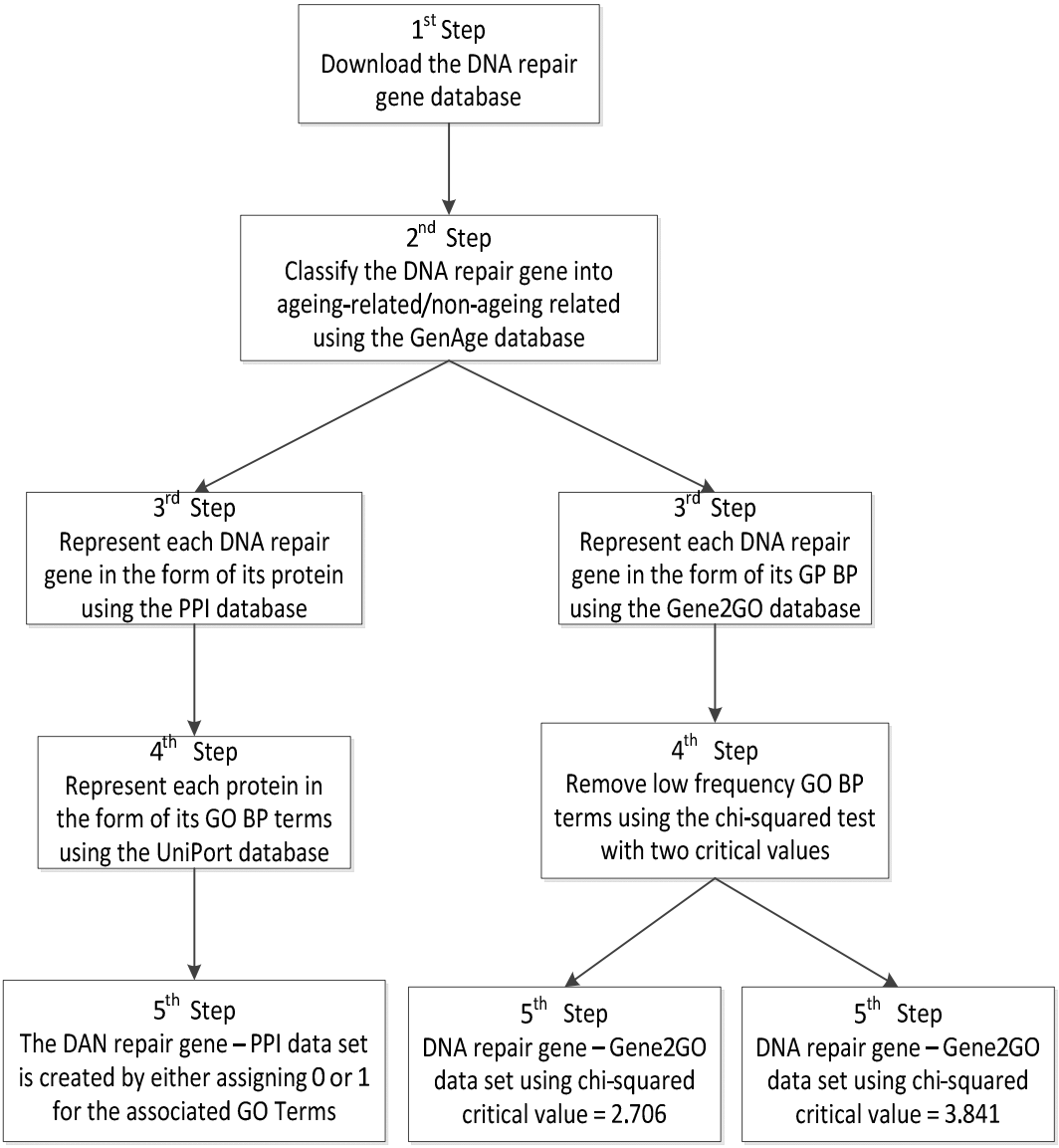


Figure 4-3 DNA Repair Gene Case Study Data Set Creation Process

The following steps explain the data set creation process for the PPI and Gene2GO approaches.

- DNA repair gene – PPI data set creation approach:
    I. Download the DNA repair gene database from the human DNA repair gene website [145].
    II. The downloaded DNA repair genes were classified into two categories: ageing-related and non-ageing-related. The DNA repair genes appearing in the GenAge [146] database were classified as ageing-related, while the other DNA repair genes that did not appear in the GenAge database were classified as non-ageing-related.

III. The protein–protein interaction was extracted from the human protein reference database [147]. The extracted protein interactions meet the following criteria:
    - At least one of the interacted proteins is located in a DNA repair gene.
    - The type of evidence for the interactions is either *in vitro* or *in vivo* experiments.

IV. The extracted protein pair was represented in the form of GO BP terms using the database available at the UniProt [148] website.

V. Finally, each DNA repair gene is represented by a set of GO BP terms associated with the proteins that represent the gene. The value of the GO BP term is equal to 1 if it appeared in the protein associated with the gene; otherwise, it is equal to 0.

Eventually, the created data set consisted of 178 records, each representing a DNA repair gene, and 3,163 GP BP terms, which represent the prediction attributes.

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

  III. Represent each DNA repair gene for its GO BP terms using the national Center for Biotechnology Information (NCBI) Gene2GO database [149].

  IV. Remove GO terms that have low frequency and possess no or very low prediction power. [142] used a predefined frequency threshold to remove low-frequency terms. In a slightly different manner, this research used the chi-squared to measure the dependency between each attribute and the label class and then remove those attributes that appeared to be independent from the label class. Two critical values were used in the chi-squared test, namely, 2.706 and 3.841.

  V. Finally, each DNA repair gene is represented by a set of GO BP terms. The value of the GO terms is equal to 1 if it is associated with the given gene; otherwise, it is equal to 0.

  Eventually, two data sets were created, one for each chi-squared critical value. The characteristics of these data sets are summarised in Table 4-2 (below):

Table 4-2 Repair Gene–Gene2GO Data Set Characteristics

Chi-Squared Critical Value	Number of Records	Initial Attributes Number	Chi-Squared Reduced Attr. Number
2.706	177	735	335
3.841	177	735	89

In summary, three data sets were created for the DNA repair gene case study. The DNA repair gene–PPI data set used the protein–protein interactions database, while the DNA repair gene–Gene2GO data sets relied on the Gene2GO database. Consequently, the SAHBN model was applied to these data sets and the results compared against the existing classification algorithms. The next subsection explains the results obtained in detail.

4.3.1.2 Experimental Results

As discussed in the previous subsection, the data sets tested in the DNA repair gene case study were pre-processed in two different ways and produced three data sets. One data set was created based on the DNA repair gene-PPI approach, and two data sets were created based on the DNA repair gene–Gene2OG approach, using different chi-squared critical values. Hereafter, the results obtained from applying the SAHBN model to these data sets are presented in detail. Additionally, the obtained results were compared with different existing classification algorithms.

- DNA repair gene–PPI data set results discussion

Table 6-1 in Appendix A presents the results obtained from the DNA repair gene–PPI data set. It shows that the SAHBN model produced a very competitive classification quality compared to the existing algorithms. For example, in the first and second experiments the classification accuracy of SAHBN was either as good as, or outperformed, the existing algorithms. Likewise, the average and harmonic accuracies of the existing algorithms did not exceed the SAHBN model. Another good example is the fourth experiment, when all the six performance criteria (i.e. precision recall, F1 measure, accuracy, average accuracy and harmonic accuracy) of the SAHBN model outperformed the existing algorithms, with the exception of one test, when the recall of the K2 algorithm was equal to the recall of the proposed model.

To further illustrate the results obtained from the DNA repair gene–PPI data set, the frequency of experiments when the SAHBN model outperformed, was equal to or less

than the existing algorithms for all six performance criteria is summarised in Table 4-3 (below).

Table 4-3 DNA Repair Gene–PPI Data Set Results Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	13	23	18	15	24	27	120
Equal to	6	9	5	10	5	2	37
Less than	14	1	10	8	4	4	41
Total	33	33	33	33	33	33	

Table 4-3 (above) shows that the total number of tests when the SAHBN model outperformed the existing algorithms is almost triple the number of tests when the SAHBN model was exceeded by the existing algorithms. Additionally, it shows that harmonic accuracy has the highest number of experiments when the proposed model outperformed the existing algorithms. This can be further explained in Figure 4-4 (below).

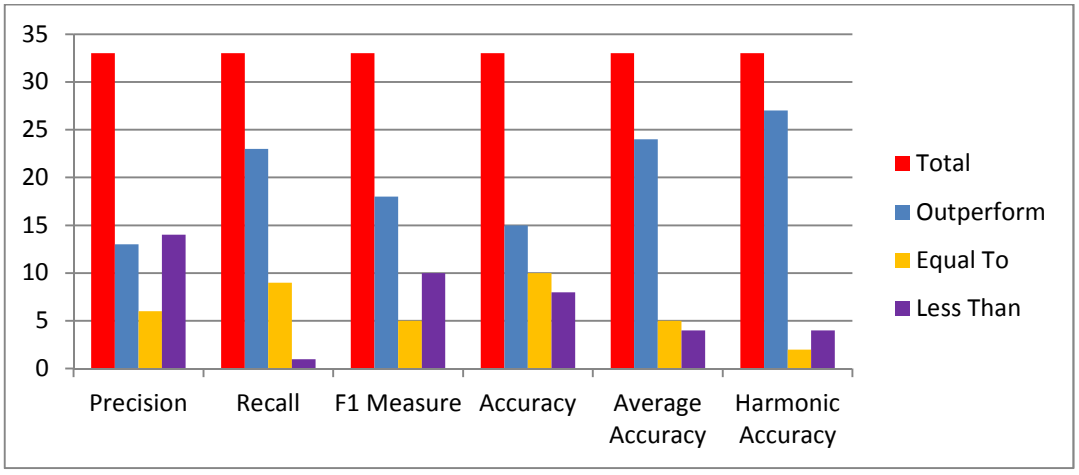


Figure 4-4 DNA Repair Gene–PPI Data Set Results Summary

Having discussed the overall results obtained from the DNA repair gene–PPI data set, hereafter the result of each experiment will be discussed individually. The result of each experiment will be briefly explained and then represented in the form of a graph.

- Experiment 1

  The first experiment used the (CfsSubSetEval+BestFirst) attributes selection method. Accordingly, the results of the first experiment show that the SAHBN model outperformed the ICSS algorithm in all performance criteria. Additionally, it exceeded the K2 and TAN algorithms in terms of recall, average and harmonic accuracies. However, for other performance criteria, it was either as good as, or outperformed by, the K2 and TAN algorithms. Figure 4-5 (below) depicts the results obtained.

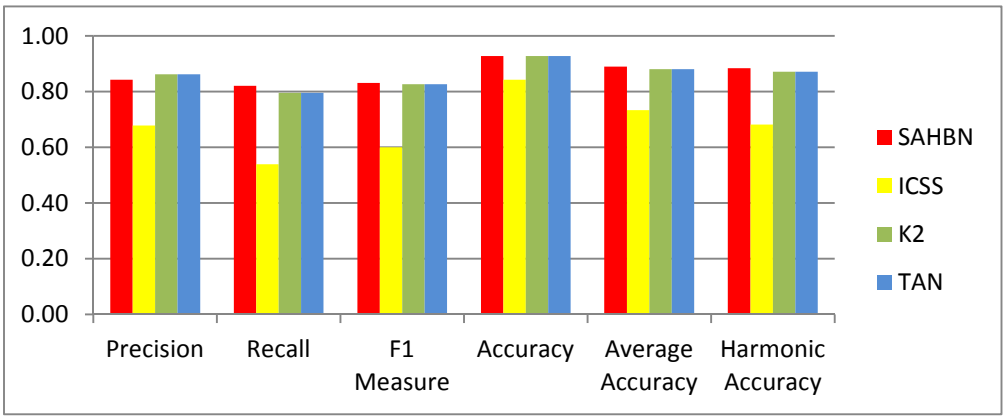


Figure 4-5 DNA Repair Gene–PPI Data Set First Experiment Results

- Experiment 2

  The second experiment used the (CfsSubSetEval + GreedyStepwise) attributes selection method. Accordingly, the results of the second experiment indicate that the SAHBN model outperformed the ICSS and TAN algorithms in terms of all performance criteria. However, it was as good as the K2 algorithm. Figure 4-6 (below) describes the obtained results.

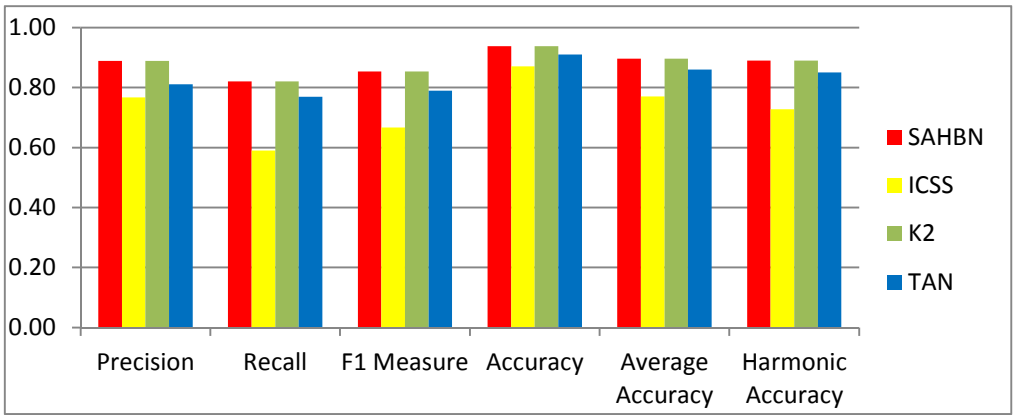


Figure 4-6 DNA Repair Gene–PPI Data Set Second Experiment Results

- Experiment 3

  The third experiment used the (ConsistencySubsetEval+BestFirst) attributes selection method. Accordingly, the results of the third experiment show that the SAHBN model outperformed the existing algorithms in terms of average and harmonic accuracies. Additionally, it was as good as the existing algorithms in terms of recall. However, SAHBN was outperformed by the existing algorithms in terms of precision, F1 measure and accuracy. Figure 4-7 (below) explains the obtained results.

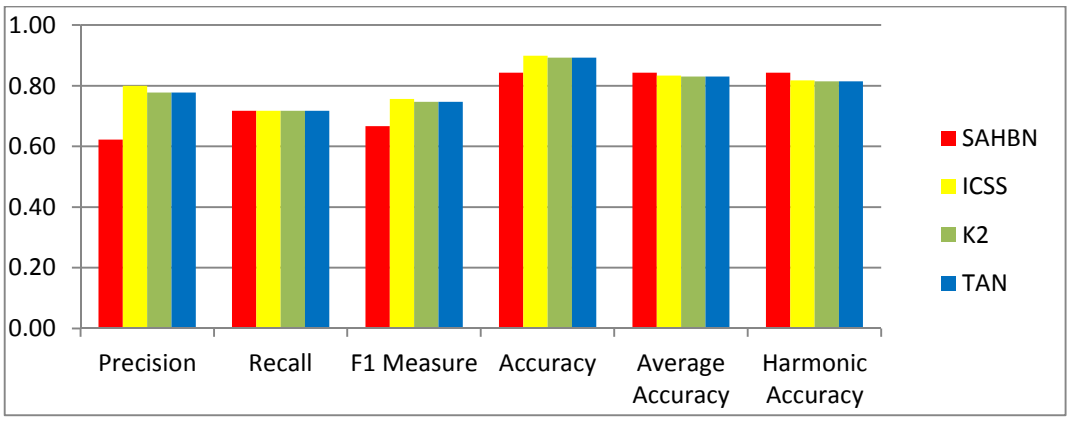


Figure 4-7 DNA Repair Gene–PPI Data Set Third Experiment Results

- Experiment 4

  The fourth experiment used the (ConsistencySubsetEval + GreedyStepwise) attributes selection method. Accordingly, the results of the fourth experiment reveal that the SAHBN model outperformed all the existing algorithms in all performance criteria, except for the recall value of the K2 algorithm, which was equal to the recall value of the SAHBN model. Figure 4-8 (below) depicts the results obtained.

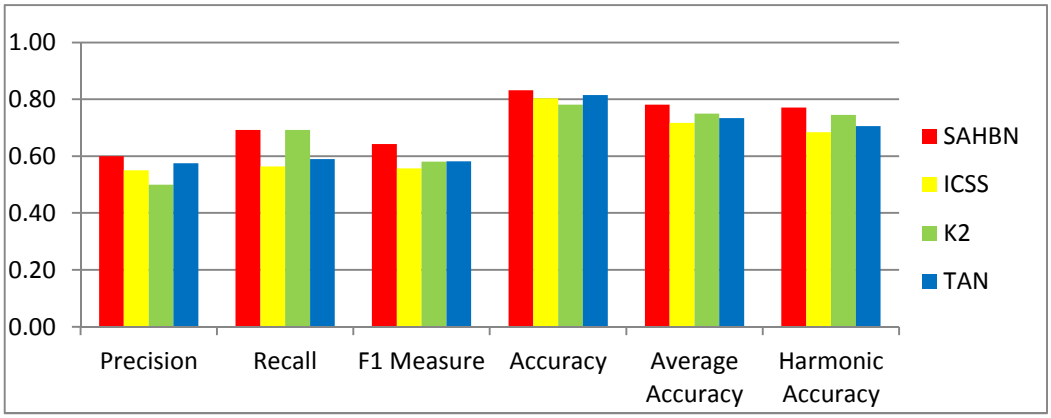


Figure 4-8 DNA Repair Gene–PPI Data Set Fourth Experiment Results

- Experiment 5

  The fifth experiment used (FilteredSubsetEval+BestFirst) attributes selection methods. Accordingly, the results of the fifth experiment indicate that the SAHBN model outperformed the ICSS algorithm in all performance criteria. Additionally, the recall, average and harmonic accuracy values of the TAN algorithm were slightly lower than the corresponding values of the SAHBN model. However, the K2 algorithm was marginally better in terms of precision. Figure 4-9 (below) depicts the results obtained.

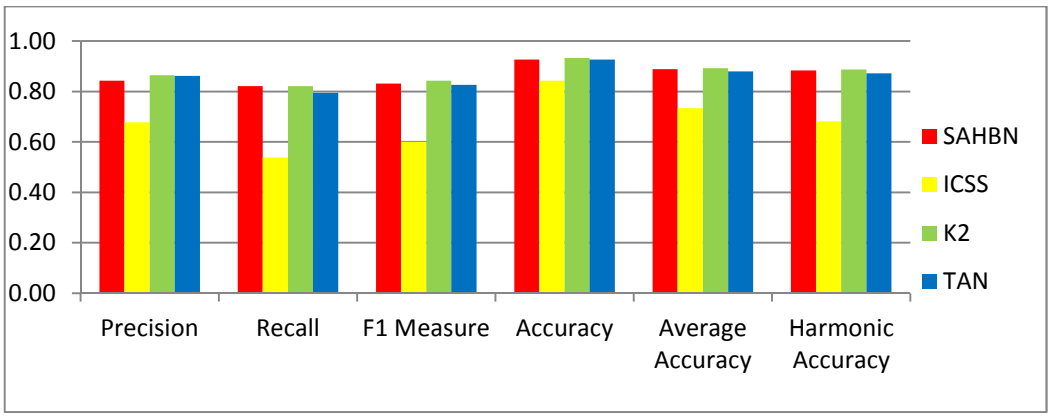


Figure 4-9 DNA Repair Gene–PPI Data Set Fifth Experiment Results

- Experiment 6

  The sixth experiment used (FilteredSubsetEval + GreedyStepwise) attribute selection methods. Accordingly, the results from the sixth experiment reveal that the SAHBN model outperformed the ICSS and TAN algorithms in all performance criteria. However, the K2 algorithm was almost as good as the SAHBN model. Figure 4-10 (below) depicts the results obtained.

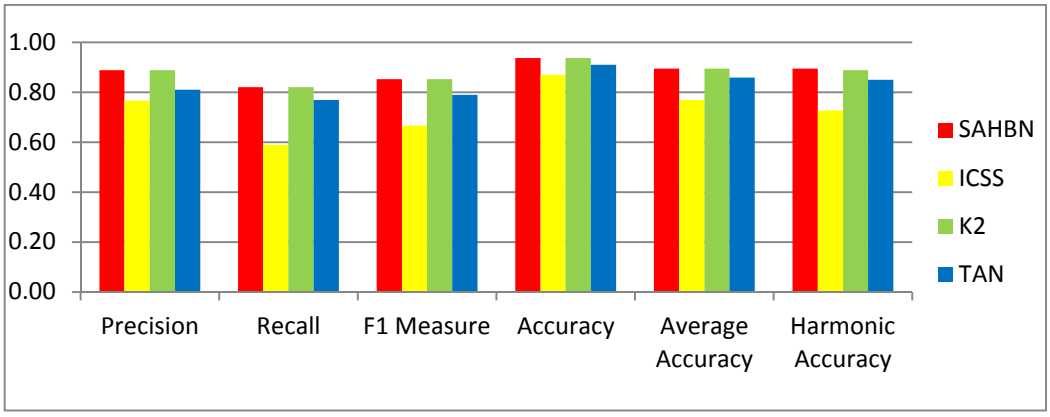


Figure 4-10 DNA Repair Gene–PPI Data Set Sixth Experiment Results

- Experiment 7

  The seventh experiment used (InfoGainAttributeEval+Ranker) attribute selection method. Accordingly, the results of the seventh experiment show that the SAHBN model significantly exceeded the ICSS algorithm in all performance criteria. However, it was slightly outperformed by the K2 and TAN algorithms. Figure 4-11 (below) depicts the results obtained.

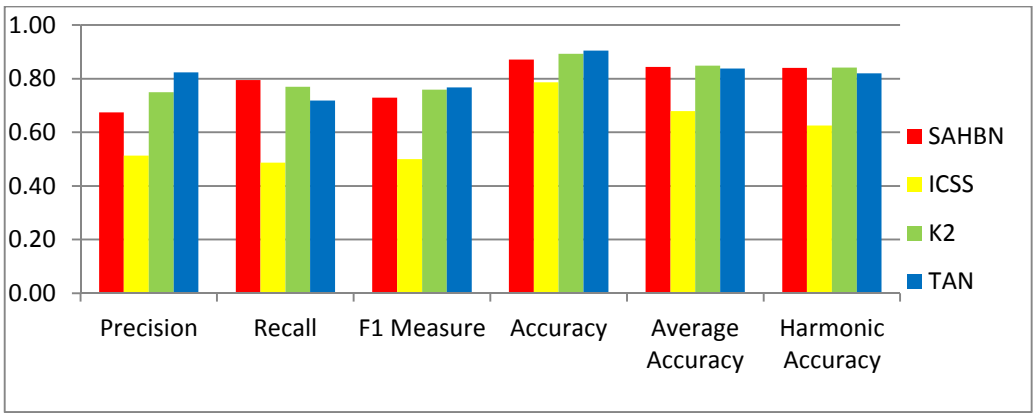


Figure 4-11 DNA Repair Gene–PPI Data Set Seventh Experiment Results

- Experiment 8

  The eighth experiment used (GainRatioAttributeEval+Ranker) attribute selection methods. Accordingly, the results of the eighth experiment indicate that the SAHBN model was either as good as, or slightly better than, the existing algorithms in terms of precision and accuracy. Additionally, it outperformed the K2 and TAN algorithms in terms of recall, F1 measure, average and harmonic accuracies. However, SAHBN was exceeded by the ICSS algorithm for the same performance criteria. Figure 4-12 (below) depicts the results obtained.

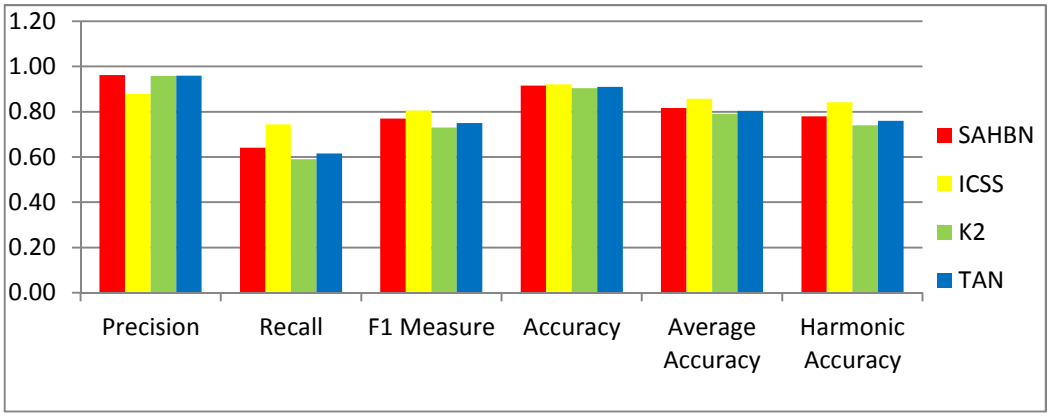


Figure 4-12 DNA Repair Gene–PPI Data Set Eighth Experiment Results

- Experiment 9

  The ninth experiment used (CorrelationAttributeEval + Ranker) attribute selection methods. Accordingly, the results of the ninth experiment show that the SAHBN model outperformed the ICSS algorithm in all performance criteria. Additionally, it exceeded the performance of the TAN algorithm in terms of recall and harmonic accuracy. However, the performance of the K2 algorithm was slightly better than the SAHBN model. Figure 4-13 (below) depicts the results obtained.

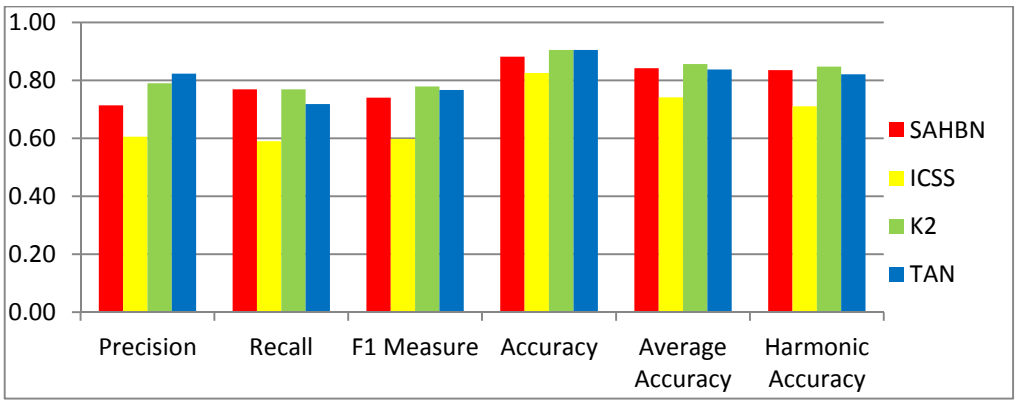


Figure 4-13 DNA Repair Gene–PPI Data Set Ninth Experiment Results

- Experiment 10

  The tenth experiment used (ReliefFAttributeEval + Ranker) attribute selection methods. Accordingly, the results of the tenth experiment show that the SAHBN model outperformed the ICSS and K2 algorithms in terms of recall, F1 measure, average and harmonic accuracies. However, it was exceeded by the TAN algorithm in all performance criteria. Figure 4-14 (below) depicts the results obtained.

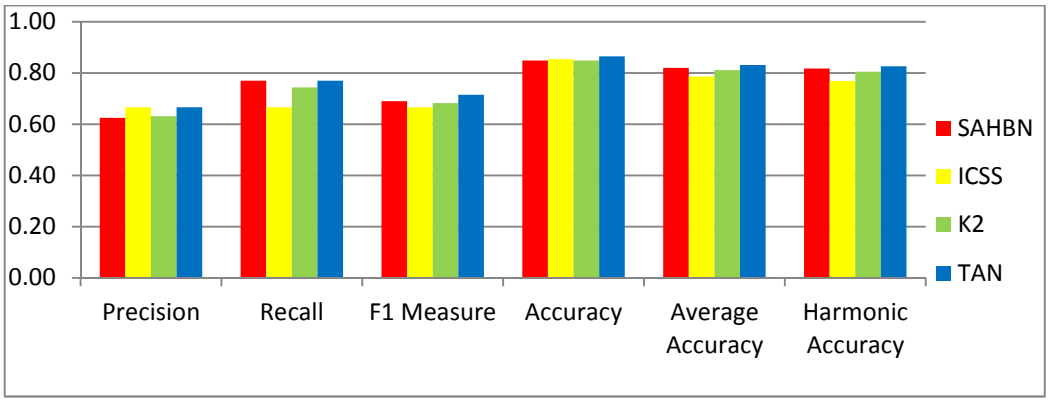


Figure 4-14 DNA Repair Gene–PPI Data Set Tenth Experiment Results

- Experiment 11

  The eleventh experiment used (OneRAttributeEval + Ranker) attribute selection methods. Accordingly, the results of the eleventh experiment show that the SAHBN model outperformed the ICSS and K2 algorithms in terms of all performance criteria. Additionally, it exceeded the TAN algorithm in terms of recall, F1 measure, average and harmonic accuracies. Figure 4-15 (below) depicts the results obtained.

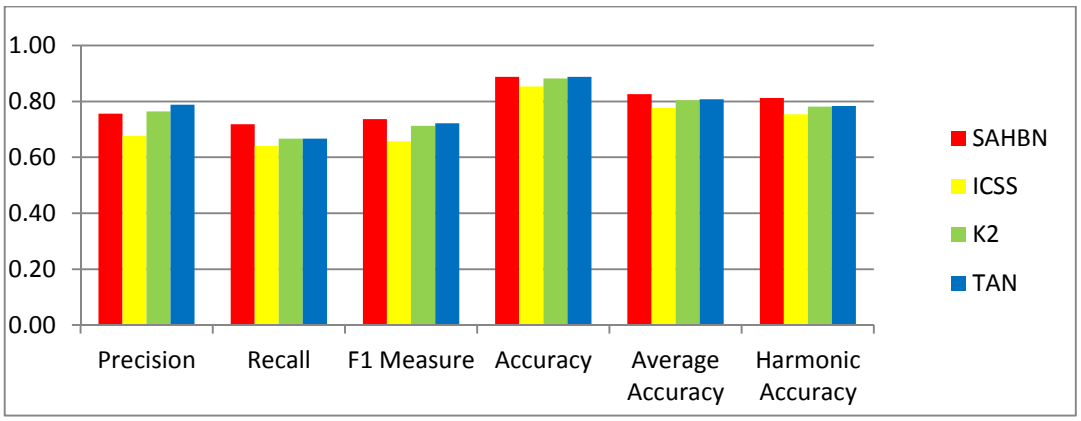


Figure 4-15 First DNA Repair Gene–PPI Data Set Eleventh Experiment Results

Having discussed the results obtained from the DNA repair gene–PPI data set, the next subsection of this chapter will explain the results of the DNA repair gene–Gene2GO data sets.

- DNA repair gene–Gene2GO (critical value = 2.706) data set

Table 6-2 in Appendix A presents the results obtained from applying the SAHBN model to the DNA repair gene–Gene2GO (critical value = 2.706) data set and compares it with the Hill-Climbing and Tabu Bayesian classification algorithms. It can be seen that six scoring methods were used for each algorithm, namely, loglik, bde, mbde, aic, bic and K2. Additionally, the process was repeated 11 times using different combinations of attribute selection method. Thus, the total number of experiments implemented for this data set is 132 experiments.

Accordingly, the implemented tests were summarised in terms of the number of experiments when the SAHBN model outperformed, was equal to or less than the existing methods. This was done for all performance criteria, and the results are explained in Table 4-4 (below).

Table 4-4 DNA Repair Gene–Gene2GO (CV = 2.706) Data Set Result Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	62	74	72	67	70	74	419
Equal to	21	10	11	22	10	10	84
Less than	49	48	49	43	52	48	289
Total	132	132	132	132	132	132	

Table 4-4 (above) reveals that the SAHBN model outperformed the existing algorithms in all performance criteria. For instance, SAHBN exceeded the existing algorithms in 74 experiments out of 132 experiments with respect to recall and

harmonic accuracy. Additionally, the number of experiments when SAHBN exceeded the existing algorithms in terms of F1 measure and average accuracy was 72 and 70 respectively. Overall, SAHBN surpassed the existing algorithms in more than 50% of the total number of tests for all performance criteria except for precision. This can be seen clearly in Figure 4-16 (below).

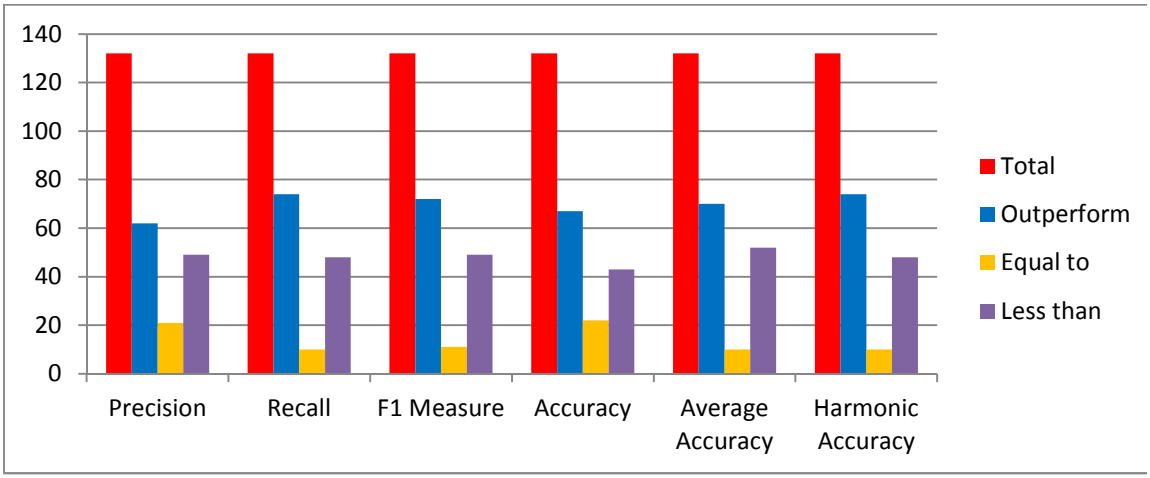


Figure 4-16 DNA Repair Gene–Gene2GO (CV = 2.706) Data Set Result Summary

Although the SAHBN model outperformed the existing algorithms in 419 experiments, its overall performance declined compared to the results obtained from the DNA repair gene–PPI data set. It seems possible that this decline is due to the chi-squared critical value used to remove attributes with low prediction power. Consequently, the same process was repeated using a higher critical value, and the results are presented in the next subsection.

- DNA repair gene–Gene2GO (critical value = 3.841) data set

  As mentioned in the previous subsection, the performance of the SAHBN model declined slightly in the DNA repair gene–Gene2GO (CV = 2.706) data set compared to its performance when it was applied to the DNA repair gene–PPI data set. Hence, a higher critical value was used to preserve only those attributes with a higher dependency on the label class. Accordingly, the results are presented in Table 6-3 in appendix A.

  The DNA repair gene–Gene2GO (CV = 2.706) data set was tested using the SAHBN model, Hill-Climbing and Tabu Bayesian classification algorithms. Likewise, the DNA repair gene–Gene2GO (critical value = 3.841) data set used the same algorithms, with the same scoring and attribute selection methods. Accordingly, 132 experiments were performed and summarised in terms of the number of times the SAHBN model

was outperformed, equal to or less than the existing algorithms, as explained in Table 4-5 (below).

Table 4-5 DNA Repair Gene–Gene2GO (CV = 3.841) Data Set Result Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	62	75	81	71	78	73	440
Equal to	14	7	5	24	11	6	67
Less than	56	50	46	37	43	53	285
Total	132	132	132	132	132	132	

The results presented in Table 4-5 indicate that the SAHBN model outperformed the existing algorithms in all performance criteria. For example, the number of experiments when SAHBN outperformed the existing algorithms in terms of F1 measure was 81 out of 132 experiments. Additionally, SAHBN exceeded the existing algorithms with respect to harmonic accuracy and average accuracies in 71, 73 and 78 experiments respectively. Overall, SAHBN outperformed the existing algorithms in 440 tests out of 792 tests. Compared to the results of the DNA repair gene–Gene2GO (CV = 2.706) data set, the total number of tests when the SAHBN model outperformed the existing algorithms increased slightly. Hence, it could be claimed that using a higher critical value in the chi-squared test led to better results. Figure 4-17 (below) clarifies the summary of the third data set results, as presented in Table 4-5.

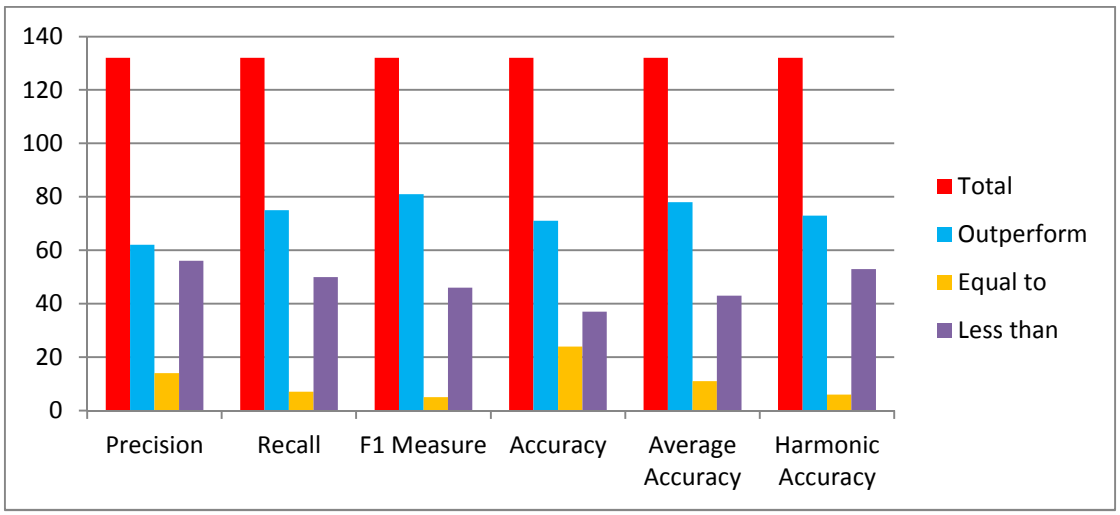


Figure 4-17 DNA Repair Gene–Gene2GO (CV = 3.841) Data Set Results Summary

In summary, in the first case study, the SAHBN model was tested using three data sets. While the first data set was created based on the protein–protein interaction database, the second and third data sets used the Gene2GO annotation database. The testing process was designed to include five classification algorithms, six scoring approaches and eleven combinations of attribute selection methods. In total 330

experiments were implemented and six performance criteria (i.e. precision, recall, F1 measure, accuracy, average accuracy and harmonic accuracy) were measured.

The analysis of these tests indicates that the proposed SAHBN model did better in the first database than the second and third data sets. Additionally, SAHBN demonstrated a very competitive performance compared to the existing classification algorithms.

In the next section the experimental results of the second case study are discussed in detail.

4.3.2 Model Organisms Case Study

As discussed in the previous pages, studying the human ageing process has faced many challenges. A long time frame and ethical difficulties are good examples of these challenges. Hence, researchers have used various model organisms to study the ageing process [137], [140]. Likewise, the classification SAHBN model proposed in this thesis was applied to the data of two model organisms and the obtained results compared with the existing classification algorithms. The data pre-processing and the experimental results are explained in the following subsections.

4.3.2.1 Data Set Characteristics and Pre-processing

In order to create the data sets used in this case study, the HAGR gene age [146] and the Gene2GO [149] databases were used. Accordingly, the actions involved in the data set creation process is summarised in Figure 4-18 (below).

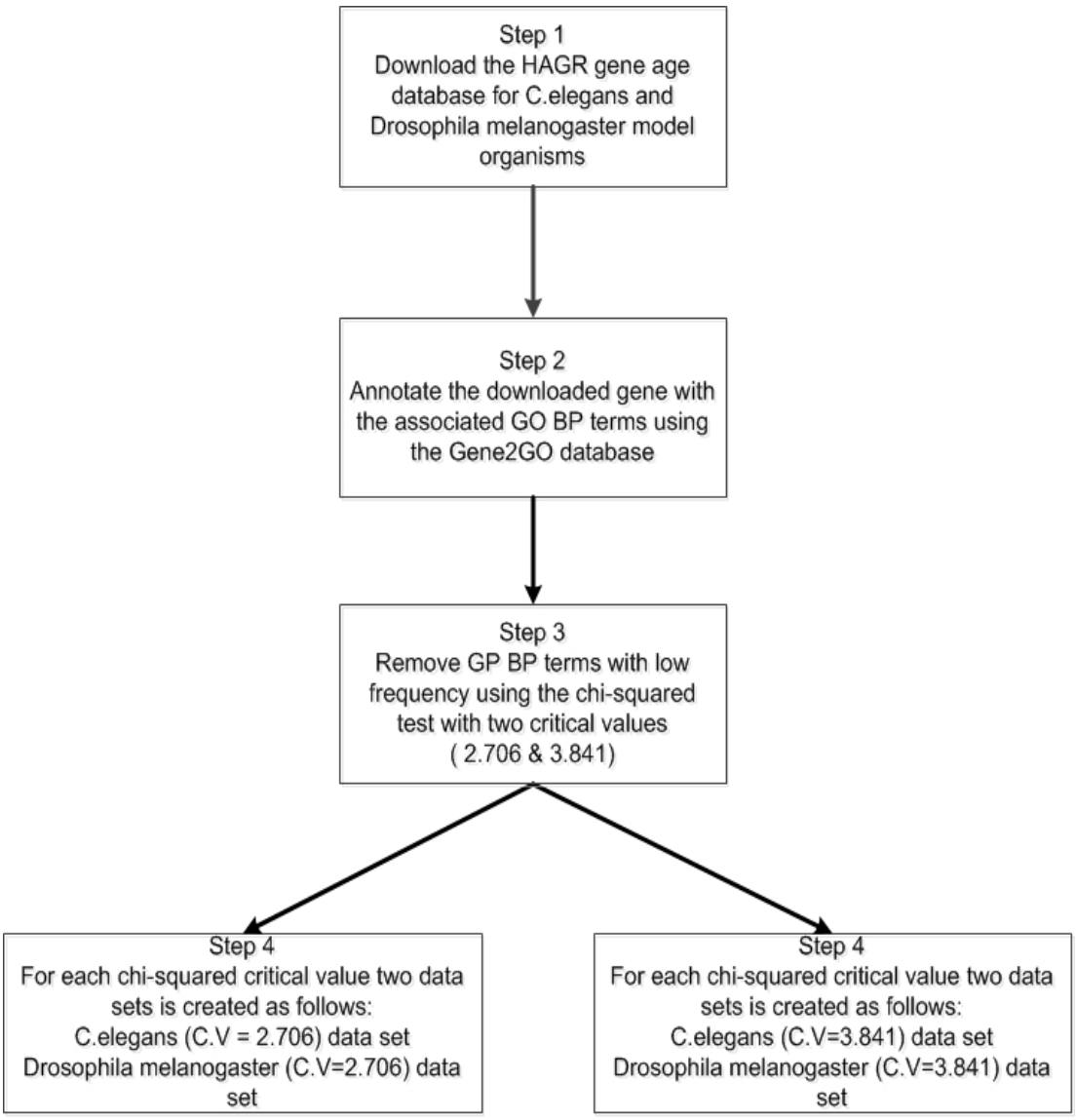


Figure 4-18 Model Organisms' Case Study Data Set Creation Process

The data set creation process is explained further in the following steps:

I. Download the HAGR gene age database for the following model organisms: 1) Caenorhabditis elegans (C.elegans); 2) Drosophila melanogaster.

II. Match each entry in the downloaded HAGR gene age database to a gene in the Gene2GO database and then extract the associated GO BP terms.

III. Remove GO terms that have low frequency and possess no or very low prediction power. [143] used a predefined threshold range from 4 to 10. In a slightly different way, this research used the chi-squared test to measure the dependency between each attribute and the label class and then to remove those attributes that appeared to be independent from the label class. Two critical values were used in the chi-squared test, namely, 2.706 and 3.841.

IV. Finally, each entry in the HAGR database is represented by a set of BP GO terms associated with its gene. The value of the GO terms is equal to 1 if it is associated with the given gene; otherwise it is equal to 0.

Eventually, two data sets were created for each model organism (one data set for each critical value). The characteristics of these data sets are summarised in Table 4-6 (below),

Table 4-6 Model Organisms' Case Study Data Set Characteristics

No.	Organism Model	Chi-Squared Critical Value	Number of Records	Initial Attributes Number	Chi-Squared Reduced Attr. Number
1	C. elegans	2.706	500	1012	65
2		3.841	500	1012	36
3	Drosophila melanogaster	2.706	115	860	44
4		3.841	115	860	15

4.3.2.2 Experimental Results

As described on the previous page, two data sets were created for each model organism using two critical values in the chi-squared test. Consequently, four data sets were created for two model organisms, namely, C.elegans and Drosophila melanogaster. The experimental results obtained from applying the SAHBN model to these data sets are explained in the following subsections.

- C.elegans–Gene2GO (CV = 2.706) data set

Table 6-4 in Appendix A presents the results obtained from applying the SAHBN model, Hill-Climbing and Tabu Bayesian algorithms to the first data set listed in Table 4-6 (above). Similarly to the previous case study, eleven attribute selection and six scoring methods were used. Additionally, the initial attributes list was reduced using the chi-squared test, with critical value equal to 2.706 to remove those attributes with low prediction power. Consequently, 124 experiments were implemented on the C.elegans–Gene2GO (CV = 2.706) data set, and the overall results were summarised in terms of when the SAHBN model outperformed, was equal to or less than the existing algorithms. Table 4-7 (below) summarises the results obtained.

Table 4-7 C.elegans–Gene2GO (CV = 2.706) Data Set Results Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	94	33	38	92	92	112	461
Equal to	0	3	3	4	6	2	18
Less than	30	88	83	28	26	10	265
Total	124	124	124	124	124	124	

It can be seen that the overall number of experiments show that the SAHBN model significantly outperformed the existing algorithms in terms of precision, accuracy, average accuracy and harmonic accuracy. However, it was exceeded by the existing algorithms with respect to the recall and F1 measure. For example, out of 124 experiments the SAHBN model outperformed the existing algorithms in 112 experiments in terms of harmonic accuracy. Likewise, the number of experiments when the accuracy of the SAHBN model was higher than the existing algorithms is 92. This can be seen clearly in Figure 4-19 (below).

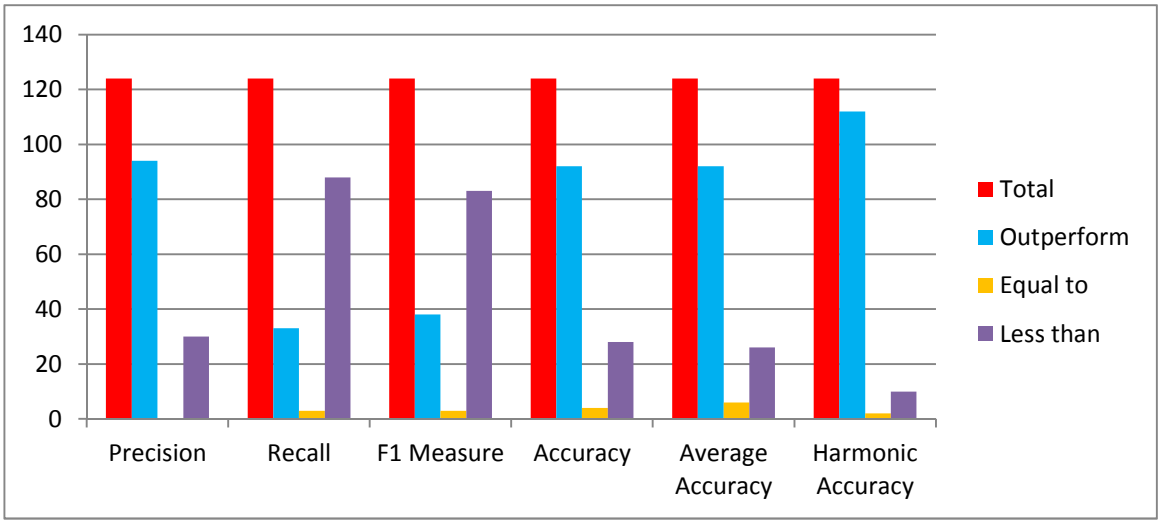


Figure 4-19 C.elegans–Gene2GO (CV = 2.706) Data Set Results Summary

Table 4-7 reveals that the SAHBN model generated good results compared to the existing algorithms. In the next subsection the experimental results of the C.elegans–Gene2GO (CV = 3.841) data set will be discussed in detail.

- C.elegans–Gene2GO (CV = 3.841) data set

In contrast to the C.elegans–Gene2GO (CV = 2.607) data set pre-processing procedure, the C.elegans–Gene2GO (CV = 3.841) data set attributes list was reduced using the chi-squared test with higher critical value (CV = 3.841), which attempts to preserve only those attributes that have higher dependency with the label class. Accordingly, the created data set was tested using the same combinations of attribute selection methods, Bayesian classification algorithms and scoring methods used to test the C.elegans–Gene (CV = 2.607) data set. Consequently, the results obtained are explained in Table 6-5 in Appendix A.

The results presented in Table 6-5 in Appendix A show that the SAHBN model outperformed the existing algorithms in many experiments. For instance, SAHBN model accuracy, average and harmonic accuracy values for the first, second, sixth, seventh, ninth

and tenth attribute selection methods were higher than their corresponding values for all the existing algorithms. Thus, and in order to further illustrate the obtained results, Table 6-5 was summarised to reflect the number of experiments when the SAHBN model outperformed, was equal to or less than the existing algorithm. Table 4-8 (below) summarises the C.elegans–Gene2GO (CV = 3.841) data set results.

Table 4-8 C.elegans–Gene2GO (CV = 3.841) Data Set Results Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	102	37	54	111	119	120	543
Equal to	0	3	0	6	3	0	12
Less than	30	92	78	15	10	12	237
Total	132	132	132	132	132	132	

The results summary of the C.elegans–Gene2GO (CV = 3.841) data set presented in Table 4-8 (above) confirms the findings of the C.elegans–Gene2GO (CV = 2.706) data set. It shows that the SAHBN model did much better in terms of precision, recall, accuracy, average and harmonic accuracies compared to the existing algorithms. However, it was beaten by the existing algorithms with respect to recall and F1 measure. Additionally, using higher critical values in the chi-squared test led to better overall quality. This can be seen in the total number of tests when the SAHBN model outperformed the existing algorithms. While the total number of tests when SAHBN outperformed the existing algorithms in the C.elegans–Gene2GO (CV = 2.706) data set was 461 tests, it increased to 543 tests in the C.elegans–Gene2GO (CV = 3.841) data set.

The findings summary of the C.elegans–Gene2GO (CV = 3.841) data set presented in Table 4-8 is depicted in Figure 4-20 (below).

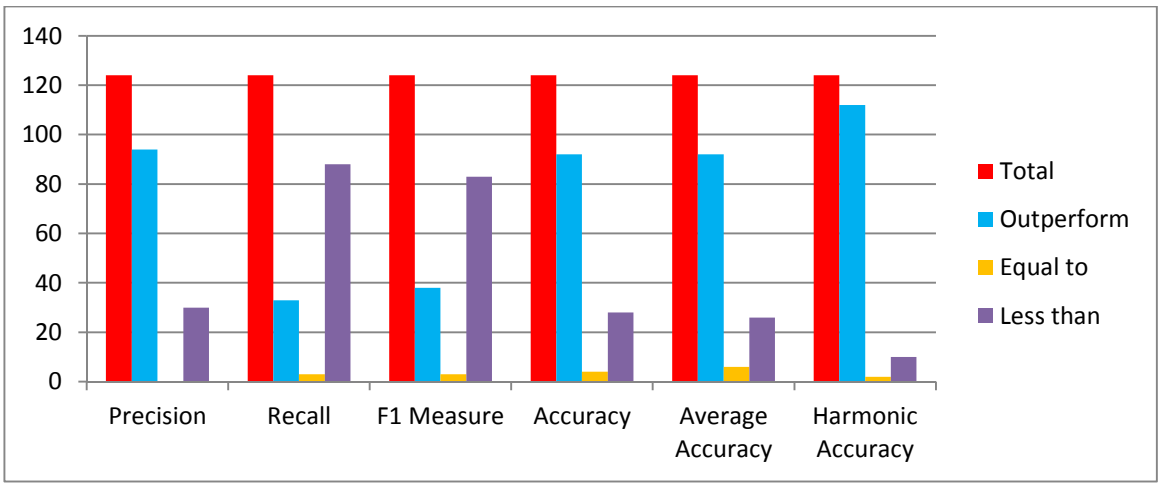


Figure 4-20 C.elegans–Gene2GO (CV = 3.841) Data Set Results Summary

Thus far, the data sets related to the C.elegans organism model have been discussed; the next subsections will explain the Drosophila melanogaster organism model and discuss the third and fourth data sets listed in Table 4-6 (above).

- D.melanogaster–Gene2GO (CV = 2.706) data set

Similar to the C.elegans–Gene2GO (CV = 2.706) and C.elegans–Gene2GO (CV = 3.841) data sets of this case study, the D.melanogaster–Gene2GO (CV = 2.706) data set was tested using eleven attribute selection methods and then the performance of the SAHBN model was compared with two Bayesian-based classification algorithms using six scoring approaches in each. Accordingly, the results are presented in Table 6-6 in appendix A.

Table 6-6 indicates that the SAHBN model demonstrated a good performance quality compared to the existing algorithms. For instance, the SAHBN model outperformed the existing algorithms in terms of accuracy, average and harmonic accuracies in the third, ninth and eleventh attribute selection methods. The results presented in Table 6-6 were further summarised in terms of when the SAHBN model outperformed, was equal to or less than the existing algorithms. The results summary is presented in Table 4-9 (below).

Table 4-9 D.melanogaster–Gene2GO Data Set Results Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	50	104	102	96	94	122	568
Equal to	3	2	2	17	14	0	38
Less than	79	26	28	19	24	10	186
Total	132	132	132	132	132	132	

Table 4-9 shows that the number of experiments when the SAHBN model outperformed the existing algorithms was significantly higher in terms of recall, F1 measure, accuracy, average and harmonic accuracies. However, the existing algorithm outperformed the

proposed SAHBN model with respect to precision. This can be seen clearly in Figure 4-21 below.

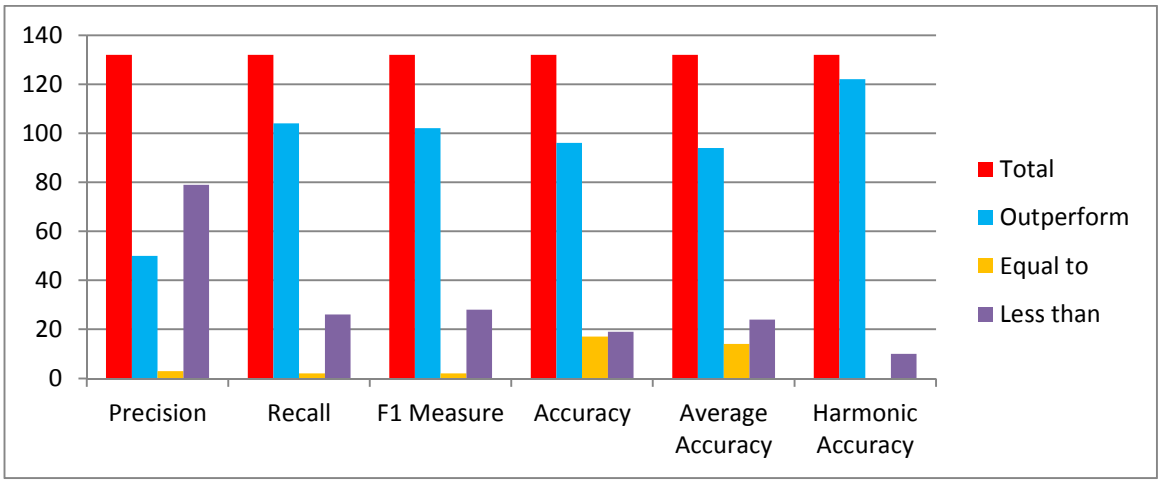


Figure 4-21 D.melanogaster–Gene2GO (CV = 2.706) Data Set Results Summary

As described in Table 4-6 on page 101, four data sets were created in the organism model case study. Thus, in the next subsection the results of the fourth data set will be discussed.

- D.melanogaster–Gene2GO (CV = 3.841) data set

Unlike the D.melanogaster–Gene2GO (CV = 2.706) data set, the attribute list of the D.melanogaster–Gene2GO (CV = 3.841) data set was reduced using a higher critical value in the chi-squared test, namely, 3.841. However, the numbers of algorithms, scoring and attribute selection methods were similar to the D.melanogaster–Gene2GO (CV = 2.706) data set. Accordingly, the obtained results are presented in Table 6-7 in appendix A.

The results presented in Table 6-7 show that the SAHBN model outperformed the existing algorithms in various experiments. For instance, the values of accuracy, average and harmonic accuracies for the third and fourth attribute selection methods of the SAHBN model were higher than the corresponding values of the existing algorithms. Similar to other data sets, the results presented in Table 6-7 were further explained in terms of when the SAHBN model outperformed, was equal to or less than the existing algorithms. Accordingly, the results summary is described in Table 4-10 (below).

Table 4-10 D.melanogaster–Gene2GO Data Set Results Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	105	37	85	110	98	105	540
Equal to	12	30	8	7	17	4	78
Less than	15	65	39	15	17	23	174
Total	132	132	132	132	132	132	

Table 4-10 (above) indicates that the SAHBN model outperformed the existing algorithms in terms of precision, F1 measure, accuracy, average and harmonic accuracy. However, it was exceeded by the existing algorithms with regard to recall. This can be seen clearly in Figure 4-22 (below).

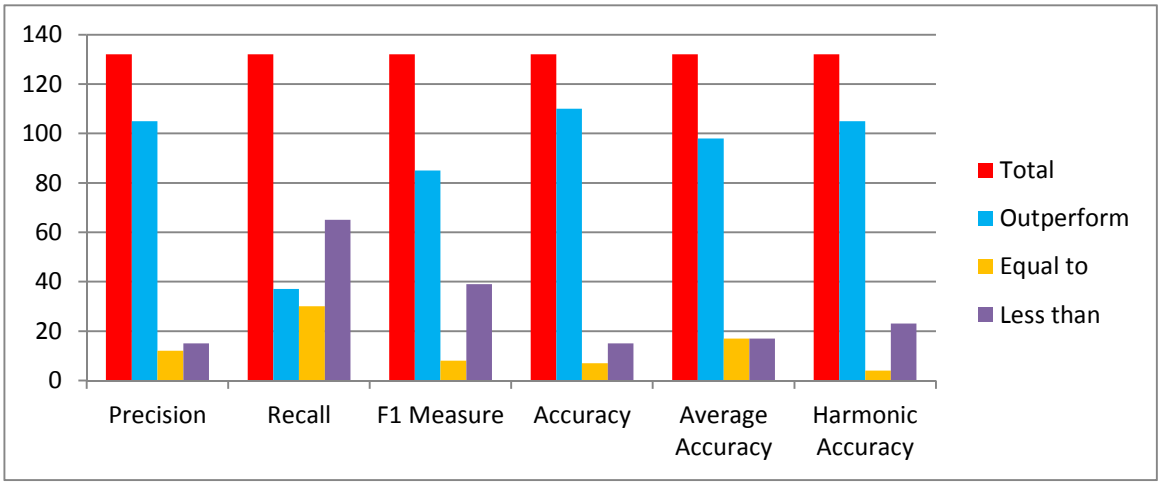


Figure 4-22 D.melanogaster–Gene2GO Data Set Results Summary

So far, this chapter has analysed the results obtained from the first and second case studies and has argued that the proposed model demonstrated a very competitive performance compared to the existing algorithms. The last part of this chapter is devoted to discussing in detail the results obtained from the third case study.

4.4 Protein Hub Case Study

Hub proteins have been defined as proteins with special topological and functional significance. There is some evidence to suggest that hub proteins are engaged in a large number of protein interactions. Additionally, they play an essential role in the organisation and function of cellular protein interaction networks (PINs). Hence, the identification of hub proteins will help in the understanding of cellular functions, identifying novel drug targets and the rational design of large-scale pull-down experiments. Therefore, it is important to identify hub proteins not only via experimental data but also using bioinformatics prediction techniques [150], [151]. Thus, the SAHBN classification model proposed in this thesis was applied to the protein data set of the *Homo sapiens* species.

4.4.1 *Homo Sapiens* Protein Hub Data Set

The term *Homo sapiens* has been used as a scientific name to refer to the human species. While Homo is the human genus, which includes Neanderthals and other species similar to humans, *Homo sapiens* is considered to be the only surviving

species of the genus Homo [152]. The following subsection gives a brief outline of the steps involved in the process of *Homo sapiens* protein hub data set creation.

4.4.1.1 Data Set Characteristics and Pre-processing

The IntAct [153] and the protein Uniprot [154] GO annotation databases were used to create the *Homo sapiens* protein data set. The data set creation process followed the following steps:

I. Download the protein–protein interaction database from the IntAct database for the *Homo sapiens* species.
II. Extract the UniPort Code for the first interacted protein.
III. Count the number of interactions for each protein.
IV. [150] used the position of the sharp turn in the accumulative protein interaction distribution plot to define hub/non-hub proteins. Likewise, this thesis used the same approach to classify the downloaded protein database into either hub or non-hub proteins.
V. Each protein is represented in the form of GO BP terms based on the annotation data provided by the UniPort database.
VI. Finally, each protein is represented by a set of GO BP terms. The value of the GO BP term is equal to 1 if it is associated with the given gene; otherwise, it is equal to 0.

Eventually, the created data set consisted of 500 records, each representing a protein, and 6,455 GO BP terms, which represent the prediction attributes.

4.4.1.2 Experimental Results

So far, the empirical implementations have only used the score-based Bayesian-based classification algorithms provided by RStudio. However, according to [155], RStudio provides other types of Bayesian-based classification algorithms, namely, constraint-based algorithms. Hence, this case study used four constraint-based and two score-based algorithms. Additionally, three scoring and five constraint methods were used with the score-based and constraint-based algorithms respectively. In total, the proposed model was compared against 26 different combinations of score- and constraint-based algorithms for each attribute selection method. Finally, similar to other case studies, 11 attribute selection methods were used. Accordingly, the obtained results are presented in Table 6-8 in Appendix A.

The results presented in Table 6-8 reveal that the SAHBN model significantly outperformed the constraint-based algorithms. Additionally, it exceeded some of the score-based algorithms in various tests. For example, the SAHBN model outperformed all other existing algorithms (score-based and constraint-based) in terms of recall, F1 measure, accuracy, average and harmonic accuracy in the fourth, fifth and sixth attribute selection methods. The results presented in Table 6-8 were further illustrated in terms of when the SAHBN model outperformed, was equal to or less than the existing algorithms. Accordingly, the summarised results are presented in Table 4-11 (below).

Table 4-11 *Homo Sapiens* Protein Hub Data Set Result Summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	167	274	272	270	270	272	1525
Equal to	5	0	2	2	2	1	12
Less than	104	2	2	4	4	3	119
Total	276	276	276	276	276	276	

Table 4-11 (above) shows that the SAHBN model outperformed the existing algorithms in all quality measures. For instance, out of 276 experiments the proposed model exceeded the existing algorithms in 270 experiments with respect to accuracy and average accuracy. Likewise, the proposed model exceeded the existing algorithms in terms of F1 measure and harmonic accuracy in 272 experiments out of 276 experiments, and in 274 experiments with regard to recall. This can be further clarified in Figure 4-23 (below).

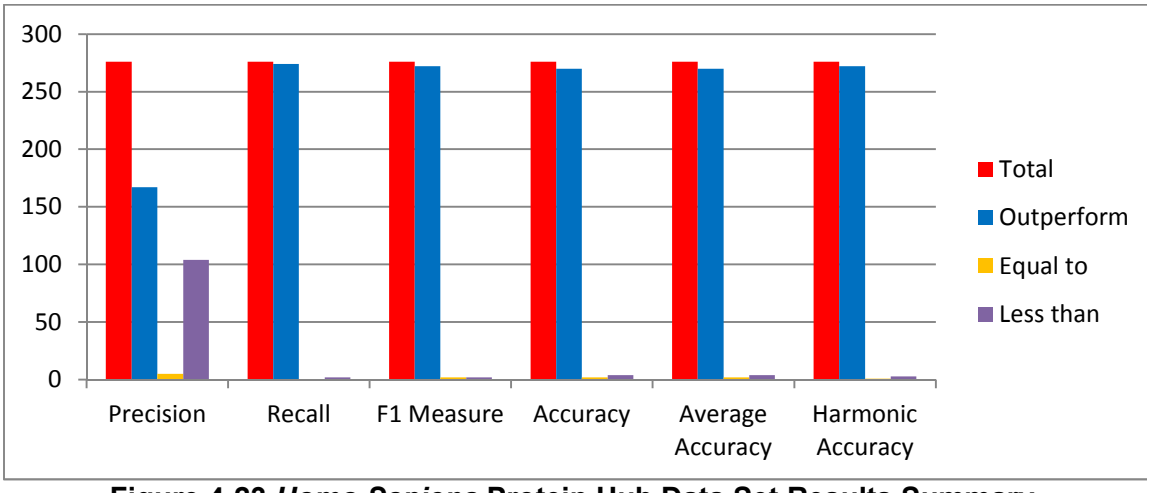


Figure 4-23 *Homo Sapiens* Protein Hub Data Set Results Summary

In summary, this chapter has discussed the experimental results obtained from applying the SAHBN model to eight data sets, which were organised in three case studies. In total 1,093 experiments were performed using 11 different combinations of attribute selection method

and five Bayesian-based classification algorithms. Additionally, three scoring and five constraint methods were used.

The findings extracted from the experimental results suggested that the SAHBN model demonstrated a very competitive performance compared to the existing algorithms. However, it cannot be claimed that SAHBN surpassed all the existing algorithms in all experiments. In the next chapter the overall experimental results will be comprehensively analysed.

4.5 Results Analysis

In order to analyse the overall performance of the proposed SAHBN model, the obtained results were summarised in two different ways:

a. **Frequency table:** this summarises the obtained results in terms of the frequency when the SAHBN model outperformed, was equal to or less than other Bayesian-based classification algorithms. This was done for all six performance criteria discussed in Section 4.2. Accordingly, the final results are presented in Table 4-12 (below).

Table 4-12 total results summary

Proposed Model	Precision	Recall	F1 Measure	Accuracy	Average Accuracy	Harmonic Accuracy	Row Total
Outperform	655	657	722	832	845	905	4616
Equal to	61	64	36	92	68	25	346
Less than	377	372	335	169	180	163	1596
Total	1093	1093	1093	1093	1093	1093	

Table 4-12 reveals that the proposed SAHBN model outperformed the existing algorithms in all quality criteria. For instance, out of 1,093 experiments SAHBN outperformed the existing algorithms in 905 tests with respect to harmonic accuracy. Additionally, it exceeded the existing algorithms in 845 and 832 tests in terms of average accuracy and accuracy, respectively. The results presented in Table 4-12 are visualized in Figure 4-24 (below).

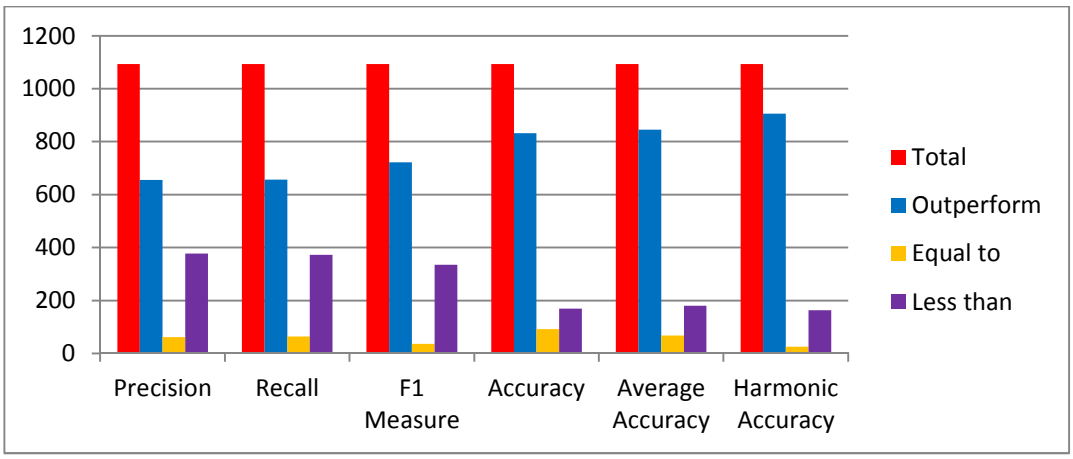


Figure 4-24 Total Results Summary

b. **Arithmetic mean table:** as a result of the fact that the proposed SAHBN model was compared with a wide range of Bayesian-based classification algorithms provided by Weka and Rstudio, the performance of SAHBN was compared against each one of these algorithms. This was done by calculating the arithmetic mean for each performance criteria for SAHBN and comparing it with the corresponding arithmetic mean of the same quality criteria of other algorithms. For example, the arithmetic mean of SAHBN's accuracy was calculated by summing the values of SAHBN accuracies in all experiments and then dividing the summed value by the number of experiments. Likewise, the accuracy arithmetic mean for algorithm "A" was calculated in the same way. Eventually, the two values were compared. This process was repeated for all 6 performance criteria and for all 35 algorithm combinations against which SAHBN was compared. Consequently, the obtained results are presented in Table 4-13 (below).

Table 4-13 Overall Performance Arithmetic Mean Results

No.	Algorithm	Precision A. Mean	Recall A. Mean	F1 Measure A. Mean	Accuracy A. Mean	Ave. Acc. A. Mean	H. Acc. A. Mean
1	**SAHBN**	**0.78**	**0.59**	**0.64**	**0.75**	**0.71**	**0.64**
2	ICSS	0.69	0.61	0.64	0.85	0.76	0.73
3	K2	0.79	0.75	0.76	0.89	0.84	0.83
4	TAN	0.80	0.72	0.75	0.90	0.83	0.82
5	hc(loglik)	0.71	**0.68**	0.63	0.70	0.65	0.48
6	hc(bde)	0.76	0.55	0.60	0.70	0.66	0.54
7	hc(mbde)	0.77	0.56	0.60	0.71	0.67	0.54
8	hc(aic)	0.70	0.58	0.54	0.65	0.59	0.36
9	hc(bic)	0.71	0.56	0.49	0.64	0.58	0.30
10	hc(K2)	0.77	0.58	0.61	0.71	0.67	0.57
11	tabu(loglik)	0.72	**0.69**	0.63	0.70	0.65	0.48

12	tabu(bde)	0.72	0.59	0.60	0.68	0.64	0.51
13	tabu(mbde)	0.72	0.58	0.60	0.68	0.64	0.51
14	tabu(aic)	0.69	0.54	0.51	0.64	0.59	0.36
15	tabu(bic)	0.69	0.56	0.50	0.63	0.57	0.31
16	tabu(K2)	0.75	**0.61**	0.63	0.71	0.67	0.58
17	gs(mi)	0.48	0.45	0.44	0.46	0.46	0.41
18	gs(mi-adf)	0.49	0.43	0.43	0.46	0.46	0.41
19	gs(mi-sh)	0.53	0.39	0.41	0.48	0.48	0.41
20	gs(X2)	0.45	0.46	0.45	0.45	0.45	0.44
21	gs(X2-adf)	0.44	0.44	0.44	0.45	0.45	0.42
22	iamb(mi)	**0.81**	0.35	0.46	0.62	0.62	0.48
23	iamb(mi-adf)	0.65	0.43	0.50	0.57	0.57	0.51
24	iamb(mi-sh)	**0.80**	0.36	0.47	0.62	0.62	0.48
25	iamb(X2)	0.60	0.41	0.45	0.52	0.52	0.44
26	iamb(X2-adf)	0.48	0.44	0.45	0.47	0.47	0.44
27	fast.iamb(mi)	0.73	0.43	0.49	0.58	0.58	0.47
28	fast.iamb(mi-adf)	0.68	0.47	0.53	0.59	0.59	0.53
29	fast.iamb(mi-sh)	0.63	0.48	0.52	0.56	0.55	0.50
30	fast.iamb(X2)	0.64	0.38	0.46	0.55	0.55	0.47
31	fast.iamb(X2-adf)	0.59	0.43	0.48	0.54	0.54	0.49
32	inter.iamb(mi)	**0.81**	0.36	0.47	0.62	0.62	0.48
33	inter.iamb(mi-adf)	0.66	0.46	0.51	0.57	0.57	0.50
34	inter.iamb(mi-sh)	**0.80**	0.35	0.46	0.62	0.62	0.48
35	inter.iamb(X2)	0.67	0.37	0.44	0.55	0.55	0.45
36	inter.iamb(X2-adf)	0.48	0.48	0.47	0.47	0.47	0.44

Table 4-13 reveals that the proposed SAHBN model was outperformed by TAN, K2 and ICSS algorithms in almost all quality criteria except for ICSS precision and F1 measure, when SAHBN either outperformed or was equal to ICSS. Additionally, it shows that SAHBN exceeded all other algorithms in almost all quality criteria, except for the following seven experiments (highlighted in red on Table 4-13):

- **Precision arithmetic mean**: for the following algorithms: iamb(mi), iamb(mi-sh), inter.iamb(mi) and inter.iamb(mi-sh).
- **Recall arithmetic mean:** for the following algorithms: hc(loglik), tabu(loglik) and tabu(K2).

The arithmetic mean results summary was further clarified in Figure 4-25 (below).

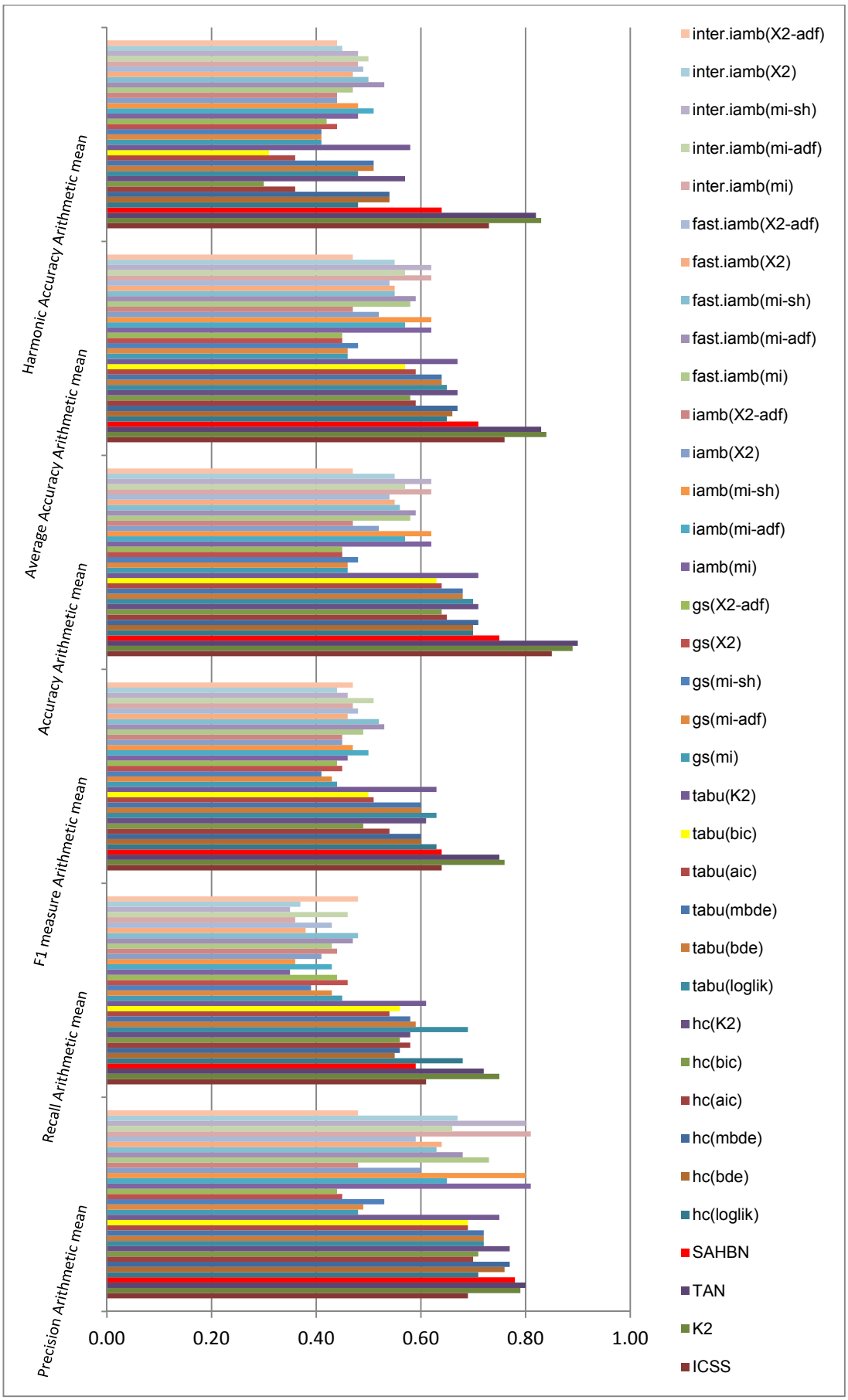


Figure 4-25 Overall arithmetic mean performance results summary

4.6 Discussion

The overall performance results presented in Table 4-12 and Table 4-13 and visualised in Figure 4-24 and Figure 4-25 reveal that the proposed SAHBN model demonstrated a very competitive performance when compared to the existing algorithms. For instance, Table 4-12 which summarise the results obtained in terms of frequency shows that SAHBN model outperformed existing algorithms in all performance criteria. However, Table 4-13 which analysis the experimental results using the arithmetic mean indicates that SAHBN model was surpassed by TAN,K2 and ICSS algorithms in all performance criteria.

TAN, K2 and ICSS algorithms were used to compare the performance of SAHBN in the first data set which consists of 178 records and 3,163 prediction attributes. Hence, it seems possible that these algorithms are able to handle data sets which have a large number of attributes and relatively small number of records.

Although SAHBN was outperformed by the TAN, K2 and ICSS algorithms, its performance exceeded all the other algorithms in Table 4-13. Hence, a deeper investigation into the aspects accommodated by the TAN, K2 and ICSS algorithms and overlooked by SAHBN is required. Furthermore, it seems that integration of the "is-a" semantic relation to check for data consistency and guide the SAHBN structure process did not cover the entire ontological knowledge for the targeted domain. Thus, further studies to integrate additional ontological knowledge are worth exploring.

In conclusion, the SAHBN model has laid the foundations for a systematic approach to integrating ontology and the HBN classifier in a flexible framework that can be applied to different domains. In fact, it is a step towards transforming the traditional BN learning process from being data-centred process which relies solely on the training data set into knowledge-centred process which accommodates a consensual knowledge in the form of ontology.

Additionally, it proposed an automatic way to consolidate ontology and BN theory in such a way that preserves the advantages of both. Finally, SAHBN highlighted the similarity between ontology and HBN concepts, which could be invested to represent the targeted domain in a consistent manner that reflects semantic relations between attributes. Overall, SAHBN investigated cutting-edge concepts in the semantic data mining area and explored some very challenging questions.

Chapter 5 Conclusion and Future Directions

Data classification is a branch of DM technology that aims to train a prediction model using a data set with a known label class. Then the trained model is used to predict the label class for unknown instances based on their observed attributes. However, recent research has argued that traditional DM techniques are a data-driven process and have not accommodated domain knowledge in the mining process. Hence, DM has faced a paradigm shift from a data-centred to a knowledge-centred process that aims to integrate domain knowledge into the mining process.

Although various approaches could be used to represent domain knowledge, the latest research in the DM field reported that ontology is the most convenient approach to represent domain knowledge for DM use because of its structural way of knowledge representation. Accordingly, the concept of semantic DM has been introduced recently. Semantic data mining has been defined as the DM task that systematically incorporates domain knowledge, especially formal semantics, into the mining process.

Despite its various advantages, ontology has been built based on description logic, which is a decidable fraction of the first-order logic. Hence, it cannot accommodate the uncertainty factor that is characteristic of many real-life situations. In contrast to ontology, Bayesian-based DM algorithms are well known for their powerful capability to work under uncertainty.

Recent research has pointed out that semantic DM is still in its early stages and ontology-based semantic DM approaches provide the most promising approach. Additionally, the challenges of developing automatic semantic data mining algorithms that systematically consolidate ontology and DM algorithms have not yet been met. Hence, this thesis proposed a flexible framework that automatically integrates ontology and the HBN classification algorithms in such a way that preserves the advantages of both. The objectives of this thesis are revisited in Section **Error! Reference source not found.**below.

5.1 Review of Research Contributions

The points presented in this section discuss the research objectives and investigate the degree to which they have been fulfilled:

1. **Conduct a comprehensive literature review centred on semantic DM, ontology-based data classification and integrating ontology with BN**: Chapter 2

reviewed the state-of-the-art literature in the area of semantic DM, ontology-based classification and integration ontology with BN. The findings of Chapter 2 highlighted the potential advantages of integrating ontology with the BN classifier. However, to the best of our knowledge, no framework that systematically consolidates ontology and a BN classifier has been proposed. Furthermore, despite the structural similarity between ontology and the HBN classifier, the advantages of integrating these two concepts have not been comprehensively investigated. Hence, the SAHBN classification model (discussed in Section 3.7) proposed in this thesis suggested an automatic, systematic and flexible framework to integrate ontology and HBN, as will be discussed in the next point.

2. **Develop an ontology-based HBN classification framework**: the proposed SAHBN classification model not only integrated ontology and HBN but also achieved the following objectives:

    - **Automatic:** unlike many of the approaches reviewed in the second chapter, the framework proposed in this thesis integrated ontology and the HBN classifier automatically without the need for human intervention.
    - **Systematic:** the SAHBN model proposed a predefined sequence of actions that have to be implemented in order regardless of the data set or the domain to which the model was applied. Hence, it is a standardised framework that is neither domain nor data-set-dependent.
    - **Flexible:** the proposed SAHBN model used the "is-a" ontological relation, which is used to represent domain knowledge as a hierarchical structure. Hence, any ontology using the "is-a" relation to represent domain knowledge can be converted into a SAHBN model and used as a classifier. Therefore, the proposed SAHBN framework is flexible and can be applied to different domains.

3. **Apply the proposed model into real-life datasets**: the proposed SAHBN classification model was applied to eight data sets organised in three case studies. In the first case study the SAHBN model was used to predict the effect of the DNA repair gene in the human ageing process. The second case study utilised the SAHBN model to classify genes in various model organisms as either an ageing-related or non-ageing-related gene. Finally, the SAHBN model was used to identify the protein hub in the third case study. In total, 1,093 experiments were

implemented and six quality criteria measured, namely, precision, recall, F1 measure, accuracy, average and harmonic accuracy. The process of SAHBN model implementation and data set pre-processing are discussed in detail in Chapter 4.

4. **Evaluate the performance of the SAHBN model against existing Bayesian-based classification algorithms**: as mentioned in the previous points the SAHBN model was tested using eight data sets, and six quality criteria were measured. Likewise, various Bayesian-based classification algorithms were applied to the same data sets and the same quality criteria were calculated. Then, the performance of the SAHBN model was compared against the existing algorithms. Additionally, the overall performance in terms of the frequency of when SAHBN outperformed, was equal to or less than the existing algorithms was summarised for all 1,093 experiments. Furthermore, the arithmetic mean for each quality criteria was measured for the SAHBN model and compared with the same quality criteria arithmetic mean for the existing algorithms. A detailed discussion of the evaluation process is presented in Chapters 4 and 5. Overall, SAHBN demonstrated a very competitive performance when compared to the existing algorithms.

In conclusion, the contribution of this thesis could be summarised in the following points:

1. Proposed an automatic, systematic and flexible framework to integrate ontology and the HBN. Based on the literature review, and to the best of our knowledge, no such framework has been proposed previously.
2. The proposed framework preserves the advantages of both ontology and Bayesian theory. It integrates the concept of Bayesian uncertainty with ontology without extending the ontology structure and adding probability-specific properties that violate the ontology standard structure.
3. The proposed model lays out a solid foundation to integrate other semantic relations such as equivalent, disjoint, intersection and union, as will be discussed in Section 5.2.

5.2 Future Directions

The future work of this thesis will focus on the following directions:

1. **Integrate more ontological relations:** at the current stage of this research only the "is-a" ontological relation has been used to check training data set consistency, the semantic relation between attributes, and to guide the HBN structure learning and

assist the CPT calculation. However, the SAHBN framework established a solid base to inject more semantic information into the mining process. Hence, the future work of this research will include the integration of ontological relations such as equivalent, disjoint, intersection and union. This can be done in two different ways, as follows:

a. The BayesOWL [74] model introduced the concept of logical nodes to the BN structure. In a similar way, the SAHBN model will construct four logical nodes to cover more ontological relations and investigate their effects on SAHBN's performance quality.
b. As was mentioned in Section 3.6, the proposed SAHBN model used the MAP probability estimation approach. Accordingly, the MAP approach used two positive real hyper-parameters to represent the imaginary true and false frequency of instances observed before starting the experiment [124]. Hence, the hyper-parameters can be utilised to reflect the semantic relation between concepts in the SAHBN model. For instance, if there is an "equivalent" relation between two ontology concepts, A and B, the hyper-parameters could be used to reflect this relation in terms of the frequency of the observed true and false instances. Consequently, the calculated CPTs will accommodate the ontological relations between concepts. Likewise, this can be done for other logical relations such as disjoint, union and intersection.

2. **Test the proposed SAHBN model using different hierarchical ontologies:** at the current stage of this research, the proposed SAHBN model has been tested using the GO. Thus, future work will include implementation of the SAHBN model in different domains using other hierarchical ontologies. In theory, the SAHBN model can be applied for any ontology using the "is-a" relation to represent the "general to specific" or "broader-than to narrow-than" relation between concepts to describe the domain knowledge.
3. **Use other GO hierarchies:**: According to [98], the structure of the GO consists of three hierarchies covering the biological process, molecular functions and cellular component concepts. However, SAHBN was tested using only the biological process hierarchy. Hence, future work will include integration of other hierarchies covering the molecular functions and cellular component concepts.

Appendix "A" Experimental Results

6.1 DNA repair gene-PPI dataset experimental results

Table 6-1 DNA repair gene-PPI dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Attr. Selection Reduction.	Sub-Super Class Reduction	Algorithm	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method										
1	CfsSubSetEval	BestFirst	3163	16	15	**SAHBN**	**0.84**	**0.82**	**0.83**	**0.93**	**0.89**	**0.88**
						ICSS algo.	0.68	0.54	0.60	0.84	0.73	0.68
						K2 algo.	0.86	0.79	0.83	0.93	0.88	0.87
						TAN algo.	0.86	0.79	0.83	0.93	0.88	0.87
2		GreedyStepwise	3163	24	23	**SAHBN**	**0.89**	**0.82**	**0.85**	**0.94**	**0.90**	**0.89**
						ICSS algo.	0.77	0.59	0.67	0.87	0.77	0.73
						K2 algo.	0.89	0.82	0.85	0.94	0.90	0.89
						TAN algo.	0.81	0.77	0.79	0.91	0.86	0.85
3	ConsistencySubsetEval	BestFirst	3163	10	9	**SAHBN**	**0.62**	**0.72**	**0.67**	**0.84**	**0.84**	**0.84**
						ICSS algo.	0.80	0.72	0.76	0.90	0.83	0.82
						K2 algo.	0.78	0.72	0.75	0.89	0.83	0.82
						TAN algo.	0.78	0.72	0.75	0.89	0.83	0.82
4		GreedyStepwise	3163	12	12	**SAHBN**	**0.60**	**0.69**	**0.64**	**0.83**	**0.78**	**0.77**
						ICSS algo.	0.55	0.56	0.56	0.80	0.72	0.68
						K2 algo.	0.50	0.69	0.58	0.78	0.75	0.74
						TAN algo.	0.58	0.59	0.58	0.81	0.73	0.71
5	FilteredSubsetEval	BestFirst	3163	16	15	**SAHBN**	**0.84**	**0.82**	**0.83**	**0.93**	**0.89**	**0.88**
						ICSS algo.	0.68	0.54	0.60	0.84	0.73	0.68
						K2 algo.	0.86	0.82	0.84	0.93	0.89	0.89
						TAN algo.	0.86	0.79	0.83	0.93	0.88	0.87

6		GreedyStepwise	3163	24	23	**SAHBN**	**0.89**	**0.82**	**0.85**	**0.94**	**0.90**	**0.90**
						ICSS algo.	0.77	0.59	0.67	0.87	0.77	0.73
						K2 algo.	0.89	0.82	0.85	0.94	0.90	0.89
						TAN algo.	0.81	0.77	0.79	0.91	0.86	0.85
7	InfoGainAttributeEval	Ranker	3163	24	22	**SAHBN**	**0.67**	**0.79**	**0.73**	**0.87**	**0.84**	**0.84**
						ICSS algo.	0.51	0.49	0.50	0.79	0.68	0.62
						K2 algo.	0.75	0.77	0.76	0.89	0.85	0.84
						TAN algo.	0.82	0.72	0.77	0.90	0.84	0.82
8	GainRatioAttributeEval	Ranker	3163	24	22	**SAHBN**	**0.96**	**0.64**	**0.77**	**0.92**	**0.82**	**0.78**
						ICSS algo.	0.88	0.74	0.81	0.92	0.86	0.84
						K2 algo.	0.96	0.59	0.73	0.90	0.79	0.74
						TAN algo.	0.96	0.62	0.75	0.91	0.80	0.76
9	CorrelationAttributeEval	Ranker	3163	24	22	**SAHBN**	**0.71**	**0.77**	**0.74**	**0.88**	**0.84**	**0.84**
						ICSS algo.	0.61	0.59	0.60	0.83	0.74	0.71
						K2 algo.	0.79	0.77	0.78	0.90	0.86	0.85
						TAN algo.	0.82	0.72	0.77	0.90	0.84	0.82
10	ReliefFAttributeEval	Ranker	3163	24	22	**SAHBN**	**0.63**	**0.77**	**0.69**	**0.85**	**0.82**	**0.82**
						ICSS algo.	0.67	0.67	0.67	0.85	0.79	0.77
						K2 algo.	0.63	0.74	0.68	0.85	0.81	0.81
						TAN algo.	0.67	0.77	0.71	0.87	0.83	0.83
11	OneRAttributeEval	Ranker	3163	24	22	**SAHBN**	**0.76**	**0.72**	**0.74**	**0.89**	**0.83**	**0.81**
						ICSS algo.	0.68	0.64	0.66	0.85	0.78	0.75
						K2 algo.	0.76	0.67	0.71	0.88	0.80	0.78
						TAN algo.	0.79	0.67	0.72	0.89	0.81	0.78

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.2 DNA repair gene-Gene2GO (CV = 2.706) dataset experimental results

Table 6-2 DNA repair gene-Gene2GO (CV = 2.706) dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Chi-Squared Test Reduction	Attr. Selection Reduction	Sub-Super Class Test Reduction	Algorithm	Score	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	735	335	35	30	**SAHBN**	**NA**	**0.92**	**0.59**	**0.72**	**0.90**	**0.79**	**0.74**
							Hill-Climbing (hc)	loglik	0.75	0.38	0.51	0.84	0.67	0.55
								bde	0.92	0.59	0.72	0.90	0.79	0.74
								mbde	0.92	0.59	0.72	0.90	0.79	0.74
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.84	0.54	0.66	0.88	0.75	0.69
							Tabu Search (tabu)	loglik	0.80	0.41	0.54	0.85	0.69	0.58
								bde	0.92	0.59	0.72	0.90	0.79	0.74
								mbde	0.92	0.59	0.72	0.90	0.79	0.74
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.84	0.54	0.66	0.88	0.75	0.69
2	CfsSubSetEval	GreedyStepwise	735	335	37	32	**SAHBN**	**NA**	**0.86**	**0.62**	**0.72**	**0.89**	**0.79**	**0.75**
							Hill-Climbing (hc)	loglik	0.87	0.33	0.48	0.84	0.66	0.50
								bde	0.93	0.64	0.76	0.91	0.81	0.78
								mbde	0.93	0.64	0.76	0.91	0.81	0.78
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.89	0.62	0.73	0.90	0.80	0.76
							Tabu Search (tabu	loglik	0.77	0.44	0.56	0.85	0.70	0.60
								bde	0.92	0.59	0.72	0.90	0.79	0.74
								mbde	0.92	0.56	0.72	0.90	0.79	0.74

								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.86	0.62	0.72	0.89	0.75	0.75
3	ConsistencySubsetEval	BestFirst	735	335	20	18	**SAHBN**	**NA**	**0.75**	**0.38**	**0.51**	**0.84**	**0.67**	**0.55**
							Hill-Climbing (hc)	loglik	0.81	0.56	0.67	0.88	0.76	0.71
								bde	0.92	0.59	0.72	0.90	0.79	0.74
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.91	0.26	0.40	0.83	0.62	0.41
								bic	0.91	0.26	0.40	0.83	0.62	0.41
								K2	0.83	0.51	0.63	0.87	0.74	0.67
							Tabu Search (tabu	loglik	0.84	0.54	0.66	0.88	0.75	0.69
								bde	0.92	0.62	0.74	0.90	0.80	0.76
								mbde	0.92	0.59	0.72	0.89	0.79	0.74
								aic	0.91	0.26	0.40	0.83	0.62	0.41
								bic	0.91	0.26	0.40	0.83	0.62	0.41
								K2	0.86	0.62	0.72	0.89	0.79	0.75
4	ConsistencySubsetEval	GreedyStepwise	735	335	20	17	**SAHBN**	**NA**	**0.60**	**0.23**	**0.33**	**0.80**	**0.59**	**0.37**
							Hill-Climbing (hc)	loglik	0.61	0.59	0.60	0.82	0.74	0.71
								bde	0.61	0.28	0.39	0.80	0.62	0.43
								mbde	0.61	0.28	0.39	0.80	0.62	0.44
								aic	0.89	0.21	0.33	0.82	0.60	0.34
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.62	0.41	0.49	0.81	0.67	0.57
							Tabu Search (tabu	loglik	0.55	0.44	0.49	0.80	0.67	0.59
								bde	0.59	0.26	0.36	0.80	0.60	0.40
								mbde	0.63	0.26	0.36	0.80	0.61	0.40
								aic	0.00	0.00	0.00	0.78	050	0.00
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.62	0.46	0.53	0.82	0.61	0.61

5	FilteredSubsetEval	BestFirst	735	335	35	30	**SAHBN**	**NA**	**0.92**	**0.56**	**0.70**	**0.89**	**0.77**	**0.72**
							Hill-Climbing (hc)	loglik	0.80	0.41	0.54	0.85	0.69	0.58
								bde	0.92	0.62	0.74	0.90	0.80	0.76
								mbde	0.92	0.62	0.74	0.90	0.80	0.76
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.79	0.59	0.68	0.88	0.77	0.73
							Tabu Search (tabu	loglik	0.79	0.38	0.52	0.84	0.68	0.55
								bde	0.92	0.62	0.74	0.90	0.80	0.76
								mbde	0.92	0.59	0.72	0.90	0.79	0.74
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.81	0.54	0.65	0.87	0.75	0.69
6	FilteredSubsetEval	GreedyStepwise	735	335	37	32	**SAHBN**	**NA**	**0.86**	**0.62**	**0.72**	**0.89**	**0.79**	**0.75**
							Hill-Climbing (hc)	loglik	0.78	0.54	0.64	0.86	0.75	0.69
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.84	0.54	0.66	0.88	0.88	0.75
							Tabu Search (tabu	loglik	0.77	0.44	0.56	0.85	0.70	0.60
								bde	0.92	0.62	0.74	0.90	0.80	0.76
								mbde	0.92	0.59	0.72	0.90	0.79	0.74
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.85	0.59	0.70	0.89	0.78	0.73
7	InfoGainAttributeEval	Ranker	735	335	37	28	**SAHBN**	**NA**	**0.77**	**0.59**	**0.67**	**0.87**	**0.77**	**0.73**
							Hill-Climbing	loglik	0.83	0.51	0.63	0.87	0.74	0.67
								bde	0.79	0.49	0.60	0.86	0.73	0.65

							(hc)	mbde	0.79	0.49	0.60	0.86	0.73	0.65
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.92	0.28	0.43	0.84	0.64	0.44
								K2	0.80	0.51	0.63	0.86	0.74	0.67
							Tabu Search (tabu)	loglik	0.85	0.44	0.58	0.86	0.71	0.60
								bde	0.79	0.49	0.60	0.86	0.73	0.65
								mbde	0.79	0.49	0.60	0.86	0.73	0.65
								aic	0.87	0.33	0.48	0.84	0.66	0.50
								bic	0.92	0.28	0.43	0.84	0.64	0.44
								K2	0.84	0.54	0.66	0.88	0.75	0.69
8	GainRatioAttributeEval	Ranker	735	335	37	32	**SAHBN**	**NA**	**1.00**	**0.33**	**0.50**	**0.85**	**0.67**	**0.50**
							Hill-Climbing (hc)	loglik	1.00	0.54	0.70	0.90	0.77	0.70
								bde	1.00	0.49	0.66	0.89	0.74	0.66
								mbde	1.00	0.51	0.68	0.89	0.76	0.68
								aic	1.00	0.15	0.27	0.81	0.58	0.27
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	1.00	0.51	0.68	0.89	0.76	0.68
							Tabu Search (tabu)	loglik	1.00	0.54	0.70	0.90	0.77	0.70
								bde	1.00	0.46	0.63	0.88	0.73	0.63
								mbde	1.00	0.51	0.68	0.89	0.76	0.68
								aic	1.00	0.15	0.27	0.81	0.58	0.27
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	1.00	0.51	0.68	0.89	0.76	0.68
9	CorrelationAttributeEval	Ranker	735	335	37	27	**SAHBN**	**NA**	**0.77**	**0.59**	**0.67**	**0.87**	**0.77**	**0.73**
							Hill-Climbing (hc)	loglik	0.85	0.59	0.70	0.89	0.78	0.73
								bde	0.79	0.49	0.60	0.86	0.73	0.65
								mbde	0.79	0.49	0.60	0.86	0.73	0.65
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.92	0.28	0.43	0.84	0.64	0.44

								K2	0.86	0.62	0.72	0.89	0.79	0.75
							Tabu Search (tabu)	loglik	0.87	0.51	0.65	0.88	0.75	0.67
								bde	0.80	0.51	0.63	0.86	0.74	0.67
								mbde	0.79	0.49	0.60	0.86	0.73	0.65
								aic	0.87	0.33	0.48	0.84	0.66	0.50
								bic	0.92	0.28	0.43	0.84	0.64	0.44
								K2	0.87	0.67	0.75	0.90	0.82	0.79
10	ReliefFAttAttributeEval	Ranker	735	335	37	25	**SAHBN**	**NA**	**0.92**	**0.31**	**0.46**	**0.84**	**0.65**	**0.50**
							Hill-Climbing (hc)	loglik	0.65	0.51	0.57	0.83	0.72	0.66
								bde	0.69	0.51	0.59	0.84	0.72	0.66
								mbde	0.73	0.49	0.58	0.85	0.72	0.64
								aic	0.88	0.38	0.54	0.85	0.69	0.55
								bic	1.00	0.15	0.27	081	0.58	0.27
								K2	0.72	0.54	0.62	0.85	0.74	0.69
							Tabu Search (tabu)	loglik	0.73	0.62	0.67	0.86	0.78	0.74
								bde	0.69	0.46	0.55	0.84	0.70	0.62
								mbde	0.69	0.56	0.62	0.85	0.75	0.70
								aic	0.88	0.38	0.54	0.85	0.69	0.55
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	0.73	0.49	0.58	0.85	0.72	0.64
11	OneRAttributeEval	Ranker	735	335	37	28	**SAHBN**	**NA**	**0.83**	**0.49**	**0.61**	**0.86**	**0.73**	**0.65**
							Hill-Climbing (hc)	loglik	0.62	0.46	0.53	0.82	0.69	0.61
								bde	0.76	0.49	0.59	0.85	0.72	0.65
								mbde	0.68	0.49	0.57	0.84	0.71	0.64
								aic	0.88	0.38	0.54	0.85	0.69	0.55
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	0.77	0.51	0.62	0.86	0.73	0.67
							Tabu Search	loglik	0.68	0.54	0.60	0.84	0.73	0.68
								bde	0.75	0.54	0.63	0.86	0.74	0.69

							(tabu)	mbde	0.73	0.56	0.64	0.86	0.75	0.71
								aic	0.88	0.38	0.54	0.85	0.69	0.55
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	0.76	0.56	0.65	0.86	0.76	0.71

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.3 DNA repair gene-Gene2GO (CV=3.841) dataset experimental results

Table 6-3 DNA repair gene-Gene2GO (CV=3.841) dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Chi-Squared Test Reduction	Attr. Selection Reduction.	Sub-Super Class Test Reduction	Algorithm	Score	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	7 35	89	24	20	**SAHBN**	**NA**	**0.86**	**0.62**	**0.72**	**0.89**	**0.79**	**0.75**
							Hill-Climbing (hc)	loglik	0.78	0.46	0.58	0.85	0.71	0.62
								bde	0.91	0.54	0.68	0.89	0.76	0.70
								mbde	0.92	0.59	0.72	0.90	0.79	0.74
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.74	0.36	0.48	0.83	0.66	0.52
							Tabu Search (tabu)	loglik	0.84	0.41	0.55	0.85	0.69	0.58
								bde	0.91	0.51	0.66	0.88	0.75	0.67
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.92	0.59	0.72	0.90	0.79	0.74
2	CfsSubSetEval	GreedyStepwise	735	89	30	24	**SAHBN**	**NA**	**0.95**	**0.46**	**0.62**	**0.88**	**0.73**	**0.63**
							Hill-Climbing (hc)	loglik	0.71	0.38	0.50	0.83	0.67	0.55
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.75	0.38	0.51	0.84	0.67	0.55
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.87	0.51	0.65	0.88	0.75	0.67
							Tabu Search (tabu	loglik	0.79	0.56	0.66	0.87	0.76	0.71
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.92	0.56	0.70	0.89	0.77	0.72

								aic	0.66	0.64	0.65	0.85	0.77	0.75
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.92	0.59	0.72	0.90	0.79	0.74
3	ConsistencySubsetEval	BestFirst	735	89	15	13	**SAHBN**	**NA**	**0.88**	**0.38**	**0.54**	**0.85**	**0.69**	**0.55**
							Hill-Climbing (hc)	loglik	0.72	0.54	0.62	0.85	0.74	0.69
								bde	0.90	0.49	0.63	0.88	0.74	0.65
								mbde	0.91	0.51	0.66	0.88	0.75	0.67
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.88	0.54	0.67	0.88	0.76	0.69
							Tabu Search (tabu	loglik	0.71	0.51	0.60	0.85	0.73	0.66
								bde	0.92	0.31	0.46	0.84	0.65	0.47
								mbde	0.92	0.31	0.46	0.84	0.65	0.47
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.83	0.51	0.63	0.87	0.74	0.67
4	ConsistencySubsetEval	GreedyStepwise	735	89	15	13	**SAHBN**	**NA**	**0.86**	**0.46**	**0.60**	**0.86**	**0.72**	**0.63**
							Hill-Climbing (hc)	loglik	0.74	0.51	0.61	0.85	0.73	0.67
								bde	0.68	0.49	0.57	0.84	0.71	0.64
								mbde	0.70	0.54	0.61	0.85	0.74	0.68
								aic	0.68	0.49	0.57	0.84	0.71	0.64
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.87	0.51	0.65	0.88	0.75	0.67
							Tabu Search (tabu	loglik	0.74	0.51	0.61	0.85	0.73	0.67
								bde	0.70	0.54	0.61	0.85	0.74	0.68
								mbde	0.68	0.49	0.57	0.84	0.71	0.64
								aic	0.68	0.49	0.57	0.84	0.71	0.64
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.84	0.41	0.55	0.85	0.69	0.58

5	FilteredSubsetEval	BestFirst	735	89	24	20	**SAHBN**	**NA**	**0.86**	**0.62**	**0.72**	**0.89**	**0.79**	**0.75**
							Hill-Climbing (hc)	loglik	0.76	0.41	0.53	0.84	0.69	0.58
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.94	0.41	0.57	0.86	0.70	0.58
							Tabu Search (tabu	loglik	0.81	0.44	0.57	0.85	0.70	0.60
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.95	0.54	0.69	0.89	0.77	0.70
6	FilteredSubsetEval	GreedyStepwise	735	89	30	24	**SAHBN**	**NA**	**0.90**	**0.49**	**0.63**	**0.88**	**0.74**	**0.65**
							Hill-Climbing (hc)	loglik	0.74	0.44	0.55	0.84	0.70	0.60
								bde	0.92	0.59	0.72	0.90	0.79	0.74
								mbde	0.92	0.62	0.74	0.90	0.80	0.76
								aic	0.78	0.36	0.49	0.84	0.66	0.52
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.83	0.51	0.63	0.87	0.74	0.67
							Tabu Search (tabu	loglik	0.80	0.51	0.63	0.86	0.74	0.67
								bde	0.92	0.59	0.72	0.90	0.79	0.74
								mbde	0.91	0.54	0.68	0.89	0.76	0.70
								aic	0.63	0.56	0.59	0.83	0.73	0.70
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.88	0.54	0.67	0.88	0.76	0.69
7	InfoGainAttributeEval	Ranker	735	89	30	22	**SAHBN**	**NA**	**0.88**	**0.54**	**0.67**	**0.88**	**0.76**	**0.69**
							Hill-Climbing	loglik	0.88	0.56	0.69	0.89	0.77	0.72
								bde	0.91	0.54	0.68	0.89	0.76	0.70

							(hc)	mbde	0.91	0.51	0.56	0.88	0.75	0.67
								aic	0.76	0.41	0.53	0.84	0.69	0.58
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.76	0.41	0.53	0.84	0.69	0.58
							Tabu Search (tabu)	loglik	0.90	0.49	0.63	0.88	0.74	0.65
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.91	0.51	0.66	0.88	0.75	0.67
								aic	0.77	0.44	0.56	0.85	0.70	0.60
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.86	0.46	0.60	0.86	0.72	0.63
8	GainRatioAttributeEval	Ranker	735	89	30	28	**SAHBN**	**NA**	**1.00**	**0.51**	**0.68**	**0.89**	**0.76**	**0.68**
							Hill-Climbing (hc)	loglik	1.00	0.44	0.61	0.88	0.72	0.70
								bde	1.00	0.46	0.63	0.88	0.73	0.63
								mbde	1.00	0.46	0.63	0.88	0.73	0.63
								aic	1.00	0.15	0.27	0.81	0.58	0.27
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	1.00	0.44	0.61	0.88	0.72	0.61
							Tabu Search (tabu)	loglik	1.00	0.46	0.63	0.88	0.73	0.63
								bde	1.00	0.51	0.68	0.89	0.76	0.68
								mbde	1.00	0.46	0.63	0.88	0.73	0.63
								aic	1.00	0.15	0.27	0.81	0.58	0.27
								bic	1.00	0.15	0.27	0.81	0.58	0.27
								K2	1.00	0.46	0.63	0.88	0.73	0.63
9	CorrelationAttributeEval	Ranker	735	89	30	23	**SAHBN**	**NA**	**0.87**	**0.51**	**0.65**	**0.88**	**0.75**	**0.67**
							Hill-Climbing (hc)	loglik	0.70	0.49	0.58	0.84	0.71	0.64
								bde	0.91	0.51	0.66	0.88	0.75	0.67
								mbde	0.91	0.51	0.66	0.88	0.75	0.67
								aic	0.78	0.36	0.49	0.84	0.66	0.52
								bic	0.89	0.21	0.33	0.82	0.60	0.34

								K2	0.81	0.44	0.57	0.85	0.70	0.60
							Tabu Search (tabu)	loglik	0.77	0.51	0.62	0.86	0.73	0.67
								bde	0.91	0.51	0.66	0.88	0.75	0.67
								mbde	0.92	0.56	0.70	0.89	0.77	0.72
								aic	0.80	0.41	0.54	0.85	0.69	0.58
								bic	0.89	0.21	0.33	0.82	0.60	0.34
								K2	0.91	0.54	0.68	0.89	0.76	0.70
10	ReliefFAttAttributeEval	Ranker	735	89	30	25	**SAHBN**	**NA**	**0.85**	**0.44**	**0.58**	**0.56**	**0.71**	**0.60**
							Hill-Climbing (hc)	loglik	0.67	0.56	0.61	0.84	0.74	0.70
								bde	0.92	0.56	0.70	0.89	0.77	0.72
								mbde	0.95	0.54	0.69	0.89	0.77	0.70
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.58	0.28	0.38	0.80	0.61	0.43
								K2	0.72	0.54	0.62	0.85	0.74	0.69
							Tabu Search (tabu)	loglik	0.55	0.54	0.55	0.80	0.71	0.67
								bde	0.95	0.54	0.69	0.89	0.77	0.70
								mbde	0.95	0.54	0.69	0.89	0.77	0.70
								aic	0.93	0.36	0.52	0.85	0.68	0.53
								bic	0.58	0.28	0.38	0.80	0.61	0.43
								K2	0.70	0.54	0.61	0.85	0.74	0.68
11	OneRAttributeEval	Ranker	735	89	30	26	**SAHBN**	**NA**	**0.83**	**0.49**	**0.61**	**0.86**	**0.73**	**0.65**
							Hill-Climbing (hc)	loglik	0.91	0.51	0.66	0.88	0.75	0.67
								bde	0.81	0.54	0.65	0.87	0.75	0.69
								mbde	0.79	0.49	0.60	0.86	0.73	0.65
								aic	0.75	0.38	0.51	0.84	0.67	0.55
								bic	0.92	0.28	0.43	0.84	0.64	0.44
								K2	0.78	0.54	0.64	0.86	0.75	0.69
							Tabu Search	loglik	0.91	0.51	0.66	0.88	0.75	0.67
								bde	0.83	0.51	0.63	0.87	0.74	0.67

							(tabu)	mbde	0.80	0.51	0.63	0.86	0.74	0.67
								aic	0.75	0.38	0.51	0.84	0.67	0.55
								bic	0.92	0.28	0.43	0.84	0.64	0.44
								K2	0.80	0.51	0.63	0.86	0.74	0.67

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.4 C. elegan-Gene2GO (CV = 2.706) dataset experimental results

Table 6-4 C. elegan–Gene2GO (CV = 2.706) dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Chi-Squared Test Reduction	Attr. Selection Reduction.	Sub-Super Class Test Reduction	Algorithm	Score	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	1012	65	12	11	**SAHBN**	**NA**	**0.80**	**0.46**	**0.58**	**0.67**	**0.67**	**0.60**
							Hill-Climbing (hc)	loglik	0.57	1.00	0.72	0.61	0.61	0.36
								bde	0.76	0.37	0.50	0.63	0.63	0.52
								mbde	0.76	0.37	0.50	0.63	0.63	0.52
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	1.00	0.19	0.32	0.59	0.60	0.32
							Tabu Search (tabu)	loglik	0.57	1.00	0.72	0.61	0.61	0.36
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.48	0.75	0.58	0.46	0.46	0.28
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	1.00	0.19	0.32	0.59	0.60	0.32
2	CfsSubSetEval	GreedyStepwise	1012	65	30	26	**SAHBN**	**NA**	**0.75**	**0.43**	**0.55**	**0.64**	**0.64**	**0.57**
							Hill-Climbing (hc)	loglik	0.57	0.99	0.73	0.63	0.62	0.40
								bde	0.78	0.41	0.54	0.65	0.65	0.56
								mbde	0.78	0.41	0.54	0.65	0.65	0.56
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.54	0.75	0.63	0.56	0.56	0.56
							Tabu Search (tabu	loglik	0.57	0.99	0.73	0.62	0.62	0.40
								bde	0.78	0.40	0.53	0.64	0.64	0.56
								mbde	0.78	0.41	0.54	0.64	0.65	0.56

								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.55	0.84	0.67	0.58	0.58	0.46
3	ConsistencySubsetEval	BestFirst	1012	65	30	27	**SAHBN**	**NA**	**0.73**	**0.52**	**0.60**	**0.66**	**0.66**	**0.63**
							Hill-Climbing (hc)							
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.56	0.95	0.70	0.59	0.59	0.37
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.81	0.51	0.63	0.69	0.69	0.65
							Tabu Search (tabu							
								bde	0.47	0.69	0.56	0.45	0.45	0.33
								mbde	0.48	0.61	0.54	0.47	0.46	0.42
								aic	0.56	0.95	0.70	0.59	0.59	0.37
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.80	0.52	0.63	0.69	0.69	0.65
4	ConsistencySubsetEval	GreedyStepwise	1012	65	33	31	**SAHBN**	**NA**	0.73	0.51	060	0.66	0.66	0.63
							Hill-Climbing (hc)							
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.56	0.95	0.70	0.59	0.59	0.37
								bic	0.54	1.00	0.70	.57	0.56	0.23
								K2	0.81	0.49	0.61	0.68	0.69	0.63
							Tabu Search (tabu							
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.56	0.95	0.70	0.59	0.59	0.37
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.80	0.49	0.61	0.68	0.68	0.63

5	FilteredSubsetEval	BestFirst	1012	65	2	2	**SAHBN**	**NA**	**0.78**	**0.37**	**0.51**	**0.63**	**0.63**	**0.53**
							Hill-Climbing (hc)	loglik	0.54	1.00	0.70	0.57	0.56	0.23
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.54	1.00	0.70	0.57	0.56	0.23
							Tabu Search (tabu	loglik	0.54	1.00	0.70	0.57	0.56	0.23
								bde	0.48	0.71	0.58	0.47	0.47	0.34
								mbde	0.47	0.65	0.54	0.45	0.45	0.36
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.54	1.00	0.70	0.57	0.56	0.23
6	FilteredSubsetEval	GreedyStepwise	1012	65	30	26	**SAHBN**	**NA**	**0.75**	**0.43**	**0.55**	**0.64**	**0.64**	**0.57**
							Hill-Climbing (hc)	loglik	0.57	0.99	0.73	0.63	0.62	0.40
								bde	0.78	0.41	0.54	0.65	0.65	0.56
								mbde	0.79	0.41	0.54	0.65	0.65	0.56
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.55	0.85	0.67	0.58	0.58	0.45
							Tabu Search (tabu	loglik	0.57	0.99	0.73	0.62	0.62	0.39
								bde	0.78	0.40	0.53	0.64	0.64	0.56
								mbde	0.78	0.41	0.54	0.64	0.65	0.56
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.56	0.89	0.69	0.59	0.59	0.42
7	InfoGainAttributeEval	Ranker	1012	65	30	26	**SAHBN**	**NA**	**0.72**	**0.49**	**0.58**	**0.65**	**0.65**	**0.61**
							Hill-Climbing							
								bde	0.76	0.37	0.50	0.62	0.63	0.52

							(hc)	mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.57	1.00	0.73	0.62	0.62	0.39
							Tabu Search (tabu)							
								bde	0.47	0.67	0.56	0.46	0.46	0.36
								mbde	0.47	0.65	0.54	0.45	0.45	0.36
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.57	1.00	0.73	0.62	0.62	0.39
8	GainRatioAttributeEval	Ranker	1012	65	30	26	**SAHBN**	**NA**	**0.76**	**0.14**	**0.23**	**0.54**	**0.54**	**0.24**
							Hill-Climbing (hc)	loglik	0.58	1.00	0.74	0.64	0.64	0.43
								bde	0.58	1.00	0.73	0.63	0.63	0.41
								mbde	0.58	1.00	0.73	0.63	0.63	0.41
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.58	1.00	0.73	0.63	0.63	0.41
							Tabu Search (tabu)	loglik	0.58	1.00	0.74	0.64	0.64	0.43
								bde	0.58	1.00	0.73	0.63	0.63	0.41
								mbde	0.58	1.00	0.73	0.63	0.63	0.41
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.58	1.00	0.73	0.63	0.63	0.41
9	CorrelationAttributeEval	Ranker	1012	65	30	27	**SAHBN**	**NA**	**0.70**	**0.51**	**0.59**	**0.64**	**0.65**	**0.62**
							Hill-Climbing (hc)							
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.51	1.00	0.67	0.51	0.51	0.04
								bic	0.54	1.00	0.70	0.57	0.56	0.23

								K2	0.57	1.00	0.73	0.62	0.62	0.39
							Tabu Search (tabu)							
								bde	0.48	0.71	0.57	0.47	0.46	0.34
								mbde	0.48	0.77	0.59	0.47	0.47	0.27
								aic	0.50	0.88	0.64	0.49	0.49	0.18
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.57	1.00	0.73	0.62	0.62	0.39
10	ReliefFAttAttributeEval	Ranker	1012	65	30	23	**SAHBN**	**NA**	**0.74**	**0.45**	**0.56**	**0.64**	**0.65**	**0.59**
							Hill-Climbing (hc)	loglik	0.59	0.98	0.74	0.64	0.64	0.46
								bde	0.76	0.40	0.52	0.63	0.64	0.55
								mbde	0.76	040	0.52	0.63	0.64	0.55
								aic	0.51	1.00	0.67	0.51	0.51	0.04
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.58	0.99	0.74	0.64	0.64	0.44
							Tabu Search (tabu)	loglik	0.59	0.98	0.73	0.64	0.64	0.45
								bde	0.76	0.40	0.52	0.63	0.64	0.55
								mbde	0.76	0.40	0.52	0.63	0.64	0.55
								aic	0.48	0.76	0.59	0.47	0.47	0.28
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.58	0.99	0.73	0.64	0.64	0.43
11	OneRAttributeEval	Ranker	1012	65	30	25	**SAHBN**	**NA**	**0.69**	**0.50**	**0.58**	**0.63**	**0.63**	**0.60**
							Hill-Climbing (hc)	loglik	0.60	0.94	0.73	0.65	0.65	0.52
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.50	0.88	0.64	0.49	0.49	0.17
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.59	0.99	0.74	0.65	0.64	0.45
							Tabu Search	loglik	0.60	0.94	0.73	0.65	0.65	0.51
								bde	0.49	0.71	0.58	0.48	0.48	0.37

							(tabu)	mbde	0.49	0.69	0.57	0.48	0.48	0.38
								aic	0.48	0.75	0.59	0.46	0.46	0.28
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.59	0.99	0.74	0.64	0.64	0.45

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.5 C. elegans-Gene2GO (CV=3.841) dataset experimental results

Table 6-5 C. elegans-Gene2GO (CV = 3.841) dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Chi-Squared Test Reduction	Attr. Selection Reduction	Sub-Super Class Test Reduction	Algorithm	Score	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	1012	36	12	11	**SAHBN**	**NA**	**0.81**	**0.46**	**0.59**	**0.67**	**0.67**	**0.61**
							Hill-Climbing (hc)	loglik	0.55	1.00	0.71	0.59	0.59	0.31
								bde	0.47	0.21	0.29	0.48	0.48	0.32
								mbde	0.53	0.13	0.21	0.50	0.51	0.23
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.54	0.89	0.67	0.56	0.56	0.37
							Tabu Search (tabu)	loglik	0.55	1.00	0.71	0.59	0.59	0.31
								bde	0.47	0.56	0.51	0.45	0.45	0.43
								mbde	0.48	0.48	0.48	0.47	0.47	0.47
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.54	0.90	0.68	0.57	0.57	0.37
2	CfsSubSetEval	GreedyStepwise	1012	36	15	14	**SAHBN**	**NA**	**0.80**	**0.48**	**0.60**	**0.68**	**0.68**	**0.62**
							Hill-Climbing (hc)	loglik	0.55	1.00	0.71	0.59	0.59	0.31
								bde	0.51	0.88	0.64	0.51	0.51	0.23
								mbde	0.50	0.87	0.64	0.50	0.50	0.21
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.69	0.36	0.48	0.60	0.60	0.50
							Tabu Search (tabu	loglik	0.55	1.00	0.71	0.59	0.59	0.31
								bde	0.52	1.00	0.69	0.54	0.53	0.13
								mbde	0.52	0.95	0.67	0.53	0.53	0.19

								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.50	0.83	0.62	0.49	0.49	0.25
								K2	0.73	0.33	0.45	0.60	0.60	0.47
3	ConsistencySubsetEval	BestFirst	1012	36	24	22	**SAHBN**	**NA**	**0.75**	**0.52**	**0.62**	**0.67**	**0.67**	**0.64**
							Hill-Climbing (hc)	loglik	0.55	1.00	0.71	0.59	0.59	0.31
								bde	0.77	0.44	0.56	0.65	0.66	0.56
								mbde	0.77	0.44	0.56	0.65	0.66	0.59
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.78	0.53	0.63	0.69	0.69	0.65
							Tabu Search (tabu	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.77	0.45	0.57	0.66	0.66	0.59
								mbde	0.77	0.45	0.57	0.66	0.66	0.59
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.76	0.37	0.50	0.62	0.63	0.52
								K2	0.79	0.54	0.64	0.69	0.70	0.66
4	ConsistencySubsetEval	GreedyStepwise	1012	36	24	22	**SAHBN**	**NA**	**0.76**	**0.52**	**0.62**	**0.68**	**0.68**	**0.64**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.77	0.45	0.57	0.66	0.66	0.59
								mbde	0.77	0.45	0.57	0.66	0.66	0.59
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.78	0.54	0.64	0.69	0.69	0.66
							Tabu Search (tabu	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.77	0.44	0.56	0.65	0.66	0.59
								mbde	0.77	0.45	0.57	0.66	0.66	0.59
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.76	0.37	0.50	0.62	0.63	0.52
								K2	0.80	0.54	0.64	0.70	0.70	0.66

5	FilteredSubsetEval	BestFirst	1012	36	2	2	**SAHBN**	**NA**	**0.78**	**0.37**	**0.51**	**0.63**	**0.63**	**0.53**
							Hill-Climbing (hc)	loglik	0.54	1.00	0.70	0.57	0.57	0.23
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.54	1.00	0.70	0.57	0.57	0.23
							Tabu Search (tabu	loglik	0.54	1.00	0.70	0.57	0.57	0.23
								bde	0.46	0.64	0.54	0.44	0.44	0.35
								mbde	0.43	0.47	0.45	0.42	0.42	0.41
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.54	1.00	0.70	0.57	0.57	0.23
6	FilteredSubsetEval	GreedyStepwise	1012	36	15	14	**SAHBN**	**NA**	**0.80**	**0.48**	**0.60**	**0.68**	**0.68**	**0.62**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.50	0.83	0.62	0.50	0.50	0.27
								mbde	0.52	1.00	0.69	0.54	0.53	0.13
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.91	0.27	0.42	0.62	0.62	0.42
							Tabu Search (tabu	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.52	1.00	0.68	0.54	0.53	0.12
								mbde	0.51	0.88	0.64	0.51	0.51	0.23
								aic	0.54	1.00	0.70	0.57	0.56	0.23
								bic	0.47	0.60	0.53	0.46	0.46	0.41
								K2	0.67	0.35	0.46	0.58	0.59	0.49
7	InfoGainAttributeEval	Ranker	1012	36	24	22	**SAHBN**	**NA**	**0.75**	**0.49**	**0.59**	**0.66**	**0.66**	**0.62**
							Hill-Climbing	loglik	0.54	0.79	0.64	0.56	0.56	0.46
								bde	0.76	0.37	0.50	0.62	0.63	0.52

							(hc)	mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.84	0.13	0.22	0.55	0.55	0.22
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.65	0.39	0.49	0.59	0.59	0.52
							Tabu Search (tabu)	loglik	0.54	0.94	0.69	0.57	0.57	0.32
								bde	0.54	1.00	0.70	0.57	0.57	0.23
								mbde	0.54	1.00	0.70	0.57	0.57	0.23
								aic	0.46	0.25	0.32	0.47	0.47	0.37
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.72	0.31	0.44	0.59	0.60	0.46
8	GainRatioAttributeEval	Ranker	1012	36	24	23	**SAHBN**	**NA**	**1.00**	**0.15**	**0.26**	**0.57**	**0.58**	**0.26**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.71	0.60	0.59	0.32
								bde	0.65	1.00	0.72	0.60	0.60	0.34
								mbde	0.56	1.00	0.72	0.60	0.60	0.34
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.55	0.88	0.68	0.58	0.58	0.42
							Tabu Search (tabu)	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.56	1.00	0.2	0.60	0.60	0.33
								mbde	0.56	1.00	0.72	0.60	0.60	0.34
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.53	0.69	0.60	0.54	0.54	0.49
9	CorrelationAttributeEval	Ranker	1012	36	24	23	**SAHBN**	**NA**	**0.74**	**0.49**	**0.59**	**0.65**	**0.66**	**0.61**
							Hill-Climbing (hc)	loglik	0.53	0.73	0.62	0.54	0.54	0.46
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.84	0.13	0.22	0.55	0.55	0.22
								bic	0.54	1.00	0.70	0.57	0.57	0.23

								K2	0.54	0.81	0.65	0.56	0.56	0.44
							Tabu Search (tabu)	loglik	0.54	0.77	0.63	0.55	0.55	0.47
								bde	0.42	0.46	0.44	0.41	0.41	0.41
								mbde	0.43	0.55	0.49	0.41	0.41	0.36
								aic	0.42	0.35	0.38	0.43	0.43	0.41
								bic	0.54	1.00	0.70	0.57	0.57	0.23
								K2	0.52	0.64	0.58	0.52	0.52	0.49
10	ReliefFAttAttributeEval	Ranker	1012	36	24	22	**SAHBN**	**NA**	**0.77**	**0.48**	**0.59**	**0.67**	**0.67**	**0.61**
							Hill-Climbing (hc)	loglik	0.55	1.00	0.71	0.59	0.59	0.31
								bde	0.78	0.42	0.55	0.65	0.65	0.57
								mbde	0.79	0.42	0.55	0.65	0.65	0.57
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.56	0.23
								K2	0.64	0.37	0.47	0.58	0.58	0.51
							Tabu Search (tabu)	loglik	0.56	1.00	0.71	0.60	0.59	0.31
								bde	0.78	0.42	0.55	0.65	0.65	0.57
								mbde	0.78	0.42	0.55	0.65	0.65	0.57
								aic	0.54	1.00	0.70	0.57	0.57	0.23
								bic	0.54	1.00	0.70	0.57	0.57	023
								K2	0.53	0.53	0.53	0.53	0.53	0.51
11	OneRAttributeEval	Ranker	1012	36	24	23	**SAHBN**	**NA**	**0.74**	**0.48**	**0.59**	**0.65**	**0.66**	**0.61**
							Hill-Climbing (hc)	loglik	0.54	0.51	0.53	0.53	0.53	0.53
								bde	0.76	0.37	0.50	0.62	0.63	0.52
								mbde	0.76	0.37	0.50	0.62	0.63	0.52
								aic	0.75	0.46	0.57	0.65	0.65	0.59
								bic	0.83	0.99	0.18	0.54	0.54	0.18
								K2	0.52	0.59	0.55	0.52	0.52	0.50
							Tabu Search	loglik	0.54	0.48	0.51	0.53	0.53	0.52
								bde	0.49	0.66	0.56	0.48	0.48	0.41

							(tabu)	mbde	0.46	0.60	0.52	0.45	0.45	0.39
								aic	0.75	0.46	0.57	0.65	0.65	.59
								bic	0.63	0.13	0.21	0.52	0.53	0.22
								K2	0.55	0.94	0.70	0.59	0.58	0.36

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.6 D.melanogaster-Gene2GO (CV=2.706) dataset experimental results

Table 6-6 D.melanogaster-Gene2GO (CV=2.706) dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Chi-Squared Test Reduction	Attr. Selection Reduction	Sub-Super Class Test Reduction	Algorithm	Score	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	860	44	27	26	**SAHBN**	**NA**	**0.81**	**0.52**	**0.63**	**0.70**	**0.70**	**0.65**
							Hill-Climbing (hc)	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	1.00	0.40	0.57	0.70	0.70	0.57
								mbde	1.00	0.40	0.57	0.70	0.70	0.57
								aic	0.38	0.21	0.27	0.43	0.43	0.31
								bic	0.46	0.19	0.27	0.48	0.48	0.30
								K2	0.97	0.53	0.69	0.76	0.76	0.69
							Tabu Search (tabu)	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	1.00	0.40	0.57	0.70	0.70	0.57
								mbde	1.00	0.40	0.57	0.70	0.70	0.57
								aic	1.00	0.10	0.19	0.55	0.55	0.19
								bic	0.52	1.00	0.69	0.54	0.54	0.13
								K2	0.97	0.52	0.67	0.75	0.75	0.68
2	CfsSubSetEval	GreedyStepwise	860	44	26	25	**SAHBN**	**NA**	**0.79**	**0.47**	**0.59**	**0.67**	**0.67**	**0.61**
							Hill-Climbing (hc)	loglik	1.00	0.34	0.51	0.67	0.67	0.51
								bde	0.83	0.43	0.57	0.67	0.67	0.59
								mbde	0.86	0.41	0.56	0.67	0.67	0.57
								aic	1.00	0.10	0.19	0.55	0.55	0.19
								bic	1.00	0.10	0.19	0.55	0.55	0.19
								K2	0.97	0.52	0.67	0.75	0.75	0.68
							Tabu Search (tabu	loglik	1.00	0.33	0.49	0.66	0.66	0.49
								bde	0.86	0.43	0.57	0.67	0.68	0.59
								mbde	0.86	0.43	0.57	0.68	0.68	0.59

								aic	1.00	0.10	0.19	0.55	0.55	0.19
								bic	0.52	1.00	0.69	0.54	0.54	0.13
								K2	0.97	0.50	0.66	0.74	0.74	0.66
3	ConsistencySubsetEval	BestFirst	860	44	14	14	**SAHBN**	**NA**	**0.92**	**0.57**	**0.70**	**0.76**	**0.76**	**0.71**
							Hill-Climbing (hc)	loglik	0.96	0.41	0.58	0.70	0.70	0.58
								bde	0.85	0.50	0.63	0.70	0.71	0.65
								mbde	0.85	0.50	0.63	0.70	0.71	0.65
								aic	0.40	0.17	0.24	0.45	0.45	0.28
								bic	0.41	0.12	0.19	0.47	0.47	0.21
								K2	0.97	0.50	0.66	0.74	0.74	0.66
							Tabu Search (tabu	loglik	0.96	0.43	0.60	0.70	0.71	0.60
								bde	0.85	0.50	0.63	0.70	0.71	0.65
								mbde	0.85	0.50	0.63	0.70	0.71	0.65
								aic	0.56	0.17	0.26	0.51	0.52	0.29
								bic	0.47	0.14	0.21	0.49	0.49	0.24
								K2	0.97	0.52	0.67	0.75	0.75	0.68
4	ConsistencySubsetEval	GreedyStepwise	860	44	16	16	**SAHBN**	**NA**	**0.96**	**0.47**	**0.63**	**0.72**	**0.72**	**0.63**
							Hill-Climbing (hc)	loglik	0.93	0.43	0.59	0.70	0.70	0.60
								bde	0.96	0.45	0.61	0.71	0.72	0.62
								mbde	1.00	0.41	0.59	0.70	0.71	0.59
								aic	0.75	0.16	0.26	0.55	0.55	0.27
								bic	0.56	0.17	0.26	0.51	0.52	0.29
								K2	0.97	0.52	0.67	0.75	0.75	0.68
							Tabu Search (tabu	loglik	0.93	0.45	0.60	0.70	0.71	0.61
								bde	1.00	0.41	0.59	0.70	0.71	0.59
								mbde	1.00	0.43	0.60	0.71	0.72	0.60
								aic	0.50	0.16	0.24	0.50	0.50	0.26
								bic	0.45	0.17	0.25	0.48	0.48	0.28
								K2	0.93	0.48	0.64	0.72	0.72	0.64

5	FilteredSubsetEval	BestFirst	860	44	27	26	**SAHBN**	**NA**	**0.83**	**0.50**	**0.62**	**0.70**	**0.70**	**0.64**
							Hill-Climbing (hc)	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	1.00	0.40	0.57	0.70	0.70	0.57
								mbde	1.00	0.41	0.59	0.70	0.71	0.59
								aic	1.00	0.14	0.24	0.57	0.57	0.24
								bic	1.00	0.10	0.29	0.55	0.55	0.19
								K2	0.97	0.48	0.64	0.73	0.73	0.65
							Tabu Search (tabu	loglik	1.00	0.29	0.45	0.64	0.65	0.45
								bde	1.00	0.41	0.59	0.70	0.71	0.59
								mbde	1.00	0.40	0.57	0.70	0.70	0.57
								aic	1.00	0.10	0.19	0.55	0.55	0.19
								bic	0.52	1.00	0.69	0.54	0.54	0.13
								K2	0.97	0.50	0.66	0.74	0.74	0.66
6	FilteredSubsetEval	GreedyStepwise	860	44	26	25	**SAHBN**	**NA**	**0.79**	**0.47**	**0.59**	**0.67**	**0.67**	**0.61**
							Hill-Climbing (hc)	loglik	1.00	0.33	0.49	0.66	0.66	0.49
								bde	0.89	0.43	0.58	0.69	0.69	0.59
								mbde	0.87	0.45	0.59	0.69	0.69	0.60
								aic	0.38	0.19	0.25	0.43	0.44	0.30
								bic	0.86	0.10	0.18	0.54	0.54	0.19
								K2	0.97	0.52	0.67	0.75	0.75	0.68
							Tabu Search (tabu	loglik	1.00	0.36	0.53	0.68	0.68	0.53
								bde	0.87	0.45	0.59	0.69	0.69	0.60
								mbde	0.86	0.43	0.57	0.68	0.68	0.59
								aic	0.43	0.16	0.23	0.47	0.47	0.26
								bic	0.49	0.79	0.61	0.48	0.48	0.26
								K2	0.97	0.48	0.64	0.73	0.73	0.65
7	InfoGainAttributeEval	Ranker	860	44	27	26	**SAHBN**	**NA**	**0.83**	**0.50**	**0.62**	**0.70**	**0.70**	**0.64**
							Hill-Climbing	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	0.60	1.00	0.75	0.67	0.67	0.50

							(hc)	mbde	0.60	1.00	0.75	0.67	0.67	0.50
								aic	0.26	0.17	0.21	0.34	0.34	0.26
								bic	0.44	0.14	0.21	0.48	0.48	0.24
								K2	0.96	0.45	0.61	0.71	0.72	0.62
							Tabu Search (tabu)	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	0.60	1.00	0.75	0.67	0.67	0.50
								mbde	0.60	1.00	0.75	0.67	0.67	0.50
								aic	0.43	0.17	0.25	0.47	0.47	0.28
								bic	0.42	0.14	0.21	0.47	0.47	0.24
								K2	0.96	0.43	0.60	0.70	0.71	0.60
8	GainRatioAttributeEval	Ranker	860	44	27	27	**SAHBN**	**NA**	**0.69**	**0.34**	**0.46**	**0.59**	**0.59**	**0.59**
							Hill-Climbing (hc)	loglik	1.00	0.38	0.55	0.69	0.69	0.55
								bde	0.54	1.00	0.70	0.57	0.56	0.22
								mbde	0.54	1.00	0.70	0.57	0.56	0.22
								aic	0.40	0.14	0.21	0.46	0.46	0.23
								bic	0.47	0.14	0.21	0.49	0.49	0.24
								K2	0.69	0.43	0.53	0.62	0.62	0.56
							Tabu Search (tabu)	loglik	1.00	0.40	0.57	0.70	0.70	0.57
								bde	0.54	1.00	0.70	0.57	0.56	0.22
								mbde	0.54	1.00	0.70	0.57	0.56	0.22
								aic	0.50	0.16	0.24	0.50	0.50	0.26
								bic	1.00	0.10	0.19	0.55	0.55	0.19
								K2	0.66	0.43	0.52	0.60	0.60	0.55
9	CorrelationAttributeEval	Ranker	860	44	27	26	**SAHBN**	**NA**	**0.87**	**0.57**	**0.69**	**0.74**	**0.74**	**0.70**
							Hill-Climbing (hc)	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	0.95	0.36	0.53	0.67	0.67	0.53
								mbde	0.95	0.36	0.53	0.67	0.67	0.53
								aic	0.50	0.16	0.24	0.50	0.50	0.26
								bic	0.64	0.12	0.20	0.52	0.53	0.21

								K2	0.96	0.47	0.63	0.72	0.72	0.63
							Tabu Search (tabu)	loglik	1.00	0.31	0.47	0.65	0.66	0.47
								bde	0.95	0.36	0.53	0.67	0.67	0.53
								mbde	0.95	0.36	0.53	0.67	0.67	0.53
								aic	0.44	0.12	0.19	0.48	0.48	0.21
								bic	0.59	0.17	0.27	0.52	0.52	0.29
								K2	0.97	0.48	0.64	0.73	0.73	0.65
10	ReliefFAttAttributeEval	Ranker	860	44	27	26	**SAHBN**	**NA**	**0.91**	**0.52**	**0.66**	**0.73**	**0.73**	**0.67**
							Hill-Climbing (hc)	loglik	0.91	0.36	0.52	0.66	0.66	0.53
								bde	1.00	0.34	0.51	0.67	0.67	0.51
								mbde	1.00	0.34	0.51	0.67	0.67	0.51
								aic	0.52	1.00	0.69	0.54	0.54	0.13
								bic	0.69	0.16	0.25	0.54	0.54	0.27
								K2	0.97	0.48	0.64	0.73	0.73	0.65
							Tabu Search (tabu)	loglik	0.95	0.36	0.53	0.67	0.67	0.53
								bde	1.00	0.34	0.51	0.67	0.67	0.51
								mbde	1.00	0.34	0.51	0.67	0.67	0.51
								aic	0.52	0.95	0.67	0.53	0.53	0.19
								bic	0.39	0.22	0.29	0.43	0.44	0.33
								K2	0.97	0.48	0.64	0.73	0.73	0.65
11	OneRAttributeEval	Ranker	860	44	27	25	**SAHBN**	**NA**	**0.83**	**0.59**	**0.69**	**0.73**	**0.73**	**0.70**
							Hill-Climbing (hc)	loglik	0.94	0.28	0.43	0.63	0.63	0.43
								bde	0.95	0.33	0.49	0.65	0.66	0.49
								mbde	0.95	0.33	0.49	0.65	0.66	0.49
								aic	1.00	0.17	0.29	0.58	0.59	0.29
								bic	0.69	0.16	0.25	0.54	0.54	0.27
								K2	0.88	0.40	0.55	0.67	0.67	0.56
							Tabu Search	loglik	1.00	0.33	0.49	0.66	0.66	0.49
								bde	0.95	0.33	0.49	0.65	0.66	0.49

							(tabu)	mbde	0.95	0.33	0.49	0.65	0.66	0.49
								aic	1.00	0.17	0.29	0.58	0.59	0.59
								bic	0.48	0.79	0.60	0.46	0.46	0.21
								K2	0.75	0.41	0.53	0.63	0.64	0.56

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.7 D.melanogaster-Gene2GO (CV=3.841) dataset experimental results

Table 6-7 D.melanogaster-Gene2GO (CV=3.841) dataset Experimental results

No.	Attributes Selection		Initial Attr. No.	Chi-Squared Test Reduction	Attr. Selection Reduction.	Sub-Super Class Test Reduction	Algorithm	Score	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	860	15	8	8	**SAHBN**	**NA**	**0.57**	**0.93**	**0.71**	**0.62**	**0.61**	**0.45**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.55	0.86	0.67	0.57	0.57	0.42
								mbde	0.58	1.00	0.73	0.63	0.63	0.42
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	1.00	0.26	0.41	0.63	0.63	0.41
							Tabu Search (tabu)	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.56	1.00	0.72	0.61	0.61	0.35
								mbde	0.56	1.00	0.72	0.60	0.60	0.32
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.75	0.25
								K2	0.54	0.66	0.59	0.54	0.54	0.51
2	CfsSubSetEval	GreedyStepwise	860	15	8	8	**SAHBN**	**NA**	**0.57**	**0.93**	**0.71**	**0.62**	**0.61**	**0.45**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.58	1.00	0.73	0.63	0.63	0.42
								mbde	0.54	0.86	0.67	0.57	0.56	0.40
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	1.00	0.26	0.41	0.63	0.63	0.41
							Tabu Search (tabu	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.56	1.00	0.72	0.61	0.61	0.35
								mbde	0.56	1.00	0.72	0.60	0.60	0.32

								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.50	0.57	0.53	0.50	0.50	0.48
3	ConsistencySubsetEval	BestFirst	860	15	5	5	**SAHBN**	**NA**	**0.59**	**1.00**	**0.74**	**0.65**	**0.65**	**0.46**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.59	1.00	0.74	0.64	0.64	0.44
								mbde	0.58	1.00	0.73	0.63	0.63	0.42
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.59	1.00	0.74	0.64	0.64	0.44
							Tabu Search (tabu	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.56	1.00	0.72	0.60	0.60	0.32
								mbde	0.56	1.00	0.72	0.61	0.61	0.35
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.59	1.00	0.74	0.64	0.64	0.44
4	ConsistencySubsetEval	GreedyStepwise	860	15	5	5	**SAHBN**	**NA**	**0.59**	**1.00**	**0.74**	**0.65**	**0.65**	**0.46**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.59	1.00	0.74	0.64	0.64	0.44
								mbde	0.58	1.00	0.73	0.63	0.63	0.42
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.58	1.00	0.73	0.63	0.63	0.42
							Tabu Search (tabu	loglik	0.56	1.00	0.72	0.60	0.60	0.32
								bde	0.56	1.00	0.72	0.60	0.60	0.32
								mbde	0.56	1.00	0.72	0.61	0.61	0.35
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.59	1.00	0.74	0.64	0.64	0.44

5	FilteredSubsetEval	BestFirst	860	15	8	8	**SAHBN**	**NA**	**0.57**	**0.93**	**0.71**	**0.62**	**0.61**	**0.45**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.56	0.93	0.70	0.60	0.60	0.41
								mbde	0.59	1.00	0.74	0.64	0.64	0.44
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.65	0.29	0.40	0.57	0.57	0.43
							Tabu Search (tabu	loglik	0.56	1.00	0.72	0.60	0.60	0.32
								bde	0.56	1.00	0.72	0.61	0.61	0.35
								mbde	0.57	1.00	0.73	0.62	0.61	0.37
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.49	0.83	0.62	0.49	0.48	0.24
								K2	0.58	0.86	0.69	0.62	0.62	0.52
6	FilteredSubsetEval	GreedyStepwise	860	15	8	8	**SAHBN**	**NA**	**0.57**	**0.93**	**0.71**	**0.62**	**0.61**	**0.45**
							Hill-Climbing (hc)	loglik	0.56	1.00	0.72	0.60	0.60	0.32
								bde	0.55	0.84	0.67	0.57	0.57	0.44
								mbde	0.59	1.00	0.74	0.64	0.64	0.44
								aic	0.52	0.88	0.65	0.52	0.52	0.27
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	1.00	0.26	0.41	0.63	0.63	0.41
							Tabu Search (tabu	loglik	0.56	1.00	0.72	0.61	0.61	0.35
								bde	0.56	1.00	0.72	0.61	0.61	0.35
								mbde	0.57	1.00	0.73	0.62	0.61	0.37
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.52	0.66	0.58	0.52	0.52	0.49
7	InfoGainAttributeEval	Ranker	860	15	5	5	**SAHBN**	**NA**	**1.00**	**0.19**	**0.32**	**0.59**	**0.59**	**0.32**
							Hill-Climbing	loglik	0.54	1.00	0.70	0.57	0.57	0.25
								bde	0.52	0.79	0.63	0.52	0.52	0.38

							(hc)	mbde	0.49	0.74	0.59	0.49	0.48	0.35
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.45	0.38	0.41	0.45	0.45	0.44
							Tabu Search (tabu)	loglik	0.54	1.00	0.70	0.57	0.57	0.25
								bde	0.37	0.17	0.24	0.43	0.44	0.28
								mbde	0.58	0.12	0.20	0.51	0.52	0.21
								aic	0.75	0.16	0.26	0.55	0.55	0.27
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.54	0.24	0.33	0.51	0.52	0.37
8	GainRatioAttributeEval	Ranker	860	15	5	5	**SAHBN**	**NA**	**0.58**	**0.26**	**0.36**	**0.53**	**0.53**	**0.39**
							Hill-Climbing (hc)	loglik	0.54	1.00	0.70	0.57	0.57	0.25
								bde	0.53	0.88	0.66	0.54	0.54	0.32
								mbde	0.53	0.79	0.63	0.54	0.54	0.41
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.52	0.76	0.62	0.52	0.52	0.41
							Tabu Search (tabu)	loglik	0.54	1.00	0.70	0.57	0.57	0.25
								bde	0.62	0.14	0.23	0.52	0.53	0.24
								mbde	0.36	0.17	0.23	0.43	0.43	0.28
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.34	0.21	0.26	0.40	0.40	0.31
								K2	0.52	0.83	0.64	0.53	0.53	0.37
9	CorrelationAttributeEval	Ranker	860	15	5	5	**SAHBN**	**NA**	**0.55**	**0.98**	**0.71**	**0.59**	**0.59**	**0.32**
							Hill-Climbing (hc)	loglik	0.55	0.98	0.71	0.59	0.59	0.32
								bde	0.62	0.22	0.33	0.54	0.54	0.36
								mbde	0.47	0.31	0.38	0.48	0.48	0.42
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25

								K2	0.49	0.38	0.43	0.49	0.49	0.46
							Tabu Search (tabu)	loglik	0.55	0.98	0.70	0.58	0.58	0.30
								bde	0.50	0.34	0.41	0.50	0.50	0.45
								mbde	0.47	0.28	0.35	0.48	0.48	0.39
								aic	0.31	0.19	0.23	0.37	0.38	0.28
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.50	0.29	0.37	0.50	0.50	0.41
10	ReliefFAttAttributeEval	Ranker	860	15	5	5	**SAHBN**	**NA**	**0.55**	**0.98**	**0.71**	**0.59**	**0.59**	**0.32**
							Hill-Climbing (hc)	loglik	0.55	0.98	0.70	0.58	0.58	0.30
								bde	0.43	0.38	0.40	0.43	0.44	0.43
								mbde	0.63	0.33	0.43	0.57	0.57	0.47
								aic	0.52	0.91	0.67	0.54	0.54	0.27
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.60	0.26	0.36	0.54	0.54	0.39
							Tabu Search (tabu)	loglik	0.55	0.98	0.71	0.59	0.59	0.32
								bde	0.50	0.26	0.34	0.50	0.50	0.38
								mbde	0.49	0.36	0.42	0.49	0.49	0.46
								aic	1.00	0.10	0.19	0.55	0.55	0.19
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.54	0.34	0.42	0.52	0.52	0.46
11	OneRAttributeEval	Ranker	860	15	5	5	**SAHBN**	**NA**	**0.55**	**0.98**	**0.71**	**0.59**	**0.59**	**0.32**
							Hill-Climbing (hc)	loglik	0.55	0.98	0.70	0.58	0.58	0.30
								bde	0.52	0.84	0.64	0.52	0.52	0.31
								mbde	0.55	1.00	0.71	0.59	0.59	0.30
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.51	0.79	0.62	0.50	0.50	0.33
							Tabu Search	loglik	0.56	1.00	0.72	0.60	0.60	0.32
								bde	0.53	0.86	0.65	0.54	0.54	0.34

							(tabu)	mbde	0.54	0.91	0.68	0.56	0.55	0.32
								aic	0.54	1.00	0.70	0.57	0.57	0.25
								bic	0.54	1.00	0.70	0.57	0.57	0.25
								K2	0.51	0.84	0.64	0.51	0.51	0.29

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

6.8 Homo-sapiens protein hub dataset experimental results

Table 6-8 Homo-sapiens protein hub dataset experimental results

No.	Attributes Selection		Initial Attr. No.	Attr. Selection Reduction.	Sub-Super Class Test Reduction	Algorithm	Score	Constraint	P	R	F1	A	AVA	HA
	Evaluation Method	Search Method												
1	CfsSubSetEval	BestFirst	6455	157	141	**SAHBN**	**NA**		**0.76**	**0.71**	**0.74**	**0.74**	**0.74**	**0.74**
						Hill-Climbing (hc)	aic	NA	0.78	0.50	0.61	0.68	0.68	0.63
							bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.91	0.60	0.72	0.77	0.77	0.73
						Tabu Search (tabu)	aic		0.78	0.50	0.61	0.68	0.68	0.63
							bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.92	0.62	0.74	0.78	0.78	0.75
						Grow-Shring (gs)	NA	mi	0.44	0.48	0.46	0.43	0.43	0.42
								mi-adf	0.44	0.53	0.48	0.43	0.43	0.41
								mi-sh	0.44	0.40	0.42	0.45	0.45	0.44
								X2	0.42	0.34	0.37	0.44	0.44	0.41
								X2-adf	0.47	0.52	0.49	0.46	0.46	0.46
						Incremental Association (iamb)	NA	mi	0.94	0.19	0.31	0.59	0.59	0.32
								mi-adf	0.45	0.37	0.41	0.46	0.46	0.44
								mi-sh	0.94	0.19	0.31	0.59	0.59	0.32
								X2	0.94	0.19	0.31	0.59	0.59	0.32
								X2-adf	0.42	0.25	0.31	0.45	0.45	0.36
						Fast Incremental Association (fast.iamb)	NA	mi	0.46	0.64	0.53	0.44	0.44	0.35
								mi-adf	0.44	0.42	0.43	0.45	0.45	0.45
								mi-sh	0.48	0.58	0.53	0.47	0.47	0.45
								X2	0.43	0.48	0.45	0.43	0.43	0.42
								X2-adf	0.77	0.28	0.41	0.60	0.60	0.42
						Interleaved	NA	mi	0.94	0.19	0.31	0.59	0.59	0.32

						Incremental Association (inter.iamb)		mi-adf	0.47	0.56	0.51	0.46	0.46	0.44
								mi-sh	0.94	0.19	0.31	0.59	0.59	0.32
								X2	0.94	0.19	0.31	0.59	0.59	0.32
								X2-adf	0.46	0.50	0.48	0.45	0.45	0.45
2	CfsSubSetEval	GreedyStep wise	6455	502	408	**SAHBN**	**NA**		0.80	0.67	0.73	0.75	0.75	0.74
						Hill-Climbing (hc)	aic	NA	0.85	0.58	0.69	0.74	0.74	0.71
							bic		0.94	0.26	0.41	0.63	0.63	0.41
							K2		0.85	0.78	0.81	0.82	0.82	0.82
						Tabu Search (tabu)	aic		0.84	0.58	0.69	0.74	0.74	0.70
							bic		0.97	0.26	0.41	0.63	0.63	0.41
							K2		0.84	0.78	0.81	0.82	0.82	0.81
						Grow-Shring (gs)	NA	mi	0.45	0.54	0.49	0.44	0.44	0.41
								mi-adf	0.89	0.10	0.17	0.54	0.54	0.18
								mi-sh	0.48	0.40	0.44	0.49	0.49	0.47
								X2	0.46	0.47	0.46	0.45	0.45	0.45
								X2-adf	0.46	0.55	0.50	0.45	0.45	0.43
						Incremental Association (iamb)	NA	mi	0.81	0.52	0.64	0.70	0.70	0.66
								mi-adf	0.44	0.44	0.44	0.44	0.44	0.44
								mi-sh	0.81	0.52	0.64	0.70	0.70	0.66
								X2	0.44	0.35	0.39	0.45	0.45	0.43
								X2-adf	0.45	0.44	0.44	0.45	0.45	0.45
						Fast Incremental Association (fast.iamb)	NA	mi	0.81	0.52	0.64	0.70	0.70	0.66
								mi-adf	0.48	0.57	0.52	0.48	0.48	0.46
								mi-sh	0.81	0.52	0.64	0.70	0.70	0.66
								X2	0.43	0.34	0.38	0.44	0.44	0.42
								X2-adf	0.44	0.41	0.43	0.45	0.45	0.44
						Interleaved Incremental Association	NA	mi	0.81	0.52	0.64	0.70	0.70	0.66
								mi-adf	0.48	0.65	0.55	0.47	0.47	0.41
								mi-sh	0.81	0.52	0.64	0.70	0.70	0.66

						(inter.iamb)		X2	0.81	0.52	0.64	0.70	0.70	0.66
								X2-adf	0.45	0.54	0.49	0.44	0.44	0.42
3	ConsistencySubsetEval	BestFirst	6455	44	36	**SAHBN**	**NA**		0.78	0.72	0.75	0.76	0.76	0.76
						Hill-Climbing (hc)	aic	NA	0.94	0.19	0.31	0.59	0.59	0.32
							bic		0.94	0.19	0.31	0.59	0.59	0.32
							K2		0.78	0.71	0.75	0.76	0.76	0.76
						Tabu Search (tabu)	aic		0.96	0.11	0.19	0.55	0.55	0.19
							bic		0.94	0.19	0.31	0.59	0.59	0.32
							K2		0.79	0.70	0.74	0.76	0.76	0.75
						Grow-Shring (gs)	NA	mi	0.45	0.50	0.47	0.44	0.44	0.44
								mi-adf	0.43	0.39	0.41	0.43	0.43	0.43
								mi-sh	0.46	0.55	0.50	0.45	0.45	0.43
								X2	0.41	0.45	0.43	0.40	0.40	0.40
								X2-adf	0.41	0.32	0.36	0.42	0.42	0.40
						Incremental Association (iamb)	NA	mi	0.85	0.19	0.31	0.58	0.58	0.31
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.85	0.19	0.31	0.58	0.58	0.31
								X2	0.41	0.29	0.34	0.44	0.44	0.39
								X2-adf	0.41	0.29	0.34	0.44	0.44	0.39
						Fast Incremental Association (fast.iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.48	0.45	0.46	0.48	0.48	0.48
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.80	0.47	0.59	0.68	0.68	0.61
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.85	0.19	0.31	0.58	0.58	0.31
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.85	0.19	0.31	0.58	0.58	0.31
								X2	0.46	0.37	0.41	0.47	0.47	0.45
								X2-adf	0.46	0.37	0.41	0.47	0.47	0.45

4	ConsistencySubsetEval	GreedyStepwise	6455	51	44	**SAHBN**	**NA**		0.74	0.70	0.72	0.73	0.73	0.73
						Hill-Climbing (hc)	aic	NA	0.77	0.56	0.65	0.70	0.70	0.67
							bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.75	0.65	0.70	0.72	0.72	0.71
						Tabu Search (tabu)	aic		0.77	0.56	0.65	0.70	0.70	0.67
							bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.77	0.64	0.70	0.72	0.72	0.71
						Grow-Shring (gs)	NA	mi	0.84	0.13	0.22	0.55	0.55	0.23
								mi-adf	0.46	0.55	0.50	0.45	0.45	0.43
								mi-sh	0.84	0.13	0.22	0.55	0.55	0.23
								X2	0.47	0.51	0.49	0.47	0.47	0.46
								X2-adf	0.46	0.56	0.51	0.46	0.46	0.43
						Incremental Association (iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.62
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.80	0.47	0.59	0.68	0.68	0.61
						Fast Incremental Association (fast.iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.78	0.60	0.68	0.71	0.71	0.70
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.41	0.32	0.36	0.43	0.43	0.40
								X2-adf	0.46	0.55	0.50	0.45	0.45	0.43
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.80	0.47	0.59	0.68	0.68	0.61
5	FilteredSubsetEval	BestFirst	6455	49	39	**SAHBN**	**NA**		0.82	0.79	0.80	0.81	0.81	0.81
						Hill-	aic	NA	0.78	0.50	0.60	0.68	0.68	0.63

						Climbing (hc)	bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.92	0.60	0.72	0.77	0.77	0.73
						Tabu Search (tabu)	aic		0.78	0.50	0.60	0.68	0.68	0.63
							bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.92	0.59	0.72	0.77	0.77	0.73
						Grow-Shring (gs)	NA	mi	0.42	0.48	0.45	0.41	0.41	0.40
								mi-adf	0.44	0.35	0.39	0.45	0.45	0.43
								mi-sh	0.40	0.40	0.40	0.40	0.40	0.40
								X2	0.44	0.35	0.39	0.45	0.45	0.43
								X2-adf	0.44	0.35	0.39	0.45	0.45	0.43
						Incremental Association (iamb)	NA	mi	0.94	0.19	0.31	0.59	0.59	0.32
								mi-adf	0.47	0.56	0.51	0.46	0.46	0.44
								mi-sh	0.94	0.19	0.31	0.59	0.59	0.32
								X2	0.94	0.19	0.31	0.59	0.59	0.32
								X2-adf	0.46	0.55	0.50	0.45	0.45	0.43
						Fast Incremental Association (fast.iamb)	NA	mi	0.46	0.50	0.48	0.46	0.46	0.45
								mi-adf	0.44	0.40	0.42	0.45	0.45	0.44
								mi-sh	0.46	0.55	0.50	0.45	0.43	0.43
								X2	0.44	0.35	0.39	0.45	0.45	0.43
								X2-adf	0.44	0.30	0.36	0.46	0.46	0.40
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.94	0.19	0.31	0.59	0.59	0.32
								mi-adf	0.46	0.55	0.50	0.45	0.45	0.43
								mi-sh	0.94	0.19	0.31	0.59	0.59	0.32
								X2	0.94	0.19	0.31	0.59	0.59	0.32
								X2-adf	0.43	0.34	0.38	0.44	0.44	0.42
6	FilteredSubsetEval	GreedyStepwise	6455	53	41	**SAHBN**	**NA**		0.82	0.80	0.81	0.81	0.81	0.81
						Hill-Climbing (hc)	aic	NA	0.76	0.43	0.55	0.65	0.65	0.57
							bic		0.76	0.43	0.55	0.65	0.65	0.57
							K2		0.84	0.63	0.72	0.76	0.76	0.74

						Tabu Search (tabu)	aic		0.44	0.48	0.46	0.43	0.43	0.43
							bic		0.46	0.55	0.50	0.45	0.45	0.43
							K2		0.84	0.64	0.73	0.76	0.76	0.74
						Grow-Shring (gs)	NA	mi	0.45	0.36	0.40	0.46	0.46	0.44
								mi-adf	0.45	0.50	0.47	0.45	0.45	0.44
								mi-sh	0.44	0.30	0.36	0.46	0.46	0.41
								X2	0.49	0.54	0.51	0.49	0.49	0.48
								X2-adf	0.44	0.48	0.46	0.43	0.43	0.42
						Incremental Association (iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.48	0.50	0.49	0.48	0.48	0.48
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.45	0.50	0.47	0.45	0.45	0.44
						Fast Incremental Association (fast.iamb)	NA	mi	0.47	0.56	0.51	0.46	0.46	0.44
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.86	0.20	0.32	0.58	0.58	0.33
								X2-adf	0.41	0.32	0.36	0.43	0.43	0.40
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.46	0.41	0.43	0.46	0.46	0.46
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.43	0.38	0.41	0.44	0.44	0.43
7	InfoGainAttributeEval	Ranker	6455	502	349	**SAHBN**	**NA**		0.84	0.80	0.82	0.82	0.82	0.82
						Hill-Climbing (hc)	aic	NA	0.96	0.29	0.44	0.64	0.64	0.45
							bic		0.89	0.34	0.49	0.65	0.65	0.50
						Tabu Search (tabu)	aic		0.96	0.29	0.44	0.64	0.64	0.45
							bic		0.89	0.34	0.49	0.65	0.65	0.50

						Grow-Shring (gs)	NA	mi	0.48	0.54	0.51	0.48	0.48	0.47
								mi-adf	0.43	0.37	0.39	0.44	0.44	0.43
								mi-sh	0.47	0.50	0.48	0.47	0.47	0.47
								X2	0.46	0.50	0.48	0.45	0.45	0.45
								X2-adf	0.45	0.54	0.49	0.45	0.45	0.43
						Incremental Association (iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.46	0.50	0.48	0.45	0.45	0.45
								X2-adf	0.43	0.48	0.46	0.43	0.43	0.42
						Fast Incremental Association (fast.iamb)	NA	mi	0.96	0.27	0.42	0.63	0.63	0.42
								mi-adf	0.77	0.52	0.62	0.68	0.68	0.64
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.77	0.50	0.61	0.67	0.67	0.63
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.80	0.47	0.59	0.68	0.68	0.61
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.80	0.47	0.59	0.68	0.68	0.61
								X2	0.44	0.39	0.41	0.44	0.44	0.44
								X2-adf	0.44	0.53	0.48	0.43	0.43	0.40
8	GainRatioAttributeEval	Ranker	6544	502	382	**SAHBN**	**NA**		0.89	0.79	0.84	0.85	0.85	0.84
						Hill-Climbing (hc)	aic	NA	0.95	0.33	0.49	0.66	0.66	0.50
							bic		0.95	0.28	0.43	0.63	0.63	0.43
						Tabu Search (tabu)	aic		0.95	0.33	0.49	0.66	0.66	0.50
							bic		0.94	0.19	0.31	0.59	0.59	0.32
						Grow-Shring	NA	mi	0.43	0.52	0.47	0.42	0.42	0.39

						(gs)		mi-adf	0.48	0.57	0.52	0.48	0.48	0.47
								mi-sh	0.94	0.19	0.31	0.59	0.59	0.32
								X2	0.46	0.51	0.48	0.46	0.46	0.45
								X2-adf	0.42	0.30	0.35	0.44	0.44	0.39
						Incremental Association (iamb)	NA	mi	0.41	0.33	0.37	0.43	0.43	0.41
								mi-adf	0.41	0.29	0.34	0.43	0.43	0.39
								mi-sh	0.47	0.56	0.51	0.47	0.47	0.45
								X2	0.44	0.40	0.42	0.45	0.45	0.44
								X2-adf	0.47	0.51	0.49	0.47	0.47	0.46
						Fast Incremental Association (fast.iamb)	NA	mi	0.49	0.53	0.51	0.48	0.48	0.48
								mi-adf	0.96	0.11	0.19	0.55	0.55	0.19
								mi-sh	0.96	0.27	0.42	0.63	0.63	0.42
								X2	0.43	0.38	0.40	0.44	0.44	0.43
								X2-adf	0.42	0.38	0.40	0.43	0.43	0.42
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.46	0.50	0.48	0.46	0.46	0.45
								mi-adf	0.44	0.32	0.37	0.46	0.46	0.42
								mi-sh	0.45	0.46	0.46	0.45	0.45	0.45
								X2	0.42	0.38	0.40	0.43	0.43	0.42
								X2-adf	0.46	0.41	0.43	0.46	0.46	0.45
9	CorrelationAttributeEval	Ranker	6544	502	336	**SAHBN**	**NA**		0.89	0.76	0.82	0.84	0.84	0.83
						Hill-Climbing (hc)	aic	NA	0.96	0.29	0.44	0.64	0.64	0.45
							bic		0.89	0.34	0.49	0.65	0.65	0.50
						Tabu Search (tabu)	aic		0.92	0.10	0.17	0.54	0.54	0.18
							bic		0.89	0.34	0.49	0.65	0.65	0.50
						Grow-Shring (gs)	NA	mi	0.47	0.60	0.53	0.47	0.47	0.43
								mi-adf	0.44	0.45	0.44	0.43	0.43	0.43
								mi-sh	0.46	0.55	0.50	0.45	0.45	0.42

								X2	0.44	0.48	0.46	0.43	0.43	0.43
								X2-adf	0.44	0.53	0.48	0.43	0.43	0.41
						Incremental Association (iamb)	NA	mi	0.78	0.35	0.48	0.63	0.63	0.51
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.78	0.35	0.48	0.63	0.63	0.51
								X2	0.47	0.53	0.50	0.46	0.46	0.46
								X2-adf	0.45	0.34	0.39	0.46	0.46	0.43
						Fast Incremental Association (fast.iamb)	NA	mi	0.96	0.27	0.42	0.63	0.63	0.42
								mi-adf	0.77	0.52	0.62	0.68	0.68	0.64
								mi-sh	0.43	0.34	0.38	0.44	0.44	0.42
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.77	0.50	0.61	0.67	0.67	0.63
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.78	0.35	0.48	0.63	0.63	0.51
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.78	0.35	0.48	0.63	0.63	0.51
								X2	0.43	0.51	0.47	0.41	0.41	0.39
								X2-adf	0.47	0.66	0.55	0.46	0.46	0.37
10	ReliefFAttAttributeEval	Ranker	6544	502	371	**SAHBN**	**NA**		0.73	0.68	0.71	0.72	0.72	0.71
						Hill-Climbing (hc)	aic	NA	0.96	0.29	0.44	0.64	0.64	0.45
							bic		0.96	0.29	0.42	0.63	0.63	0.42
						Tabu Search (tabu)	aic		0.96	0.29	0.44	0.64	0.64	0.45
							bic		0.96	0.27	0.42	0.63	0.63	0.42
						Grow-Shring (gs)	NA	mi	0.47	0.52	0.50	0.47	0.47	0.47
								mi-adf	0.46	0.55	0.50	0.45	0.45	0.43
								mi-sh	0.46	0.41	0.43	0.46	0.46	0.45
								X2	0.44	0.35	0.39	0.45	0.45	0.43
								X2-adf	0.43	0.39	0.41	0.44	0.44	0.43

						Incremental Association (iamb)	NA	mi	0.96	0.27	0.42	0.63	0.63	0.42
								mi-adf	0.94	0.19	0.31	0.59	0.59	0.32
								mi-sh	0.86	0.20	0.32	0.58	0.58	0.33
								X2	0.46	0.53	0.49	0.45	0.45	0.44
								X2-adf	0.47	0.51	0.49	0.47	0.47	0.46
						Fast Incremental Association (fast.iamb)	NA	mi	0.83	0.19	0.31	0.58	0.58	0.32
								mi-adf	0.46	0.58	0.51	0.44	0.44	0.40
								mi-sh	0.46	0.59	0.52	0.45	0.45	0.41
								X2	0.86	0.20	0.32	0.58	0.58	0.33
								X2-adf	0.42	0.51	0.46	0.41	0.41	0.38
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.96	0.27	0.42	0.63	0.63	0.42
								mi-adf	0.94	0.19	0.31	0.59	0.59	0.32
								mi-sh	0.86	0.20	0.32	0.58	0.58	0.33
								X2	0.94	0.19	0.31	0.59	0.59	0.32
								X2-adf	0.44	0.48	0.46	0.43	0.43	0.42
11	OneRAttributeEval	Ranker	6544	502	336	**SAHBN**	**NA**		0.83	0.80	0.81	0.82	0.82	0.82
						Hill-Climbing (hc)	aic	NA	0.94	0.31	0.46	0.64	0.64	0.47
							bic		0.89	0.34	0.49	0.65	0.65	0.50
						Tabu Search (tabu)	aic		0.94	0.30	0.46	0.64	0.64	0.46
							bic		0.89	0.34	0.49	0.65	0.65	0.50
						Grow-Shring (gs)	NA	mi	0.43	0.34	0.38	0.44	0.44	0.42
								mi-adf	0.45	0.34	0.39	0.46	0.46	0.43
								mi-sh	0.46	0.41	0.43	0.46	0.46	0.46
								X2	0.45	0.54	0.49	0.44	0.44	0.41
								X2-adf	0.45	0.32	0.37	0.46	0.46	0.42
						Incremental Association	NA	mi	0.78	0.35	0.48	0.63	0.63	0.51
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61

						(iamb)		mi-sh	0.78	0.35	0.48	0.63	0.63	0.51
								X2	0.45	0.59	0.51	0.44	0.44	0.38
								X2-adf	0.44	0.49	0.47	0.44	0.44	0.43
						Fast Incremental Association (fast.iamb)	NA	mi	0.96	0.27	0.42	0.63	0.63	0.42
								mi-adf	0.77	0.52	0.62	0.68	0.68	0.64
								mi-sh	0.47	0.52	0.49	0.47	0.47	0.46
								X2	0.80	0.47	0.59	0.68	0.68	0.61
								X2-adf	0.77	0.50	0.61	0.68	0.68	0.63
						Interleaved Incremental Association (inter.iamb)	NA	mi	0.78	0.35	0.48	0.63	0.63	0.51
								mi-adf	0.80	0.47	0.59	0.68	0.68	0.61
								mi-sh	0.78	0.35	0.48	0.63	0.63	0.51
								X2	0.44	0.36	0.40	0.45	0.45	0.43
								X2-adf	0.47	0.59	0.52	0.46	0.46	0.43

P: Precision, R: Recall, F1: F1 measure, A: Accuracy, AVA: Average Accuracy, HA: Harmonic Accuracy.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, 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," *Constr. Manag. Econ.*, 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 Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 755–769, 2010.

[6] R. Paul, T. Groza, J. Hunter, and A. Zankl, "Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain.," *J. Biomed. Semantics*, vol. 5, no. 1, p. 8, 2014.

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

[8] C. Antunes and A. Silva, "New Trends in Knowledge Driven Data Mining a position paper," *Proc. 16th Int. Conf. Enterp. Inf. Syst.*, pp. 346–351, 2014.

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

[10] G. Mansingh and L. Rao, "The Role of Ontologies in Developing Knowledge Technologies," *Knowl. Manag. Dev. Springer US*, pp. 145–156, 2014.

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

[12] 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.

[13] P. Ristoski and H. Paulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey,” *Web Semant. Sci. Serv. Agents*, 2016.

[14] L. Cao, “Actionable knowledge discovery and delivery,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 149–163, 2012.

[15] V. Narasimha, P. Kappara, R. Ichise, and O. P. Vyas, “LiDDM: A Data Mining System for Linked Data,” in *Workshop on Linked Data on the Web. CEUR Workshop Proceedings*, 2011, vol. 813.

[16] H. Paulheim, “Exploiting Linked Open Data as Background Knowledge in Data Mining,” *DMoLD, 1082*, 2013.

[17] M. Egaòa Aranguren, J. T. Fernández-Breis, and M. Dumontier, “Special issue on Linked Data for Health Care and the Life Sciences,” *Semant. Web*, vol. 5, no. 2, pp. 99–100, 2014.

[18] A. Lausch, A. Schmidt, and L. Tischendorf, “Data mining and linked open data – New perspectives for data analysis in environmental research,” *Ecol. Modell.*, vol. 295, pp. 5–17, 2015.

[19] B. Brumen, M. Herièko, A. Sevènikar, J. Zavrnik, and M. Hölbl, “Outsourcing medical data analyses: Can technology overcome legal, privacy, and confidentiality issues?,” *J. Med. Internet Res.*, vol. 15, no. 12, 2013.

[20] C. R. Kothari, *Research methodology: Methods and techniques*. New Age International, 2004.

[21] B. F. Crabtree and W. L. Miller, *Doing qualitative research*. Sage Publications, 1999.

[22] D. Silverman, *Doing qualitative research: A practical handbook*. SAGE Publications Limited, 2013.

[23] L. Rokach and O. Maimon, *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2010.

[24] M. Bramer, *Principles of Data Mining*. London: Springer London, 2013.

[25] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[26] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[27] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," *Knowl. Eng. Rev.*, vol. 21, no. 1, pp. 1–24, 2006.

[28] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and others, "Knowledge discovery and data mining: towards a unifying framework.," in *AAAI Press*, 1996, pp. 82–88.

[29] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.

[30] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, vol. 81. WORLD SCIENTIFIC, 2014.

[31] L. Cao, P. S. Yu, C. Zhang, and H. Zhang, *Data mining for business applications*. Boston, MA: Springer US, 2009.

[32] R. Athauda, C. Fernando, and M. Tissera, *Data Mining Applications: Promise and Challenges*. INTECH Open Access Publisher, 2009.

[33] L. Cao, C. Zhang, Q. Yang, D. Bell, M. Vlachos, P. S. Yu, B. Taneri, E. Keogh, N. Zhong, M. Z. Ashrafi, D. Taniar, E. Dubossarsky, and W. Graco, "Domain-driven, actionable knowledge discovery," *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 78–88, 2007.

[34] Z. Ma, F. Zhang, L. Yan, and J. Cheng, *Fuzzy Knowledge Management for the Semantic Web*. Springer, 2014.

[35] A. Maedche and S. Staab, "Ontology Learning for The Semantic Web," *IEEE Intell. Syst.*, vol. 16, pp. 72--79, 2001.

[36] V. Sugumaran and G. Jon Atle, *Applied semantic web technologies*. CRC Press, 2011.

[37] G. Antoniou, P. Groth, F. van Harmelen, and R. Hoekstr, *A Semantic Web Primer*. The MIT Press, 2012.

[38] L. M. Campbell and S. MacNeill, "The Semantic Web , Linked and Open Data. A Briefing Paper," *World Wide Web Internet Web Inf. Syst.*, p. 6, 2010.

[39] C. Bizer, "The emerging web of linked data," *IEEE Intell. Syst.*, vol. 24, no. 5, pp. 87–92, 2009.

[40] L. Yu, *A Developer's Guide to the Semantic Web*. Springer Science & Business Media, 2011.

[41] A. Gerber, A. Van Der Merwe, and A. Barnard, "A functional semantic web architecture," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5021 LNCS, pp. 273–287, 2008.

[42] T. Heath and C. Bizer, "Linked Data Evolving the Web into a Global Data Space," *Synth. Lect. Semant. web theory Technol.*, vol. 1, pp. 1--36, 2011.

[43] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Int. J. Semant. Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[44] E. H. Y. Lim, J. N. K. Liu, and R. S. T. Lee, *Knowledge Seeker – OntologyModelling for Information Search and Management*. Verlag - Berlin An.: springer, 2013.

[45] J. Domingue, D. Fensel, and J. Hendler, *Handbook of Semantic Web Technologies*. Springer Science & Business Media, 2011.

[46] A. Gómez-Pérez, M. Fernández-López, O. Corcho, and A. Gomez-Perez, *Ontological Engeenering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Publishing Company, Incorporated, 2010.

[47] M. Uschold and M. Gruninger, "Ontologies: Principles, methods and applications," *Knowl. Eng. Rev.*, vol. 11, no. 2, pp. 93–136, 1996.

[48] A. Malhotra, E. Younesi, M. Gundel, B. Muller, M. T. Heneka, and M. Hofmann-Apitius, "ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease," *Alzheimer's Dement.*, vol. 10, no. 2, pp. 238–246, 2014.

[49] K. K. Breitman, M. a Casanova, and W. Truszkowski, *Semantic Web: Concepts, Technologies and Applciations*. Springer Science & Business Media, 2007.

[50] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," *Proc. 2015 IEEE 9th Int. Conf. Semant. Comput. IEEE ICSC 2015*, pp. 244–251, 2015.

[51] 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.

[52] J. Polpinij and A. K. Ghose, “An ontology-based sentiment classification methodology for online consumer reviews,” in *Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008*, 2008, pp. 518–524.

[53] J. A. Motta, L. Capus, and N. Tourigny, “VENCE: A new machine learning method enhanced by ontological knowledge to extract summaries,” in *SAI Computing Conference (SAI), 2016*, 2016, pp. 61–70.

[54] T. Mabotuwana, M. C. Lee, and E. V. Cohen-Solal, “An ontology-based similarity measure for biomedical data - Application to radiology reports,” *J. Biomed. Inform.*, vol. 46, no. 5, pp. 857–868, 2013.

[55] Z. Kastrati, A. S. Imran, and S. Y. Yayilgan, “An Improved Concept Vector Space Model for Ontology Based Classification,” in *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2015, pp. 240–245.

[56] S. Albitar, B. Espinasse, and S. Fournier, “Semantic Enrichments in Text Supervised Classification: Application to Medical Domain,” *Proc. Twenty-Seventh Int. Florida Artif. Intell. Res. Soc. Conf.*, no. i, pp. 425–430, 2013.

[57] L. Pereira, R. Rijo, C. Silva, and M. Agostinho, “Using text mining to diagnose and classify epilepsy in children,” in *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services, Healthcom 2013*, 2013, pp. 345–349.

[58] Z. Elberrichi, B. Amel, and T. Malika, “Medical Documents Classification Based on the Domain Ontology MeSH,” *arXiv Prepr. arXiv1207.0446*, 2012.

[59] J. Wang, G. Wu, and X. Hu, “An ontology-based dimensionality reduction algorithm for biomedical literature classification,” in *Conference Anthology, IEEE*, 2013, pp. 1–5.

[60] C. M. Wijewickrema and R. Gamage, “An ontology based fully automatic document classification system using an existing semi-automatic system,” 2013.

[61] N. M. De Mel, H. H. Hettiarachchi, W. P. D. Madusanka, G. L. Malaka, A. S. Perera, and U. Kohomban, “Machine learning approach to recognize subject based sentiment

values of reviews," in *2016 Moratuwa Engineering Research Conference (MERCon)*, 2016, pp. 6–11.

[62] M. Allahyari, K. J. Kochut, and M. Janik, "Ontology-based text classification into dynamically defined topics," in *Semantic Computing (ICSC), 2014 IEEE International Conference on*, 2014, pp. 273–278.

[63] N. Sanchez-Pi, L. Martí, and A. C. Bicharra Garcia, "Improving ontology-based text classification: An occupational health and security application," *J. Appl. Log.*, vol. 17, pp. 48–58, 2016.

[64] F. Kiamarzpour, R. Dianat, M. Sadeghzadeh, and others, "Improving the methods of email classification based on words ontology," *arXiv Prepr. arXiv1310.5963*, 2013.

[65] F. Franca, S. Schulz, P. Bronsert, P. Novais, and M. Boeker, "Feasibility of an ontology driven tumor-node-metastasis classifier application: A study on colorectal cancer," in *Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on*, 2015, pp. 1–7.

[66] S. Rudolph, "Foundations of description logics," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6848 LNCS, pp. 76–136.

[67] K. Wolstencroft, P. Lord, L. Tabernero, A. Brass, and R. Stevens, "Protein classification using ontology classification," in *Bioinformatics*, 2006, vol. 22, no. 14.

[68] F. Ongenae, T. Dhaene, F. De Turck, D. Benoit, and J. Decruyenaere, "Design of a probabilistic ontology-based clinical decision support system for classifying temporal patterns in the ICU: A sepsis case study," in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2010, pp. 389–394.

[69] Y. Kassahun, R. Perrone, E. De Momi, E. Berghöfer, L. Tassi, M. P. Canevini, R. Spreafico, G. Ferrigno, and F. Kirchner, "Automatic classification of epilepsy types using ontology-based and genetics-based machine learning," *Artif. Intell. Med.*, vol. 61, no. 2, pp. 79–88, 2014.

[70] M. Belgiu, I. Tomljenovic, T. J. Lampoltshammer, T. Blaschke, and B. Höfle, "Ontology-based classification of building types detected from airborne laser scanning

data," *Remote Sens.*, vol. 6, no. 2, pp. 1347–1366, 2014.

[71] E. F. Luque, D. L. Rubin, and D. A. Moreira, "Automatic classification of cancer tumors using image annotations and ontologies," in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2015, vol. 2015–July, pp. 368–369.

[72] G. Costa, Paulo C G, Laskey, Kathryn B & AlGhamdi, "Bayesian Ontologies in AI Systems," *Fourth Bayesian Model. Appl. Work.*, 2006.

[73] T. Lukasiewicz and U. Straccia, "Managing uncertainty and vagueness in description logics for the Semantic Web," *Web Semant.*, vol. 6, no. 4, pp. 291–308, 2008.

[74] Z. Ding, Y. Peng, and R. Pan, "BayesOWL : Uncertainty Modelling in Semantic Web Ontologies," *Soft Comput. Ontol. Semant. Web*, vol. 204, pp. 3–29, 2006.

[75] Y. Sun, "A Prototype Implementation of BayesOWL ,Doctoral Dissertation. University of Mayryland Baltimore County." Doctoral Dissertation. University of Mayryland Baltimore County, 2009.

[76] Z. Ding, "BayesOWL," 2008. [Online]. Available: http://www.csee.umbc.edu/~ypeng/BayesOWL/index.html. [Accessed: 17-Oct-2016].

[77] S. Zhang, Y. Sun, Y. Peng, and X. Wang, "A Practical Tool for Uncertainty in OWL Ontologies," *Proc. 10th IASTED Int. Conf.*, vol. 674, no. 7, p. 235, 2009.

[78] Z. Ding and Y. Peng, "A Bayesian approach to uncertainty modelling in OWL ontology, MARYLAND UNIV BALTIMORE DEPT OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING." MARYLAND UNIV BALTIMORE DEPT OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING, 2006.

[79] S. Zhang, Y. Sun, Y. Peng, and X. Wang, "BayesOWL : A Prototype System for Uncertainty in Semantic Web," *Sci. Technol.*, 2009.

[80] P. Mitra, N. F. Noy, and A. R. Jaiswal, "OMEN: A probabilistic ontology mapping tool," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, vol. 3729 LNCS, pp. 537–547.

[81] A. S. Larik and S. Haider, "Efforts to blend ontology with Bayesian networks: An

overview," in *ICACTE 2010 - 2010 3rd International Conference on Advanced Computer Theory and Engineering, Proceedings*, 2010, vol. 2.

[82] P. Mitra, N. F. Noy, and a. R. A. R. Jaiswal, "Ontology mapping discovery with uncertainty," in *Proc. 4th International Semantic Web Conference (ISWC)*, 2005, vol. 3729, pp. 537–547.

[83] B. Andrea and T. Franco, "Extending ontology queries with Bayesian network reasoning," *Proc. - 2009 Int. Conf. Intell. Eng. Syst. INES 2009*, pp. 165–170, 2009.

[84] A. Devitt, B. Danev, and K. Matusikova, "Constructing Bayesian Networks Automatically using Ontologies," *Appl. Ontol.*, vol. 1, no. 1, 2006.

[85] H. T. Zheng, B. Y. Kang, and H. G. Kim, "An ontology-based bayesian network approach for representing uncertainty in Clinical Practice Guidelines," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5327 LNAI, pp. 161–173.

[86] S. Fenz, A. M. Tjoa, and M. Hudec, "Ontology-based generation of bayesian networks," in *Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, CISIS*, 2009, pp. 712–717.

[87] K. B. Laskey, "MEBN: A language for first-order Bayesian knowledge bases," *Artif. Intell.*, vol. 172, no. 2–3, pp. 140–178, 2008.

[88] P. C. G. Da Costa, K. B. Laskey, and K. J. Laskey, "PR-OWL: A bayesian ontology language for the Semantic Web," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5327 LNAI, pp. 88–107.

[89] E. M. Helsper and L. C. Van der Gaag, "Building Bayesian networks through ontologies," *ECAI2002, Proc. 15th Eur. Conf. Artif. Intell.*, pp. 680–684, 2002.

[90] Z. Ding, Y. Peng, R. Pan, and Y. Yu, "A Probabilistic Extension to Ontology Language OWL," *Proc. 37th Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 0, no. C, pp. 1–10, 2004.

[91] Y. Yang and J. Calmet, "OntoBayes: An Ontology-Driven Uncertainty Model," *Proc. Int. Conf. Comput. Intell. Model. Control Autom.*, vol. 1, pp. 457–463, 2005.

[92] M. Ben Ishak, P. Leray, and N. Ben Amor, “Ontology-based generation of object oriented bayesian networks,” in *CEUR Workshop Proceedings*, 2011, vol. 818, pp. 9–17.

[93] G. Bucci, V. Sandrucci, and E. Vicario, “Ontologies and Bayesian networks in medical diagnosis,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 1–8, 2011.

[94] S. Fenz, “An ontology-based approach for constructing Bayesian networks,” *Data Knowl. Eng.*, vol. 73, pp. 73–88, 2012.

[95] M. S. Sayed and N. Lohse, “Ontology-driven generation of Bayesian diagnostic models for assembly systems,” *Int. J. Adv. Manuf. Technol.*, vol. 74, no. 5–8, pp. 1033–1052, 2014.

[96] Y. Zhou, N. Fenton, and C. Zhu, “An empirical study of Bayesian network parameter learning with monotonic influence constraints,” *Decis. Support Syst.*, 2016.

[97] M. J. Druzdzel and F. J. Díez, “Combining knowledge from different sources in causal probabilistic models,” *J. Mach. Learn. Res.*, vol. 4, pp. 295–316, 2003.

[98] 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.

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

[100] P. Gaudet, N. Škunca, J. C. Hu, and C. Dessimoz, “Primer on the Gene Ontology,” *arXiv Prepr. arXiv1602.01876*, 2016.

[101] 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.

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

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

[104] 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.

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

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

[107] M. F. Zibran, "Chi-squared test of independence," *Handb. Biol. Stat.*, pp. 1–7, 2015.

[108] M. L. McHugh, "The Chi-square test of independence," *Biochem. Medica*, vol. 23, no. 2, pp. 143–149, 2012.

[109] D. S. Shafer and Z. Zhang, *Introductory Statistics*. Washington, DC, USA:Flat World Knowledge, 2012.

[110] P. J. F. Lucas, L. C. Van Der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and health-care," *Artif. Intell. Med.*, vol. 30, no. 3, pp. 201–214, 2004.

[111] B. Networks, F. Faltin, and R. Kenett, "Bayesian Networks," *Encycl. Stat. Qual. Reliab.*, vol. 1, no. 1, p. 4, 2007.

[112] C. Bielza and P. Larrañaga, "Bayesian networks in neuroscience: A survey," *Front. Comput. Neurosci.*, vol. 8, no. October, p. 131, 2014.

[113] R. E. Neapolitan, *Learning Bayesian Networks*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2003.

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

[115] K. B. Korb and A. E. Nicholson., *Bayesian Artificial intelligence.*, vol. 180, no. 8. CRC press, 2010.

[116] T. J. T. Koski and J. M. Noble, "A review of bayesian networks and structure learning," *Math. Appl.*, vol. 40, no. 1, pp. 53–103, 2012.

[117] P. Larrañaga, H. Karshenas, C. Bielza, and R. Santana, "A review on evolutionary

algorithms in Bayesian network learning and inference tasks," *Inf. Sci. (Ny).*, vol. 233, pp. 109–125, 2013.

[118] L. Liu, "Survey of contemporary Bayesian Network Structure Learning methods," 2015.

[119] Y. Zhou, "Structure Learning of Probabilistic Graphical Models : A Comprehensive Survey," *arXiv: 1111.6925*, 2007.

[120] 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.

[121] 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.

[122] 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.," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pp. 3807–3810, 2008.

[123] 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.

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

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

[126] 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.

[127] G. Heinrich, “Parameter estimation for text analysis,” *Univ. Leipzig, Tech. Rep*, 2008.

[128] S. Purcell, “Maximum Likelihood Estimation.” [Online]. Available: http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html. [Accessed: 25-Oct-2016].

[129] C. Walck and P. P. Group, “Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists,” *Hand-b. Stat. Distrib. Exp.*, no. September, pp. 26–35, 2007.

[130] R. Levy, “Probabilistic models in the study of language,” *Online Draft. Nov*, 2012.

[131] 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.

[132] Rs. Team, “RStudio: Integrated Development Environment for R.” RStudio, Inc., Boston, MA, 2015.

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

[134] C. Elkan, “Evaluating Classifiers,” *Univ. San Diego, California, retrieved [01-11-2012] from http//cseweb. ucsd. edu/∼ elkan B*, vol. 250, pp. 1–11, 2012.

[135] J. D. Kelleher, B. Mac Namee, and A. D’Arcy, “Fundamentals of Machine Learning for Predictive Data Analytics.” Mit Pr, 2015.

[136] H. E. Wheeler and S. K. Kim, “Genetics and genomics of human ageing.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 366, no. 1561, pp. 43–50, 2011.

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

[138] T. B. Kirkwood, “The origins of human ageing.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 352, no. 1363, pp. 1765–72, 1997.

[139] P. Rashidi and A. Mihailidis, “A survey on ambient-assisted living tools for older adults,” *IEEE J. Biomed. Heal. Informatics*, vol. 17, no. 3, pp. 579–590, 2013.

[140] 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 Trans. Comput. Biol. Bioinforma.*, vol. 12, no. 2, pp. 262–275, 2015.

[141] J. P. de Magalhães, A. Budovsky, G. Lehmann, J. Costa, Y. Li, V. Fraifeld, and G. M. Church, "The Human Ageing Genomic Resources: Online databases and tools for biogerontologists," *Aging Cell*, vol. 8, no. 1. pp. 65–72, 2009.

[142] 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.

[143] 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.

[144] R. D. Wood, M. Mitchell, J. Sgouros, and T. Lindahl, "Human DNA repair genes.," *Science*, vol. 291, no. 5507, pp. 1284–9, 2001.

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

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

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

[148] "UniProt." [Online]. Available: http://www.uniprot.org/. [Accessed: 08-Dec-2016].

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

[150] M. Hsing, K. G. Byler, and A. Cherkasov, "The use of Gene Ontology terms for predicting highly-connected'hub'nodes in protein-protein interaction networks," *BMC Syst. Biol.*, vol. 2, no. 1, p. 1, 2008.

[151] S. A. Bakar, J. Taheri, and A. Y. Zomaya, "Identifying hub proteins and their

essentiality from protein-protein interaction network," in *Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on*, 2011, pp. 266–269.

[152] J. J. Hublin, "The origin of Neandertals," *Proc. Natl. Acad. Sci.*, vol. 106, no. 38, pp. 16022–16027, 2009.

[153] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, "IntAct: an open source molecular interaction database.," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D452-5, 2004.

[154] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D262-6, 2004.

[155] M. Scutari, M. M. Scutari, and H.-P. MMPC, "Package 'bnlearn,'" 2016.