Deep Learning Black Box Problem
Master Thesis Research Project

Jabbar Hussain

Subject: Information Systems
Corresponds to: 30 hp
Presented: VT 2019
Supervisor: Prof. Dr. Andreas Hamfelt
Examiner: Prof. Dr. Pär Ågerfalk

Department of Informatics and Media
Abstract

Application of neural networks in deep learning is rapidly growing due to their ability to outperform other machine learning algorithms in different kinds of problems. But one big disadvantage of deep neural networks is its internal logic to achieve the desired output or result that is un-understandable and unexplainable. This behavior of the deep neural network is known as “black box”.

This leads to the following questions: how prevalent is the black box problem in the research literature during a specific period of time? The black box problems are usually addressed by so-called rule extraction. The second research question is: what rule extracting methods have been proposed to solve such kind of problems?

To answer the research questions, a systematic literature review was conducted for data collection related to topics, the black box, and the rule extraction. The printed and online articles published in higher ranks journals and conference proceedings were selected to investigate and answer the research questions. The analysis unit was a set of journals and conference proceedings articles related to the topics, the black box, and the rule extraction.

The results conclude that there has been gradually increasing interest in the black box problems with the passage of time mainly because of new technological development. The thesis also provides an overview of different methodological approaches used for rule extraction methods.

Keywords:
Deep Learning, Artificial Neural Network, Black Box, Inductive Logic Programming, Knowledge Discovery, First-Order Logic, Machine Learning and a list of keywords are given in Appendix A.
Acknowledgment

I wish to acknowledge the efforts of my supervisor Prof. Dr. Andreas Hamfelt for his contribution, guidance, and support in conducting the research work and to prepare the report. I would also like to thank all of my teachers at Uppsala University for their encouragement and continuous support during my studies. I am also very thankful to my parents, family members, and friends for their support.
Table of Contents

1 Introduction: ............................................................................................................................ 8
  1.1 Problem Description: .............................................................................................................. 8
  1.2 Deep learning .......................................................................................................................... 9
  1.3 Problem Area (the black-box) ................................................................................................. 10
  1.4 Research Questions ............................................................................................................... 10
  1.5 Literature review: .................................................................................................................... 11
    1.5.1 Decompositional Techniques: ........................................................................................ 12
    1.5.2 Pedagogical Techniques: .............................................................................................. 12
    1.5.3 Eclectic Techniques: ...................................................................................................... 13
  1.6 Methodology .......................................................................................................................... 13
  1.7 Delimitation ........................................................................................................................... 14
  1.8 Ethical and social consideration ........................................................................................... 14

Background and Theory .............................................................................................................. 15

2 Artificial Neural Networks ................................................................................................... 15
  2.1 Feed Forward Neural Network: ............................................................................................ 15
    2.1.1 Single-layer perceptron: .............................................................................................. 16
    2.1.2 Multi-layer perceptron (MLP): .................................................................................. 16
  2.2 Regulatory Feedback Neural Networks: ............................................................................... 16
  2.3 Radial basis function Neural Network: .............................................................................. 16
  2.4 Recurrent Neural Network (RNN): .................................................................................... 17
  2.5 Kohonen Self Organizing Neural Networks: ....................................................................... 17
  2.6 Modular Neural Network: .................................................................................................... 18
  2.7 Convolution Neural Network (CNN, or ConvNet): ............................................................. 18

3 Neural Networks Learning: ................................................................................................. 20
  3.1 Error-Correction Learning: ................................................................................................. 20
  3.2 Hebbian Learning: ............................................................................................................... 21
  3.3 Competitive Learning: ........................................................................................................ 22
  3.4 Boltzmann Learning: .......................................................................................................... 22
  3.5 Supervised Learning: .......................................................................................................... 22
  3.6 Unsupervised Learning: ...................................................................................................... 23
  3.7 Reinforcement Learning: .................................................................................................... 23
4 Deep Learning and Deep Neural Networks:

4.1 Deep Belief Network (DBN):

4.2 Deep Belief Network Based on Dynamic Supervised Learning:

4.3 Restricted Boltzmann Machine (RBM):

4.4 Knowledge-Based Artificial Neural Network (KBANN):

5 Methodology:

Search Criteria:

6 Results, discussion and Analysis, Research Questions-Answers:

6.1 Approaches Related to Black Box and Rule Extraction Methods:

6.2 Models related to classification and prediction:

6.3 Discussions and Analysis:

6.4 Research Questions and Answers:

7 Conclusion:

References:

Appendix A - Abbreviations:
List of Figures:

Figure 2.1 A single-layer perceptron ............................................................... 16
Figure 2.2 A multilayer perceptron ................................................................. 16
Figure 2.3 A radial basis neural network ......................................................... 17
Figure 2.4 A recurrent neural network ............................................................ 17
Figure 2.5 The structure of SOM ................................................................. 18
Figure 2.6 A block diagram of MNN ............................................................. 18
Figure 2.7 A diagram of CNN ................................................................. 19
Figure 3.1 A taxonomy of learning process .................................................. 20
Figure 3.2 Error-Correction Learning .......................................................... 21
Figure 3.3 The structure of supervised learning ........................................... 23
Figure 3.4 A block diagram of unsupervised learning ................................. 23
Figure 3.5 The structure of reinforcement learning ...................................... 24
Figure 4.1 Deep Neural Network ............................................................... 25
Figure 4.2 A structure of a deep belief network based on dynamic supervision ........................................ 26
Figure 4.3 The structure of a restricted Boltzmann machine (RBM) .................. 27
Figure 6.1 The number of articles published about deep learning black box problems from 2000 to 5/3/2018 ................................................................. 52
List of Tables:

Table 5.1 Search Criteria and the first Journals and conference proceedings database .................. 32
Table 5.2 The third refined list of articles ......................................................................................... 33
Table 5.3 The “Scientific Journal Rankings (SJR) - SCImago” list ..................................................... 34
Table 5.4 Shows the number of articles with respect to the journal’s ranking .................................. 36
Table 6.1 Summary of the neural networks and rule extraction algorithms models/approaches used for knowledge extraction: ........................................................................................................ 40
Table 6.2 Summary of the models/approaches used for classification: ........................................... 45
Table 6.3 Summary of the models/approaches used for prediction: ................................................ 46
Table 6.4 Summary of the models/approaches used for classification and prediction both together in one system: ........................................................................................................................................ 47
Deep learning black box problem

This chapter presents the introduction that includes: problems description, deep learning, problem area-the black box, research questions, literature review, methodology, delimitation, and ethical and social consideration.

1 Introduction:

This thesis examines previous research about Deep Neural Networks (DNNs) and Rule Extraction Methods. It is necessary to understand these topics in order to be able to develop methods for synthesizing rules (a symbolic representation) of the knowledge hidden in neural networks.

The research presented herein is preparatory to and part of a larger research effort to apply Meta interpretative inductive logic programming to extract rules from Deep Neural Networks (DNNs).

Data mining is called “knowledge discovery” when it refers to databases. Knowledge discovery implies the process of automated extraction of unknown, hidden and fruitful information from large databases by using the tools from various fields such as statistics, machine learning, artificial intelligence, database management, etc.

1.1 Problem Description:

On May 6, 2010 at 14:32, a flash crash occurred in security prices on the stock exchange in the United States of America. As a result, in few minutes, stock market indices of three major stocks included S&P 500, Nasdaq Composite, and DJIA crashed down and went up again. The effect registered variations of around 1000 points [1] that resulted in a loss of trillions of dollars in just a few minutes. The investigators reported a technical problem in the computers responsible for carrying out stock market transactions. These computers couldn’t work as the developers had expected. After 13 minutes, at 14:45, an automatic security mechanism activated and it stopped the process.

The exchanges representatives and regulators hold a meeting after the session to reverse the process as much as possible and to help the victims by canceling the transactions those had happened at prices too distant to those before the crisis.

To understand the problem that had occurred, and to prevent it from happening again, the engineers who were working at ATSs (Automatic Trading Systems) found some problems in the memory registers:

“The computers had applied the action rules they were programmed with, but they had done it in an unforeseen context in which these rules, which usually produce desired results, were catastrophic” [2].

The investigators also mentioned that the problem occurred, was already known as qualification problem [3] in the field of artificial intelligence. So, it was not a new problem. The situation could become more complicated, if the engineers had not been able to understand both the problem and the abnormal behavior of computers.
These kinds of machine models exist and are known as black-box models in the field of artificial intelligence. These machines return outputs, or make decisions, by running internal processes, that is un-understandable by human beings. These black-box models are widely used due to their high degree of achievements.

Another example of black box problem is: autonomous cars make decisions by means of neural networks. The liability presupposes that the decision system has been properly validated. As an example, it does not have rules like “if blue ahead then free passage” that do not discriminate between the sky and a blue truck. This has already caused a fatal accident for test driver.

Similarly, to decide the creditworthiness by deep learning systems those are black boxes. The decision must be explained to avoid liability for discrimination etc.

According to the U.S. Defense Advanced Research Projects Agency (DARPA):

“There is an inherent tension between machine learning performance (predictive accuracy) and explainability; often the highest performing methods (e.g., deep learning) are the least explainable, and the most explainable (e.g., decision trees) are less accurate” [4].

In the modern computerized world, in which the economy, stock market, defense, communications, energy, transportation, food, etc, depend on artificial intelligence. Sometimes, it is difficult to trust and to use the artificial intelligence models that are not very understandable. So, to create the reliable solutions, it is very important to open and uncover the black box behavior of the deep neural networks.

1.2 Deep learning

A feed-forward neural network is said to be deep if it consists of more than one hidden layer. To solve such complex problems, researchers are trying to explore and develop the representation of artificial neural networks (ANNs) that are understandable for human beings. Recent research explores different approaches that can shed a light on the internal connections and relationships present within the networks.

The strategy of using deep neural networks to tackle complex problems is made possible by deep learning. Traditional machine learning is composed of one input layer, one hidden layer, and one output layer; these are shallow networks. More than one hidden layer qualifies a network as a deep learning network.

Deep learning is also known as “deep machine learning, deep structured learning or hierarchical learning. It comes under the domain of machine learning through artificial neural networks (ANNs) [5]”. Neural networks are used in classification, prediction, and pattern recognition in datasets. They are also used to infer knowledge from them in the form of rules. It is often hard to understand the outcomes of the rules constructed by the neural networks. A significant amount of work has been done on the development of rule extraction algorithms in order to uncover the black box behavior of the networks. More details are given under the section “Deep Learning and Deep Neural Networks” at page 25.
1.3 Problem Area (the black-box)

Different neural network models are used for the problem related to classification, prediction, fault detection, analysis, supervision, estimation of immeasurable variables, etc. The use of artificial neural networks (ANNs) is very popular, especially with the problems related to the classification and prediction of data [35]. The predictive accuracy attained by ANNs is often higher than that of human experts or other related methods.

These ANNs models can be simple neural network models with one hidden layer, such as shallow neural network or complex structured neural networks with more than one hidden layer, those are known as deep neural works. More than one hidden layer qualifies a network in the domain of deep neural networks [8]. As mentioned above the predictive accuracy of artificial neural network is very high but generally, the network’s internal logic is unexplainable, and incomprehensible due to the complex architectures of artificial neural networks [37]. This behavior of deep neural networks is therefore known as the “Black Box problem”. The black box behavior of deep neural networks is the basic obstacle for the artificial neural networks that restricts their usage. To open the black box behavior and convert it into the white box is a very challenging and complex task.

1.4 Research Questions

The following examples have been discussed under the topic “1.1 Problem Description: ”.

a) A flash crash occurred in security prices on the stock exchange in the United States of America.

b) Autonomous cars make decisions by means of neural networks. The system does not discriminate between the sky and a blue truck. That has already caused a fatal accident for test driver.

c) To decide the creditworthiness by deep learning systems those are black boxes. The decision must be explained to avoid liability for discrimination etc.

In all the above three examples, deep neural networks were used. In-fact, sometimes, the system slightly improves its class boundaries and learns automatically in routine tasks during the processing in DNNs, that affects the final results of the system after a certain period of time. Because the researchers are unable to explain the internal logic of the system, so such problems are called the Black Box problems.

To prevent from mishaps and big loses as discussed at page 8-9 under the topic “1.1 Problem Description: ”. It is required to open and understand the internal logic of the decision made by deep neural networks.

As, in the modern computerized world, in which the economy, stock market, defense, energy, food, communications, transportation, etc depend on artificial intelligence. Sometimes, it is difficult to trust and to use the artificial intelligence models that are not very understandable. So, to create the reliable solutions, it is very important to open and uncover the black box behavior of the deep neural networks.
This leads to investigate first: how prevalent (mean frequent, usual, common, general, universal) are the black box problems in a specific period of time? The black box problems are usually addressed by so-called rule extraction. It also leads to investigate second: what rule extracting methods have been proposed to solve such kind of problems in a specific period of time?

So, finally the study aims to investigate the following research questions in existing research publications. Two research questions are:

1) How prevalent is the black box problem in the research literature during the period 2000 to 2018?
2) What rule extracting methods have been proposed in this regard from 2010-2018?

A question raised here about the time frames for the research questions, especially the time frames 2000-2018 for the first and 2010-2018 for the second research question respectively.

To write a master thesis of 30 Cr, there is a limited time frame of one semester (4-5 months). That is why, I limited my research work from 2000-2018 for the first research question. The second research question was more complicated and requires more time to investigate. It was a huge workload to search all the rule extracting methods with their supporting methods those have been proposed from 2000-2018, in a limited time frame of one semester. So, I also limited the time frame from 2010-2018 for the second research question. These were the main reasons to set the limited time frames for both research questions.

Both research questions are related to deep neural networks and rule extraction methods. As the black box problems are usually addressed by so-called rule extraction. The answer to the second research question provides an overview of different methodological approaches used for rule extraction methods. Several rule extraction methods have been used for different kinds of problems in various fields. The usage of these rule extraction methods have some advantages and disadvantages that depend upon the type of the problem. The researchers who want to work in the area of black box and rule extraction methods, studying these questions may contribute an adequate knowledge relevant for synthesizing rules of the knowledge hidden in neural networks.

1.5 Literature review:

Everyday large amounts of data are generated on the internet, banks, retail stores, and insurance companies. It can be structured, unstructured and semi-structured form of data. Researchers used different approaches for knowledge extraction in the form of rules from these datasets. Neural networks attained the highest accuracy in some of the problems related to classification and prediction. Sometimes, the meaning of rules generated by such networks is difficult to understand and therefore often known as black box problem. To open this black box problem and convert it into the white box, a significant amount of work have been devoted to the development of rule extraction algorithms to understand the behavior of the networks. It is done by translating the internal knowledge of the neural network into a set of symbolic rules [8]. Rule extraction methods are often based on sample learning with different techniques such as neural networks, decision trees, and genetic algorithms. These techniques have some advantages and disadvantages as well. For example, decision trees have difficulties to mine large datasets. Standard genetic algorithms
have difficulties to express relationships of complementing and competing among rules. Neural networks perform well in noisy data and have the ability to classify the patterns. These approaches extract the rules either in the form of the conjunctive normal form (if and then) or in the form of subset selection form (M of N).

Researchers have proposed several rule extraction algorithms. The aims of the rule extraction algorithms are:

a). To generate symbolic representations, understandable by the domain experts.

b). To produce symbolic representations, able to form the networks again from which they are extracted.

c). Requiring neither special training methods nor restrictions on network structure.

The rule extraction algorithms are based on various rule extraction techniques. Some of them are known as decompositional, pedagogical, and eclectic techniques [9].

1.5.1 Decompositional Techniques:

Decomposition algorithms divide the network into the neuron level. The aggregated sum of the result of each neuron represents the network as a whole [9]. By analysis of the activation and weights of the hidden layers of the networks, these techniques extract the symbolic rules. Some examples of decompositional techniques are given below.

Sang Park et al. [10] proposed a rule extraction method to extract rules by interpreting the strengths of connection weights directly in a trained network. This is a generalized analytic rule extraction method for feedforward neural networks. Neuro Linear [11] is a decompositional technique used to extract oblique classification rules from single hidden layer neural networks. Kim and Lee [12] proposed an algorithm that applies to multilayer perceptrons networks with two hidden layers. The algorithm is based on feature extraction and feature combination. FERNN [13] is a decompositional algorithm for feedforward neural networks. It is used for classification rule extraction which returns oblique rules and these rules can be further simplified to disjunctive normal form DNF-rules or M-of-N (subset selection form) rules under certain conditions. Another decompositional approach called Tsukimotosmethod [14] used to extract low order rules from each node in the network. A Recursive Rule Extraction algorithm (Re-RX) was proposed by Setiono et al., [15]. That is used to generate hierarchical classification rules from datasets having both discrete and continuous attributes.

1.5.2 Pedagogical Techniques:

Pedagogical Techniques focus toward the finding of an output for a corresponding input by mapping the closest possible relationship of input-output of the network [9]. The internal structure of the weights of ANN is not subjected to analysis. This approach contains various techniques, such as sampling approach, validity interval analysis, and reverse engineering of the artificial neural network. Some of the examples of pedagogical techniques are given below:
Saito and Nakano [16] proposed a medical diagnosis expert system to extract rules by changing the inputs and then observing the effect on the output of networks. It is based on a multilayer artificial neural network. Hayward and Diederich [17] have proposed a method known as RULENEG to extract rules by stepwise negation, from the trained artificial neural networks. It is based on binary inputs. It extracts rules both in conjunctions and disjunction form. Craven and Shavlik proposed TREPAN [18] method that uses sampling and queries to extract rules in the form of decision tree. Filer et al. proposed a method called Interval Analysis (IA) to extract rules in the form of M-of-N [19]. Augusta and Kathirvalavakumar [20] proposed a method called RxREN to extract classification rules using a pedagogical approach from the trained neural networks. The method uses a reverse engineering technique to remove and discover the insignificant and significant input neuron respectively. It also uncovers the technological principles of each significant input neuron of ANN.

1.5.3 Eclectic Techniques:

Eclectic Techniques are the combination of both, decompositional and the pedagogical techniques [9]. Some of the examples of eclectic techniques are given below:

Tickle et al. proposed an eclectic technique called DEDEC [21] that used to extract a set of rules from the set of individual cases. The proposed technique identifies the minimal set of information to differentiate an object from other objects for rule extraction. Keedwell et al. proposed an eclectic technique [22] that uses a genetic algorithm for rule extraction from ANN. Kaikhah and Doddameti proposed an eclectic technique [23] that used neural networks for trends discovery in the large datasets. That monitors the weights for clustering and pruning the activation values of the hidden neurons. It also used the control parameters for controlling occurrences probability and rules accuracy for data analysis. Hruschka and Ebecken proposed an eclectic technique called Rex-CGA (Clustering Genetic Algorithm) [24] that extracts rules from multilayer perceptrons. This technique used CGA to discover clusters of hidden unit activation values and produced logical rules from these clusters. Kahramanli and Allahverdi proposed an eclectic technique that used artificial immune systems for rules extraction from trained adaptive neural networks [25]. The decompositional process of the technique is hard that gave complex results and large descriptions. So, this technique has a drawback of computation limitations and time. While the pedagogical processes of the technique are faster but are unable to capture all of the valid rules accurately. Those rules describe the behavior of the networks. The eclectic technique can be slower but more accurate as compared to a pedagogical approach.

1.6 Methodology

Briony J Oates (2006) [26, pp 93], discussed different kinds of survey methods, but for the above-mentioned research questions, the systematic literature survey is supposed to be the right one. Yair Levy and Timothy J. Ellis (2007) [28], and B. Kitchenham, S Charters (2007) [29], presented the guidelines for systematic literature reviews in their articles. Lochan Winton, Ruth Alicia (2017) [30], did a systematic literature review in her research project “Technology and the Internet as a key component in psychosocial care of Somatic diseases”.

A literature review was conducted for data collection related to topics, the black box, and the rule extraction. The printed and online published documents related to black box problem and rule extraction methods were investigated. The analysis unit was a set of journals and conference proceedings articles on the required topics. Those articles published in journals and conference proceedings with higher ranks were selected for the investigation to answer the research questions.

For data collection related to black box topic, two more subtopics were included in the search criteria, those are neural networks, and deep learning, because they are indirectly involved with the black box problem. For data collection related to rule extracting methods, three more subtopics were included in the search criteria, those are, first-order theory, inductive logic programming, and relation knowledge because they are also indirectly involved with the rule extracting methods. For more detail, see details in chapter 5-Methodology page 29.

1.7 Delimitation

Different topics of interest and various research approaches have been used in black box problem. I cannot claim that the analysis of published papers is exhaustive, as I have limited my search to the Thomson Reuters Web of Knowledge (WOK) databases of journals and conference proceedings articles. However, I expect these studies to contribute an adequate coverage of the work that led to the development of methods for synthesizing rules of the knowledge hidden in neural networks.

1.8 Ethical and social consideration

It is not needed to write about the ethical and social consequences of the results because this is a literature study about black box problem and rule extraction methods. Ethically and socially, it will not put any effect to anyone.
Background and Theory  
*This chapter presents artificial neural networks, and types of artificial neural networks.*

2 Artificial Neural Networks

Artificial neural networks are roughly equal to the biological neural networks. Different scientists have defined it in different ways. Perhaps, it was Laurene Fausett who put forth a precise and comprehensive definition of an artificial neural network. According to him:

"An artificial neural network is an information processing system that has certain performance characteristics in common with biological neural networks. Artificial neural networks have been developed as a generalization of mathematical models of human cognition or neural biology"[31].

The artificial neural network is characterized by its architecture (neuron connection pattern), training and learning methods (methods for weights determination), and the activation function.

History:

The first neural network was designed by McCulloch and Pits in 1943 and found that a combination of simple neurons in the neural system increases the computational power. In 1949, the first learning law of artificial neural network was designed by Donald Hebb, a psychologist at McGill University. Rosenblatt (1958,1962), Block (1962), Minsky and Papert (1969), introduced a lot of neural networks and Rosenblatt was the inventor of “Perceptron”. Bernard Widrow and Marcian (1960) introduced a learning rule which is closely related to the perceptron learning rule.

Teuvo Kohonen (1972), dealt with associative memory neural network and in 1982, he developed the self-organizing map. James Anderson of Brown University also dealt with associative memory. In the 1970s, Grossberg and Carpenter together Developed Adaptive Resonance Theory ART, a theory of self-organizing neural networks. In 1974, Backpropagation had been discovered by Werbos, but couldn’t get wide publicity. Before it became widely used, a similar method in optimal control theory was also discovered independently by David Parker (1985) and Le Cun (1986). In the 1980s John Hopfield, from California Institute of Technology and David Tank, a researcher at AT&T, developed neural networks with fixed weights and adaptive activations known as Hopfield Nets.

Types of Artificial Neural Networks:

Artificial neural networks consist of neurons, which semantically communicate among each other. The way they communicate among each other is an area of ongoing research nowadays. There are several types of artificial neural networks. Some of them are given below.

2.1 Feed Forward Neural Network:

These are the simplest type of networks. They are among the first in the history of artificial neural networks. In feedforward networks, the information is unidirectional, which move from the input layer to the output layer through hidden layers with no cycle or loops in the network [27]. The simplest form of feedforward neural networks is perceptron. These neural networks consist of binary units such as McCulloch-Pitts neurons.


2.1.1 Single-layer perceptron:

Single-layer perceptrons are the simplest form of feedforward neural network [27]. They have one input layer and one output layer of processing units with no feedback connections. The inputs are fed directly to the output through a series of weights. Each node in the output layer calculates the sum of the products of weights and inputs. The output neuron fires if the value of the calculated sum is higher than the threshold, otherwise it will not fire. The activation value can be “1” in case of fire and deactivation value can be “0” when it does not fire. Figure 2.1 shows a single-layer perceptron.

![Figure 2.1 A single-layer perceptron.](image)

2.1.2 Multi-layer perceptron (MLP):

They have multiple layers of neurons, connected with each other in a feedforward way [27]. They have one input layer, one output layer and one or more hidden layers of processing units with no feedback connections. Multilayer perceptrons use the sigmoid function as an activation function. Some examples of MLP networks are auto encoder, probabilistic, time delay, and convolution neural networks. Figure 2.2 shows a multilayer perceptron.

![Figure 2.2 A multilayer perceptron.](image)

2.2 Regulatory Feedback Neural Networks:

Regulatory feedback neural network facilitates negative feedback by performing inference [27]. The feedback is not used for learning or training purposes but to find the optimal activation of nodes. The approach is similar to the nonparametric method. These neural networks are best where mathematically equivalent classification is required.

2.3 Radial basis function Neural Network:

These networks consist of three layers, input layer for input vector (features), hidden layer consists of radial basis functions (RBFs), and the output layer is a linear combination of the outcome of the radial basis functions [27]. Radial basis functions calculate the distance of points with respect to a central point and act as an activation function. Figure 2.3 shows an example of the radial basis neural network.
2.4 Recurrent Neural Network (RNN):

Recurrent neural networks are based on the principle “save the outcome and feedback to input [32]”. In RNN, the first layer acts similar to the feed forward neural networks, with a weighted sum of the product of weights and input vector. Once computed, the recurrent network process starts. During this step, each neuron saves ex-step information like a memory cell for later use. If the prediction is incorrect, the learning rate and error correction are used to make changes during backpropagation, to achieve the right prediction. This is the basic concept in RNN. Figure 2.4 shows a recurrent neural network.

2.5 Kohonen Self Organizing Neural Networks:

The name of Kohonen Self Organizing neural network came from the algorithm called self-organizing Kohonen’s map, developed by Teuvo Kohonen in the 1970s [32]. The main inspiration for the self-organizing map was speech recognition; especially the auditory nerve in the ear was a model for the self-organizing map (SOM). These are the feedforward neural networks and suitable for higher dimensional input reduction to two-dimensional response. SOM is an unsupervised learning approach, in which the neurons work competitively. Each neuron computes a weighted sum of the input vector and in the next step neurons enter in a competitive phase. Figure 2.5 shows the structure of the SOM.
Figure 2.5 The structure of SOM.

2.6 Modular Neural Network:

Modular Neural Networks consists of different independent networks [32]. Each network has its own contribution to the final output without interfering with the operations of other’s networks. Each network has a unique set of inputs. The MNNs are suitable for complex tasks and the complexity decreases with the breakdown of large tasks into smaller processes. So, the number of connections decreases and the computational speed increase due to no interference among each other. MNN is the hot and rapidly growing field in ANN research. Figure 2.6 shows a block diagram of MNN.

Figure 2.6 A block diagram of MNN

2.7 Convolution Neural Network (CNN, or ConvNet):

These networks are similar to feedforward neural networks. These networks have learnable weights and biases [32]. They are widely used in image and signal processing in the field of computer vision and machine learning. Each neuron gets some inputs, conducts a dot product and optionally follows it with a nonlinearity layer. The whole network shows a single differentiable score function. The architectural design of CNN includes an input layer, an output layer and between them hidden layers. Some of them are convolution layers, fully connected layers, pooling layers, nonlinearity
layers, rectified layers, rectified linear units, and normalization layers, etc. The most common layers are:

**Convolution layer:** Convolution layers perform the convolution operation to the input, and then pass the result to the next layer. It processes the data for its respective field and imitates the individual neuron’s response to visual stimuli.

**Pooling layer:** These layers reduce the size of free parameters and permit the network to go deep by using few parameters in image processing. Convolution networks may have local or global pooling layers, which merge the neuron’s clusters output into a single neuron in the next layer.

**Fully connected layer:** Fully connected layers have the same principle as MLP (Multi-Layer Perceptrons), in which every neuron in one layer is connected to every neuron in another layer. Figure 2.7 shows a diagram of CNN.

![Figure 2.7 A diagram of CNN.](image)
3 Neural Networks Learning:

This chapter presents different types of neural networks learning techniques.

The definition of learning adapted from Mendel and McClaren in the context of the neural network is “Learning is a process by which the free parameters of a neural network are adapted through a process of stimulation by the environment in which the network is embedded. The learning type is determined by the manner in which changes in the parameters take place” [32, pp 50].

“Learning algorithm is a prescribed set of well-defined rules for the solutions of a learning problem” [32, pp 50]. There is no single or unique learning algorithm. There is a kit of tools represented by different learning algorithms where each has different advantages. The four basic rules for learning algorithms are Error-correction learning, Hebbian learning, Competitive learning, and Boltzmann learning. The three basic classes of learning paradigms are supervised learning, reinforcement learning, and self-organized (unsupervised) learning. Figure 3.1 shows the taxonomy of the learning process [32, pp 47].

![Figure 3.1 A taxonomy of learning process.](image)

3.1 Error-Correction Learning:

Error-Correction Learning is also known as delta rule learning or widrow-Hoff rule learning. It is used with a supervised learning approach. In this technique, the system output value is compared with the desired output value, and the error between them is used to direct the system training. Each neuron in the system is connected to all other neurons in the surrounding layers. So, there may be different routes that depend upon the number of neurons in the layers. The error values can be used in the most direct route for directly weight adjustment. Mostly, Error-correction learning uses the algorithm known as “back propagation algorithm”. During training iterations, error-correction learning algorithms are used to minimize the error signal. Mathematically, the error signal $e_k(n)$ can be written as [32, pp 47].

$$e_k(n) = d_k(n) - y_k(n)$$

Here $d_k(n)$ denotes the target output, and $y_k(n)$ denotes the output signal of neuron $k$. Here $k$ is a node in the output layer at a time $n$ and neuron $k$ is driven by a signal vector $x(n)$, produced by one or more layers of neurons. Figure 3.2 shows Error-Correction Learning.
3.2 Hebbian Learning:

Hebbian learning is one of the most famous and the oldest of all the learning rules. According to this rule, when an axon of cell A is near enough to excite a cell B, some growth process or metabolic changes take place in one or both cells. Hebb proposed this change as a basis of associative learning which would result in an enduring modification in the activity pattern of a spatially distributed assembly of nerve cells [32, pp 49].

To understand this rule, it can be rephrased as a two-part rule as follows [32, pp 50].

1) If both neurons are activated simultaneously on either side of a synapse, then synapse strength is increased selectively.

2) If both neurons on either side of a synapse, are activated asynchronously, then that synapse is weakened or eliminated selectively.

Such a synapse is known as a Hebbian synapse. The modifications in Hebbian learning have a local and time-dependent mechanism. The Hebbian synapse depends on presynaptic and postsynaptic signal’s exact time of occurrence. The Hebbian synapse produces an input-specific local synaptic modification as the information bearing signals are in spatiotemporal contiguity. The change in a synapse is always with respect to the signals on both sides of the Hebbian synapse and the learning is dependent on the “true interaction” between presynaptic and postsynaptic signals. The interaction or dependence can be naturally statistical or deterministic. In synaptic efficiency, the condition for a change is the conjunction of both presynaptic and postsynaptic signals, and the co-occurrence of their activities is sufficient to produce synaptic modification within a short interval of time. Therefore it is referred to as a conjunctional synapse. Mathematically, change in the adjustment applied to the synaptic weight “\( w_{kj} \)” at time step “\( n \)” can be expressed as

\[
\Delta w_{kj}(n) = F(y_k(n), x_j(n))
\]

The function \( F(y_k(n), x_j(n)) \) denotes both postsynaptic and presynaptic signals. The \( x_j(n) \) and \( y_k(n) \) signals are mostly dimensionless. Here \( x_j \) and \( y_k \) denote presynaptic and postsynaptic signal respectively. The \( w_{kj} \) is the synaptic weight of the neuron.
3.3 Competitive Learning:

During competitive learning a competition occurs among the output neurons of the network to become an active (fired) neuron. At any one time, only a single output neuron is active in competitive learning while several output neurons may be active at the same time in Hebbian learning. This makes competitive learning attractive for input patterns classification. In competitive learning, there are three basic rules [32, pp 53].

a. All neurons are the same except for some with randomly distributed synaptic weights and their response is different for input patterns.
b. Each neuron’s strength is limit imposed.
c. A mechanism allows the neurons to compete for the right in response to a given subset of inputs and only one neuron per group is active at a time. The competition winner neuron is called winner-takes-all neuron. Similarly, the individual network neurons are called feature detectors, which are those who learn to specialize onset of a similar pattern [33, pp 53].

3.4 Boltzmann Learning:

The name of this learning rule is for the honor of Ludwig Boltzmann. It is a stochastic learning algorithm derived from an idea that has root to statistical mechanics. In Boltzmann learning the neurons operate in a binary manner and have recurrent structure [32, pp 55]. If the neuron operates as “on”, it is denoted by +1 while in case of “off”, it is denoted by -1. The system is characterized by an energy function denoted by E. The value of the energy function E is determined by a particular state occupied by the individual neurons. Mathematically it is calculated as [32, pp 55].

\[ E = -\frac{1}{2} \sum_{j \neq k} w_{kj} x_k x_j \]

The \( x_j \) and \( x_k \) denote the states of neuron \( j \) and \( k \) respectively, \( w_{kj} \) denotes the synaptic weight connecting from neuron \( j \) to \( k \), and \( j \neq k \) mean, none of the neurons has self-feedback in the machine.

3.5 Supervised Learning:

When the desired output is already known, a neural network is said to learn supervised. In supervised learning, the training set consists of both input and desired output pairs. System performance can be measured in: mean square error, the sum of squared errors over the training set and error surface with free parameters as coordinates [32, pp 57].

During supervised learning, one input pattern is propagated through the network from the input layer to the output layer, which results in an output pattern. The generated output pattern is compared to the target pattern. An error value is computed with respect to the difference between input and output pattern value. This error value is known as learning effort of the network which is
under the supervision of “imaginary supervisor”. If the error value is greater, more weights will be changed [32, pp 57]. Figure 3.3 shows the structure of supervised learning.

![Figure 3.3 The structure of supervised learning](image)

### 3.6 Unsupervised Learning:

In unsupervised learning, the networks have no targets outputs. The results of the learning process cannot be determined that how will it look like? Unsupervised learning works well for purposes such as pattern classification.

“There is no specific function to be learned by the network. Rather, provision is made for a learning task independent measure of the quality of representation that the network is required to learn” [32, pp 65].

During unsupervised learning weights of the network are arranged within a certain range with respect to the input values. The main goal is to group similar units closely together in specific areas of a certain range. There is no imaginary supervisor or teacher in unsupervised learning [32, pp 65]. Figure 3.4 shows a block diagram of unsupervised learning [32, pp 65].

![Figure 3.4 A block diagram of unsupervised learning.](image)

### 3.7 Reinforcement Learning:

In Reinforcement Learning, there is a continuous interaction of “online learning an input-output mapping through a process of trial and error design to maximize the scalar performance index
known as reinforcement signal” [32, pp 59]. There is a “critic” in the learning system that receives primary reinforcement signal from the environment and then converts the received signal into a heuristic reinforcement signal that is a higher quality reinforcement signal [32, pp 59]. Figure 3.5 shows the structure of reinforcement learning.

![Figure 3.5 The structure of reinforcement learning.](image-url)
4 Deep Learning and Deep Neural Networks:

This chapter presents deep learning and deep neural networks, and few examples of DNN.

A feed-forward neural network is said to be deep if it consists of more than one hidden layer. The strategy of using deep neural networks to tackle the complex problems goes through deep learning. Deep neural networks can concisely represent several complex functions as compared to shallow neural networks and support vector machine [34]. It can be shown that a shallow net requires several numbers of neurons to compute some possible functions while deep net can do the same job with a fewer number of neurons [34]. Different problems have different complexity, depth helps in learning and practicing. It has been proved that several real world problems have been solved with deep learning architecture in a more efficient way. The idea is similar to the biological brain that has a hierarchy of layers of neurons connected via billions of connections known as deep architecture. Figure 4.1 shows an example of Deep Neural Network [87].

![Deep Neural Network](image)

Figure 4.1 Deep Neural Network

Deep neural networks are distinguished from single hidden layer neural networks by the number of hidden layers. Deep learning neural networks have more than one hidden layer in which data passes through multistep processes [8].

Deep learning is based on the representation of learning data. It is a part of machine learning also known as hierarchical learning or deep structured learning. It can be supervised, unsupervised or semi-supervised learning [6].

Semi-supervised learning is a class of supervised learning that also uses unlabeled data for training, typically a small amount of labeled data with a large amount of unlabeled data [36].

Neural networks have shown higher degree of achievements in different kinds of problems. But they left behind a very complex problem. It is the incomprehensible internal logic of the hidden layers in deep networks, which makes it a black box problem. That is not easy to understand for a human being.
Deep neural networks are being applied to different kind of problem such as image detection, segmentation, classification, fault detection, speech recognition, industrial controlling, in the aviation industry etc. To achieve the highest level of accuracy and reliability, it is needed to understand and open the black box behavior of the network into a white box. Some examples of deep networks are given below.

4.1 Deep Belief Network (DBN):

In machine learning, a “Deep Belief Network (DBN) is a generative graphical model, or alternatively a class of deep neural network, composed of multiple layers of latent variables (hidden units), with connections between the layers but not between units within each layer” [37]. DBN is looking like a simple composition of, unsupervised networks such as, “Restricted Boltzmann Machines (RBMs)” or autoencoders, which consists of some visible and invisible layers.

Deep belief networks can be greedily trained such as one layer at a time, and known to be one of the first deep learning algorithms. These networks have been used in different real-life applications such as drug discovery and electroencephalography.

4.2 Deep Belief Network Based on Dynamic Supervised Learning:

In DBN networks, the training method is unsupervised for each RBM layer. However, with backpropagation (BP) technique, it is reverse supervised fine-tuned. That improve training speed, without the presence of supervised training between RBM layers. The RBM training results are estimated and backpropagation training is used with supervision, that depends on the estimated results. Figure 4.2 shows the structure of a deep belief network based on dynamic supervision [33].

![Figure 4.2 A structure of a deep belief network based on dynamic supervision.](image)

The evaluation criteria is the average error and its standard deviation. Backpropagation algorithm is used to fine-tune the RBM network parameters if the value of average error or its standard deviation is greater than the given threshold. It improves the recognition performance by reducing the risk of errors propagating in RBM layers. The DBN network training time is limited because all RBM layers do not require to be trained by supervision, which also improves recognition performance of the network.
4.3 Restricted Boltzmann Machine (RBM):

Hinton and Sejnowski introduced a restricted Boltzmann machine (RBM), which is a generative stochastic neural network. The system is composed of some visible and invisible units where both are binary variables.

RBNs have input visible layers, hidden layers and the connections between layers, but not within the layer. In RBNs, each sub-network's hidden layer works as a visible layer for the next layer. The network is fast and gets layer by layer training with unsupervised training method. Figure 4.3 shows the structure of a Restricted Boltzmann Machine network (RBM) [33].

![Figure 4.3 The structure of a restricted Boltzmann machine (RBM).](image)

Here v and h denote visible and hidden layers, w denotes the weight value between visible and hidden layers, and a and b are their offset values, respectively. The network energy function is defined by

\[
E_v(v, h) = -\sum_{i=1}^{n_v} a_i v_i - \sum_{j=1}^{n_h} b_j h_j - \sum_{j=1}^{n_h} \sum_{i=1}^{n_v} h_j w_{ji} v_i
\]

The hidden layer can be obtained using the joint probability distribution equation given below.

\[
P(h_j = 1|v) = \frac{1}{1 + \exp(-b_j - \sum_{i=1}^{n_v} w_{ji} v_i)}
\]

The node value of a visible layer can be obtained from a known node value of a hidden layer using the equation given below.

\[
P(v_i = 1|h) = \frac{1}{1 + \exp(-a_i - \sum_{j=1}^{n_h} w_{ij} h_j)}
\]
The training goal of RBM is to achieve the parameters $\theta$ ($w, a, b$), that maximizes $P(v, h)$ the joint probability. Markov chain Monte Carlo (MCMC) algorithm is used to solve this problem. The node values of the hidden and visible layers are continuously updated using this approach until the Markov chain approach reach to its stable stage that represents the maximum value of the joint probability $P(v, h)$ [38]. Then the derivatives of the maximum and initial joint probability distributions are obtained. The following equation is used to update weight value $\theta$.

$$
\theta^{(t+1)} = \theta^{(t)} + \eta \frac{\partial \log P(v, h)}{\partial \theta}
$$

Here $\eta$ denotes the learning speed and $\tau$ denotes the iteration number. The following equation is used to evaluate the slope of the joint probability distribution.

$$
\eta \frac{\partial \log P(v, h)}{\partial \theta} = < h^o \left( v^o - v^1 \right) > + < v^1 (h^h - h^e) > + \cdots
$$

$$
= < h^e v^1 > - < h^e v^c >
$$

Here $v^0$ is the input vector of the visible layer when $t = 0$. Vector $h^o$ of the hidden layer is obtained by $v^0$. Here $v^1$ is the vector of the visible layer when time $t = 1$. While $v^c$ and $h^c$ are the visual and hidden layers vectors, respectively when time $t = 1$.

Hinton used Contrastive Divergence (CD) guidelines [39] to speed up calculations without compromising the accuracy. To calculate the “difference” between two probability distributions, Kullback-Leibler (KL) distance is used by using the following equation.

$$
CD_n = KL(P_0||P_\infty) - KL(P_n||P_\infty)
$$

Here KL (P||P’) denotes the “difference” of two probability distributions and $CD_n$ is the position measurement of $P_n$ between $P_0$ and $P_\infty$.

### 4.4 Knowledge-Based Artificial Neural Network (KBANN):

Knowledge-Based Artificial Neural Network (KBANN) is composed of phases:

“These phases consist of expression and abstraction of the domain knowledge at neural networks, the training of neural networks and finally, the extraction of rules from trained neural networks [48].”

The KBANN was used to explore the neural network black box and also used to generate symbolic rules with the predictive power similar to the neural network itself.

KBANN consider the contribution of inputs as a group for classification as compared to rule-based algorithms that support individual contribution. For example, C5.0 is a rule-based algorithm that considers the individual contribution of inputs, one at a time as the tree is grown.
5 Methodology:

This chapter presents the methodology of the research work. It includes planning and designing of the survey, and the search criteria.

Planning and Designing of the Survey:

A literature review [26] was conducted where the analysis unit was a set of journals and conference proceedings articles using the Thomson Reuters Web of Knowledge (WOK) databases of journals and conference proceedings articles as at 5th March 2018 on publications related to topics such as Black Box, and Rules Extraction Methods. The literature survey was used in the data generation process, in which online published articles related to the above-mentioned topics with higher ranking journals and conference proceedings were considered. Generally, in the planning and designing of the literature survey, six activities are involved. These activities are: data requirement, data generation method, sampling frame, sampling technique, response rate and non-responses, and sample size.

Establishing data requirements:

The two research questions are:

1. How prevalent is the black box problem in the research literature during the period 2000 to 2018?
2. What rule extraction methods have been proposed in this regard from 2010-2018?

The data I decided to generate for the first research question included:

- Direct topic related data such as black box problem and
- Indirect topic related data such as deep learning, and neural networks.

The data I decided to generate for the second research question included:

- Direct topic related data such as rules extraction methods and
- Indirect topic related data such as relational knowledge, inductive logic programming, and the first order theory.

Data generation methods:

As the research questions were related to the work done in the past relevant to the black box and rule extraction methods. Therefore, I decided to use the document survey. In this survey, some documents were directly retrieved from the existing literature and some were indirectly retrieved.

Sampling frame:

To answer the research questions data were collected related to the topics the black box and the rule extraction during the period 2000 to 2018. The first database was refined and created. In the first list, 5571 articles were retrieved.

Sampling technique:

A probability sampling technique, known as systematic sampling technique was used to find the
second refined list of the articles. In which at least three combinations of the keywords were used to select and ensure the second refined list of those articles which contain both ANN and some rule extraction method used to solve the problem. So, the 2nd refined list contains 86 articles.

Then a third refined list was created by excluding the articles with nonavailability of ranking of their journals or conference proceedings and the articles with the dual presentation. The 3rd refined list consists of 67 articles.

To ensure the high-quality, the fourth and final refined articles list was created which contain the articles with higher ranking journals or conference proceedings. The fourth refined list contains 30 articles. For more detail, see topic 5.1 at page 31.

Response rate and non-responses:
The response rate and non-responses are not required in this research work because it is a literature survey. The response rate and non-responses are often considered in surveys such as questionnaires, and interviews.

Sample size:
There are a lot of databases of the journals and conference proceedings for research articles. Due to the limited time frame for the master research project which is one semester, it was impossible to consult all databases of journals and conference proceedings articles. Therefore, the survey is limited to the thirty articles those were selected from higher ranked journals and conference proceedings registered with Web of Knowledge (WOK) databases.

The research question 1 was answered by using Table 5.2. While the tables: 6.2, 6.3, and 6.4 are the answers to the research question 2. Tables 6.2, and 6.3 show the summary of the approaches used for classification and prediction respectively. Table 6.4 shows the summary of the approaches used for classification and prediction both together in one system such as in a modular approach.

The studies of the summaries in these tables hope to lead to adequate knowledge to develop methods for synthesizing rules of the knowledge hidden in neural networks. There are few questions and answers related to the research work given below.

Question 1: Why is the chosen methodology appropriate for the research questions?
Answer: The research questions were answered from data collected in the literature review using the Thomson Reuters Web of Knowledge (WOK) databases of journals and conference proceedings articles on publications related to the above-mentioned topics.

Question 2: Why did you choose WOK as your journal databases?
Answer: There are a lot of databases of journals and conference proceedings articles existed in the market for research articles. Due to the limited time frame for the master research project which is only one semester, it was impossible to consult all databases of journals and conference proceedings articles. Because, Web of Knowledge (WOK) databases are registered with the Uppsala University library, so it was kind of easy to access the Web of Knowledge (WOK) databases through the university library.
Question 3: As the main research questions are concerned with the topics black box and rule extraction methods, why the other related subtopics were included in the research criteria?

Answer: The subtopics related to black box are neural networks and deep learning. These subtopics are included in the search criteria because they are indirectly involved with the black box problem. The subtopics related to rule extraction methods are first-order theory, inductive logic programming, and relation knowledge. These subtopics are included in the search criteria because they are indirectly involved with the rule extraction methods.

Question 4: Why were the combinations of three keywords used to select the second refined list?

Answer: At least three combinations of the keywords were used to select and ensure the second refined list of those articles which contain both ANN and some rule extraction method used to solve the problem.

Question 5: Why are the articles from both conference proceedings and journals selected? This is confusing the reader. Please clarify.

The conference proceedings and journals are two different venues for publication. In computer science proceedings for high ranked conferences are meticulously peer reviewed just like articles in (high ranked) journals, but the latter go through several rounds of reviewing before finally getting accepted. In many other scientific disciplines conference proceedings are not even peer reviewed, peer reviewing is reserved for journal articles. It is very common in computer science to publish in conference proceedings because of the faster publishing cycle, however, in other fields, it might be the gathering of the people to talk abstracts about ongoing research.

In computer science, the acceptance rates of conferences papers are around 10% in high ranked conferences. The A+ conference papers published in high ranked conferences, are held in high regard within the computer science community. It is far more competitive as compared to many of the best journals. In basic sciences, journal publications are held in high regard than the proceedings, however, the conference proceedings are even more important in some other majors such as in computer science area.

The publications I referred to have high rank, and gone through proper peer reviewing, regardless of whether they are from conference proceedings or journals, they were selected from the right track.

Question 6: What do you expect to gain from the literature study?

Answer: This literature study may contribute an adequate knowledge relevant for synthesizing rules of the knowledge hidden in neural networks.

Search Criteria:

I took into account the different approaches used to solve different problems in the most prominent published articles in well-known journals and conference proceedings. A collection of articles were retrieved by using the different combinations of the keywords “rules extraction”, “deep neural networks”, “black box”, “inductive logic programming”, “neural networks”, “relational
knowledge”, and “first-order theory”. Table 5.1 shows the search criteria and the first refined list of the articles.

**Subject** = Artificial Intelligence

**Keywords**= Deep learning, Neural Networks, Relation Knowledge, Rules Extraction, Inductive Logic Programming, Black Box and First Order Theory.

Table 5.1 Search Criteria and the first Journals and conference proceedings database.

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Search Criteria</th>
<th>Results</th>
</tr>
</thead>
</table>
| 1      | TS=(Rule Extraction)  
| 2      | TS=(Inductive Logic programming) AND TS=(Neural Network)  
| 3      | TS=(Inductive Logic programming) AND TS=(Deep Learning)  
| 4      | TS=(Inductive Logic programming) AND TS=(Relational Knowledge)  
| 5      | TS=(Inductive Logic Programming (ILP)) AND TS=(Relational First Order Theory)  
| 6      | TS=(Inductive Logic Programming (ILP)) AND TS=(Relation knowledge) AND TS=(First Order Theory)  
| 7      | TS=(Inductive Logic Programming) AND TS=(Rule Extraction) AND TS=(Neural Network)  
| 8      | TS=(Inductive Logic Programming(ILP)) AND TS=(Rule Extraction)  
First Refined list:
A total number of 5571 articles were retrieved in the first refined list.

Second Refined list:
In the second step, the database was screened and refined to find the second refined list. At least three combinations of the keywords were used to select and ensure the second refined list of those articles which contain both ANN and at least one rule extraction method used to solve the problem. The second, refined list contained 86 articles.

Third Refined list:
A third, refined list was created by excluding 7 articles due to the nonavailability of their journals and conference proceeding ranking and 12 articles have a dual presentation. The 3rd refined list, which includes 67 articles, is given in table 5.2, which represents a yearly quantity of articles.

Table 5.2 The third refined list of articles.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Articles</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>19</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67</td>
</tr>
</tbody>
</table>

Scimago Journal & Country Rank (SJR):
To ensure high quality among the refined articles, Table 5.3 shows the ranking of the journals and conference proceedings of the selected articles with respect to “SJR: Scientific Journal Rankings-SCImago”, given below. In the fourth refined list, those articles were selected which were published in higher ranks journals and conference proceedings. The “Rank” value in Table 5.3 represents the higher rank of the journals and conference proceedings in descending order. Table 5.3 has 53 rows because some rows contain more articles.
Table 5.3 The “Scientific Journal Rankings (SJR) - SCImago” list.

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Journal or Conference proceeding</th>
<th>Rank</th>
<th>Year</th>
<th>Arts.Qty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lecture Notes in Computer Science (2), Artificial Neural Networks and Machine Learning – ICANN</td>
<td>251</td>
<td>2005,10, 16, 16</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Plos One</td>
<td>218</td>
<td>2016</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>IEEE Transactions on Neural Networks and Learning Systems (2)</td>
<td>161</td>
<td>2015,18</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Journal of Machine Learning Research</td>
<td>147</td>
<td>2004</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Chemical Engineering Journal</td>
<td>141</td>
<td>2013</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Expert Systems with Applications</td>
<td>131</td>
<td>2011</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Proceedings of SPIE</td>
<td>125</td>
<td>2017</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Machine Learning</td>
<td>124</td>
<td>2001,14, 17</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Neural Networks</td>
<td>117</td>
<td>2016</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>Accident Analysis and Prevention</td>
<td>108</td>
<td>2016</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Sensors</td>
<td>104</td>
<td>2011,17</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>Neurocomputing</td>
<td>94</td>
<td>2010,11</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>Applied Soft Computing</td>
<td>89</td>
<td>2014,14</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Journal of Systems and Software</td>
<td>82</td>
<td>2015</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>Engineering Applications of Artificial Intelligence</td>
<td>76</td>
<td>2011</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>International Journal of Approximate Reasoning</td>
<td>76</td>
<td>2017</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>Knowledge-Based Systems</td>
<td>74</td>
<td>2013,15, 16,16,17</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>Journal of Hydrologic Engineering</td>
<td>67</td>
<td>2014</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>Advanced Engineering Informatics</td>
<td>60</td>
<td>2017</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>IEEE Transactions on Semiconductor Manufacturing</td>
<td>55</td>
<td>2017</td>
<td>1</td>
</tr>
<tr>
<td>No.</td>
<td>Journal Title</td>
<td>Volume</td>
<td>Year(s)</td>
<td>Number</td>
</tr>
<tr>
<td>-----</td>
<td>-------------------------------------------------------------------------------</td>
<td>--------</td>
<td>-----------------</td>
<td>--------</td>
</tr>
<tr>
<td>22</td>
<td>Soft Computing</td>
<td></td>
<td>53</td>
<td>2010,12</td>
</tr>
<tr>
<td>23</td>
<td>International Journal of Adaptive Control and Signal Processing</td>
<td></td>
<td>53</td>
<td>2010</td>
</tr>
<tr>
<td>24</td>
<td>Brain Stimulation</td>
<td></td>
<td>49</td>
<td>2017</td>
</tr>
<tr>
<td>25</td>
<td>2012, 2016, 2016 International Joint Conference on Neural Networks (IJCNN)</td>
<td></td>
<td>49</td>
<td>2012,16,16</td>
</tr>
<tr>
<td>26</td>
<td>Knowledge And Information Systems</td>
<td></td>
<td>47</td>
<td>2016</td>
</tr>
<tr>
<td>27</td>
<td>Journal of Intelligent Information Systems</td>
<td></td>
<td>45</td>
<td>2004</td>
</tr>
<tr>
<td>28</td>
<td>Neural Processing Letters</td>
<td></td>
<td>42</td>
<td>2012</td>
</tr>
<tr>
<td>29</td>
<td>Cognitive Systems Research</td>
<td></td>
<td>38</td>
<td>2018</td>
</tr>
<tr>
<td>30</td>
<td>Etri Journal</td>
<td></td>
<td>37</td>
<td>2017</td>
</tr>
<tr>
<td>31</td>
<td>International Journal of Information Technology &amp; Decision Making</td>
<td></td>
<td>31</td>
<td>2017</td>
</tr>
<tr>
<td>32</td>
<td>Procedia Computer Science</td>
<td></td>
<td>29</td>
<td>2017</td>
</tr>
<tr>
<td>33</td>
<td>IFIP Advances in Information and Communication</td>
<td></td>
<td>29</td>
<td>2011</td>
</tr>
<tr>
<td>34</td>
<td>Frontiers in Artificial Intelligence and Applications</td>
<td></td>
<td>25</td>
<td>2016</td>
</tr>
<tr>
<td>36</td>
<td>Lecture Notes in Electrical Engineering</td>
<td></td>
<td>19</td>
<td>2016</td>
</tr>
<tr>
<td>37</td>
<td>Intelligent Automation and Soft</td>
<td></td>
<td>17</td>
<td>2012</td>
</tr>
<tr>
<td>38</td>
<td>Digital Signal Processing</td>
<td></td>
<td>17</td>
<td>2016</td>
</tr>
<tr>
<td>39</td>
<td>Analytic Methods in Accident Research</td>
<td></td>
<td>14</td>
<td>2016</td>
</tr>
<tr>
<td>40</td>
<td>International Journal of Cognitive Informatics and Natural</td>
<td></td>
<td>14</td>
<td>2017</td>
</tr>
<tr>
<td>41</td>
<td>2016 16th International Conference on Control, Automation and Systems (ICCAS)</td>
<td></td>
<td>10(2010)</td>
<td>2016</td>
</tr>
<tr>
<td>42</td>
<td>2016 IEEE International Conference on Image Processing (ICIP)</td>
<td></td>
<td>10(2014)</td>
<td>2016</td>
</tr>
<tr>
<td>43</td>
<td>Intelligent Information and Database Systems, ACIIDS</td>
<td></td>
<td>8(2009)</td>
<td>2011</td>
</tr>
<tr>
<td>44</td>
<td>Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)</td>
<td></td>
<td>8</td>
<td>2013</td>
</tr>
<tr>
<td>45</td>
<td>IEEE Industrial Electronics Society</td>
<td></td>
<td>8(2014)</td>
<td>2016</td>
</tr>
</tbody>
</table>
Table 5.4 Shows the number of articles with respect to the journal’s ranking.

<table>
<thead>
<tr>
<th>Sr.#</th>
<th>Ranking Range</th>
<th>Articles Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>141-251</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>106-140</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>71-105</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>36-70</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>1-35</td>
<td>22</td>
</tr>
</tbody>
</table>

Fourth Refined list:
To ensure a high-quality list, a final refined articles list was created including articles whose journals and conference proceedings rank exist between 71 and 251 as shown in Tables 5.3 & 5.4. The final/4th refined list of the journals and conference proceedings ranking was created with respect to “Scientific Journal Rankings (SJR) - SCImago”. The final list included 30 articles. These articles were used to investigate the thesis research questions.
6 Results, discussion and Analysis, Research Questions Answers:

This chapter presents results (summary-tables) of the models related to topics the black box and rule extraction methods, discussions and analysis, and the answers to the research questions.

The final list contains thirty articles, and following are the most relevant approaches used for the problems related to topics the black box and rule extraction methods. These approaches are:

6.1 Approaches Related to Black Box and Rule Extraction Methods:

Manoel V.M. França, Gerson Zaverucha, and Artur S. d’Avila Garcez, (2014) [41], extended a neural-symbolic system known as C-IL$^2$P and introduced a fast method and algorithm called CILP++ to deal with first-order logic (FOL) with the help of Bottom Clause Propositionalization (BCP).

According to the authors, C-IL$^2$P is a neural-symbolic system that uses background knowledge in the form of propositional logic programs. The system is based on Bottom Clause Propositionalization (BCP). The CILP++ system (The integrated system, handles first-order logic knowledge) is an open-source and is a distributed neural-symbolic system for relational learning. The ILP systems Progol and Aleph also used Bottom clauses which are the boundaries in the hypothesis search space. According to the authors, at most standard configurations, CILP++ obtained accuracy similar to Aleph but behind Aleph. However, CILP++ system proved itself faster on early stopping configurations. CILP++ has shown itself better than RSD (Relational Subgroup Discovery) [42], a well-known propositionalization method. When BCP and RSD are using C4.5 (An algorithm used to generate a decision tree developed by Ross Quinlan.) as a learner, the results are closely similar to CILP++ system. However, in general, BCP obtained satisfactory runtime results.

The addition of mRMR (minimum Redundancy and Maximum Relevance) [33] with CILP++ is very useful. The mRMR decreases the number of features with a little loss of accuracy that improves the readability when knowledge extraction is required. The results were indicating that more than 90% of the reduction in features could be attained with a little loss of accuracy. The authors claimed that propositionalizing first order example with BCP while using any effective learning algorithm should present a faster and reasonably correct method to deal with first-order data.

Zeng Q, Huang HL, Pei X, Wong SC and, Gao M, (2016) [44], developed a Neural Network (NN) for modeling the nonlinear relationship between risk factors and crash frequency. The authors proposed a model that consists of two algorithms: one is N2PFA (Neural Network Pruning for Function Approximation) algorithm and the other is a modified rule extraction algorithm. Both algorithms are used to improve the generalization capacity of the system and also to treat the black-box characteristic of the neural network. The authors acquired crash dataset from TIS (Traffic Information System) that was managed under the Transport Department of Hong Kong for the proposed model. The acquired crash dataset was also used to compare the results with NB (negative binomial) model. The results showed that both trained and optimized NNs performed better to some
extent as compared to NB (negative binomial) models with respect to fitting and predictive performance. The authors also claimed that there was an act of dropping certain numbers of input and hidden nodes in the optimized NNs; however, a better approximation performance was achieved. That clearly shows the ability of the N2PFA algorithm to detect unimportant inputs and hidden nodes. It also shows the ability of the N2PFA algorithm to enhance the generalization capacity of the model. According to the authors, the proposed optimized neural network produced ten rules. In these rules, the explanatory variables have different coefficients that set up the truth regarding their nonlinearity related to the crash frequency. Mostly, the explanatory variables showed reasonable results. The proposed optimized model not only attained the better fitness performance but also displayed the effects of risk factors effectively as compared to the statistical model NB (negative binomial) and the previous NN (neural network) models.

Kamruzzaman and Sarkar (2011) [6] proposed a connectionist approach ESRNN (Extraction of Symbolic Rules from ANNs) for symbolic rule extraction from trained ANNs. The ESRNN is a four-phase algorithm that uses back propagation and has a recursive nature. First and second phases are concerned with the determination of suitable network architecture with the help of weight freezing based constructive and pruning algorithms. The third phase is concerned with the discretizing of the continuous activation values of the hidden nodes with the help of an efficient heuristic clustering algorithm. In the fourth phase of the algorithm, symbolic rules were extracted with the help of frequently occurred pattern based rule extraction algorithm. It was achieved by examining the discretizing activation values of the hidden nodes. The authors claimed that using ESRNN, high-quality rules were extracted from the given data sets. The extracted rules were concise, order insensitive, comprehensible, and did not involve any weight values. The authors claimed for the high accuracy in both pruned network and fully connected networks.

Authors also claimed that ESRNN helped to reduce the number of rules without loss of classification quality. With respect to the number of rules, average number of conditions for a rule, and the accuracy, the ESRNN system outperformed the “Pareto-based multi objective machine learning (PMML), CART, NN RULES, OC1, C4.5, NN-C4.5, DT RULES, Partial RE, BIO RE, RULES, Full RE, RULES-2, X2R, and PRISM” [35].

Mohamed, and Marghny, (2011) [45], proposed an approach to train a simple neural network by constructive learning and showed the convergence rate of error in ANN, with and without threshold. The authors applied Genetic Algorithm to approximate rules extraction from neural networks. The authors claimed that the rules extracted by using genetic algorithm are better than those extracted from various methods based on the decision trees. The implementation of the genetic algorithm is simple but for the large neural networks considerable computing resources are required. To check the method’s predictive ability and to compare it with standard classifiers, the method was tested on various public domain datasets (Monk datasets, breast cancer datasets, iris datasets, and lenses datasets). The results showed high accuracy. The performance of the network with threshold was better than the performance without threshold.

Dissolved gas analysis (DGA) is very popular for fault diagnosis in transformer. The predictive accuracy of artificial neural networks is higher as compared to other machine learning tools and human beings but their results are difficult to interpret, therefore known as a black box problem. To
get a better insight of the generated solution, it is required to extract knowledge from trained neural networks.

Bhalla D, Bansal, Bansal Raj Kumar, and Gupta HO, (2012) [7], applied a pedagogical technique to extract the rules from function approximating artificial neural network for fault diagnosis by using inputs of dissolved gases within the oil of the transformer. Their applied method evaluates the linear equations by activation functions approximation of the hidden unit within the neural network and divides the input space into sub-region, where each sub-region has a linear equation. The authors claimed that simple and useful rules can be extracted using the proposed approach.

According to the authors, the result of the ANN that takes input as the percentage concentration of five gases is better than the results of the ANN that takes inputs as the gas ratios. They also claimed that the generated “Rule Sets”, proved the ability to compute the inputs linear functions. The faults developed inside a transformer can be truly predicted by using the linear equations, partial discharge, discharge type, and the thermal. The linear equations can also be used for fault classification where neural network software is unavailable (such as in field sites).

The authors tested the system’s performance by applying it to the case studies. The “Rule Sets Fault Prediction” showed the same results with respect to the internal examination reports and the report generated by IEC (International Electrotechnical Commission) [46] / IEEE [41]. In all case studies, the authors claimed the true accuracy of advance warning regarding the incipient fault development within the transformer. They also claimed the promising results of the proposed method for transformer fault prediction and advance warning, even in the absence of one of the gases such as C2H4, H2, and C2H6.

Ozbakir L, and Delice Y, (2011) [47], proposed a new approach by combining artificial neural networks (ANN) with swarm intelligence to obtain understandable and accurate classification rules from databases. Apart from the high accuracy in classification and prediction, one of the disadvantages of ANNs is their black-box structure. In order to get rid of this disadvantage, Ozbakir L, and Delice Y, proposed a new approach to discover the hidden knowledge within ANNs.

The authors used the Particle Swarm Optimization (PSO) algorithm to extract the accurate and comprehensible classification rules by transforming the behaviors of trained artificial neural networks. PSO with time-varying inertia weight and acceleration coefficients, is configured to investigate the best attribute-value combination by optimizing artificial neural network output function. The results showed that the comprehensible, readable, and highly true/accurate classification rules could be obtained in this way. The authors set up a detailed experimental design to obtain the best performance of an artificial neural network for classification. The system was also compared with traditional and population-based algorithms to check two different performances measure the test accuracy and the number of rules. In addition, the selected datasets for experiments had different characteristics such as binary–n-ary classes, and continuous-categorical attributes. Apart from PSO-miner that slightly extends higher the number of rules, the testing accuracy values showed a notable difference as compared to other algorithms mentioned in the article. The efficiency and accuracy of the proposed system were found very encouraging to classification problems. Table 6.1 shows the summary of the neural networks and rule extraction algorithms models/approaches used for knowledge extraction.
Table 6.1 Summary of the neural networks and rule extraction algorithms models/approaches used for knowledge extraction:

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Ref.</th>
<th>Proposed Model</th>
<th>Objective</th>
<th>Approaches used</th>
<th>Results</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manoel V.M. França, Gerson Zaverucha, and Artur S. d’Avila Garcez</td>
<td>2014</td>
<td>[41]</td>
<td>CILP++</td>
<td>The task of learning first-order logic rules.</td>
<td>C-IL2P, BCP, RSD, C4.5, mRMR</td>
<td>CILP++ obtained accuracy similar to Aleph and stood behind Aleph, but it proved itself faster on early stopping configurations on most standard configurations, CILP++ showed superior as compared to RSD, using C4.5 as a learner, CILP++ showed similar results with BCP and RSD. While BCP attained overall satisfactory runtime results. Adding mRMR, a reduction of more than 90% of features can be achieved with a little loss of accuracy.</td>
<td>Propositionalizing first order example with BCP while using back-propagation should present a faster and reasonably correct way to deal with first-order data</td>
</tr>
<tr>
<td>Zeng Q, Huang HL, Pei X, Wong SC and, Gao, M</td>
<td>2016</td>
<td>[44]</td>
<td>N2PFA and a Modified Rule Extraction Algorithm</td>
<td>To explore the nonlinear relationship between crash frequency and risk factors</td>
<td>N2PFA, MLP, CGA, PSO, Modified of Piecewise Linear Function for Rules Extraction.</td>
<td>The optimized NNs perform better to some extent than NB (negative binomial) models with respect to fitting and predictive performance.</td>
<td>These methods may also be useful in the highway safety analysis. The proposed NN techniques can also predict risk factor’s effects.</td>
</tr>
<tr>
<td>Kamruzzaman, and Sarkar</td>
<td>2011</td>
<td>[6]</td>
<td>ESRNN</td>
<td>Symbolic rules extraction from trained ANNs</td>
<td>BP, DNC, CC FNNC, Pruning Algorithm, Heuristic Clustering Algorithm, RE, black-box technique (SVM, ANN, and RF), white-box technique (Ripper and C4.5), ALPA</td>
<td>ALPA is the first to be applicable to any black-box model while using advanced algorithmic concepts. The experiments have shown the suitability for rule extraction from SVMs, ANNs, and RFs, which holds considerable promise for the broad applicability of ALPA to other black-box techniques and domains.</td>
<td>The publicly available ALPA software, compatible with WEKA, can spur further investigation into development and application of rule extraction.</td>
</tr>
<tr>
<td>Mohamed, and Marghny</td>
<td>2011</td>
<td>[45]</td>
<td>GA, ANN</td>
<td>To extract approximate rules from neural networks</td>
<td>GA, ANN</td>
<td>Genetic algorithm technique determines rules better than the various methods based on the decision tree.</td>
<td>The implementation of the genetic algorithm is simple but for the large neural networks, considerable computing resources are required.</td>
</tr>
<tr>
<td>Bhalla D, Bansal, Bansal Raj Kumar, and Gupta HO</td>
<td>2012</td>
<td>[7]</td>
<td>REFANNs Pedagogical Approach For Rule Extraction</td>
<td>A pedagogical approach for rule extraction from function approximating ANN with application to develop fault diagnosis using the concentrations of the dissolved gases within the transformer oil, as the inputs.</td>
<td>DGA, IEC/IEEE, REFANN</td>
<td>The result of the ANN that takes input as the percentage concentration of five gases is better than the results of the ANN that take the gas ratios as inputs</td>
<td>The advance warning can be predicted regarding developing an incipient fault within the transformer even in the absence or not traced off one of the following gas such as C2H4, H2, and C2H6.</td>
</tr>
<tr>
<td>Ozbakir L, Delice Y</td>
<td>2011</td>
<td>[47]</td>
<td>ANN with swarm intelligence</td>
<td>To obtain comprehensible and accurate classification rules from databases</td>
<td>ANN, PSO</td>
<td>PSO-miner slightly extends higher the number of rules, testing accuracy values show the notable difference as compared to other algorithms such as NB Tree, Decision table, PART, C4.5.</td>
<td>The efficiency and accuracy of the new combined approach of the PSO-miner integrated with ANNs are found very encouraging for classification problems</td>
</tr>
</tbody>
</table>
6.2 Models related to classification and prediction:

In some of the above-mentioned articles as well as in the following articles, it is concluded that the researchers used various neural networks approaches for the purpose for classification and prediction. At the end, tables 6.2, and 6.3 show the summary of some of the approaches used for classification, and prediction respectively and the table 6.4 shows both approaches together in one system.

Zhao Wei, Meng Qing-Hao, Zeng Ming, and Qi Pei-Feng, (2017) [50], proposed a data-processing model for electronic noses (e-noses) system for the classification of different Chinese liquor brands. This model uses deep learning (DL) method based on stacked sparse auto-encoder (SSAE) to learn the features from the gas sensor response data. After learning the features, the results are used to build a new SSAE-BPNN (back propagation neural network). The authors also built “SSAE-SVM” model by combining the SSAE with support vector machine (SVM). The method based on SSAE don’t need preprocessing and feature extraction (generation and reduction), that make easier the data processing procedures for e-noses.

For identification, and classification of different brands of Chinese liquors and also to compare it with traditional methods, seven types of strong-flavored Chinese liquors were used to self-designed e-nose as experimental material. The experimental results showed that SSAE-BPNN method gave 96.67% results which is the highest classification performance as compared to SSAE-SVM and two other kinds of traditional methods.

Alom Md. Zahangir, Awwal Abdul A. S., Lowe-Webb R., and Taha TM, (2017) [49], used a deep learning approach called convolution neural network (CNN) for composite optical laser beam images classification and compared the results with other techniques such as deep neural network (DNN), support vector machine (SVM), and deep belief network (DBN).

According to the authors, the laser-beam images were taken from the National Ignition Facility (NIF). It is the largest, and the most energetic laser in the world, having around 40 000 optics that exactly guide, amplify, reflect, and focus 192 laser beams onto a fusion target. NIF uses the advanced radiographic capability (ARC) (a four petawatt lasers) for backlighting X-ray illumination production to capture implosion dynamics of NIF experiments with temporal resolution in picoseconds.

The authors, claimed experimental results with a testing accuracy of 95.86%, 86.56%, 86.66%, and 96.67%, by using CNN, SVM, DNN, and DBN respectively. Here Deep Belief Network (DBN) showed the best results as compared to other methods.

Yajun Zhang, Zongtian Liu and Wen Zhou, (2016) [43], introduced the Chinese emergency event recognition model (CEERM), based on deep learning. According to them, the event recognition models already present in the market are rules based and with shallow neural networks. These models had some limitations. For example, extracting features are difficult with the rule-based methods and these methods also have poor portability problem. Likewise, the methods based on shallow neural networks result in low recognition precision because of their tendency to converge.
too quickly to a local minimum. They also have a problem with complex function approximations and the vanishing gradient problem.

The authors used the term “segmentation system” for sentence segmentation and arranged the words into five categories. These categories are; trigger words, objects, participants, time and location. Each word is vectorized for the six feature layers. These layers are; part of speech, length, dependency grammar, the distance between trigger word and core word, location, and the trigger word frequency.

They used a deep belief network (DBN) for training a feature vector set and obtained deep semantic features of words while using a back-propagation neural network to analyze those features for trigger words identification. During testing, they achieved robust results with a maximum F-measure value of 85.17% with recognition performance (F-measure is a weighted harmonic mean of Recall & Precision (R & P)). They even got much better results with an accuracy of 88.11% with recognition performance, using dynamic supervised deep belief network (DBN) that monitor the training performance. This is a supervised learning approach that was used to do fine-tuning to the restricted Boltzmann machine layer. While with dynamic supervised deep belief network (DBN), the training time was increased by 25.35% but it enhances the recognition performance and controls the training time effectively.

Li Huaxiong, Zhang, Zhou, and Huang, (2017) [48], proposed a deep neural network (DNN)-based sequential three-way decision (3WD) model for image data analysis [48]. Three-way decision (3WD) models are widely used in the fields of approximate reasoning, and decision-making. The 3WD models deal with boundary region decisions. By applying a 3WD model to image data, the main problem is to find a suitable granular feature extraction method.

Due to the powerful capacity for representation, DNN is a feature learning method and an effective feature extraction method. DNN is a cost-blind, time-consuming and two-way decision method that is unable to provide boundary decisions if enough information is not available that make it unfit for real-world cost-sensitive decision problems.

Therefore the authors made an extension to 3WD model by adding a nonlinear feature extraction method known as DNN-based sequential 3WD method that considered the test cost and misclassification cost in different decision phases. The authors designed experiments based on the PIE database. The experiments proved the effectiveness of the DNN-based sequential 3WD method with granular features in the input images.

The growths in deep learning have been similar to a modular system that put forth itself well to learn representations, by the use of a layerwise combination of the unsupervised learning with the supervised learning for fine-tuning. The authors, Son N. Tran, Arthur S., and d’Avila Garcez (2018) [41], tried to investigate the following questions.

“Whether such modularity can be useful to the insertion of background knowledge into deep networks, whether it can improve learning performance if such modularity is available, and extract
knowledge from trained deep networks, whether such modularity can offer a better understanding of the representations learned by such networks [47, PP 248]."

The authors proposed a new inference known as confidence rules for deep networks. This inference combines together both the symbolic representation and quantitative reasoning. The inference rule was modeled to treat hierarchical reasoning. In deep networks, this model was proved to be an appropriate representation for the modular training. The representation of knowledge using confidence rules can be inserted or extracted using Deep Belief Networks.

According to the authors, in a single-layer DBNs (Restricted Boltzmann Machines) the confidence rules provide a low-cost representation. The authors also said that the information loss at times can be reproduced if the modular layer-wise approach is used to extract knowledge from DBNs. For example, if the training is done with complex image data. The modular training can produce an improvement in the performance during insertion, learning, and extraction of the knowledge.

According to the authors, it is concluded that there is a declaration of construction of a hierarchical reasoning system which is able to integrate the symbolic and subsymbolic representation. The work done can be extended to deeper networks with the principle of layer-wise extraction and insertion. However, due to addition in computational overhead for knowledge extraction, the use of the confidence rules for model selection may not be useful in very deep networks.

The following tables show the summaries of some of the models/approaches used for classification and prediction.

- Table 6.2 Shows the summary of the models/approaches used for classification.
- Table 6.3 Shows the summary of the models/approaches used for prediction.
- Table 6.4 Shows the summary of the models/approaches used for and both classification and prediction together in one system such as in modular approaches.
Table 6.2 Summary of the models/approaches used for classification:

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year/ Domain</th>
<th>Ref.</th>
<th>Categories</th>
<th>Proposed Model</th>
<th>Objective</th>
<th>Approach used</th>
<th>Results</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohamed, and Marghny</td>
<td>2011/ Data Mining</td>
<td>[45]</td>
<td>Classification</td>
<td>GA, ANN</td>
<td>To extract approximate rules from neural networks</td>
<td>GA, ANN</td>
<td>Genetic algorithm technique determines rules better than the various methods based on the decision tree.</td>
<td>The implementation of the genetic algorithm is simple but for the large neural networks, considerable computing resources are required.</td>
</tr>
<tr>
<td>Alom Md. Zahangir, Awwal Abdul A. S., Lowe-Webb R., and Taha TM</td>
<td>2017/ Image processing, Computer Vision, Machine Learning</td>
<td>[49]</td>
<td>Classification</td>
<td>CNN</td>
<td>Composite optical laser beam images classification</td>
<td>CNN</td>
<td>The experiments showed the results with a testing accuracy of 95.86%, 86.56%, 86.66%, and 96.67%, using CNN, SVM, DNN, and DBN respectively.</td>
<td>Deep Belief Network showed the best results as compared to other methods, (CNN, SVM, and DNN).</td>
</tr>
<tr>
<td>Zhao Wei, Meng Qing-Hao, Zeng Ming, and Qi Pei-Feng</td>
<td>2017/ Data Mining/ Electronic Nose</td>
<td>[50]</td>
<td>Classification</td>
<td>SSAE-BPN, SSAE-SVM</td>
<td>To learn the features from the gas sensor response data</td>
<td>SSAE-BPN, SSAE-SVM</td>
<td>The results showed that SSAE-BPNN gave 96.67% results, the highest classification performance as compared to SSAE-SVM and two kinds of traditional methods.</td>
<td>The method based on SSAE doesn’t need preprocessing and feature extraction (generation and reduction).</td>
</tr>
</tbody>
</table>
### Table 6.3 Summary of the models/approaches used for prediction:

| Authors                          | Year/Domain                        | Ref. | Proposed Model                                                                 | Objective                                                                                                                                                                                                 | Approach used                                                                                           | Results                                                                                                                                                                                                                           | Comments                                                                                                                                                                                                                         |
|----------------------------------|------------------------------------|------|--------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Zeng Q, Huang HL, Pei X, Wong SC, Gao M | 2016/Data Mining/Risk actors and Crash Frequency | [44] | N2PFA and a Modified Rule Extraction Algorithm                               | To explore the nonlinear relationship between crash frequency and risk factors                                                                                                                               | N2PFA, MLP, CGA, PSO, Modified of Piecewise Linear Function for Rules Extraction.                               | The optimized NNs performed better to some extent than NB (negative binomial) models with respect to fitting and predictive performance.                                                                                       | These methods may also be useful in highway safety analysis. The proposed NN techniques can also predict risk factor’s effects.                                                                                     |
| Li Huaxiong, Zhang, Zhou, and Huang | 2017/Approximate Reasoning and Decision Making/Image Data Analysis | [48] | An extended 3WD model by adding a nonlinear feature extraction method to suggest a DNN-based sequential 3WD method. | Propose a model for data classification and image recognition                                                                                                                                                 | 3WD Model, A nonlinear feature extraction method to suggest a DNN-based sequential 3WD method.                  | PIE database based experiments prove the effectiveness of the sequential 3WD method with granular features in the input images.                                                                                     | DNN is an effective feature learning method and feature extraction method, but a cost-blind, time-consuming and two-way decision method, unable to provide boundary decisions if enough information is not available, which make it unfit for real-world cost-sensitive decision problems. |
| Kamruzzaman, and Sarkar           | 2011/Data Mining                    | [6]  | ESRNN                                                                            | Symbolic rules extraction from trained ANNs                                                                                                                                                              | BP, DNC, CC FNNC, Pruning Algorithm, Heuristic Clustering Algorithm, RE, blackbox technique (SVM, ANN, and RF), whitebox technique (Ripper and C4.5), ALPA | ALPA is the first to be applicable to any black-box model while using advanced algorithmic concepts. The experiments have shown the suitability for rule extraction from SVMs, ANNs, and RFs, which holds considerable promise for the broad applicability of ALPA to other black-box techniques and domains. | The publicly available ALPA software, compatible with WEKA, can spur further investigation into the development and application of rule extraction.                                                                                                                                 |


Table 6.4 Summary of the models/approaches used for classification and prediction both together in one system:

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year/ Domain</th>
<th>Ref.</th>
<th>Categories</th>
<th>Proposed Model</th>
<th>Objective</th>
<th>Approach used</th>
<th>Results</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhalla D, Bansal, Bansal Raj Kumar, and Gupta HO</td>
<td>2012/ Electrical Power and Energy Systems</td>
<td>[7]</td>
<td>ANN- Classification and REFANN- Prediction</td>
<td>REFANNs Pedagogic al Approach For Rule Extraction</td>
<td>Rule extraction from function approximating ANN with application to develop fault diagnosis.</td>
<td>DGA, IEC/IEEE, REFANN</td>
<td>The result of the ANN that takes input as the percentage concentration of five gases is better than the results of the ANN that take the gas ratios as inputs.</td>
<td>The advance safety warning can be predicted within the transformer, even in the absence or not traced off one of the following gas such as C2H4, H2, and C2H6.</td>
</tr>
<tr>
<td>Ozbakir L, and Delice Y</td>
<td>2011/ Data Mining/ Engineering Applications of Artificial Intelligence</td>
<td>[47]</td>
<td>ANN- Classification based training and PSO- Prediction.</td>
<td>ANN with Swarm Intelligenc e</td>
<td>To obtain comprehensible and accurate classification rules from databases</td>
<td>ANN, PSO</td>
<td>PSO-miner slightly extends higher the number of rules, testing accuracy values show the notable difference as compared to other algorithms such as NB Tree, Decision table, PART, C4.5.</td>
<td>The efficiency and accuracy of the new combined approach of the PSO-miner integrated with ANNs are found very encouraging for classification problems.</td>
</tr>
<tr>
<td>Manoel V.M. França, Gerson Zaverucha, and Artur S. d’Avila Garcez</td>
<td>2014/ Data Mining</td>
<td>[41]</td>
<td>The NN is trained from examples. Inside, a kind of classification based training and then CILP++ extracts the features and can be categorized as Prediction.</td>
<td>CILP++</td>
<td>The task of learning first-order logic rules and feature extraction.</td>
<td>C-IL2P, BCP, RSD, C4.5, mRMR, ANN,Neural-symbolic integration</td>
<td>CILP++, accuracy similar to or stood behind Aleph, but faster on early stopping configurations on most standard configurations, CILP++, superior as compared to RSD, using C4.5 as a learner, CILP++ showed similar results with BCP and RSD. Adding mRMR, 90% of features can be achieved with a little loss of accuracy.</td>
<td>Propositionalizing first order example with BCP while using back propagation should present a faster and reasonably correct way to deal with first-order data.</td>
</tr>
</tbody>
</table>
6.3 Discussions and Analysis

As mentioned before, the use of artificial neural networks (ANNs) is very popular especially in the problems related to the classification and prediction. The predictive accuracy attained by ANNs is often higher than that of human experts or other related methods. The approaches mentioned in the above tables, have some limitations, drawbacks, and are specific to certain domain areas. That means, it still needs a lot to do and a complete understanding of the black box behaviour of ANNs in the coming future would be a revolution in the field of artificial intelligence. In the tables 6.2, 6.3, and 6.4:

Mohamed, and Marghny, (2011) [45], applied Genetic Algorithm to approximate rule extraction from neural networks.

The authors said that the predictive capability of the approach used, showed high accuracy on various public domain datasets such as monk datasets, breast cancer datasets, iris datasets, and lenses datasets. However, the implementation of the genetic algorithm is simple but for the large neural networks considerable computing resources are required. That is not only a very complex task but also a major drawback of Genetic Algorithm. Perhaps, for the small networks, genetic algorithm determines rules better than the methods based on the decision tree. The greatest value of Genetic Algorithm is the determination of heuristic rules from medium terms.


The results showed that the Deep Belief Network (DBN) showed the best results 96.67% as compared to CNN, DNN, and SVM methods. It was a comparison between the rule-and feature-based classification. The authors suggested that the result can be even more improved by increasing the number of training samples to train the networks. But it’s a suggestion, and it is also specific to composite optical laser beam images classification.

Li Huaxiong, Zhang, Zhou, and Huang, (2017) [48], proposed a deep neural network (DNN)-based sequential three-way decision (3WD) model for image data analysis [48].

DNN is a feature learning method and an effective feature extraction method. But DNN is a cost-blind, time-consuming and two-way decision method that is unable to provide boundary decisions if enough information is not available that make it unfit for real-world cost-sensitive decision problems.

Therefore the authors made an extension to 3WD model by adding a nonlinear feature extraction method known as DNN-based sequential 3WD method that considered the test cost and misclassification cost in different decision phases. The authors compared the proposed sequential three-way decision (3WD) model with sequential cost-blind two-way decision model and sequential cost-sensitive two-way decision model. The results showed that an increase to the number of decision steps, decrease the costs of misclassification in all above three models. The authors designed experiments were based on the PIE database. The costs of 3WD model were lower as compared to two other sequential two-way decision models due to the incorporation of cost-
sensitive strategy and boundary decisions in 3WD sequential model. The experiments proved the effectiveness of the DNN-based sequential 3WD method with granular features in the input images.

Zhao Wei, Meng Qing-Hao, Zeng Ming, and Qi Pei-Feng, (2017) [50], proposed a data-processing model known as **SSAE-BPNN (Stacked Sparse Auto-Encoders-Back Propagation Neural Network)** for electronic noses (e-noses) system for the classification of different Chinese liquor brands. The model has one advantage, as it is based on SSAE, it don’t need preprocessing and feature extraction that make easier for data processing procedures for e-noses. The experimental results showed that SSAE-BPNN method gave the highest classification performance of 96.67% as compared to SSAE-SVM and two other kinds of traditional methods i.e., kernel entropy component analysis (KECA) and Principal component analysis (PCA) based methods.

Zeng Q, Huang HL, Pei X, Wong SC and, Gao M, (2016) [44], developed a model that consists of two algorithms: one is N2PFA (Neural Network Pruning for Function Approximation) algorithm and the other is a modified rule extraction algorithm. The purpose of the model was to create a nonlinear relationship between risk factors and crash frequency.

According to the authors, the proposed optimized neural network produced ten rules. That showed its limitation. In these rules, the explanatory variables have different coefficients that set up the truth regarding their nonlinearity related to the crash frequency. The N2PFA algorithm identified the insignificant factors and improves the generalization capacity of the proposed model by achieving the better approximation performance, when optimized NNs dropped off a certain amount of input and hidden nodes. The proposed optimized model not only attained the better fitness performance but also displayed the effects of risk factors effectively as compared to the statistical model NB (negative binomial) and the previous NN models. These methods may also be useful in highway safety analysis that showed its further possible application.

Bhalla D, Bansal, Bansal Raj Kumar, and Gupta HO, (2012) [7], applied a pedagogical technique to extract the rules from function approximating artificial neural network for fault diagnosis by using inputs of dissolved gases within the oil of the transformer. The result of the ANN that takes input as the percentage concentration of five gases is better than the results of the ANN that takes the gas ratios as inputs. According to the authors, the networks having too many hidden nodes are not ideal for rule extractions because such networks produce a large number of rules as an output. The advance safety warning can be predicted within the transformer, even in the absence or not traced off one of the following gas such as C2H4, H2, and C2H6. That was also a very impressive achievement. The linear equations used in the system, can be used for fault classification by the site engineers if neural network software is not available. That means, the researchers fully understood the black box behavior of the system to this specific problem.

Ozbakir L, and Delice Y, (2011) [47], proposed a new approach by combining artificial neural networks (ANN) with swarm intelligence to obtain understandable and accurate classification rules from databases by transforming the behaviors of trained artificial neural networks.

PSO with time-varying inertia weight and acceleration coefficients is configured to investigate the best attribute-value combination by optimizing artificial neural network output function. Apart from PSO-miner that slightly extends higher the number of rules, the testing accuracy values showed a
notable difference as compared to other algorithms mentioned in the article. The authors discussed the higher CPU timing of the system and the bigger size of the generated rule set. It would be the challenging tasks for future research to decrease the CPU times and also to decrease the size of the rule set without losing the efficiency and without compromising the predictive capability of the system. The efficiency and accuracy of the proposed system were found very encouraging to classification problems. The results showed that the comprehensible, readable, and highly true/accurate classification rules could be obtained in this way.

Kamruzzaman and Sarkar (2011) [6] proposed a connectionist approach ESRNN (Extraction of Symbolic Rules from ANNs) for symbolic rule extraction from trained ANNs. The proposed approach is a neural network based data mining model to mining classification rules from the databases. The ESRNN is a four-phase algorithm that uses back propagation and has a recursive nature. In the fourth phase of the algorithm, symbolic rules were extracted with the help of frequently occurred pattern based rule extraction algorithm. The authors claimed that using ESRNN, high-quality rules were extracted from the given data sets. The authors claimed for the high accuracy in both pruned network and fully connected networks. Authors also claim that ESRNN helped to reduce the number of rules without loss of classification quality. That was also a very impressive achievement.

Manoel V. M. França, Gerson Zaverucha, and Artur S. d’Avila Garcez, (2014) [41], introduced a fast method and algorithm called CILP++ to deal with first-order logic (FOL) with the help of Bottom Clause Propositionalization (BCP). It is an open-source and distributed neural-symbolic system for relational learning. An addition of mRMR (minimum Redundancy and Maximum Relevance) to the system, more than 90% of the reduction in features could be attained with a little loss of accuracy that improves the readability. The use of mRMR also helps to reduce the size of the network. The authors also claimed that CILP++ can improve the trade-off between accuracy and efficiency as compared to propositionalization methods.

The CILP++ deals with first-order logic (FOL). The authors claimed that propositionalizing first-order example with Bottom Clause Propositionalization (BCP) using backpropagation offer a reasonably accurate and faster way to deal with first-order data. It can also lead the researchers; how to deal with second-order logic (SOL) in future? The translation of background knowledge into artificial neural networks can be investigated further in CILP++. The capacity of CILP++ to deal with large relational datasets can also be investigated in future.

The tables 6.2, 6.3, and 6.4 showed that different techniques have been used in various problem domain areas, with some advantages and disadvantages. The implementation of some algorithms is simple for the smaller networks but for the large neural networks considerable computing resources are required. The structure of some rule extraction models have limited number of hidden layers and as well as neurons (nodes).

It is concluded that the selection of neural network approaches, rule extraction algorithms, and their supporting methods are based on the requirement of the problem, the nature of the problem and the data available for training and for input purpose.
A lot of databases of Journals and Conference proceedings articles exist in the market. Due to limited time frame it was impossible to consult all those databases. In order to continue further research more articles from different databases can be investigated in future. The old methods can be replaced with modern approaches. A comprehensive knowledge of the problem to be investigated helps the researchers to select the appropriate artificial neural network approaches, rule extraction methods, and the supporting methods from the existing models present in the market. It also helps the researchers in the development of new approaches.

A detailed explanation of each approach and its supporting methods are beyond the scope of the thesis questions. To study and discuss each method and its supporting methods is a time consuming process. This is the black box problem. It is not completely open yet, that is why it is an interesting field of research right now that need to be explored. When it will be completely explored? It is a difficult question yet. The models mentioned in the above tables 6.2, 6.3, and 6.4 can guide the researchers to select the appropriate supporting methods, artificial neural networks, and rule extraction approaches for the specific problem domain areas.

If anybody wants to explore these black box problems, studying this thesis may contribute an adequate knowledge relevant for synthesizing rules of the knowledge hidden in neural networks.
6.4 Research Questions and Answers:

This part presents the answers to the two research questions.

**Question 1:** How prevalent is the black box problem in the research literature during the period 2000 to 2018?

**Answer:**

Table 5.2 (Chapter-5) and Figure 6.1 show the yearly quantity of articles published for deep learning black box problems. Figure 6.1 shows that from the year 2001-2005, a few articles were published and from 2006-2009 the quantity was at a very low level, then it increased gradually from 2010-2012. From 2013-2015, the number of articles published went down once again and then it went up again from 2016 until February 2018. The year 2016 stood peaked among all the years from 2001 until Feb 2018 with respect to the number of articles published, regarding deep learning black box problems.

The overall study shows that the number of deep learning black box problems are increasing gradually with the passage of time. To open the black box of such problems and convert it into the white box, would be a revolution in the domain of deep learning.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Articles</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>19</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67</td>
</tr>
</tbody>
</table>

Table 5.2 (Chapter-5) The third refine list of Articles.

![Graph of Deep Learning Black Box Problems](image_url)

Figure 6.1 The number of articles published about deep learning black box problems from 2000 to 5/3/2018
**Question 2:** What rule extracting methods have been proposed from 2010-2018?

**Answer:**

The tables 6.2, 6.3, and 6.4 present detailed information about the published articles, such as authors name, published year, problem domain, reference number, categories, proposed models, objectives, approaches used, results and the comments regarding the models.

The rule extraction models/approaches summarized in the tables 6.2, 6.3, and 6.4 are the answers to the second research question. The tables show the summaries of the most relevant approaches used for knowledge extraction for the problem related to classification and prediction. Perhaps, it is beyond the scope of this thesis to provide the detailed explanation of the each models/approaches mentioned in this thesis.

Table 6.2 shows a summary of the rules extraction methods used for classification such as Genetic Algorithms (GA), Convolution Neural Network (CNN), and Stacked Sparse Auto-Encoders (SSAE-BPNN) Back Propagation Neural Network, and etc.

Table 6.3 shows a summary of the rules extraction methods used for prediction such as Neural Network Pruning for Function Approximation (N2PFA), Multilayer Perceptrons (MLP), Conjugate Gradient Algorithm (CGA), Particle Swarm Optimization Algorithm (PSO), Three-Way Decision (3WD), and etc.

Table 6.4 shows a summary of the rules extraction methods in modular approaches with respect to both classification and prediction together in the one system such as artificial neural network (ANN) with Rule Extraction from Function Approximation of Neural Networks (REFANN), ANN with Particle Swarm Optimization Algorithm (PSO), and etc.

The analysis of the summaries showed that different techniques have been used in various problem domain areas. It was also found that some rule extraction approaches work with a limited number of hidden layers and also with a limited number of nodes (neurons).

It is concluded that the selection of neural network approaches, rule extraction algorithms, and their supporting methods are based on the requirement of the problem, the nature of the problem and the data available for training and for input purpose.

Further research can be done by searching more articles from different databases. Some methods can be replaced with better methods. It is necessary to have a comprehensive knowledge of the problem that is required to be investigated. It will help the researchers not only to select the appropriate ANN approach, rule extraction methods, and the other supporting methods from the existing approaches but also leads the researchers to add the new ones.

The study lays the foundation for the development of methods for synthesizing rules of the knowledge hidden in neural networks.
7 Conclusion:

Knowledge discovery is a process of automated extraction of unknown, hidden, and fruitful information from large databases. The prediction accuracy of artificial neural networks for the problems related to data classification and prediction is often higher as compared to human experts or other related methods. But their results are un-understandable, due to their complex architectural structure. Therefore, it is known as the “Black Box problem”.

This study has created a referenced work based on principled collection methods. The first aim of the research was to investigate, that how prevalent is the black box problem in the research literature during the period 2000 to 2018?

The overall study proved that the number of black box problems is increasing gradually with the passage of time.

The second aim of the research was to investigate, what rule extracting methods have been proposed in this regard from 2010-2018?

The tables 6.2, 6.3, and 6.4 showed some of the rules extraction methods used for knowledge discovery in different problem domains. For example:

The table 6.2 in chapter-6 showed the summary of the rules extraction methods along with their results used for classification such as Genetic Algorithms (GA), Convolution Neural Network (CNN), and Stacked Sparse Auto-Encoders (SSAE-BPNN) Back Propagation Neural Network.

The table 6.3 in chapter-6 showed the summary of the rules extraction methods along with their results used for prediction such as Neural Network Pruning for Function Approximation (N2PFA), Multilayer Perceptrons (MLP), Conjugate Gradient Algorithm for training the network (CGA), Particle Swarm Optimization Algorithm (PSO), Three-Way Decision (3WD), etc.

The table 6.4 in chapter-6 showed the summary of the rules extraction methods with respect to the classification and the prediction together in the one system such as artificial neural network (ANN) with Rule Extraction from Function Approximation of Neural Networks (REFANN), ANN with Particle Swarm Optimization Algorithm (PSO), etc.

The analysis of the summaries showed that different techniques have been used in various problem domain areas. It is concluded that the selection of neural network approaches, rule extraction algorithms, and their supporting methods are based on the requirement of the problem, the nature of the problem and the data available for training and input purpose.

Some rule extraction approaches work only with a limited number of hidden layers and also with a limited number of neurons. A significant work has been done to understand the internal connections and relationships present within the network. It is still needed a lot to do. A full understanding of such black box problems in the coming future would be a revolution in the domain of deep learning neural networks and artificial intelligence.
**Future Work:**

The future research could be;

1. If you have time to explore more articles, how do you think it would affect the outcome of this project?
2. Will it be possible to see a pattern, for example, this method seems to be most appropriate to use on this type of model in general?
3. Perhaps some methods can be replaced with better methods?
4. How can we develop methods for synthesizing rules of the knowledge hidden in neural networks?

It is necessary to have a comprehensive knowledge of the problem that is required to be investigated and solved. It will enable the researchers not only to select the appropriate ANN, rule extraction methods, and the other supporting methods from the existing approaches but also leads the researchers to add the new ones.

The study lays the foundation for the development of methods for synthesizing rules of the knowledge hidden in neural networks.
References:


[8]. M. Gethsiyal Augusta and T. Kathirvalava kumar, “Rule Extraction from Neural networks - a Comparative Study”, Pattern Recognition and Medical Engineering, 978-1-4673-1039-0/12/$31.00 ©2012 IEEE.


[35]. https://deeplearning4j.org/neuralnet-overview#concept
[36]. http://rinuboney.github.io/2016/01/19/ladder-network.html


Appendix A - Abbreviations:

1. ALPA - Active Learning Component and the Pedagogical Approach
2. ANN - Artificial Neural Network
3. ANN-DT - Artificial Neural Network-Decision Tree
4. BCP - Bottom Clause Propositionalization
5. BB - Black-Box
6. C4.5 - An algorithm used to generate a decision tree developed by Ross Quinlan
7. CC - Cascade Correlation algorithm
8. C-IL2P - Neural-Symbolic System
9. CILP++ - An extension of C-IL2P, A fast method, and an algorithm for ILP learning with artificial neural networks (ANNs)
10. CGA - Conjugate Gradient Algorithm for training the network
11. CNN, or ConvNet - Convolutional Neural Network
12. DGA - Dissolved Gas Analysis
13. DNC - Dynamic Node Creation
14. ESRNN - Extraction of Symbolic Rules from ANNs
15. FNNC - Feedforward Neural Network Construction Algorithm
16. GA - Genetic Algorithms
17. HCA - Heuristic Clustering Algorithm such as k-means, k-medoids etc.
18. IEC - International Electrotechnical Commission
19. KBANN - Knowledge Base ANN
20. MLP - Multilayer Perceptron
21. mRMR - minimum Redundancy and Maximum Relevance
22. N2PFA - Neural Network Pruning for Function Approximation
23. NB - Negative Binomial
24. PSO - Particle Swarm Optimization Algorithm
25. RSD - Relational Subgroup Discovery
26. RA - Rule Extraction Algorithm
27. REFANNs - Rule Extraction from Function Approximation of Neural Networks
28. RF - Random Forest
29. SVM - Support Vector Machine
30. Stacked Sparse Auto-Encoders (SSAE-BPNN) Back Propagation Neural Network
31. TIS - Traffic Information System
32. 3WD - Three-Way Decision