1.1 Introduction

In case of statistics, once a conceptual model is built (null hypothesis), we go for the validation of that hypothesis which leads us to the final acceptance or rejection of the null hypothesis. In contrast, the intelligent computational techniques for machine learning (Knowledge discovery) works almost in an opposite way wherein the first step does not start with the null hypothesis. Rather, we just have a data set and we don’t really know what and which pattern we are looking for. So, here we start by applying the interestingness criteria (notion) over the dataset in an attempt to get some interesting patterns forming the basis of the hypothesis thus the name “Hypothesis discovery”. Inspired by Machine learning and Statistics, the process of data mining has been provided that extracts the nuggets of knowledge (potentially useful information) from the huge amount of complex and nonlinear data. All these techniques help in cognitive modelling of human intelligence wherein the linguistic, commonsense and intuitive skills are being learned.

Machine learning techniques such as Bayesian technique, logistic regression, Decision tree, Artificial Neural Network etc. have been used so far in problem solving, decision making and prediction. These techniques are evaluated on the basis of some important evaluation measures such as accuracy. However, apart from this, there is an important aspect of model evaluation i.e. cognitive development and functioning that help in extracting hidden nuggets of knowledge efficiently. This involves the process of evaluation, reasoning, computation, comprehension etc. and such knowledge extracted can be best presented in the form of effective if-then rule strategy. All this is possible when the model tries to learn from the data/information (training data) provided. The research work focusses on such comprehensive study and experiments of various computational techniques, their hybrid structures improvements and applications.

The study and analyzation of Collective System/Model of some important computational techniques in the field of artificial intelligence and machine learning focusing on complex and nonlinear data is referred to as Softcomputing. The learning capability of artificial neural networks, vast data representation of fuzzy controllers and evolutionary behavior of genetic algorithms made researchers to build hybrid models that could act as a general predictor model for difference fields especially medical diagnostics.
Before predicting the class using the supplied dataset, the data needs to be purified i.e. remove irrelevant and noisy features or attributes. The softcomputing technique of building hybrid model for feature selection is also one of the challenge before extracting knowledge out of data using predictors. This involves filtering out the strong features that best predicts the output of data instance.

One of the challenging tasks in the medical field is an effective disease diagnosis based on complex, non-linear and uncertain data. Since learning from such data is difficult for medical experts (doctors), an automatic diagnostic system is required to simulate the experts that are better in performance as well. This research focusses on such intelligent control techniques that makes human-like decisions for the disease diagnosis.

1.2. Problem Definition:

The abundance of available data that is retrieved from or is related to the human condition and behavior challenges the research field in processing and analyzing it. The aim is two-fold: on the one hand, to extract knowledge that will help understand human behavior, creativity, way of thinking, reasoning, learning, decision making, socializing and even biological processes; on the other hand, to exploit the extracted knowledge by incorporating it into intelligent systems that will support humans in their everyday activities. The new techniques, appropriate to deal with such information, e.g. behavior, learning and biological information need to be proposed and existing ones adapted to its special characteristics is a significant challenge on its own [Han and Gao, 2008]. This involves managing and computationally analyzing realistic data for discovering the underlying existing knowledge.

In addition to this the extracted knowledge should be appropriate enough to match the real outcomes. To ensure this, evaluation on the existing or new computational techniques shall be performed keeping in view various important evaluation measures like Accuracy, error rate etc. Several issues like misclassifications (usually the most interesting one for the application that is being sorted out) by certain algorithms would also be taken into consideration.
1.3. Objectives:

The research will focus on the design and development of efficient knowledge extraction system. The objectives of this research project are as under:

1. Analysis of various existing knowledge discovery techniques.

2. Study of the work that has been previously done in this research area with focus on all the extents.

3. Based on the literature available, we shall provide a better method or model, which would be efficient enough in providing better knowledge than traditional methods.

4. Comparing various existing computational techniques taking into consideration various evaluation measures like Accuracy, error rate etc.

5. Comparing the new method with competing traditional methods focusing on various evaluation measures.

6. Implement and test the selected techniques with special focus on realistic data.

1.4. Methodology:

There are two major research methodologies that can guide this research. Given the nature of the problem to be addressed, an important part of the research will adopt an exploratory research approach where problem is not clearly defined and several data mining techniques will be tested to check what sort of useful patterns can be detected within a given scenario.

On the other hand, experimental/scientific methodology shall also be used where the data-based research comes up with conclusions capable of being verified by observation or experiment and thus arbitrates between competing models.

This research work is structured into four major tasks:

1) Problem definition

2) Design of system prototype

3) Implementation of the system prototype and
4) Testing the implemented prototype in a real-environment.

Figure below shows a work flow that is to be followed in this research:

![Work Flow for the Research](image)

**Fig 1.1. Work Flow for the Research.**

### 1.5. Contributions

The contribution which we have made during the research project is as under:

1. An attempt has been made to provide a detailed performance analysis of various machine learning techniques.

2. Ideas of various AI paradigms like evolutionary computing, random search procedures, and global optimization in neural network learning, fuzzy logic and use of heuristics based on human experiences have been reviewed out.

3. An attempt has been made to provide enhanced and better methods or new methods to extract some hidden knowledge and patterns out of the dataset.

4. Algorithms to model highly complex and nonlinear systems have been enhanced.

5. Efficient feature selection technique has been provided that helps in maintaining the efficient number of features (not small or large enough) that would increase accuracy and decrease the error.

6. Mathematical solutions have been provided to resolve certain issues and search new knowledge from the realistic dataset.
7. Appropriate rules formed by Generalization of inferencing systems has been provided.

8. Case studies have been introduced and used in analysis for evaluating the improved intelligent systems based on softcomputing.

9. An attempt to provide a general framework has been made representing the dataflow architecture to model the overall task of the diagnostic simulating robot for efficient knowledge discovery taking into consideration the cost efficiency.

10. Develop a powerful model for diagnosis based on computational system wherein the appropriate parameters are selected optimally and has less error and better accuracy

1.6. Outline

Chapter Two provides review of research work and contributions in the field of artificial intelligence, machine learning and data mining. Literature relevant to the analytic aspects of the data has been discussed keeping in view various algorithms based on computational technique. Further, we shall look into the technical aspects, various algorithms and concepts that make up the whole process of soft computing.

Chapter Three discusses various machine learning techniques by focusing on the two important aspects – classification and clustering. The techniques are being evaluated using various datasets and then compared on the bases of some evaluation measures.

Chapter Four discusses the fuzzification process and the various clustering techniques in fuzzy controller that help in input output space partitioning. Further, the membership types, Fuzzy Inference System (FIS) structures and the rule base structure has been discussed and compared to resolve the various issues that arise in fuzzification process. The section will provide a discussion of results acquired from the experiments done.

Chapter Five discusses various processes behind genetic algorithm, the parameters and options used. Different computational experiments have been conducted to check the performance of the genetic algorithm by changing the values of parameters. It shows how the little variations in the values for particular parameter changes the objective function value for the genetic algorithm
Chapter Six reviews various hybrid models of softcomputing, their advantages over the other and the drawbacks of each. Further computational techniques have been discussed in relation to how it helps in solving the various issues that arise in a complex nonlinear system. Proper Neuro-Fuzzy models have been applied on the medical dataset and compared in an attempt to search the best technique used for input/output space partitioning.

Chapter Seven discusses the various issues related to feature selection and parameter optimisation. Evaluation for improvements and comparisons have been done using Matlab simulator. The modifications have been made in an attempt to improve the overall performance. BCGA-Eco and GNFS have been proposed in order to improve the accuracy and decrease the error in the prediction system.