CHAPTER I

1. INTRODUCTION

The tremendous developments in the digital information technologies have expedited the development of massive volumes of multivariate data in an extensive range of uses like image processing, engineering, biological and medical diagnosis, Jaesung Lee et al., 2013 and Swati Shilaskar et al., 2013. In the data era, the growth of data has developed easier and economical. It has been estimated that the amount of stored information doubles in every twenty months, Witten H et al. 2005 and Oded Maimon et al., 2005. Unfortunately, the growth of data volume is far away from the human’s ability to manage and understand them. The advancement in the field of software accessories like hardware tools facilitates the collection of data automatically. While reducing the cost of such system design will lead to rapid increase in the collection and storing of data. Still there is a continuous increase in data collection and storage depending on the resources utilized. The cost also becomes high when more voluminous of data has been stored in a public storage. This happens because of extracting useful and potential knowledge to extract the relevant information which becomes more difficult and it takes more time, human power, technology for gathering such data precisely to form a data warehouse.

To overwhelm this problem, there is a trend of adapting machine learning with data mining S. Mitra 2004, H. H. Hsu et al., 2006, X. Q. Zhu et al., 2007 and artificial intelligence based techniques to handle and extract the useful knowledge from the vigorously growing volume of data. Data mining provides tools for analyzing large amounts of data. Though, applications with a large number of features or attributes make it interesting to abstract the valuable evidence from huge data. Feature selection is one of the preprocessing steps in data mining that selects a subset of the most important features and removes irrelevant, redundant and noisy features to solve the high dimensionality problems. This thesis proposes two novel techniques for handling the problem of data preprocessing in health care data. The data preprocessing contains two different phases: missing value handling using Fuzzy K-nn, Bezdek JC and selection of informative subsets known as optimal feature subset selection using Dempster Shafer theory.
1.1 An Overview of Data Mining

Extraction of already existing useful information and the process of extracting or digging valuable information from a large warehouse is known as Data mining Chang Pin Wei et al. The data used for observation is analyzed in various dimensions and determining the informative answers in a most summarized way to get the useful knowledge is its primary process. Data mining techniques are mainly helpful in organizations to set the marketing strategies. In addition, the unwanted data are eliminated to save the resources. It also helps in understanding the behavioral patterns of the customers and based on their easiness the knowledge are driven practical. The data mining approaches are very faster than the traditional approaches in covering their customers with the power of forecasting and prediction.

Mining techniques catch the attention greatly when there is a need for tuning the available voluminous data in a proper way to get useful and informative information. This resultant information can be further used for processing user queries, making decision, managing information and other various applications in the real time. The overall processing of discovering the informative patterns from the data is known as Knowledge Discovery in Databases. The KDD is greatly helpful in identifying the potential information with previously unknown fact of well organized data Sayyad Rasheeduddin., 2013. The KDD method is an essential part of the data mining. Which solves the problems in various domains more precisely, Gurmeetkaur et al., 2014. The raw dataset collected from the real world is converted to new form of knowledge information with the help of KDD. The repetitive process of KDD steps are as follows in the Figure1.1

Figure 1.1: Steps in KDD process
• **Learning the application domain:** The initial step which is very helpful for finding the information available in the given data warehouses.

• **Making a target dataset:** Based on the nature of the application area, the subset variable or the samples are analyzed. The data is collected from various resources and they need combination of actual representation of the dataset.

• **Data cleaning/preprocessing:** Data to be preprocessed to improve the quality of data and it is done by cleaning, noise removal and removing the outlier and finding various strategies of the missing value handling.

• **Reduction or Transformation of Data:** The growth in the dataset leads to process irrelevant, redundant information. The useful information to be gathered by choosing feature subset selection features depending on the ultimate task of the prediction process. The number of features to be reduced should be done more carefully.

• **Selecting the Data Mining algorithms:** To discover the hidden knowledge of the given dataset, the data mining is used. But depending on the goals, factors of the dataset and the values it holds are to be considered for choosing the type of the algorithm in the dataset.

• **Pattern evaluation and knowledge presentation:** In this step, the determination of the retrieved patterns based on the goals to be accomplished are evaluated and interpreted in this task. This is done after the previous step.

In KDD, the mining algorithms are chosen based on the application goal and extract the information based on it to emphasis the knowledge and increase the efficiency of the algorithm and to improvise the decision making quality. The steps which contribute in modeling the learning process is integration of data, preprocessing of the dataset and the nature of induction form the modeled algorithm. Depending on the requirement of the domain problem, different approaches are exposed to determine the different patterns which are implemented in mining algorithms. The most common technique used in the process of mining process is classification, rule mining, preprocessing, clustering and the analysis of evaluation. Priyadharsini.C , Dr. Antony Selvadoss & Thanamani ,2014.
Data Preprocessing:

In real world the data is full of noise, contains redundant information, missing value and inconsistent. Due to these factors the quality of the data is greatly affected. There is an importance in improving the quality of data arises and the data preprocessing is thus essential in this circumstances. The very crucial step in the mining process is transformation and preparation of the initial dataset for performing improvised pattern recognition. Data set with poor quality will always result in incorrect and unreliable output.

Figure 1.2 shows the process of data preprocessing in the data mining process.

Figure 1.2: Data preprocessing in Data Mining

Data preprocessing methods are divided into the following categories.

- **Data cleaning**: In this method, identification of noisy dataset and filling them, finding outliers and removing it, determining inconsistencies exist in the dataset and resolving are the important task. The whole process of data mining may leads to confusion due to the presence of dirty data.

- **Data Integration**: Multiple resources are used for collection of data and they are integrated and stored in data warehouse. The different source or forms of data collected are databases and cubes or flat files.
- **Data Transformation**: To perform efficient clustering or classification of dataset, the data has to be fall under a common range of value. The transformation of dataset by the process of generalization, normalization, smoothing, and attribute construction process are done during data transformation.

- **Data Reduction**: Without giving the integrity of the dataset and yielding, the good quality of data to gain knowledge the data reduction is a must. The dimensions of the dataset or the features are reduced for finding the optimal solution with limited time period and overcome the cost factor.

**Association Rule Mining**: In this process, extraction of useful information based on correlation, frequent pattern mining, finding association among the dataset instances or other repositories are performed. The relationship among the existing instances is determined.

**Classification**: The process of creating new models to distinguish and describe the data classes in data mining are known as classification technique. The prediction of type of class is done in it both using the label and without label. The training and testing phases are involved to learn more about the interesting patterns from the given dataset.

**Clustering**: The processes of finding similarity among the datasets are done using the clustering technique. The main objective of this work is minimizing the intracluster similarity and maximizing the intercluster similarity for unknown class variables. The unsupervised algorithms are used under this category. The classification algorithm used instances with labels for performing the prediction process while clustering is performed for unknown class labels.

**Outlier Detections**: Detection of outliers which doesn’t contribute in the process of knowledge discovery has to be identified and eliminated. Many different data mining approaches are discarded such outliers which are known as exceptions or noise. The outliers in these cases are the dataset which is out of fit for analysis.

**Evolution Analysis**: The behavior pattern changes frequently with different regularities for determining the evolution of dataset. The customer decision making process on different stages and providing strategy for marketing is done with the help of this analysis.
1.2 Data Mining in Clinical Data Management

In the medical data filed, maintenance of the patient dataset is a toughest and most challenging task which can be done thru the Data pre-processing techniques Goel A et al, 2016. In general, the medical dataset are voluminous and there is a need arises to handle them in a formatted manner and be secure and easy to retrieve based on the need, faster and credential information has to be mined effectively with proper and optimal utilization of storage as well as computation time.

Electronic health records Kharya S, Soni S et al., 2016 and Li D, Gu H, Zhang L, 2010 did the very first attempt that has maintained confidentially and securely to manage the patient records. Concurrently they are many medical institutions and healthcare centers strictly following it. The interconnected network of medical records is another innovation approach that has spurred a revival in the field of medical science which allows to share the clinicians about the patient information for the usage of getting an advice from the experts of the concern field and sharing the memory space accessible from other networks in order to provide better and instant backup facility. The intense of research now strongly focuses towards Knowledge-Based Systems/Clinical Decision Support Systems.

Several existing research works have been described on the consequences of mining medical data by solicitation of data mining techniques that include feature selection, outlier detection and classification/prediction. Each of the algorithms is assessed and the technique that yields the finest classification accuracy is nominated. The rules created by the classification algorithm and the medical data records on which the data mining techniques are implemented constitute the awareness of knowledge base which is the core component of any data mining framework.

Ensuing this, several medical databases comparative to the specific ailment under the revision can be given as classifier process input and the correctness in grouping can be established from the decision of the clinical system. Hence, such classifier schemes recommendation support to the medical consultants in predicting the progression of a disease based on the prevailing symptoms, recommending drugs, recognizing the necessity for hospitalization and anticipating conceivable time for recovery. Other data mining claims
associated to clinical repetition contain concerning the numerous side-effects of usage, checking common indications to help diagnosis, persuasive the most current drug mixtures for handling subpopulations that react obstinately from the predictable population to influenced drugs, and determining proactive steps that can lessen the hazard of illness. Data mining techniques thus bring improved impost of patient desires, healthier confirmation about clinical truthfulness, superior knowledge about patient consequences and implication between interferences and consequences (Madadipouya1K et al., 2015) and (Liao Z, Lu X, Yang T, Wang H 2009).

1.3 Applications of Data Mining in Healthcare Sector.

In, health care system, huge amount of information is used for storing personal information of patients, resources and information of hospitals, history of patients about the treatments provided for them etc. These information are collected for extracting useful knowledge for decision making and saving the cost.

In order to evaluate the efficiency of medical treatment, treatment effective mining techniques are developed. The mining process involves in comparison, contrasting symptoms treatment of group of patients with same type of symptoms. Health care management is necessary to identify and track prolong diseases in case of high risk, proper designing techniques are needed to perform and detect diseases in its initial stages. The effective functioning of health care management is greatly improved by integrating data mining applications. Data mining is used in the field of customer relationship management to enhance the relationship between organization and bank, seller and their customers with the help of mining techniques where the maintenance of customer relationship greatly improved.

In case of fraudulent detection data mining techniques, the abnormal activities of the transactions are recognized as unusual patterns and avoids mischiefs act. It also discovers false insurance and unsuitable health claims in the organizations and public sectors. The pharmaceutical analysis of new products and their services can be developed with the help of data mining approaches. To overcome the problem of competition exists among the firms and organization can be performed using proper decision making with the aid of discovering hidden knowledge in pharmacy data set. The development of hospitals with modern equipment and
enriched services to the patients are done using mining techniques adopted in medical fields. It also involves the pattern reorganization among the hospital staffs, managements and patients.

1.4 Feature Selection – An Overview

Feature selection plays an important role in data preprocessing techniques for data mining. Vipin Kumar and Sonajharia Minz, 2014. It is a process of finding a subset of features from the original set and forming patterns in a given dataset to obtain the optimal feature subset according to the given goal of processing and criterion. It reduces the number of features, removes irrelevant, redundant or noisy and brings immediate effects for applications: speeding up a data mining algorithm, improving mining performance such as classification accuracy, and improving results comprehensively Edgar Gabriel, Viswanath Venkatesan and Shishir Shah., 2010. Many researchers have focused on several feature selection algorithms at various points of view, but many of these are similar in content Krishnapuram R, Keller JM., 1999.

The following lists are conceptually different and cover a range of definitions.

**Improving classification accuracy**: The aim of feature selection is to choose a subset of features for improving classification accuracy or decrease the size of the features without significantly decreasing the accuracy of the classifiers (Polat K, Şahan S, Güneş S., 2007).

**Approximating original class distribution**: The goal of feature selection is to select a subset, such that the resulting class distribution, given only the values of the selected features, is as close as possible to the original class distribution given all feature values. (Polat K, Şahan S, Güneş S., 2007).

**Classical**: Select a subset of ‘M’ features from ‘N’ original features, M < N, such that the value of a criterion function is optimized over all subsets of size ‘M’ (Varde AS, Massey KL, Wood HC., 1991).

**Idealized**: Identify the minimally sized feature subset that is necessary and sufficient to describe the target concept (Isa IS, Saad Z, Omar S et al., 2010).

The first definition highlights the classification accuracy of a classifier using only the selected features, whereas the second definition emphasizes the class distribution given the training set. Hence, the feature selection attempts to select the small sized subset of features according to the following criteria (Krishnapuram R, Keller JM., 1999).
- The classification accuracy does not significantly decrease.
- The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution given all feature values.

1.5 Components in Feature Selection

Feature selection algorithms to search through the subsets of features and try to find the best among the competing $2^N$ candidate subsets according to some evaluation functions. However this procedure is exhaustive and it may be too costly even for a medium sized feature sets ($N$). Other methods based on heuristic or random search methods attempt to reduce the computational complexity by compromising the performance M. R. Nazari-Kousarrizi, F. Seiti, and M. Teshnehlab., 2012. There are four basic steps in a typical feature selection method Deepika Koundle, Savita Gupta, Sukhwinde, 2012 such as subset generation, evaluation function, stopping criterion and a validation procedure as depicted in the Figure 1.3.

![Figure 1.3: General Feature Selection process](image)

1.5.1 Subset Generation

The subset generation is basically a process of heuristic search, with each state in the search space specifying a candidate subset for evaluation. The nature of this process is determined by two basic issues. First, the search starting point which in turn influences the search direction. The search may start with an empty set and successively add features (i.e., forward search), or start with a full set and successively remove features (i.e., backward search), or start with both sides and add and remove features simultaneously (i.e., bi-directional).
Sometimes, search may also start with a randomly selected subset. Second, decides a search strategy. For a dataset with \( N \) features, there exist \( 2^N \) candidate subsets. This search space is exponentially prohibitive for exhaustive search with even a moderate \( N \). Therefore, three different search strategies have been explored (Edgar Gabriel, Viswanath Venkatesan and Shishir Shah., 2010): Complete, Heuristic and Random search.

A. Complete search

The complete search which explores every possible combination of ‘\( D \)’ features from the original ‘\( N \)’ features. This strategy will always achieve the best feature subset under given evaluation criterion, but generally with a huge computational cost, especially when ‘\( N \)’ is large. The order of the search space is \( O(2^N) \), a smaller number of subsets are evaluated Rajdev Tiwari., Manu Pratap Singh., 2010.

B. Heuristic Search

The heuristic searches are many variations to the greedy hill-climbing approach such as sequential forward search, sequential backward search and bi-directional search. All these approaches add or remove features one at a time. Algorithms with heuristic searches are simple to implement and fast in producing results as the order of the search space is usually \( O(N^2) \) or less.

C. Random search

In this search, features are either iteratively added or removed randomly. The evaluation function analyzes the suitability of a feature subset produced by the generation procedure and compares this with the previous best candidate feature subset, replacing it if found to be better. Consequently, the performance of the search process will depend on the resources available. The order of the search space is \( O(2^N) \), but this method typically searches a smaller number of subsets than \( 2^N \) by setting a maximum number of iterations possible.

1.5.2 Subset Evaluation criteria

As mentioned earlier, each newly generated subset needs to be evaluated by an evaluation criterion. The goodness of a subset is always determined by a certain criterion. The evaluation criteria can be broadly categorizes into two groups based on their dependency on data mining algorithms, namely classifier-independent and classifier-dependent criteria.
A. Classifier-Independent Criteria

The classifier-independent criteria are evaluating the goodness of a feature subset based on the inherent property of the data without involving any classification algorithms. Some popular classifier-independent criteria are distance measures, information measures, dependency measures and consistency measures (Senol, et al., 2009, Nikita Singh, et al 2009).

**Distance measures** are also known as separability or discrimination measures. For a two class problem, a feature X is preferred to another feature Y if X induces a greater difference between the two class conditional probabilities than Y.

**Information measures** typically determine the information gain from a feature. The information gain from a feature X is defined as the difference between the prior uncertainty and expected posterior uncertainty using X. The information gain of feature X is more than feature Y, then feature X is said to be better than feature Y. Entropy measures is such type of measures [Zhang K, Chai Y et al.,2010].

**Dependence measures** are also known as correlation measures or similarity measures. It measures the degree of relation between two variables by changing the value of one variable with observing the value of another variable. A feature X is preferred to another feature Y if the correction between feature X and class C is higher than the correction between Y and C.

**Consistency measures** are characteristically different from the above measures. In order to evaluate a given subset of features, its consistency rate is calculated by considering only the features of this subset. This refers to the number of instance pairs with same feature values, but belonging to different classes. Consequently, these measures find out the minimally sized subset that satisfies the acceptable inconsistency rate.

B. Classifier-dependent Criteria

The classifier-dependent criteria require a predetermined induction algorithm in feature selection and use the performance of the algorithm applied to the selected feature subset. It usually provides higher performance than the classifier-independent criteria, but it also tends to be more computationally expensive, and may not be suitable for other classification algorithms Maksood FZ, Achuthan G., 2016.
1.5.3 Stopping criteria

Since the number of subsets can be enormous, some sort of stopping criterion is necessary. A stopping criterion is tested every iteration to determine whether the feature selection process should continue or not. It may be based on a generation procedure/evolution function Mark A. Hall., 1999.

Stopping criterion based on generation procedure includes:

- Whether a predefined number of features are selected.
- Whether a predefined number of iterations reached.

Stopping criterion based on an evaluation function can be:

- Whether addition (or deletion) of any feature does not produce a better subset.
- Whether an optimal subset according to some evaluation function is obtained.

1.5.4 Result Validation

The result validation procedure is not a part of the feature selection process itself, but a feature selection method must be validated. It tries to test the validity of the selected feature subset by carrying out different tests, and comparing the results with previously known results, or with the results of competing feature selection algorithms using various sources of datasets. A more general way to validate the result is to evaluate the classification accuracy of the full set of features and the selected subset of features.

1.6 Problem Definition

In the field of medical research, diagnosing disease in their earlier stages is a very challenging issue for a long period of time. The primary work of the doctors is to diagnose disease very accurately. In the past years many engineering techniques and tools are used in the disease diagnosis which assists the medical personnel a lot. Even though there are many different traditional techniques are available for disease diagnosis the invention of soft computing based techniques paved a way for information technology to be introduced in the medical filed and which it yields new dimension of earlier prediction of diseases. With the advancement in storage devices now the past records of the patient is easily available. To predict the future behavior of the patient by studying the past history is a very good technique to diagnosis a disease.
The rapid growth in digital technologies have make possible to capable of generating and collecting a large amounts of data in a range of applications like medical diagnosis. In general details of the patients and their medical history are voluminous and they are stored as medical database or data mart or data warehouse. For the process of analyzing this kind of voluminous information in order to discover knowledge from them and it is really impossible by the manual operations and the time taken for diagnosing such disease is really very important factor which save the lives of many patients if they are analyzed and diagnosed in the earlier stage itself.

Hence, these factors necessitate a novel generation of computational techniques and tools are essential to support the mining of beneficial knowledge from the swiftly aggregate volume of data. Medical Data mining provides tools for such investigation of large amount of data and extracting potentially useful information for decision making process.

1.7 Motivation

In modern days, continuous monitoring of different medical parameters like blood pressure, temperature, heart rate, respiratory rates, ECG, EEG, and EMG can be done easily and all these data can be stored for future analysis and diagnosis. But the huge amount of data needs a lot of storage space which is costly. And another aspect is that all data are not of importance, only a few relevant data are important which possess a big problem in disease diagnosis.

The feature selection has been studied in the machine learning community in many years consequently; a novel technique is needed to assess relationships among the rapidly growing data and to reduce the redundant features. One of the approaches is the measure information content of the features using an information theory concept. It is generally more suitable to measure the information content of features in large of volume of data is the main motivation behind this thesis.

The motivation of the proposed work is to address the following:

- To remove the irrelevant and redundant features from entire database
- To reduce the dimensionality of the feature space
- To reduce the classification effort for new data collected
- Improve accurate decision making
1.8 Objectives of this Research

An information theory based feature selection and uncertainty based missing value handling plays an important role in effective feature selection. Any improvement of the feature selection problem represents an important advance in the pattern recognition with its implications to many other areas. The most important advancement, which is exploited in this thesis, is the ability to capture the interactions among all the selected features and relevancy in an efficient way. In particular, the thesis investigates the feature subset that contains highly correlated with the class, yet uncorrelated with each other. The objective of this thesis have been

i. To study and analyze the state-of-the-art feature selection algorithms in detail.
ii. To identify existing feature selection algorithms.
iii. To design and implement the concept of feature relevancy measures in the database classification problems.
iv. To describe the effectiveness of the redundant feature selection and develop a new techniques to overcome redundancy.

To develop and implement a prototype to prove the above concepts.

1.9 Contribution

The existing techniques used for disease diagnosis are still suffering from false detection rate due to the raw nature of dataset. The insufficient and incomplete dataset may leads to false alarms in diagnosis of accurate results. The preprocessing of dataset is less concentrated in the existing work which leads to major setback in overall process. The proposed work of our contribution is that instead of just finding the pattern recognition alone the dataset which is used for disease diagnosis has to be enriched with the following process.

- Data preprocessing methods to overcome the problem of handling missing values.
- Implementing best feature selection approach to determine the optimistic attributes which contributes mainly in pattern recognition.
- Designing the pattern discovery technique in case of incomplete and vagueness identification in the real time datasets.
Proposed Methodology of Medictiv Mining in case of Uncertainty using Fuzzy K-NN and Dempster Shafer Theory

Medical Dataset

Data Normalization

Enhanced Fuzzy K-NN based Missing Value Handling

Feature Subset Selection using Dempster Shafer Theory

Figure 1.4: Proposed Architecture of this research work
This research work aims at handling and enriching the information of the medical dataset in two different phases. This work uses the uci machine learning repository for collecting disease dataset for simulating the proposed work. In the first phase, the selected raw dataset is normalized for converting the different range of values to fall under the specific range. The minmax normalization is applied for the process of data normalization. Then to handle the missing value fuzzy K-NN based imputation method is adapted with the concept of finding k-nearest neighbor with the similar membership value and filling the predicted value based on it. The comparison of the proposed method is done with mean imputation, K-NN and Weighted K-NN. The performance result shows that the proposed work outperforms the existing algorithms.

In the second phase, to handle the voluminous dataset in the field of medical dataset ten different clinical dataset are analyzed to find the useful attributes and to eliminate the redundant and useless attributes from the dataset. To find the potential attributes this research work contributes the Dempster Shafer theory for handling the uncertainty which exist in the selection of feature subset by introducing the belief and plausibility as its factor for finding the optimal subset. The resultant features are validated using J48 and SVM classifier. Based on the simulation observation it is identified that the proposed Dempster Shafer approach performs better than the existing Normalized Mutual Information Feature Subset and join the Mutual Information Feature Subset.

Thus these two phases integrated under one roof to perform a enriched data preprocessing for accurate and enhanced prediction of disease in earlier stages itself.

1.10 Outline of the Thesis

In the chapter 1, overview of data mining, steps involved in the process of KDD, the importance of data mining in the field of medical science is explained in detail. The overall architecture of the research work with description of missing value handling and the feature subset selection are explained. The objective, motivation and contribution of this research work is neatly represented in this chapter.

In the chapter 2, the existing approaches in the field of disease diagnosis are elaborately explained. The various approaches and difficulties in handling the raw dataset of diseases are shown clearly. The literature survey exploits that even though there are lot of existing techniques are available to detect the disease, they still suffer to produce high accuracy due to its poor and
incomplete quality of dataset. So this research work concentrates on enhancing the quality of the medical datasets by performing missing value handling and selection of optimal features.

In the Chapter 3, the problem of handling missing values in the raw dataset of disease’s are well explained. The proposed method uses enhanced fuzzy K-NN approach to impute the missing values in the given dataset. The method of replacing value by filling with mean value of the concern attributes of other instances, conventional K-NN and weighted KNN methods are compared with the proposed method and the simulation results are clearly explained.

In the chapter 4, the problem of handling voluminous data in medical dataset is put forth. The proposal of Dempster Shafer theory based feature subset selection, importance of feature subset selection and the existing approaches on feature subset selection are clearly given in this work.

Chapter 5 concluded overall all observation of the research work with the outcome of two different phases of data preprocessing to enhance the quality of the medical dataset. The future enhancement of this proposed research work is also discussed in this chapter.

1.11 Conclusion

This era is believed as information age with the sophisticated technologies such as computers, satellite and other network techniques. It has collected tremendous amount of information due to its robustness. Regrettably, these massive collections of data are stored on unequal structures and they rapidly became devastating. This initial disorder has led to the formation of structured databases and database management systems. The challenges in extracting useful knowledge from data draws up researches in databases, pattern recognition, machine learning, data visualization and high-performance computing, to deliver advanced business intelligence. The ongoing rapid growth of data and the widespread use of databases have created an immense need for knowledge discovery in the database based methodologies. This chapter covers essential background of data mining, classification algorithms and its accuracy estimation methods used in the thesis.