# E-Intelligence form design and Data Preprocessing in Health Care

by

Padma Pedarla

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2004

# Abstract

Clinical data systems continue to grow as a result of the proliferation of features that are collected and stored. Demands for accurate and well-organized clinical data have intensified due to the increased focus on cost-effectiveness, and continuous quality improvement for better clinical diagnosis and prognosis. Clinical organizations have opportunities to use the information they collect and their oversight role to enhance health safety.

Due to the continuous growth in the number of parameters that are accumulated in large databases, the capability of interactively mining patient clinical information is an increasingly urgent need to the clinical domain for providing accurate and efficient health care. Simple database queries fail to address this concern for several problems like the lack of the use of knowledge contained in these extremely complex databases. Data mining addresses this problem by analyzing the databases and making decisions based on the hidden patterns.

The collection of data from multiple locations in clinical organizations leads to the loss of data in data warehouses. Data preprocessing is the part of knowledge discovery where the data is cleaned and transformed to perform accurate and efficient data mining results. Missing values in the databases result in the loss of useful data. Handling missing values and reducing noise in the data is necessary to acquire better quality mining results.

This thesis explores the idea of either rejecting inappropriate values during the data entry level or suggesting various methods of handling missing values in the databases. E-Intelligence form is designed to perform the data preprocessing tasks at different levels of the knowledge discovery process. Here the minimum data set of mental health and the breast cancer data set are used as case studies. Once the missing values are handled, decision trees are used as the data mining tool to perform the classification of the diagnosis of the databases for analyzing the results. Due to the ever increasing mobile devices and internet in health care, the analysis here also addresses issues relevant hand-held computers and communicational devices or web based applications for quick and better access.

# Acknowledgements

My deepest gratitude goes to my supervisors, Dr. K. Ponnambalam and Dr. Romy Shioda, for their guidance, and valuable suggestions in completing my thesis. I would like to specially thank Dr. K. Ponnambalam for his moral support and continuous encouragement throughout my program.

I would also like to greatly thank my thesis readers, Dr. Hamid Tizhoosh and Dr. Jose Arocha, for reviewing my thesis and providing insightful comments and suggestions.

My sincere appreciation goes to department secretary Vicky Lawrence for her valuable information when I needed most. I would also like to thank all my friends for their help and support throughout my program and stay in waterloo.

My deepest thanks to GOD for the blessings, strength and courage given to me in achieving this degree. I would finally thank all my family members for their deepest love, understanding and continuous support, without whom I wouldn’t have completed this program.

# Table of Contents

# List of Figures

# List of Tables

# Table of Acronyms

PDA:- Personal Digital Assistant.

KDD :- Knowledge Discovery in Databases.

RAI :- Resident Assessment Instrument.

MDS:- Minimum Data Set.

MH:- Mental Health

RAP :- Resident Assessment Protocol.

AI:- Artificial Intelligence

CRUISE :- Classification Rule with Unbiased Interaction Selection and Estimation.

CSV:- Comma Separated Format.

WEKA :- Waikato Environment for Knowledge Analysis.

NORM:- Multivariate Normal Distribution

DTREG:- Decision Tree Regression

FACT :- Fast Algorithm for Classification Trees.

# Chapter 1
# Introduction

Health-related data is extremely rich for discovering knowledge. The amount of information from heterogeneous sources, like the internet-based databases for diagnosis, symptoms, and procedures, are the parts of the health trail. The culmination of heterogeneous clusters of data would aid healthcare professionals in providing quality healthcare.

Clinical information is a powerful asset in increasing health care efficiency. Ever increasing amounts of clinical data are becoming the norm in both inpatient and outpatient care. The best possible health care is critically dependent on capturing the most complete data possible for accurate risk assessments. The diversity and complexity of clinical data makes the procurement, storage and analysis of this information one of the most exciting challenges. As a result, health informatics cuts across scientific boundaries.

Health Informatics focuses on the application of computer information systems to health care and public health. Vast quantities of clinical data, like information about patients and their medical conditions, are captured and stored in clinical data warehouses.

Evaluation of this stored data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our knowledge of in disease development and management. Thus, data mining is introduced to search the large quantities of data for these patterns and relationships

Data mining, an active area of research, is the process of finding hidden patterns in a given data by integrating and analyzing data from databases. Data mining techniques are used by various applications to predict useful information from stored data. The data mining process involves transferring originally collected data into data warehouses, cleaning the data to remove errors and check for consistency of formats and then searching the data using statistical queries, neural networks or other machine learning methods.

The data mining process consists of three stages:

(1) Initial exploration: The first stage in data mining, this stages starts with the data preparation.

(2) Model building or pattern identification: This stage involves considering various prediction models and choosing the best one based on their predictive performance.

(3) Deployment: The final stage involves using the model selected in the previous stage and applying it to new data in order to generate predictions and estimations of the expected outcomes.

In data mining applications, the preparation of data for analysis constitutes at least 80% of total effort. Data should be organized before performing data mining techniques for receiving better results.

Medical data repository requires a great deal of preprocessing in order to be useful. Numerical and textual information may be interspersed and different symbols can be used for the same meaning, redundancy often exists in data, erroneous, misspelled medical terms are common and the data is frequently sparse.

Data cleaning and handling missing values are the main stages in data preprocessing. Missing observations occur in many areas of research and evaluation. These observations may results in the loss of useful data during the data mining process.

## 1.1 Motivation:

Numerous data mining tools and methods are available for different types of applications. Machine learning is one kind of data mining tool that can be used for health related databases. Machine learning in health systems is developed to support healthcare workers in the tasks of manipulation of data and knowledge on daily basis. It is expected to be used for diagnosis and prognosis.

We will be dealing with text formatted data, which is stored in the form of numeric (binary) and text format in the database. The Thesis deals with the classification of a new case from the information provided by means of interactive medical forms. This Thesis mainly focuses on low level implementation of forms and data preprocessing, mainly handling missing values during entry level and before applying data mining techniques. Once errors in data have been removed, data mining tasks can be applied to extract the patterns in data.

This Thesis describes the E-Intelligence form design that can be implemented for any questionnaire type data forms. Various methods for handling missing values are discussed and checked for accurate results. Further, implementation of future user-interface controller is described, where already existing raw databases of any format can be inputted and various data mining techniques including missing values handling methods can be applied to these databases with results displayed to the user. The user also has an option of viewing the replaced values for the missing attributes.

With the utilization of mobile electronic devices, implementing the user-interface software on mobile or hand held devices such as PDAs should be taken into account. Once the user-interface is implemented and various data mining tools are accessible by PDAs, this software may replace paper-based forms and can be used to check / access data in an easier format.

In our research, we make use of two case studies for our research, both related to health care.  The system we are trying to automate is the MDS-MH, which is the minimum data set for mental health. The MDS-MH system can be considered as the minimum number of questions that need to be answered for the proper assessment of mental health patients. Firstly, we develop the E-Intelligence form design (MDS-MH) to enter the required data. Before the data is stored in the database it checks for inappropriate or missing data. This case study is related to mental health care and has 455 attributes for classification. Once the data preprocessing is done during entry level any data classification method can be used for performing the classification of patient assessment.

The second case study is related to breast cancer. This data is used to explain the various missing data handling methods that are used in this Thesis. The results also check for the accuracy of the methods. These methods are performed later on the subset of MDS-MH data. Few already existing attribute values are eliminated to compare the accuracy of the replaced values. These results are discussed in Chapter 4 of this Thesis.

## 1.2 Goals and Objectives:

Missing data may be inadvertent and uncontrolled but the overall result is that data cannot be analyzed accurately.  The main objective of this research is to address the impact of missing data on the data mining process and to show that handling missing values and reducing noise results in better quality data mining results. The goal of this thesis is to develop a form that is designed to minimize the missing and erroneous values during data entry step and further clean data using data mining and fuzzy logic techniques.

Firstly, during entry level, missing values and noisy data are reduced before they are stored into the database. For this case study, appropriate form design for the MDS-MH data is created using the Visual Basic software. Different tasks are taken into account in handling the missing values and reducing noisy data. Further, the implementation of a user-interface controller is discussed, for quick and easy usage of these testing methods on the raw data of any format. The designed form is named E-Intelligence form due to its features of handling the data electronically and the intelligence behavior to reduce missing and noisy data to produce quality data. E-Intelligence form refers to Electronic - Intelligence form.

Later, several methods for handling missing values are discussed on breast cancer data and checked for accuracy of the methods. A subset of MDS-MH data consisting of 18 attributes and 200 instances is considered for checking the accuracy of these replaced values.

Once the data preprocessing task is done, a data classification method namely, a decision tree technique, is used to classify the data for diagnosis in breast cancer case study and for patient assessment in MDS-MH case study, and check for better quality data. In future, implementation of the user-interface can be downloaded into PDA for convenience.

### 1.3 Thesis Outline:

Chapter 2 presents the literature review on knowledge discovery; data mining and different tools; text mining; health informatics; data preprocessing and its tasks; missing values and various methods used to handle them, and form design.

Chapter 3 provides the system architecture and the model used in this Thesis. The details of design used for developing the E-Intelligence form and the logic used to perform the intelligent data preprocessing tasks are discussed followed by the implementation of the user-interface controller for web/PDA application. Various tasks that are used in this Thesis to handle the missing values are also discussed using the breast cancer data. This data is considered as the case study to explain how these methods can perform and give better results compared to data without handling the missing values.

Chapter 4 discusses the implementation of the E-Intelligence form for the MDS-MH data. Various cases that are used to perform the data preprocessing tasks during the data entry are discussed in detail. Later, a subset of MDS-MH data is considered and a few existing attribute values are eliminated to check the accuracy of the replaced missing values using the methods. Summary of the results are provided, followed by the conclusion of the accuracy of the methods for handling missing values. The summary, conclusions and future work are presented in Chapter 5.

# Chapter 2

# Background and Literature Review

Computers changed our life enormously and have become influential machines in our daily life. They play important roles for a person in work, knowledge/information and entertainment. They are being used in various fields including education, banking, agriculture, military and medical applications. Humans invented computers that reduce the daily stress load and store important information that is necessary in future. In the medical field, for better health care a physician needs to understand the relationships between the patients and their life style and work environment, gather information from various sources like laboratory, x-rays, other physical examinations, and has to analyze and synthesize this data and develop a treatment plan. Computers play an important role in data procurement and storage. With the increasing use of computers, the size of stored data is also increasing.

As data increases at a phenomenal rate, users are expecting even more sophisticated information. There are various query languages to find the stored information, like SQL, but they cannot find the hidden information in the databases or can acquire knowledge from the stored data for future development. These tools can only perform analysis at a primitive level. This resulted in the evolution of data mining tools to find the hidden information in databases and find better techniques to search for useful information. Various data mining algorithms are introduced to classify the data based on similarities between the training and the testing sets.

Data mining has been used in various applications including fraud-detection in banking, weather prediction, financial analysis, and predicting diagnosis among others. These applications can be classified into sets of problems having similar characteristics. The same approaches and models can be used for similar applications, but the parameters can be different from industry to industry and from application to application. For example, the approach for fraud-detection in banking can be used for error-detection in medical insurance.

## 2.1 The Knowledge Discovery Process

Knowledge discovery (KDD) is the process of finding useful information and patterns in the data [9]. They work with various databases and use various data mining methodologies like fuzzy logic; neural networks etc. to retrieve useful hidden information. They have been used in various applications in fields like banking, automobile, agriculture, medicine etc.

The main goal of KDD is to find any high-level knowledge from the low-level data. It uses training sets to find the patterns, knowledge or useful information hidden in databases. It focuses on the overall process of knowledge discovery from storing data to accessing data, finding required algorithms to run efficiently and producing results, and how to provide the results to the end-user in an understandable way.

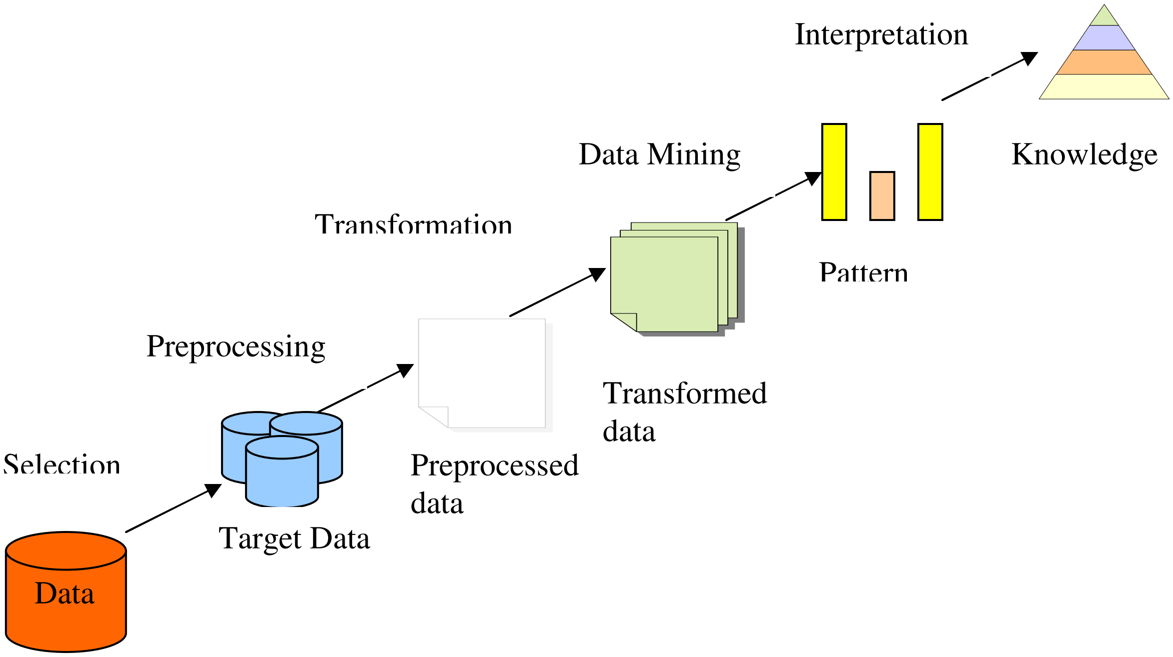


**Figure 1 Overview of the steps involved in the KDD process [17]**

Figure [1] shows the overview of the KDD process. It consists of several steps from selecting data to providing understandable knowledge to the user. The following are the steps involved in KDD:

STEP 1:- GOAL: First is to understand the main application problem. Keeping the end-user in mind, one has to identify the main goal of applying the process. For example: the necessity of a physician is to predict a new patient's diagnosis. The main goal here is to classify the patient to provide or lead towards a successful diagnosis.

STEP2:- SELECTION: Second is to select the necessary data to solve the problem. The main part of the KDD process is the selection of raw data that is necessary for the discovery. There might be unnecessary data attributes provided, but only few data attributes are needed in the process. Selecting the necessary data attributes for the discovery places an important part which leads to finding useful knowledge and accurate results.

STEP3:- PREPROCESSING: Handling missing values, eliminating noise and duplicate records in the data sets is the main part of this process. Missing values in the data lead to loss of useful information, which might not result in discovering useful knowledge. Noise in data and duplicate records mislead the process in obtaining accurate knowledge. Therefore, data cleaning and the preprocessing are necessary to produce better quality results. There are various tools to apply in these tasks.

STEP 4:- TRANSFORMATION: In this step data is transformed or consolidated into forms appropriate for data mining [40]. Data transformation involve: Smoothing, where the noise from data is removed; Aggregation, where aggregation operations are applied to data for the analysis of the data; Generalization, where raw data is replaced with high-level concepts; Normalization, where the attribute data are scaled to fall within a specified range; and Feature Selection, where new attributes are constructed and added from the given set of attributes.

STEP 5:- ANALYSIS AND MODEL SELECTION: This step matches the goals of the task to a particular data mining method. It analyzes the main problem and decides which models and parameters, like classification, clustering, or similarity etc. is appropriate. Depending on the model, different data mining algorithms and methods are chosen that are needed for searching data patterns.

STEP 6:- DATA MINING: This is the step where the main task of data mining is done. Data mining methods are performed to achieve the goal by finding the interesting patterns in the data. Better data mining results are obtained if the preceding steps are performed in the correct way.

STEP 7:- INTERPRETATION: This is the step where the mining results are interpreted and even the process can start again from step 1 if there are any errors or for providing further accurate results. This step also involves the visualization of extracted patterns and results. Finally, the knowledge discovery process takes the raw results from data mining and transforms them into useful and understandable information for the end-users.

This Thesis mainly concentrates on the data preprocessing step. The software designed here reduces the missing values and noise in the data during the entry level and make the data non-erroneous before they are stored in databases. Various methods are also discussed to handle the data preprocessing task of handling missing values. The data preprocessing task and how the missing values are handled are further discussed in the following section.

### 2.1.1 Data Preprocessing

Data preprocessing is the main step in the KDD process [8]. Data in the real world is dirty. Real world data is often incomplete and noisy, say wrong values or duplicate records. This results in poor quality data which in turn results in poor quality mining results. Quality decisions are based on quality data and data warehouses needs consistent integration of quality data, which has no missing or noise data. In order to get quality data, the data in the database need to be checked for accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility.

Tasks in data preprocessing are:

Data Cleaning: Filling in missing values, smooth the noisy data, identify or remove outliers, and resolve inconsistencies

Data Integration: Integration of multiple databases or files

Data Transformation: Normalization and aggregation

Data Reduction (Feature Selection): Obtains reduced representation in volume but produces the same or similar analytical results

Data Discretization: Part of data reduction but with particular importance, especially for numerical data

We described these tasks in detail:

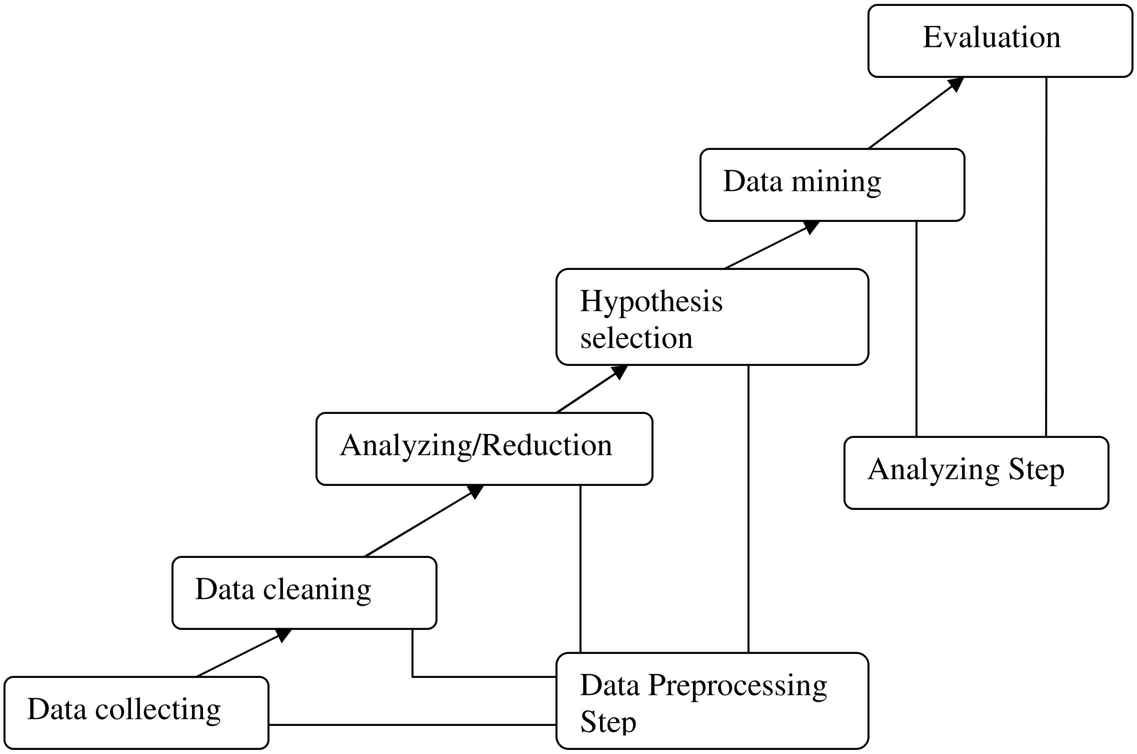


**Figure 2 Data preprocessing tasks [8]**

(1) Data cleaning tasks: Handles missing and noisy data.

(a) Missing data need to be inferred. Missing data may be due to: data not entered due to misunderstanding, data inconsistent with other recorded data and thus deleted, data may not considered important at the time of entry, data changes not recorded. Missing data can be handled by various ways: ignoring the records, filling the missing values manually, or using a global constant, or attribute mean or most probable value by inference based on Bayesian formula or decision trees.

(b) Noisy data: Noisy data is due to random errors or variance in a measured variable. Incorrect attribute values may be due to: faulty data collection instruments, data entry problems, data transmission problems, duplicate records, incomplete data and inconsistent data. Noisy data can be handled by binning method, or clustering or combined computer and human inspection by regression method.

(2) Data Integration: Data integration combines data from multiple databases into a single database. During the data integration, one has to detect and resolve data errors. Errors might be due to different values from different sources, different attribute formats, or attribute names might be different in different databases. One has to handle these kinds of redundant data to make sure the final database after integration consists of quality data. Correlation analysis can be used for handling these kinds of redundant data. Data integration reduces redundancies and inconsistencies and improves the speed of mining and produces quality information.

(3) Data Transformation: Smoothes the data (remove noise from data), summarizes and generalizes the data and constructs new attributes from the given ones. Normalization is done using min-max normalization, or z-score normalization or by decimal scaling. Sometimes map to higher dimension and categorical data to numerical data.

(4) Data Reduction (Feature/ subset selection): Data warehouses store vast amounts of data. Mining takes long time to run this complete and complex data set. Data reduction reduces the data set and provides a smaller volume data set, which yields similar results as the complete data sets. Data reduction strategies include: Data cube aggregation, dimensionality reduction, and luminosity reduction.

(5) Discretization and concept hierarchy: Discretization reduces the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Concept hierarchies reduce the data by collecting and replacing low-level concepts by higher-level concepts.

This Thesis mainly focuses on data cleaning task, handling missing and noisy data. The following sections describe the form design and how the missing values are handled in data entry and in existing data.

### 2.1.1.1 Form Design

The collection of data at clinical site is becoming an important component of clinical trials. To ensure accurate and timely collection and communication of information, collection of data at various points is necessary. Data collection is the key component to support both the research and day–to–day patient treatment.

There are several ways of data collection:

(a) Electronic data collection: Some systems use electronic data collection when manual data entry is not feasible or cost effective, such as data collected from laboratory instruments and databases

(b) Manual data entry: The data entry unit is staffed by experienced staffs that are familiar with special conventions and the quality needs of clinical data processing.

(c) Computer-Assisted Telephone Interviews: This system provides powerful interviewing feature such as dynamic lists, form-based questions, automated consistency checks, online context-sensitive interviewer help, external directory and coding lookup

(d) Optical Scanning: Scannable forms may be created to serve as interview instruments or case report forms

Forms are tools used for gathering the data that is necessary for further study. Data are transcribed from a patient's medical record onto a paper form, and then keyed into the database application. Today there are many research studies done on transcribing the data from the patient's medical record into a form [28].

Forms are used to collect patient data. These forms meet data collection requirements of the study. They are practical and easy to use, and contribute to the accuracy and timeliness of data collection.

Data/ Questions on the electronic forms must meet the needs of different people: 1) the form filler, who enters the patient's information and 2) the database designer, who designs the application to receive the data.

During the form design, one has to set the specifications for the fields in the database application: one field for each question and the fields will be organized into the tables of related information.

### 2.1.1.2 Missing Values

*(a) Missing values in data Entry:*

The data mining analysis mainly depends on the accuracy of the database and on the chosen sample data to be used for model training and testing. In real-world data, tuples (instances) with missing values for some attributes are a common occurrence. Missing values or incomplete data are becoming serious problems in many fields of research.

Mostly missing values occur during observations in many areas of research [7]. When data are collected by surveys and questionnaires, missing values occur due to entry problems or respondents refusing to answer questions. Capturing patient data in computer-based database places an important role for quality data mining results [20]. Appropriate data entry software should be used to reduce missing values and erroneous data during entry.

*(b) Missing values in existing data:*

Various missing value methods can be used for handling missing data for existing databases and for data left unknown during or not applicable during entry.

Methods of handling missing data in existing data are:
(1) Ignoring the instance: The record containing the missing value attribute is ignored/ omitted. This results in loss of a lot of information. There are two methods of doing this using listwise or pairwise deletion (Appendix E).
(2) Manual Replacement: Manually search for all missing values and replace them with appropriate values. Mostly, these are done when the replacing missing values are known.
(3) Using a global constant: Replacing all missing values with some constant like "unknown" or "?".
(4) Using attribute mean/mode: Replacing the missing values with mean or mode of non-missing values of that attribute or of same class.
(5) Using the most probable value: Replacing by the most probable value, using decision trees or Bayesian methods.
(6) Expectation maximization (EM) method: It proceeds in two steps. First step compute the expected value of the complete data record likelihood and the second step, substitute the missing values by the expected values, obtained from the first step, and maximize the likelihood function. These steps are continued until convergence is obtained.
(7) Multiple Imputation: This process creates data matrices, containing actual raw data values to fill the gaps in an existing database. These matrices are further analyzed and the results are combined into a single summary finding.

In today's market, there are several software that use various data mining tools for handling the missing values. This Thesis describes four software that can handle the missing values: WEKA, CRUISE, NORM and DTREG.

- WEKA: - WEKA is a software developed by Waikato University, New Zealand [16]. It is implemented in Java, and has implemented package classes that can be used.

These classes perform various machine learning algorithms that can be used for any data set. The data set can be preprocessed and fed into a learning scheme and can analyze the resulting classifier, with just using the classes. This software uses mean – mode method to handle the missing values, where the numeric missing values are replaced with their means and nominal values are replaced with their mode values.

- CRUISE: - CRUISE is a software used for the tree structured classification. It is developed by Hyunjoong Kim, and Wei-Yin Loh. CRUISE stands for "Classification Rule with Unbiased Interaction Selection and Estimation", improved from the algorithm FACT [21]. It has several algorithms for the construction of classification trees. Missing values in CRUISE are treated by global imputation or node wise imputation. There are mainly two kinds of node splits during tree construction for classification: univariate split and linear combination split. Univariate split handles the missing values by fitting (available cases) and imputing, whereas linear split has four choices of handling missing values: to fit (complete cases) and impute, for node wise imputation and fit, to fit (available cases) and impute, or for global imputation at root node and fit. These missing values are handled during the construction of the tree.
- NORM: - NORM uses multiple imputation method to handle the incomplete multivariate data, assuming multivariate normal distribution of the data. It is a software that performs the pre and post processing of data, but cannot handle any statistical methods for classification or regression [6]. This model generates imputations. There are two procedures that are performed in this model: EM and data augmentation (DA). The EM procedure estimates the means, variances and covariances (or correlations) of the variables, and the DA method (Appendix B) uses these values as the starting values and generate multiple imputations for the missing values. NORM can also perform the post processing like transformations and rounding and imputing all the variables in one file to be used for performing other statistical methods.
- DTREG: - DTREG is a tool that is used for modeling the data with categorical variables [33]. It is referred as "Decision Tree Regression". DTREG builds

classification and regression decision trees that can model data, their relationships and predict values for future observations. It uses two methods for handling the missing values: use surrogate splits method, or put rows in most probable group. In surrogate splits, if the target variable is missing, they are replaced by the surrogate splits. In this process, the association between the primary splitter and each alternative predictor is computed and ranked as surrogate variables. The highest association to the primary splitter that has no missing value for the row is used for the missing value. Whereas in the other case if the value of the splitting variable is missing, the row is put into the child group that has the greatest likelihood of receiving unknown cases.

There are other software programs that can perform various missing handling methods. Due to the authorization and cost constraints, the above four free software are considered for the analysis of data.

## 2.1.2 Data Mining

Data mining is among the most important tools that is used in the knowledge discovery process. It can be considered the heart of the KDD process. It is an analytical process that is designed to discover the hidden information from data warehouses.

Data mining is described as the "Use of algorithms to extract the information and patterns derived by the KDD process" [9].

Given a medical/healthcare database, Data mining can generate new medical improvements by providing:

- Automated prediction of trends and behavior: The process of finding predictive information in the databases. A typical example is predicting the diagnosis of a new patient. Data mining uses the past data and classifies the patient to the corresponding diagnosis.

- Automated discovery of previously unknown patterns: Data mining mines through databases and finds hidden patterns. An example of pattern discovery is error-detection, like finding fraud authorizations and inconsistent data in medical insurance.

Data mining tasks are categorized as predictive and descriptive:

Predictive:

- Classification
- Regression
- Time Series Analysis

Descriptive:

- Clustering
- Summarization
- Association rules
- Sequence Discovery

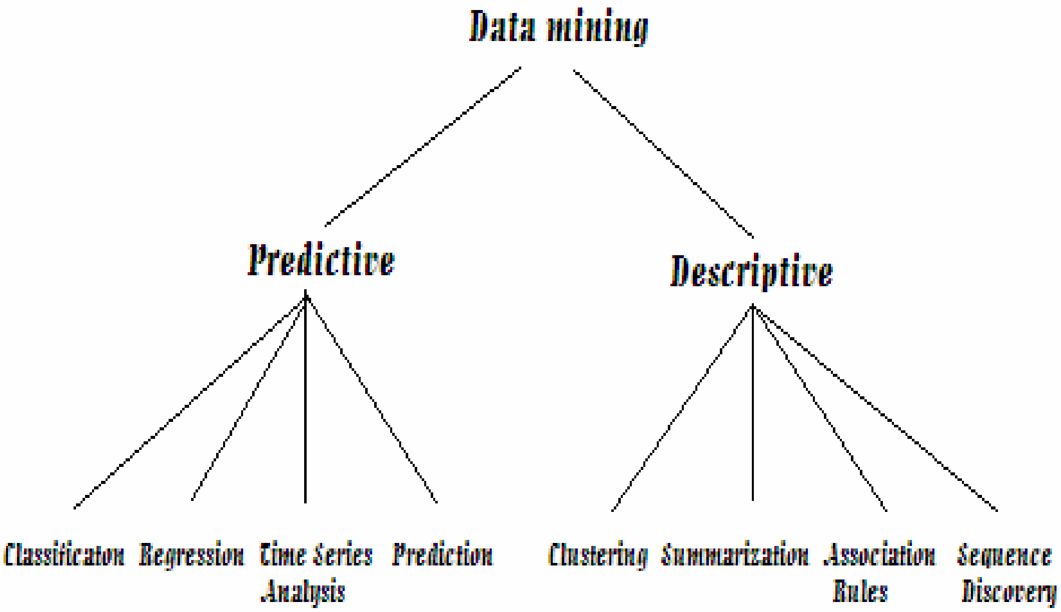


**Figure 3 Data mining models and tasks [9]**

Though they are many different data mining problems, three major problems are popular [9]:

- Classification learning: - It can be described as predictive modeling. The main goal is to induce a model from a training set and to predict (make decisions) the class of a new set. The classes are predefined in this process. It finds a rule or a formula from the data in the training set (Appendix E) for organizing data into classes. The reliability of the rule is then evaluated using the test set of data. For instance, classification of patient assessment for the MDS-MH data.
- Association learning: - The goal of this learning is to find trends across the large number of databases that can be used to understand and exploit natural patterns. It creates rules that describe how often events occurred together. For example, association rules generated for a super market, which is planning to

display juice on sale to find the frequently purchased products along with juice.

- Clustering: - Clustering breaks a large database into different subgroups or clusters. It's an art of finding groups in data. It is a descriptive technique that groups similar entities together and dissimilar entities into another group. The main drawback of this learning is it has no predefined classes. It is also termed as unsupervised learning. For instance, determining the target mailings of various catalogs bases on the customer interest and income.

The most commonly used techniques in data mining are:

- Decision trees [25]: Tree-shaped structure to predict the future trends. Each branch represents set of rules (decisions) for the classification of data. It is popular algorithm that is used for classification.
- Artificial neural networks [4]: These are non-linear predictive models that learn through training and predict new observations from such learning. They are represented as network architectures.
- Nearest neighbor method [9]: To classify a new case, the algorithm computes the nearest distance from the new case to each case in the training data. It can easily build class models for predication of data.
- Rule Induction [9]: It is the method of discovering knowledge by inducing rules (If-then rules) about the data. This process is in unstructured format and difficult to maintain.

We describe the decision tree method in detail in the following section.

### 2.1.2.1 Decision Trees

There are several data mining tools available, but only few are fit to solve the problems considered here. Depending on the algorithms used the results may vary. Most of the software today, like WEKA, CRUISE and DTREG that are used in this Thesis, uses decision trees for classification to yield better results.

Decision trees can be the fastest and easiest of the data mining algorithms that are used to solve the task of classifying the cases into classes. This algorithm is used in many applications from predicting buyers/non-buyers in database marketing to automatically diagnosing patients in the medical field.

This Thesis considers only decision trees for the classification due to the following reasons:

1. Decision trees are classifiable: The MDS-MH case study used in this Thesis is multiple class classification problem.
2. Decision trees yields good results for categorical data: Data attributes of the MDS-MH are categorical.
3. Simplicity and interpretability features of decision trees to produce trees.

When there exists a model which needs to make a decision based on several factors, a decision tree suits in helping to identify the factors and provide various outcomes of the decision. Decision tree model results are presented in tree structured format (see Figure [4]), i.e. it graphically represents the relationships in data. It can be both a predictive and descriptive model is used for classification as well as regression tasks.

Decision trees are used to classify the cases/instances by categorizing them down the tree from root node to leaf node. Each node represents attributes of the test cases and its leaf node represents the class values of the observation case. The decision rule is split on the attribute value. This process is iterative which starts from classifying the case from root node of the tree and testing the rules specified by the node and moving down the tree branch to the corresponding test node value. It repeats the same process at the branch node (which becomes root node) until the final leaf node is reached. The test cases on each node are mostly If-Then rules, but some uses separate and conquer strategy. Decision trees can work on any type of data and the main target attribute is Boolean (yes/no) or categorical type of data.

For instance, consider a decision tree used to classify for a patient's risk depending on blood pressure rate.

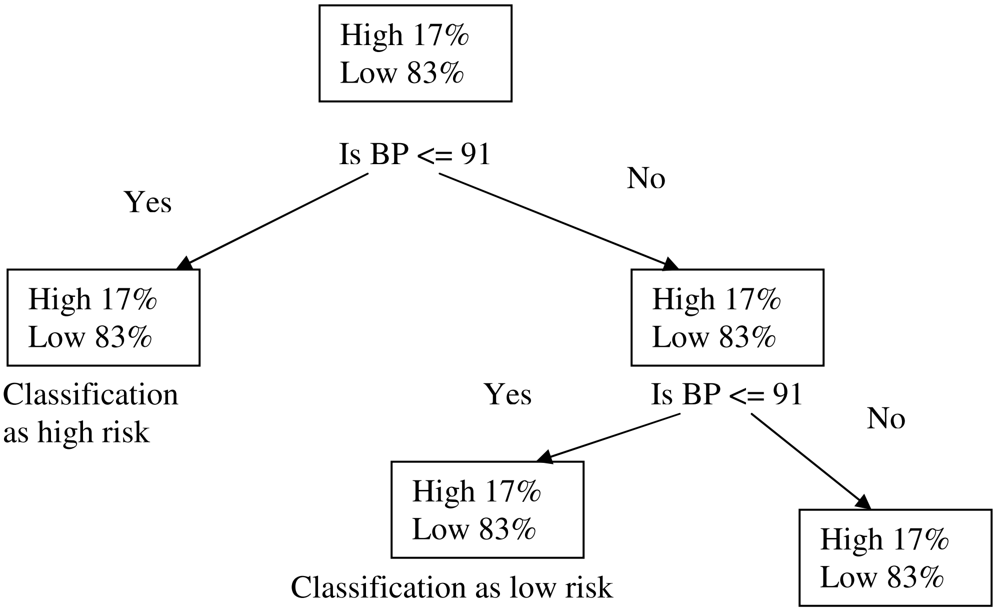


**Figure 4 Decision tree for a patient's classification of risk [39]**

The tree is build for classifying the patient's risk. The splitting predicates for BP are {=Yes, =No}. The tree is built in a top-down fashion by examining the training data. The higher splitting predicates are based on whether the patient blood pressure rate is less than or equal to 91. The right most leaf node can be further build.

There are a number of algorithms that are based on decision trees. Some of the most common and effectives algorithms are CART, FACT and C4.5. FACT is based on CART. In the tools that are used in this Thesis, WEKA is based on the C4.5 learning algorithm and CRUISE is advanced version of FACT. The C4.5 is the upgraded version of the basic ID3 algorithm [37].

Once the construction of the tree is done it can be used for predicting a new case starting at the root node and following the path of the tree structure to the leaf node. The main important concern of this model is finding when to stop the growth of the tree and what the maximum depth of the tree. Tree pruning (Appendix F) is considered into account for attaining quality predicting results, i.e. once the tree is know, before performing the data mining task only necessary data is considered to avoid over fitting. A tree is grown to learn the training data where as pruned to avoid over fitting the data (Appendix F).

## 2.2 Health Informatics

Health care is an enormous part of a country's economy. In general, health care is referred to the delivery of medical services provided by doctors, nurses and allied health professionals. Health information, especially clinical information, proliferates on a daily basis and is extremely variable and difficult to assess. As a result, there is a need for finding criteria that can be used to evaluate the quality of the hidden information. Thus, focus on using computers to maintain the information has begun.

The use of computers in health care started in early 1970's [19]. The initial use of computers focused on administrative process like hospital billing, financial applications and physician billing. Later, computers took over from paper based medical records in the processing of organizing, storing, integrating and retrieving medical and patient information.

Clinical information about a patient derives from a variety of sources, like the patient, the attending physician, consultants, laboratory etc. Paper-based systems are inefficient for managing enormous amount of medical and patient information that can affect patient care. For example, physicians must learn and retain tremendous information about antibiotics, appropriate dose and frequency for the patients. If the medical record is hand written and poorly organized, it is difficult for the physicians to locate the information they need.

Computers play an important role in eliminating medical errors in the development of medical alerts and protocols that make health care professionals aware of the potential for a medical error. Computer technology also reduces the number of mortality and waiting time to see the physician as well as the cost and time during registration process.

Evolution of data mining in health applications helps the professionals in making clinical decisions. Medical decision-support systems deal with medical data and knowledge domain in diagnosing patient's conditions as well as recommending suitable treatments for the particular patients.  Several data mining techniques and algorithms are used to mine quality data in clinical databases. Also different data preprocessing techniques are applied during the data inputting stage for better mining results.

The informatics part of health care can take care of the structuring; searching, organizing and decision making with the emergence in health informatics came many important research ideas and fields of study. One among them is the Resident assessment instrument (RAI) [13].

## 2.2.1 Resident Assessment Instrument (RAI):

The RAI is a comprehensive and multidisciplinary mental health assessment system that is used for assisting adult facilities, mainly long-term residential facilities [13]. These assessments information is used to gather an entry resident's psychiatric, social, environmental and medical information. Assessment forms are basically structured in the form of questionnaires. This information is used to provide acute and long term care of a resident.

The forms used by these RAI systems are termed as the minimum data set (MDS) which can be considered as the minimum number of questions that are required to make a proper diagnosis of a patient with respect to certain problem. The data in the forms can be updated with respect to the patient’s health and gradually improve the care that is provided to the patient.

Gathering and updating the information will provide better treatment for the patient. The end result after using these forms is to produce quality of resident care and at the same time to promote quality of the resident’s life. The information of a patient can be gathered from the patient or a person representing the patient.

The RAI consists of three basic components:

- Minimum data set (MDS)
- Resident Assessment protocols (RAP).
- Utilization Guidelines

MDS can be considered, as an accurate and complete documentation of a patient with long term needs. These are questionnaire kind of forms that are used for classification or categorization of a new patient. The patients’ needs with respect to care, problems and conditions of medication are mentioned within this documentation.

In this Thesis, we are concentrating on the MDS-MH assessment data for psychiatric patient that is obtained from RAI-MH, where MH refers to mental health. The assessment form deals with all the information that is required to give proper assessment of patients with long term mental problems. The assessment information will give information regarding which of the four categories will a patient be admitted looking at the various attributes in the assessment form. The four categories of patient classification are the following:

- Acute Care
- Longer term patient
- Forensic patient
- Psychogeriatric patient

The RAI-MH has data obtained from 43 hospitals with around 4000 patients of cases are being considered as a case study. There are 455 attributes that will be used for the classification of patients into the four major categories in mental healthcare.

Some of the sections that are present in the minimum data set for mental health (MDS-MH) are

- Name and identification numbers
- Referral items
- Mental health service history
- Assessment information
- Metal state indicators
- Substance use and excessive behavior
- Harm to self and others
- Behavior disturbance
- Self care
- Medications
- Health conditions and possible medication side effects
- Service utilization and Treatment

The advantage of the MDS-MH is some of the patient attributes are based on time series of data. Each and every attribute is related to another. Thus we can refer to an attribute of importance to the Clinician over a particular period of time to check on the improvements and changes that need to be made with respect to the patients care.

## 2.3 Summary

This chapter presented a brief literature review of the different components that are required for the architecture of the system. It also presents an overview of the different components, such as the MDS-MH and machine learning. With the background knowledge of data preprocessing, missing values and form design we can show how MDS-MH are converted to electronic forms while reducing erroneous data for better quality data mining results. The next chapter is more focused on the detailed design of the E-Intelligence form and different types of the handling missing values methods.

# Chapter 3

# System Architecture and Model

The goal of this Thesis is to minimize erroneous data and handle missing values during different stages in data mining.  The main task is to reduce the missing values during the entry level of data, before the data is stored in data warehouses. Various data mining and fuzzy logic concepts can be used to reduce the noise in data. This chapter discusses the architecture and model of the E-Intelligence form design, and various handling missing values methods that can be performed on existing databases.

## 3.1 System Architecture

The Thesis concentrates on the data preprocessing task of the KDD process. Two main problems addressed here are:

1. Handling data during entry step
2. Handling missing values for the existing databases

The system architecture of the E-Intelligence form and the support software are shown in Figure 4:

1) GUI
2) Controller
3) Web/PDA application

The three components are discussed in further detail in the subsequent sections.

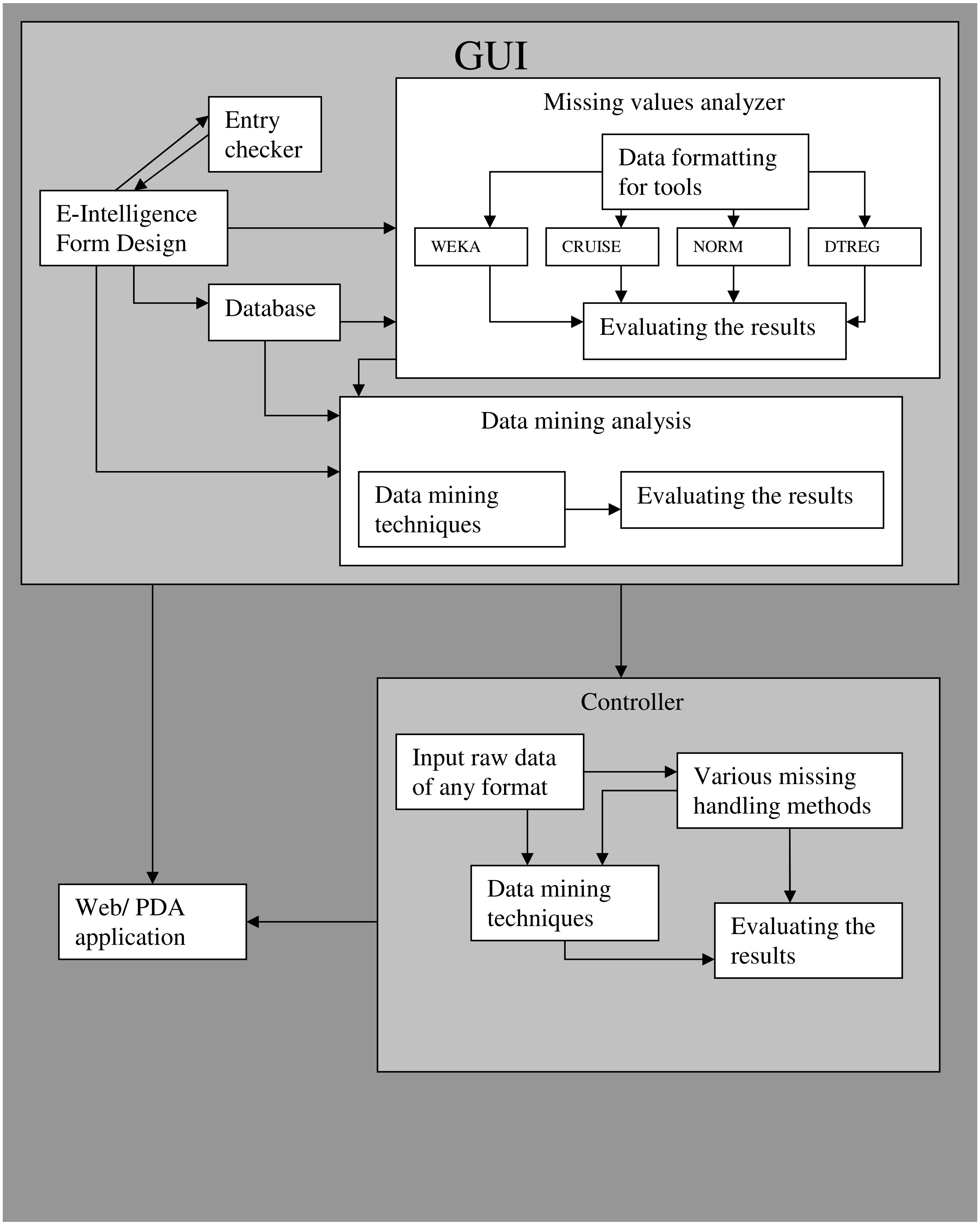


**Figure 5 Architecture of the E-Intelligence form design**

## 3.1.1 GUI

The GUI is the main component of the overall architecture, which performs various tasks of the KDD process from data preprocessing to data mining.

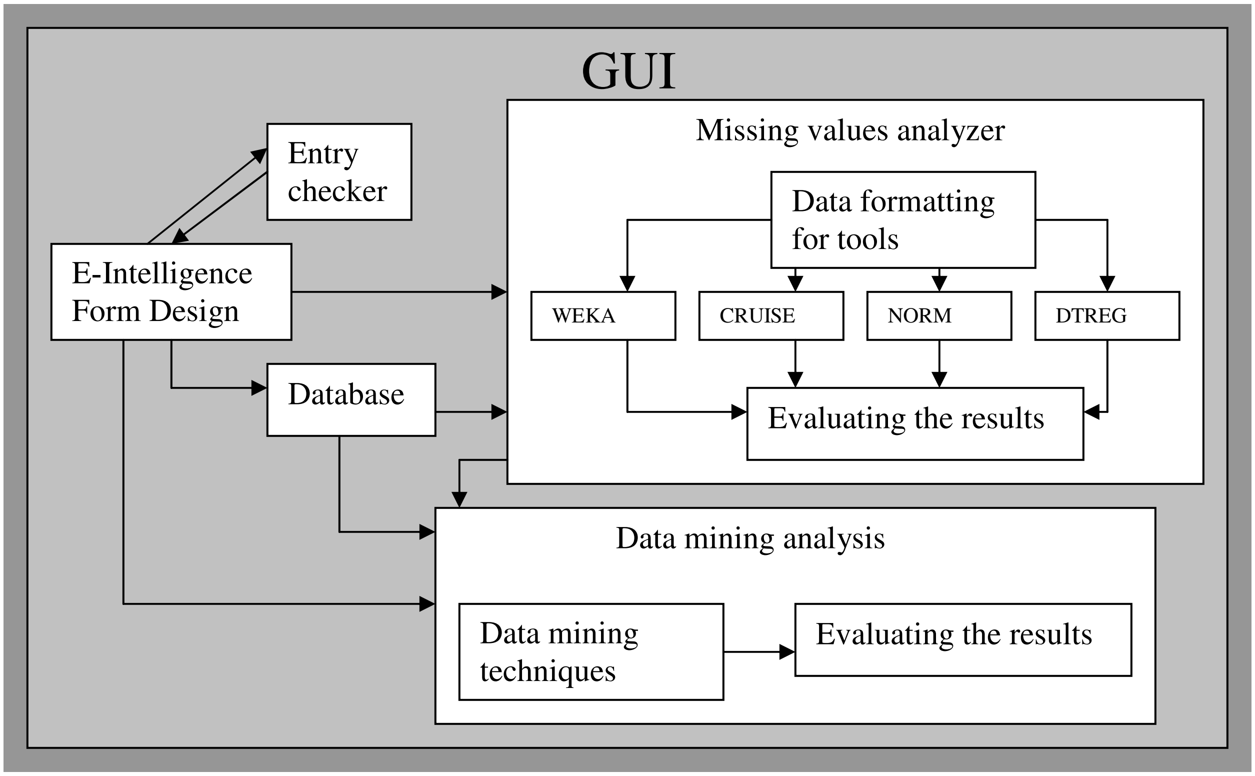


**Figure 6 Architecture of the GUI**

The software design of the E-Intelligence form consists of the following components:

1. E-Intelligence form design: This component creates the electronic forms that can be designed for any questionnaire type of data forms which have some intelligence (finding missing values, reducing errors, and consistency) with each field.
2. Entry checker/ logic layer: This is the component where the logic to handle the data is used to reduce the missing values during entry level. The data is made clean as

possible by avoiding the noisy data. Logic is applied to check the consistency between the attributes for accuracy. This is domain dependent. It can also be accessed for the offline data, that is existing data to reduce the error data, duplicate records and inconsistent data.

3. Database layer: This component is where the data is stored and maintained in the database for future usage. Before the data is stored in the database, the entry checker checks for the duplicate records using various constraints and conditions and, if any exist, it eliminates them. The data now stored is error free and only has unknown values.
4. Missing values analyzer: This is the component that performs various missing data handling methods on the existing raw data. Software like WEKA, CRUISE, NORM and DTREG are used to perform these missing values tasks. This is an offline process, which can be performed on the existing data. It is not performed on the online data during the entry step.
5. Data mining analysis: This component can be used for performing various data mining techniques like classification, regression etc to analyze the data. Decision tree technique is used to classify the data sets in this Thesis. This is also an offline, post-collecting step.

The flow of the software starts at the E-Intelligence form. During form entry, data preprocessing tasks are performed like reducing missing values and erroneous data by the Entry checker before the data is stored in the database. After data is stored in the database, once again the data preprocessing tasks can be performed and missing values, if any, are replaced using different tools. Later, any data mining technique can be applied for analyzing the data and the results obtained can be evaluated with the results of the data having missing values or erroneous data.

The model can also perform the missing values handling methods on the existing raw data sets. The E-Intelligence form has an option of selecting the missing values analyzer to perform this task. It can also perform only data mining techniques for the existing data sets.

The flow of the software is shown in the Figure [5]. The flow can change from performing data preprocessing tasks and then applying data mining techniques to only applying the data mining techniques. The results of both the cases can be compared for the accuracy and efficiency of analyzer results.

Various data mining techniques and missing value handling methods use different formats of data. The main idea of this Thesis is to develop a user-interface for quick and better results. This software can help the user in selecting the raw data sets of any format and then apply different missing values replacing techniques and check for accuracy of the results. The results are output to the user in the format that is understandable and shows the replaced missing attribute values. The user can also have the option of evaluating the results from the analysis of data that have missing values. Once the data preprocessing is done, various data mining techniques can be applied and the results can be displayed to the user. The software used in this Thesis performs all these tasks but cannot provide the results or analyze the results obtained form different software in one single file.

### 3.1.2 User-Interface Controller

This is the second component of the architecture. This is a user-interface, which is mainly designed for the easiest and quickest access of data and different tasks like handling missing value tasks and data mining tasks.

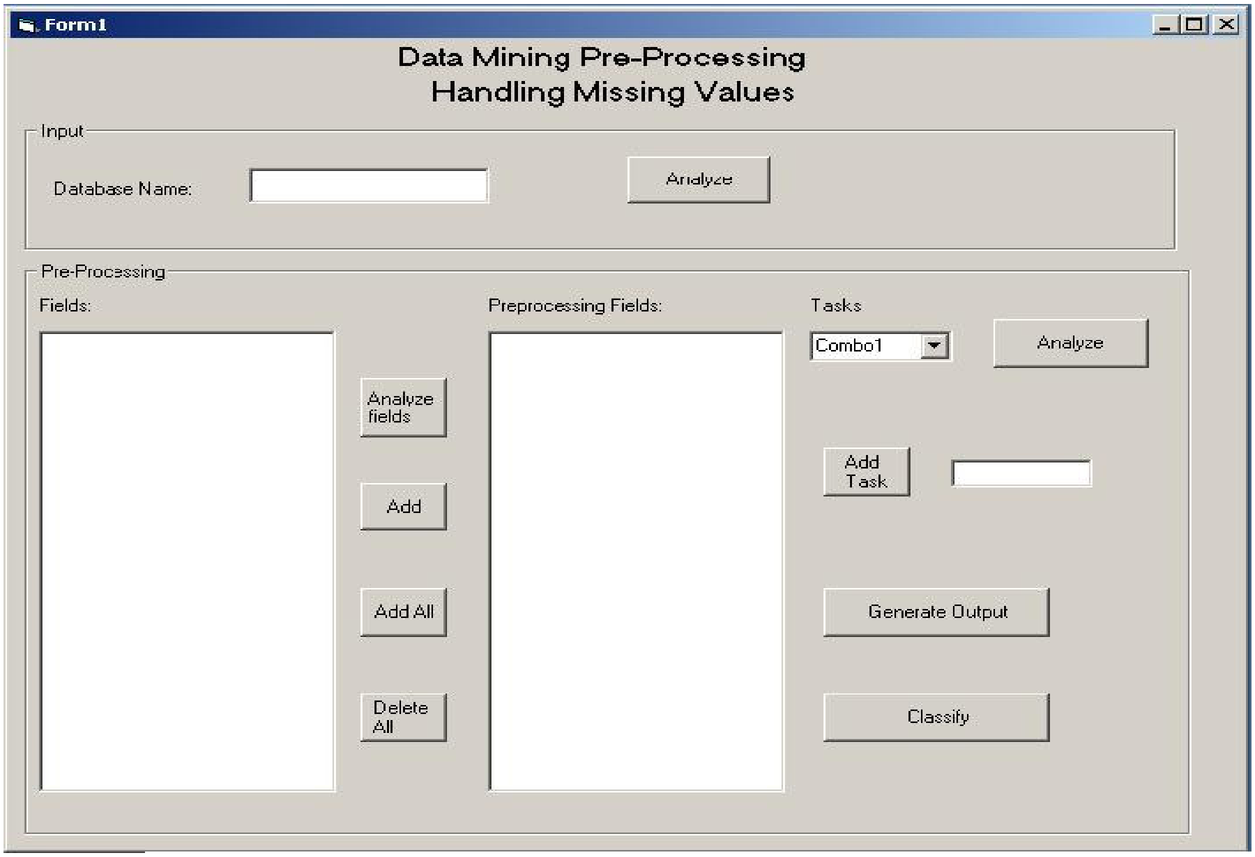


**Figure 7 User-interface controller**

The above user-interface controller can be activated from the E-Intelligence software and is designed for handling missing values using various techniques. The user can enter the raw data of any format in the database name input box for performing the data mining. The selected database is analyzed and the attributes data types are automatically listed in the preprocessing fields frame. Once the fields are selected, the user can perform different missing values tasks. Before performing the tasks, the fields should be formatted respectively to perform the corresponding task. Once all the fields have performed the missing values task and corrected data is ready, the new output for the database can be generated. These results also indicate which attribute values are replaced and by what value. The results obtained can be checked and evaluated after performing the data mining classification on the new data.

The user also has the option of adding new algorithms for handling missing values to perform on the selected data. Due to the complexity of databases for analyzing the replaced missing values, and as the software used in this Thesis are restricted for database sizes, and due to time constraints, this system has not been yet implemented. But different missing value methods are discussed and explained using the breast cancer data as the case study and a subset of MDS-MH data is considered for checking the replaced missed values accuracy of the software by eliminating some of the existing attribute values.

### 3.1.3 Web/PDA Application

The effectiveness of mobile and point-of-care decision support indicates the significant impact that personal computers and handheld computers have on the daily activities of different medical specialists. The study, conducted by Skyscape, shows that more than 50 percent of medical professionals indicated PDA use reduces their medical errors by more than 4 percent. Considering, the fast growing new technologies, further, this Thesis model can be moved towards the development of the software on a mobile computing device like a PDA. Once the software is developed on a personal computer it can be installed on PDA for quicker and convenient access.

Due to the main drawbacks of this device, namely, low memory and low computation power, it is not advised to store all the data and the data mining techniques. Another approach is to run the techniques on the PC and save the results on the PDA.

Thus, instead of storing all the data, data preprocessing tasks (missing values handling methods) and the data mining algorithms are run on the computers and only the results or the rule set for classification or missing data are saved on the PDA. We then input the data directly on the PDA and the inputs will be tested with the rule set and providing the answers. With the help of Internet service we can send the data to the server where computation can take place and output the results from the server. Due to the increasing usability of Internet, this software can also be implemented on the Web. The programs are similar for both the models. Next section describes the form design.

Flowchart of PDA and Server side program:

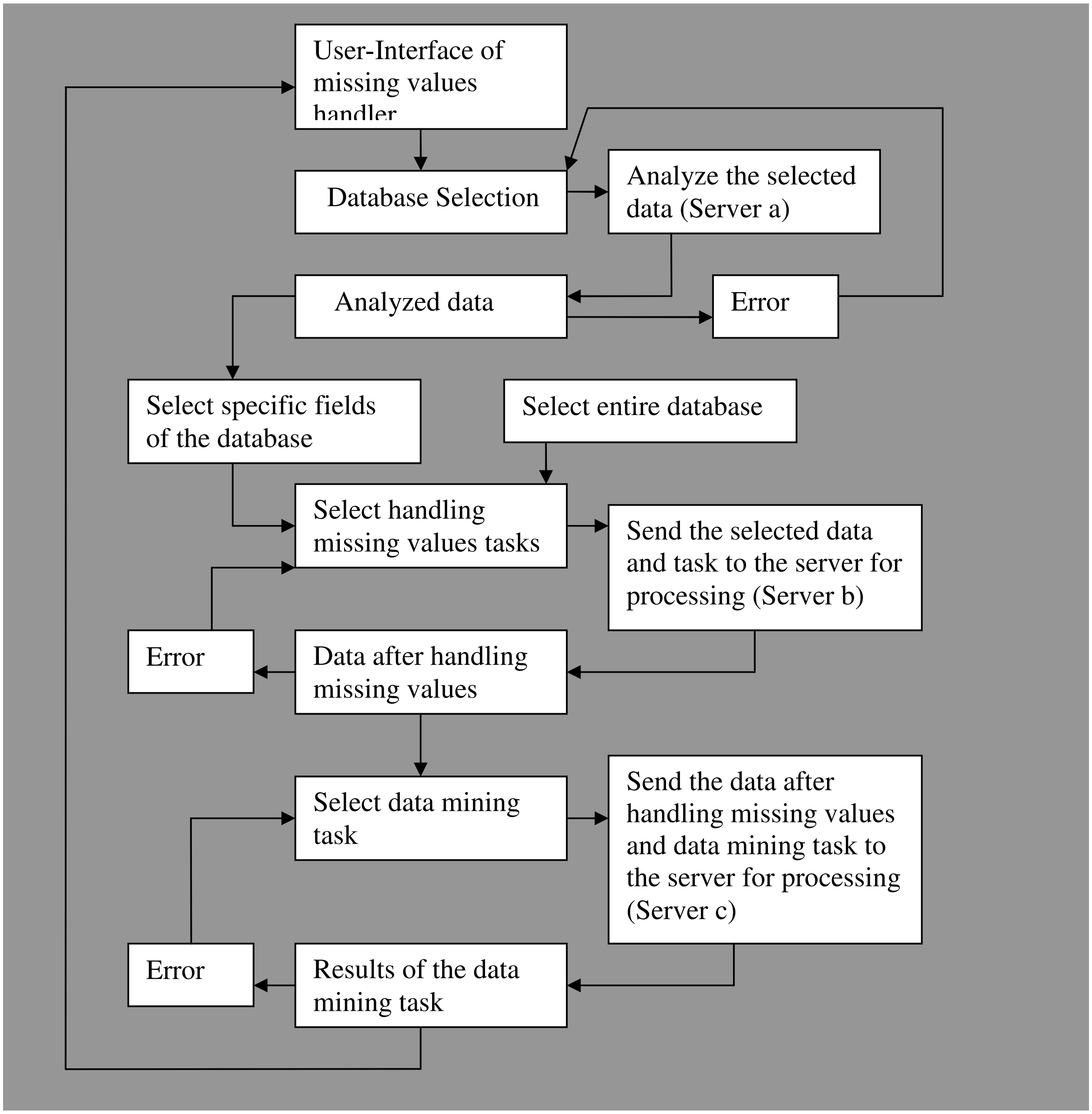

User-Interface of missing values handler
Database Selection
Analyze the selected data (Server a)
Analyzed data
Error
Select specific fields of the database
Select entire database
Select handling missing values tasks
Send the selected data and task to the server for processing (Server b)
Error
Data after handling missing values
Select data mining task
Send the data after handling missing values and data mining task to the server for processing (Server c)
Error
Results of the data mining task

**Figure 8 Flowchart of PDA program**

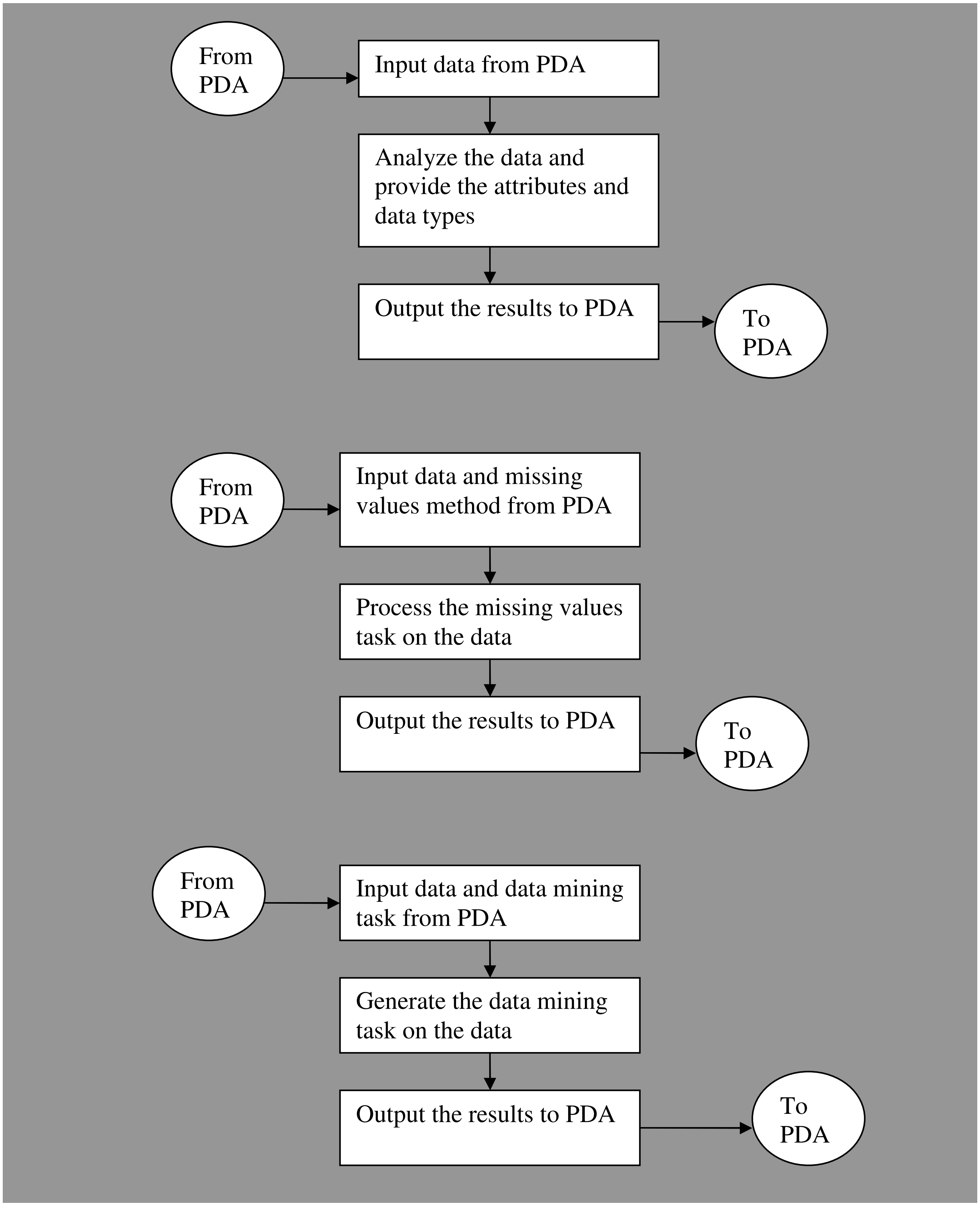

From PDA
Input data from PDA
Analyze the data and provide the attributes and data types
Output the results to PDA
To PDA
From PDA
Input data and missing values method from PDA
Process the missing values task on the data
Output the results to PDA
To PDA
From PDA
Input data and data mining task from PDA
Generate the data mining task on the data
Output the results to PDA
To PDA

**Figure 9 Flowchart of server program**

## 3.2 E-Intelligence Form Design

Clinical data consists of large quantities of information about patients and their medical conditions. Accessing this data by various people for further improvement of treatment is most necessary on a daily basis. For example, if a physician wants to know what kind of medication the current patient is taking or how much dosage the patient is taking, he/she has to access the previous clinical records of the patient. With technology advancements, computer-based patient record software makes the collection and access of health care data more manageable. For further discovery of trends and patterns hidden within the data, in improving the medical treatment and providing quality health care, analyzing and evaluating the stored data is essential.

There are several steps to be taken for maintaining the patient record software. The most important cases are entering the patient data and storing the data in the databases. This Thesis mainly concentrates on the handling the data with form design. The underlying concept of this application design is to reduce errors and make sure the data has no missing values, apart from the cases of unknown data, during data entry. The logic used in this design can be applied for any database application that has need for some underlying intelligence between the attributes.

Form design plays an important role in the medical field. A range of forms have been designed and used on a daily basis in various medical departments like nursery, laboratory, pharmacy, and clinical treatment. Several electronic forms are designed that satisfy the needs of the medical assistants. These forms perform tasks like storing, retrieving and modifying the data. Some of these also check for the attribute data types and eliminates duplicate records. But most of the forms cannot handle the technique of learning forms, reducing missing values, noise in the data and finding attribute consistencies. The E-Intelligence form created uses intelligence in handling data, by finding erroneous data and inconsistent data.

Several software tools like JAVA, C++, HTML, Developer 2000, VB, VB.Net, etc are studied for the design and implementation of the data set. Visual Basic is selected for the application design as it was easy to use, design and implement.

The E-Intelligence form is designed to reduce the missing values and noisy data during data entry. Duplicate records and consistency between attributes are also checked before the data is stored in databases. This form is used for storing, and retrieving the patient's data. Patient's data is maintained using the case record number, which acts as the primary key for the database. Patient's data can also be retrieved for viewing or editing the details by one of the attributes: case record number, health card number or patient's name. This form also has the option of selecting various missing values handling methods or data methods to perform on the databases of any format.

The underlying logic of the E-Intelligence form design can be used to any questionnaire kind of data. The E-Intelligence form commutes with the Entry checker layer for handling the data. After each and every data entry value, the entry checker checks for the accuracy. Various cases are considered to reduce the noise in the data and missing values before the data is stored in the databases. For instance, during the data entry of birthdate attribute of a patient, it checks if the patient's age is accurate like checking if the day and month are consistent. It also checks for the year, for instance, a patient who is born in the same present year cannot be a right age since considering this form for a MDS-MH data, a newly born child cannot be admitted in a mental assessment care. Other tasks like eliminating the duplicate records and reducing the errors in data by checking the consistency between the attributes are also implemented in this layer.

The E-Intelligence form can also transfer the process to the controller, which performs all the tasks in a single user-interface like selecting the existing database of any format, applying various missing value handling algorithms or data mining algorithms and analyzing the results. The user also has the option of adding new algorithms to perform on the data. Due to the increasing need of internet and mobile devices in the medicine field for acquiring good and quality health care, this user-interface controller can be developed as a web application or can be fed into a PDA.

## 3.3 Entry Checker

Every attribute of each record should contain valid and meaningful information. In the real world, we cannot assume that source data will be complete. In order to avoid the problem of missing dimension key values, it is necessary to consider the starting place where data could be missed namely, data entry.

There are mainly two kinds of missing values: user defined missing values and system missing values. User defined missing values are values that are indeed missing or that does not apply in particular cases. System missing values occur when no value can be obtained for a variable during data transformations. Most of the missing values occur due to missed entry or due to unknown data. Every survey will have missing data for some questions, but too many missing data leads to poor quality data mining.

In most occasions of missing value: some parts of data may be missing. For example, patient's name, gender, birthdate etc. Here the application design is implemented to handle missing values during data entry. The Entry checker level handles the data preprocessing task for the E-Intelligence form design. During the entry of the form each new patient has unique record number, primary key for which the value is entered directly and each attribute's value is checked before they are stored in the database. Another case considered during data entry is eliminating noisy data. Entering wrong data in the databases misleads the data mining results.

There is a two-way process between the E-Intelligence form and the Entry checker layer. Each and every data entry is checked for accuracy by the Entry checker and if any value is found erroneous, a message is sent to the user indicating the wrong entry and immediate action is taken. Until the data value entered for the particular attribute is correct, the user cannot access or enter to the next field. For attributes where the value is compulsory say name of the patient, date of birth etc., the user cannot pass to the next question until correct values are entered for these fields. The only data values that are missing in these cases are unknown values.

The Entry checker layer also has fuzzy logic concept to check the consistency between attributes. The main usage of this E-Intelligence form design is to reduce the noisy data. Various rule based statements are coded to perform this consistency check and the accuracy of the attribute values is analyzed. For instance, the birthdate attribute and the admission date attribute are checked for consistency. A patient's birthdate and admission date cannot be similar incase of a mental assessment care form. Before the data is stored in the databases, Entry checker also handles the duplicate records. Duplicate records mislead data mining results.

Other preprocessing tasks like transformation and reduction of data can also be done in this layer. The user-interface of E-Intelligence form has a selection feature to select these tasks. The user inputs the data and fields that need to be transformed or reduced to the Entry checker, which perform the corresponding tasks. If the tasks are done correctly, the user is prompted with a message of successful task performance else the corresponding error is messaged to the user, to take action. The results of this data can be saved in the database as different data file or can update the existing file.

## 3.4 Missing Values Handling Software

Most of the data mining techniques ignore the records with missing values. Instead, using efficient methods to fill these missing values will extend the applicability in terms of accuracy for many data mining methods. The accuracy of the tools will be increased by a larger training set, leading to betters rules and decision trees which will contribute improved results in terms of classification of the data set.

Each data mining algorithms requires the data to be submitted in a specified format. The generation of raw data into machine understandable format can also be termed as preprocessing of the data.

- *Machine understandable format*

Raw data is usually stored in ASCII, Excel ® or other database types of files. There are times when the raw data does not process any format. Raw data cannot be used directly for processing, with most data mining algorithms. They first need to be processed into a machine understandable format. Some of the data mining algorithms require attributes that are separated by comma; others might require attributes to be separated by space etc. Thus usually the set of data must be made into different formats for the tools to understand the data. This is termed as the machine understandable format.

Having data already in a format understandable by the algorithms can result in improvement in efficiency. In most cases the rows represent a single case and columns represent the attributes that are present within this case. In some of the databases that are available online, most of them are in the CSV format where all the attributes are separated by commas and two commas simultaneously stands for a missing data attribute. Sometimes when attributes are missing, we will find a question mark in place of the missing attribute instead of finding an empty space.

This section looks into a few missing value handling methods like mean-mode, complete cases, available cases, multiple imputation, and surrogate splits, for the cases where the data already exists but have missing values. To evaluate these methods, WEKA, CRUISE, NORM and DTREG software are considered and were tested with the breast cancer data from the University of Wisconsin [40].

The attribute and their data types of the breast cancer data are given below:

| # | Attribute | Domain |
|---|---|---|
| 1. | Sample code number | id number |
| 2. | Clump Thickness | 1 - 10 |
| 3. | Uniformity of Cell Size | 1 - 10 |
| 4. | Uniformity of Cell Shape | 1 - 10 |
| 5. | Marginal Adhesion | 1 - 10 |
| 6. | Single Epithelial Cell Size | 1 - 10 |
| 7. | Bare Nuclei | 1 - 10 |
| 8. | Bland Chromatin | 1 - 10 |
| 9. | Normal Nucleoli | 1 - 10 |
| 10. | Mitoses | 1 - 10 |
| 11. | Class: | (2 for benign, 4 for malignant) |

Examples of a few cases in the data set are as mentioned below.

```
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2
1016277,6,8,8,1,3,4,3,7,1,2
1017023,4,1,1,3,2,1,3,1,1,2
1017122,8,10,10,8,7,10,9,7,1,4
```

In the database the attribute “ID number” will not contribute any information towards the machine learning in determining whether the person has cancer or not so from hence forth that column will be removed from all the cases within the database. The Wisconsin database consists of 699 cases. Using the below software, missing values in the data are handled and then classified as ‘a’ for benign or ‘b’ for malignant.

Tools that will be tested and the corresponding missing values handling methods are listed below:

1. WEKA
    a. Mean – Mode method
2. CRUISE
    a. Fit (complete cases) and impute
    b. Nodewise imputation
    c. Fit (available cases) and impute
    d. Global imputation at root node
3. NORM
    a. Multiple Imputation
4. DTREG
    a. Surrogate splits

The above software handles the missing values during data preprocessing stage and data mining stage. WEKA and NORM handles the missing values before processing the data mining task, while CRUISE and DTREG handle the missing values during the data mining processing, building the decision trees for classifying the diagnosis/patient assessment. WEKA can also handle the missing values using J4.8.

## 3.4.1 WEKA Software

WEKA system, developed by University of Waikato in New Zealand is used to implement the data mining algorithms. The system is written in Java, an object oriented language. Java provides a uniform interface to many different learning algorithms, along with methods of pre and post processing and for evaluating the results of learning algorithms on any given data set. Using WEKA system we can process the data set, feed into a learning scheme and analyze the resulting classifier and its performance.

The most common method of filling the attributes efficiently without too much computation is replacing all the missing values with the mean or the mode of that attribute. WEKA includes a predesigned algorithm, ReplaceMissingValuesFilter class, for handling the missing values. This uses mean or mode method of handling missing values. It replaces the missing values with constants. For numeric/continuous attributes the constant is the attribute’s mean value and for nominal/categorical ones, its mode. This algorithm overwrites the three main methods defined in the Filter: inputformat (), input () and batchFinished ().

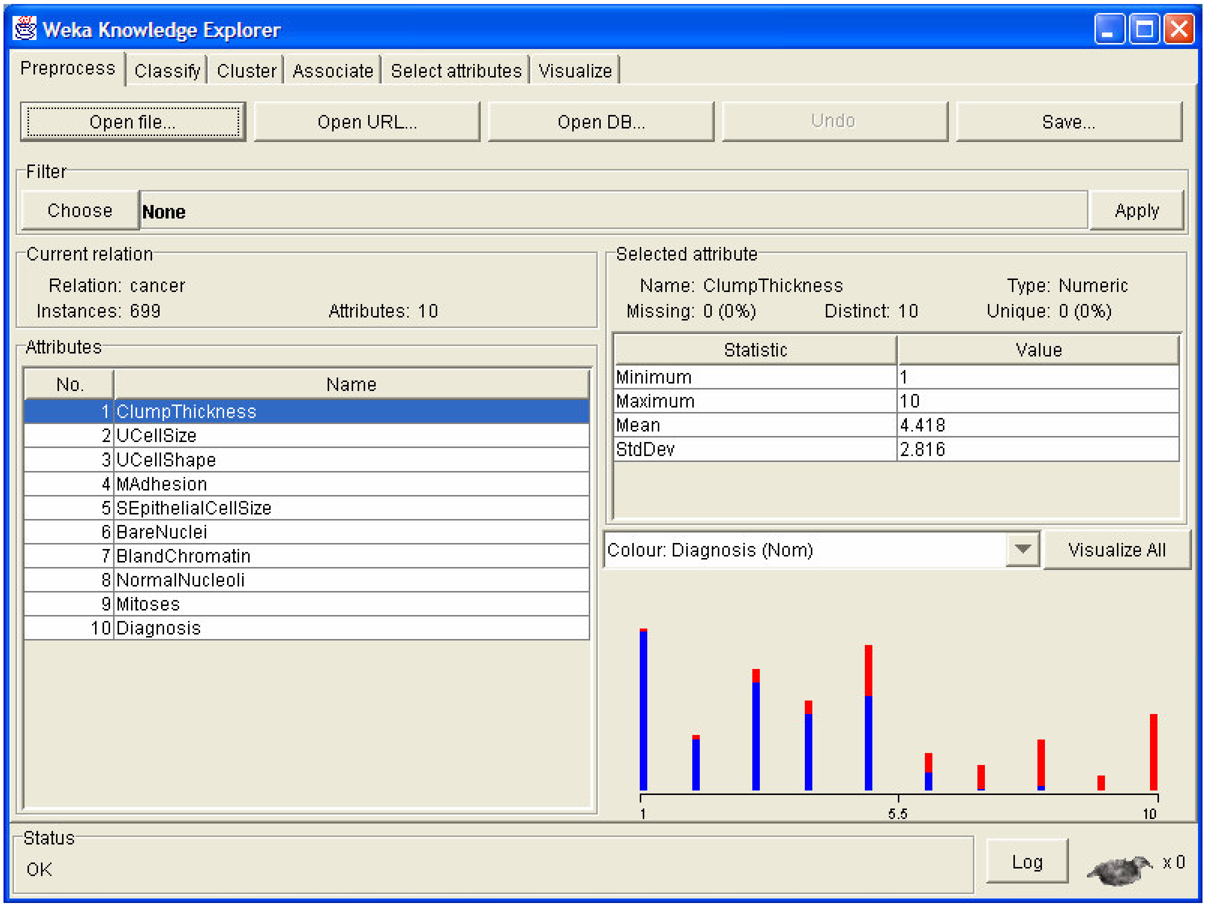
WEKA software provides various data mining techniques for classification, clustering, and association. This Thesis focuses on classification technique. J4.8 decision tree algorithm, in analogous to C4.5 is used to implement the classification. C4.5 has ad-hoc procedures built in to handle the missing values. CART uses surrogate splits whereas C4.5 uses fractional cases. The latter handles the data with missing values by replacing them with the most common or the most probable value.

Most of the data mining tools consider data in the CSV format for running the data mining algorithms. The data that is used for WEKA should be made into the following format shown in the table below and the file will have the extension dot ARFF (.arff). The last attribute where the classification of the patient is done is made into a categorical format that is the classification attribute ‘diagnosis’ takes string values ‘a’ when tumor is benign and ‘b’ when tumor is malignant. The missing values are replaced by ‘?’ mark.

```
@relation 'cancer'
@attribute 'ClumpThickness' real
@attribute 'UCellSize' real
@attribute 'UCellShape' real
```

```
@attribute 'MAdhesion' real
@attribute 'SEpithelialCellSize' real
@attribute 'BareNuclei' real
@attribute 'BlandChromatin' real
@attribute 'NormalNucleoli' real
@attribute 'Mitoses' real
@attribute 'Diagnosis' {'a','b'}
@data

35,1,4,120,198,0,1,130,1,1.6,2,1,3,a
43,1,4,120,177,0,3,120,1,2.5,2,1,3,a
62,0,4,160,164,0,3,145,0,6.2,3,4,3,a
```

Once the breast cancer data is transferred to .arff format, it is loaded into WEKA. The WEKA explorer is shown in Figure [10]. First the data is preprocessed and then the learning algorithm is applied for classification.



None

Relation: cancer
Instances: 699
Attributes: 10

Name: ClumpThickness
Type: Numeric
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)

| No. | Name |
|---|---|
| 1 | ClumpThickness |
| 2 | UCellSize |
| 3 | UCellShape |
| 4 | MAdhesion |
| 5 | SEpithelialCellSize |
| 6 | BareNuclei |
| 7 | BlandChromatin |
| 8 | NormalNucleoli |
| 9 | Mitoses |
| 10 | Diagnosis |

| Statistic | Value |
|---|---|
| Minimum | 1 |
| Maximum | 10 |
| Mean | 4.418 |
| StdDev | 2.816 |

Colour: Diagnosis (Nom)

OK

**Figure 10 WEKA software**

As shown in Figure [10], the attributes of the data set are displayed in row format on the left hand side of the screen and the data type and the number of missing values for each attribute along with the graphs that represents the attribute distributions are displayed on the right hand side.

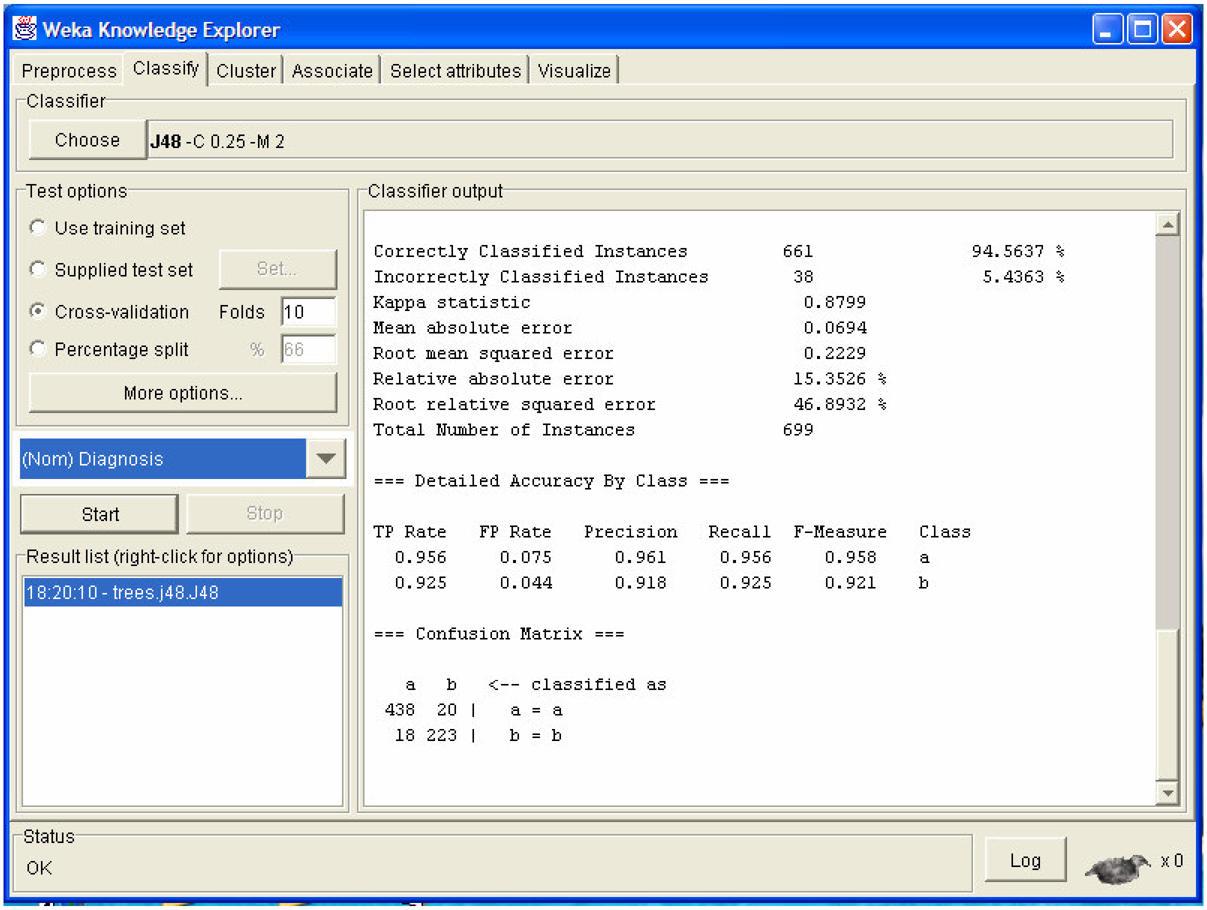
Here only one set of data is considered, that is training data set. Once the data set is loaded, users can select the preprocessing tasks. In WEKA, the preprocessing tools are located under the filter package. WEKA uses the mean – mode method of handling missing values, which is located under the filter package hierarchy: unsupervised->attribute->ReplaceMissingValues. Once the preprocessing tool is chosen it can be applied on the data set and the results can be saved and also viewed in the WEKA explorer.

The data set of breast cancer database has only 16 missing values (2%) for BareNuclei attribute (numeric data type), where 14 cases are of class ‘a’ and the other two cases are of class ‘b’. These missing values can be handled using the ReplaceMissingValues tool. Once the missing values are handled, the resulting data set show zero missing values for the BareNuclei attribute.

Experiments for the breast cancer case study without handling the missing values and after handling the missing values are shown as follows:

1. J4.8 ad-hoc procedure:

   Figure [11] shows the classifying results of the breast cancer data before handling the missing values using mean-mode method. Decision trees are used to classify the data. WEKA uses J4.8, the modified version of CART, to perform the decision tree method. The missing values are handled using the J4.8 built in ad-hoc procedure. The results shown in the Figure [11] indicates that out of 699 instances, 661 are correctly classified.



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Diagnosis

18:20:10 - trees.j48.J48

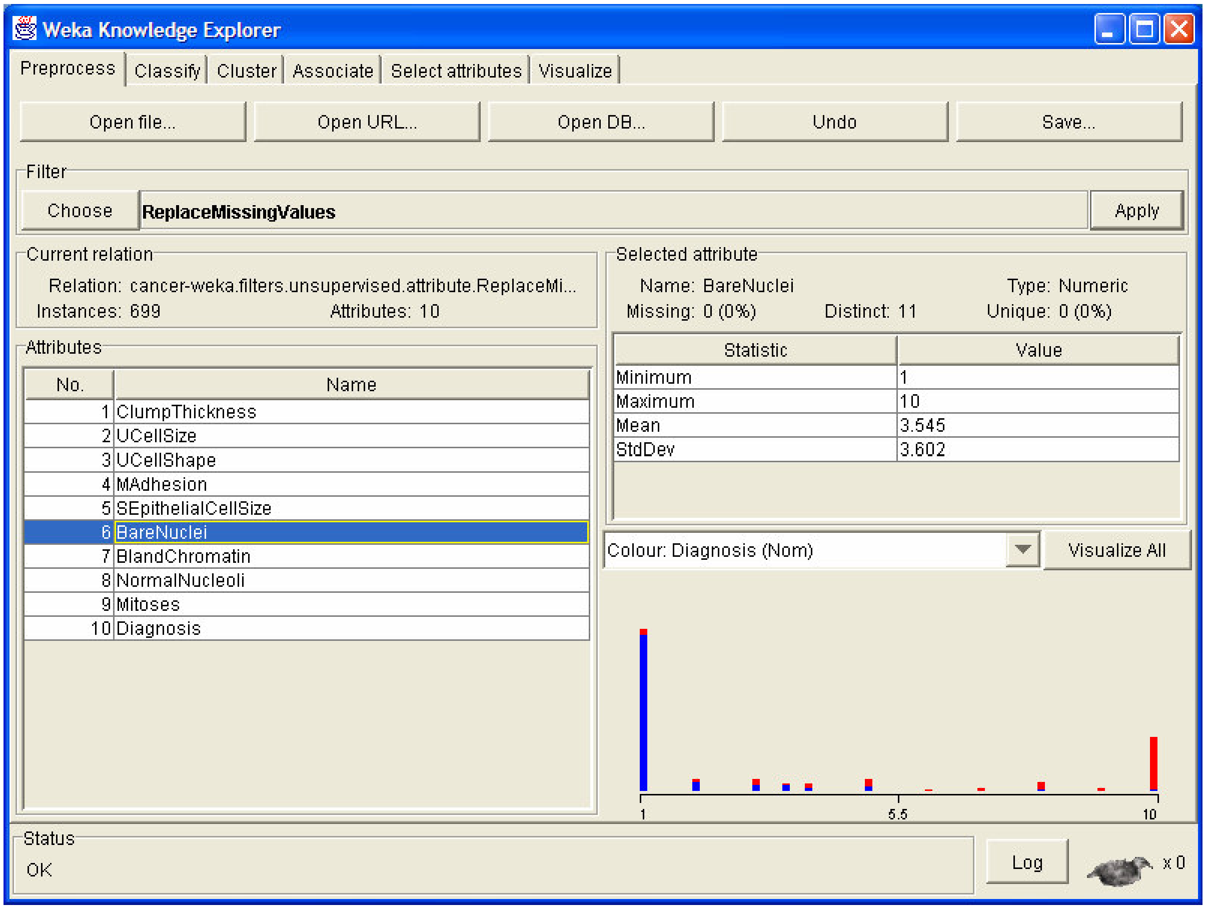
```
Correctly Classified Instances         661               94.5637 %
Incorrectly Classified Instances        38                5.4363 %
Kappa statistic                          0.8799
Mean absolute error                      0.0694
Root mean squared error                  0.2229
Relative absolute error                 15.3526 %
Root relative squared error             46.8932 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.956     0.075      0.961     0.956     0.958    a
  0.925     0.044      0.918     0.925     0.921    b

=== Confusion Matrix ===

   a   b   <-- classified as
 438  20 |   a = a
  18 223 |   b = b
```

Status: OK

**Figure 11 Classification output for J4.8 ad-hoc procedure of WEKA software**

2. Handling the missing values using mean-mode method:

The breast cancer training data set consist of 16 missing values for BareNuclei attribute. This data is to be preprocessed to handle the missing values for obtaining accurate and efficient data mining results. Figure [12] shows the resulting data set after handling the missing values. Since BareNuclei is a numeric data type the missing values are replaced with the attributes mean value.



Filter

ReplaceMissingValues

Current relation

Relation: cancer-weka.filters.unsupervised.attribute.ReplaceMi...
Instances: 699 Attributes: 10

Attributes

| No. | Name |
|---|---|
| 1 | ClumpThickness |
| 2 | UCellSize |
| 3 | UCellShape |
| 4 | MAdhesion |
| 5 | SEpithelialCellSize |
| 6 | BareNuclei |
| 7 | BlandChromatin |
| 8 | NormalNucleoli |
| 9 | Mitoses |
| 10 | Diagnosis |

Selected attribute

Name: BareNuclei Type: Numeric
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

| Statistic | Value |
|---|---|
| Minimum | 1 |
| Maximum | 10 |
| Mean | 3.545 |
| StdDev | 3.602 |

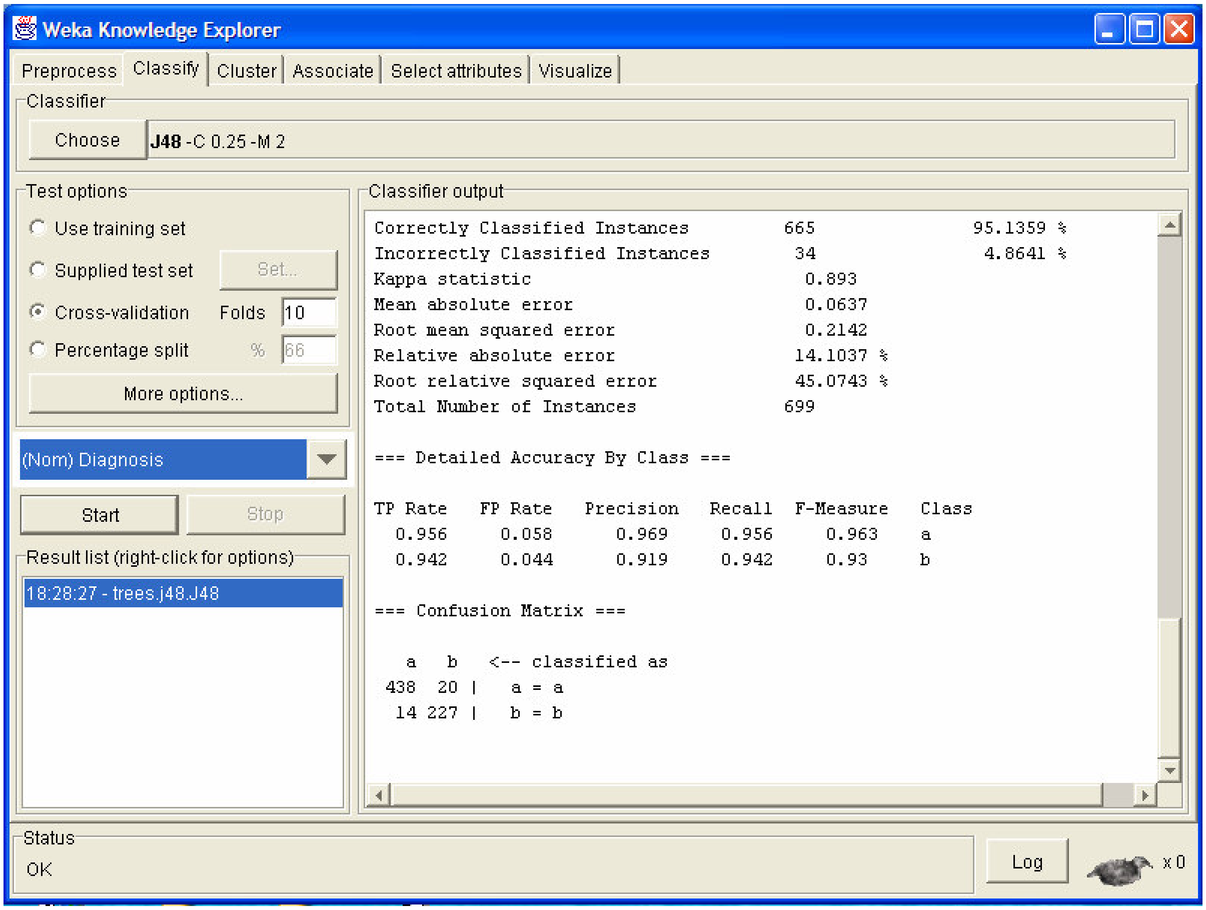
Colour: Diagnosis (Nom)

Status

OK

**Figure 12 Preprocessing results after replacing the missing values of WEKA software**

After preprocessing, the data set is trained using the decision tree (J4.8) by supplying the testing set for classification. Figure [13] shows the classification results of the breast cancer data after handling the missing values. The results shows 665 instances are classified correctly out of 699.



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Diagnosis

18:28:27 - trees.j48.J48

```
Correctly Classified Instances         665               95.1359 %
Incorrectly Classified Instances        34                4.8641 %
Kappa statistic                          0.893
Mean absolute error                      0.0637
Root mean squared error                  0.2142
Relative absolute error                 14.1037 %
Root relative squared error             45.0743 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.956     0.058      0.969     0.956     0.963    a
  0.942     0.044      0.919     0.942     0.93     b

=== Confusion Matrix ===

   a   b   <-- classified as
 438  20 |   a = a
  14 227 |   b = b
```

Status
OK

**Figure 13 Classification results after replacing the missing values using mean-mode method of WEKA software**

The results show there is slight difference before handling the missing values and after handling the missing values. There are also differences in the confusion matrix, mean error, root mean squared error, relative absolute error, root relative squared error, kappa static rate etc.

Accuracy obtained for handling the missing values by J4.8 method and mean-mode method:

| WEKA | Handling missing values using J4.8 | Handling the missing values using mean-mode method |
|---|---|---|
| | 94.5637% | 95.1359% |

**Table 1: Accuracy obtained with respect to the WEKA software**

Confusion matrix for the above results:

| | Handling missing values using J4.8 | | Handling the missing values using mean-mode method | |
|---|---|---|---|---|
| WEKA | A | B | A | B |
| | 438 | 20 | 438 | 20 |
| | 18 | 223 | 14 | 227 |

**Table 2: Confusion matrix of WEKA software**

The error rates for the classification of breast cancer data:

| | Kappa Static | Mean absolute error | Root mean squared error | Relative absolute error | Root relative squared error |
|---|---|---|---|---|---|
| J4.8 | 0.8799 | 0.0694 | 0.2229 | 15.3526% | 46.8932% |
| Mean-mode method | 0.893 | 0.0637 | 0.2142 | 14.1037% | 45.0743% |

**Table 3: Error rates of WEKA software**

The above table shows that there is some difference in the error rates for classifying the data by handling the missing values using J4.8 and mean-mode method. The confusion matrix does show that the results classifying the diagnosis to ‘a’ or ‘b’ is also differed from the ones handling the missing values using mean-mode method.

Classifying the data, which is handled by mean-mode method, using CRUISE linear combination split method, the results shows the correctly classified instances are 676 out of 699.

Accuracy obtained by CRUISE classification and WEKA's J4.8 classification after handling the missing values using mean-mode method:

| Classification after handling the missing values using mean-mode method | WEKA (J4.8) | CRUISE |
|---|---|---|
| | 95.1359% | 96.7095% |

**Table 4: Accuracy obtained by CRUISE and WEKA classification after handling the missing values using mean-mode method**

Confusion matrix obtained by CRUISE software:

| CRUISE | A | B |
|---|---|---|
| | 447 | 11 |
| | 12 | 229 |

**Table 5: Confusion matrix obtained by CRUISE software after handling the missing values using mean-mode method**

Classifying the date using DTREG software after handling the missing values using mean-mode method, the misclassification percent of the training data is 4.292%.

Accuracy obtained by DTREG software after handling the missing values using mean-mode method:

| Classifying after handling the missing values using mean-mode method | DTREG |
|---|---|
| | 95.708% |

**Table 6: Accuracy obtained by DTREG classification after handling the missing values using mean-mode method**

Accuracy obtained for the three software:

| Classification after handling the missing values using mean-mode method | WEKA (J4.8) | CRUISE | DTREG |
|---|---|---|---|
| | 95.1359% | 96.7095% | 95.708% |

**Table 7: Accuracy obtained by WEKA (J4.8), CRUISE and DTREG classification after handling the missing values using mean-mode method**

There is slight difference between the accuracy obtained by the three software. But CRUISE software indicates better accuracy compared to other two methods

## 3.4.2 CRUISE Software

CRUISE is a tree-structured classification. It is a modification of FACT decision tree that splits each node into many sub nodes. It has many ways to deal with the missing values. It can treat them by global imputation or node wise imputation. The CRUISE algorithm can structure the nodes in two ways, either by the univariate split or by the linear combination split. The default setting of the tool is univariate split.

The univariate split method can only handle the missing values using the available case solution (Appendix A). The CRUISE tool only supports the linear split method to handle the missing values using the four options:

a. Fit (complete cases) and impute
b. Nodewise imputation

c. Fit (available cases) and impute

d. Global imputation at root node

During pruning the tree also handles the missing values using the following:

1. Root node imputation
2. Nodewise mean/ mode imputation
3. Estimate the class and impute
4. Go down if possible
5. Alternate split
6. Proxy split

CRUISE takes three input files: data file, test file (if available) and description file. Data file consists of the training data set. The test set can be provided for the classification of the training data set. In this case, only two input files are considered: data file and description file. The description file consists of the name of the data file, missing value, followed by attribute names and their data types.

The description file appears as follows, where “?” indicates the missing value reference:

```
BCWLarge.txt
?
Column, varname, vartype
1,ClumpThickness,n
2,UCellSize,n
3,UCellShape,n
4,MAdhesion,n
5,SEpithelialCellSize,n
6,BareNuclei,c
7,BlandChromatin,c
8,NormalNucleoli,n
```

```
9,Mitoses,c
10,Diagnosis,d
```

The data file is a CSV format file

```
5,1,1,1,2,1,3,1,1,a
5,4,4,5,7,10,3,2,1,a
3,1,1,1,2,2,3,1,1,a
6,8,8,1,3,4,3,7,1,a
4,1,1,3,2,1,3,1,1,a
```

The following are the overall accuracy obtained when running the sets of experiments on the various data sets

| CRUISE | Univariate Split | Linear Split |
|---|---|---|
| | 94.993% | 96.9957% |

**Table 8: Accuracy obtained with respect to the CRUISE software**

The confusion matrix will provide a rough estimate on how the data is classified for the different sets of Examinations

| CRUISE | Univariate Split | | Linear Split | |
|---|---|---|---|---|
| | A | B | A | B |
| | 438 | 20 | 448 | 10 |
| | 15 | 226 | 11 | 230 |

**Table 9: Confusion matrix of Cruise Software**

The linear split use different methods for handling the missing values. The above results are drawn from Case 1 (see below). Applying all the four methods on the breast cancer data:

Case 1: Fit complete cases and impute

During pruning:

Test 1: Root node imputation

Test 2: Nodewise mean/ mode imputation

Test 3: Estimate the class and impute

Test 4: Go down if possible

Test 5: Alternate split

Test 6: Proxy split

Similarly, all the six methods of handling missing values during pruning the tree are performed for the remaining three cases:

Case 2: Nodewise imputation and fit

Case 3: Fit (available cases) and impute

Case 4: Global imputation at root node and fit

The correctly classified testing set instances (out of 699) for all four cases:

| | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Test 6 |
|---|---|---|---|---|---|---|
| Case 1 | 676 | 676 | 676 | 676 | 676 | 676 |
| Case 2 | 678 | 678 | 678 | 678 | 678 | 678 |
| Case 3 | NA | NA | NA | NA | NA | NA |
| Case 4 | 678 | 678 | 678 | 678 | 678 | 678 |

**Table 10: Correctly classified instances for all four cases in CRUISE software**

The confusion matrix provided for all the four cases:

| | Test 1 | | Test 2 | | Test 3 | | Test 4 | | Test 5 | | Test 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case 1 | 447 | 11 | 447 | 11 | 447 | 11 | 447 | 11 | 447 | 11 | 447 | 11 |
| | 12 | 229 | 12 | 229 | 12 | 229 | 12 | 229 | 12 | 229 | 12 | 229 |
| Case 2 | 448 | 10 | 448 | 10 | 448 | 10 | 448 | 10 | 448 | 10 | 448 | 10 |
| | 11 | 230 | 11 | 230 | 11 | 230 | 11 | 230 | 11 | 230 | 11 | 230 |
| Case 3 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| Case 4 | 448 | 10 | 448 | 10 | 448 | 10 | 448 | 10 | 448 | 10 | 448 | 10 |
| | 11 | 230 | 11 | 230 | 11 | 230 | 11 | 230 | 11 | 230 | 11 | 230 |

**Table 11: Confusion matrix for all the four cases in CRUISE software**

Case 3 can be applied for only the univariate method and Case 2 and Case 4 give the same results. Eliminating Case3, the accuracy of the three cases is shown below:
Accuracy of the three cases:

| Accuracy | Case 1 | Case 2 | Case 4 |
|---|---|---|---|
| | 96.7095% | 96.9957% | 96.9957% |

**Table 12: Accuracy for all the four cases in CRUISE software**

There is a slight difference between Case 1 and Case 2 & Case 4.

### 3.4.3 NORM Software

NORM is a software that handles the missing values using multiple imputation, i.e. each missing datum is replaced by m>1 simulated values. It can also perform the pre-processing tasks like transformation for generating efficient results and post-processing like combining all the imputations into one set for performing the data mining tasks on a single data set. NORM cannot handle all the statistical analysis like classification or regression but it just performs only the pre-processing tasks like handling missing values. Once the output of the data set is generated, it can be used by other statistical packages for performing the data mining tasks.

The process consists of two algorithms to handle the missing values: EM algorithm for estimating the mean, variances, and covariances (or correlations) of the attributes and the data augmentation for generating multiple imputations of the missing values (Appendix B).

NORM accepts only ASCII format data types with extension ".dat". It takes one input fie, containing the training data values, and the attributes of the data file must be given in another file with the same name as the data values file name and extension ".nam". The missing values must be denoted by single numeric value like -9, -99 or 1000 but not any non-numeric type. The main drawback of this software is it gives better results for continuous data.

The description file appears as follows:

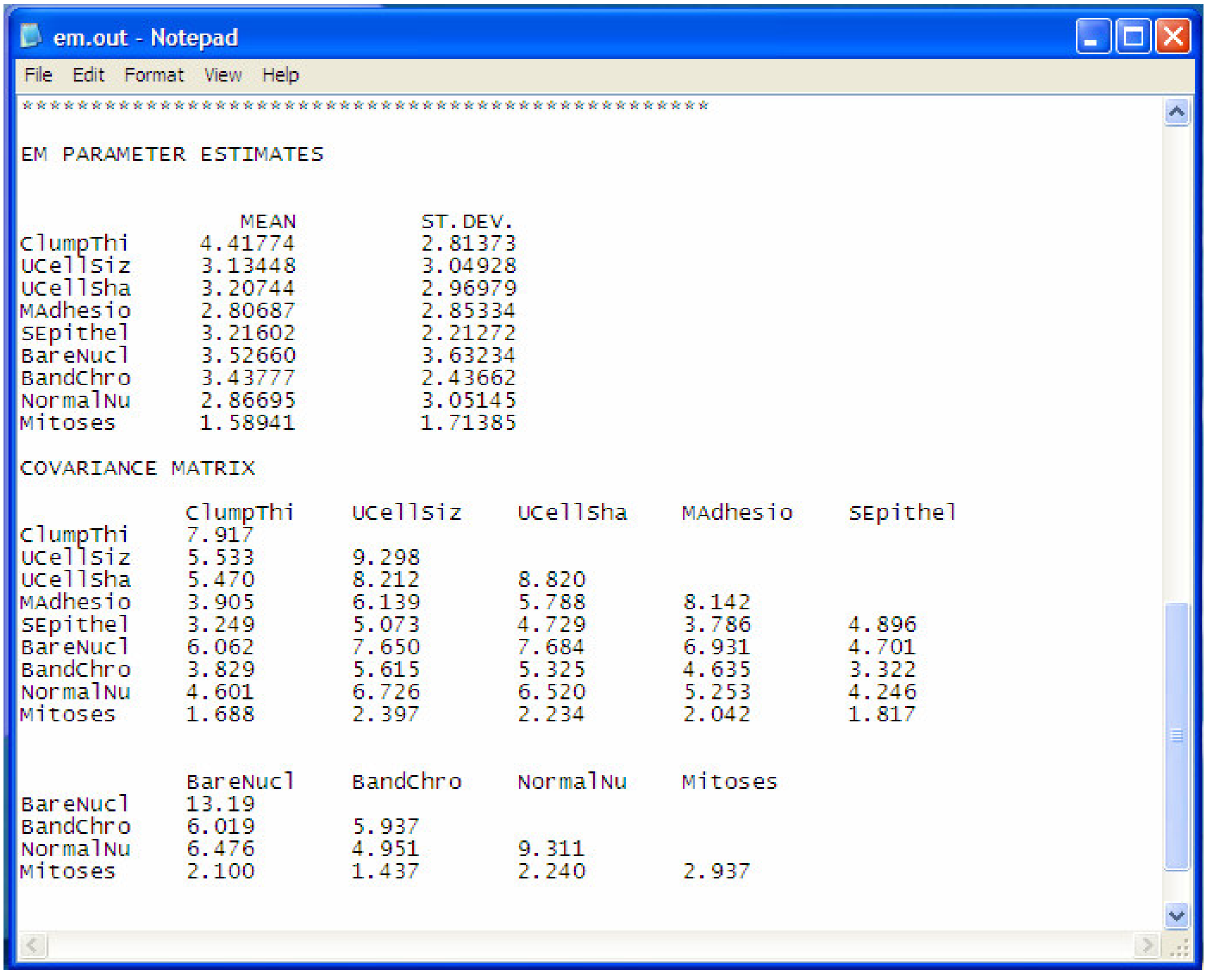
ClumpTh
UCSize
UCShape
MAdh
SEpSize
BareNuc
BlChrom
NorNuc
Mitoses

The data file in ascii format:

```
5 1 1 1 2 1 3 1 1
5 4 4 5 7 10 3 2 1
3 1 1 1 2 2 3 1 1
6 8 8 1 3 4 3 7 1
```

As noticed above, the diagnosis attribute is deleted due to its non-numeric data type, as NORM does not support non-numeric data type. The values can be transformed to numeric, say ‘a’ to 1 and ‘b’ to 2, since this is categorical data and has no missing values this can also be eliminated.

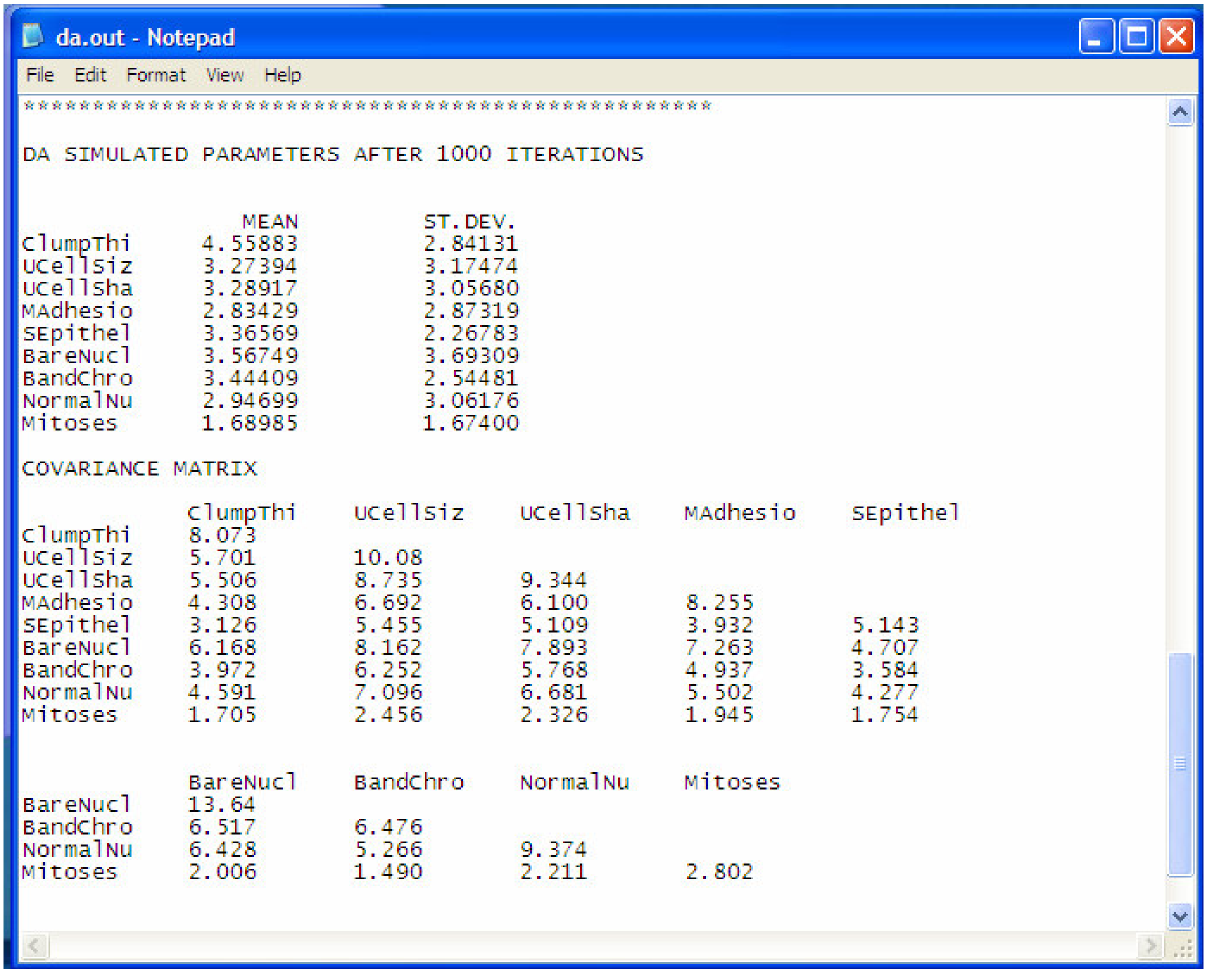
Once the data file is inputted, pre-processing tasks like transformations, selecting a few variables, and modifying the attribute names can be done before performing the process. The summary of the data file and the attribute names and the data types can be viewed. Once the data file is selected before performing the multiple imputation of missing values, EM algorithm is executed for providing good starting values for data augmentation procedure and also helps the DA to converge quickly. The EM algorithm is used to estimate the mean, variances and covariances using all the cases in the data set (including missing values). Figure [14] shows the EM algorithm output.



```
**************************************************

EM PARAMETER ESTIMATES


                MEAN         ST.DEV.
ClumpThi     4.41774         2.81373
UCellSiz     3.13448         3.04928
UCellSha     3.20744         2.96979
MAdhesio     2.80687         2.85334
SEpithel     3.21602         2.21272
BareNucl     3.52660         3.63234
BandChro     3.43777         2.43662
NormalNu     2.86695         3.05145
Mitoses      1.58941         1.71385

COVARIANCE MATRIX

            ClumpThi    UCellSiz    UCellSha    MAdhesio    SEpithel
ClumpThi    7.917
UCellSiz    5.533       9.298
UCellSha    5.470       8.212       8.820
MAdhesio    3.905       6.139       5.788       8.142
SEpithel    3.249       5.073       4.729       3.786       4.896
BareNucl    6.062       7.650       7.684       6.931       4.701
BandChro    3.829       5.615       5.325       4.635       3.322
NormalNu    4.601       6.726       6.520       5.253       4.246
Mitoses     1.688       2.397       2.234       2.042       1.817


            BareNucl    BandChro    NormalNu    Mitoses
BareNucl    13.19
BandChro    6.019       5.937
NormalNu    6.476       4.951       9.311
Mitoses     2.100       1.437       2.240       2.937
```

**Figure 14 EM algorithm output for NORM**

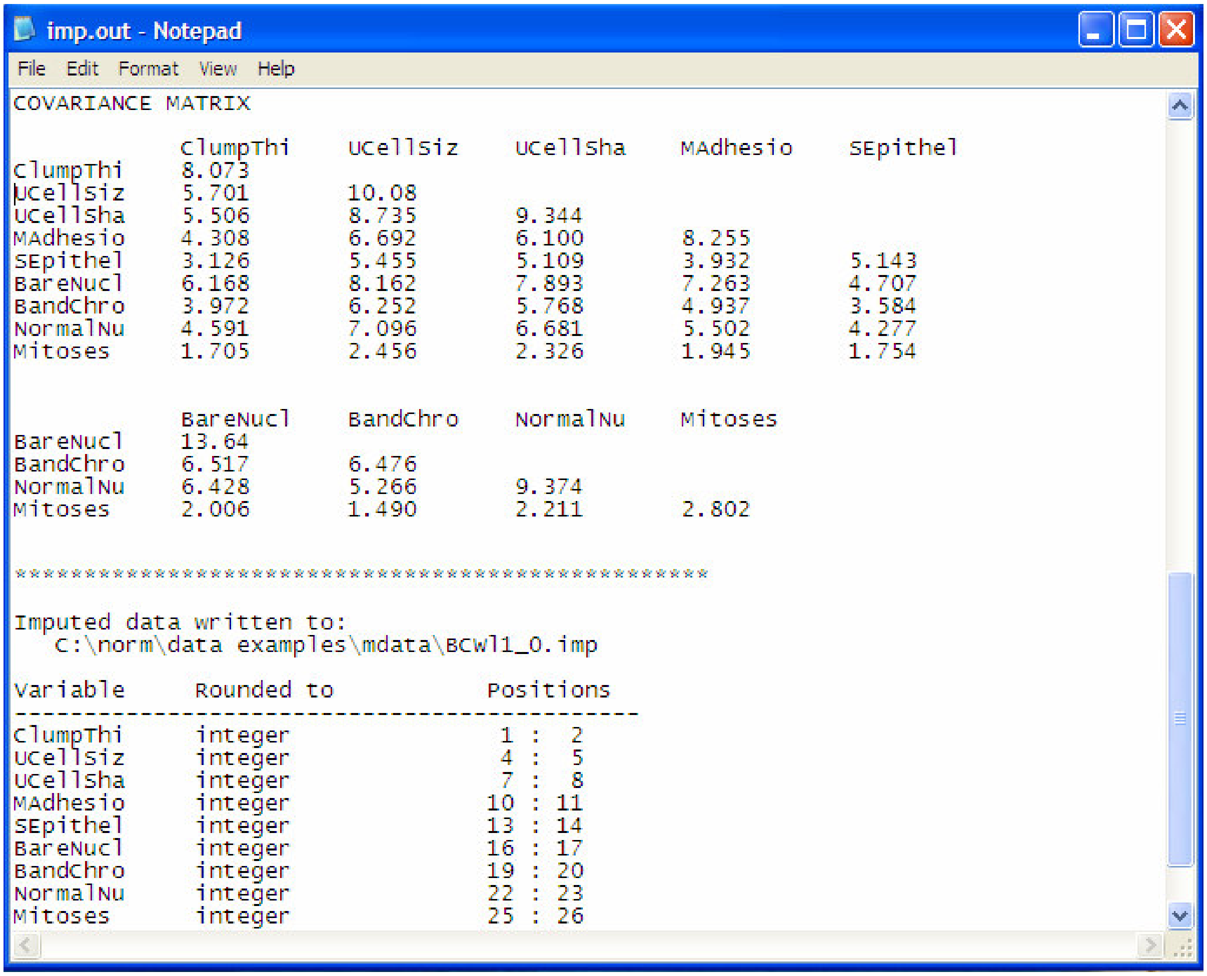
Once the EM algorithm is executed, the DA algorithm simulates the random values of the parameters and missing data from their posterior distribution by taking the starting values for the parameters generated by EM algorithm. The DA results can be seen in Figure [15].



```
**************************************************

DA SIMULATED PARAMETERS AFTER 1000 ITERATIONS


                   MEAN          ST.DEV.
ClumpThi     4.55883         2.84131
UCellsiz     3.27394         3.17474
UCellsha     3.28917         3.05680
MAdhesio     2.83429         2.87319
SEpithel     3.36569         2.26783
BareNucl     3.56749         3.69309
BandChro     3.44409         2.54481
NormalNu     2.94699         3.06176
Mitoses      1.68985         1.67400

COVARIANCE MATRIX

            ClumpThi    UCellsiz    UCellsha    MAdhesio    SEpithel
ClumpThi    8.073
UCellsiz    5.701       10.08
UCellsha    5.506       8.735       9.344
MAdhesio    4.308       6.692       6.100       8.255
SEpithel    3.126       5.455       5.109       3.932       5.143
BareNucl    6.168       8.162       7.893       7.263       4.707
BandChro    3.972       6.252       5.768       4.937       3.584
NormalNu    4.591       7.096       6.681       5.502       4.277
Mitoses     1.705       2.456       2.326       1.945       1.754


            BareNucl    BandChro    NormalNu    Mitoses
BareNucl    13.64
BandChro    6.517       6.476
NormalNu    6.428       5.266       9.374
Mitoses     2.006       1.490       2.211       2.802
```

**Figure 15 DA output for NORM**

Finally the NORM software can also generate the imputation file containing the imputed data set for performing other data mining tasks. Figure [16] shows the final results of the procedure.



```
COVARIANCE MATRIX

            ClumpThi    UCellSiz    UCellSha    MAdhesio    SEpithel
ClumpThi    8.073
UCellSiz    5.701       10.08
UCellSha    5.506       8.735       9.344
MAdhesio    4.308       6.692       6.100       8.255
SEpithel    3.126       5.455       5.109       3.932       5.143
BareNucl    6.168       8.162       7.893       7.263       4.707
BandChro    3.972       6.252       5.768       4.937       3.584
NormalNu    4.591       7.096       6.681       5.502       4.277
Mitoses     1.705       2.456       2.326       1.945       1.754


            BareNucl    BandChro    NormalNu    Mitoses
BareNucl    13.64
BandChro    6.517       6.476
NormalNu    6.428       5.266       9.374
Mitoses     2.006       1.490       2.211       2.802


**************************************************

Imputed data written to:
   C:\norm\data examples\mdata\BCW11_0.imp

Variable     Rounded to             Positions
---------------------------------------------
ClumpThi     integer                 1 :  2
UCellSiz     integer                 4 :  5
UCellSha     integer                 7 :  8
MAdhesio     integer                10 : 11
SEpithel     integer                13 : 14
BareNucl     integer                16 : 17
BandChro     integer                19 : 20
NormalNu     integer                22 : 23
Mitoses      integer                25 : 26
```

**Figure 16 Final imputed results for NORM**

Once the data is ready after performing the pre-processing tasks (handling missing values), any software can be applied for data mining tasks. In this case, WEKA software is used to classify the resulting data set. The results shows 661 correctly classified instances and the error rates are much less compared to the mean – mode method:

The accuracy obtained for NORM method, classification using WEKA (J4.8):

| | Accuracy |
|---|---|
| NORM | 94.5637% |

**Table 13: Accuracy obtained with respect to the NORM software using WEKA classification**

The confusion matrix for NORM method:

| NORM | A | B |
|---|---|---|
| | 436 | 22 |
| | 16 | 225 |

**Table 14: Confusion matrix of NORM software using WEKA classification**

The error rates for NORM method:

| | Kappa Static | Mean absolute error | Root mean squared error | Relative absolute error | Root relative squared error |
|---|---|---|---|---|---|
| NORM | 0.8804 | 0.069 | 0.2265 | 15.26% | 47.646% |

**Table 15: Error rates obtained for NORM software using WEKA classification**

Classifying the data using CRUISE software, using linear combination split, the results indicate the correctly classified instances as 677.

Accuracy obtained using CRUISE software:

| Classification after handling missing values using NORM | CRUISE |
|---|---|
| | 96.8526% |

**Table 16: Accuracy obtained with respect to the NORM software using CRUISE classification**

Confusion matrix using CRUISE software:

| After handling the missing values using NORM | A | B |
|---|---|---|
| | 447 | 11 |
| | 11 | 230 |

**Table 17: Confusion matrix of NORM software using CRUISE classification**

Classifying the data using DTREG software, the misclassification percent of the training data is 4.292, which is similar to the classification of WEKA's mean-mode data using DTREG.

Accuracy obtained by DTREG software:

| Classification after handling the missing values using NORM | DTREG |
|---|---|
| | 95.708% |

**Table 18: Accuracy obtained with respect to the NORM software using DTREG classification**

Accuracy obtained by the three software:

| Classification after handling the missing values using NORM | WEKA (J4.8) | CRUISE | DTREG |
|---|---|---|---|
| | 94.5637% | 96.8526% | 95.708% |

**Table 19: Accuracy obtained for the three classification methods after handling the missing values using NORM**

The above results indicate the accuracy obtained for classification using CRUISE software is more compared to the other two software.

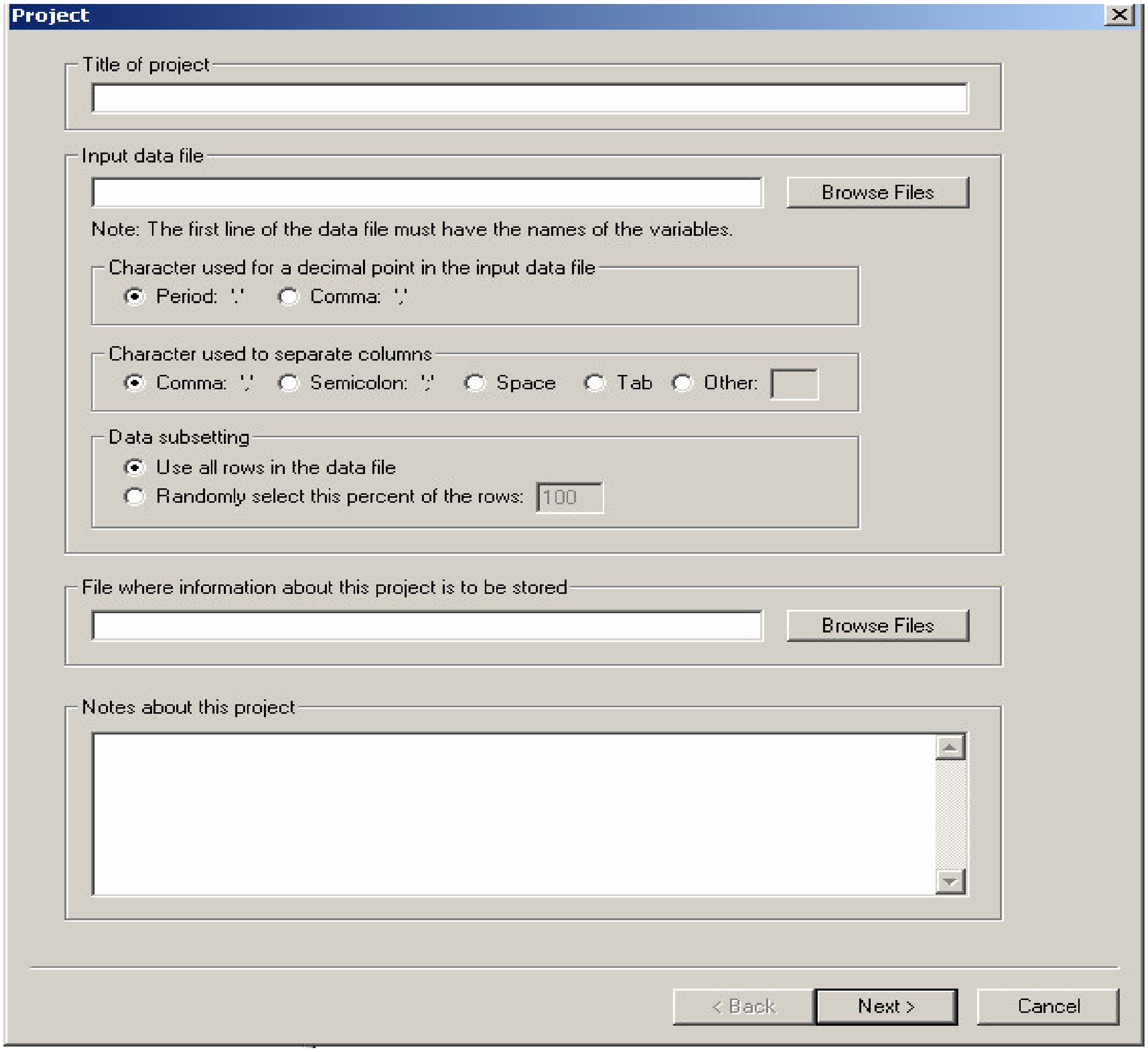
## 3.4.4 DTREG Software

DTREG builds classification and regression decision trees that model data relationships and predict values for future observations. DTREG accepts a data set where one of the variables is chosen as a target variable whose value is to be modeled and predicted as a function of the predictor variable. It analyzes the data and generates a decision tree by performing binary splits on the predictor variable.

DTREG offers two methods for salvaging rows with missing values on the predictor variable that is used for splitting the group:

1. Surrogate splits: Surrogate variables are predictor variables that are not as good as the primary splitters but serves similar splits. DTREG compares which rows are sent to left and right child groups by the primary splitter. The predictors/attributes whose splits most closely mimic the split by the primary splitter are the surrogate splitters. The association between the primary splitter and each alternate predictor are calculated and the surrogate splitter variables are ranked in the decreasing order of the association. When DTREG encounters a missing value on the primary splitter, it replaces the value with the surrogate splitter variable (no missing value), which is having the highest association.
2. Put rows in the most probable group: If the value of the splitting variable is missing, the row is put into the child group that has the highest likelihood of receiving unknown or random cases.

Figure [17] shows the DTREG software. The results of the case study, the breast cancer data set for DTREG software are discussed further:



Project

Title of project

Input data file

Note: The first line of the data file must have the names of the variables.

Character used for a decimal point in the input data file

● Period: '.' ○ Comma: ','

Character used to separate columns

● Comma: ',' ○ Semicolon: ';' ○ Space ○ Tab ○ Other:

Data subsetting

● Use all rows in the data file

○ Randomly select this percent of the rows: 100

File where information about this project is to be stored

Notes about this project

**Figure 17 DTREG software**

DTREG accepts ASCII format files. Mostly comma separated format files (CSV) are used as input files. The decimal point indication, either period or comma, and the character used to separate the columns must be specified allowing the input file. The attributes of the data set must be specified at the first line of the file.

Data file format for DTREG input:

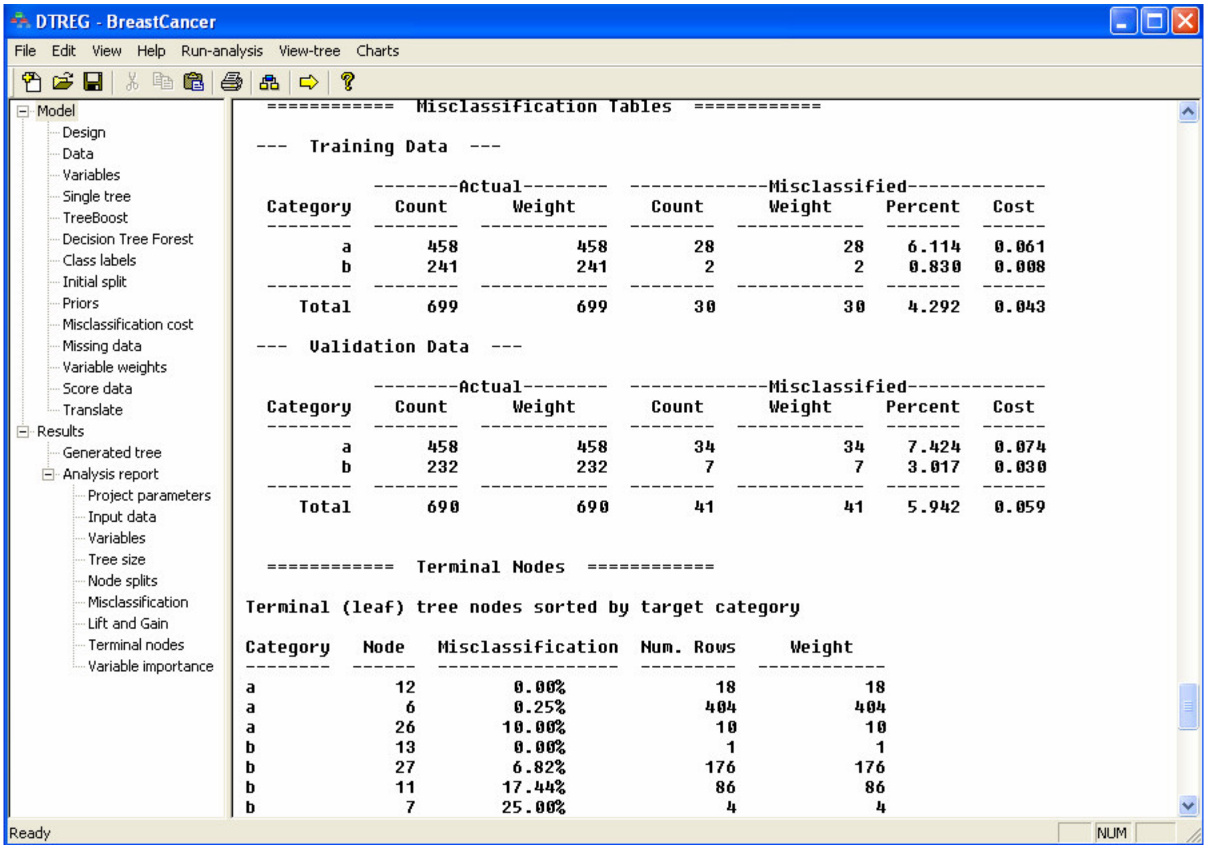
```
“ClumpThickness”, “UCellShape”, “MAdhesion”, “SEpithelialCellSize”, “BareNuclei”,
“BlandChromatin”, ”'NormalNucleoli”, “Mitoses”,  “Diagnosis”
5,  1,  1,  1,  2,  1,  3,  1,  1, a
 5,  4,  4,  5,  7, 10,  3,  2,  1, a
 3,  1,  1,  1,  2,  2,  3,  1,  1, a
 6,  8,  8,  1,  3,  4,  3,  7,  1,  a
 4,  1,  1,  3,  2,  1,  3,  1,  1,  a
 8, 10, 10,  8,  7, 10,  9,  7,  1, b
```

The following notations can be used for the missing values:

- The question mark character
- A single period
- Empty space between two commas

Once the input data is loaded, the user has the option of selecting the target variable, which needs to be categorized and the predictor variables. In the breast cancer case the target variable is diagnosis attribute. The remaining variables are considered as the predictors.

The resulting analysis of the DTREG is shown in Figure [18].



```
 ============  Misclassification Tables  ============

---  Training Data  ---

           --------Actual--------  -------------Misclassified-------------
 Category    Count       Weight      Count      Weight     Percent   Cost
 --------  --------  ------------  --------  ------------  -------  ------
        a       458           458        28            28    6.114   0.061
        b       241           241         2             2    0.830   0.008
 --------  --------  ------------  --------  ------------  -------  ------
    Total       699           699        30            30    4.292   0.043

---  Validation Data  ---

           --------Actual--------  -------------Misclassified-------------
 Category    Count       Weight      Count      Weight     Percent   Cost
 --------  --------  ------------  --------  ------------  -------  ------
        a       458           458        34            34    7.424   0.074
        b       232           232         7             7    3.017   0.030
 --------  --------  ------------  --------  ------------  -------  ------
    Total       690           690        41            41    5.942   0.059


 ============  Terminal Nodes  ============

Terminal (leaf) tree nodes sorted by target category

Category   Node   Misclassification  Num. Rows     Weight
--------  ------  -----------------  ---------  ------------
a             12       0.00%                18            18
a              6       0.25%               404           404
a             26      10.00%                10            10
b             13       0.00%                 1             1
b             27       6.82%               176           176
b             11      17.44%                86            86
b              7      25.00%                 4             4
```

Figure 18 Analysis results of breast cancer data set for DTREG

During handling the missing values process, both the options surrogate splitters and putting the rows in the probable group are selected. The results in Figure [18] indicate the correctly classified instances for the training data.  The DTREG software uses 699 cases of training and 690 cases for validation data. The misclassification percent of the training data is 4.292.

Accuracy obtained for the validation data:

|  |  |
|---|---|
| DTREG | Accuracy |
|  | 95.708% |

**Table 20: Accuracy obtained for DTREG software**

Figure [19] shows the tree structure results of the breast cancer for the DTREG software:

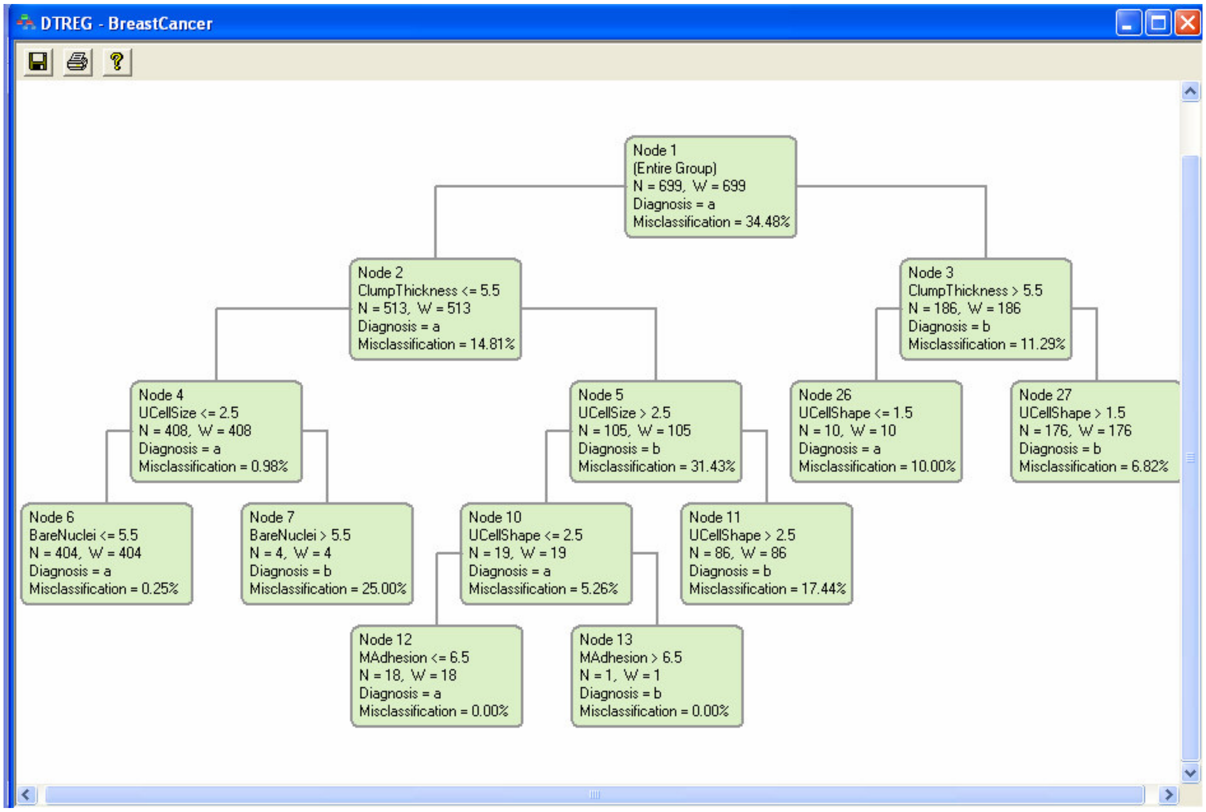


**Figure 19 Decision tree of breast cancer for DTREG**

## 3.5 Conclusion for the Breast Cancer Case Study

The above section discussed the methods of handling the missing values used by four software: WEKA, CRUISE, NORM and DTREG. The results of the breast cancer data set used for the case study indicates there is a slight difference between the accuracy obtained for all the four software in handling the missing values. Among the four, CRUISE software indicates better accuracy results for classification of diagnosis. The results might differ for other data. In most cases it depends on the number of missing values in the data and the instances given. But the results above indicate no matter which method is used for handling the missing values, the results might depend on the classification method is used.

Accuracy obtained for breast cancer data using the four software:

| Handling missing values | Classification | | |
|---|---|---|---|
| | WEKA (J4.8) | CRUISE | DTREG |
| WEKA | 95.1359% | 96.7095% | 95.708% |
| NORM | 94.5637% | 96.8526% | 95.708% |
| CRUISE | NA | 96.9957% | NA |
| DTREG | NA | NA | 95.708% |

**Table 21: Accuracy obtained by the four software for breast cancer data.**

## 3.6 Summary:

There are various methods of handling missing values and data mining techniques, some algorithms might work better than the other while running one type of data as compared to the rest. We have seen the model used in developing the E-Intelligence form design and how the underlying logic of the Entry checker can reduce the missing values during entry level and make sure the data is less erroneous before storing into databases. Software that performs different methods of handling the missing values are introduced in this chapter. Breast cancer data is used to show how these software can work. In the next chapter, we will discuss how the E-Intelligence form for MDS-MH data set is designed and tested with different missing value tools.

# Chapter 4

# Experiments and case study

In Chapter 3, a discussion on the model of E-Intelligence form design to reduce the missing values is presented. The E-Intelligence form is designed for MDS-MH data, but the underlying logic of this form design can be applied to any data. This chapter discusses the E-Intelligence form design and the logic used in Entry checker for the MDS-MH. A subset of MDS-MH data is considered as a case study and all the four methods of handling missing values, discussed in Chapter 3, are performed on this set and checked for accuracy.

## 4.1 User Interface for E-Intelligence Form

The main user interface of E-Intelligence form has two features: one for data entry step, which is online process and another for existing data, offline, which performs tasks like data mining and data preprocessing, mainly handling missing values using different software, data transformation and data reduction.

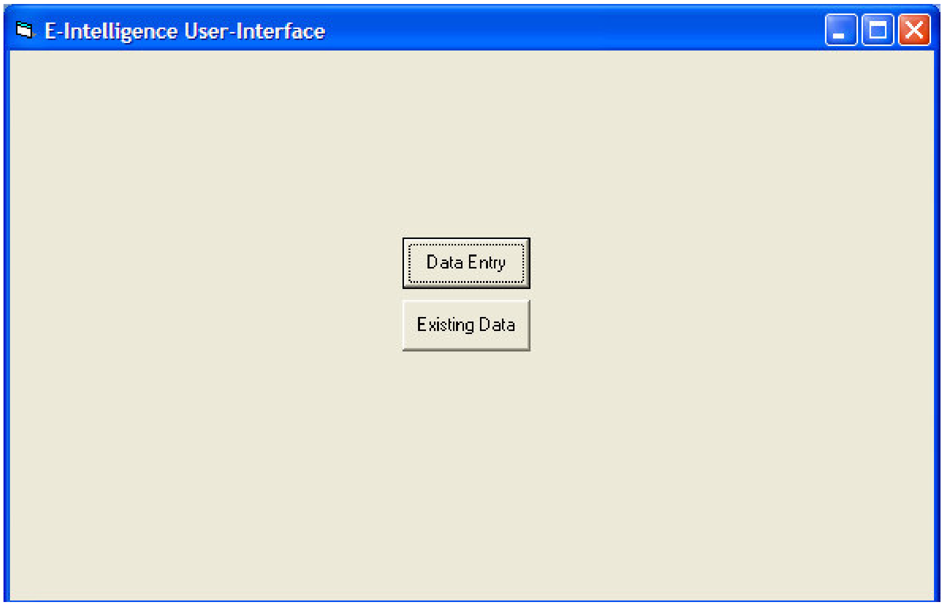


**Figure 20 User-Interface for the E-Intelligence form design**

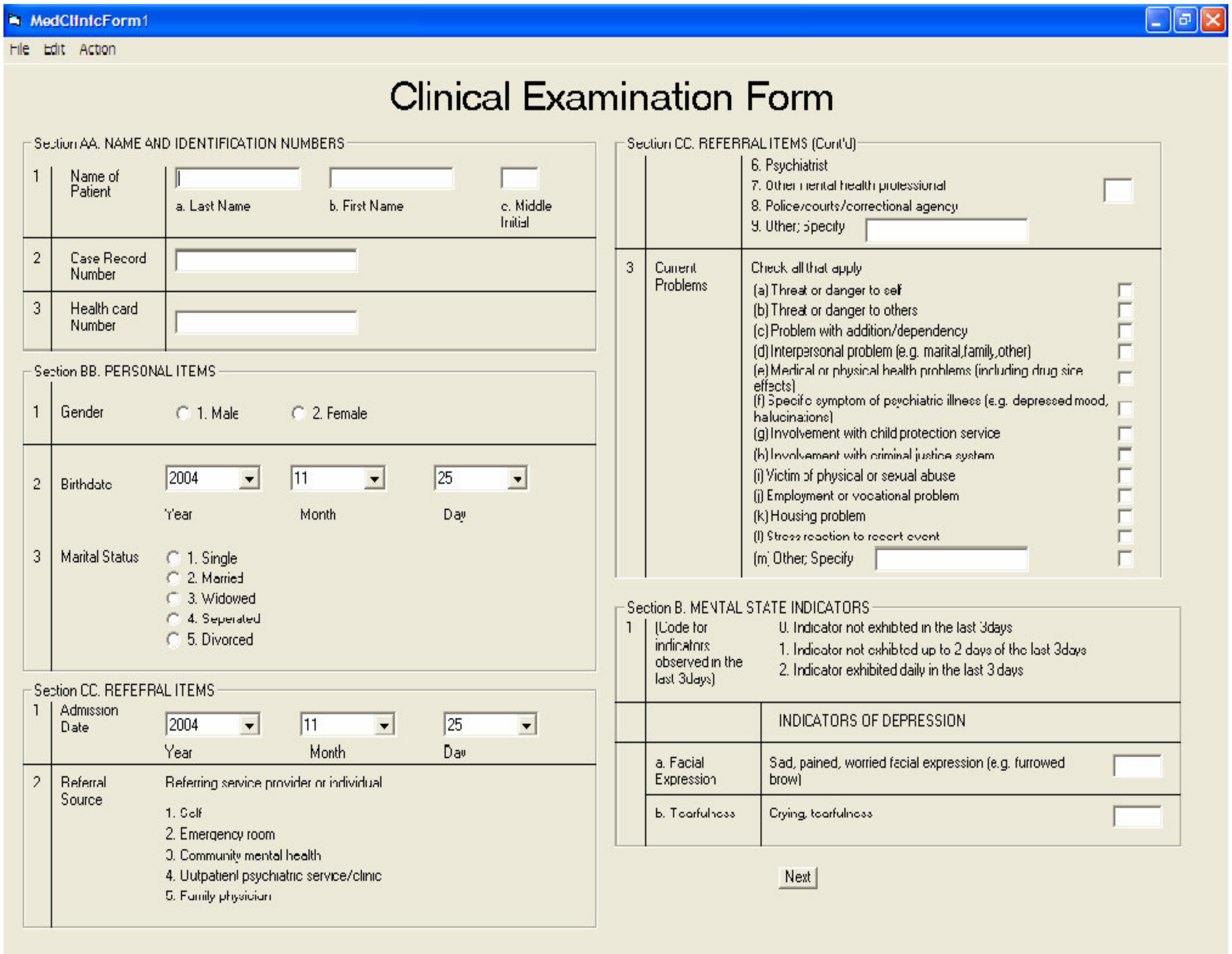
The “Data Entry” feature guides the user to the E-Intelligence form of the MDS-MH data set where the user can enter the data, which is an online process. The “Existing Data” feature transforms the user to a new form where the users have options of performing the data mining process, and data preprocessing tasks like handling missing values using the four software: WEKA, CRUISE, NORM and DTREG, data transformation and data reduction on the existing data. Here, the tasks of data transformation and data reduction are implemented for transforming MDS-MH categorical data to numerical and vise versa and feature selection of attributes among the 455 attributes. The underlying logic in these tasks can be used for any kind of data set.

### 4.1.1 E-Intelligence Form Design for MDS-MH data

E-Intelligence form is mainly designed for MDS-MH data, an assessment made with psychiatric patients obtained from RAI-MH. This data consists of 455 attributes and currently has 4000 patient instances. The data can be used to classify the patient into the four major categories, described in Chapter 2, in mental healthcare. The advantage of MDS-MH data is most of the attributes are linked together.

Forms are designed similar to the MDS-MH paper questionnaire. Clinical data are stored in text, Microsoft ® Excel ® and various other formats. Considering the storage capacity and for an easier implementation, Microsoft ® Access ® is used as the MDS-MH database.

Next the E-Intelligence user-interface containing the first Form, paper1, of MDS-MH system is presented:



## Clinical Examination Form

Section AA. NAME AND IDENTIFICATION NUMBERS

|  |  |  |
|---|---|---|
| 1 | Name of Patient | a. Last Name<br>b. First Name<br>c. Middle Initial |
| 2 | Case Record Number |  |
| 3 | Health card Number |  |

Section BB. PERSONAL ITEMS

|  |  |  |
|---|---|---|
| 1 | Gender | ○ 1. Male ○ 2. Female |
| 2 | Birthdate | 2004 Year<br>11 Month<br>25 Day |
| 3 | Marital Status | ○ 1. Single<br>○ 2. Married<br>○ 3. Widowed<br>○ 4. Seperated<br>○ 5. Divorced |

Section CC. REFEFRAL ITEMS

|  |  |  |
|---|---|---|
| 1 | Admission Date | 2004 Year<br>11 Month<br>25 Day |
| 2 | Referral Source | Referring service provider or individual<br>1. Self<br>2. Emergency room<br>3. Community mental health<br>4. Outpatient psychiatric service/clinic<br>5. Family physician |

Section CC. REFERRAL ITEMS (Cont'd)

|  |  |  |
|---|---|---|
|  |  | 6. Psychiatrist<br>7. Other mental health professional<br>8. Police/courts/correctional agency<br>9. Other; Specify |
| 3 | Current Problems | Check all that apply<br>(a) Threat or danger to self □<br>(b) Threat or danger to others □<br>(c) Problem with addition/dependency □<br>(d) Interpersonal problem (e.g. marital,family,other) □<br>(e) Medical or physical health problems (including drug side effects) □<br>(f) Specific symptom of psychiatric illness (e.g. depressed mood, halucinations) □<br>(g) Involvement with child protection service □<br>(h) Involvement with criminal justice system □<br>(i) Victim of physical or sexual abuse □<br>(j) Employment or vocational problem □<br>(k) Housing problem □<br>(l) Stress reaction to recent event □<br>(m) Other; Specify □ |

Section B. MENTAL STATE INDICATORS

|  |  |  |
|---|---|---|
| 1 | (Code for indicators observed in the last 3days) | 0. Indicator not exhibted in the last 3days<br>1. Indicator not exhibted up to 2 days of the last 3days<br>2. Indicator exhibited daily in the last 3 days |
|  |  | INDICATORS OF DEPRESSION |
|  | a. Facial Expression | Sad, pained, worried facial expression (e.g. furrowed brow) |
|  | b. Tearfulness | Crying, tearfulness |

**Figure 21 E-Intelligence user-interface form of MDS-MH data set**

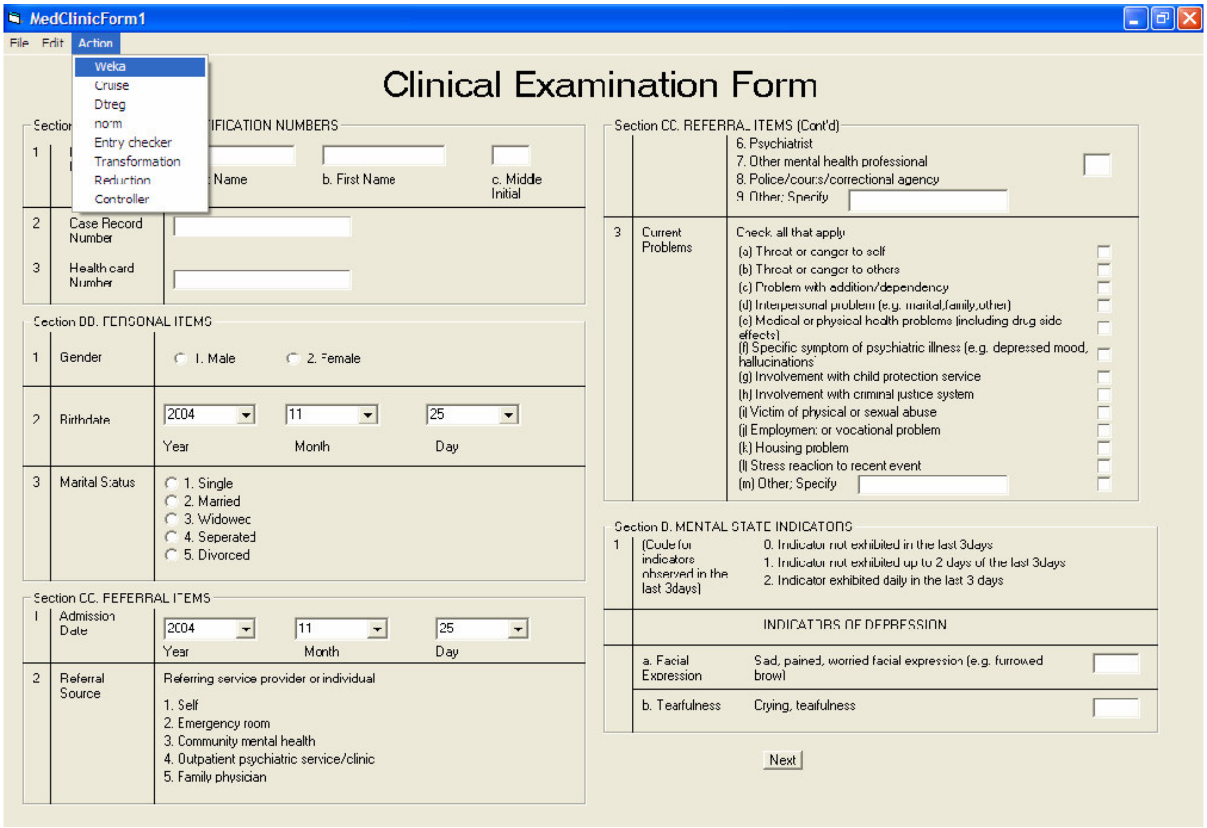
These forms are used for entering, storing and retrieving patient's data. Patient's data is stored and can be identified using the case record number, which acts as the primary key for the database. Patient's data can also be retrieved for viewing or editing the details by one of the attributes: case record number, health card number or patient's name.

Most of the existing forms in today's medical field satisfy initial tasks like entering, storing and retrieving the patient's data. The E-Intelligence form designed for the MDS-MH is different for its intelligence behavior of finding inconsistent and erroneous data. This form handles the data preprocessing tasks that can be performed during the different stages of data mining process. This intelligence behavior helps in yielding efficient and accurate classification results.

This E-Intelligence form communicates with the Entry checker online and reduces the missing values, noise in data and duplicate records during the data entry. The form performs online data transformation. It can automatically convert the categorical data to numerical data before storing into the database. The user can also do other data preprocessing tasks from the form, like data transformation and data reduction on the existing/stored data. The E-Intelligence form can also guide the user with different missing value handling methods and data mining techniques to perform on the existing data. Figure [22] shows the form which performs the actions menu.

The File menu in the tool bar of the E-Intelligence form has open, save, close and print options, which are used for opening an existing data (instance/record), saving an entry record, closing the recent opened data file and printing the data forms respectively. The Edit menu has options like clearing the data of an entered attribute, and Go to, which can transfer to a particular form (page) of the MDS-MH data. Figure [22] shows the form which performs the actions menu.

The Action option in the menu has Entry checker, controller, data transformation, data reduction, WEKA, CRUISE, NORM and DTREG. These software are offline processes. They can perform only the stored/existing data but for the newly entering data in the data entry step. For each and every data entry in the E-Intelligence form the Entry checker checks for reducing missing values, and noise in data. Entry checker option can also be used on existing instances/records.



Clinical Examination Form

Section ... IFICATION NUMBERS

| | | | | |
|---|---|---|---|---|
| 1 | | ... Name | b. First Name | c. Middle Initial |
| 2 | Case Record Number | | | |
| 3 | Health card Number | | | |

Section DD. PERSONAL ITEMS

| | | |
|---|---|---|
| 1 | Gender | ○ 1. Male ○ 2. Female |
| 2 | Birthdate | 2004 Year 11 Month 25 Day |
| 3 | Marital Status | ○ 1. Single<br>○ 2. Married<br>○ 3. Widowec<br>○ 4. Seperated<br>○ 5. Divorced |

Section CC. REFERRAL ITEMS

| | | |
|---|---|---|
| 1 | Admission Date | 2004 Year 11 Month 25 Day |
| 2 | Referral Source | Referring service provider or individual<br>1. Self<br>2. Emergency room<br>3. Community mental health<br>4. Outpatient psychiatric service/clinic<br>5. Family physician |

Section CC. REFERRAL ITEMS (Cont'd)

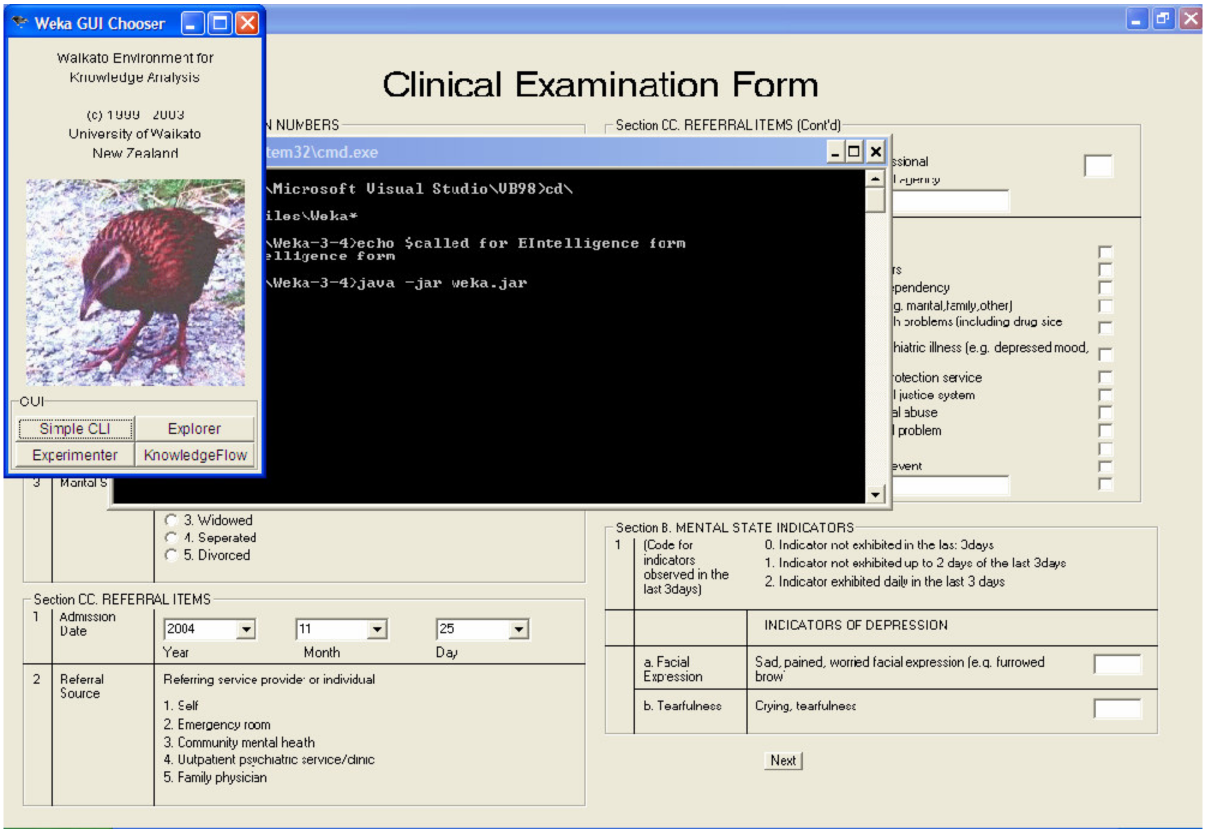
| | | | |
|---|---|---|---|
| | | 6. Psychiatrist<br>7. Other mental health professional<br>8. Police/courts/correctional agency<br>9. Other: Specify | □ |
| 3 | Current Problems | Check all that apply | |
| | | (a) Threat or danger to self | □ |
| | | (b) Threat or danger to others | □ |
| | | (c) Problem with addition/dependency | □ |
| | | (d) Interpersonal problem (e.g. marital,family,other) | □ |
| | | (e) Medical or physical health problems (including drug side effects) | □ |
| | | (f) Specific symptom of psychiatric illness (e.g. depressed mood, hallucinations | □ |
| | | (g) Involvement with child protection service | □ |
| | | (h) Involvement with criminal justice system | □ |
| | | (i) Victim of physical or sexual abuse | □ |
| | | (j) Employment or vocational problem | □ |
| | | (k) Housing problem | □ |
| | | (l) Stress reaction to recent event | □ |
| | | (m) Other; Specify | □ |

Section D. MENTAL STATE INDICATORS

| | | | |
|---|---|---|---|
| 1 | (Code for indicators observed in the last 3days) | 0. Indicator not exhibited in the last 3days<br>1. Indicator not exhibited up to 2 days of the last 3days<br>2. Indicator exhibited daily in the last 3 days | |
| | | INDICATORS OF DEPRESSION | |
| | a. Facial Expression | Sad, pained, worried facial expression (e.g. furrowed brow) | □ |
| | b. Tearfulness | Crying, tearfulness | □ |

**Figure 22 E-Intelligence user-interface form of MDS-MH data set displaying the Action menu**

The main aim of this E-Intelligence form design is to provide the user with better usage of different data mining algorithms and techniques. The controller is designed such that the user can perform all the data mining tasks by selecting a data set and applying different tasks on them. The data can be of any format and the data mining algorithms like classification, regression etc. and missing values handling methods are fed into and can be applied on the input data. The results are generated to the user by providing the replaced attributes. The user also has an option of adding new algorithms to the software. Given the available resources and the time constraints, only a selection of missing value handling algorithms are chosen, for example, WEKA, CRUISE, NORM and DTREG. Figure [23] shows the WEKA software running from E-Intelligence form. It can be seen from the Figure [23], the command prompt indicating the message “called from E-Intelligence form”.



```
\Microsoft Visual Studio\VB98>cd\

iles\Weka*

\Weka-3-4>echo $called for EIntelligence form
elligence form

\Weka-3-4>java -jar weka.jar
```

Clinical Examination Form

N NUMBERS

3 Marital S

○ 3. Widowed
○ 4. Seperated
○ 5. Divorced

Section CC. REFERRAL ITEMS

| | | | | |
|---|---|---|---|---|
| 1 | Admission Date | 2004<br>Year | 11<br>Month | 25<br>Day |
| 2 | Referral Source | Referring service provider or individual<br>1. Self<br>2. Emergency room<br>3. Community mental heath<br>4. Outpatient psychiatric service/clinic<br>5. Family physician | | |

Section CC. REFERRAL ITEMS (Cont'd)

ssional □
l agency

rs □
pendency □
g. marital,family,other) □
h problems (including drug sice □
hiatric illness (e.g. depressed mood, □
otection service □
l justice system □
al abuse □
l problem □
□
event □

Section B. MENTAL STATE INDICATORS

| | | |
|---|---|---|
| 1 | (Code for indicators observed in the last 3days) | 0. Indicator not exhibited in the las: 3days<br>1. Indicator not exhibited up to 2 days of the last 3days<br>2. Indicator exhibited daily in the last 3 days |
| | | INDICATORS OF DEPRESSION |
| | a. Facial Expression | Sad, pained, worried facial expression (e.g. furrowed brow |
| | b. Tearfulness | Crying, tearfulness |

**Figure 23 WEKA software running from the E-Intelligence user-interface**

Typically, any piece of software cannot handle all types data formats hence, data transformation is necessary for performing. MDS-MH data is mainly divided into categories with multiple choices. Most of the attribute values are 0,1,2,3 or 4 etc, which indicates particular choice. All of the software we use considers these values as numeric. Software like WEKA and NORM, replaces the missing values with mean-mode and multiple imputations respectively, which have no meaning. This misleads the data mining results. To avoid this, data transformation can be done by selecting the data transformation option available in the Action menu. Selecting this option, the user has to provide the data file, fields, data type to be changed and to which data type it has to be changed. By providing this information, the control goes to the Entry checker layer where these actions are performed.

Data can also be reduced to provide more efficient results. The attributes of MDS-MH are reduced to 255. Considering the case of diagnosing the patient to the categorized health care option, few attributes are eliminated as being irrelevant. The user can also have an option of reducing the data size to attain more efficient and accurate results. Choosing this option, the user is prompted with a window for providing the attribute names that are needed to be eliminated. Once the attribute names are given, the control goes to Entry checker to perform this action. The Entry checker also prompts the user for saving the reduced data file.

The data stored in the database from the E-Intelligence form is less erroneous and most of the missing values are reduced. The only missing values the database containing are unknown values. The user has an option of handling these missing values using different software like WEKA, CRUISE, NORM and DTREG. On selecting an option, the user will be directed to the selected software user-interfaces for performing the techniques.

## 4.2 Logic used in Entry Checker for Handling data

Entry checker is the layer where all the data preprocessing tasks, such as handling missing values, noise data, duplicate records, data transformation and reduction are performed. Reducing missing values and noise data can be performed at various levels of the data mining process. E-Intelligence form design of the MDS-MH data communicates with the Entry checker to reduce the missing values and noise in data during the data entry.

Different cases performed by the Entry checker to handle the data during the entry level are:

Case 1: Essential Info: Attributes like name of the patient, gender, birthdate, marital status, and admission date are considered as essential. The missing values for these attributes have no meaning for the records. Say, for instance, the name or the health card attribute values are missing for a record, it is hard to identify whose data the instance is referring to. Since in practice, health card number sometimes can be filled later, name attribute is considered compulsory in this design. During the E-Intelligence form entry of these attributes, the Entry checker forces the user that these attributes cannot have missing values.

Case 2: Wrong data types should be eliminated. For example, name of the patient can not be numeric. Attributes like gender, marital status and others should have numeric data types and cannot be entered other than the specified choices, that is gender attribute value should be either 1 or 2 indicating male or female but it cannot be 3 or 4, which has no meaningful value.

Case 3: Date attribute values are considered as numeric during entry stage and stored as date data type in the database. Special scenario is considered for the date attributes. Patient's birthdate cannot be greater than his/her admission date. Admission year cannot be the same as birth of year in MDS-MH. Month cannot be greater than 12 or less than 1 and day cannot be greater than 31 or less than 1. Consistency between month and year must be considered.

Month = 1, 3, 5, 7, 8, 10 and 12 can have day <=31

Month = 4, 6, 9 and 11 can have day<=30

Month = 2 can have day<=28

Leap year is also considered when.

Month=2, can have day<=29

Case 4: Attributes that are unknown are considered different from inapplicable. The user is prompted to make sure if the attribute value is unknown or not entered by mistake. Unknown values are stored in the database with an indication of "-1". Existing MDS-MH data has "99" for missing values. Various handling missing values methods can be applied to these attributes before processing the data mining techniques.

Case 5: Consistency between attributes should also be considered an issue. For example, anxiety disorder patient have certain mental state indicators. Machine intelligence techniques are necessary for the extraction of expert rules. Medical experts are needed for identifying the contradictions among the rules for this case study. The major problem here is discovering sufficient, complete and comparable sets of expert rules and data-driven rules and analysis of these new rules by a medical expert. In such cases, fuzzy logic represents a feasible and useful alternative. Earliest fuzzy systems were constructed using knowledge provided by human experts and were linguistically correct. Now they are replaced with new data-driven fuzzy modeling techniques [31]. Fuzzy logic can be used in this case for finding the boundaries and rules. Different “if-then” rules are developed to identify the attribute consistencies and check for the values accuracy.

Case 6: Before the data is stored in the database it checks for duplicate records.

1. Patient having the same name, birthdate, gender, and health card number cannot be added into the database.
2. Patient having same birthdate, gender, and health card number but different name cannot be added into the database.
3. Patient having the same birthdate and a few other same attribute values are checked once again before adding into the database.
4. Patient having same attribute values other than birthdate, health card number are checked once again before adding into the database.

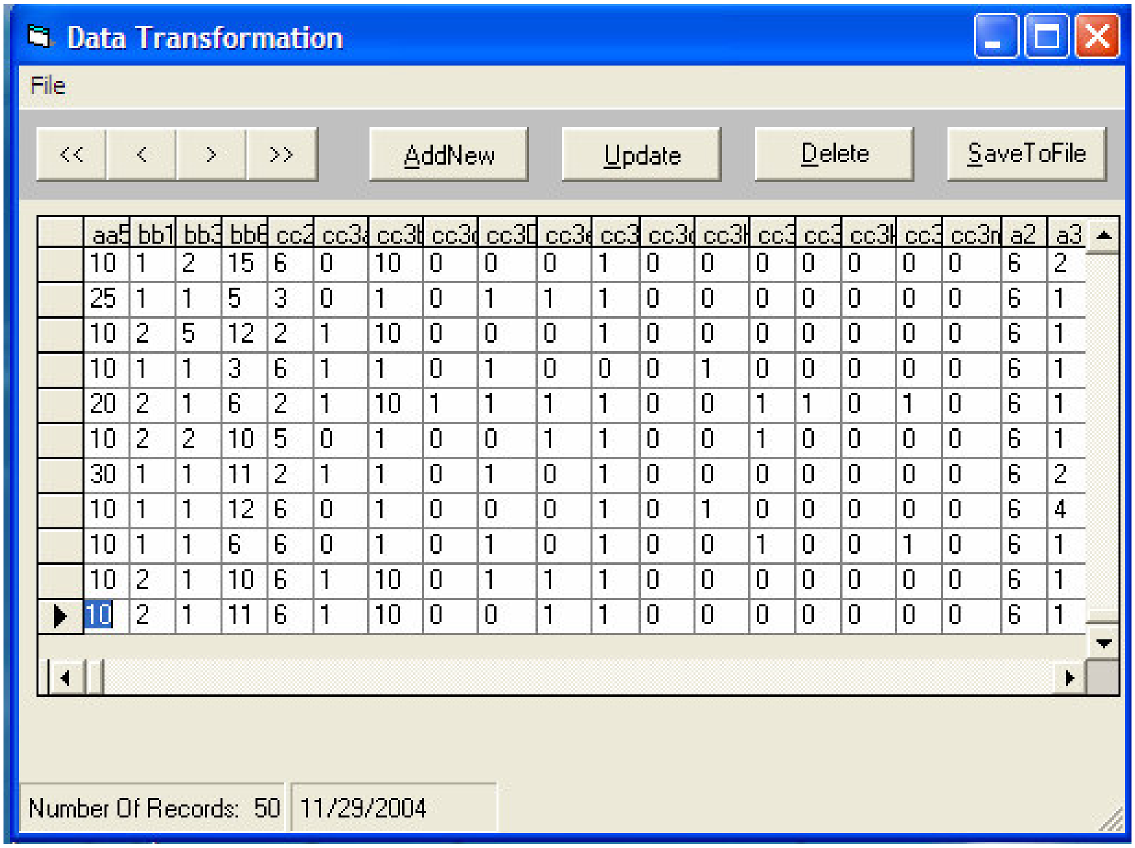
Point 3 and 4 are similar apart, point 4 considered all the attributes but point 3 considered few attribute cases only.

Case 7: The application has also the option of retrieving and modifying the stored data. Data can be retrieved using a case record number, patient’s name, health card number or birthdate. Once the data is modified, it can be saved which replaces the old existing data record with the new modified one.

Case 8: The data that is stored in the database is less erroneous and most of the attribute values are complete. The only missing values now indicate unknown data. These can be handled using different missing handling methods. But the data now is stored in .dbf format, which needs to be converted to any format that can be read by the chosen software.

Case 9: Already stored MDS-MH data can also be selected to reduce the noise in the data. The E-Intelligence form prompts the Entry checker to perform the data preprocessing task on the existing data, like reducing noise in data by checking the attribute values errors and finding the consistency between attributes and checks for their accurate values.

Case 10: The Entry checker layer can also perform other data preprocessing tasks like data transformation and reduction. To obtain better data mining results, the data is needed to be transformed and reduced. Say for instance, in MDS-MH data the patient's classification categorization is stored as 1, 2, 3 or 4, these attribute values can be transformed to dependent data types as 'a', 'b', 'c', and 'd'. The data is transformed for NORM software. The categorical data is transformed to binary data since NORM software can handle mainly numerical/continuous data. Some attributes can be reduced to perform the data mining task to attain efficient and accurate results. This reduction can also be done in Entry checker layer for an existing data. Figure [24] shows the data transformation task of the E-Intelligence software.



| | aa5 | bb1 | bb3 | bb6 | cc2 | cc3a | cc3b | cc3c | cc3D | cc3e | cc3 | cc3c | cc3h | cc3 | cc3 | cc3k | cc3 | cc3n | a2 | a3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 1 | 2 | 15 | 6 | 0 | 10 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 2 |
| | 25 | 1 | 1 | 5 | 3 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 1 |
| | 10 | 2 | 5 | 12 | 2 | 1 | 10 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 1 |
| | 10 | 1 | 1 | 3 | 6 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | 1 |
| | 20 | 2 | 1 | 6 | 2 | 1 | 10 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 6 | 1 |
| | 10 | 2 | 2 | 10 | 5 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 6 | 1 |
| | 30 | 1 | 1 | 11 | 2 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 2 |
| | 10 | 1 | 1 | 12 | 6 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | 4 |
| | 10 | 1 | 1 | 6 | 6 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 6 | 1 |
| | 10 | 2 | 1 | 10 | 6 | 1 | 10 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 1 |
| ▶ | 10 | 2 | 1 | 11 | 6 | 1 | 10 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 1 |

Number Of Records: 50 11/29/2004

**Figure 24 Data transformation task of E-Intelligence user-interface for MDS-MH data**

Cases 11: Once the data is stored in the database, different data mining techniques can be applied for diagnosis. The application has an option that directs to different software for performing the data mining tasks.

Data preprocessing is the main step in the knowledge discovery process. Using this Entry checker one can obtain or select quality data for attaining efficient and accurate mining results. The underlying logic used in this Entry checker for performing data preprocessing tasks can be used for any type of databases which has need for some intelligence in the data entry form.

## 4.3 Experiments for Handling Missing Values on MDS-MH data

Various missing value handling methods and software used to perform these methods are presented and explained with the breast cancer case study in Chapter 3. A subset of MDS-MH data is considered to check for the accuracy obtained by these methods. Several cases are considered in checking the accuracy obtained for the methods used in WEKA, CRUISE, NORM and DTREG software to handle the missing values.

The MDS-MH data set is reduced to a subset of 18 attributes and 200 instances. There are four experiments that are performed on this data set to classify the patients into the four categories of mental health by handling the missing values using the four software. The main objective of these experiments is to show how the accuracy is decreasing with an increase in the number of missing values. MDS-MH data is categorical data. NORM software cannot handle nominal values. Imputations can be performed on real value data types only and the data is transformed to binary format.

The three experiment cases are as follows:

- Experiment 1: - Considering a subset of MDS-MH data as mentioned above for 50 instances that have no missing values and classifying this data using the three software WEKA, CRUISE NORM and DTREG.

- Experiment 2: - Eliminate few attribute values in the 200 instances randomly and handle these missing values using the three software and check for the accuracy.

- Experiment 3: - Eliminate most of the values for a single attribute and handle them using the three methods by the software.

## 4.3.1 Experiments using WEKA

A subset of MDS-MH data is considered and converted into .arff file. The values of the categorization attributes are transformed from numeric to categorical. The values of MDS-MH data are numeric values but they represented as nominal values. WEKA uses mean-mode method for handling the missing values. Due to this representation, WEKA considers the data as numeric values and replaces the missing values with their mode values, which has no meaning and misleads the data mining results. So, the data should be transformed to nominal values and then perform the ReplacingMissingValues filter, which uses mode method to replace the missing attribute values. WEKA does have a data transformation filter, to transform the data to different formats like binary, nominal, integer. Diagnosing the data after using the data transformation (to nominal values) and ReplacingMissingValues filter, the results are shown below:

Accuracy obtained for the three experiments:

| Accuracy | Exp 1 | Exp 2 | Exp 3 |
|---|---|---|---|
| | 67% | 67.5% | 68% |

**Table 22: Accuracy obtained for the three experiments by WEKA (J4.8) software**

Most of the missing values are under class ‘a’. Experiment3 has 195 instances (98%) missing for variable ‘cc3a’. The confusion matrix gives a rough estimate on how the data is classified.

Confusion matrix for the three experiments transformation:

| Exp 1 | | | | Exp 2 | | | | Exp3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | A | B | C | D | A | B | C | D |
| 76 | 13 | 11 | 1 | 77 | 13 | 9 | 2 | 77 | 15 | 8 | 1 |
| 10 | 31 | 4 | 5 | 9 | 31 | 5 | 5 | 10 | 34 | 5 | 1 |
| 7 | 3 | 21 | 0 | 7 | 3 | 21 | 0 | 9 | 3 | 19 | 0 |
| 4 | 7 | 1 | 6 | 4 | 7 | 1 | 6 | 5 | 6 | 1 | 6 |

**Table 23: Confusion matrix obtained for the three experiments by WEKA software**

There are no much differences between the accuracies and the confusion matrix for the four experiments. Classifying the data using CRUISE software, after handling the missing values using mean-mode method:

Accuracy obtained for the three experiments using CRUISE software after handling the missing values using mean-mode method of WEKA:

| Accuracy | Exp 1 | Exp 2 | Exp 3 |
|---|---|---|---|
| | 74% | 74% | 69.5% |

**Table 24: Accuracy obtained for the three experiments by CRUISE software, after handling missing values using mean-mode method of WEKA**

Confusion matrix for the three experiments after handling the missing values using mean-mode method of WEKA:

[illegible]

| Exp 1 | | | | Exp 2 | | | | Exp3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | A | B | C | D | A | B | C | D |

Experiment1 has only one case since there are no missing values to handle. The accuracy obtained for this experiment is 74%. Case 3 in Cruise is eliminated since it can work only for univariate method. The below results are performed before transforming the data.

Accuracy obtained for Experiment1 using the four cases:

| | Case 1 | Case 2 | Case 4 |
|---|---|---|---|
| Exp 2 | 71% | 73.5% | 73.5% |
| Exp 3 | 84.5% | 84.5% | 50.5% |

**Table 27: Accuracy obtained for the two experiments by CRUISE software**

Confusion matrix for the first experiment:

| Experiment 1 | A | B | C | D |
|---|---|---|---|---|
| | 73 | 17 | 10 | 1 |
| | 3 | 41 | 4 | 2 |
| | 1 | 1 | 29 | 0 |
| | 1 | 11 | 1 | 5 |

**Table 28: Confusion matrix for the Experiment1 by CRUISE software**

Confusion matrix for the other two experiments:

|  | Exp 2 |  |  |  | Exp 3 |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | A | B | C | D | A | B | C | D |
| Case 1 | 70 | 19 | 12 | 0 | 91 | 0 | 10 | 0 |
|  | 3 | 42 | 5 | 0 | 0 | 50 | 0 | 0 |
|  | 0 | 1 | 30 | 0 | 3 | 0 | 28 | 0 |
|  | 2 | 15 | 1 | 0 | 0 | 18 | 0 | 0 |
| Case 2 | 73 | 17 | 10 | 1 | 91 | 0 | 10 | 0 |
|  | 3 | 41 | 4 | 2 | 0 | 50 | 0 | 0 |
|  | 1 | 1 | 29 | 0 | 3 | 0 | 28 | 0 |
|  | 1 | 12 | 1 | 4 | 0 | 18 | 0 | 0 |
| Case 4 | 73 | 17 | 10 | 1 | 101 | 0 | 0 | 0 |
|  | 3 | 41 | 4 | 2 | 50 | 0 | 0 | 0 |
|  | 1 | 1 | 29 | 0 | 31 | 0 | 0 | 0 |
|  | 1 | 12 | 1 | 4 | 18 | 0 | 0 | 0 |

**Table 29: Confusion matrix obtained for the two experiments by CRUISE software**

The results obtained from the three experiments are more accurate compared to WEKA software apart from the last case of Exp3. Some of the cases indicate the same results, due to the attribute values. The missing value handling methods are done during the building of the tree as most of the attributes values are alike and there is only a slight difference in the accuracy.

## 4.3.3 Experiments using NORM

NORM software is used to perform the data processing task, handling the missing values. Since experiment1 has no missing values, it is eliminated in this case. The data set of experiment2 and experiment3 are converted to ".dat" files and the missing values of these files are handled using the NORM software. Since NORM can handle mainly continuous data the categorical data of these files are transformed to binary. After handling the missing values, the data is classified for patient assessment using the three classification methods of WEKA, CRUISE and DTREG.

Accuracy obtained for classification using WEKA (J4.8) method:

| Classification using WEKA (J4.8) after handling the missing values with NORM | Exp 2 | Exp 3 |
|---|---|---|
| | 68% | 67.5% |

**Table 30: Accuracy obtained for the two experiments by WEKA software, after handling the missing values using NORM**

Confusion matrix obtained by WEKA software:

| Exp 2 | | | | Exp3 | | | |
|---|---|---|---|---|---|---|---|
| A | B | C | D | A | B | C | D |
| 77 | 13 | 10 | 1 | 78 | 15 | 7 | 1 |
| 9 | 31 | 5 | 5 | 10 | 33 | 5 | 2 |
| 6 | 3 | 22 | 0 | 10 | 3 | 18 | 0 |
| 4 | 7 | 1 | 6 | 5 | 6 | 1 | 6 |

**Table 31: Confusion matrix obtained for the two experiments by WEKA software, after handling the missing values using NORM multiple imputation method**

Classifying using CRUISE software for the two experiments:

Accuracy obtained by CRUISE software after handling the missing values using NORM:

| Classification using CRUISE after handling the missing values with NORM | Exp 2 | Exp 3 |
|---|---|---|
| | 73.5% | 70% |

**Table 32: Accuracy obtained for the two experiments by CRUISE software, after handling the missing values using NORM**

Confusion matrix obtained by CRUISE software:

| Exp 2 | | | | Exp3 | | | |
|---|---|---|---|---|---|---|---|
| A | B | C | D | A | B | C | D |
| 72 | 17 | 11 | 1 | 67 | 22 | 12 | 0 |
| 3 | 41 | 4 | 2 | 2 | 44 | 4 | 0 |
| 1 | 1 | 29 | 0 | 1 | 1 | 29 | 0 |
| 1 | 11 | 1 | 5 | 2 | 15 | 1 | 0 |

**Table 33: Confusion matrix obtained for the two experiments by CRUISE software, after handling the missing values using NORM multiple imputation method**

Classifying the data using DTREG software for the two experiments:

Accuracy obtained by DTREG software after handling the missing values using NORM:

| Classification using CRUISE after handling the missing values with NORM | Exp 2 | Exp 3 |
|---|---|---|
| | 67% | 56% |

**Table 34: Accuracy obtained for the two experiments by DTREG software, after handling the missing values using NORM**

From the experiments results, CRUISE software shows better accuracy compared to the other two.

## 4.3.4 Experiments using DTREG

The data set is converted to ".csv" file and trained using the DTREG software. It uses surrogate splits and put them into a most probable group are the methods to handle the missing values. Both the missing values handling methods are used to perform on this data set.

Accuracy obtained for the three experiments before transforming the data:

| Exp 1 | Exp 2 | Exp 3 |
|---|---|---|
| 68% | 68% | 52% |

**Table 35: Accuracy obtained for the three experiments by DTREG software**

There is no much difference between experiment1 and experiment2. The accuracy is much less for experiment3 compared to CRUISE software.

## 4.4 Conclusion for the MDS-MH Case Study

The results obtained from the three experiments for the MDS-MH data, performed on the four software, indicate some difference. The software when used for performing on the breast cancer data gave better results compared to the MDS-MH data. This is due to the data values in the MDS-MH data. Every method gives different results for different data types. However, the results from experiment3 indicates the more the missing values for a particular attribute, the less the accuracy results are. It also indicates by handling the missing values the results are more accurate than without handling.

Accuracy obtained for Experiment1 using the four software:

| Handling missing values | Classification | | |
|---|---|---|---|
| | WEKA (J4.8) | CRUISE | DTREG |
| WEKA | 67% | 74% | 68% |
| NORM | NA | NA | NA |
| CRUISE | NA | 74% | NA |
| DTREG | NA | NA | 68% |

**Table 36: Accuracy obtained by the four software for experiment1**

Accuracy obtained for Experiment2 using the four software:

| Handling missing values | Classification | | |
|---|---|---|---|
| | WEKA (J4.8) | CRUISE | DTREG |
| WEKA | 67.5% | 74% | 68% |
| NORM | 68% | 73.5% | 67% |
| CRUISE | NA | 73.5% | NA |
| DTREG | NA | NA | 68% |

**Table 37: Accuracy obtained by the four software for experiment2**

Accuracy obtained for Experiment3 using the four software:

| Handling missing values | Classification | | |
|---|---|---|---|
| | WEKA (J4.8) | CRUISE | DTREG |
| WEKA | 68% | 69.5% | 55% |
| NORM | 67.5% | 70% | 56% |
| CRUISE | NA | 84.5% | NA |
| DTREG | NA | NA | 52% |

**Table 38: Accuracy obtained by the four software for experiment3**

The results obtained shows CRUISE software gives much better results for any kind of data. In any case, handling missing values indicate better results compared to the ones without handling. The E-Intelligence form can thus be used to reduce the number of missing values.

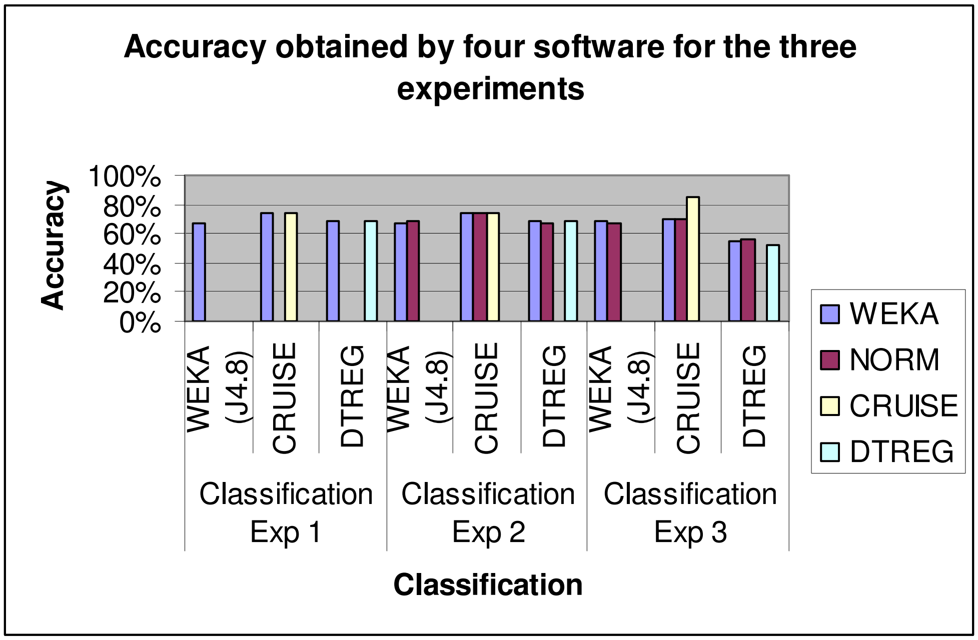


**Figure 25 Accuracy chart for the MDS-MH data**

## 4.5 Summary

In most cases, missing values are removed by deleting the records that contain missing information. Recently, the methods for analyzing the incomplete data are increasing. Different methods of handling missing values give different results. Chapter 4 discusses how the E-Intelligence form for the MDS-MH data is developed, data checked at entry and other preprocessing tasks like data transformation and reduction are performed. A subset of MDS-MH data is tested to demonstrate how to use the software for handling the missing values and other data preprocessing tasks.

# Chapter 5

# Conclusion and Future Work

## 5.1 Conclusion

Electronic applications in clinical health care are an emerging area with great economic and health importance. They have all the potential to reduce the cost of health care while increasing the quality. Various data mining tools and algorithms are used in health informatics to provide accurate and efficient decision making for quality health care. Data preprocessing is one of the main task in knowledge discovery for obtaining high quality mining results. This Thesis describes various methods of handling missing values that are applied to data preprocessing.

We have implemented the E-Intelligence form design for MDS-MH data set that consists of 455 attributes. The E-Intelligence form handles the data during different stages of the KDD process. The interaction between the E-Intelligence form and the Entry checker shows how the data is handled during the entry level. Different cases in Entry checker are discussed for rectifying the missing values during the data entry level. Other data preprocessing tasks such as data reduction and data transformation are also performed with this software. Identifying duplicate records and erroneous data are also implemented before the data is stored in the databases. The only missing values obtained in the database now are the unknown values. Different software are used to handle the missing values during the data mining stage. How the software redirects to different software to perform the data preprocessing task, handling missing values and data mining task is also described in this Thesis.

In this Thesis we used four software tools, which used different methods of handling missing values. The breast cancer data set was used as a case study to explain how these methods work. Later a subset of MDS-MH data was taken that consisted of 18 attributes and

200 instances and four experiments were performed with the software to find the accuracy obtained by handling the missing values and reducing the missing values. The multiple imputation method of NORM did not work well for the MDS-MH data compared to the in success breast cancer case study. This is due to the categorical data format of the MDS-MH data. All the software did give better accuracy results for breast cancer compared to the MDS-MH data. This can be due to the number of missing values and the data type. But in any case, handling the missing values is expected to give results with better accuracy when compared to the ones without handling. Overall, the reducing the missing values during data entry and by using different missing values methods results in better performance. They also depend on the data mining method used to classify the instances. For both the case studies CRUISE method has given better accurate results. This indicates, CRUISE method can be used for any kind of data.

## 5.2 Future Work.

This Thesis focused on the E-Intelligence form design of MDS-MH for handling missing values and erroneous data during entry by developing electronic forms. Further, the work on reducing errors of the already existing paper based MDS-MH forms can be done by electronic scanning. Before storing the data into the database, the errors, inconsistent attribute values and duplicate records can be identified and rectified using the same logic applied in the form design during the electronic paper scanning process.

This Thesis work can be extended in developing the user-interface where the user can have additional features like selecting or upgrading different missing values tasks and data mining tools for different kinds of databases. The user can also have the option of selecting only a few data type fields of the database and can perform the missing value methods to obtain accurate method for a particular data type. The user can also derive the conclusion of which method provides good results for a particular kind of data set.

Due to the role of mobile computing in today’s information retrieval system, PDA systems like handheld Blackberry, Microsoft ® pocket PC’s etc. can be connected to the internet for accessing data. The use of these systems is increasing in health care for quick and readily available health care. Further work on setting up the user-interface software on PDA systems or through Internet can be done.

# Bibliography

[1] Carroll M.R., M.S, "*Database design techniques for clinical research*", Oct 20. 2003

[2] "*Changing role of health care benefits*", 2001 Survey Report, Watson Wyatt, worldwide

[3] Charlos. A.P, "*Evolutionary fuzzy modeling human diagnostic decisions*", Swiss Federal institute of technology at Lausanne – EPFL, Ann. N. Y. Acad. Sci. 1020: 190-211 (2004)

[4] Christopher M.B., "*Neural networks for pattern recognition*", Oxford University Press, London, Jan 1996.

[5] Colleen M.E., Monique F., Robin W.C., "*Influence of missing values on Artificial Neural Network Performance*", V. Patel et al. (EDS), Amsterdam: IOS Press, 2001 IMIA

[6] Darmawan G.W., University of South Australia, *Norm software review: handling missing values with multiple imputation methods*, Evaluation Journal of Australasia, Vol. 2, No. 1, Aug 2002

[7] "*Dealing with missing values in data warehouse*", 1998, Stone bridge Technologies Inc., Dallas

[8] Dunamel A., Nultens M.C., Devos P., Picavet M., Beuscart R., "*A preprocessing method for improving data mining techniques*", Application to a large medical diabetes database, Pub Med, Stud Health Technol Inform. 2003;95:269-74

[9] Dunham M.H., "*Data mining: Introductory and advanced topics*", Prentice Hall, 2002

[10] Fayyad U., Piatetsky-Shapiro G., and Smyth P., *From data mining to knowledge discovery: An overview*. In Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramaswamy Uthuruswamy, editors, Advances in Knowledge discovery and data mining. AAAI/MIT Press, 1996.

[11] Greenes R.A., Pappalardo A.N., Marble C.W. and Barnett G.O., "*Design and Implementation of clinical data management system*", Pub Med, Comput Biomed Res. 1969 Oct;2(5):469-85

[12] Han J., and Kamber M., "*Data Mining: concepts and techniques*", SanFrancisco: Morgan Kaufman Publishers, 2001

[13] Hirdes J.P., Perez E., Curtin-Telegdi N., Predergast P., Morris J.N., Ikrgami N., Fries B.E., Phillips C., RAI-Mental Health (RAI-MH) ⑦ Training manual and resource Guide Version 1.0, 1999.

[14] Hirdes J.P., Fries B.E., Morris J.N., *et al.* "*Integrated Health Information Systems Based on the RAI/MDS Series of Instrument*" Healthcare Management Forum 12(4):30-40, 1999.

[15] Hirdes J.P., Marhaba M., Smith T.F., et al. (2001) Development of the Resident Assessment Instrument - Mental Health (RAI-MH), Hospital Quarterly, 4(2), 44-51

[16] Holmes G., Donkin A. and Witten I.H. (1994) WEKA: A Machine Learning Workbench *Proc Second Australia and New Zealand Conference on Intelligent Information Systems,* Brisbane, Australia

[17] Huang, H. etc. "Business rule extraction from legacy code", *Proceedings of 20th International Conference on Computer Software and Applications,* IEEE COMPSAC'96, 1996, pp.162-167

[18] John R.G., Cindy G., "*Those missing values in Questionnaires*" , Plymouth Meeting, PA

[19] Johnson R., Johnson R.L., "*Health care technology: A history of clinical care innovation*", HCT Project Vol 1

[20] Jonathan C.P., David F.L., Linda K.G., Joseph W.H., Marvin L.H and Edward H., "*Medical Data mining: Knowledge discovery in a clinical data warehouse*", Pub Med, Proc AMIA Annu Fall Symp. 1997;:101-5

[21] Kim H., and Loh W.-Y. (2003), Classification trees with bivariate linear discriminant node models, *Journal of Computational and Graphical Statistics*, vol. 12, pp. 512-530

[22] Marti A.H., "Untangling Text Data mining": *Proceeding of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20-26, 1999

[23] Martin R., "Supervised learning in Multilayer Perceptrons – from Backpropogation to Adaptive learning techniques", *Int. Journal of computer standards and interfaces* (16), 1994

[24] Michie. D., Spiegelhalter D.J. and Taylor C.C., "*Machine learning, Neural and Statistical Classification*", Ellis Horwood, 1994

[25] Mitchell T., "Decision tree learning", in Mitchell T., *Machine Learning*, The McGraw-Hill companies, Inc., 1997, pp. 52-78

[26] Nilsson N.J., "*Introduction to Machine learning*", Stanford University, CA, 1996

[27] Noam H.A., "*Rising to the challenge: Strategies for the new health care enterprise*", Pub Med, J Health Inf Manag. 2003 Summer, 17(3):9-11

[28] Rector A.L., Nowlan W.A., Kay S., et al. Foundations for an electronic medical record. *Methods of Information in Medicine,* 1991: 179-186

[29] Rudolf K,, Christian B., and Detlef N., "*Problems and prospects in Fuzzy data analysis*", pp:95-109, 2000, London, UK

[30] Shinomoto S., “Information classification scheme of feedforward networks organized under unsupervised learning”, 1990 Network: Comput. Neural sys. 1, 135-147 Issue 2 (April 1990)

[31] Sun Y., Karray F., “Hybrid soft computing techniques for heterogeneous data classification”, 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02, IEEE, 1511-16 vol.2, 2002, NJ, USA

[32] Svend J., “Take good care of data”, Dept of Epidemiology and Social Medicine, University of Aarhus, May 2003, Version 10 June 2003

[33] Sherrod H.P., “*Decision tree regression analysis for data mining and modeling*”, 2003-2004

[34] Willams G.J., and Huang Z., *Modelling the KDD process: A four stage process and four element process*, 1996 June, CSIRO

[35] Zhao F., and Yun T., *A Data Preprocessing Framework for Supporting Probability-Learning in Dynamic Decision Modeling in Medicine*, Pub Med, Proc AMIA Symp. 2000;:933-7

[36] Björg A., Svanfríður H., and Þóra J., *Mining the Wisconsin Prognostic Breast Cancer Data by using the WEKA software*: The final project in pattern recognition

[37] Quinlan J. R., C4.5 Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA

[38] Little R. J. A, Rubin D. B., Statistical Analysis with Missing Data, 1987, New York: John Wiley

[39] Darlene Goldstein, ‘Supervised Learning, Classification, Discrimination’, *Lecture notes on Supervised Learning*. Retrieved 10 Dec 2004 from http://statwww.epfl.ch/davision/teaching/Microarrays/

[40] William H. W., *Wisconsin Breast Cancer Database*, University of Wisconsin Hospitals, Madison, Wisconsin, USA, 1992.

[41] Anand V., ‘Issues in Decision Learning’, *Lecture notes on Decision trees*. Retrieved 10 Dec 2004 from http://www.speech.sri.com/people/anand/771/html/html.html

# Appendix A

## Univariate splits in CRUISE:

It uses available case solution for handling the missing values that occur in the learning sample. Here each variable is evaluated using only the cases non-missing in that variable at the node. The procedures used for the two cases in CRUISE are as follows:

1. For first method: computes the *p*-value of each *X* in the algorithm from the non-missing cases in *X*.
2. For second method: computes the *p*-value of each pair of variables in the algorithm from the non-missing cases in the pair.
3. If *X** is the selected split variable, it uses the cases with non-missing *X** values to find the split points
4. If *X** is a numeric variable, it uses the node sample class mean to impute the missing values in *X**. IF *X** is categorical use the class mode
5. Pass the imputed sample through the split.
6. Delete the imputed values and restore their missing values.

To process the future case for which the selected variable is missing at node *t*, it splits on an alternate variable.

## Linear combination splits in CRUISE:

It handles the missing values by imputing them with the node mean or mode values. It uses the strategy used in FACT and QUEST. If X and s are the selected variable and the split, the procedure is as follows:

1. If X is non-missing in the case, it uses s to predict its class. It then imputes all the missing values in the case with the means and modes of the numerical and categorical variables, respectively, for the predicted class.

2. If X is missing in the case, it imputes all the missing values with the grand means or modes in t, ignoring the class.

After the class is sent to the sub node, its imputed values are deleted and their missing status restored.

# Appendix B

## EM algorithm used in NORM:

It is a general method of obtaining maximum-likelihood estimation of parameters from incomplete data. This method is an iteration of two-steps:

- E-step: Given the observed data it replace the missing statistics by their expected values using the estimation values for the parameters
- M-step: Given the statistics obtained from the E-step, it updates the parameters by their maximum-likelihood estimates.

## DA algorithm used in NORM:

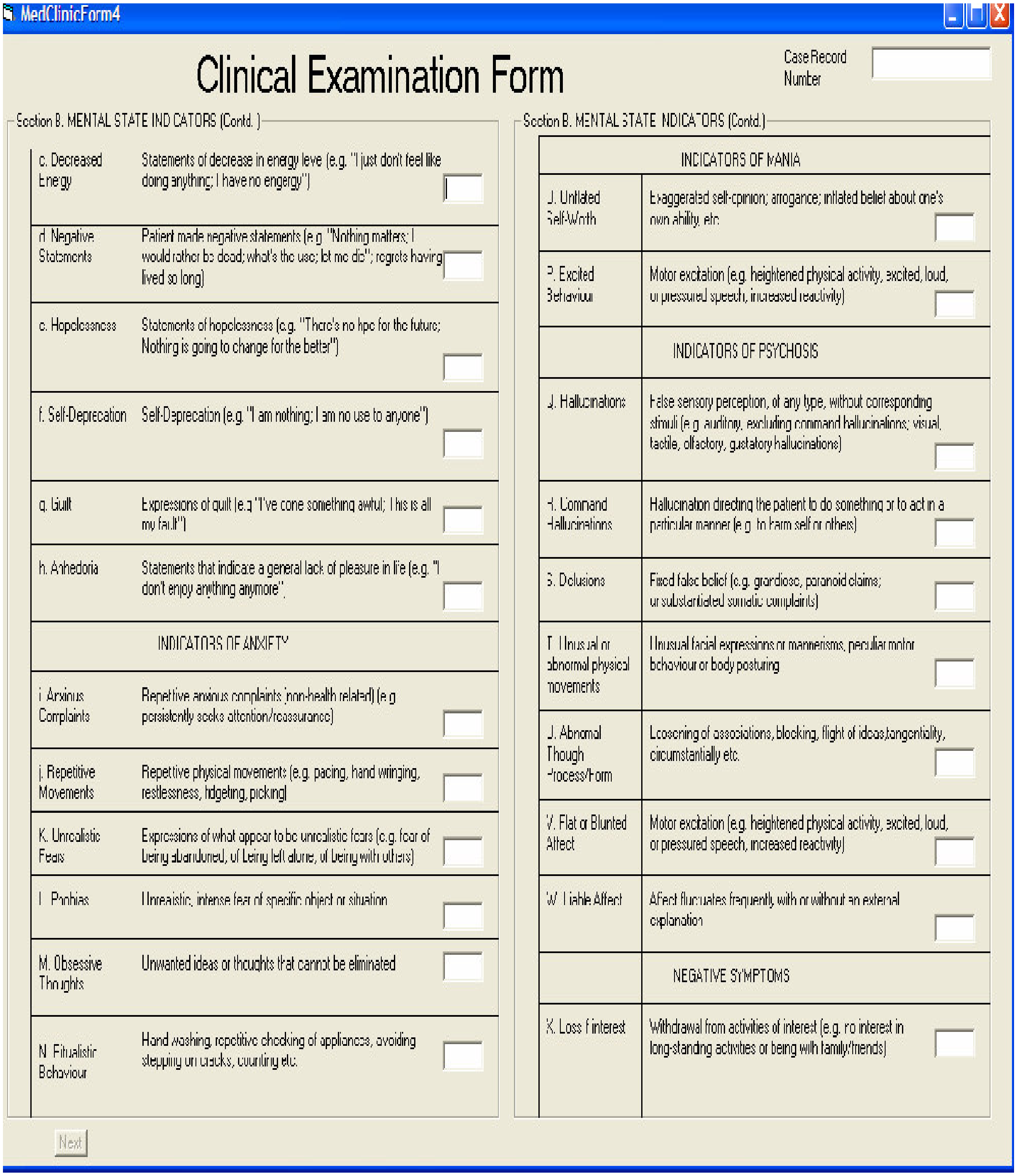
DA is an iterative simulation technique. It performs the following steps:

I-step: Given the observed data and the assumed values for the parameters, it imputes the missing data by drawing them from their conditional distribution.

P-step: Given the observed data and the most recently imputed values for the missing data, it simulates new values for the parameters by drawing them from a Bayesian posterior distribution.

# Appendix C

Forms designed for MDS-MH data set:



## Clinical Examination Form

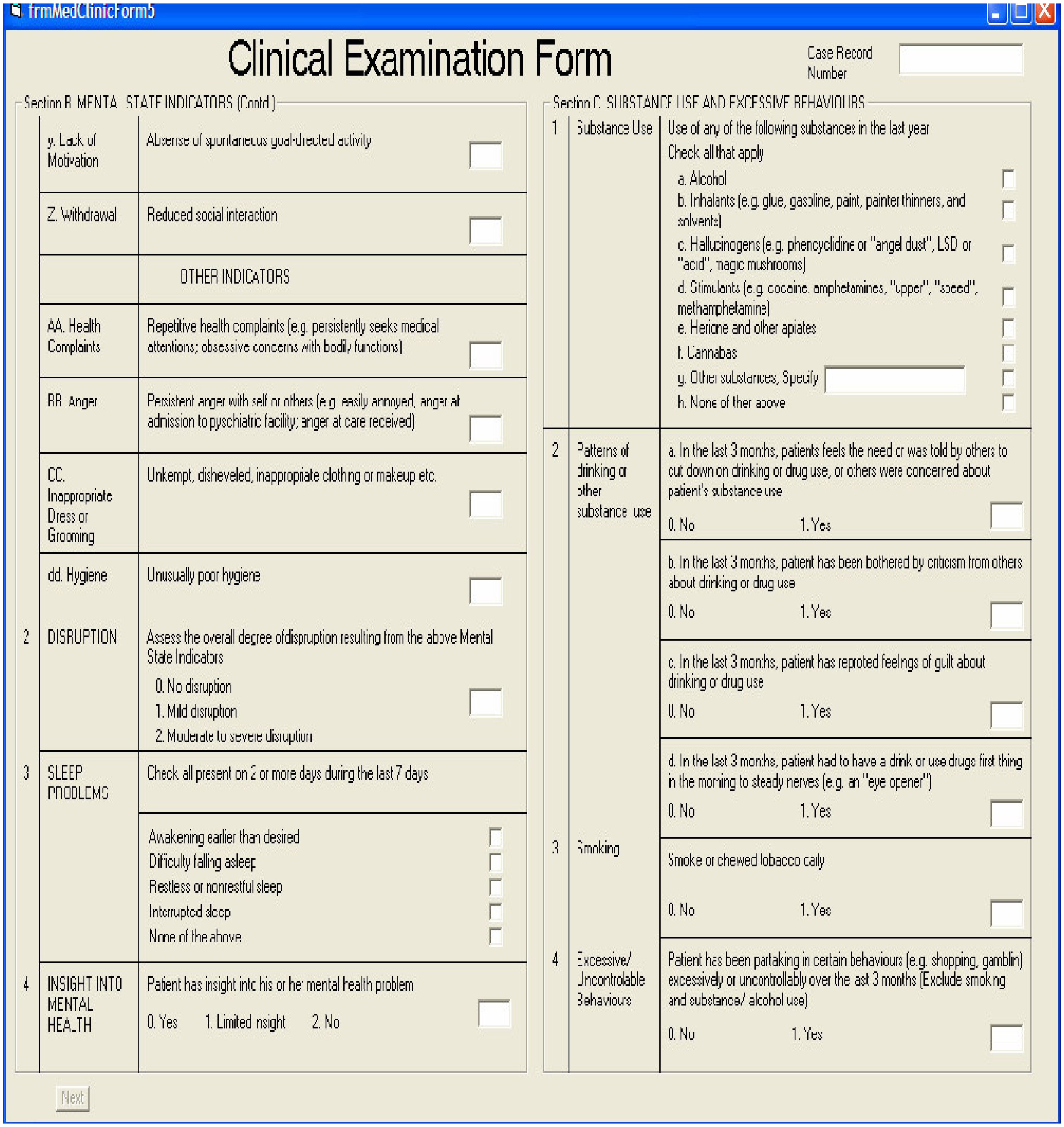
Case Record Number

Section B. MENTAL STATE INDICATORS (Contd.)

| | | |
|---|---|---|
| c. Decreased Energy | Statements of decrease in energy leve (e.g. "I just don't feel like doing anything; I have no engergy") | |
| d. Negative Statements | Patient made negative statements (e.g. "Nothing matters; I would rather be dead; what's the use; let me die"; regrets having lived so long) | |
| e. Hopelessness | Statements of hopelessness (e.g. "There's no hpe for the future; Nothing is going to change for the better") | |
| f. Self-Deprecation | Self-Deprecation (e.g. "I am nothing; I am no use to anyone") | |
| g. Guilt | Expressions of guilt (e.g "I've done something awful; This is all my fault") | |
| h. Anhedoria | Statements that indicate a general lack of pleasure in life (e.g. "I don't enjoy anything anymore") | |
| | INDICATORS OF ANXIETY | |
| i. Anxious Complaints | Repetitive anxious complaints (non-health related) (e.g persistently seeks attention/reassurance) | |
| j. Repetitive Movements | Repetitive physical movements (e.g. pacing, hand wringing, restlessness, fidgeting, picking) | |
| K. Unrealistic Fears | Expressions of what appear to be unrealistic fears (e.g. fear of being abandoned, of being left alone, of being with others) | |
| L. Phobias | Unrealistic, intense fear of specific object or situation | |
| M. Obsessive Thoughts | Unwanted ideas or thoughts that cannot be eliminated | |
| N. Ritualistic Behaviour | Hand washing, repetitive checking of appliances, avoiding stepping on cracks, counting etc. | |

Section B. MENTAL STATE INDICATORS (Contd.)

| | | |
|---|---|---|
| | INDICATORS OF MANIA | |
| O. Inflated Self-Worth | Exaggerated self-opinion; arrogance; inflated belief about one's own ability, etc. | |
| P. Excited Behaviour | Motor excitation (e.g. heightened physical activity, excited, loud, or pressured speech, increased reactivity) | |
| | INDICATORS OF PSYCHOSIS | |
| Q. Hallucinations | False sensory perception, of any type, without corresponding stimuli (e.g. auditory, excluding command hallucinations; visual, tactile, olfactory, gustatory hallucinations) | |
| R. Command Hallucinations | Hallucination directing the patient to do something or to act in a particular manner (e.g. to harm self or others) | |
| S. Delusions | Fixed false belief (e.g. grandiose, paranoid claims; unsubstantiated somatic complaints) | |
| T. Unusual or abnormal physical movements | Unusual facial expressions or mannerisms, peculiar motor behaviour or body posturing | |
| U. Abnormal Though Process/Form | Loosening of associations, blocking, flight of ideas,tangentiality, circumstantially etc. | |
| V. Flat or Blunted Affect | Motor excitation (e.g. heightened physical activity, excited, loud, or pressured speech, increased reactivity) | |
| W. Liable Affect | Affect fluctuates frequently with or without an external explanation | |
| | NEGATIVE SYMPTOMS | |
| X. Loss of interest | Withdrawal from activities of interest (e.g. no interest in long-standing activities or being with family/friends) | |



# Clinical Examination Form

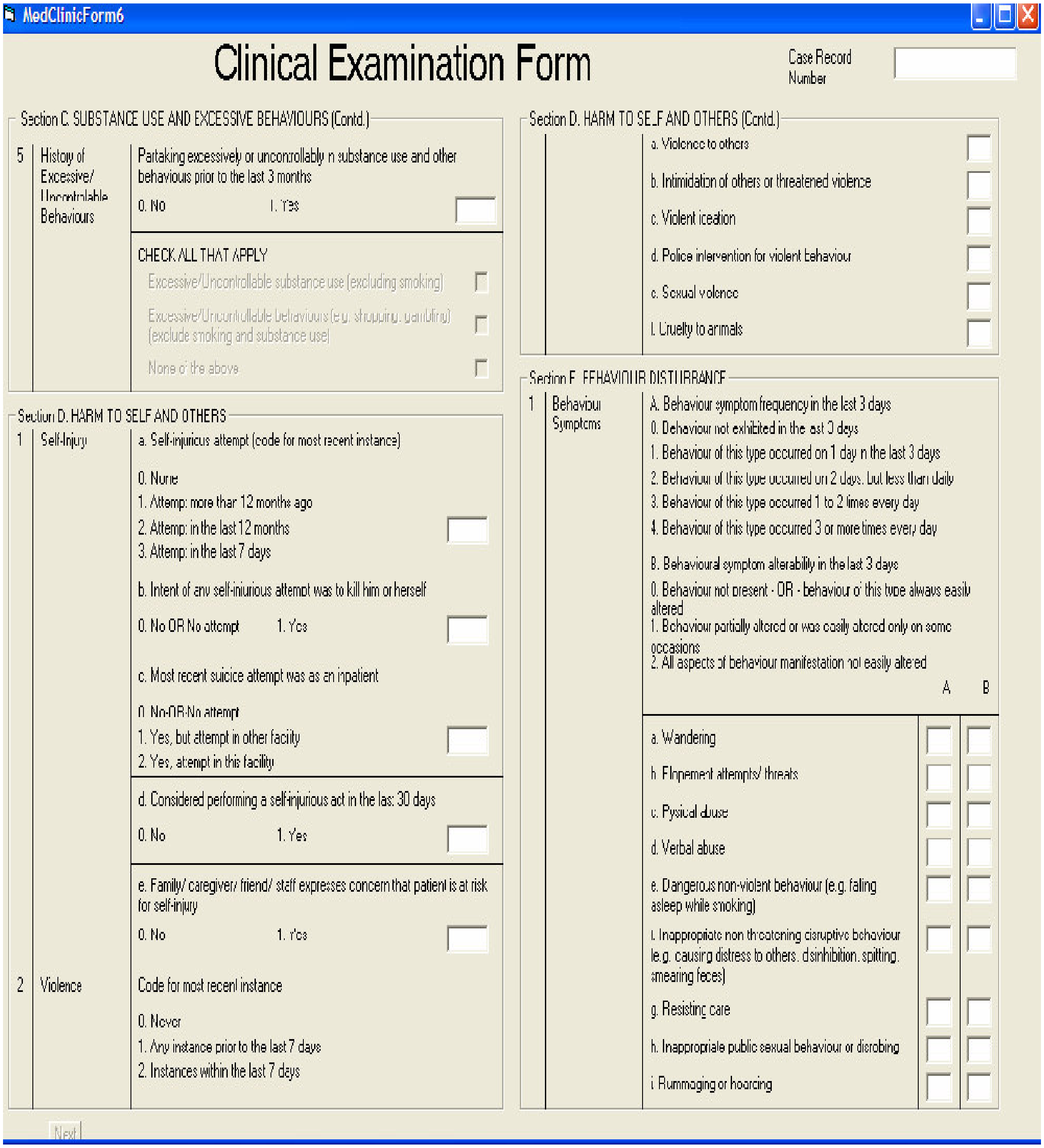
Case Record Number ☐

Section B MENTAL STATE INDICATORS (Contd.)

| | | | |
|---|---|---|---|
| | y. Lack of Motivation | Absense of spontaneous goal-directed activity | ☐ |
| | Z. Withdrawal | Reduced social interaction | ☐ |
| | | OTHER INDICATORS | |
| | AA. Health Complaints | Repetitive health complaints (e.g. persistently seeks medical attentions; obsessive concerns with bodily functions) | ☐ |
| | BB. Anger | Persistent anger with self or others (e.g. easily annoyed, anger at adnission to pyschiatric facility; anger at care received) | ☐ |
| | CC. Inappropriate Dress or Grooming | Unkempt, disheveled, inappropriate clothng or makeup etc. | ☐ |
| | dd. Hygiene | Unusually poor hygiene | ☐ |
| 2 | DISRUPTION | Assess the overall degree ofdisruption resulting from the above Mental State Indicators<br>0. No disruption<br>1. Mild disruption<br>2. Moderate to severe disruption | ☐ |
| 3 | SLEEP PROBLEMS | Check all present on 2 or more days during the last 7 days | |
| | | Awakening earlier than desired | ☐ |
| | | Dificulty falling asleep | ☐ |
| | | Restless or nonrestful sleep | ☐ |
| | | Interrupted sleep | ☐ |
| | | None of the above | ☐ |
| 4 | INSIGHT INTO MENTAL HEALTH | Patient has insight into his or her mental health problem<br>0. Yes 1. Limited insight 2. No | ☐ |

Section C SUBSTANCE USE AND EXCESSIVE BEHAVIOURS

| | | | |
|---|---|---|---|
| 1 | Substance Use | Use of any of the following substances in the last year<br>Check all that apply | |
| | | a. Alcohol | ☐ |
| | | b. Inhalants (e.g. glue, gasoline, paint, painter thinners, and solvents) | ☐ |
| | | c. Hallucinogens (e.g. phencyclidine or "angel dust", LSD or "acid", magic mushrooms) | ☐ |
| | | d. Stimulants (e.g. cocaine. amphetamines, "upper", "speed", methamphetamine) | ☐ |
| | | e. Herione and other apiates | ☐ |
| | | f. Cannabas | ☐ |
| | | g. Other substances, Specify ☐ | ☐ |
| | | h. None of ther above | ☐ |
| 2 | Patterns of drinking or other substance use | a. In the last 3 months, patients feels the need or was told by others to cut down on drinking or drug use, or others were concerned about patient's substance use<br>0. No 1. Yes | ☐ |
| | | b. In the last 3 months, patient has been bothered by criticism from others about drinking or drug use<br>0. No 1. Yes | ☐ |
| | | c. In the last 3 months, patient has reproted feelngs of guilt about drinking or drug use<br>0. No 1. Yes | ☐ |
| | | d. In the last 3 months, patient had to have a drink or use drugs first thing in the morning to steady nerves (e.g. an "eye opener")<br>0. No 1. Yes | ☐ |
| 3 | Smoking | Smoke or chewed tobacco daily<br>0. No 1. Yes | ☐ |
| 4 | Excessive/ Uncontrolable Behaviours | Patient has been partaking in certain behaviours (e.g. shopping, gamblin) excessively or uncontrollably over the last 3 months (Exclude smoking and substance/ alcohol use)<br>0. No 1. Yes | ☐ |



# Clinical Examination Form

Case Record Number □

Section C. SUBSTANCE USE AND EXCESSIVE BEHAVIOURS (Contd.)

5 History of Excessive/ Uncontrolable Behaviours

Partaking excessively or uncontrollably n substance use and other behaviours prior to the last 3 months

0. No 1. Yes □

CHECK ALL THAT APPLY

Excessive/Uncontrollable substance use (excluding smoking) □

Excessive/Uncontrollable behaviours (e.g. shopping, gambling) (exclude smoking and substance use) □

None of the above □

Section D. HARM TO SELF AND OTHERS

1 Self-Injury

a. Self-injurious attempt (code for most recent instance)

0. None
1. Attemp: more thar 12 months ago
2. Attemp: in the last 12 months □
3. Attemp: in the last 7 days

b. Intent of any self-injurious attempt was to kill him or herself

0. No OR No attempt 1. Yes □

c. Most recent suicice attempt was as an inpatient

0. No OR No attempt
1. Yes, but attempt in other facilty □
2. Yes, atempt in this facility

d. Considered performing a self-injurious act in the las: 30 days

0. No 1. Yes □

e. Family/ caregiver/ friend/ staff expresses concern that patient is at risk for self-injury

0. No 1. Yes □

2 Violence

Code for most recent instance

0. Never
1. Any instance prior to the last 7 days
2. Instances within the last 7 days

Section D. HARM TO SE_F AND OTHERS (Cntd.)

a. Violence to others □

b. Intimidation of others or threatened violence □

c. Violent iceation □

d. Police intervention for violent behaviour □

e. Sexual violence □

f. Cruelty to animals □

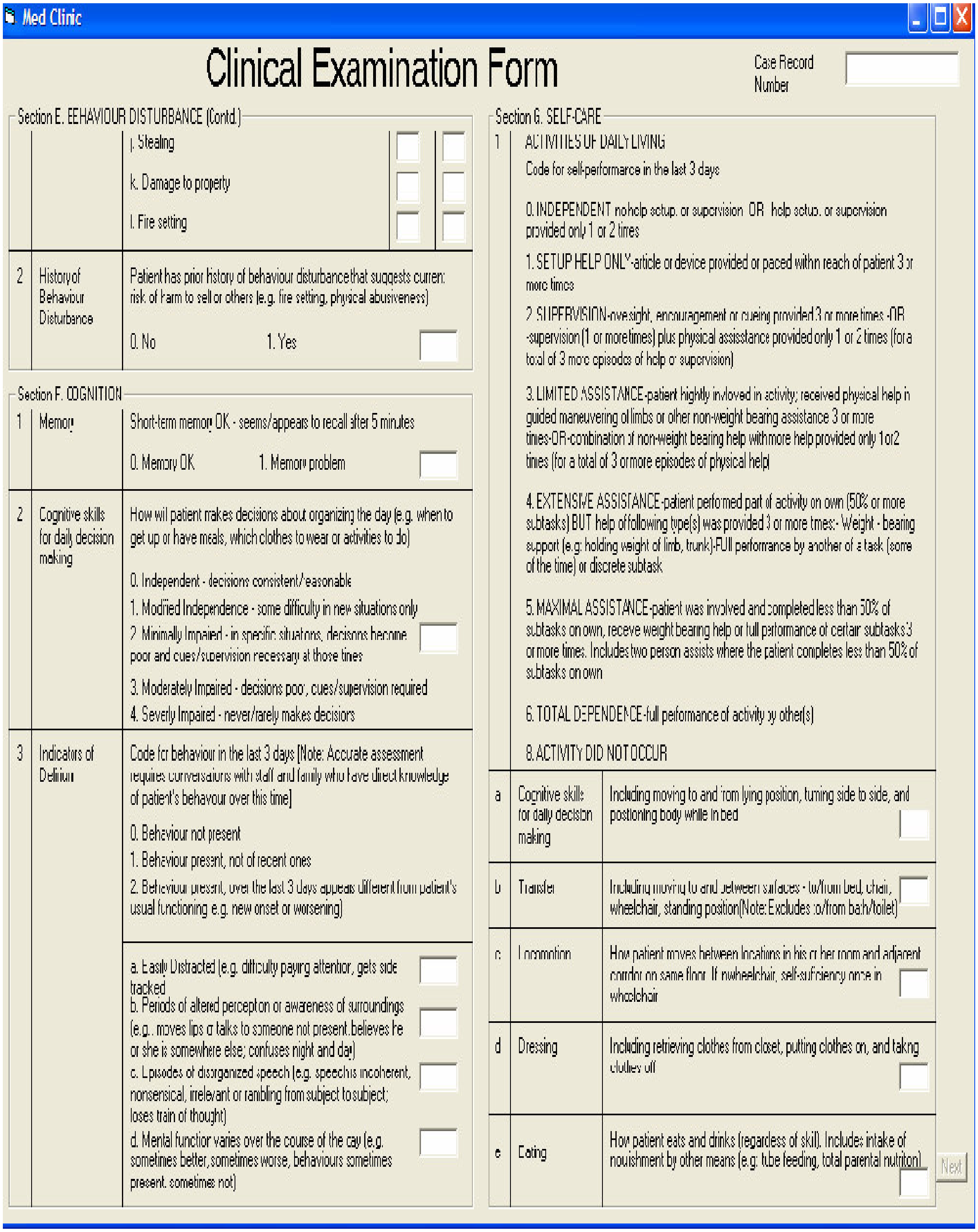
Section E. BEHAVIOUR DISTURBANCE

1 Behaviour Symptoms

A. Behaviour symptom frequency in the last 3 days
0. Behaviour not exhibited in the last 3 days
1. Behaviour of this type occurred on 1 day n the last 3 days
2. Behaviour of this type occurred on 2 days, but less than daily
3. Behaviour of this type occurred 1 to 2 times every day
4. Behaviour of this type occurred 3 or more times every day

B. Behavioural symptom alterability in the last 3 days
0. Behaviour not present - OR - behaviour of this type always easily altered
1. Behaviour partially altered or was easily altered only on some occasions
2. All aspects of behaviour manifestation not easily altered

| | A | B |
|---|---|---|
| a. Wandering | □ | □ |
| b. Elopement attempts/ threats | □ | □ |
| c. Pysical abuse | □ | □ |
| d. Verbal abuse | □ | □ |
| e. Dangerous non-violent behaviour (e.g. falling asleep while smoking) | □ | □ |
| f. Inappropriate non threatening disruptive behaviour (e.g. causing distress to others, disinhibition, spitting, smearing feces) | □ | □ |
| g. Resisting care | □ | □ |
| h. Inappropriate public sexual behaviour or disrobing | □ | □ |
| i. Rummaging or hoarding | □ | □ |



# Clinical Examination Form

Case Record Number □

Section E. EEHAVIOUR DISTURBANCE (Contd.)

| | | | | |
|---|---|---|---|---|
| | | j. Stealing | □ | □ |
| | | k. Damage to property | □ | □ |
| | | l. Fire setting | □ | □ |
| 2 | History of Behaviour Disturbance | Patient has prior history of behaviour disturbance that suggests curren: risk of harm to sell or others (e.g. fire setting, physical abusiveness)<br>0. No 1. Yes | □ | |

Section F. COGNITION

| | | | |
|---|---|---|---|
| 1 | Memory | Short-term memory OK - seems/appears to recall after 5 minutes<br>0. Memory OK 1. Memory problem | □ |
| 2 | Cognitive skills for daily decision making | How wil patient makes decisions about organizing the day (e.g. when to get up or have meals, which clothes to wear or activities to do)<br>0. Independent - decisions consistent/easonable<br>1. Modified Independence - some difficulty in new situations only<br>2. Minimally Impaired - in specific situatons, decisons become poor and cues/supervision recessary at those tines<br>3. Moderately Impaired - decisions poor, cues/supervision required<br>4. Severly Impaired - never/rarely makes decisions | □ |
| 3 | Indicators of Delirium | Code for behaviour in the last 3 days [Note: Accurate assessment requires conversations with staff and family who have direct knowledge of patient's behavour over this time]<br>0. Behaviour not present<br>1. Behaviour present, not of recent ones<br>2. Behaviour present, over the last 3 days appears different from patient's usual functioning e.g. new onset or worsening) | |
| | | a. Easily Distracted (e.g. difficulty paying attention, gets side tracked | □ |
| | | b. Periods of altered percepton or awareness of surroundings (e.g.. moves lips or talks to someone not present.believes he or she is somewhere else; confuses night and day) | □ |
| | | c. Episodes of disorganized speech (e.g. speechis incoherent, nonsensical, irrelevant or rambling from subject to subject; loses train of thought) | □ |
| | | d. Mental functior varies over the course of the cay (e.g. sometimes better, sometimes worse, behaviours sometimes present. sometimes not) | □ |

Section G. SELF-CARE

1 ACTIVITIES OF DAILY LIVING

Code for self-performance in the last 3 days

0. INDEPENDENT nohelp setup. or supervision OR help setup. or supervision provided only 1 or 2 times

1. SETUP HELP ONLY-article or device provided or paced withn reach of patient 3 or more times

2. SUPERVISION-oversight, encouragement or cueing provided 3 or more times -OR -supervision (1 or more times) plus physical assisstance provided only 1 or 2 times (for a total of 3 more episodes of help or supervision)

3. LIMITED ASSISTANCE-patient hightly invloved in activity; received physical help in guided maneuvering of limbs or other non-weight bearing assistance 3 or more times-OR-combination of non-weight bearing help withmore help provided only 1or2 times (for a total of 3 ormore episodes of physical help)

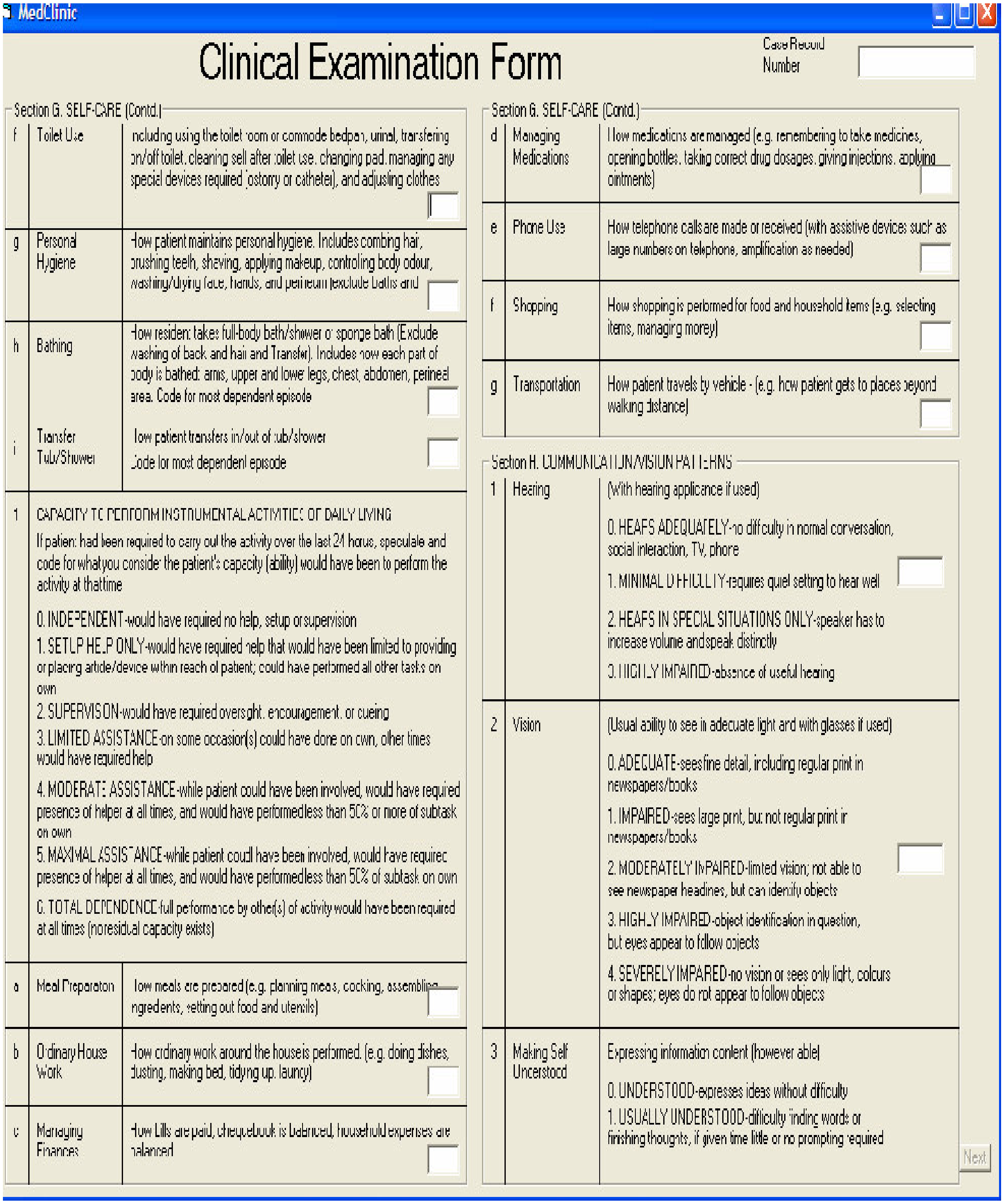
4. EXTENSIVE ASSISTANCE-patient performed part of activity on own (50% or more subtasks) BUT help offollowing type(s) was provided 3 or more tmes:- Weight - bearing support (e.g: holding weight of limb, trunk)-FUll performance by another of a task (some of the time) or discrete subtask

5. MAXIMAL ASSISTANCE-patient was involved and completed less than 50% of subtasks on own, receve weight bearing help or full performance of certain subtasks 3 or more times. Includes two person assists where the patient completes less than 50% of subtasks on own

6. TOTAL DEPENDENCE-full performance of activity by other(s)

8. ACTIVITY DID NOT OCCUR

| | | | |
|---|---|---|---|
| a | Cognitive skills for daily decision making | Including moving to and from lying position, turning side to side, and postioning body while in bed | □ |
| b | Transfer | Including moving to and between surfaces - to/from bed, chair, wheelchair, standing position(Note: Excludes to/from bath/toilet) | □ |
| c | Locomotion | How patient moves between locations in his or her room and adjacent corridor on same floor. If nwheelchair, self-sufficiency once in wheelchair | □ |
| d | Dressing | Including retrieving clothes from closet, putting clothes on, and takng clothes off | □ |
| e | Eating | How patient eats and drinks (regardless of skil). Includes intake of nouishment by other means (e.g: tube feeding, total parental nutriton) | □ |



# Clinical Examination Form

Case Record Number ☐

Section G. SELF-CARE (Contd.)

| | | | |
|---|---|---|---|
| f | Toilet Use | Including using the toilet room or commode bedpan, urinal, transfering on/off toilet, cleaning self after toilet use, changing pad, managing any special devices required (ostomy or catheter), and adjusting clothes | ☐ |
| g | Personal Hygiene | How patient maintains personal hygiene. Includes combing hair, brushing teeth, shaving, applying makeup, controling body odour, washing/drying face, hands, and perineum (exclude baths and | ☐ |
| h | Bathing | How resident takes full-body bath/shower or sponge bath (Exclude washing of back and hair and Transfer). Includes how each part of body is bathed: arms, upper and lower legs, chest, abdomen, perineal area. Code for most dependent episode | ☐ |
| i | Transfer Tub/Shower | How patient transfers in/out of tub/shower<br>Code for most dependent episode | ☐ |

1 CAPACITY TO PERFORM INSTRUMENTAL ACTIVITIES OF DAILY LIVING

If patient had been required to carry out the activity over the last 24 horus, speculate and code for whatyou consider the patient's capacity (ability) would have been to perform the activity at thattime

0. INDEPENDENT-would have required no help, setup orsupervision

1. SETUP HELP ONLY-would have required help that would have been limited to providing or placing article/device within reach of patient; could have performed all other tasks on own

2. SUPERVISION-would have required oversight, encouragement, or cueing

3. LIMITED ASSISTANCE-on some occasion(s) could have done on own, other times would have required help

4. MODERATE ASSISTANCE-while patient could have been involved, would have required presence of helper at all times, and would have performedless than 50% or more of subtask on own

5. MAXIMAL ASSISTANCE-while patient coudl have been involved, would have required presence of helper at all times, and would have performedless than 50% of subtask on own

6. TOTAL DEPENDENCE-full performance by other(s) of activity would have been required at all times (noresidual capacity exists)

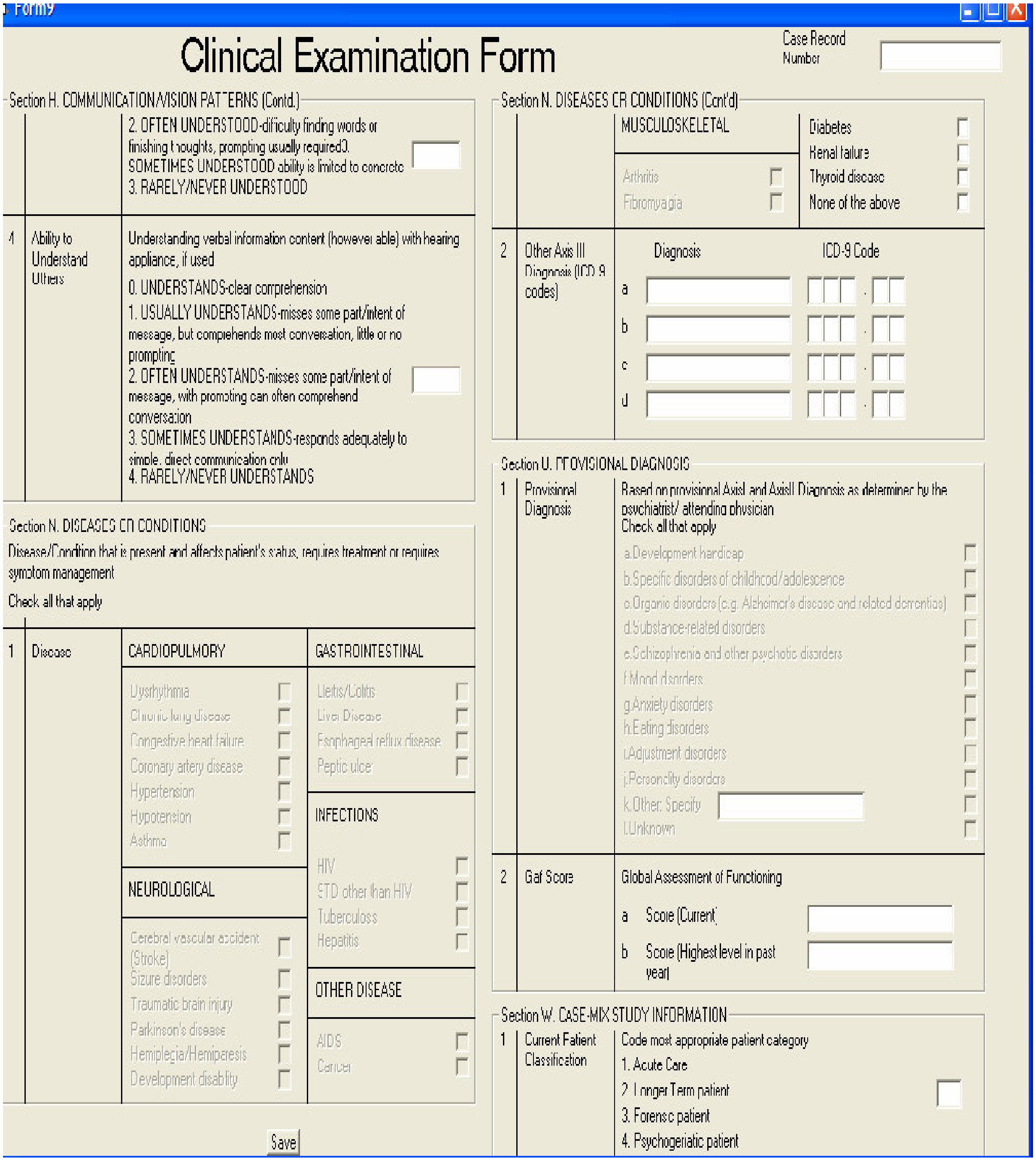
| | | | |
|---|---|---|---|
| a | Meal Preparaton | How meals are prepared (e.g. planning meals, cooking, assembling ingredents, setting out food and utensils) | ☐ |
| b | Ordinary House Work | How ordinary work around the house is performed. (e.g. doing dishes, dusting, making bed, tidying up, launcy) | ☐ |
| c | Managing Finances | How bills are paid, chequebook is balanced, household expenses are balanced | ☐ |

Section G. SELF-CARE (Contd.)

| | | | |
|---|---|---|---|
| d | Managing Medications | How medications are managed (e.g. remembering to take medicines, opening bottles, taking correct drug dosages, giving injections, applying ointments) | ☐ |
| e | Phone Use | How telephone calls are made or received (with assistive devices such as large numbers on telephone, amplification as needed) | ☐ |
| f | Shopping | How shopping is performed for food and household items (e.g. selecting items, managing money) | ☐ |
| g | Transportation | How patient travels by vehicle - (e.g. how patient gets to places beyond walking distance) | ☐ |

Section H. COMMUNICATION/VISION PATTERNS

| | | | |
|---|---|---|---|
| 1 | Hearing | (With hearing applicance if used)<br>0. HEARS ADEQUATELY-no difficulty in normal conversation, social interaction, TV, phone<br>1. MINIMAL DIFFICULTY-requires quiet setting to hear well<br>2. HEARS IN SPECIAL SITUATIONS ONLY-speaker has to increase volume andspeak distinctly<br>3. HIGHLY IMPAIRED-absence of useful hearing | ☐ |
| 2 | Vision | (Usual ability to see in adequate light and with glasses if used)<br>0. ADEQUATE-seesfine detail, including regular print in newspapers/books<br>1. IMPAIRED-sees large print, but not regular print in newspapers/books<br>2. MODERATELY IMPAIRED-limted vision; not able to see newspaper headines, but can identify objects<br>3. HIGHLY IMPAIRED-object identification in question, but eyes appear to follow objects<br>4. SEVERELY IMPARED-no vision or sees only light, colours or shapes; eyes do not appear to follow objects | ☐ |
| 3 | Making Self Understood | Expressing information content (however able)<br>0. UNDERSTOOD-expresses ideas without difficulty<br>1. USUALLY UNDERSTOOD-difficulty finding words or finishing thoughts, if given time little or no prompting required | |



# Clinical Examination Form

Case Record Number

Section H. COMMUNICATION/VISION PATTERNS (Contd.)

2. OFTEN UNDERSTOOD-dificulty finding words or finishing thoughts, prompting usually required3. SOMETIMES UNDERSTOOD ability is limited to concrete
3. RARELY/NEVER UNDERSTOOD

4 Ability to Understand Others

Understanding verbal information content (however able) with hearing appliance, if used

0. UNDERSTANDS-clear comprehension
1. USUALLY UNDERSTANDS-misses some part/intent of message, but comprehends most conversation, little or no prompting
2. OFTEN UNDERSTANDS-misses some part/intent of message, with prompting can often comprehend conversation
3. SOMETIMES UNDERSTANDS-responds adequately to simple, direct communication only
4. RARELY/NEVER UNDERSTANDS

Section N. DISEASES OR CONDITIONS

Disease/Condition that is present and affects patient's status, requires treatment or requires symptom management

Check all that apply

1 Disease

CARDIOPULMORY

- Dysrhythmia □
- Chronic lung disease □
- Congestive heart failure □
- Coronary artery disease □
- Hypertension □
- Hypotension □
- Asthma □

NEUROLOGICAL

- Cerebral vascular accident (Stroke) □
- Sizure disorders □
- Traumatic brain injury □
- Parkinson's disease □
- Hemiplegia/Hemiparesis □
- Development disablity □

GASTROINTESTINAL

- Uletis/Colitis □
- Liver Disease □
- Esophageal reflux disease □
- Peptic ulcer □

INFECTIONS

- HIV □
- STD other than HIV □
- Tuberculoss □
- Hepatitis □

OTHER DISEASE

- AIDS □
- Cancer □

Save

Section N. DISEASES OR CONDITIONS (Cont'd)

MUSCULOSKELETAL

- Arthritis □
- Fibromyagia □

- Diabetes □
- Renal failure □
- Thyroid disease □
- None of the above □

2 Other Axis III Diagnosis (ICD 9 codes)

| | Diagnosis | ICD-9 Code |
|---|---|---|
| a | | . |
| b | | . |
| c | | . |
| d | | . |

Section U. PROVISIONAL DIAGNOSIS

1 Provisional Diagnosis

Based on provisional AxisI and AxisII Diagnosis as determiner by the psychiatrist/ attending physician
Check all that apply

- a.Development hardicap □
- b.Specific disorders of childhcod/adolescence □
- c.Organic disorders (e.g. Alzheimer's disease and related dementias) □
- d.Substance-related disorders □
- e.Schizophrenia and other psychotic disorders □
- f.Mood disorders □
- g.Anxiety disorders □
- h.Eating disorders □
- i.Adjustment disorders □
- j.Personality disorders □
- k.Other: Specify □
- l.Unknown □

2 Gaf Score

Global Assessment of Functioning

a Score (Current)

b Score (Highest level in past year)

Section W. CASE-MIX STUDY INFORMATION

1 Current Patient Classification

Code most appropriate patient category
1. Acute Care
2. Longer Term patient
3. Forensc patient
4. Psychogeriatic patient

# Appendix D

## Data Mining:

Is an application of different intelligent algorithms to find patterns in data.

## Machine Learning:

The use of algorithms to generate a model from data. Machine learning can be used in data mining for prediction or classification.

# Appendix E

## Training Data:

A data set used to estimate or train a model based on the data set on hand.

## Testing Data:

A data set used to test prediction accuracy of a model.

## Validation:

The process of testing the models with a data set different from the training data set.

## Listwise Deletion (case deletion/ complete case analysis):

It omits the cases containing the missing values for at least one variable.

## Pairwise Deletion (available case method):

For each pair of instances, ignore all variables that are missing in any instance.

# Appendix F

## Over Fitting:

*Def:* Assume a hypothesis space H. A hypothesis h in H over fits a dataset D if there is another hypothesis h1 in H where h has better classification accuracy than h1 on D but worse classification accuracy than h1 on D1 [41].

## Tree Pruning:

To avoid over fitting, the tree is to be pruned, that is reduced. It uses separate training/ validation sets for building/ pruning the tree.

***Pre-Pruning**:* Stop the growing tree.

***Post-Pruning**:* Grow full tree and then prune.