Ensemble Classifier Modelling for Dealing with Missing Values

By

Mohammad Rajib Hasan

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR
THE
MASTER OF PHILOSOPHY,
AT
AUCKLAND UNIVERSITY OF TECHNOLOGY
AUCKLAND, NEW ZEALAND
JANUARY 2020

To my lovely wife, kids,

Parents, siblings and supervisors

Table of content

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma from a university or other institution of higher learning.

Mohammad Rajib Hasan
AUT, New Zealand.

Abstract

An ensemble classifier method for life critical data classification is considered one of the most capable classifiers where data suffers from missing values. The execution of a decision tree classifier can be expanded by the ensemble method as it is found to be the most superior method for single classifiers. Notwithstanding, the performance of an ensemble classifier relies upon the data quality and missing values. In this study, we discover that better classification accuracy is often achieved by missing value imputation. Medical experts do not have confidence in missing value imputation (filling up the missing values by any of the statistical methods) as each case/attribute is unique and possesses different possibilities. Missing value imputation in life critical data may lead to the wrong diagnosis and thus medical decision making may be influenced wrongly, which is dangerous and life threatening. This study, therefore, proposes a new ensemble model that can accomplish a preferred accuracy of over 96 percent without missing value imputation. The relevancy of features like HPV, HIV, AIDS, and smoking with cervical cancer is a long debate. This study successfully selected some of these influential features and validated their relevancy in terms of accuracy with statistical error root squared mean error and mean absolute error. This study also considers true-positive and false-positive rates in accuracy. Finally, this study concluded that missing value imputation in life critical data may not be necessary to obtain better accuracy. Selection of base classifiers in the ensemble method should be the prior concern over missing value imputation.

Acknowledgement

I am grateful to the Almighty for providing me with the capability to finish this difficult task. I would like to express my deepest gratitude to my supervisor, Professor Ajit Narayan, for all these years of support, supervision, and guidance. Without him, this thesis would never have been done. I feel honoured and fortunate to have been his student. My gratitude goes also to Associate Professor Nurul I Sarkar for his help and guidance.

I wish to convey my warm gratitude to Enrico Haemmerle for his patience, and constructive and bold decisions when the research needed management attention. Thank you, Enrico, for your dedication, enormous help, and for the guidance you provided towards my thesis. I also thank Hamid GholamHosseini, who held my hand and helped me in the research publication process.

This research was supported by the AUT Scholarship and stipend. I recognize that this research would not have been possible without its financial assistance and express my gratitude to AUT. I would like to thank my friends and colleagues at AUT for their support and encouragement over the last years.

Finally, I would like to express my love and gratitude to my wife, Tahlina Ahmed, and our children for giving full support during my M. Phil study. Their love has motivated me to complete the MPhil programme.

Auckland, New Zealand
July 2019
Mohammad Rajib Hasan

List of Publications

Hasan, M. R., Gholamhosseini, H., Sarkar, N. I., & Safiuzzaman, S. M. (2018). Intrinsic motivated cervical cancer screening intervention framework. *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. doi:10.1109/r10-htc.2017.8289009

Hasan, M. R., Gholamhosseini, H., & Sarkar, N. I. (2017). A new ensemble classifier for multivariate medical data. *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*.

List of Abbreviations and Acronyms

DT	Decision Tree
DS	Decision Stump
RF	Random Forest
RNT	Random Tree
RT	Rep Tree
ANN	Artificial Neural Network
SVM	Support Vector Machine

Chapter 1
Introduction

The introduction of this research details the purpose of the study and the research problem, offers a justification for the study, and defines the research objectives and research questions with research scopes. This research proposes Ensemble Classifier Modelling for Dealing with Missing Values. Finally, the organization of this research will be explained with the thesis organization.

1.1 Background

Classifier technology such as Decision Tree (DT) and ensemble classifiers plays major roles in data mining. Decision Stump (DS), M5P, Random Forest (RF), Random Tree (RNT), Rep Trees (RT) are examples of DT, which have been employed in this study. Note that, Random Forest is a decision tree, and as well as it is an ensemble classifier. Other classifiers considered in this study are Artificial Neural Network (ANN), Support Vector Machine (SVM) and bagging. These are the well-known classifiers in data mining to model data. The ensemble modelling approach has been employed in this study to obtain better accuracy in cervical cancer data where the data is multivariate and imbalanced. The main purpose of this study is to avoid or eliminate dependency on data pre-processing techniques but obtain high accuracy because data pre-processing may lead to wrong medical diagnoses, and medical professionals do not favour data pre-processing in such cases. Classifiers are widely used for exploratory knowledge discovery where comprehensible knowledge representation is preferred. An extensive literature review has been conducted from reputed and indexed articles/journals like SCOPUS, IEEE, Science Direct, etc. starting from 2011.

DT classifiers can also be described like a combination of mathematical and machine learning techniques to aid the description,

categorization and generalization of a given set of data, and it is a common way to organize classification schemes. Using various types of DT classifiers or ensemble DT classifiers, the main goal is to get higher classification accuracy. However, its accuracy depends on the type of DT classifiers being used, and sometimes each learning technique (single classifier) produces a different hypothesis but no perfect hypothesis. Therefore, there is a need to study multiple classifiers (known as ensemble learning).

The main attraction of classifiers lies in an intuitive representation that is easy to understand and comprehend. Accuracy, however, is dependent on the quality of the data and learning algorithms. One of the methods to improve the accuracy of a classifier is the use of Ensemble Learning. DT classifiers inductive inference is considered attractive for many real-life applications, and this Data Mining technology is well suited for many medical settings. However, this real-life data, especially medical data, has lots of missing values. Existing data mining technology offers to deal with these missing values by statistical means, median value imputation or deletion. Though these statistical methods may deal with missing values and may improve accuracy, they are not an option favoured by the medical profession as it often changes the diagnosis result.

To be more specific, one of the key difficulties is medical professionals do not believe data pre-processing because ignoring or filling up the missing values with a statistical approach may change real-life diagnosis outcomes. For instance, the medical doctors test the antibodies to hepatitis B before offering hepatitis B vaccines (US department of veterans affairs, 2018a) because if a person is already exposed to the hepatitis B virus, then the person may get protection from an injection of hepatitis B immunoglobulin (HBIG), which is different from the hepatitis B vaccine (US department of veterans affairs, 2018b). Similarly, the doctor cannot rely on a statistical method to alter cervical cancer data to predict cervical cancer assessments.

Many researchers have mentioned that imputation in data is critical (Nanni, Lumini, & Brahnam, 2012). To my best knowledge, I have identified three different research papers working on an ensemble approach to deal with missing values, yet they have shortcomings as they are imputing missing values in different ways (shown in Table 1):

Table 1.1: Missing value dealing method

Author	Paper Title	Missing value dealing method
(Nanni et al., 2012)	A classifier ensemble approach for the missing feature problem	Multiple imputation method based on random subspace
(Khan, Ahmad, & Mihailidis, 2018)	Bootstrapping and multiple imputation ensemble approaches for missing data	Single imputation method such as Expectation Maximization Imputation, Gussian Random imputation, Bagging single imputation, multiple imputation
(Hassan, Atiya, El-Gayar, & El-Fouly, 2007)	Regression in the presence of missing data using ensemble methods	Generating missing values based on their probability density

Hence, data scientists need to find a way where they will not be dependent on data pre-processing techniques, specially on missing value imputation to achieve high accuracy on extracting influential features that are closely related.

1.1.1 Data Mining

Data mining (DM) is a process of inferring knowledge from data. Classification/clustering analyzes a set of data and generates a set of grouping rules, which can be used to classify future data (Kesavaraj & Sukumaran, 2013). DM is the process of extracting information from a data set and transforming it into an understandable structure which is the machine learning process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and databased systems (Vellido, 2014). DM involves six common classes of tasks: anomaly detection, Association rules learning, Clustering, Regression, Summarization and Classification (Silwattananusarn & Kulthidatuamsuk, 2012). There are many classification techniques in DM from the statistical approach (Siraj, Omer, & Hasan, 2012) to the machine learning approach. The DT classifier is the DM technique which plays a major task in DM and is widely used in various fields. Decision Tree falls within the section of Machine Learning (Siraj & Abdoulha, 2007) (see Fig. 1.1).

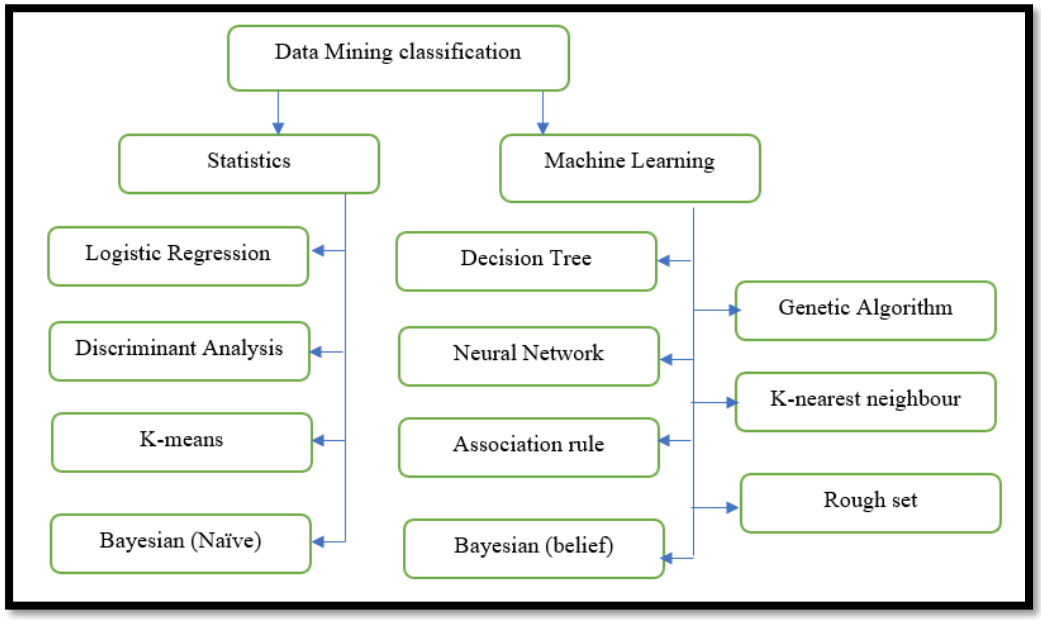

Figure 1.1 Data Mining Classification

Source: adopted from (Siraj & Abdoulha, 2007)

Decision Tree Classification (DTC) is one of the significant DM techniques in Medical Diagnosis and Decision Trees are very useful in diagnosing a patient problem by the physicians (Lavanya, DRani, 2011). For example, DT classifiers are used extensively for diagnosis of breast tumours in ultrasonic images, ovarian cancer and heart sound diagnosis (Ha & Joo, 2010). In medical diagnoses, the role of data mining approaches is increasing rapidly. Particularly, DTCs are very helpful in classifying data, which is important in the decision-making process for medical practitioners. Further, to enhance the DTCs accuracy, various pre-processing techniques and ensemble classifiers have been developed, and are being used widely in the medical domain (Mahila & Pradesh, 2012).

In this research, we claim data mining or classification modelling techniques are the best solution to finding best accuracy in imbalanced data. However, we have noticed that modelling may produce low accuracy when data is multivariate, suffering from missing values. To resolve this issue, we have proposed a novel ensemble model, "Ensemble_RH", which can achieve better accuracy without employing data preprocessing (Hasan, Gholamhosseini, & Sarkar, 2017).

1.1.2 Decision Tree

In the data mining community, decision tree algorithms are very popular since they are relatively fast to train. DT algorithms are very popular due to their characteristics such as fast to train, produce transparent models (Mahila & Pradesh, 2012) and are more properly known as a classification tree (Chandra, 2011). DT is used to learn a classification which facilitates decision making in sequential decision problems, and it is a form of multiple variable (or multiple effects) analyses, including prediction, explanation, description, or classification of an outcome or target (Witten, Frank, & Hall, 2011).

The incorporation of machine learning into medical diagnosis is a new tendency with many medical applications. Many medical diagnostic procedures can be categorized as intelligent data classification tasks such as

in medical data classification, and DT has been widely used both to represent and to conduct decision processes (López-Vallverdú, Riaño, & Bohada, 2012). Hence this research is focusing on DT classifiers for medical data settings.

1.1.3 Decision Tree Classifiers

In medical decision making, there are many situations where a decision must be made effectively and reliably (Podgorelec, Kokol, Stiglic, & Rozman, 2002). Decision analysis is a tool that clinicians can use to choose an option that maximizes the overall net benefit to a patient. It is an explicit, quantitative, and systematic approach to decision making under conditions of uncertainty (A. Lee et al., 2009). Conceptual simple decision-making models with the possibility of automatic learning are the most appropriate for performing such tasks. For instance, DT classifiers are reliable and effective decision-making techniques that provide high classification accuracy with a simple representation of gathered knowledge, and they have been used in different areas of medical decision making (Podgorelec et al., 2002).

D T classifiers are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition, to name only a few. Perhaps the most important feature of DTCs is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret (Chourasia, 2013). Hence, DT with an ensemble model may offer better accuracy for complex cervical cancer data. DT classifiers are comprised of two types; single classifiers and multiple classifiers.

a. Single classifier

A single-classifier obtains prediction accuracy by training, and it makes use of all the available samples (Ko & Sabourin, 2013); nevertheless, it suffers in prediction accuracy in the presence of concept drifts (Wang & Yu, 2002).

b. Multiple classifier

A multiple classifier is a set of classifiers whose individual predictions are combined in some way to classify new examples. This is also known as ensemble classifier learning (Stefanowski, 2008).

1.1.4 Ensemble learning

Ensemble models are considered a more advanced data mining technique where multiple classifiers are combined to produce better predictions and more robust models (Holst & Manga, 2013). Ensemble learning refers to the procedures employed to train multiple learning machines and combine their outputs, treating them as a combination of DT classifiers to decision makers.

Ensemble learning is effective for a variety of classification models (Wang, Yin, Pei, Yu, & Yu, 2006; Wang & Yu, 2002). It is known as ensemble classifier learning. It is a learning algorithm that constructs a set of classifiers and then classifies new data points by taking a (weighted) vote of their predictions (Devroye, 2008). The original ensemble method is Bayesian averaging, but more recent algorithms include error-correcting output coding, bagging, and boosting (Objectives, 2011). In this study, we will use two ensemble approaches: Random Forest and bagging.

- **Bagging**

  Bootstrap aggregating is often abbreviated as bagging. It involves having each model in the ensemble vote with equal weight. To promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set. As an example, the random forest algorithm combines random decision trees with bagging to achieve very high classification accuracy (Darwish, 2013). Bagging can benefit accuracy by missing value imputation. The desired output is a categorical one (Jordanov, Petrov, & Petrozziello, 2018), not a numeric one.

1.2 Problem Statement

Ensemble methods have been widely used to improve the generalization performance of machine learning and data mining systems. In the past decade, there have been numerous studies on generating different kinds of ensemble models, and the benefits of ensemble methods have been confirmed in much literature (Huanhuan Chen, Yao, & Tino, 2011). Ensemble (multiple) classifiers are often more accurate than one single classifier (Liang, 2014). The principle is that the decision obtained from the ensemble method, with individual predictions combined appropriately, should have better overall accuracy, on average, than any individual DT classifier (Brown, 2010).

However, ensemble methods depend on combining classification models (Huanhuan Chen et al., 2011). Diversity among the base classifiers is deemed to be important when constructing a classifier ensemble. Numerous algorithms have been proposed to construct a good classifier ensemble by seeking both the accuracy of the base classifiers and the diversity among them. However, there is no generally accepted definition of accuracy, and measuring the accuracy should not be done with dependency on data preprocessing explicitly (Gangadhara, Anusha, & Dubbaka, 2010); (Tang, Suganthan, & Yao, 2006). It is widely believed that the success of ensemble accuracy without employing data preprocessing is in great need for systematic ensemble study, and understanding and application of base classifiers in ensemble models (Huanhuan Chen et al., 2011).

Classification algorithms are considered as one of the most promising in medical data classification (Hasan, Siraj, & Sainin, 2015b), (Krawczyk & Schaefer, 2012) for selecting suitable features. Among them, decision tree and ensemble learning is a better learning method to extract the features based on improved accuracy where multimodal medical data with a high relative dimensionality is present (Hasan, Golamhosseini, Sarkar, & Safiuzzaman, 2017). Decision tree and ensemble methods such as bagging have favourable properties to select suitable features from the data sets with high dimensionality (Hasan, Gholamhosseini, et al., 2017) or missing values (Nanni et al., 2012).

In medical data mining a suitable feature based on accuracy, selection could offer more understandings of medical data but most of the researchers have overlooked the high error in the feature. In such a case, the features in medical data may be selected based on accuracy, but the features are not really related to the diagnosis outcome. Cervical cancer data from UCI poses such a trend with high dimensionality and typically suffers from one or more of the above conditions due to the difficulty and cost of acquiring clinical data (Hasan, Gholamhosseini, et al., 2017), (C. H. Lee & Yoon, 2017). Keeping in mind that data pre-processing may change the medical diagnosis results, this study employed several classification approaches to cervical cancer data without involving data preprocessing techniques. Based on our empirical study, decision tree and ensemble methods are suitable to be applied to select features from cervical cancer data sets which have been supported by (Hasan, Siraj, et al., 2015b). In this study, decision tree, ANN, SVM, and bagging algorithms are employed due to the nature of cervical cancer data, which is has many missing values.

1.3 Research Question

This is not a thesis dealing with the advantages or disadvantages of missing value imputation or imputation techniques. This research is dedicatedly focused on the research question:

(i) How is quantifiably acceptable accuracy obtained with a systematic ensemble approach without missing value imputation?

1.4 Research Scope

The scope of this study is within the above-mentioned objectives that are to propose the Ensemble Classifier Model for cervical cancer data based on the better accuracy obtained from bagging (decision stump), bagging (REPTree), bagging (Random Tree) and from other classifiers such as M5P, Random Forest, RepTree, Artificial Neural Network, Support Vector Machine. The data set is cervical cancer data obtained from the UC

Irvine (known as UCI) machine learning repository that is openly accessible at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms(Hall et al., 2009b); (Zhao & Zhang, 2007) which will be used to achieve the objectives of this research as it can handle both types of classifier and multiple classifiers (ensemble).

1.5 Organization of the research

In this study, Chapter One provides an overview of the research including the problem statement, the objective, the scope, method and the contribution of the research. Chapter Two reviews ensemble learning and decision tree classifiers. Chapter Three discusses the research methodology, and Chapter four discusses the Ensemble Classifier Modelling for Dealing with Missing Values.

1.6 Summary of the chapter

In this chapter, the background of the research has been described under the background section, and the problem statement describes the focus of the research and the research needs. The research objective has been generated based on the problem defined and the research question has also been generated. In addition, this chapter described the research scope and the research contribution which highlights the significance of this research.

Chapter 2

Literature Review

The literature review and conceptual framework can maximize the chances of spanning the abyss and reaching something substantive. In this chapter previous related work on the base classifier and ensemble classifier for medical data are discussed, which reflect the proposed research topic: Ensemble Classifier Modelling for Dealing with Missing Values.

2.1 Data mining

Data Mining (DM) is the process to extract information from a data set (Kesavaraj & Sukumaran, 2013) and transform it into an understandable structure which is the machine learning process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems (Vellido, 2014). The Decision Tree (DT) classifier is the DM technique which plays a vital role in DM and is widely used in various fields (Silwattananusarn & Kulthidatuamsuk, 2012). There are many classification techniques in DM from the statistical approach to the machine learning approach (Siraj et al., 2012). Decision tree falls within the section of machine learning (Siraj & Abdoulha, 2007) (see Fig. 2.1).

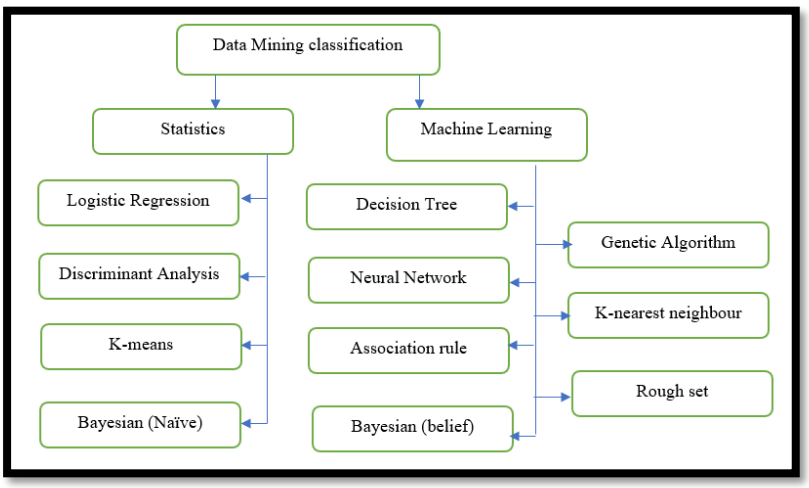

Figure 2.1 Data Mining Classification (*Source: adopted from* (Siraj & Abdoulha, 2007))

Decision Tree Classification is one of the significant DM techniques in medical diagnosis and decision trees are very much useful to diagnose a patient problem by physicians (Lavanya & Rani, 2012). Decision tree classifiers are used extensively for diagnosis of breast tumours in ultrasonic images, ovarian cancer, heart sound diagnosis and so on (Ha & Joo, 2010).

In medical diagnoses, the role of data mining approaches is increasing rapidly. Particularly Decision Tree Classifiers (DTCs) are very helpful in classifying data, which is important in the decision-making process for medical practitioners (Lavanya & Rani, 2012). Further, to enhance the DTCs accuracy various pre-processing techniques and ensemble classifiers have been developed, which are being used widely in the medical domain (Mahila & Pradesh, 2012).

2.2 Ensemble Learning

Ensemble classifiers refer to the procedures employed to train multiple learning machines and combine their outputs, treating them as a combination of DTC and decision makers (Brown, 2010). It is effective for a variety of classification methods (H. Wang, Yin, Pei, Yu, & Yu, 2006). It is a learning algorithm that constructs a set of classifiers and then classifies new data points by taking a (weighted) vote of their predictions (Devroye, 2008).

Ensemble learning like bagging and boosting, which combine the decisions of multiple hypotheses are some of the strongest existing machine learning methods (Melville, 2003). An ensemble is itself a supervised learning algorithm because it can be trained and then used to make predictions (Fensterstock, Salters, & Willging, 2013). The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the methods from which it is built (Rudin, 2007). Thus, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single method would, but in practice, some ensemble techniques (especially bagging) tend to

reduce problems related to over-fitting of the training data (Li, Wang, Ma, & Song, 2013).

An ensemble classifier provides a good way to have a near-optimal classifying system for any problem (Alizadeh & Parvin, 2011). One of the most challenging problems in the classifier ensemble is introducing a suitable ensemble of base classifiers (Parvin, MirnabiBaboli, & Alinejad-Rokny, 2015). Every ensemble needs to identify the base classifier for a specific problem. It means that if a group of classifiers is to be a successful ensemble, the base classifier needs to be chosen accordingly, which may produce less error (Parvin et al., 2015). However, this research did not identify the base classifier that may be best for specific classification problem in classification approach. Therefore, during ensemble creation, a method is needed to ensure that the ensemble classifiers produce best accuracy. Adhvaryu & Panchal (2012) have identified several methods to estimate classifier accuracy such as holdout method, cross validation, boosting, bagging, and random forest from the previous study (Adhvaryu & Panchal, 2012).

Ensemble methods have been widely used in the literature (Huanhuan Chen et al., 2011) to improve the generalization performance of machine learning and data mining systems. In the past decade, there have been numerous studies on generating different kinds of ensemble models, and the benefits of ensemble methods have been confirmed in the literature (Huanhuan Chen et al., 2011). The ensemble is often more accurate than any of the base classifiers (Liang, 2014). However, many cases can occur where an ensemble may not produce a better result and a question may arise in this situation: how can this issue be resolved? The principle is that the decision obtained from the ensemble method, with individual predictions combined appropriately, should have better overall accuracy, on average, than any individual DTC (Brown, 2010). Nevertheless, as mentioned earlier, that ensemble may not always follow its basic principle.

Previous research has proven that it may be better to ensemble many instead of all of the classifiers at hand. Thus, classifier selection became a

crucial problem for ensemble learning. To select a better classifier set from a pool of classifiers, the classifier ensemble is the most important property to be considered.

(Yin, Huang, Hao, Iqbal, & Wang, 2014) managed to identify the classifier ensemble problem with accuracy and uses diverse ensemble learning. In the experiments, 10-fold cross validation of the data sets is performed and four ensemble methods are compared: Bagging (Bag), LS Estimation Combination (LSE), Sparsity Learning (SPA), and Sparsity and Diversity Learning (S&D). Each ensemble contains 100 neural network classifier components (with back-propagation in Matlab), which are similar to the components in bagging. That is to say, mainly focus remained on neural network ensembles in the experiments. (Yin et al., 2014).

The main two research gaps from the above literature review are:

- (Parvin et al., 2015) uses a clustering approach for ensemble creation. The approach was classifier selected based on clustering (CSBC) on training results. To partition the cluster a modified method of bagging and K-means has been used. However, as the training result always produces high accuracy, there is a need to investigate how the ensemble method behaves when decision tree classifiers (classification approach) are involved both in bagging and boosting in terms of testing results. Furthermore, another gap may arise how to handle the variance in result during ensemble creation.

- (Stefanowski & Pachocki, 2013) compared four algorithm bagging, boosting, DECORATE and random forests by Query by Committee Based Active Learning on J48. This study is inspired by earlier promising results from (Melville & Mooney, 2004), and the empirical result shows that the ensemble method improves accuracy. However, there may be an additional lead of research on how the ensemble methods work for other decision tree classifiers such as Random forest, Random tree, J48 grafts, and LMT. Furthermore,

further research may be involved with the improvement of accuracy in the ensemble method by selecting a better base classifier.

2.2.1 Bootstrap aggregating

Bootstrap aggregating was invented by Breiman in 1999 (Breiman, 1999). It is abbreviated as bagging (Kusum & Rupali, 2013) and known as one of the earliest ensemble algorithms (Zhang & Ma, 2012). It involves having each method in the ensemble vote with equal weight. In order to promote method variance, bagging trains each method in the ensemble using a randomly drawn subset of the training set (Che, Liu, Rasheed, & Tao, 2011). As an example, the random forest algorithm combines random decision trees with bagging to achieve very high classification accuracy (Darwish, 2013).

Kulkarni (2014) conducted research on ensemble techniques of bagging, boosting and Ada-Boost. The experiment observed that the performance of ensemble classifiers is better than individual classifiers and bagging often performs better (Kulkarni & Kelkar, 2014).

Faraz (2012) conducted a study on retinal vessel segmentation using an ensemble classifier of bagged decision tree based on supervised classification using an ensemble classifier of bagged decision trees. The performance, effectiveness, and robustness along with its simplicity and speed in training as well as classification, make this ensemble based method a suitable tool to be integrated into a complete retinal image analysis system for clinical purposes and in particular for large population studies (Fraz et al., 2012).

Ye (2013) conducted an empirical comparison of bagging-based ensemble classifiers. The comparison was done empirically on four bagging based ensemble classifiers: the ensemble adaptive neurofuzzy inference system (ANFIS), the ensemble support vector machine (SVM), the ensemble extreme learning machine (ELM) and the random forest. The empirical results also showed that bagging is the most favourable ensemble classifier among them (Ye & Suganthan, 2013).

2.2.2 Bagging Algorithm

Given a set S of s samples, bagging works as follows. For iteration t (t = 1, 2…T), a training set St is sampled with a replacement from the original set of samples, S. Since sampling with replacements is used, some of the original samples of S may not be included in St, while others may occur more than once. Each bootstrap sample Si contains approx. 63.2% of the original training data. Remaining (36.8%) are used as a test set. A classifier Ct is learned for each training set, St. To classify an unknown sample, X, each classifier Ct returns its class prediction, which counts as one vote. The bagged classifier, C*, counts the votes and assigns the class with the most votes to X. Bagging can be applied to the prediction of continuous values by taking the average value of each vote, rather than the majority.

In case of classification into two possible classes, a classification algorithm creates a classifier H: D → {-1, 1} on the base of a training set of example descriptions D. The bagging method creates a sequence of classifiers H_m, m = 1... M in respect to modifications of the training set. These classifiers are combined into a compound classifier. The prediction of the compound classifier is given as a weighted combination of individual classifier predictions:

$$H(d_i) = \sin\left(\sum_{m=1}^{M} \propto_m \quad H_m(d_i)\right) \quad \text{.....(4)}$$

The meaning of the above formula can be interpreted as a voting procedure. An example di is classified to the class for which the majority of classifiers vote.

2.2.3 Random Forest

The early development of random forests was influenced by the work of (Amit & Geman, 1997), and it has been introduced by (Breiman, 2001). It is a variant of bagging algorithms whose base classifiers are decision tree (Arabnia & Tran, 2011). Like bagging, random forests use

bootstrap sampling and un-weighted aggregation of committees for the final classification. Random forests tend to perform very well, especially for those data sets containing many attributes (Che et al., 2011).

(Elshazly, Elkorany, Hassanien, & Azar, 2013) conducted a study on the performance of two novel ensemble classifiers, Random Forest (RF) and Rotation Forest (ROT), for biomedical data sets tested with five medical data sets.

Prediction performance is evaluated using an accuracy measure. It was observed that ROT achieved the highest classification accuracy in most tested cases (Elshazly et al., 2013).

(Saghir & Megherbi, 2013) studied experimental results of the codon-based attribute reduction and binning prediction algorithms, using a random forest classifier and a Bayes classifier, respectively, which are presented along with their comparison to their DNA-based k-means counterparts. The findings showed that the classification/prediction accuracy achieved is between 59% and 92% for various data sets using a random forest classifier and between 44% and 64% using a Naïve Bayes classifier. The random forest classifier did better in classification in all the data sets compared to Naïve Bayes.

(Guidi, Pettenati, Miniati, & Iadanza, 2013) described an automatic classifier of patients with heart failure designed for a tele monitoring scenario. The result showed that analyzing the data with its direct evolution, that is the random forest algorithm, showed improvements both in accuracy and in limiting critical errors.

(Tripoliti, Fotiadis, & Manis, 2012) conducted research on the automated diagnosis of diseases based on classification: dynamic determination of the number of trees in the random forest algorithm. He proposed a new method for the automated diagnosis of diseases based on the improvement of the random forest classification algorithm. The proposed method produces an ensemble not only accurate but also diverse, ensuring the two important properties that should characterize an ensemble

classifier. The method is based on an online fitting procedure, and it is evaluated using eight biomedical data sets and five versions of the random forest algorithm (40 cases). The method decided correctly the number of trees in 90% of the test cases.

Random forests differ from bagging in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging." The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the *B* trees, causing them to become correlated. Typically, for a data set with *p* features, $\sqrt{p}$ features are used in each split.

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way. The following technique was described in Breiman's original paper and is implemented in the R package random forest.

The first step in measuring the variable importance to a data set is:

$$D_n = \{X_i\ Y_i\}_{i=1}^{n} \qquad \ldots\ldots\ldots\ldots \qquad (5)$$

This step is to fit a random forest to the data. During the fitting process, the out-of-bag error for each data point is recorded and averaged over the forest (errors on an independent test set can be substituted if bagging is not used during training).

To measure the importance of the j-th feature after training, the values of the j-th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the j-th feature is computed by averaging the difference in the out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences. Features which

produce large values for this score are ranked as more important than features which produce small values.

2.3 Ensemble approach to missing value imputation

Though missing value imputation is not my research focus, it is still better to identify whether any research on my focus has been conducted by other researchers. To my best knowledge, I have identified four ensemble approaches that dealt with missing values in the ensemble method. (Nanni et al., 2012) agreed that missing value imputation is critical. It does not matter which approach we are using: normalization, mean imputation, fusion. (Hassan et al., 2007) dealt with regression in the presence of missing data using ensemble methods. He imputes missing values based on probability density. (Khan et al., 2018) proposed bootstrapping and multiple imputation ensemble approaches for missing values. All these three types of research could not achieve the acceptable accuracy though missing value imputation has been carried out as a data preprocessing technique.

(Conroy, Eshelman, Potes, & Xu-Wilson, 2016) conducted research on the dynamic ensemble approach to robust classification in the presence of missing data and mentioned that missing value imputation is commonly based on their mean, median. He used two complicated stages of ensemble approach and measured features into a predictive model for ICU patient's data. Though this research looks like it did not impute missing values, it imputes a secondary layer of resilience to missing data and assigned weight, which are more dangerous in medical diagnosis accuracy. This research did not consider other popular machine learning methods such as SVM, ANN and so one is limited in investigating the impact of classification in the presence of the missing data decision stump and NB classifiers.

2.4 Data modelling study in medical science

Modelling multivariate cervical cancer data suffers from different classification problems due to missing values, outliers and attribute characteristics (Holst & Manga, 2013). Using classification techniques such as decision tree learning is influenced by these factors. Identifying the most

suitable decision tree learning algorithm, especially in medical sequential decision making to obtain an accurate model, remains challenging (Ha & Joo, 2010). (Małgorzata, 2012) focus on how data mining techniques are applied to predict breast cancer in a Wisconsin data set. He explores the applicability of decision trees (Random tree, ID3, CART, C4.5, and Naive Bayes) to predict the presence of breast cancer. Among the classifiers, random tree outperforms of all the other algorithms with the highest accuracy rate. To handle missing values a Linear Interpolation technique has been employed and data cleaning has been initiated in the data pre-processing phase.

Sun et.al. (2014) presents a novel machine learning method for the construction of cancer progression models based on the analysis of static tumour samples. He demonstrated the reliability of the method with simulated data and describes the application to breast cancer data. The findings support a linear, branching model for breast cancer progression. The author did not include the feature of high dimensionality. The proposed method can reconstruct tumour progression but is not able to identify the cancer risk factor (Sun, Yao, Nowak, & Goodison, 2014).

Ludwig (2018) investigates a fuzzy decision tree algorithm applied to the classification of gene expression data. The fuzzy decision tree algorithm is compared to a classical decision tree algorithm as well as other well-known data mining algorithms commonly applied to classification tasks. Based on the five data sets analyzed, the fuzzy decision tree algorithm outperforms the classical decision tree algorithm. However, compared to other commonly used classification algorithms, both decision tree algorithms are competitive, but they do not reach the accuracy values of the best-performing classifier (Ludwig, Picek, & Jakobovic, 2018).

Chaurasia (2017) presents a diagnosis system for detecting breast cancer based on RepTree, RBF Network and Simple Logistic. This research demonstrated that the Simple Logistic can be used for reducing the dimension of feature space and proposed Rep Tree and RBF Network model

can be used to obtain fast automatic diagnostic systems for other diseases. However, the correct classification rate of the proposed system is only 74.5% while in medial data, we expect to obtain a classification accuracy of nearly 100% (Chaurasia & Pal, 2017).

2.5 Problem with modelling medical data

Life critical data such as medical data classification is acknowledged as an area of increasing importance, yet also poses many difficulties (Hasan, Bakar, Siraj, Sainin, & Hasan, 2015), (Krawczyk & Schaefer, 2012). Multiple classifier systems are considered as one of the most promising in medical data classification (Hasan, Siraj, et al., 2015b), (Krawczyk & Schaefer, 2012). Ensemble learning is a better learning method for improving accuracy where multimodal medical data with a high relative dimensionality is present (Hasan, Golamhosseini, et al., 2017), (Wu, Shen, & Sabuncu, 2016), (Tay, Chui, Ong, & Ng, 2013). The ensemble method has favourable properties that make them suitable for data sets with high dimensionality (Hasan, Gholamhosseini, et al., 2017), (Dittman, Khoshgoftaar, & Napolitano, 2015), (Blagus & Lusa, 2015), (Ojha, Jackowski, Abraham, & Snášel, 2015), (Moon et al., 2007) or missing values (Nanni et al., 2012). Data from medical studies typically suffers from one or more of the above conditions, due to the difficulty and cost of acquiring clinical data (Hasan, Gholamhosseini, et al., 2017), (C. H. Lee & Yoon, 2017), (Kang, 2013). Ensemble methods are therefore suitable to be applied to medical data sets. Table 1 shows the summaries of the research findings and gaps.

Table 1: Summary of literature review in the proposed field of study contains such a critical review (cervical cancer detection framework).

Author	**Key findings**	**Research gap**
(Tan & Gilbert, 2003)	The author employed 14 decision tree classifiers for three different medical data sets: Wisconsin's breast cancer data, Pima Indian diabetes data, and	Though this research revealed the best performers among the decision tree classifiers, it was limited to the decision tree classifiers

	hepatitis data. The results revealed that classifiers such as FT, LMT, NB tree, Random Forest and Random Tree are the five best single classifiers as they constantly provide better accuracy in their classifications.	only. It could be better if the performance of other classifiers is considered and this research did not focus on ensemble methods.
(Hasan, Siraj, et al., 2015b)	This research employed two prominent ensembles, Adaboost and Bagging, with base classifiers such as Random Forest, Random Tree, j48, j48grafts and Logistic Model Regression (LMT) that have been selected independently. The empirical study shows that the performance varies when different base classifiers are selected and even in some places overfitting issues have also been noted. The evidence shows that ensemble decision tree classifiers using AdaBoost and Bagging improve the performance of selected medical data sets	The author employed a popular ensemble model and noticed over-fitting. How to deal with this overfitting and no mention about data pre-processing. The biased issues in the results are overlooked.
(Wu et al., 2016)	The author employed machine learning probabilistic modelling for an ultrasound, magnetic resonance imaging (MRI), computed tomography (CT), histology, and microscopy images.	The author involved the probability of modelling techniques which requires images but, in our research, we are trying to avoid cervical pap smear images.
(Tay et al., 2013)	The author identified that there may be an over-fitting issue in multimodal	The researcher often suggests that Adaboost

	areal bone mineral density data and tried to improve regression accuracy. He employed meta leaner filtering instead of bagging techniques to build feature-wise ensembles.	may perform better where boosting does not improve performance. However, the bias issue is overlooked which is the main issue when dealing with the ensemble model.
(Dittman et al., 2015)	The author focused on bagging and boosting on balanced bioinformatics data and found that bagging performs well.	To our best knowledge a single classifier (such as a random tree, complex tree, and j48 tree) often obtained better accuracy without employing ensemble. The researcher did not compare whether any single classifier works well with balanced data. Most of the medical data is unbalanced. Hence, the performance analysis is needed if the data is unbalanced
(Blagus & Lusa, 2015)	The author used to boost when the number of variables is more than the number of the samples which is a high dimensional two class problem. He identified the over-fitting issue when base classifiers are not chosen	Though this research suggested some base classifiers for boosting, it did not explain the biasness and did not focus on bagging while other research shows

	accordingly and explains why the over-fitting occurs.	that bagging may perform well when boosting does not perform well.

2.6 Decision Tree

A decision tree classification model is represented by a tree-like structure, where each internal node represents a test of a feature, with each branch representing one of the possible test results and each leaf node representing a classification. Depending on which construction algorithms are applied, decision tree models may vary (Che et al., 2011).

2.7 Decision Tree classifiers

Decision Tree Classifiers are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition (Aymerich, Alonso, Cabañas, & Comabella, 2011). Perhaps, the most important attribute of DTCs is their capability to break down a complex decision-making process into a collection of simpler decisions(López-Vallverdú et al., 2012), thus providing a solution which is often easier to interpret (Chourasia, 2013). DTC comprises of two types; Single classifiers and multiple classifiers.

A decision tree classification method is represented by a tree-like structure, where each internal node represents a test of a tree, with each branch representing one of the possible test results (Arabnia & Tran, 2011), and each leaf node representing a classification(Mohamed, Salleh, & Omar, 2012). Decision tree models may vary depending on which construction algorithms are applied (Che et al., 2011).

Decision Tree (DT) algorithms are very popular due to their characteristics such as being fast to train (Mahila & Pradesh, 2012) and more properly known as a classification tree (Chandra, 2011). DT is used to learn

a classification which facilitates decision making in sequential decision problems, and it is a form of multiple variable (or multiple effects) analyses, including prediction, explanation, description, or classification of an outcome or target (Witten et al., 2011). DT has been widely used both to represent and to conduct decision processes (López-Vallverdú et al., 2012), and hence the research is focusing on Decision Tree classifiers (DTC) for medical data settings.

Multiple classifiers is a set of classifiers whose individual predictions are combined in some way to classify new examples. It is also known as ensemble classifier learning (Stefanowski, 2008).

In medicine, decision processes may be of several kinds and for different purposes (López-Vallverdú et al., 2012): screening, diagnosing, prognosing, drug and therapy prescription, and others. Through the years, multiple computer-based structures have been proposed to formalize these decision processes. They range from such statistical approaches as Bayesian Networks (Arsene, Dumitrache, & Mihu, 2011); (Lucas, van der Gaag, & Abu-Hanna, 2004); (Velikova, de Carvalho Ferreira, & Lucas, 2007); (López-Vallverdú et al., 2012), or probabilistic models (Husmeier, Dybowski, & Roberts, 2004) to symbolic approaches as decision trees, decision tables or decision rules (Yeh, Cheng, & Chen, 2011). Among them, decision trees have been particularly successful (Arsene et al., 2011) and widely used both to represent and to conduct decision processes. Medical decision trees can be provided by experts or automatically induced from medical databases (Fauci et al., 2009).

Among the computerized methods that can be applied to the analysis of metabolism in formation in the spectral data, decision trees can be considered especially appropriate because they can be used to deal with problems that are rich in data but complex to interpret (Goodacre, Vaidyanathan, Dunn, Harrigan, & Kell, 2004). Decision trees are widely used in pattern recognition, machine learning and data mining applications (Aymerich et al., 2011). Many methods have been developed for constructing decision trees from collections of examples, some of the more

commonly used algorithms (Olafsson, Li, & Wu, 2008); (Rokach & Maimon, 2008); (Rokach & Maimon, 2005) being ID3, C4.5, and CART (Aymerich et al., 2011). Although decision tree techniques are interpretable, efficient, problem-independent and able to deal with large-scale applications, they have proven to be subject to high variance, which leads to the classification accuracy and signposts the need for further research (Aymerich et al., 2011).

Often the medical decision maker faces problems with a sequential decision problem involving decisions that lead to different outcomes depending on chance. If the decision process involves many sequential decisions, then the decision problem becomes difficult to visualize and to implement (Ishwaran & Rao, 2011). Decision trees are indispensable graphical tools in such settings as they allow for an intuitive understanding of the problem and can aid in decision making.

However, the medical decision maker may not know what the decision rule is and would like to discover the decision rule by using data. In such settings, decision trees are often referred to as classification trees (Ishwaran & Rao, 2011). Classification trees apply to data where the outcome is a classification label, such as the disease status of a patient, and the medical decision maker would like to construct a decision rule that predicts the outcome using dependent variables available in the data as the data set available is just one sample of the underlying population. In this case it is desirable to construct a decision rule that is accurate not only for the data at hand but over external data as well i.e., the decision rule should have good prediction performance (Ishwaran & Rao, 2011). At the same time, it is helpful to have a decision rule that is understandable. That is, it should not be so complex that the decision maker is left with a black box. Decision trees offer a reasonable way to resolve these two conflicting needs (Ishwaran & Rao, 2011).

DTCs are the classification methods which are very useful to diagnose a patient problem by physicians. DTCs are used extensively for diagnosis of breast tumours in ultrasonic images, ovarian cancer and heart

sound diagnosis (Lavanya, Pradesh, & Rani, 2011). Decision tree approaches and the ensembles of decision tree-based classifiers have been widely applied in cancer classification, including breast cancer, central nervous system embryonic tumours, colon tumours, leukemia, lung cancer, ovarian cancer, pancreatic cancer, and prostate cancer (Che et al., 2011).

(Nai-Arun & Sittidech, 2014) conducted research on diabetic data and popular ensemble learning; bagging and boosting were applied using the three base classifiers in the study. The research found that the better method with the highest accuracy was bagging with a base classifier decision tree algorithm (95.312%). The experiments also showed that ensemble classifier methods performed better than the base classifiers alone.

(Farid, Maruf, & Rahman, 2013) introduced a new approach of boosting using decision trees for classifying noisy data. The proposed approach considers a series of decision tree classifiers and combines the votes of each classifier for classifying known or unknown instances. The weights of training instances were updated based on the misclassification error rates that are produced by the training instances in each round of classifier construction. They tested the performance of proposed boosting algorithms with existing decision tree algorithms by employing benchmark data sets from the UCI machine learning repository. Experimental analysis proved that the proposed boosting approach achieved high classification accuracy for different types of data sets.

(Kelarev, Stranieri, Yearwood, & Jelinek, 2012) was concerned with the detection and monitoring of Cardiovascular Autonomic Neuropathy (CAN), in diabetes patients. Using a small set of attributes identified previously, the author carried out an empirical investigation and comparison of a few ensemble methods based on decision trees for a novel application of the processing of sensor data from diabetes patients for pervasive health monitoring of CAN. The experiments relied on an extensive database collected by the Diabetes Complications Screening and included a couple of essential ensemble methods. The results showed that the novel application

of the decision trees in ensemble classifiers for the detection and monitoring of CAN in diabetes patients achieved better results.

(M. Wang, Gao, Wang, & Miu, 2012) found that there is a difficulty for a base classifier to resolve the problem of high dimension in the hyper spectral image classification applications. A combination of multiple classifiers can make full use of the complementary of the existing classifiers, thus owning better classification performance. A novel multiple classifier based on the C 5.0 decision tree has been proposed, which reduces the hyper spectral dimension through a wavelet-PCA transformed algorithm, and the proposed method can reduce the dimension of attributes and improve the classification performance efficiently.

(Floares & Birlutiu, 2012) conducted research on classification methods that are able to discriminate between normal and cancer samples based on the molecular bio markers discovered, which focused on transparent and interpretable methods for data analysis. They built molecular classifiers using decision tree methods in combination with boosting and cross-validation to distinguish between normal and malign samples. The approach is designed to avoid over fitting and overoptimistic results. We performed an experimental evaluation of a data set related to the urothelial carcinoma of the bladder. We identified a set of tumour microRNAs bio markers, which, integrated into an ensemble of decision tree classifiers, can discriminate between normal and cancer samples with the better published accuracy.

(Xiaochen & Xue, 2011) conducted research on an ID3 algorithm, and he overcame the existing bias of the ID3 algorithm. And then, ADABOOST Algorithm and improved ID3 Algorithm were constituted as a multi-decision-tree classifier, and it was applied in the Master Data Management System to form the redundant data judgment module which responsibility is judging the redundant data. The result shows that the accuracy of this classifier is better than the pure Decision-Tree classifier, and the training duration of this classifier is shorter than the original Decision-Tree-ID3 based ADABOOST classifier. It greatly reduces manual

labour after applying it in the Master Data Management System and saves the consumption of human and material resources.

(Oh, Lee, & Zhang, 2011) conducted research on biomedical data and found that the imbalanced data problem occurs frequently and causes poor prediction performance for minority classes. This is because the trained classifiers are mostly derived from the majority class. (Oh et al., 2011) also described an ensemble learning method combined with active sample selection to resolve the unbalanced data problem and evaluated three methods (an active example selection algorithm, an ensemble learning method, and an incremental learning method on six real-world unbalanced data sets in biomedical domains), showing that the proposed method outperforms both the random under sampling and the ensemble with under sampling methods.

(Lavanya & Rani, 2012) studied decision tree classifiers and the experiments were conducted to find the best classifier for Medical Diagnosis. The experimental results show that CART is the best algorithm for classification of medical data. It is also observed that CART performs well for classification on medical data sets of increased size (Lavanya, DRani, 2011).

2.8 WEKA for Ensemble Learning

Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms for data mining tasks (Zhao & Zhang, 2007) which were developed by the Machine Learning Group at the University of Waikato (Hall et al., 2009a) in New Zealand in1993 (Markov & Russell, 2006).

In WEKA the algorithms can either be applied directly to a data set or called from Java code. Weka contains tools for data pre-processing, regression, clustering, association rules, and classification, including decision tree algorithms and ensemble learning algorithms, which can be used for biologists to classify their biological data (Che et al., 2011). It is

also well-suited for developing new machine learning schemes (Machine Learning Group at the University of Waikato, 2013).

2.9 Summary of the chapter

This chapter discussed some machine learning literature, particularly on data mining classification including ensemble learning using bagging, boosting, and random forest. A data modelling study on medical science literature focused on the use of decision trees. In all the literature, we have noticed missing value imputation. Our research proposes decision tree and ensemble techniques and, in particular, we propose decision tree and ensemble techniques without missing value imputation.

Chapter 3

Design of the study

This chapter describes the research methods for the ensemble classifier model for medical data based on Decision Tree. The main objectives are to explain the investigation steps of the ensemble classifiers which reflect the best classification accuracy of the decision tree.

3.1 Design of the Study

The study design combines several phases including phase 1: data analysis and knowledge acquisition, phase 2: Methods to build the ensemble model, phase 3: Primary data analysis and knowledge acquisition, and phase 4: Output/contribution as shown in Fig 3.1. Throughout the study, MATLAB and WEKA have been used as a data analysis tool.

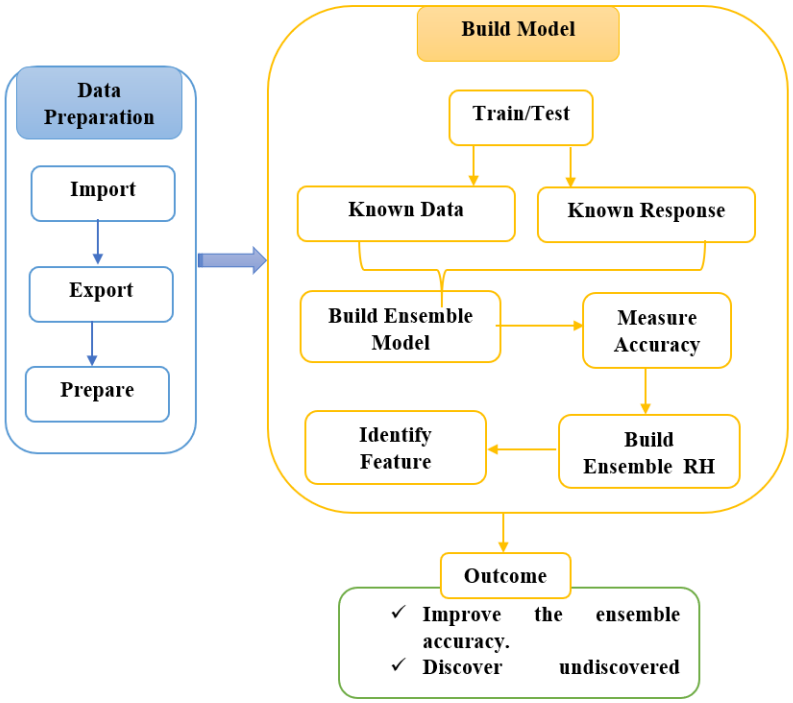

Figure 3.1 Research Design

3.1.1 Phase 1: Data analysis and knowledge acquisition

- **Data description**

Data has been obtained from UC Irvine (known as UCI) machine learning repository that is openly accessible fromhttps://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. There are 858 instances and 35 attributes in this data. There are missing values in 23 features (total 35 features) in a different instance. STDs: Time since the last diagnosis and first diagnosis features has 92% missing values, which is 787 out of 858 instances. Therefore, the number of missing values is very high in this cervical cancer data.

Table 3.1 Missing values in the feature

Feature name	Number of missing values	Missing value (%)
Number of sexual partners	26	3
First sexual intercourse	71	1
Number of pregnancies	56	7
Smokes	13	2
Smokes (Year)	13	2
Hormonal contraceptives	108	13
Hormonal contraceptives (years)	108	13
IUD	117	14
IUD (Years)	105	12
STDs (number)	105	12
STDs: Condylomatosis	105	12
STDs: Cervical Condylomatosis	105	14
STDs: Vaginal Condylomatosis	105	12
STDs: Vulvo perinial Condylomatosis	105	12
STDS: Syphilis	105	12
STDs: pelvic inflammatory diseases	105	12
STDs: Genital herps	105	12
STDs: Mulluscum congiosum	105	12
STDs: AIDS	105	12
STDs: HIV	105	12
STDs: Hepatitis B	105	12
STDs: Time since first diagnosis	787	92
STDs: Time since first diagnosis	787	92

In this research, we have fed all the 858 instances and 35 attributes/ features into our preliminary analysis. Moreover, we have tried to select suitable features by feature selection techniques (Ensemble modelling). The interesting and challenging part of this research is, we do not involve any data pre-processing techniques because it may change the result of the

medical diagnosis. Finally, the most challenging part of this research is "How to get better accuracy when data is suffering from outliers, missing values and so on." From the data, the relationship between the attributes with cervical cancer will be identified.

- **Justification for using cervical cancer data**

3.1.2 Classification techniques

Since this study focuses on feature selection and modelling medical data, various classification techniques from a decision tree such as a complex tree, simple tree, ensemble method (bagged tree and boosted tree) are selected and applied (see Fig 3.2). From the literature survey, we found that decision tree and ensemble of the decision tree are better performers in medical data sets.

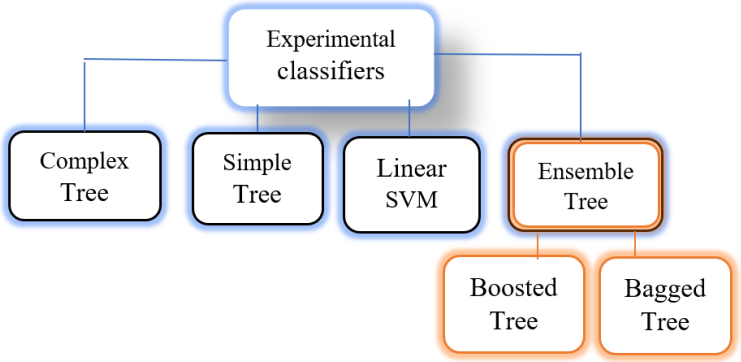

Figure 3.2 Methods to build Ensemble_RH

(i) Simple tree

The simple Tree is a decision tree with few leaves that can make a few fine distinctions between classes with a maximum of four splits.

(ii) Complex Tree

The complex tree is a decision tree classifier with many leaves that can make many fine distinctions between classes with a maximum split of 100.

(iii) Boosted Tree

The boosted tree creates an ensemble model with the medium tree (maximum 20 splits) by the AdaBoost algorithm. Compared to bagging, boosting uses little time and memory but might need more numbers of the ensemble.

(iv) Bagging

Bagging is a boot strap aggregated ensemble of fine decision trees (max number of the split is 100), often very accurate but slow and memory intensive.

3.1.3 Phase 2: Methods to build the ensemble model

- **Decision tree, bagging and boosting**

Phase 2 deals with testing the existing ensemble models such as bagging, boosting, and random forest from the data. Initially, we employed several "decision tree algorithms" such as a simple decision tree, a complex tree, and ensemble decision tree (Fig 3.2). Based on article review from previous study by(Hasan, Bakar, et al., 2015); (Hasan, Siraj, & Sainin, 2015a), we have chosen Adaboost and bagging as an ensemble method.

Fig 3.2 depicts the classification techniques employed in this research. In this research, we have chosen the decision tree classifiers simple tree and complex tree. For the ensemble method, we have chosen a complex tree and bagging. To improve the performance of the week ensemble classifier (i.e bagging) we have made a brand-new ensemble model with bagging and the complex tree which we called new Ensemble_RH.

- **Ensemble model theory**

Figure 3.3 shows how the combination of two or more classifiers can form an ensemble method. The classifiers can be the same (in this study the decision tree classifier) or can be different (such as, as we can combine the decision tree classifier with a support vector machine or SVM). When the weak classifiers are combined, it improves the performance by decreasing the variance (in the bagging method) or by stacking (in boosting method). However, we may combine two or best classifiers; alternatively, one best classifier with a weak classifier; alternatively, two weak classifiers to improve classification accuracy. Fig. 3.3 explains that two build an ensemble classifier, and we may combine two or any number of classifiers.

In this research, we are employing ensemble methods. Bagging and boosting is the ensemble method which combines multiple classifiers and it is the most robust machine learning method in medical settings (Hasan, Siraj, et al., 2015b). Several studies dealt with a single classifier and only one class problem when all information was available, but in our case, the data is multivariate, and suffers from missing values, outliers, and multi-classes.

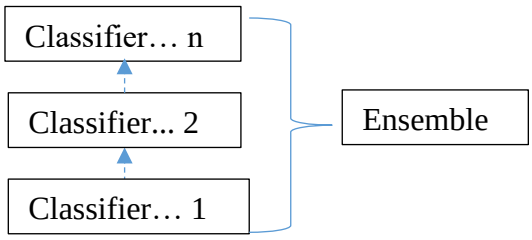

Figure 3.3 Ensemble model

The ensemble classifier is a combination of multiple classifiers (See Fig. 3.1) whose accuracy varies according to the accuracy of every single classifier (Hasan, Bakar, et al., 2015), (Hasan, Siraj, et al., 2015b). Since the ensemble method is itself a classifier, the combination of two or more ensemble classifiers can create a new ensemble method as follows. In

summary, the brand-new ensemble method can be represented mathematically,

$$Ensemble_{new} = \frac{Ensemble_1 + Ensemble_2 + \ldots\ldots\ldots\ldots Ensemble_n}{N} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where,

$$Ensemble_1 = First\ Ensemble\ method\ (i.e\ bagging)$$

$$Ensemble_2 = Second\ Ensemble\ method\ (i.e\ Boosting)$$

$$Ensemble_n = n\ number\ of\ Ensemble\ methods\ where\ n = a\ positive\ real\ number$$

$$N = Number\ of\ Ensemble\ methods$$

Alternatively, the formation of the new ensemble that can be written for this study is

$$Ensemble_{new} = \frac{Ensemble_1 + Decision\ tree_1 + \ldots\ldots\ldots\ldots Algorithm_n}{N} \ldots\ldots\ldots\ldots\ldots (2)$$

Where,

$$Ensemble_1 = First\ Ensemble\ method\ (i.e\ bagging\ or\ boosting)$$

$$Decision\ tree_1 = It\ could\ be\ Simple\ tree\ or\ complex\ tree\ or\ any\ other\ tree$$

$$Algorithm_n = It\ could\ be\ either\ n\ number\ of\ single\ decision\ tree\ or\ ensemble\ tree\ where\ n\ is\ a\ real\ positive\ number$$

$$N = Number\ of\ algorithms$$

More specifically this study employed the following equation to develop Ensemble_RH which is

$$Ensemble_{RH} = \frac{Ensemble_{bagging} + complex\ tree}{N} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (3)$$

Where,

$$Ensemble_{bagging} = Ensemble\ bagging\ method$$

$$complex\ tree = Decision\ tree\ algorithm$$

$$N = Number\ of decision\ tree\ algorithms$$

The primary research challenge here is which decision tree or which ensemble method should be chosen. We considered the test results for the preliminary analysis and selected a complex tree and bagged tree to build the new ensemble model known as Ensemble_RH because both obtained near perfect accuracy (see the section: a preliminary result). In the future, this study will explore mixing classifiers based on equations 1 and 2 above.

3.2 Justification of using decision tree with an Ensemble method

We have chosen a bigger correlation coefficient with less error. Deep learning obtained more error than the decision tree in all the error analysis. The run time required by deep learning is higher than decision tree analysis. Since the ensemble method works well in medical data settings and obtains a bigger correlation coefficient with fewer error ensemble methods it is applied in this study. Compared to ensemble learning, deep learning requires more data. It is often noticed that deep learning doesn't perform well when the data is small. Since cervical cancer is a very sensitive topic, so getting more data is not an easy task. Some literature declares that deep learning is similar in ways to ensemble-based learning. Deep learning may be thought of as an ensemble of neural networks.

3.2.1 Root Mean Squared Error

Root-mean-squared error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator, and the values observed. Fig 4.1 shows the RMSE between deep learning and decision tree. Decision tree obtained an RMSE

of 0.07 with 851 sec run time while deep learning obtained an RMSE of 0.08 with 9 sec.

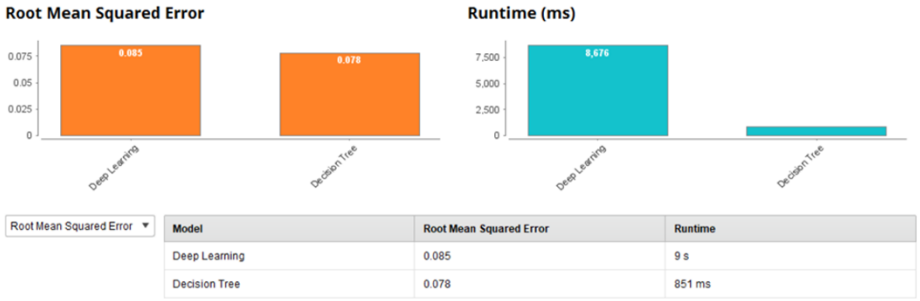

Fig. 4.1 RMSE and run time comparison between decision tree and deep learning

3.2.2 Absolute Error (AE)

Absolute error is the magnitude of the difference between the exact value and the approximation. The relative error is the absolute error divided by the magnitude of the exact value. The percent error is the relative error expressed in terms of per 100. Fig 4.2 shows the AE between deep learning and decision tree. Decision tree obtained an AE of 0.02 with 851 sec run time while deep learning obtained an AE of 0.03 with 9 sec.

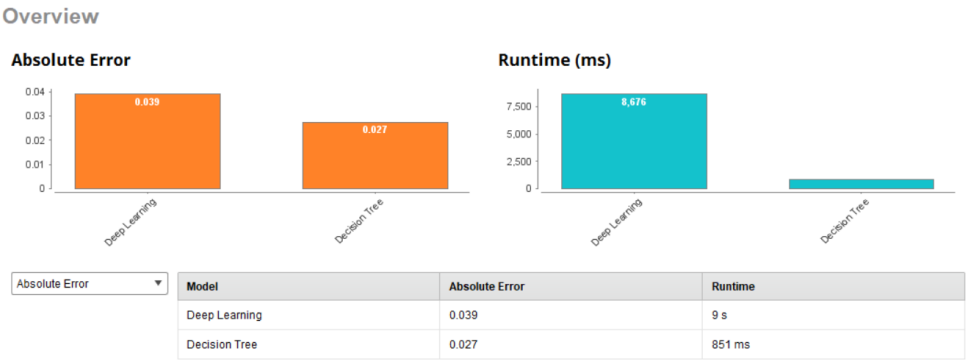

Fig. 4.2 AE and run time comparison between decision tree and deep learning

3.2.3 Relative Error Lenient (REL)

The average lenient relative error is the average of the absolute deviation of the prediction from the actual value divided by the maximum of the actual value and the prediction. Fig 4.3 shows the REL between deep learning and decision tree. Decision tree obtained an REL of 87.2% with 851 sec run time while deep learning obtained an REL of 100% with 9 sec.

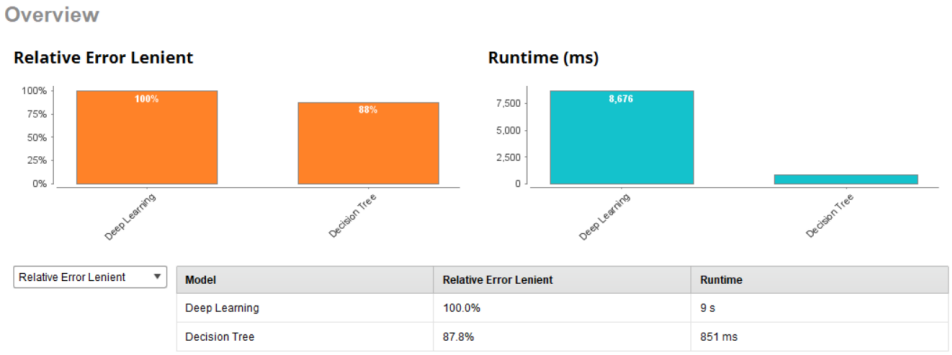

Fig. 4.3 REL and run time comparison between decision tree and deep learning

3.2.4 Squared Error (SE)

The mean squared error (MSE) or meant squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated. Fig 4.4 shows the SE between deep learning and decision tree. Decision tree obtained an SE of 0.006 with 851 sec run time while deep learning obtained an SE of 0.007 with 9 sec.

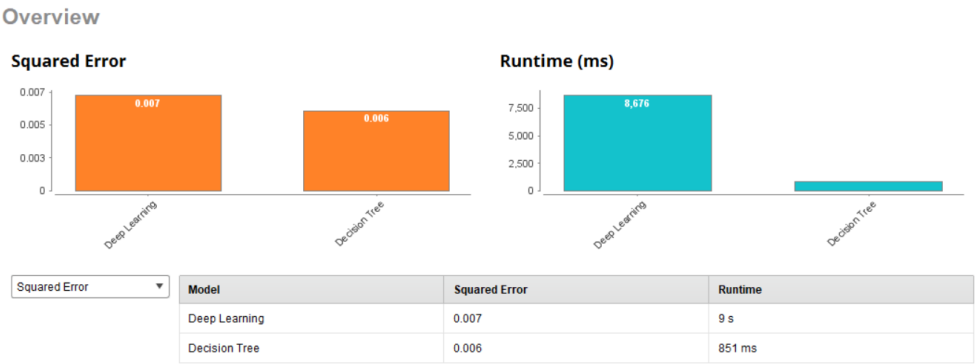

Fig. 4.4 SE and run time comparison between decision tree and deep learning

3.2.5 Squared Correlation (SC)

This returns the squared correlation coefficient between the label and prediction attributes. Fig 4.5 shows the SC between deep learning and decision tree. Decision tree obtained an SE of 0.001 with 851 sec run time while deep learning obtained an SE of 0.005 with 9 sec.

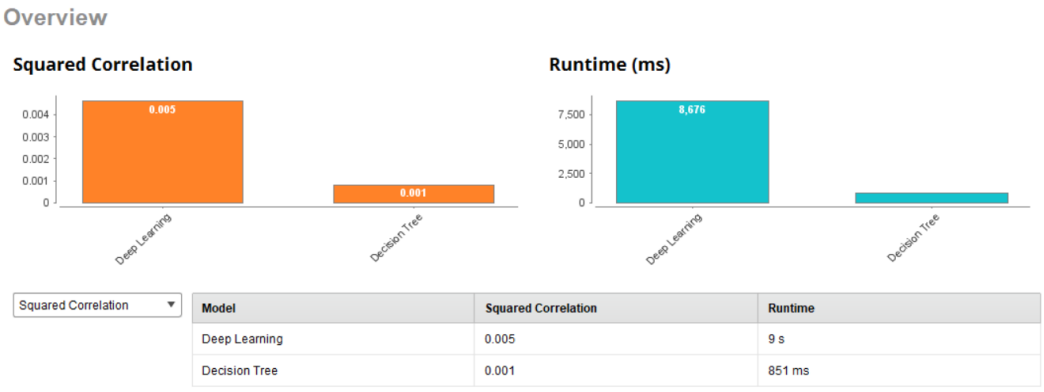

Fig. 4.5 SC and run time comparison between decision tree and deep learning

3.3 Summary

In summary of the comparison between deep earning and decision tree performance, we can conclude that, for our data, decision tree is more suitable as it always obtained less error and less run time compared to deep learning.

At this stage in this study, the ensemble classifier model for medical data based on Decision Tree will be built that appear to have high quality from a data analysis perspective. In this phase, the performance between the single DT classifier and ensemble classifier will be compared and discussed. Finally, the justification for the study will be concluded here.

In this chapter, the methodology has been discussed step by step. The research design and methods to build Ensemble_RH were explained with the research steps for this study.

Chapter 4

Experiments

In this study, we have employed default parameters for machine learning and NN algorithms. This study is not comparison study. More than 95% accuracy with ensemble method with less statistical error is considered acceptable accuracy in this study. To evaluate the performance of the proposed ensemble classifier method, a number of experiments with different classifiers was carried out. The results were then compared with the results of the proposed ensemble classifier method. Finally, the ensemble method was enhanced with ensemble_RH and validated the outcome with statistical methods such as root squared error mean absolute error.

4.1 Experimental Results

The preliminary results are divided into subsections: modelling the classification techniques, proposed modelling technique, and a summary of the preliminary results. In this section simple tree, complex tree, linear SVM, boosted tree and bagged tree have been employed to model classification. In the figure, the left side of the x-axis is the accuracy in terms of a correlation coefficient in percentage; the right side of the x-axis is the average error where the error can be from 0 to 1000. The highest coefficient accuracy with the least error is the optimal feature selection. 66% of training and 34% testing regime have been employed in this study.

4.2 Modelling the classification Techniques

Several classification techniques have been employed using experimental data. Fig 4.1 is a training analysis of decision tree classifiers, which depicts the medical data modelling problem for this study. We have chosen simple tree (95.6%), linear SVM (51.7%), boosted tree (97%), bagged tree (94.97%) and complex tree (97.9%). We noticed that linear SVM obtained the lowest accuracy of 51.17% while other classifiers obtained an accuracy close to 95%. Surprisingly, two or more classifiers obtained nearly similar accuracy: complex tree and boosted tree obtained closely 97%. A similar pattern was also observed in simple tree and bagged tree, which is closely 95%. Most of the literature identified that the different performance may be obtained due to multivariate, missing values, outliers, and multi-classes. Some literature suggests that employing an ensemble model may validate accuracy. However, most of the previous studies focused on only one classifier problem and either eliminated missing values or filled up missing values with the statistical method. However, medical experts do not agree with these methods of data handling (Hasan, Gholamhosseini, et al., 2017). In this research, we have tried to propose an ensemble method without involving data pre-processing techniques that are in line with the medical professionals' views.

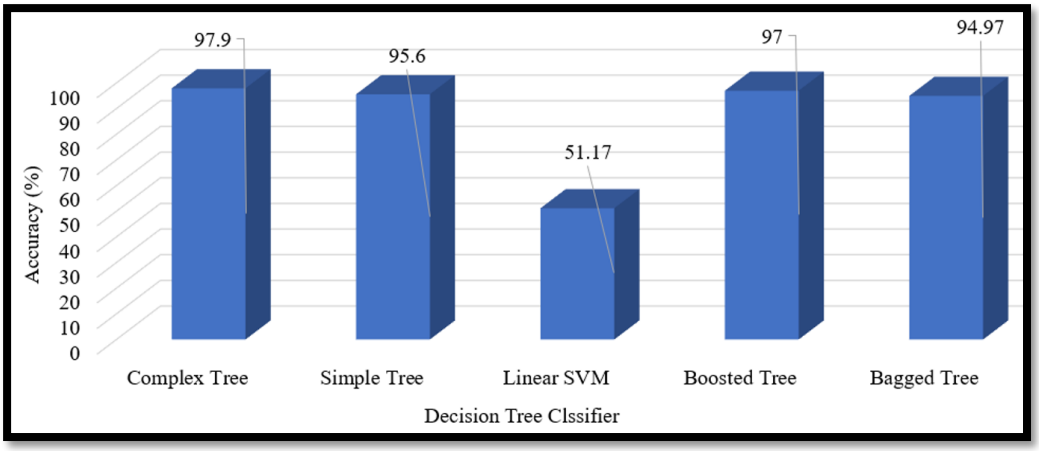

Figure 4.1 The accuracy of different decision tree classifiers

Fig 4.2 Explains the test results of the same data used above by decision tree classifiers. We have noticed a similar pattern of results like Fig. 4.1 but a huge change of performance of Linear SVM, which is 95.1%.

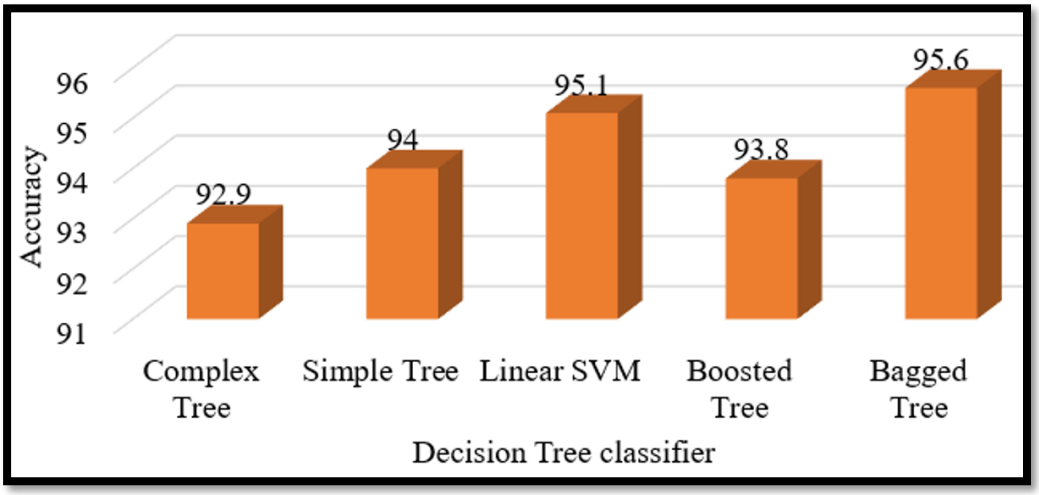

Figure 4.2 The accuracy of different decision tree Classifiers

Fig 4.3 is the performance between test and train analysis. The interesting point here is Linear SVM improved accuracy nearly 40% more than the test. The point may be noted that complex tree performance has been reduced by more than 5% in testing results while it was the highest performer (97.9%) during test analysis.

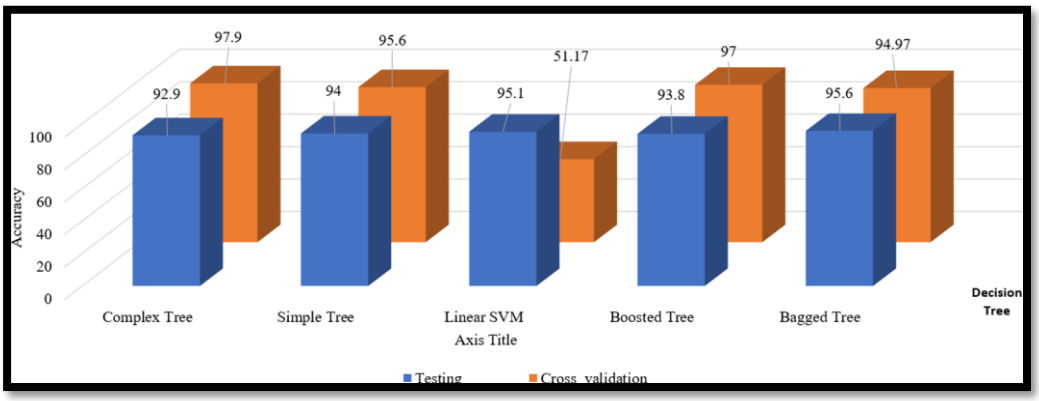

Figure 4.3 Comparison between test and training analysis

4.3 Missing values

Fig. 4.4 depicts that the number of maximum missing values is 117, and the minimum is 0. This means we cannot ignore any feature, and this is very high dimensional data.

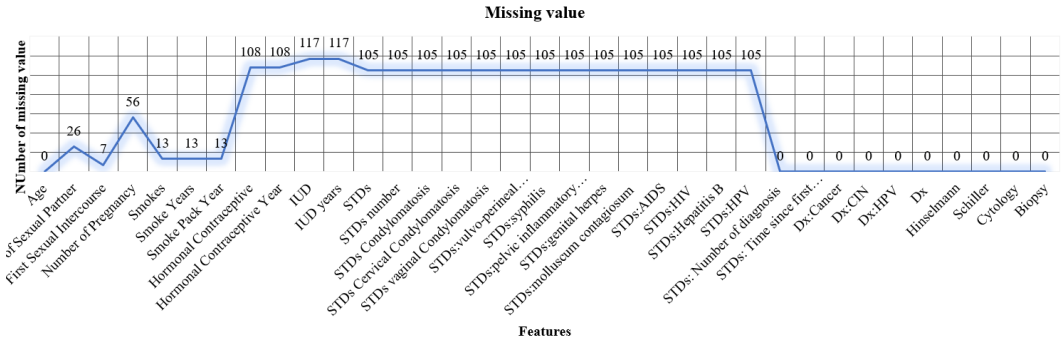

Figure 4.4 Feature success rate and missing values

4.4 Experimental Feature Selection

This study will use the feature selection method of data modelling technique and identify the relationship in real-life problems by employing knowledge acquisition, symptom mining, and case-based reasoning. Fig. 4.5 clarifies all the influential features that may be closely related to cervical cancer. The classification accuracy is for features obtained by test and train. During feature selection, ‘biopsy’ is chosen as a predictor and other features as a predictor. Feature ‘biopsy’ is chosen because in this study, we are proposing an intervention framework which will predict cervical cancer earlier than the cancerous stage.

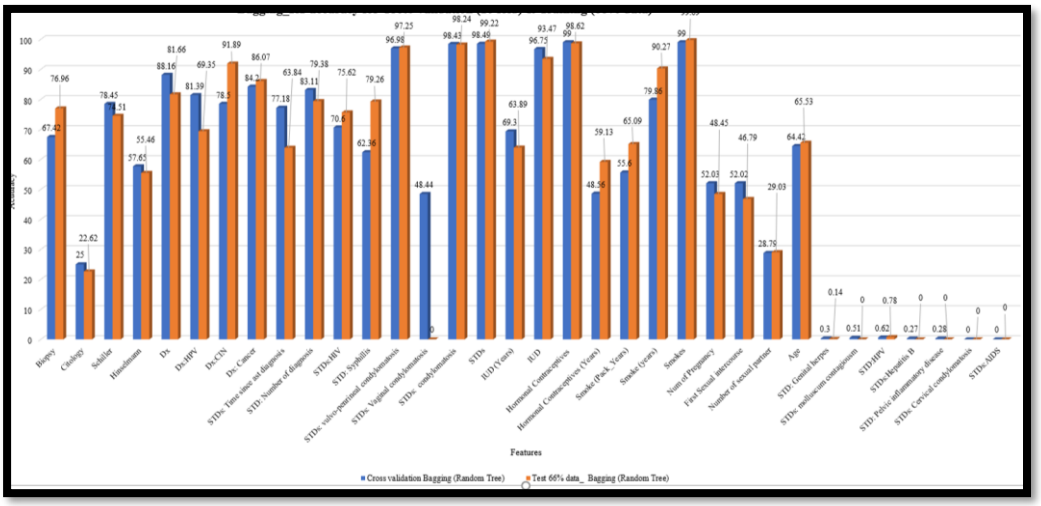

Figure 4.5 Influence of all features

4.5 Relevancy of STD features with cervical cancer

The interesting point from Fig 4.5 is it reveals that all features may not be equally influential in cervical cancer such as, we can notice here that all STD features are not important. Hence, this study has a deeper look at the features of STD. Fig. 4.6 shows that during the test analysis, all features in STDs are not very influential in cervical cancer except STDs (98.49%), STDs: vulvo-perineal condylomatosis (96.98%), and STDs: condylomatosis (98.43).

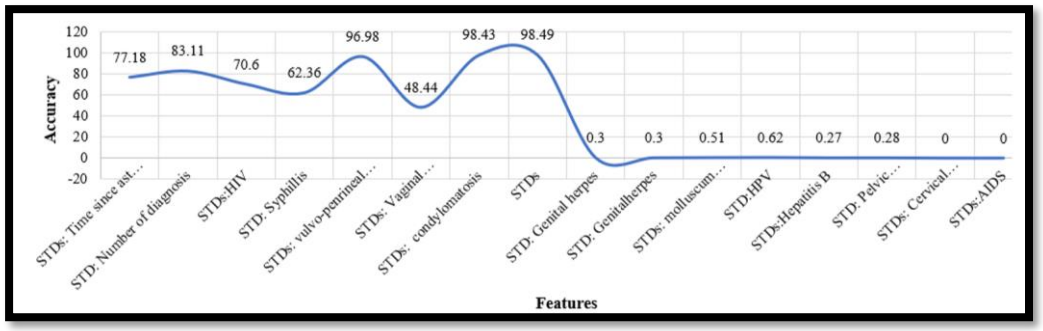

Figure 4.6 Influence of STD features based on training analysis

In Fig. 4.6, it is clearly shown that STDs: vaginal condylomatosis has an influence of 48.4% while in the test analysis of the same feature extraction Fig. 4.7) we have identified that STDs: vaginal condylomatosis has no relation (0%) with cervical cancer.

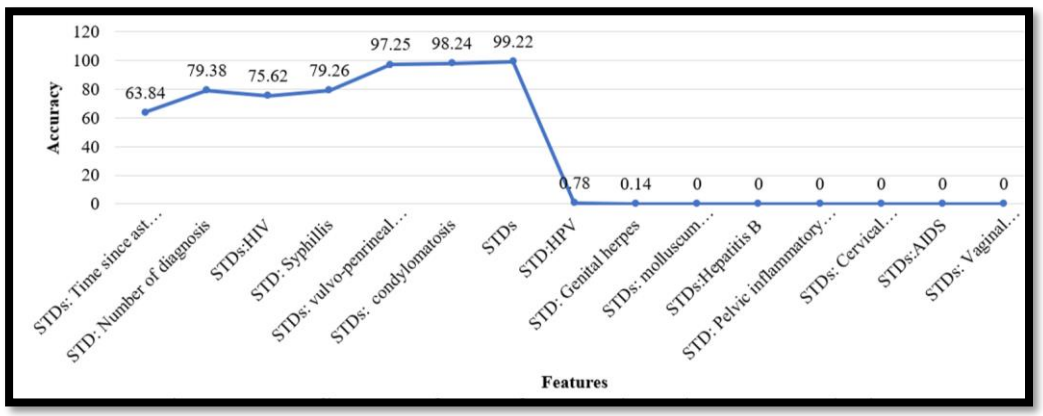

Figure 4.7 Influence of STD features based on test analysis

4.6 Relevancy of HIV and AIDS Features with cervical cancer

Fig. 4.6, and Fig. 4.7 show that cervical cancer may be influenced by STDs: HIV but not by STDs: AIDS which seems unusual but in reality, is not. Hence, data mining knowledge alone may not be sufficient to extract the relationship between features for an accurate intervention framework to identify the risk factor of having cervical cancer. For this reason, this research proposed that expert rule mining and case-based reasoning is important to identify and clarify the relationship between features. Once we have mined the rules of STs: HPV, STDs: HIV, and STDs: AIDS from an expert and utilized case-based reasoning, we have the answer for these unusual results from data mining. We found that HPV causes cervical cancer, but it requires a minimum of 10 years. If HPV is detected early cervical cancer is preventable. Similarly, if anyone is infected with HIV; he or she may be diagnosed as an AIDS patient after 10 years. Hence, we see an HPV infected person requires a minimum 10 years to suffer from cervical cancer while it is preventable if detected early, and an HIV infected person requires 10 years to suffer from AIDS. On the other hand, the lifespan of AIDS patients is normally no more than two years. So the AIDS patients normally do not survive for another 10 years as they may get cervical cancer after HPV infection. Hence, this research is a bridge between data mining and expert rule mining and case based reasoning.

4.7 Relevancy of HPV Features with cervical cancer

Fig 4.8 employs several machine learning algorithms and shows the relevancy of the HPV feature to cervical cancer. The lowest relevancy rate is 63.07% by M5P algorithm, and the highest is above 80% with several machine learning algorithms: Decision stump, RepTree, ANN, SVM, and bagging.

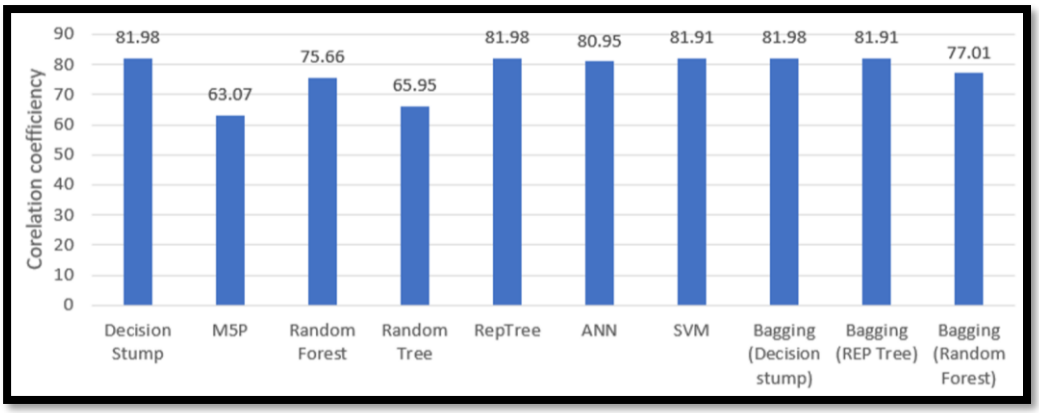

Figure 4.8 Influence of STD features based on test analysis

4.8 Relevancy of Smoking Features with cervical cancer

Fig. 4.9 reveals that smoking may be one of the relevant features that may influence cervical cancer yet it is controversial (refer to the literature review section). It depicts the influence of having cervical cancer by smoking patterns such as what is the influence of a pack of cigarettes per day for a year, more than a packet per day for a year and chain smoking. Our empirical study shows that if anyone continues smoking a packet of cigarettes per annum she may have a chance of 24% to suffer from cervical cancer, smoking more than one packet in a year increases the chance of suffering from cervical cancer to 34%, and they may have an increased chance by 42% of having cervical cancer if they are a chain smoker.

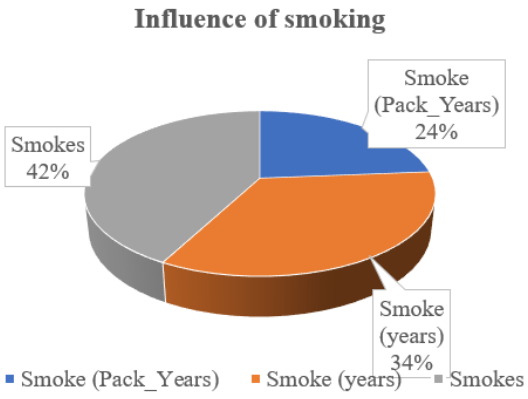

Figure 4.9 Influence of Smoking

4.9 Justification of the feature relevancy by the average error of Root Mean Squared Error and Mean Absolute Error)

Several machine learning algorithms have been employed to identify the relevancy of the features and the validation of feature relevancy has been done by statistical error methods.

Fig 4.10 shows that biopsy features are required for cervical cancer diagnosis as the error is very less. The lowest error is 0.09 and the highest error is 0.2 which clearly statistically validate the feature is relevant.

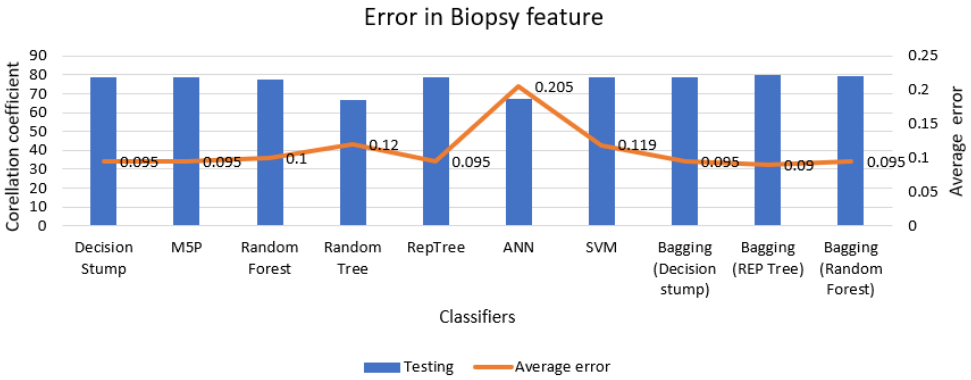

Figure 4.10 Accuracy and average Error in Biopsy feature

Fig 4.11 shows that HIV features are required for cervical cancer diagnosis as the error is very less. The lowest error is 0.04 and the highest error is 0.07 which clearly statistically validate the feature is relevant.

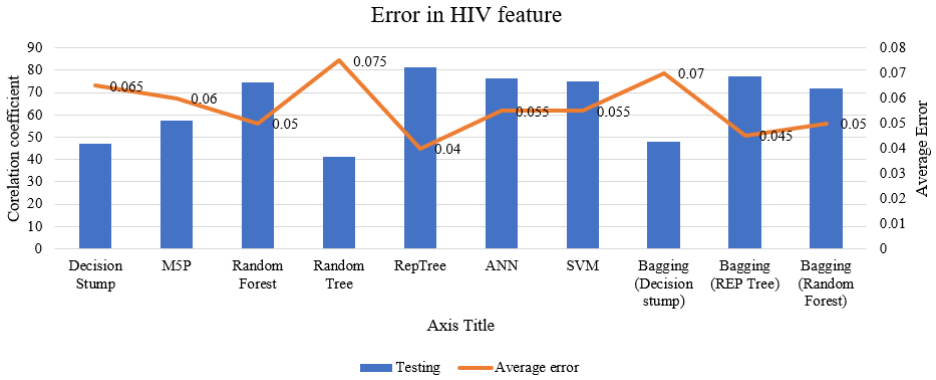

Figure 4.11 Accuracy and average Error in HIV feature

Fig 4.12 shows that HPV features are required for cervical cancer diagnosis as the error is very less. The lowest error is 0.05 and the highest error is 0.8 which clearly statistically validate the feature is relevant.

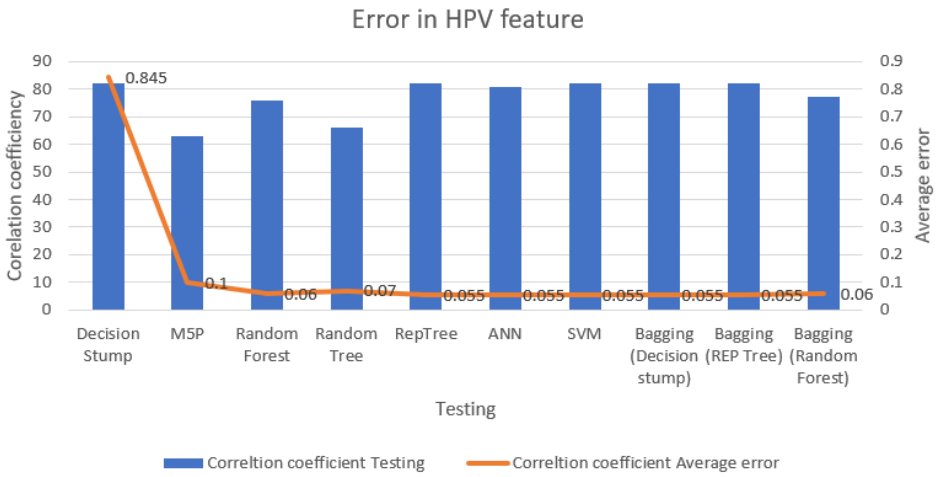

Figure 4.12 Accuracy and average Error in HPV feature

Fig 4.13 shows that "Number of sexual partners" features are not required for cervical cancer diagnosis as the statistical error are too high. The lowest error is 1.1 and the highest error is 2.2 which clearly statistically validate the feature is not relevant.

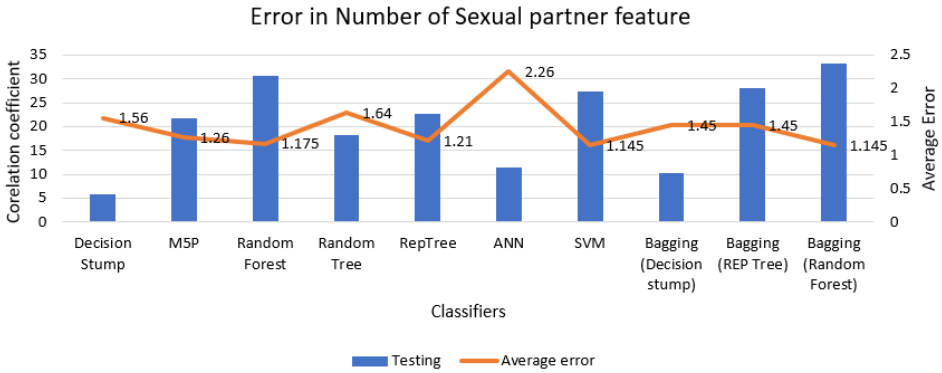

Figure 4.13 Accuracy and average Error in Number of sexual partner feature

Fig 4.14 shows that "Number of pregnancies" features are not required for cervical cancer diagnosis as the statistical error are too high.

The lowest error is 0.5 and the highest error is 1.4 which clearly statistically validate the feature is not relevant.

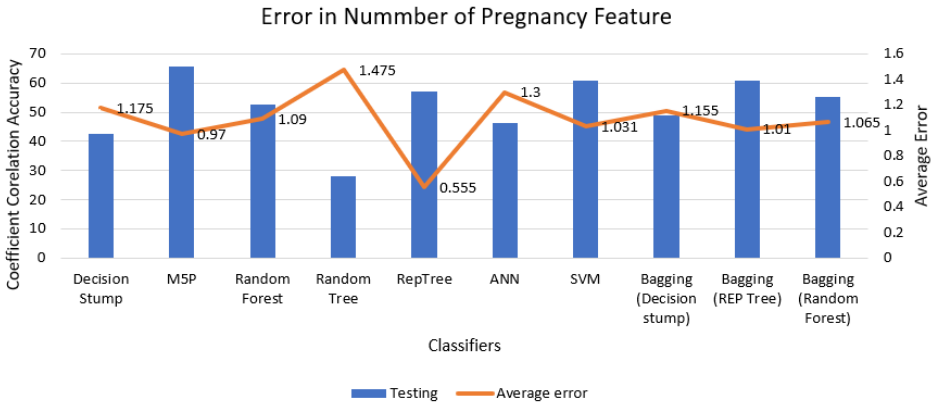

Figure 4.14 Accuracy and average Error in Number of pregnancy feature

Fig 4.15 shows that "AIDS" features are presenting unusual statistical error which is 0 in all case. More clear explanation of this feature has been described later in this study.

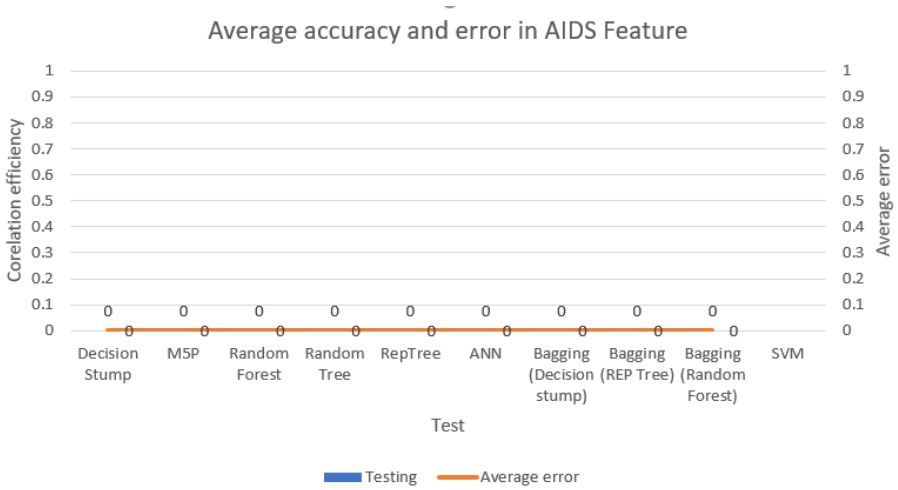

Figure 4.15 Accuracy and average Error in AIDS feature

4.10 Justification of the feature selection accuracy by True Positive (TP) and False Positive (FP) error

From the earlier section in 4.9, we have identified a few features that are related to cervical cancer, and a few that are not. Table 4.1 shows the True Positive (TP) and False Positive (FP) rates and it shows similar findings to section 4.9. In the previous section, I claimed that not all STD features are related such as STD: HIV and STD: HPV and obtained high TP,

but the rate of FP is also very high. Number of sexual partners and number of pregnancies obtained high accuracy with high error in a previous analysis in section 4.9. The TP and FP analysis also supports that the identification from the previous analysis was correct. My analysis shows that smoking may be related to cervical cancer, but it obtained a bit high FP though TP is perfect. From the literature survey, the relevancy of smoking with cervical cancer is still debatable. My findings from an ensemble perspective and statistical methods give a hint that medical researchers need to have a deep look into smoking features to identify their relevancy with cervical cancer.

Table 4.1: True Positive (TP) and False Positive (FP) error

Features	TP Rate	FP Rate
STD all	1	0.051
STDs: condylomatosis	1	0.068
STDs:Vulvo-periniul condolytomasosis	0.999	0.001
STD: HIV	0.999	0.889
STD: HPV	1	1
Smokes	1	0.618
Number of sexual partners	0.34	0.217
Number of Pregnancy	0	0.011

4.11 Classification algorithm performance for cervical cancer data

Several classification techniques have been employed using experimental data. It is noticed that the linear SVM achieved the least accuracy of 50.8% (Fig. 4.16 and Fig 4.17).

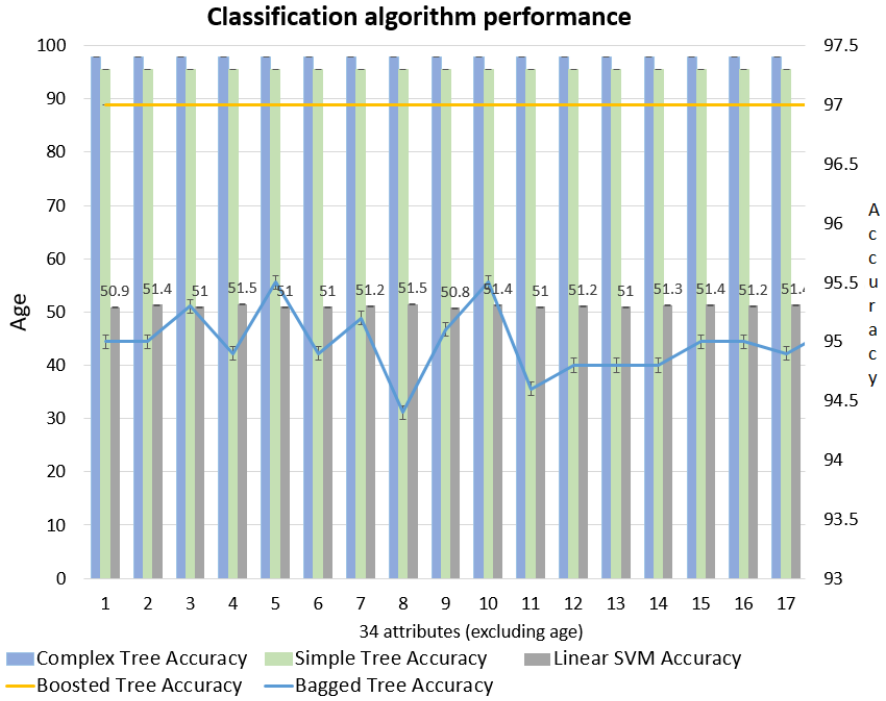

Figure 4.16 Classification algorithm performance (1st half)

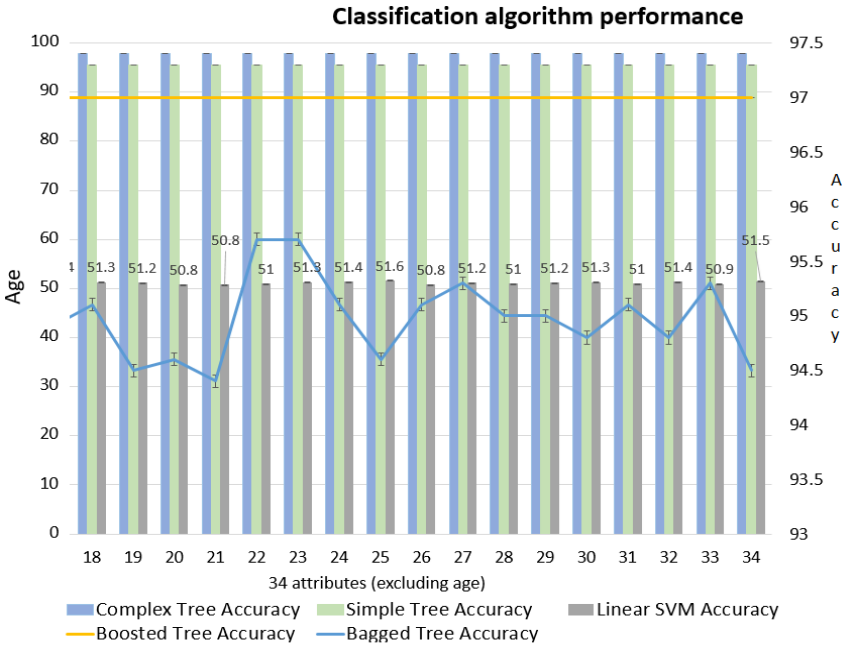

Figure 4. 17 Classification algorithm performance (2nd half)

In this stage, four decision tree classifiers have been applied: complex decision tree (97.9%), simple decision tree (95%), simple tree (95.6%), boosted tree (97%), bagged tree (max 95.7%). Fig. 4.18 uncovers

that the best performance is acquired by the bagged tree (min 95.5 ~ max 95.7%) among the other classifiers (because the accuracy is free from bias). However, a consistent accuracy of the simple tree classifier (95.6%), complex tree classifier (97.9%) and boosted tree classifier (97%) reveal that the outcome is biased to get a positive result. This is a direct result of the data suffering from missing values, and it is multivariate data. Hence, among the classifiers, strange constant accuracy has been noticed, which is irregular.

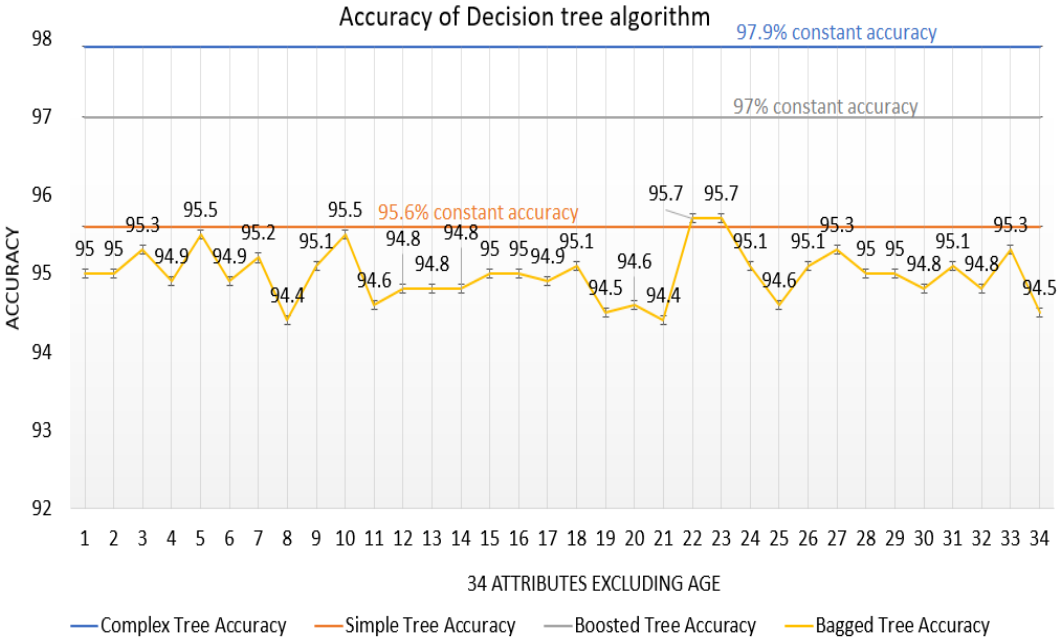

Figure 4.18 Decision tree accuracy

From the analyses, it is noted that the ensemble bagged tree performs better with an accuracy of more than 95% (see Fig. 4.19).

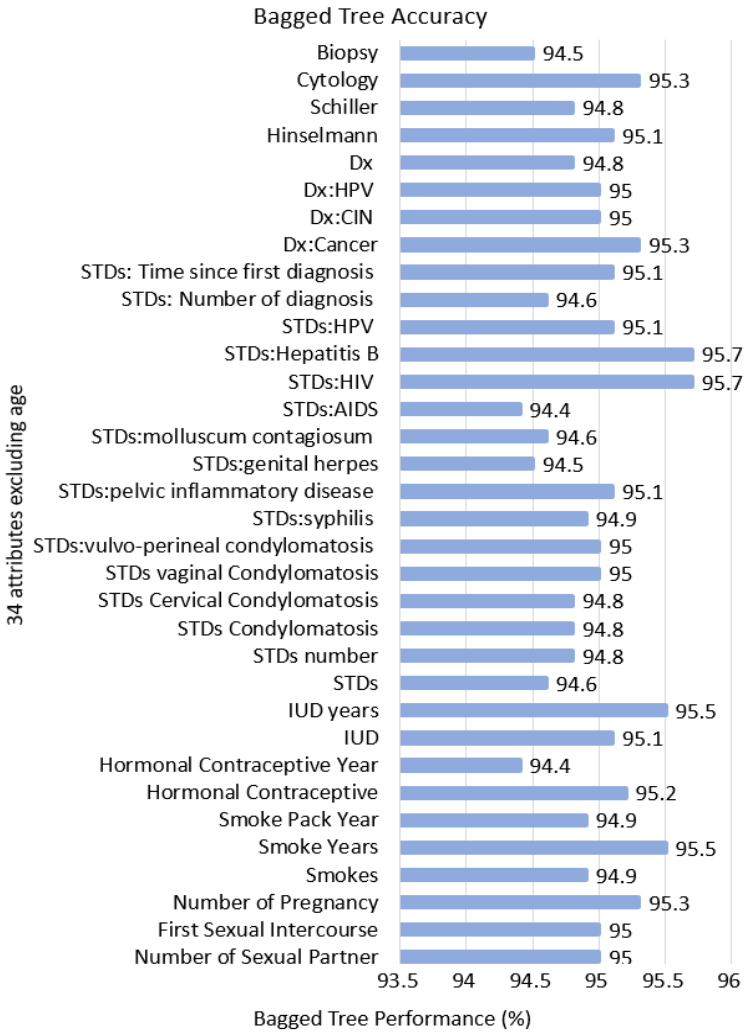

Figure 4.19 Bagged tree

4.12 Proposed modelling technique

The bagged tree is an ensemble method, which has acceptable and satisfactory performance as it obtained better accuracy. However, there is always room to enhance efficiency. In this study, the proposed Ensemble_RH can be represented as:

$$Ensemble_{RH} = \frac{Ensemble_{bagging} + complex\ tree}{N}$$

Where,

The algorithm is mixed with an ensemble method (bagging) and a single decision tree (complex tree).

N= Number of algorithms employed

The new Ensemble_RH has improved the accuracy more than the bagged tree, which is over 96% (see Fig. 4.20 (first half) and Fig. 4.21 (second half)). We claim that Ensemble_RH has the potential to improve classification accuracy.

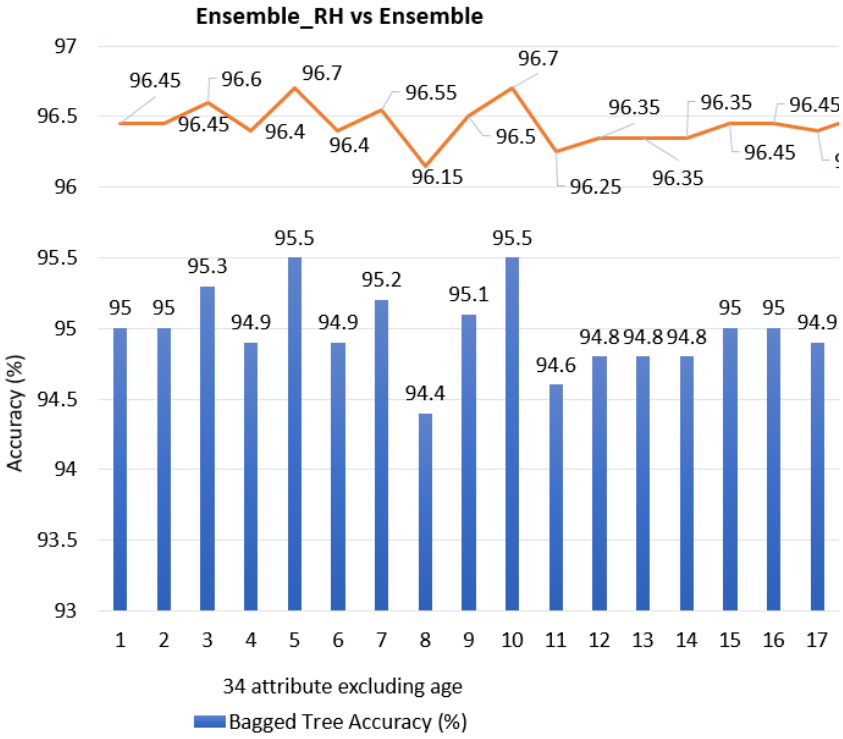

Figure 4.20 Bagged tree vs Ensemble_RH (1st half)

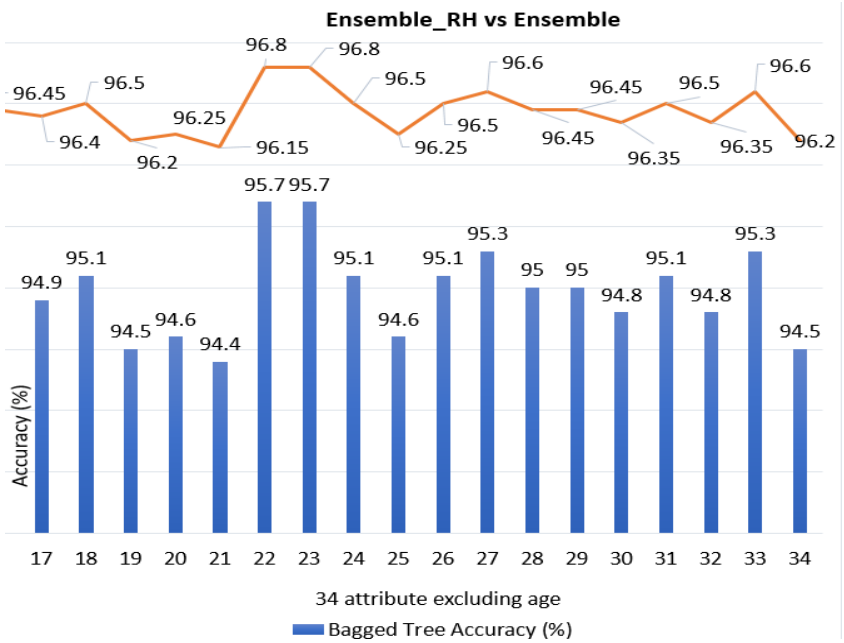

Figure 4.21 Bagged tree vs Ensemble_RH (2nd half)

Chapter 5

Contributions

This chapter discusses the contributions in this study.

5.1 Main Contributions

In this study, through a literature survey and experimental analysis, we have identified that decision tree classification is a better model for medical data modelling. A single classifier often does not perform well due to misclassification when data is missing and multivariate. The most interesting part was to find a suitable ensemble model for multivariate medical data. In the future, this study may try to explain why the performance of the single classifier is not as good as the ensemble classifier.

The outcomes are preliminary and meant to demonstrate that ensemble learning without imputation is no worse than ML with imputation. The aim was not to show that the accuracy was significantly improved.

5.1.1 Ensemble model without pre-processing

Based on the literature review it was found that ensemble is one of the best methods for early detection of cervical cancer. In this study, we have employed a decision tree classifier. We proposed a novel ensemble method called ‘Ensemble_ RH',’ which offers expected 96.3% accuracy in the experimental analysis while the ensemble method bagged tree obtained 94.97%. The most challenging part was to evade data pre-processing as it is not a favourable and agreeable option by physicians. Our proposed ensemble model outperformed without employing data pre-processing techniques.

5.1.2 The performance of Ensemble models (bagging vs Ensemble_RH) without preprocessing

The bagged tree is an ensemble method, which has an acceptable and satisfactory performance as it obtained 94.97%, which is the second lowest performer. However, there is always room to enhance efficiency. Fig. 4.16 shows that the new Ensemble_RH obtained 96.43%, which is 1.46% more than the bagged tree performance of 94.97% when a bagged tree is combined with a complex tree.

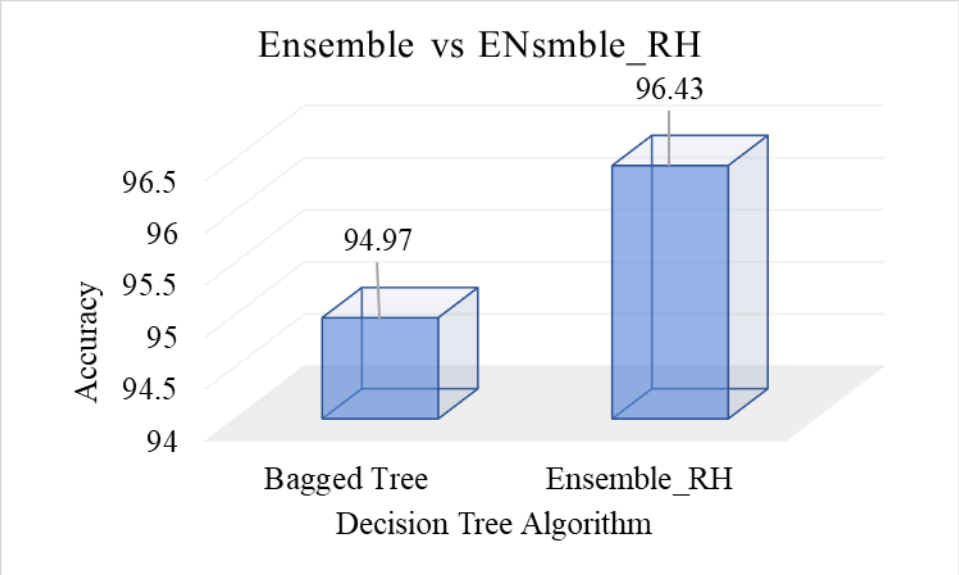

Figure 5.1 Bagged tree vs Ensemble_RH

5.1.3 Suitable feature selection/ knowledge discovery

From the data we have identified related features based on feature accuracy where “biopsy” was the response and other features were the predictor. We identified that both smoking (a feature from data) and non-smoking (a feature from data) may have a direct or indirect relationship with cervical cancer. It is also noticed that all the features in STDs are not related

to cervical cancer. We explained the relationship between the features when data mining knowledge is not enough to explain the unusual relationship.

This study resolves the contradictory relationship between HPV, HIV and AIDS features. This study also explains how the HPV feature is related to AIDS but not with AIDS, which seems a contradiction as HIV and AIDS are closely related. From the data analysis in Chapter 4 and the review of the literature, we can conclude that HIV-infected patients need a minimum of 10 years to be infected by AIDS and AIDS patients live normally one or two more years while HPV infected patient need no more than 10 years to be infected by cervical cancer. Reasonably if an AIDS patient is infected by HPV, they will not survive 10 years to develop cervical cancer. Since HIV infection takes 10 years to develop AIDS and HPV takes no more than 10 years to develop cervical cancer, the patient with HIV has a chance to be infected by HPV. These support the findings from this research that the feature HPV may be related to HIV but not to AIDS.

Chapter 6

Conclusion and Future work

This chapter includes the implications of using the ensemble model in classification technology for cervical cancer data sets, a new approach of creating an ensemble model, modelling Ensemble_RH, conclusion and future work. Section 6.1 presents considerable findings from Chapters 4 and 5. The thesis is summarized in Section 6.4. Finally, several conceivable future advancements of this research are presented in Section 6.5 for possible future research directions.

6.1 Implications of using ensemble modelling for the cervical cancer data set

Modelling cervical cancer data allows us to identify which information is more relevant and contributes more to cervical cancer risk. This study aims to address the current challenges to obtain better accuracy in modelling life critical data (i.e. multivariate cervical cancer data) without missing value imputation in ensemble modelling. Cervical cancer data is multivariate with many missing values. In the view of machine learning, obtaining better classification accuracy is very difficult when data suffers from missing values and outliers. It is also not always right to depend on accuracy only because sometimes classification technology may offer better accuracy but a huge statistical error such as Root Squared Mean error (RSME) and so on. In terms of the medical view, cervical cancer is the second most common cancer (Scarinci et al., 2011) of the cervix, which is the lower part of the uterus or womb (Southern Cross Medical Library, 2013). For both views in machine learning and medical data, cervical cancer is very important to be looked at to obtain better accuracy and identify some features that may be relevant to cervical cancer with the ensemble approach is a research interest. The more challenging part in this data set is to improve accuracy without employing any preprocessing techniques and with minimum statistical error in the features. This study proposed a new

ensemble method, Ensemble_RH, which achieved better accuracy than other single classifiers (discussed in Chapters 4 and 5).

6.2 The novelty of ensemble approach and proposed new ensemble model Ensemble_RH

Life critical data like cervical cancer data where missing values are present and missing value imputation is not a favourable option by the medical professionals (as it changes the diagnosis outcome), my research shows that choosing the right base classifiers in the ensemble and the proposed new ensemble, Ensemble_RH, offers better accuracy without missing value imputation. To my best knowledge, most of the existing researchers employ data preprocessing to improve accuracy, which is dangerous as it often changes the medical diagnosis. Figure 6.1 shows a systematic study of different classifiers may improve ensemble accuracy. The choice or selection of base classifiers in the ensemble model may improve the accuracy. Random forest, itself an ensemble classifier, obtained 77.7%. When we introduce bagging and random tree as a base classifier, it improved the accuracy to 79.51%, which is 1.81% higher. Though RepTree is a single classifier, we noticed a similar trend happened with RepTree. Rep Tree alone obtained 78.7 but bagging with RepTree obtained 80.2%. This study claimed that a systematic approach could improve the accuracy of the ensemble model without missing value imputation, and Fig 6.1 shows that the proposed Ensemble_RH obtained 96.43%, which shows a dramatic increase of 16.23%.

This study avoids using data preprocessing techniques for medical data as medical doctors usually do not prefer it as it may change the diagnosis result even though it may have better accuracy in terms of data mining view. The novelty of this research is it did not impute missing values yet obtained accuracy as high as 96.43% in the test analysis.

Figure 6.1 Comparison between classifiers, Ensemble and Ensemble_RH

6.3 Feature selection with an Ensemble approach

Feature selection from medical data is acknowledged as an area of increasing importance, yet also poses many difficulties (Hasan, Bakar, et al., 2015), (Krawczyk & Schaefer, 2012). One of the key difficulties is medical professionals do not believe data pre-processing because ignoring or filling up the missing values with a statistical approach may change real-life diagnosis outcomes. For instance, medical doctors test the antibodies to hepatitis B before offering hepatitis B vaccines (US department of veterans affairs, 2018a) because if a person is already exposed to the hepatitis B viruses, then the person may get protection from an injection of hepatitis B immunoglobulin (HBIG), which is different from the hepatitis B vaccine and (US department of veterans affairs, 2018b) similarly, the doctor cannot rely on a statistical method to alter cervical cancer data to predict cervical cancer assessment. For this reason, data scientists need to find a way so that they will not be dependent on data pre-processing techniques to achieve high accuracy on extracting influential features that are closely related.

This research aimed to identify some features that cause cervical cancer or are influential in cervical cancer. Cervical cancer data has been obtained from UC Irvine (known as UCI) machine learning repository that is openly accessible by https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. There are 858 instances and 35 attributes in this

data. Keeping in mind that data pre-processing may change the medical diagnosis results, this study employed several classification approaches on cervical cancer data without involving data preprocessing techniques. Among the classification algorithms decision tree (Decision Stump, M5P, Random Forest, Random Tree, REP Tree), Artificial Neural network (ANN), Support Vector Machine (SVM) and bagging as an ensemble method have been utilized to extract the influential and related features to the Human Papilloma Virus (HPV) without the dependency of data preprocessing techniques.

6.4 Conclusion

Selecting an efficient classifier for medical data is considered one of the most important parts of today's machine learning aided diagnosis. The performance of single classifiers such as decision tree classifier can be increased by the ensemble method. However, this approach relies on data quality and missing values. To my best knowledge, this is the first systematic approach of ensemble study specially on life critical data where we can achieve good accuracy without missing value imputation. Missing value imputation may offer good accuracy in data science, but it is not an acceptable option to build an expert system to assist medical doctors as missing value imputation often leads to the wrong diagnosis.

6.5 Future research direction

This research is the first systematic approach of ensemble learning without missing value imputation specially on decision trees on ensemble learning. In the future, we may consider ensemble learning with other classifiers such as ANN and SVM. It would be a good idea to generate a new data set without missing values, then apply the concept of this study with 5% to 30% missing values and observe the performance of the ensemble model. At present, we could not apply deep learning as cervical cancer data is small and unsuitable for deep learning without preprocessing. Applying Ensemble_RH on big data or large data would be a good idea. In

the future, we may identify the performance of Ensemble_RH and deep learning for big data. A best challenging idea is to investigate when the ensemble model fails to obtain better accuracy and find out whether deep learning could take over in that situation.

6.6 Future research on missing value imputation

In future, we may need to find the best way I can apply my proposed method or find the best data pre-processing techniques to deal with missing value efficiently. In future, I may test my methods for other life critical data that has more missing values. At that moment, I may conclude that missing value imputation for all life critical data is not necessary to improve ensemble classification accuracy.

References

Adhvaryu, P. S., & Panchal, P. M. (2012). A Review on Diverse Ensemble Methods for Classification. *Journal of Computer Engineering (IOSRJCE)*, *1*(4), 27–32.

Alizadeh, & Parvin. (2011). Surface matching degree. *Australian Journal of Basic and Applied Science*, *5*(9), 653–660.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*(1), 545–1588.

Arabnia, H. R., & Tran, Q.-N. (2011). *Software Tools and Algorithms for Biological Systems* (First; Q.-N. T. Hamid R. Arabnia, Ed.). New York: Springer Press.

Arsene, O., Dumitrache, I., & Mihu, I. (2011). Medicine expert system dynamic Bayesian network and ontology based. *Expert Systems with Applications*, *38*, 15253–15261.

Aymerich, F. X., Alonso, J., Cabañas, M. E., & Comabella, M. (2011). Decision tree based fuzzy classifier of 1 H magnetic resonance spectra from cerebrospinal fluid samples. *Fuzzy Sets and Systems*, *170*(1), 43–63. https://doi.org/10.1016/j.fss.2011.01.003

Blagus, R., & Lusa, L. (2015). Boosting for high-dimensional two-class prediction. *BMC Bioinformatics*, *16*(1), 1–17. https://doi.org/10.1186/s12859-015-0723-9

Breiman, L. (1999). *Using adaptive bagging to debias regressions*.

Breiman, L. (2001). *Some infinity theory for predictor ensembles*.

Brown, G. (2010). Ensemble Learning. In *Encyclopedia of Machine Learning* (pp. 1–24). Springer Press.

Chandra, B. (2011). *Heterogeneous Node Split Measure for Decision Tree Construction*. (1996), 872–877.

Chaurasia, V., & Pal, S. (2017). *Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2994925

Che, D., Liu, Q., Rasheed, K., & Tao, X. (2011). Software Tools and Algorithms for Biological Systems. *Advances in Experimental Medicine and Biology*, *696*(1), 191–199. https://doi.org/10.1007/978-1-4419-7046-6

Chourasia, S. (2013). Survey paper on improved methods of ID3 decision tree classification. *International Journal of Scientific and Research Publications*, *3*(12), 1–4.

Conroy, B., Eshelman, L., Potes, C., & Xu-Wilson, M. (2016). A dynamic ensemble approach to robust classification in the presence of missing data. *Machine Learning*, *102*(3), 443–463. https://doi.org/10.1007/s10994-015-5530-z

Darwish, D. (2013). Data Mining: Concepts, Models, Methods, and Algorithms. *International Journal of Computer Science*, *10*(4), 103–111.

Devroye, L. (2008). Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, *9*(1), 2015–2033.

Dittman, D. J., Khoshgoftaar, T. M., & Napolitano, A. (2015). Selecting the Appropriate Ensemble Learning Approach for Balanced Bioinformatics Data. *International Florida Artificial Intelligence Research Society*, 329–334.

Elshazly, H. I., Elkorany, A. M., Hassanien, A. E., & Azar, A. T. (2013). Ensemble classifiers for biomedical data: Performance evaluation. *2013 8th International Conference on Computer Engineering & Systems (ICCES)*, 184–189. https://doi.org/10.1109/ICCES.2013.6707198

Farid, D. M., Maruf, G. M., & Rahman, C. M. (2013). A new approach of Boosting using decision tree classifier for classifying noisy data. *2013*

International Conference on Informatics, Electronics and Vision (ICIEV), 1–4. https://doi.org/10.1109/ICIEV.2013.6572718

Fauci, A. S., Braunwald, E., Kasper, D. L., Hauser, S. L., Longo, D. L., & Jameson, J. L., & Al., E. (2009). *Featuring the complete contents of Harrison's principles of internal medicine* (17th ed.). McGraw Hill. Harrison's Online.

Fensterstock, B. A., Salters, J., & Willging, R. (2013). On the Use of Ensemble Models for Credit Evaluation. *The Credit and Financial Management Review*, *1*(1), 1–14.

Floares, A., & Birlutiu, A. (2012). Decision tree models for developing molecular classifiers for cancer diagnosis. *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN.2012.6252781

Fraz, M. M., Remagnino, P., Hoppe, a., Uyyanonvara, B., Rudnicka, a., Owen, C. G., & Barman, S. a. (2012). Retinal vessel segmentation using ensemble classifier of bagged decision trees. *IET Conference on Image Processing (IPR 2012)*, B9–B9. https://doi.org/10.1049/cp.2012.0458

Gangadhara, K., Anusha, S., & Dubbaka, R. (2010). *C OMPARING C OMPOUND D IVERSITY A ND O RDINARY D IVERSITY M EASURES U SING*. University of Boras.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol*, *22*, 245–252.

Guidi, G., Pettenati, M. C., Miniati, R., & Iadanza, E. (2013). Random Forest for automatic assessment of heart failure severity in a telemonitoring scenario. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2013*, 3230–3233.

https://doi.org/10.1109/EMBC.2013.6610229

Ha, S. H., & Joo, S. H. (2010). A Hybrid Data Mining Method for the Medical Classification of Chest Pain. *World Academy of Science, Engineering and Technology*, *4*(1), 499–504.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009a). The WEKA Data Mining Software: An Update; SIGKDD Explorations. *Machine Learning Group at the University of Waikato*, *11*(1), 1–6.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009b). The WEKA Data Mining Software. *SIGKDD Explorations Paper*, *11*(1).

Hasan, M. R., Bakar, N. A. A., Siraj, F., Sainin, M. S., & Hasan, M. S. (2015). Single Decision Tree Classifiers ' Accuracy on Medical Data. *Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015*, (188), 671–676.

Hasan, M. R., Gholamhosseini, H., & Sarkar, N. I. (2017). A new ensemble model for multivariate medical data. *International Telecommunication Networks and Applications Conference*, In press. Melbourne, Australia.

Hasan, M. R., Golamhosseini, H., Sarkar, N. I., & Safiuzzaman, S. M. (2017). Intrinsic motivated cervical cancer screening intervention framework. *Humanitarian Technology Conference*, 506–509.

Hasan, M. R., Siraj, F., & Sainin, M. S. (2015a). Improving ensemble decision tree performance using Adaboost and Bagging. *AIP Conference Proceedings*, *1691*(September), 1–7. https://doi.org/10.1063/1.4937027

Hasan, M. R., Siraj, F., & Sainin, S. (2015b). Improving ensemble decision tree performance using Adaboost and Bagging. *Innovation and Analytics Conference and Exhibition (IACE 2015): Proceedings of the 2nd Innovation and Analytics Conference & Exhibition*, *1691*(1).

Hassan, M. M., Atiya, A. F., El-Gayar, N., & El-Fouly, R. (2007). Regression in the presence missing data using ensemble methods. *IEEE International Conference on Neural Networks - Conference Proceedings*, 1261–1265. https://doi.org/10.1109/IJCNN.2007.4371139

Holst, K., & Manga, A. (2013). SAP Predictive Analysis – Real Life Use Case Predicting Who Will Buy Additional Insurance. *SAP COMMUNITY NETWORK*, *1*(1), 1–104.

Huanhuan Chen, Yao, X., & Tino, P. (2011). Ensemble Learning through Diversity Management: Theory, Algorithms, and Applications. *The 2011 International Joint Conference on Neural Networks*, *1*(1), 1–6.

Husmeier, D., Dybowski, R., & Roberts, S. (2004). *Probabilistic modelling in bioinformatics and medical informatics* (D. Husmeier, R. Dybowski, & S. Roberts, Eds.). Springer.

Ishwaran, H., & Rao, J. S. (2011). Decision Ttree: IintrodDuction. *Classification and Regression Trees*, *1*, 323–328.

Jordanov, I., Petrov, N., & Petrozziello, A. (2018). Classifiers Accuracy Improvement Based on Missing Data Imputation. *Journal of Artificial Intelligence and Soft Computing Research*, *8*(1), 31–48. https://doi.org/10.1515/jaiscr-2018-0002

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kelarev, a. V., Stranieri, A., Yearwood, J. L., & Jelinek, H. F. (2012). Empirical Study of Decision Trees and Ensemble Classifiers for Monitoring of Diabetes Patients in Pervasive Healthcare. *2012 15th International Conference on Network-Based Information Systems*, *1*(1), 441–446. https://doi.org/10.1109/NBiS.2012.20

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013 Fourth International Conference on Computing,*

Communications and Networking Technologies (ICCCNT), *1*(1), 1–7. https://doi.org/10.1109/ICCCNT.2013.6726842

Khan, S. S., Ahmad, A., & Mihailidis, A. (2018). *Bootstrapping and Multiple Imputation Ensemble Approaches for Missing Data*. (Mi), 1–17.

Ko, A. H.-R., & Sabourin, R. (2013). Single Classifier-based Multiple Classification Scheme for weak classifiers: An experimental comparison. *Expert Systems with Applications*, *40*(9), 3606–3622. https://doi.org/10.1016/j.eswa.2012.12.067

Krawczyk, B., & Schaefer, G. (2012). Dealing with the Difficult Learning Situation. *Neural Network Applications in Electrical Engineering (NEUREL)*, *1*(1), 12–15.

Kulkarni, S., & Kelkar, V. (2014). *Classification of Multispectral Satellite Images Using Ensemble Techniques of Bagging , Boosting and Ada-Boost*. 253–258.

Kusum, M., & Rupali, M. (2013). A REVIEW ON VARIOUS CLASSIFICATION ALGORITHMS FOR AN INCREMENTAL SPAM FILTER. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, *2*(11), 325–331.

Lavanya, DRani, K. U. (2011). Performance Evaluation of Decision Tree Classifiers on Medical Datasets. *Nternational Journal of Computer Applications*, *26*(4), 2–5.

Lavanya, & Rani, K. U. (2012). ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA. *International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012*, *2*(1), 17–24.

Lee, A., Joynt, G. M., Ho, A. M. H., Keitz, S., McGinn, T., & Wyer, P. C. (2009). Tips for teachers of evidence-based medicine: making sense of decision analysis using a decision tree. *Journal of General Internal Medicine*, *24*(5), 642–648. https://doi.org/10.1007/s11606-009-0918-8

Lee, C. H., & Yoon, H.-J. (2017). Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, *36*(1), 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3

Li, K., Wang, Z., Ma, C., & Song, S. (2013). SNS Privacy Protection based on the ELM Integration and Semi-supervised Clustering. *Journal of Software*, *8*(1), 160–167. https://doi.org/10.4304/jsw.8.1.160-167

Liang, G. (2014). *Ensemble Predictions : Empirical Studies on Learners ' Performance and Guohua Liang*. University of Technology, Sydney.

López-Vallverdú, J. A., Riaño, D., & Bohada, J. a. (2012). Improving medical decision trees by combining relevant health-care criteria. *Expert Systems with Applications*, *39*(14), 11782–11791. https://doi.org/10.1016/j.eswa.2012.04.073

Lucas, P., van der Gaag, L., & Abu-Hanna. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, *30*(3), 201–214.

Ludwig, S. A., Picek, S., & Jakobovic, D. (2018). *Operations Research Applications in Health Care Management*. https://doi.org/10.1007/978-3-319-65455-3

Machine Learning Group at the University of Waikato. (2013). Weka 3: Data Mining Software in Java.

Mahila, S. P., & Pradesh, A. (2012). Ensemble Decision Tree Classifier For Breast Cancer Data. *International Journal of Information Technology Convergence and Services*, *2*(1), 17–24.

Małgorzata, Ć.-J. (2012). Boosting, Bagging and Fixed Fusion Methods Performance for Aiding Diagnosis. *Biocybernetics and Biomedical Engineering*, *32*(2), 17–31. https://doi.org/10.1016/S0208-5216(12)70034-7

Markov, Z., & Russell, I. (2006). An introduction to the WEKA data mining system. *ACM SIGCSE Bulletin*, *38*(3), 367–368.

https://doi.org/10.1145/1140123.1140127

Melville, & Mooney. (2004). Creating Diversity in Ensembles Using Artificial Data. *Special Issue on Diversity in Multiclassifier Systems*, *1*(1), 1–15.

Melville, P. (2003). *Creating Diverse Ensemble Classifiers* (Vol. 2003). The University of Texas at Austin.

Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2012). A comparative study of Reduced Error Pruning method in decision tree algorithms. *2012 IEEE International Conference on Control System, Computing and Engineering*, *1*(1), 392–397. https://doi.org/10.1109/ICCSCE.2012.6487177

Moon, H., Ahn, H., Kodell, R. L., Baek, S., Lin, C.-J., & Chen, J. J. (2007). Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine*, *41*(3), 197–207. https://doi.org/10.1016/j.artmed.2007.07.003

Nai-Arun, N., & Sittidech, P. (2014). Ensemble Learning Model for Diabetes Classification. *Advanced Materials Research*, *931–932*(1), 1427–1431. https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427

Nanni, L., Lumini, A., & Brahnam, S. (2012). A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine*, *55*(1), 37–50. https://doi.org/10.1016/j.artmed.2011.11.006

Objectives, C. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. In Mehmed Kantardzic (Ed.), *Data Mining: Concepts, Models, Methods, and Algorithms* (Second, pp. 235–248). John Wiley & Sons, Inc.

Oh, S., Lee, M. S., & Zhang, B. (2011). *Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification*. *8*(2), 316–325.

Ojha, V. K., Jackowski, K., Abraham, A., & Snášel, V. (2015). Dimensionality reduction, and function approximation of poly (lactic-co-glycolic acid) micro-and nanoparticle dissolution rate. *International Journal of Nanomedicine*, *10*, 1119.

Olafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. *European Journal of Operational Research*, *187*, 1429–1448.

Parvin, MirnabiBaboli, M., & Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence*, *37*(1), 34–42. https://doi.org/10.1016/j.engappai.2014.08.005

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of Medical Systems*, *26*(5), 445–463.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees. *IEEE Trans. Syst Man Cybern. Part C Appl. Rev*, *35*, 476–487.

Rokach, L., & Maimon, O. (2008). Data Mining With Decision Trees: Theory and Applications. *Series in Machine Perception and Artificial Intelligence, World Scientific*, *1*, 78–88.

Rudin, C. (2007). *Ensembles—Combining Multiple Learners For Better Accuracy*. New York University.

Saghir, H., & Megherbi, D. B. (2013). A random-forest-based efficient comparative machine learning predictive DNA-codon metagenomics binning technique for WMD events & applications. *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, 171–177. https://doi.org/10.1109/THS.2013.6698995

Scarinci, I. C., Garcia, F. a R., Kobetz, E., Partridge, E. E., Brandt, H. M., Bell, M. C., … Philip, E. (2011). Cervical Cancer Prevention: New Tools and Old Barriers. *Cancer*, *116*(11), 2531–2542. https://doi.org/10.1002/cncr.25065.Cervical

Silwattananusarn, T., & Kulthidatuamsuk, A. P. (2012). Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to. *Nternational Journal of Data Mining & Knowledge Management Process (IJDKP)*, *2*(5), 13–24.

Siraj, F., & Abdoulha, M. A. (2007). *Mining Enrolment Data Using Predictive and Descriptive Approaches* (A. Karahoca, Ed.). Intech Open.

Siraj, F., Omer, E. A. O. A., & Hasan, R. (2012). Data Mining and Neural Networks : The Impact of Data Representation. In Adem Karahoca (Ed.), *Advances in Data Mining Knowledge Discovery and Applications* (1st ed., pp. 1–20). https://doi.org/10.5772/3349

Southern Cross Medical Library. (2013). Cervical cancer - causes, symptoms, treatment, prevention. Retrieved July 10, 2017, from https://www.southerncross.co.nz/group/medical-library/cervical-cancer-causes-symptoms-treatment-prevention

Stefanowski, J. (2008). *Multiple classifiers*. Catania-Troina.

Stefanowski, J., & Pachocki, M. (2013). Comparing Performance of Committee Based Approaches to Active Learning. *Recent Advances in Intelligent Information Systems*, *1*(1998), 457–470.

Sun, Y., Yao, J., Nowak, N. J., & Goodison, S. (2014). Cancer progression modeling using static sample data. *Genome Biology*, *15*(8), 440. https://doi.org/10.1186/s13059-014-0440-0

Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, *2*(3 Suppl), S75-83.

Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, *65*(1), 247–271. https://doi.org/10.1007/s10994-006-9449-2

Tay, W., Chui, C., Ong, S., & Ng, A. C. (2013). Expert Systems with

Applications Ensemble-based regression analysis of multimodal medical data for osteopenia diagnosis. *Expert Systems With Applications*, *40*(2), 811–819. https://doi.org/10.1016/j.eswa.2012.08.031

Tripoliti, E. E., Fotiadis, D. I., & Manis, G. (2012). Automated diagnosis of diseases based on classification: dynamic determination of the number of trees in random forests algorithm. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *16*(4), 615–622. https://doi.org/10.1109/TITB.2011.2175938

US department of veterans affairs. (2018a). Do I need to be tested for hepatitis B before getting the vaccination? - Hepatitis B for Patients. Retrieved February 2, 2018, from Virua Hepatitis website: https://www.hepatitis.va.gov/patient/hbv/vaccine-before-getting.asp

US department of veterans affairs. (2018b). What should you do if exposed to the hepatitis B virus? - Hepatitis B for Patients. Retrieved February 1, 2015, from Viral Hepatitis website: https://www.hepatitis.va.gov/patient/hbv/vaccine-exposure.asp

Velikova, M., de Carvalho Ferreira, N., & Lucas, P. (2007). Bayesian network decomposition for modeling breast cancer detection. *Artificial Intelligence in Medicine, AIME 2007*, *4594*, 346–350.

Vellido, A. (2014). *Intelligent Data Analysis and Data Mining or Data Analysis and Knowledge Discovery*.

Wang, H., Yin, J., Pei, J., Yu, P. S., & Yu, J. X. (2006). Suppressing model overfitting in mining concept-drifting data streams. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06*, 736. https://doi.org/10.1145/1150402.1150496

Wang, H., & Yu, P. S. (2002). Mining Concept-Drifting Data Streams using Ensemble Classifiers. *Proceedings of the 12th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining - KDD '02, 1–10.

Wang, M., Gao, K., Wang, L., & Miu, X. (2012). A Novel Hyperspectral Classification Method Based on C5.0 Decision Tree of Multiple Combined Classifiers. *2012 Fourth International Conference on Computational and Information Sciences*, *0*(1), 373–376. https://doi.org/10.1109/ICCIS.2012.33

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques* (Third Edit; I. H. Witten, E. Frank, & M. A. Hall, Eds.). Burlington, USA: Elsilvier.

Wu, G., Shen, D., & Sabuncu, M. R. (2016). *Machine Learning and Medical Imaging*. Elsevier Inc.

Xiaochen, D., & Xue, H. (2011). Multi-Decision-Tree Classifier in Master Data Management System. *Beijing Natural Science Foundation*, *1*(1), 1–4.

Ye, R., & Suganthan, P. N. (2013). Empirical Comparison of Bagging-based Ensemble Classifiers. *International Journal of Information Technology Convergence and Services*, *1*(1), 917–924.

Yeh, D., Cheng, C., & Chen, Y. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, *37*(7), 8970–8977.

Yin, X.-C., Huang, K., Hao, H.-W., Iqbal, K., & Wang, Z.-B. (2014). A novel classifier ensemble method with sparsity and diversity. *Neurocomputing*, *134*, 214–221. https://doi.org/10.1016/j.neucom.2013.07.054

Zhang, C., & Ma, Y. (2012). *Ensemble Machine Learning: Methods and Applications* (C. Zhang & Y. Ma, Eds.). New York: Springer New York.

Zhao, Y., & Zhang, Y. (2007). Comparison of decision tree methods for

finding active objects. *Accepted for Publication in Advances of Space Research*, *1*(1), 1–10.