MACHINE LEARNING TECHNIQUE IN APPLICATION AND COMPARISON IN PEDIATRIC FRACTURE HEALING TIME

KEDIJA

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

University of Malaya

MACHINE LEARNING TECHNIQUE IN APPLICATION AND COMPARISON IN PEDIATRIC FRACTURE HEALING TIME

KEDIJA

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTES OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

University of Malaya

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Kedija

Matric No: SGR150082

Name of Degree: Master of Science

Title of Project Research Thesis ("this Work"): MACHINE LEARNING TECHNIQUE IN APPLICATION AND COMPARISON IN PEDIATRIC FRACTURE HEALING TIME

Field of Study: Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

University of Malaya

MACHINE LEARNING TECHNIQUE IN APPLICATION AND COMPARISON IN PEDIATRIC FRACTURE HEALING TIME

ABSTRACT

Machine learning methods have been used in this study to analyze and predict the required healing time among pediatric orthopedic patients particularly for lower limb fracture. Random forest (RF), Self-Organizing Feature map (SOM), decision tree (DT), support vector machine (SVM) and Artificial Neural Network (ANN) were used to analyze the data obtained from the pediatric orthopedic unit in University Malaya Medical Centre. Radiographs of long bones of lower limb fractures involving the femur, tibia and fibula from children under twelve years, with ages recorded from the date and time of initial injury. Inputs assessment included the following features: type of fracture, angulation of the fracture, contact area percentage of the fracture, age, gender, bone type, type of fracture, and number of bone involved; all of which were determined from the radiographic images. Leave one out method was used to enhance machine learning models as dataset that was available for this project were limited in numbers. RF is used to select variables affecting bone healing time. To our best knowledge there is no study reported using machine learning method to predict paediatric orthopaedics fracture healing time. Findings from this study identified contact area percentage of fracture, type of fracture, number of fractured bone and age as important variables in explaining the fracture healing pattern. SVM model for predicting fracture healing time outperformed ANN and RF models. Based on the outcomes obtained from the models it is concluded that RF, Decision Tree, SVM, ANN and SOM techniques can be used to assist in analysis of the healing time efficiently.

Keywords: ANN, SVM, SOM, pediatric orthopedic

University of Malaya

TEKNIK PEMBELAJARAN MACHINE DALAM PERMOHONAN DAN PERBANDINGAN DALAM PELAJAR PEDIATRIK PELAJAR MASA

ABSTRAK

Kaedah pembelajaran mesin telah digunakan dalam kajian ini untuk menganalisis dan meramalkan masa penyembuhan yang diperlukan di kalangan pesakit ortopedik pediatrik terutamanya untuk patah kaki bawah. Perhitungan rawak hutan (RF), peta ciri sendiri (SOM), keputusan pokok, mesin vektor sokongan (SVM) dan Rangkaian Neural Buatan (ANN) digunakan untuk menganalisis data yang diperolehi dari unit ortopedik pediatrik di Pusat Perubatan Universiti Malaya. Radiografi tulang panjang yang melibatkan femur, tibia dan fibula daripada kanak-kanak di bawah dua belas tahun, dengan umur yang direkodkan dari tarikh dan masa kecederaan awal. Penilaian input termasuk ciri-ciri berikut: jenis fraktur, angsi patah tulang, peratusan kawasan sentuhan fraktur, umur, jantina, jenis tulang, jenis patah tulang, dan jumlah tulang yang terlibat; semuanya telah ditentukan dari imej radiografi. Meninggalkan satu kaedah digunakan untuk meningkatkan model pembelajaran mesin memandangkan dataset yang tersedia untuk projek ini adalah terhad. RF digunakan untuk memilih pembolehubah yang mempengaruhi masa penyembuhan tulang. Model-model yang digunakan untuk ini tidak dilaporkan secara meluas dalam bidang ortopedik pediatrik. Keputusan dari beberapa aplikasi menunjukkan bahawa peratusan kawasan hubungan patah, jenis fraktur , jumlah tulang ang patah dan usia telah dikenalpasti sebagai pembolehubah penting dalam menjelaskan corak penyembuhan patah tulang. Model SVM memberikan keputusan yang lebih baik berbanding dengan ANN dan RF.Berdasarkan hasil yang diperoleh dari model, disimpulkan bahawa teknik RF, Decision Tree, SVM, ANN dan SOM dapat digunakan untuk membantu dalam menganalisis masa penyembuhan secara efisien.

Kata kunci: ANN, SVM, SOM, ortopedik pediatrik

University of Malaya

ACKNOWLEDGEMENTS

First and foremost I would like to thank God for helping me in the process of putting this book together I realized how true this gift of writing is for me. I could never have done this without the faith I have in you, the Almighty.

To my mother: Noora M. Berhan: when it comes to you I became speechless! I can barely find the words to express all the wisdom, love and support you've given me. You are my number 1fan and for that I am eternally grateful. I always try my best to make you be proud of me. I Love you unconditionally Mother.

Father: Seid Yesuf: thank you a lot for the support and dua’s you always make for me.

I would like to thank my supervisor Dr. Sorayya Malek for keeping her door always open whenever I ran into a trouble spot or had a question about my research or writing. She consistently steered me in the right the direction whenever he thought I needed it. She patiently guided me throughout this research and she offers her help and advice unlimitedly.

I would also like to thank Dr. Roshan Gunalan who contributed to this research.

Without his passionate participation this could not have been successfully conducted.

At the end, I must express my very profound gratitude to my beloved family specially my sisters; Huda: I reach this level because of you, I will never forgot the efforts you made to make accomplish my dreams, so here I would like to show you my appreciation. Thank you dear. Iman: thanks for your love. And friends (especially Maram) for providing me with unfailing support and continuous encouragement of many moments of crisis throughout my years of study .This accomplishment would not have been possible without them.

University of Malaya

TABLE OF CONTENTS

University of Malaya

University of Malaya

LIST OF FIGURES

University of Malaya

LIST OF TABLES

University of Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	:	Artificial Neural Network
CV	:	Cross-Validation
DT	:	Decision Tree
γ	:	Gamma
LMO-CV	:	Leave-More-Out CV
LOO-CV	:	Leave-One-Out CV
MSE	:	Mean Square Error
NN	:	Neural network
MTYR	:	Number of selected variable
NTREE	:	Number of trees
OOB	:	Out-Of-Bag
RBF	:	Radial Basis Function
RF	:	Random Forest
RSS	:	Residual Sum of Squares
RMSE	:	Root Mean Square Error
SOM	:	Self-Organizing Feature Map
σ2	:	Sigma2
SVM	:	Support Vector Machine
SVR	:	Support Vector Regression

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Introduction

A fracture can be considered as a break in the continuity of the bone affecting the bone's cortex. It came in two ways; an incomplete or complete break in the bone's continuity (University of Rochester Medical Center, 2015). Children's fractures (0 to 12 years old) differ in features comparing to adults' fractures. Skeletal trauma accounts for 15% of all injuries in children (Staheli, 2008). There are several types of fracture such as a transverse fracture which occurs when the fracture goes through at right angles to the long bone's shaft. Spiral type of fractures goes through an angle oblique to the long bone's shaft of the long bone. The bone structure of Children has a thick periosteal layer, subsequent in an incomplete type of fracture which is called buckle/ torus fracture.

Children's lower limb fractures take half the time to fully recover compared to the adults corresponding fracture (Ogden, 2000). In cases of paediatric, it is important to evaluate skeletal trauma as it can additionally signal a non-unintentional injury or abnormal restoration, in which it may suggest an underlying medical condition that affect the time required for bone fracture healing. While rates have been published for a normal adults' bone restoration manner, not much is known about the rates of healing among the paediatric cases. Paediatric bone body structure suggests that more youthful individuals heal at a quicker rate in comparison to adults (Ogden, 2000).

Lower limb long bones's can be divided into three parts; femur, tibia and fibula. The femur is the body's longest bone. Its fundamental duty is to carry (power) physical action starting the hip joint to the tibia. The tibia is considered as the second largest bone in the body. It broadens at the proximal and distal boundies, articulating at both the knee

University of Malaya

and ankle joints. The fibula and tibia together forms the bones of the leg. Long bone fractures are defined with reference to the direction of the line's fracture in relation to the shaft of the bone. Limited literature are reported on the assessment on classification of paediatric fracture recovery time using of radiographic/ x-ray fracture and statistical approach to determine the healing rates. Fracture healing time correlated with the events leading to the injury, may help in injuries that are recovering differently or might point to non-accidental injury (Tseng et al., 2013). Predicting healing time is useful and should be used as a tool in the treatment process for general practitioners and medical officers and in the follow-up period.

Several machine learning techniques have been applied in clinical settings to predict disease classify big quantity of data into a valuable format. Machine learning methods have shown higher accuracy for diagnosis than classical statistical methods.

Machine learning classifiers uses medical data of each patient and predict the existence of diseases based on hidden patterns found in the data. The most commonly used machine learning methods for analysing complex medical data are Support vector machines (SVM), Random Forests (RF), Decision Tree (DT) and Artificial Neural Networks (ANN). SVM is based on mapping data to a higher dimensional space through a kernel function, and choosing the maximum-margin hyper-plane that separates training data to improve accuracy by the optimization of space separation. RF grows many classification trees built from a random subset of predictors and bootstrap samples. RF can handle high dimensional data in training faster compared to other methods. ANN comprises several layers and connections which mimic biological neural networks to construct complex classifiers. ANN has been applied to many problems of non-linear pattern classification. DT consists of tests or attribute nodes linked to two or more subtrees and leafs or decision nodes labelled with a class that represents the decision (Mantzaris et al., 2008). SVM, RF,

University of Malaya

ANN and DT are popular option in in medicine and Bioinformatics for task that involves selecting informative variables or genes and predicting diseases more accurately.

Machine learning methods such as ANN and RF have been applied in orthopaedic field in our previous study to predict fracture healing time (Malek et al., 2016). Zhao et al., (2003) have used several machine learning methods such as SVM, RF ANN and logistic regression (LR) on osteoporosis risk assessment for postmenopausal women for the measurement of bone mineral density. In this study SVM have been applied for screening femoral neck in postmenopausal women and compared the result to a conventional clinical decision tool, osteoporosis self-assessment tool (OST). Sapthagirivasan and Anburajan, (2013) applied SVM kernel classifier-based computer-aided diagnosis (CAD) system for osteoporotic risk detection with 90% accuracy rate. Umadevi and Geethalakshmi, (2012) reported automatic detection of fractures in in long bones tibia using Back Propagation Neural Network, K-Nearest Neighbour, Support Vector Machine. SVM is also applied for fracture risk prediction (Burges, 1998; Cristianini & Shawe-Taylor, 2000) and hip fracture prediction (Jiang et al., 2014).

Tseng et al. (2013) found ANN outperforms conditional logistic regression in an age-and sex-matched case control study about morbidity and mortality among patients who have hip bone fractures. They examined the factors that may influence the hip risk and evaluate the risk by using a logistic regression model (CLR) and ensemble artificial neural network (ANN). They made a comparison between those two models of machine learning to assess the risk and the factors that are related to hip fractures.

Shaikh et al., (2014) developed an expert system in detecting and diagnosing osteoporosis using ANN. Mantzaris et al., (2008) successfully predicted the presence

University of Malaya

of osteoporosis using two different ANN techniques: Multi-Layer Perceptron (MLP) and Probabilistic Neural Network (PNN).

Previous work done has reported estimation of paediatric fracture healing time using supervised and unsupervised ANNs. Multilayer perceptron (MLP) using back-propagation for supervised ANN and Kohonen self-organizing feature map (SOM) was used for the unsupervised learning (Malek et al., 2016). SOM usage also has been stated in investigation of osteoporosis dataset (Kilmer et al., 1997). SOM method has been used to classify the dataset for the problem of osteoporosis classification of high and low osteoporosis risk. SOM is an excellent tool in the visualization of high dimensional data (Kohonen, 1988). SOM decreases the dimensions of data of a high level of complexity and plots the data similarities through clustering technique (Hollmén, 1996).

Multilayer perceptron (MLP), a supervised ANN learning method, is the most frequently used machine learning technique. However, this technique provides little insight to the significance of variables against the predictor. Transparency is very important in areas, such as medical decision support. This can be achieved by using classification and regression trees (Tseng et al., 2013). DT has been already successfully used in medicine (Zorman et al., 2001). In orthopaedic field it has been used for decision analysis of Operative Versus Nonoperative Treatment of Jones Fractures. In paediatric orthopaedic, DT have been used to determine foot disorder groups and biomechanical parameters related to symptom on the basis of the paediatric clinical data by developing a prediction model of the decision tree (Mantzaris et al., 2008). RF is a machine learning method that is a specific instance of bagging. RF method is classification and regression method based on the

University of Malaya

aggregation of large number of decision trees built using several bootstrap samples was developed by Breiman (2001).

The two by-products of RF method are out-of-bag (OOB) estimates of generalization error and variable importance measures. RF method has been demonstrated to have better accuracy compared to other supervised learning methods such as MLP and SVM. RF has been applied in various applications in computational biology and medicine where the relationship between response and predictors is complex and the predictors are strongly correlated.

However, application of SVM, ANN, RF and DT in orthopaedics, especially in paediatric orthopaedic field has yet to be reported which is the aim of this study. Variables that had been chosen for this study were selected due to its high importance with the predictor which is in this study the healing weeks as the variable importance's quantification is a critical issue for understanding data in applied problems.

1.2 Objectives

- To identify variables that affect the time required for lower bone healing fracture using Machine learning methods.
- To developed machine learning methods to predict lower limb healing time.
- To compare different machine learning methods in predicting lower limb healing time.

1.3 Problem statement:

Machine learning techniques had been used widely in medical field. Especially in hip fractures, however it has not reported that machine learning had been used in pediatric orthopedic field. This study aims to provide assistance to orthopedics in the task of predicting the required healing time for the children. Therefore, a systematic research approach is required to fill this knowledge gap between machine learning and pediatric orthopedic field. This study addresses the implementation of machine learning in the field of medicine.

1.4 Research scope

Understanding the importance of machine learning in the medical field is crucial to implement it on many areas that need improvements in terms of analysis, risk assessment and expected healing time (for bone fractures). Many research had been done in medicine using machine learning techniques to analyze or predict the outcome.

This study is expected to predict the fully recovery time for children with lower limb bone fracture. It attempts to investigate application of machine learning application in estimating pediatric fracture of lower limb bones healing time.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

2.1 Paediatric Orthopaedic

Any break or discontinuity on the shaft of a bone is considered to be a fracture. It can occur in both partial and complete fracture on the bone. Children's (paediatrics) fracture has different features compared to adults' bone fractures (Staheli, 2008). There are several types of bone fractures such as spiral, transverse and torus.

Transverse fracture is defined as a fracture that passes at right angles to the long bone's shaft. While torus which is known as (Buckle) as well is defined as an incomplete fracture that occur as a result of a thick periosteal layer in the children's bone. The time required for the bone (Lower Limb) to be fully recovered in children is most likely half of the time of adults corresponding fracture (Ogden, 2000).

In paediatric cases, it is essential to evaluate whether the skeletal trauma signals a non-accidental injury or abnormal healing, as this indicates an underlying medical condition affecting bone healing. While rates have been published for a normal bone healing process in adults, very little is known about healing rates in the paediatric population. Paediatric bone physiology indicates that younger individuals heal at a faster rate as compared to adults (Ogden, 2000).

The long bones of the lower limb are classified into three parts:

- Femur: the longest bone in human body, so it can transmit forces from the hip to tibia
- Tibia: second largest bone, expand at the proximal and distal ends
- Fibula: together with tibia it forms the leg of the long bones.

University of Malaya

2.1.1 Lower Limb Anatomy

Lower limb bones can be divided into four sections; the femur, tibia, fibula and foot. The femur is the only bone in the thigh. It is classed as a long bone, and is the longest bone in the human body. The function of the femur is to transmit forces from the tibia to the hip joint. It serves as the place of origin and attachment of many muscles and ligaments. The tibia is the main bone of the leg, or commonly known as the shin. It expands at the proximal and distal ends, articulating at the knee and ankle joints respectively. It is the second largest bone in the body; this is due to its function as a weight bearing structure. The bones of the leg are made up of fibula and tibia.

The fibula is lateral to tibia, and is much thinner. The main function of fibula is to act as an attachment for muscles. The fibular shaft has three surfaces; anterior, lateral and posterior. Distally, the lateral surface continues inferiorly, and is called the lateral malleolus. The lateral malleolus is more prominent than the medial malleolus, and can be palpated at the ankle on the lateral side of the leg. Figure 2.1 below shows a complete lower limb anatomy of bone.

University of Malaya

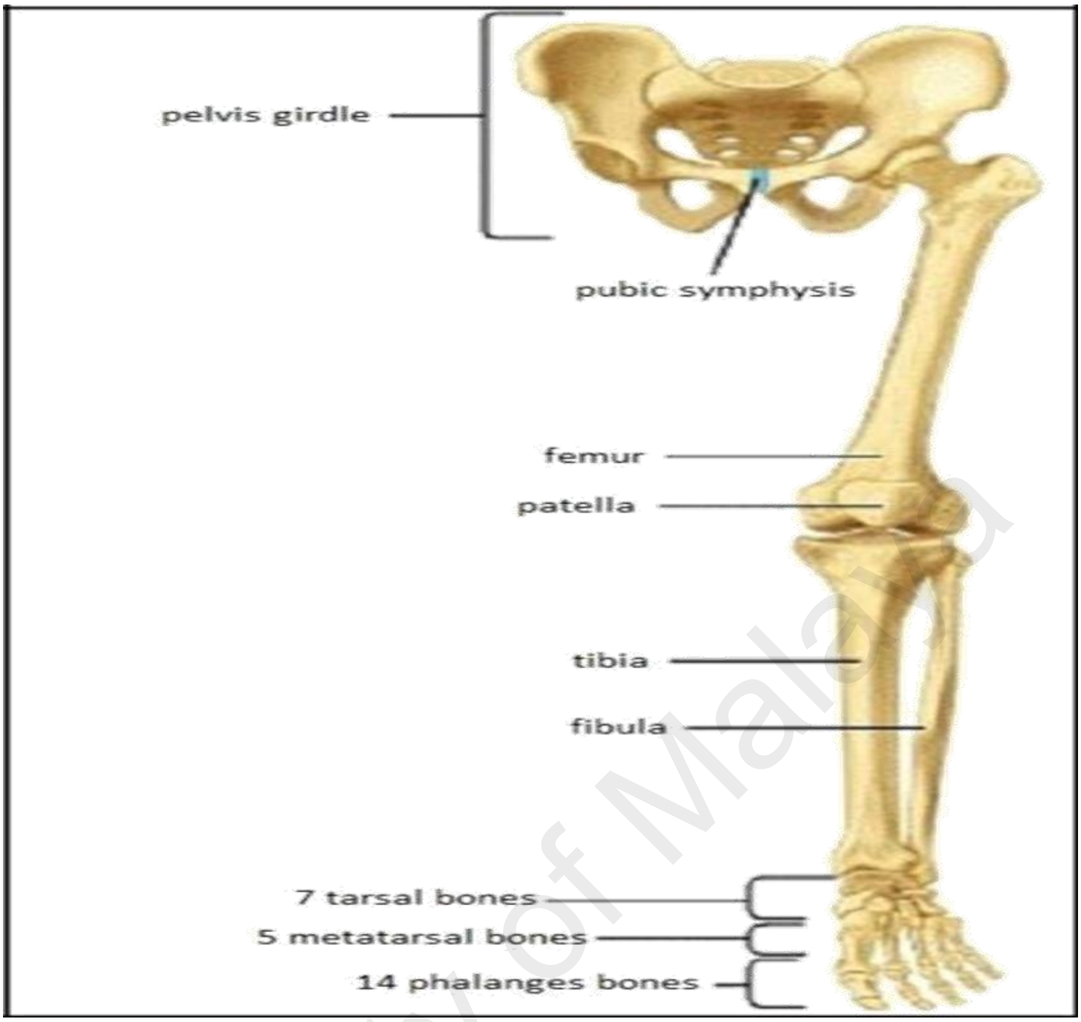

Figure 2.1: Anatomy of lower limb

2.1.2 Fracture Anatomy

Region of a long bone can be distinguished into 3 distinct zone: epiphysis, metaphysis, and diaphysis as shown in Figure 2.2 below. In development, the epiphysis and metaphysis are separated by a fourth zone, known as the epiphyseal plate, or physis. This segment of the bone is cartilaginous and is the region from which the bone grows longitudinally. By adulthood, all epiphyseal plates have closed down, and a bony scar is all that remains of this important structure. Long bones include the femur, tibia, fibula, humerus, radius, ulna, metacarpals, metatarsals, and phalanges.

University of Malaya

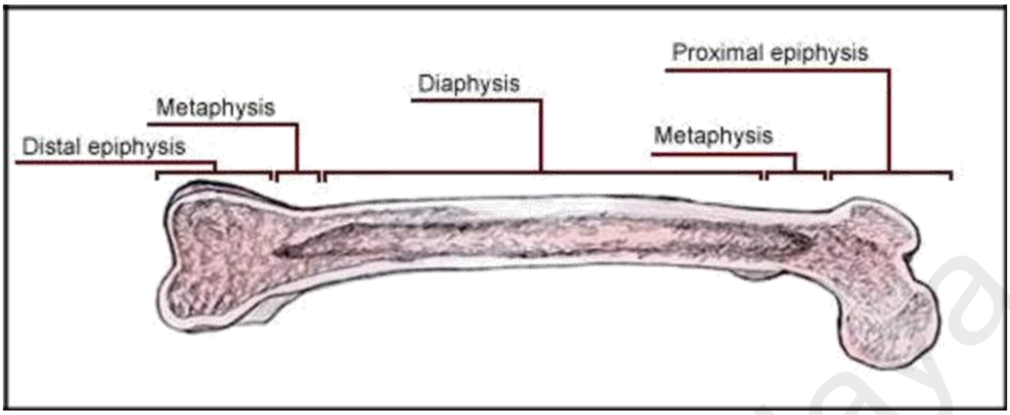

Figure 2.2: Region of a long bone

University of Malaya

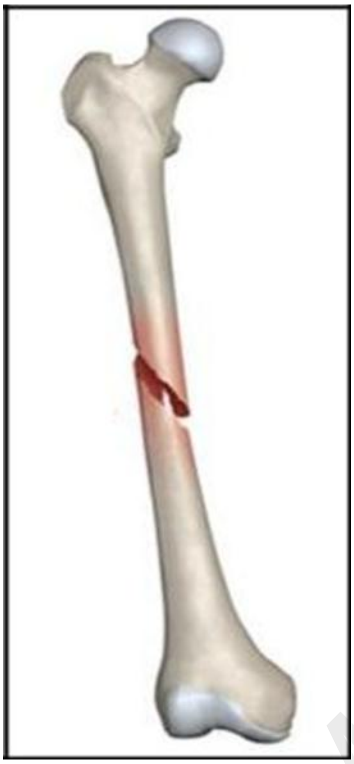

Figure 2.3 : Spiral fracture

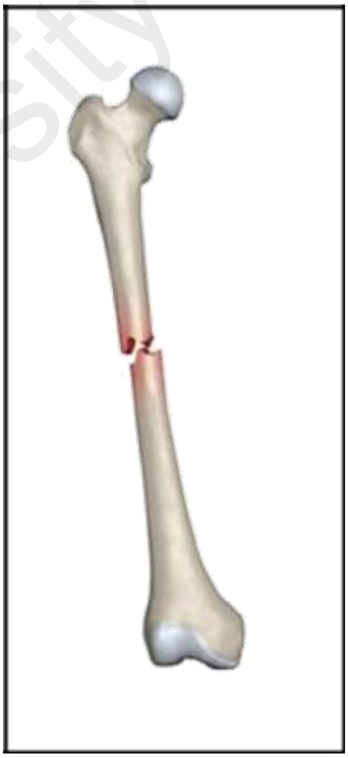

Figure 2.4: Transverse fracture

University of Malaya

Figure 2.5: Torus fracture

Long bone fractures are described with reference to the direction of the fracture line in relation to the shaft of the bone. Above Figures shows several type of fracture from radiograph samples. Figure 2.3 is a spiral fracture on tibia bone. The fracture line spirals along the shaft of the long bone as a result from twisting injury. Figure 2.4 indicates transverse fracture occurred at tibia as the fracture passes at right angles to the shaft of the long bone. Figure 2.5 shows an example of oblique type of fracture occurred at metatarsal region due to the fracture passes at an angle oblique to the shaft of the long bone.

University of Malaya

2.1.3 Healing Rates

There are few types of unity for bone fracture determination. Mal-union can be defined as united fractured in deformed structure due to tilted, twisted or shortened. A varus deformity in the leg usually leads to osteoarthritis of the knee or ankle ((Simonis et al., 2003). Delayed union consume longer time to heal. Perkin's timetable provides a guide to elucidate the period taken for fresh fracture to recover as shown in Table 2.1.

The decision function in SVM depends on the inner product between two vectors rather than on input vectors alone. SVMs can be extended to non-linear problems by means of a kernel function K that satisfies the Mercer conditions (symmetric semi-definite positive function). The kernel induces an implicit non-linear function ϕ which maps the sample point's xi $\in$ X into a high dimensional (even infinite) feature space T where one constructs the optimal hyperplane that separates the mapped point's ϕ (xi). This is equivalent to a non-linear separating surface in X. SVMs kernel methods is constructed to use a kernel for a particular problem that could be applied directly to the data without the need for a feature extraction process. This is particularly important in problems where a lot of structure of the data is lost by the feature extraction process (Sackinger et al., 1992).

University of Malaya

Table 2.1: Perkins classification of fracture healing time (in weeks)

PERKIN'S CLASSIFICATION	Spiral		Transverse	
	Union	Consolidation	Union	Consolidation
Upper Limb	3	6	6	12
Lower Limb	6	12	12	24

The Table above shows the required healing time for the fracture to unite and be fully healed. Lower limb fracture of children usually takes place half of the time given in the figure 2.3 approximately 3 to 6 weeks for spiral and 6 to 12 weeks for transverse fracture. Non-union occurred at least nine months from the initial accident. There is no evidence on X-rays changes of union over the last three months. In certain cases some fresh fractures take 18 months to heal (Simonis et al., 2003).

This study examined fractures occur on children aged from 0 to 12. Children's bone fractures are called paediatric Orthopaedic. It has different characteristics comparing to the fractures of the adults. Factures of the lower limb divided generally into numerous types: a transverse, spiral and torus. Transverse fracture occurs as the fracture passes at right angles to the shaft of the long bone. Lower limb fracture in children usually takes half the time of the corresponding fracture in adults. Lower limb long bones can be divided into three sections; the femur, tibia and fibula. The femur is the longest bone in the body.

Few articles reported the evaluation of classification on paediatric fracture healing on the basis of radiographic fracture and statistical approach to determine healing rates. Correlating healing time with the chronologic history of injury, may aid in injuries that are healing abnormally or may indicate non-accidental injury. The system for predicting

University of Malaya

healing time required should serve as a tool in the process of treatment for general practitioners and medical officers and in the follow-up period.

2.2 Random Forest

In this study RF method have been used for fracture healing time prediction and variable selection. RF is used not only for prediction, but also to assess variable selection and importance. RF methodology is used in this study to construct a prediction rule for a supervised learning problem and to assess and rank variables based on their capability to predict the output response. Variable importance measures are automatically computed for each predictor in the RF algorithm to assess and rank the variables. RF variable importance measure can recognize predictors involved in interactions for example predictors which can predict the response only in association with one or several other predictor(s). After practical authentication, the resulting prediction rule can then be applied, for instance, in clinical practice (Díaz-Uriarte & De Andres, 2006). RF is a combination of tree predictors where each tree depends on the value of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). RF is an improvement over bagged trees by having a small tweak that decorrelates the trees (James et al., 2013b). Breiman proposed RF which adds an additional layer of randomness to bagging (Breiman, 2001). RF provides estimators of Bayes classifier, which is the mapping minimizing the classification error or regression function. In bagging, successive trees do not depend on earlier trees. There are independently constructed using a bootstrap sample of the data set. Bagging results in improved accuracy over prediction using a single tree and can provide estimates of generalization error of the combined ensemble of trees and its strength and correlation (Genuer et al., 2010)

University of Malaya

Bagging provides summary of the importance of each predictor using the RSS (for bagging regression trees) or the Gini index (for bagging classification trees) (James et al., 2013b)

The out of bag (OOB) sample is the set of observations not used for building the existing tree but it is used to estimate the prediction error and then to evaluate variable importance (Genuer et al., 2010). An estimate of the error rate can be obtained, based on the training data. At each bootstrap iteration, data that is not in the bootstrap sample is used for prediction. The error is calculated and it is named the OOB estimate of error rate.

RF is an ensemble method that builds many decision trees from boostrapping samples which are then clustered together by classification or regression method with additional randomness added (Breiman, 2001; Liaw & Wiener, 2002). At each node in RF, only a subset of predictor are randomly chosen from the full set of predictors, p, (Genuer et al., 2010) which is denoted by *mtry* and the best split is done by Gini index node of impurity. Gini index of impurity is a measure of the class label conveyance at each node and is calculated only among the subset of predictors. The value of Gini impurity are 0 and 1 where 0 indicates when all the predictors at the node are of the same class (Khalilia et al., 2011). The decision on selecting the best split is based on the lowest Gini impurity value among the predictors to reduce the error rate, at each nodes of the tree. The default value of *mtry*=$p^{1/2}$ is set for classification and *mtry*=p/3 for regression) . Pruning is not required in RF therefore the trees generated are maximal, low-bias and low correlation among the trees (Díaz-Uriarte & De Andres, 2006).

University of Malaya

RF performance is superior compared to performance over single tree classifiers such as CART, and yield generalization error rates that compare acceptably to other statistical and machine learning methods (Biau et al., 2008). RF are noted to be the best general-purpose classifiers present (Breiman, 2001).

2.2.1 Algorithms of Random forest

The algorithm of the RF for regression and classification are as followed (Breiman, 2001):

1) Each tree of RF is grown a bootsrap sample of the training set.

2) At each node, n number of variables are chosen randomly out of N predictors when growing a tree.

3) The value of n starts with $n=\sqrt{N}$ and then increase it it until the smallest error of the OOB is obtained. At each node, one variable with the best split is used from all value of n.

4) Test set error estimate is obtained from growing a tree from a boostrap data (Verikas et al., 2011) which then be used to estimate the variable importance which is a useful byproducts of RF. In RF for regression, the test error estimate is defined by the Root Mean Square Error (RMSE).

2.2.1.1 Variable Importance

RF by products are , out-of-bag estimates of generalization error (Bylander, 2002) and variable importance measures (Svetnik et al., 2003) Each tree in RF is grown from a bootstrapped sample, on average about one-third of the observations in the data set will

University of Malaya

not be used to grow the tree. These is considered as out-of-bag observation (OOB) for that tree (Archer & Kimes, 2008).

In the RF framework, score of importance of a given variable is the increasing in mean of the error of a tree (MSE for regression and misclassification rate for classification) in the forest when the observed values of this variable are randomly permuted in the OOB samples. Classification problems the score of importance is based on the average loss of entropy criterion, the Gini entropy used for growing classification trees. The Gini criterion is used to select the split with the lowest impurity at each node. For each tree in the forest, the predicted class for each observation is obtained. The class with maximum number of votes among the trees in the forest is the predicted class of an observation. Specifically, at each split the decrease in the Gini impurity in the forest forms split yields the Gini variable importance measure (Archer & Kimes, 2008).

For regression problems, making a distinction between various variance decomposition based indicators which are dispersion importance, level importance or theoretical importance quantifying explained variable or changes in the response for a given change of each regressor (Grömping, 2009).

The RF algorithm estimates the importance of a variable by looking at how much prediction error increase when OOB data is permuted while others are left unaffected Genuer et al. (2010). OOB sample is the set of observations which are not used for building the current tree It is used to estimate the prediction error and then to evaluate variable importance.

The OBB estimate for the generalization error is the error rate of the OBB classifier on the training set. Macready and Wolpert (1996) worked on regression type problems and proposed a number of methods for estimating the generalization error of OBB predictors. Tibshirani (1996) used out-of-bag estimates of variance to estimate generalization error for arbitrary classifiers. Breiman (1996) proved that the OBB estimate is as accurate as using a test set of the same size as the training set. Hence, using the OBB error estimate removes the need for separate test set. In each bootstrap training set, about one-third of the instances set aside as OBB set. The error rate decreases as the number of combinations increases and OBB estimates are unbiased compared to in cross-validation. Strength and correlation can also be estimated using OBB methods which are helpful in understanding and improving accuracy rate (Breiman, 2001). Furthermore, OBB provides reasonable estimation compared to test set error and is it is default output of RF procedure (Genuer et al., 2010).

2.2.1.2 Variable selection

Rakotomamonjy (2003) introduced methods for variable selection using SVM with descending elimination of variables. Ye et al., (2011) discussed in “Efficient variable selection in support vector machines via the alternating direction method of multipliers” an alternative way of eliminating less important variables that it won’t affect the dataset during the analysis stage.

Díaz-Uriarte and De Andres (2006) proposed a strategy based on recursive elimination of variables. This is done by computing RF variable importance and at each step, 20% of the variables having the smallest importance are eliminated and a new forest is built with the remaining variables. The set of variables leading to the smallest OOB error rate are selected. The proportion of variables to eliminate is an arbitrary parameter of their method and does not depend on the data choose an ascendant strategy based on a sequential introduction of variable by computing SVM-based variable importance. A sequence of SVM models invoking at the beginning the k most important

University of Malaya

variables, by step of 1. When k becomes too large, the additional variables are invoked by packets. The set of variables leading to the model of smallest error rate are selected.

Genuer et al., (2010) proposed preliminary elimination and ranking using RF. Their method of variable is being applied in this study.

- Compute the RF scores of importance, cancel the variables of small importance;
- Order the m remaining variables in decreasing order of importance. Step 2. Variable selection:
- For interpretation: construct the nested collection of RF models involving the k first variables, for k = 1 to m and select the variables involved in the model leading to the smallest OOB error;
- For prediction: starting from the ordered variables retained for interpretation, construct an ascending sequence of RF models, by invoking and testing the variables stepwise. The variables of the last model are selected.

2.3 ANN

Artificial Neural Networks (ANNs) is a machine learning method that processes information by adopting the way on how the neurons of human brains work (Daliakopoulos et al., 2005), which consists of a set of nodes that imitating the neuron and carries activation signals of different strength. If the strength of the combined signals are strong enough, the signal will be propagated to the other neuron in the system. ANN is a black box as it does not provide any insights on the structure of the function being approximated. There are two approaches on ANN development which are supervised and unsupervised learning algorithm. Supervised learning involves learning relationship function between inputs and output from the examples presented in the training data sets, whereas unsupervised learning involves learning patterns in the

University of Malaya

input training data sets when no specific output values are supplied (Mohri et al., 2012). Both supervised and unsupervised ANN is adopted in this study. In supervised learning at each network layer, error is minimized between the layer's response and the actual data. The actual output of the network is compared with the expected output for that particular input. This results in an error value. The connection weights in the network are gradually adjusted until the correct output is produced. Kohonen self-organizing feature map (SOM) is used for the unsupervised learning in this study because it has several important properties that can be used within the knowledge discovery and exploratory data analysis process. Specific architecture like Hopefield network or Kohonean network is implemented by connecting the neurons in which they learn through process of self-organization (Navarro & Bennun, 2014).

MLP is a supervised ANN learning method that consists of three layers which are input layer, hidden layer and output layer (Figure. 2.6). The number of neurons of input layer is equal to the selected features. One hidden layer is preferred as classifiers but it also can have multiple hidden layers (Hekim, 2012). More hidden layers can be added to increase the capability of the network and is useful for nonlinear systems (Naghsh-Nilchi & Aghashahi, 2010). There is no fixed number of hidden layers that is needed in ANN. If computational complexity and processing time would increase if large number of hidden layer is used and classification errors can occur if the number of neuron is too small.

University of Malaya

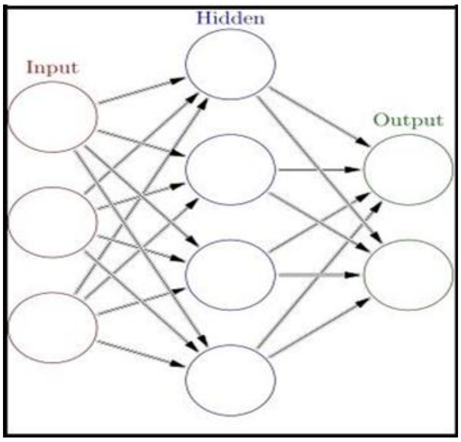

Figure 2.6: General Architecture of ANN

A MLP model with insufficient or excessive number of neurons in the hidden layer can lead to poor generalization and overfitting problem. There is no systematic method for determining the number of neurons in the hidden layer. It is only found by trial and error (Subasi & Ercelebi, 2005). To optimize ANN model performance, there are three data sets that are used for the ANN model development which are training set, test set and validation set. The RMSE is measured and the test set was used to evaluate the generalization ability of the network. The validation set was used to assess the performance model once the training phased has been completed. The process of cross-validation removes the risk of the neural network memorizing the data.

Resilient back-propagation multilayer perceptron is adopted in study as the supervised ANN. It consists of an input layer, one or more hidden layers comprising the computational nodes, and an output layer as illustrated in Figure 2.7.

University of Malaya

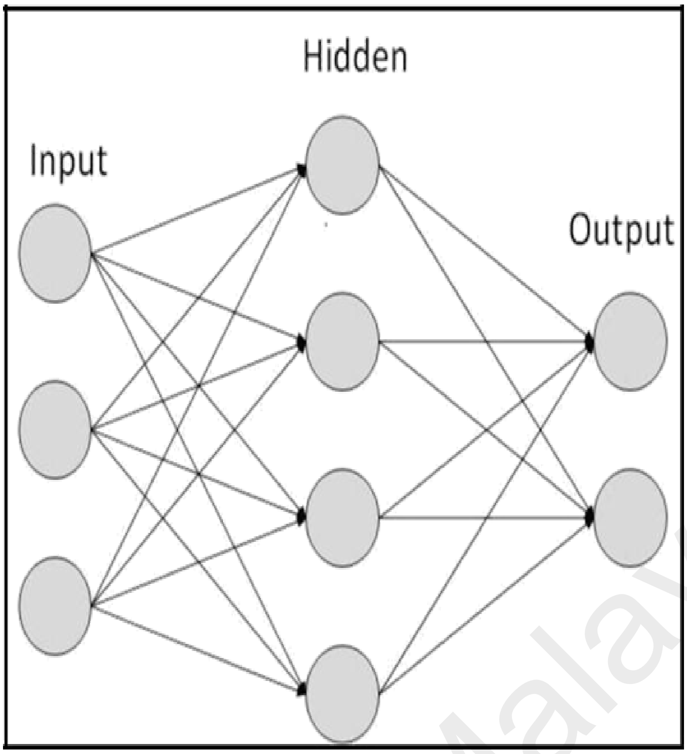

Figure 2.7: Architecture of the resilient backpropagation of ANN

Training a SVM for classification, regression or novelty detection involves solving a quadratic optimization problem that can transforms data from 2-dimensional input space to 3-dimensional feature space using alternative mapping (Ventura Dde et al., 2009)

The resilient backpropagation was trained in this study to build neural network predictions model. It is based on the traditional backpropagation algorithm that adjusted the weights of a neural network in order to find a local minimum of the error function. Therefore, the gradient of the error function is calculated with respect to the weights in order to find a root. In particular, the weights are modified going in the opposite direction of the partial derivatives until a local minimum is reached (Rojas, 1996). Weight backtracking is a technique of undoing the last iteration and adding a smaller value to the weight in the next step. Without the usage of weight backtracking, the algorithm can jump over the minimum several times (Riedmiller & Braun, 1993). In this

University of Malaya

study, the standard root mean square error (RMSE) was used to assess network performance. RMSE indicates the absolute fit of the model to the data or how close the observed data points are to the model's predicted values.

There are approaches need to be conducted in order to get the best results. Results will be depending on the variable selected. If the variable is selected wrongly, it will affect the result. In this study, the backward elimination method was used. Backward elimination is an approach in which it involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable that improves the model by being deleted and repeating this process until no further improvement is possible. The inputs are ranked using average %IncMSE before carrying out backward elimination process.

ANN package in R is built to train multi-layer perceptron in the context of regression analyses to approximate functional relationships between input variables and output variables. ANN in R (neuralnet) contains backpropagation algorithm, resilient backpropagation, activation and error function (Günther & Fritsch, 2010).

2.4 Decision Tree

Decision trees, together with rule based classifiers, represent a group of classifiers that perform classification by a sequence of simple, easy-to-understand tests whose semantics are intuitively clear to domain experts (Stiglic et al., 2012). There are two types of DT predict responses to data that is classification trees and regression trees. DT been used in the task of analysis of decision, to assist in determining a reaching goal strategy (Quinlan, 1987).

University of Malaya

A DT each internal node represents a "test" on an attribute, each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. The aim of decision tree is to predict a response, follow the decisions in the tree from the root (beginning) node down to a leaf node. DT is used as a visual and analytical decision support tool, where the expected values of competing options are calculated.

DT is indispensable graphical tools when the decision process involves many sequential decisions that are difficult to visualize and to implement. They allow for intuitive understanding of the problem and can aid in decision making. A DT is a graphical model describing decisions and their possible outcomes. DT generate a set of conditions that are highly interpretable and easy to implement (James et al., 2013a). DT can effectively handle missing data and implicitly conduct feature selection. Observations given are used to make a prediction using mean or the mode of the training observations to which it belongs. The set splitting rules used to segment the predictor space can be summarized in a tree which is called as DT methods (Kuhn & Johnson, 2013).

In this study DT tree type of regression tree is used. Common technique for constructing regression trees is the classification and regression tree (CART) methodology (Breiman et al., 1984). In constructing regression tree, the model development begins with the entire data set which is then searched to determine every distinct value of every predictor. This is done to find the predictor which is then split into value that partitions the data into two groups.

University of Malaya

The splitting is done such to ensure the overall sums of squares error are minimized using the following formula:

$$\text{SSE}=\sum_{i\in S1}(y_i \; -y1) \; + \sum_{i\, \epsilon S2}(yi-y2)\,2$$

Where ¯y1 and ¯y2 are the averages of the training set outcomes within groups S1 and S2, respectively. Then within each of groups S1 and S2, this method searches for the predictor and split value that best reduces SSE. This method is also known as recursive partitioning. To avoid over-fit the training set when the tree has reached maximum depth or is very large pruning needs to done to obtain a smaller tree (Kuhn & Johnson, 2013).

Tree pruning is used to produce good predictions on the training set. Large tree is grown and pruned to obtain subtree. Subtree with the lowest error test rate will be selected as the best pruned tree. Test error rate using cross-validation or validation set approach can be estimated using subtree (James et al., 2013a). Recent publications concerning DT analysis in the medical field indicate its usefulness for defining prognostic factors in various diseases such as prostate cancer, diabetes, melanoma, colorectal carcinoma and liver failure. The results of DT analysis are presented in the form of a flow chart, which is easy to use in clinical practice (Hiramatsu et al., 2011).

University of Malaya

2.5 Self-Organizing Map (SOM)

2.5.1 Background

Kohonen's self-organizing map is unsupervised mathematical model of topological mapping. SOMs learn on their own through unsupervised competitive learning, where it attempts to map their weights to conform to the given input data. The nodes in a SOM network attempt to become like the inputs presented to them which is called learning. The topological relationships between inputs data are preserved when mapped to a SOM network which suitable for representing complex data. SOMs are also known as vector quantization also considered as a data compression technique. SOMs provide a way of representing multidimensional data in a much lower dimensional space into one or two dimensions.

SOM is a nonlinear generalization of the principal component analysis and has found much application in data exploration particularly in data visualization, vector quantization and dimension reduction. SOM is inspired by biological neural networks; it is a type of artificial neural network which uses unsupervised learning algorithm with the additional property that it preserves the topological mapping from input space to output space making it a great tool for visualization of high dimensional data in a lower dimension. It is originally developed for visualization of distribution of metric vectors.

The quality of learning of SOM is determined by the initial conditions: initial weight of the map, the neighbourhood function, the learning rate, sequence of training vector and number of iterations (Pal & Pal, 1993)

2.5.2 The SOM Algorithm

Fagbohungbe et al. (2012) proposed the following self-organizing maps algorithms. The Self-Organizing Map algorithm can be broken up into 6 steps.

1) Each node's weights are initialized.

2) A vector is chosen at random from the set of training data and presented to the network.

University of Malaya

3) Every node in the network is examined to calculate which ones' weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).

4) The radius of the neighbourhood of the BMU is calculated. This value starts large. Typically it is set to be the radius of the network, diminishing each time-step.

5) Any nodes found within the radius of the BMU, calculated in the previous step, are adjusted to make them more like the input vector. The closer a node is to the BMU, the more its' weights are altered.

SVM can perform both linear and non-linear classification. Supervised approach is used when the data are labelled while unsupervised approach is implemented with the unlabelled data. SVM find the optimal separating hyperplane that separates the two groups of data points with the largest margin usin the following formula ℓ2-norm penalized optimization problem (Vapnik & Vapnik, 1998) .

$$\min \frac{1}{n} \sum_{i=1}^{n} \left(1 - yi(\beta 0 + x_i^T \beta)\right) + \frac{\gamma}{2} ||\beta||\beta$$

Where the loss function (1−·) + := max (1−·, 0) is called the hinge loss, and ≥ 0 is a regularization parameter, which controls the balance between the ‘loss’ and the ‘penalty’. SVM uses classifiers in which this classifier is a binary classifier algorithm that looks for an optimal hyperplane as a decision function in a high-dimensional space (Cristianini & Shawe-Taylor, 2000).

Support vector classification (SVC) is the algorithm searches for the optimal separating surface, i.e. the hyperplane that is, in a sense, equidistant from the two classes 1 and 0. SVC is outlined first for the linearly separable case. Kernel functions are used to construct non-linear decision surfaces. Slack variables are introduced to allow for training errors for noisy data, when complete separation of the two classes may not be desirable (Burbidge et al.,

University of Malaya

2001). Regression in SVM is carried out by using a different loss function called the -insensitive loss function $ky-f(x)k = \max\{0, ky- f(x)k - \}$. This loss function ignores errors that are smaller than a certain threshold > 0 thus creating a tube around the true output. The primal becomes: minimize $t(w, \xi) = 1\ 2\ kwk\ 2 + C\ m\ Xm\ i=1\ (\xi i + \xi * i)$ subject to $(h\Phi(xi), wi + b) - yi \leq - \xi i$ (13) $yi - (h\Phi(xi), wi + b) \leq - \xi * i$ (14)

$\xi * i \geq 0$ ($i = 1, \ldots, m$). To estimate the accuracy of SVM regression scale parameter of a Laplacian distribution on the residuals needs to be computed $\zeta = y - f(x)$, where $f(x)$ is the estimated decision function (Wu et al., 2004).

2.6 Support Vector Machine (SVM)

SVM is a supervised training algorithm that can be useful in the purpose of classification and regression (Vapnik & Vapnik, 1998). SVM had been widely applied in pattern recognition for data analysis and to test the performance of the provided dataset. SVM can be used to analyse data for classification and regression using algorithms and kernels in SVM (Cortes & Vapnik, 1995). SVM is a powerful tool in data mining, in which it works to discover patterns on a given dataset, which will help to enhance our understanding the analysed data and improve its prediction.

Support Vector Regression (SVR) is applied in this study which has the same principle as Support Vector Machine (SVM) for classification problem. SVR is the adapted form of SVM when the dependent variable is numerical rather than categorical. SVR is a non-parametric technique and the output model from SVR does not depend on distributions of the underlying dependent and independent variables. The SVR technique depends on kernel functions and uses the principle of maximal margin as a convex optimization problem. Application of a loss function ε-insensitive and the

University of Malaya

parameter *C* which is called the regularization constant and reflects the balance cost parameter to avoid over-fitting in SVM for regression. Error (ϵi) with the value of 0.1

is used in this study to developed SVM for regression and a C value of 1 is used which are identified based on trial and error. The kernel radial basis function (RBF) is adopted in this study a kernel of a general purpose when there is no a priori knowledge about the data is required. SVR uses a cost parameter, to avoid over-fit. The error and cost parameter is set to the value of epsilon = 0.1 cost C = 1. SVR model in this study have been constructed using R package "e1071". SVR technique uses kernel functions to construct the model. However, there is no systematic kernel selection procedure available. In this study, radial basis function (RBF) was selected. RBF kernel can reduce the computational complexity of the training procedure while giving good performance under general smoothness assumptions. Optimal parameter values for these kernel was set automatically using function available in R, these included the regularization parameter gamma (γ) and the RBF kernel function parameter sig^2 (σ^2) (Cristianini & Shawe-Taylor, 2000; Suykens et al., 2002)

2.6.1 Non-linear SVM

The decision function in SVM depends on the inner product between two vectors rather than on input vectors alone. SVMs can be extended to non-linear problems by means of a kernel function K that satisfies the Mercer conditions (symmetric semi-definite positive function). The kernel induces an implicit non-linear function ϕ which maps the sample point's $xi \in X$ into a high dimensional (even infinite) feature space T where one constructs the optimal hyperplane that separates the mapped point's ϕ (xi). This is equivalent to a non-linear separating surface in X. SVMs kernel methods is constructed to use a kernel for a particular problem that could be applied directly to the

University of Malaya

data without the need for a feature extraction process. This is particularly important in problems where a lot of structure of the data is lost by the feature extraction process (Sackinger et al., 1992) .

Training a SVM for classification, regression or novelty detection involves solving a quadratic optimization problem that can transforms data from 2-dimensional input space to 3-dimensional feature space using alternative mapping (Ventura Dde et al., 2009)

SVM uses binary classifier algorithm that looks for an optimal hyperplane as a decision function in a high-dimensional space (Sackinger et al., 1992).

Some widely used kernels in SVM are:

- Polynomial: $K(x, z) = (hx.zi + 1) d$, where $d \in N$ is the degree.
- Quadratic Kernel: $k(x, z) = (x \top z) 2$ or $(1 + x \top z) 2$
- Radial Basis Function (RBF): $K(x, z) = \exp - 1 2\sigma 2 kx - zk 2$, where $\sigma \in R * +$ is the bandwidth. (Rai, 2011).

2.6.2 Model evaluation

Test set error estimate is obtained from growing a tree from a boostrap data (Verikas et al., 2011) which then be used to estimate the variable importance. These are the important byproducts of RF. As compared to k-NN, SVM and Neural Network (NN), RF performs very well as it gives the insights on which variable are more important (KJ Archer, 2008) based on the test error estimates and the variable importance construction. In RF for regression, the test error estimate is defined by the Root Mean Square Error (RMSE).

University of Malaya

RMSE is a measurement for the difference (or residuals) in the value of the observed result and the predicted result. To obtain this prediction error, the standard deviation of the prediction and observed data are calculated before RMSE is generated.

The lower the value of the RMSE, the better quality of the prediction model generated. The formula of the RMSE as referred to Armstrong and Collopy, (1992).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

p is the observed value, **a** is the predicted value, **n** is the total number of the dataset, and **i** is the individual reading of the value. RMSE alone is not sufficient to determine predictive model performance. Therefore two-tailed paired t-test and Mann-Whitney u-test are adopted to further evaluate the model in this study. T-test is the statistical significant indicates whether two groups' (predicted and actual in this case) averages show difference. Mann-Whitney u-test is a non-parametric test that is used to compare two sample means of the same population whether they are the same or not. The significant level of the t-test is 0.05 with the confidence level of 95%. If the p-value of the predictive model is more than 0.05, hence we accept the null hypothesis and reject the alternative hypothesis which means that there is no difference between the two groups of values and if the p-value is less than the chosen significance level, we reject the null hypothesis and accept the alternative hypothesis hence there is difference between the two groups of value. The same condition is applied to the u-test in determining the association between predicted and actual values (Greenland et al., 2016)

University of Malaya

2.6.3 Validation method of model

Data pre-processing is done to partition the datasets into testing and training dataset. This is to avoid over fitting of the predictive model result. K-fold cross validation method is a technique of generalizing the results of statistical analysis to an independent dataset in predicting the performance of the predictive model (Geisser & Johnson, 1993). In K-fold cross validation original sample will be randomly divided into k subsamples. One of the k subsamples is kept as the testing data while the other subsamples are used as the training data. The processes of building the predictive models are repeated k times in each iteration, each of the subsamples are used as the validation data. The generated test error estimate generated then will be averaged to produce a single estimation. K is an unfixed parameter, but usually 10 fold cross validation is used (McLachlan, 2004).

University of Malaya

CHAPTER 3: METHODOLOGY

3.1 Data Collection

A collection of four years of patient data and radiographs from the years 2009, 2010, 2011 and 2014 respectively were obtained from the University Malaya Medical Centre pediatric orthopedic unit, Orthopedic Department in Kuala Lumpur, Malaysia. Radiographs of fractured bones (femur, tibia and fibula) from infants and young children of ages less than 12 years were included, with ages recorded from the time of initial injury. The individuals in the 57 samples of children age 12 and below from the time of injury are from the surrounding Selangor population. The radiographs examined consisted of images from each individual. The radiograph images were analyzed by a Pediatric Orthopedic surgeon. Any individuals demonstrating comorbidity or any systemic disorder, which may affect the bone healing rate, were excluded from the study. Data was retrieved based on radiography and patient records. Through radiograph examinations, variables such as bone involve, region of bone, type of fracture and measurement parameters such as angulation of the fracture (in coronal and sagittal planes) and contact area of the fracture were obtained. Diameter of the fractured bone, in two views, anterior and lateral was also analyzed. The time interval between injury and the union of the bone, age and sex of the patient and other demographic factors are also identified. Those parameters were selected based on the recommendations of the orthopaedics specialists. They select those parameters due to their importance in affecting the healing rate as well as their ease to demonstrate and analyze from the radiography images. Healing time was defined as the time in which the bone achieved union based on radiographic evidence. Healing was defined at the time the radiograph showed that the fracture line was absent, and that the cortices between the fracture sides were well formed. The remodeling of the bone, thereafter, was not taken into consideration for this study.

University of Malaya

3.2 Data Analysis

The parameters used in this study, were lateral (sagittal plane) and anterior (coronal plane) angle and contact area and age. The measurement is taken based on radiograph features (Malone et al., 2011). Angulation describes the direction of the distal bone and degree of angulation in relation to the proximal bone. Loss of alignment or displacement is usually accompanied by some degree of angulation, rotation or change in bone length. Contact area is described as how much the bone is in contact with each other, taking into account the amount of contact in two radiograph views, in anterior and lateral views. Diameter of the bone is the measurement between the two cortices of the fractured bone, in two views. Angulation and contact area are interrelated to each other (Staheli, 2008). Besides continuous variables categorical variables used in this study are type of fracture, bone involved, race, gender, bone segment, fracture segment. Summary statistics of the continuous variable used in this study given in Table 3.1 and Table 3.2 displays categorical variables used in this study.

Table 3.1: Summary statistics of data used in this study

Variable	Min	Max	Median	Standard deviation
Age (years)	0.16	13	8.5	3.92
Lateral Contact Area (%)	0	100	100	32.08
Lateral Diameter (mm)	6.9	41.6	15.3	7.30
Lateral Contact Area (mm)	0	41.6	12.6	8.95
Lateral Angulation (degree)	0	35	2.5	6.61
Anterior Diameter mm	5.9	42.3	14.8	7.81
Anterior Contact Area (mm)	0	42.3	12	9.77
Anterior Angulation (degree)	0	46	3	7.86
Anterior Contact Area (%)	0	100	91.3	33.60
Healing Weeks (output)	3	12	8	2.81

University of Malaya

Table 3.2: Summary of categorical variables used on this study

Variable	**Categories**
Type of fracture	(1= Transerve, 2=Spiral, 3= Torus)
Bone Involved	(1=Femur, 2= Tibia /Fibula)
Race	(1 = Malay, 2= Chinese =3 = Indians)
Gender	(1= male , 2= Female)
Bone Part	(1= Proximal, 2= Diaphyseal, 3= Distal)
Bone Segment	*(1= Metaphysis, 2= Diaphysis, 3= Epiphysis)*
Number of fractured Bone	(1= both bones , 0 = one bone)

3.3 Development of the models

All the models developed in this study have been implemented on R software after downloading the necessary packages. The models selected in this study RF, SVM, ANN, SOM and DT have not been applied in orthopaedic paediatric field before hence it is a novel study.

Model Validation

Validation methods of model include substitution method, retention method (or called holdout method) and cross-validation (CV) method. There are leave-one-out CV (LOO-CV), leave-more-out CV (LMO-CV) and *k*-fold CV. In the absence of a very large designated dataset in this study. We proposed the usage of Leave-one-out method.

University of Malaya

This method has a couple of major advantages such as; it has far less bias and provides an approximately unbiased estimate for the test error. In this method a single observation (x1, y1) is used for the testing set, and the remaining observations {(x2, y2), . . . , (xn, yn)} is made up the training set. The RF, ANN and SVM model is built on the n − 1 training observations and a prediction ˆy1 is made for the excluded observation, using its value x1. Mean square error value (MSE) is calculated as follows for the single observation. MSE1 = (y1 − ˆy1)2. This procedure is repeated by selecting (x2, y2) for the testing data, training the RF method on the remaining n − 1 observations sets {(x1, y1), (x3, y3), . . . , (xn, yn)}, and computing MSE value. Repeating these approach n times produces n squared errors. The leave-one-out method estimate for the test MSE is the average of these n test errors (James et al., 2013b).

3.3.1 Model evaluation criteria

Root mean square error (RMSE) is used in this study as a model to assess the development of RF, SVM and ANN. RMSE is used to measure the average level of prediction error which indicates the absolute fit of the model to the data or ow close the observed data points are to the model's predicted values. It is shown in the following formula where X is the observed value, Y is the Predicted value, n is the number of reading used and 1 is the individual reading of the value. T-test and U-test are used to compare the values of the predicted and actual adherence level. If *p*-value is more than 0.05, there is significance different.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X\text{obs,i} - Ymodel,i)^2}{n}}$$

Model evaluation

All the models have been evaluated RMSE and t-test, U-test.

RMSE

The standard deviation of the prediction errors is known as RMSE. It is a measure of how far from the regression line data points are; RMSE is also a measure of how spread out these residuals are.

T-test

t-test' is a statistical significance indicates whether or not the difference between two groups' averages most likely reflects a "real" difference in the population from which the groups were sampled. In other word t-test evaluate the means of two groups to determine if they are statistically different from each other

U-test

It is a Non-parametric alternative test to the independent sample t-test. It is used to compare two sample means that come from the same population, and used to test whether two sample means are equal or not.

University of Malaya

3.4 Random Forest development

The principle of RF is to combine many binary decision trees constructed using several bootstrap samples from a learning sample and choosing randomly at each node a subset of predictors. In RF each tree is a standard classification or regression tree (CART). At each node in RF a given number of input predictors (indicated as mtyr) are randomly chosen and the best split using decrease of Gini impurity is calculated only within the subset. No pruning step is performed therefore all the trees of the forest are maximal trees. This random selection of features at each node decreases the correlation between the trees in the forest thus decreasing the forest error rate.

The random subspace selection method has been demonstrated to perform better than bagging alone when there are many redundant features contribute to discrimination between classes. The main difference between bagging and random forests is the choice of predictor subset size. For instance, if a random forest is built using all predictors mtyr

= p, then this equal to bagging a well know tree based method. Leave on out method was used for sample splitting into training and testing set.

The RF method for regression is implemented in this study using randomForest function built in R kernel software version.

University of Malaya

The step of running RF algorithm used in this study is as follows:

1) Each tree is built using bootstrap sample of the training set.
2) When growing a tree, at each node m variables are randomly selected out of the p available variables. (mtyr argument) is used to indicate the number of m variables to be selected.
3) Minimum error is obtained by using out-of-bag routine to estimate generalization error (OOB) on data set.
4) RF method is then applied for testing data for prediction.
5) This step is repeated using different mtyr argument starting from (mtyr = 17). Where the default value of regression RF is p/3 = 5.
6) The RF is repeated with different ntree argument starting from ntree=500; ntree= 1000 and ntree= 2000. This is done to examine the sensitivity to method argument mtyr and ntree to better determine important variables and the stability of the variable importance scores.

Model assessment for the developed RF method adopted in this study is root mean square error (RMSE), where RMSE is a measure of the average level of prediction error. It indicates the absolute fit of the model to the data or how close the observed data points are to the model's predicted values. It is shown in the following formula where y is the observed value, ỹ is the predicted value, n is the number of readings used, and j is the individual reading of the value. T test is also used to compare the values of the predicted and actual healing week.

3.4.1 Variable importance

In this study, variables importance measures were used as implemented in the *randomForest* on R package. Mean decrease accuracy was used to determine the important variables to predict adherence level for arthritis patients. Mean decrease in accuracy using the out-of-bag observation (OOB) is based on when on the out of bag samples when a given variable is excluded from the model. About one-third of the observations in the data set will be not used to grow the tree and it is considered as OOB observations for the tree. The mean decrease accuracy is defined as the difference between the OOB error (MSE for regression) that obtained through random permutation of the predictor of interest and the OOB error from original dataset. High permutation value leads in increasing the OOB error. The variables selected from RF model are used to develop all other model in this study.

After the k-fold process, variables importance was generated from RF. The average of %IncMSE were counted and used to rank the variables. After that, backward elimination was performed. The variables were deleted one by one according to their ranking that was generated by RF. The highest RMSE error was selected as one of the important variables that can affect healing time. After backward elimination was performed, variables importance was obtained. Predictions model were developed by using all variables and the selected variables. Then, predicted and actual healing time was compared using RMSE, t-test and u-test from both models.

3.5 Artificial Neural Network Model Development

The architecture of ANN was determined by trial and error. The geometry of ANN used for all input variables (Figure 3.1) in this study is 17-8-1 that represents input layer with 17 nodes, 8 nodes in the hidden layer and one output layer. Figure 3.2 illustrates ANN architecture for selected variables from RF model with 3-2-1. The resilient backpropagation was trained to build neural network predictions model in this study using sigmoid transfer function with the value of 0.01 for learning rate. It is based on the traditional backpropagation algorithm that adjusted the weights of a neural network in order to find a local minimum of the error function.

University of Malaya

Therefore, the gradient of the error function is calculated with respect to the weights in order to find a root. In particular, the weights are modified going in the opposite direction of the partial derivatives until a local minimum is reached (Rojas, 1996). Weight backtracking is a technique of undoing the last iteration and adding a smaller value to the weight in the next step. Without the usage of weight backtracking, the algorithm can jump over the minimum several times (Riedmiller & Braun, 1993). Figure 3.1 illustrates ANN architecture used in this study constructed using all input variables.

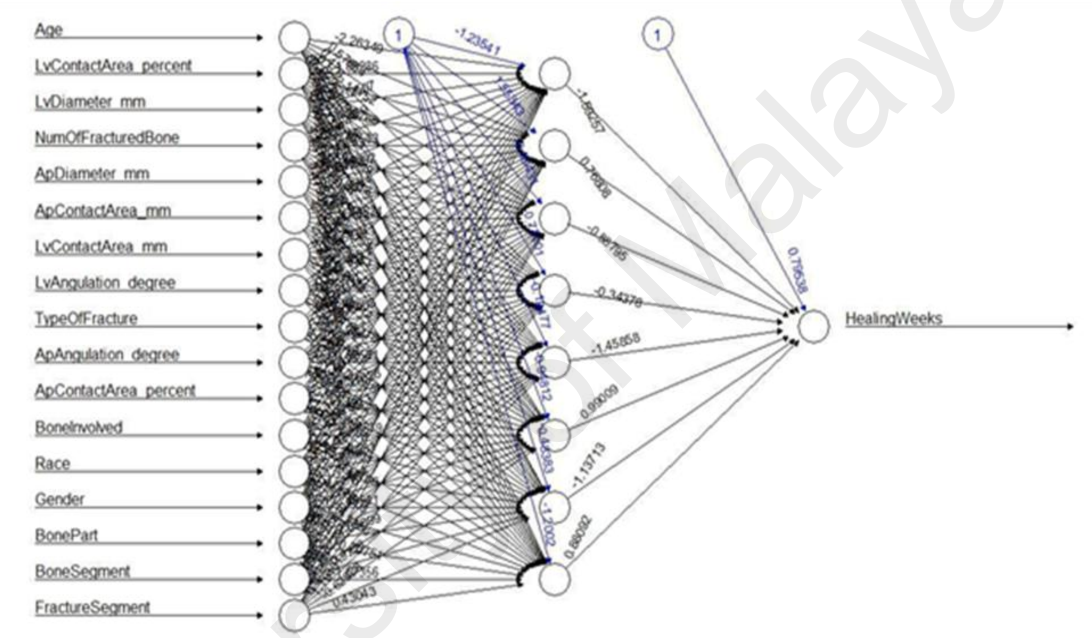

Figure 3.1: Artificial Neural Network architecture for all variables

University of Malaya

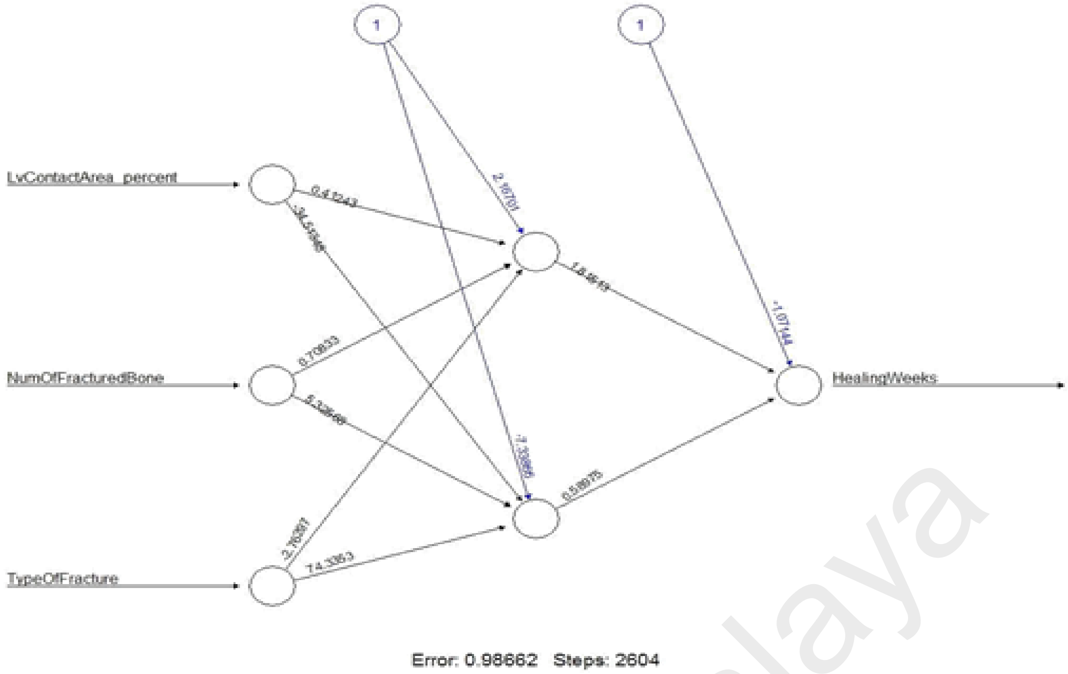

Figure 3.2: Artificial Neural Network architecture for selected variables from RF

3.6 SOM

Self-organizing map was generated by using toolbox in MATLAB Ver. (R2013, Math Works). SOM usually used as a tool to support natural language interfaces in the information system design process and as the test for the conceptual model comprehensibility. SOM also was reported as an excellent tool in visualization of high dimensional data. In this study, SOM was applied to cluster and map adherence level with respect to input variables.

The quality measures that are used in SOM are quantization error and topographic error which are used to measure how the map fit with the input data and how well topology of the data is preserved. The best map is expected to have the smallest average quantization error. The Euclidean distance between the inputs are calculated and visualized as distance matrix or known as U-matrix. The U-matrix represents the distance between neurons.

University of Malaya

The distance was calculated and present with different colorings between the adjacent nodes. The clusters can be detected because there were light areas as clusters and dark areas as cluster separators.

In the Kohonen network, every node in the input layer is represented as vector (xi) and is connected to each neuron (j). This connectivity is constituted as weights, wij(t), adaptively varying at each iteration of t. The weights are arbitrarily assigned into small value at first. As the input vector is sent through the network, each neuron computes the summed distance between the weight and input. The winning neuron is selected based on neuron that responds greatly to a given input vector. The winning neuron has the weight vector which has the shortest distance to the input vector. The winning neuron and may be its neighbouring neurons are allowed to learn by altering the weights in a way to additionally decrease the Euclidian distance among the weight and the input vector via the following equation.

$$dj\ (x) = \sum_{i=1}^{D} (Xi - Wj, i)^2$$

Z_j is assigned 1 for the winning and neighboring neurons while 0 is assigned for the other neurons, and represents the fractional to increase of the alteration (Provost & Kohavi, 1998). U-matrix is the acronym for unified distance matrix.

The U-matrix representation of SOM visualizes the distances between neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent nodes (Ritter & Kohonen, 1989). The Euclidian distance between the inputs are calculated and visualized as distance matrix (U-matrix). Kohonen Self Organizing Feature Maps (SOM) as introduced by Kohonen was applied to ordinate, cluster and map fracture healing time with respect to input variables. SOM reduces data dimensions by producing a map of 1 or 2 dimensions which plot the

University of Malaya

similarities of the data by grouping similar data items together. Thus SOM reduce dimensions and display similarities. This enables the discovery or identification of features or patterns of most relevance through data reduction and projection.

The winning neuron is selected based on neuron that responds greatly to a given input vector. The winning neuron has the weight vector which has the shortest distance to the input vector. The winning neuron and may be its neighbouring neurons are allowed to learn by altering the weights in a way to additionally decrease the Euclidian distance among the weight and the input vector via the equation.

The SOM in this study have been developed using variables selected from the variable importance measures resulted from the previous RF method. Variables are ranked in the decreasing order from the RF scores of importance. The SOM is built using variables with higher scores until a suitable quantization and topographic error have been achieved.

University of Malaya

3.7 Decision Tree

Decision tree is a tree-based method usually used for interpretation because it is easy to implement and simple. In this study, decision tree was generated by using R 3.3.2. The packages that were used to build the decision tree are *caret* and *rpart.*

The data is divided into 70% for testing and 30% for training.

The following are the steps to build decision tree:

1. Determine the size of the tree based on its relative error.
2. After that, the number of tree split, errors and standard deviation were calculated.
3. Variables importance were obtained
4. Decision tree were built for unpruned tree and pruned tree.

Tree pruning is used to produce good predictions on the training set. The lowest error test rate will be selected as the best pruned tree.

Decision tree was built by using *rpart* package. In order to grow the decision tree, the algorithm that were used are 'data' to specify the data frame and the 'method' used for regression tree is "anova". Then, to examine the decision tree the algorithm that were used in this study are 'printcp(fit)' used to display cp table, cp means cost-complexity parameter, 'plotcp(fit)' is used to plot cross-validation results that generated automatically in the algorithm, rsq.rpart(fit) used to plot approximate R-squared and relative error for different splits. The labels are only appropriate for "anova" method. Print(fit) used to print results, summary(fit) used to display detailed results including surrogate splits, plot(fit) used to plot the decision tree, text(fit) used to label the decision tree plot, post(fit,file=) used to create postscript plot of decision tree and prune(fit, cp=) used to prune the tree to avoid overfitting the data.

3.8 Support Vector Machine (SVM)

Support Vector Regression (SVR) is applied in this study which has the same principle as Support Vector Machine (SVM) for classification problem. SVR is the adapted form of SVM when the dependent variable is numerical rather than categorical. SVR is a non-parametric technique and the output model from SVR does not depend on distributions of the underlying dependent and independent variables. The SVR technique depends on kernel functions and uses the principle of maximal margin as a convex optimization problem. SVR uses a cost parameter, to avoid over-fit. The error and cost parameter is set to the value of epsilon = 0.1 cost C = 1. SVR model in this study have been constructed using R package “e1071”. SVR technique uses kernel functions to construct the model. The commonly used kernel functions are: a) Linear, b) Polynomial, c) Sigmoid and d) Radial Basis. In this study, the constructed SVR model used Radius

Basis Function (RBF) kernel. The kernel trick allows the SVR to find a fit and then data is mapped to the original space.

University of Malaya

CHAPTER 4: RESULTS

The RF model has been used to determine variable importance. Table 4.1 below list out the variable importance based on mean square error percentage of error increase average value.

Table 4.1: List of variables importance based on %IncMse

Variables	%IncMSE
Number Of Fractured Bone	16.78
Lv Contact Area (%)	10.36
Type Of Fracture	9.44
Age (years)	8.69
Ap Contact Area (mm)	6.24
Lv Diameter (mm)	5.09
Ap Contact Area (%)	5.00
Ap Diameter (mm)	3.36
Lv Contact Area (mm)	3.33
Bone Part	3.22
Bone Involved	3.12
Ap Angulation (degree)	1.72
Race	0.93
Bone Segment	0.89
Fracture Segment	0
Gender	-0.44
Lv Angulation (degree)	-2.44

University of Malaya

Each variable has been backwardly eliminated to determine the best model of each run. In RF structure mtry was used to select the number of variables available for splitting at each tree node. The default value of this parameter depends on which R model is used: RF for classification models, the default is the square root of the number of predictor variables (rounded down). For regression models, it is the number of predictor variables divided by 3 which is used in this study.

Figure 4.1 presented on the graph represent the RMSE when each particular variable is eliminated based on their importance generated from RF variable importance method. The elimination is using back ward elimination method eliminating from the least important variable to the most important variable. The variables are deemed important when the error rate increases when the variable is eliminated. The variables type of fracture, lateral contact area in percentage and number of fractured bones are identified as identified from RF as variables that effects healing time.

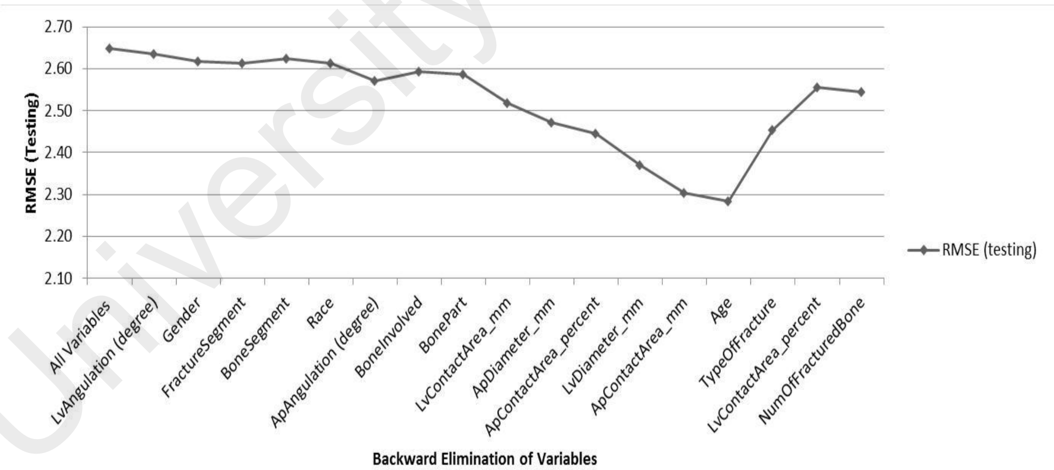

Figure 4.1: RMSE values for testing using RF method with backward elimination

The Figure below illustrate the predicted value of the healing time for testing dataset of all the models against the actual time required for the fracture recovery using all the 17 variables identified in this study as well as using the selected variables.

Figure A illustrates the predicted value of the healing time for testing dataset against the actual time required for the fracture recovery using all the 17 variables identified in this study. The RMSE for RF is recorded is 2.65 with a t-test value of 0.81 and U-test of 0.95 for testing dataset.

Figure B shows the predicted value of the healing time against the actual time required for the full recovery by using the best three variables identified from the backward elimination process. The variables are: number of fracture, lateral contact area in percentage and type of fracture. The RMSE recorded for this model is 2.28 with t-test value of 0.86 and U-test value of 0.72. The model performance is almost similar to using all 17 variables.

Figure D illustrates ANN model built using all 17 input variables with RMSE of 2.37 for testing dataset. Better prediction results are achieved when using selected variables from RF method as depicted in Figure C the RMSE recorded using selected variables for testing dataset is 1.99.

SVM model built using all input variables and significant variable is illustrated in Figure. E and F respectively. The RMSE for testing dataset is reported as 1.82 using all variables and 1.96 using selected variables identified from backward elimination. There is a negligible difference between SVM model constructed using all variables and selected variables. Therefore it can be concluded that SVM model using limited number variables are preferable compared to using all 17 input variables.

University of Malaya

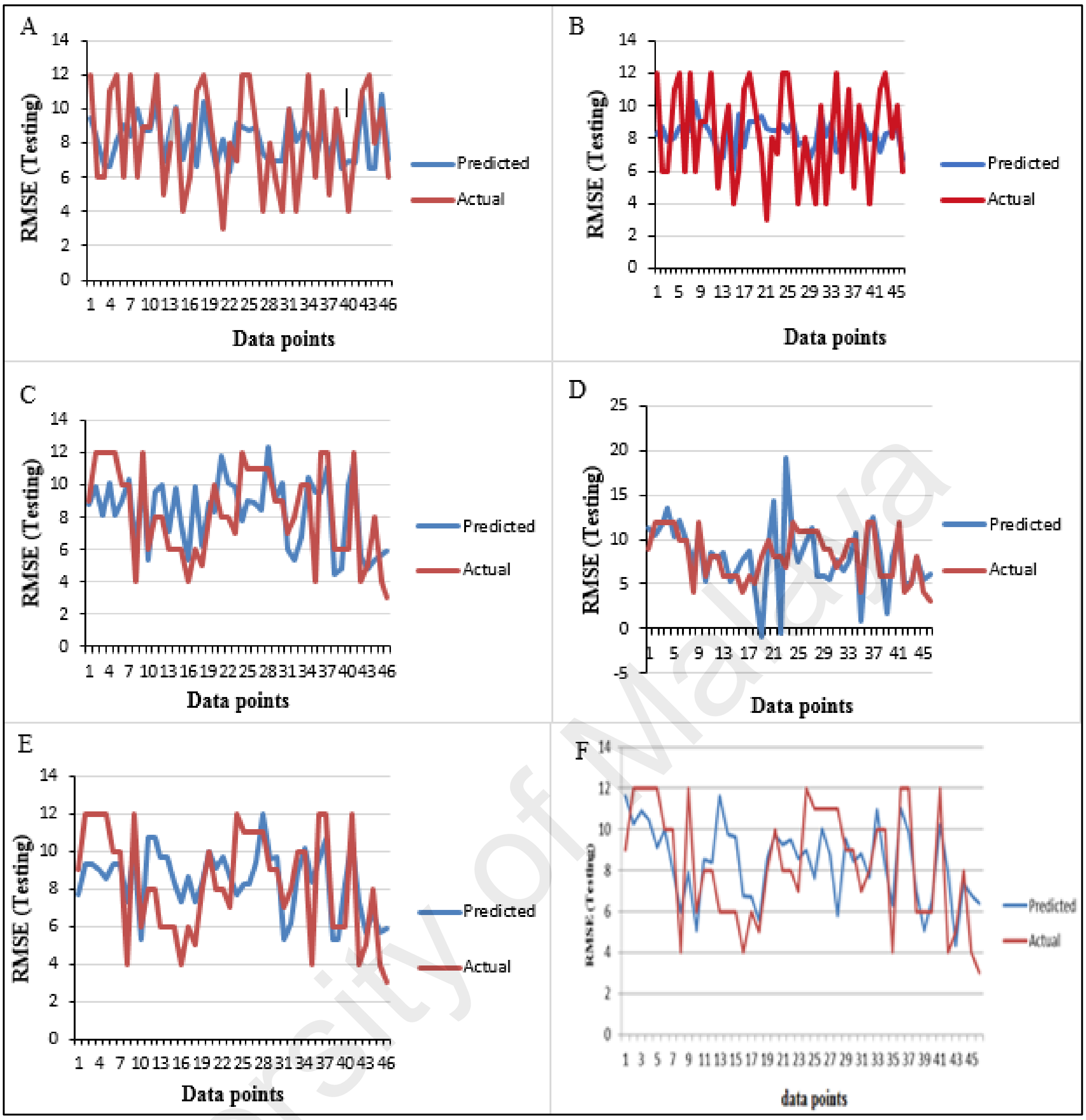

Figure 4.2: Predicted value vs actual value of healing weeks for all the models

Table 4.2 summarizes RF, SVM and ANN model performance using RMSE and Table 4.3 illustrates t-test and U-test results for models developed using all 17 input variables and selected variables in this study. SVM have outperformed both ANN and RF model.

However, it is noted that there is not much differences reported between the performances of all the models.

University of Malaya

Table 4.2: Performance's comparison of model using all and selected variables

RMSE	RF		SVM		ANN	
	Testing	Training	Testing	Training	Testing	Training
All variables Used	2.64	2.65	1.82	0.01	2.37	0.8
Selected Variables: Type of Fracture Lateral Contact Area Number Of Fractured Bone	2.28	2.29	1.96	0.38	1.99	0.93

Table 4.3: T-Test and U-Tests results

Model used	T-Test	U-Test
RF (All variables used)	0.81	0.95
RF (Selected variables)	0.86	0.72
SVM (All variables used)	0.81	0.95
SVM (Selected variables)	0.73	0.91
ANN (All variables used)	0.79	0.81
ANN (Selected variables)	0.61	0.80

The visualization of data in this study has been performed using two methods that are DT and SOM. Figure 4.3 illustrates the best number of split to generate DT which is 2. The DT is generated for the best number of split and it is illustrated in Figure 4.4.

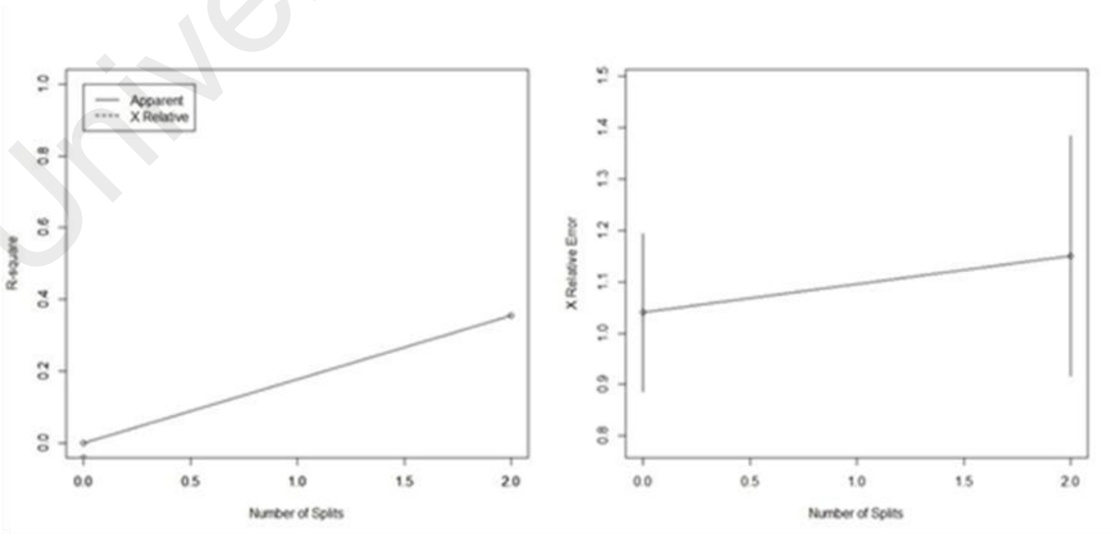

Figure 4.3: Number of splits for Decision Tree

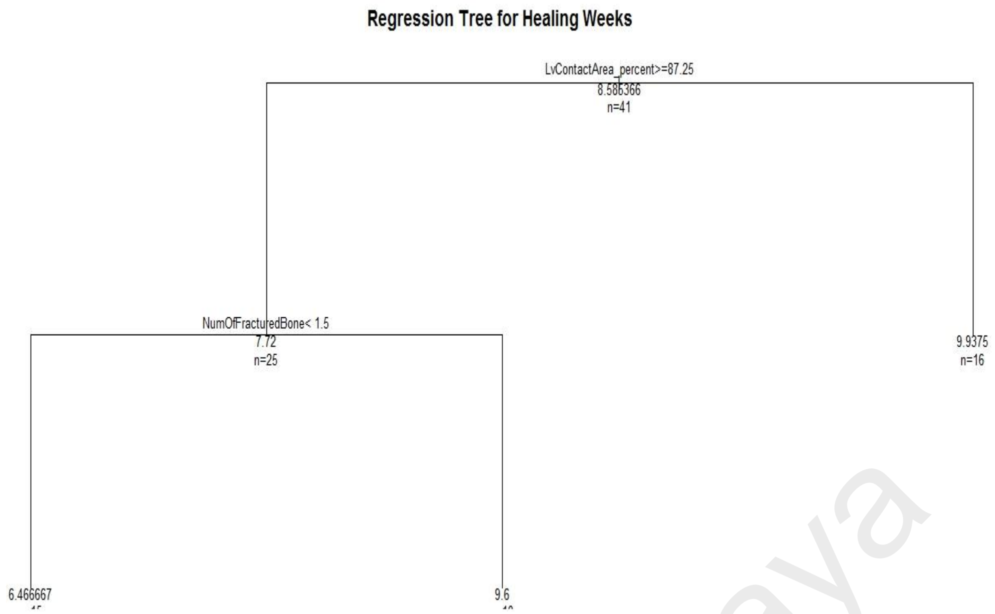

Figure 4.4: DT tree for regression for predicting healing week

Decision Tree in Figure 4.4 suggests that when the lateral contact area percentage is 87 % or more and the number of fractured bones is less than two bones the estimated healing time is around 6 weeks. The healing time is longer around 9 – 10 weeks when more number of bones are fractured and the contact area is less. This finding conformed to SOM illustrated in Figure 4.5. The quantization error of 0.17 and topographical error of 0.00 is obtained for the SOM map which is considered as a very good value for quantization (the smaller the value the better the quantization).

University of Malaya

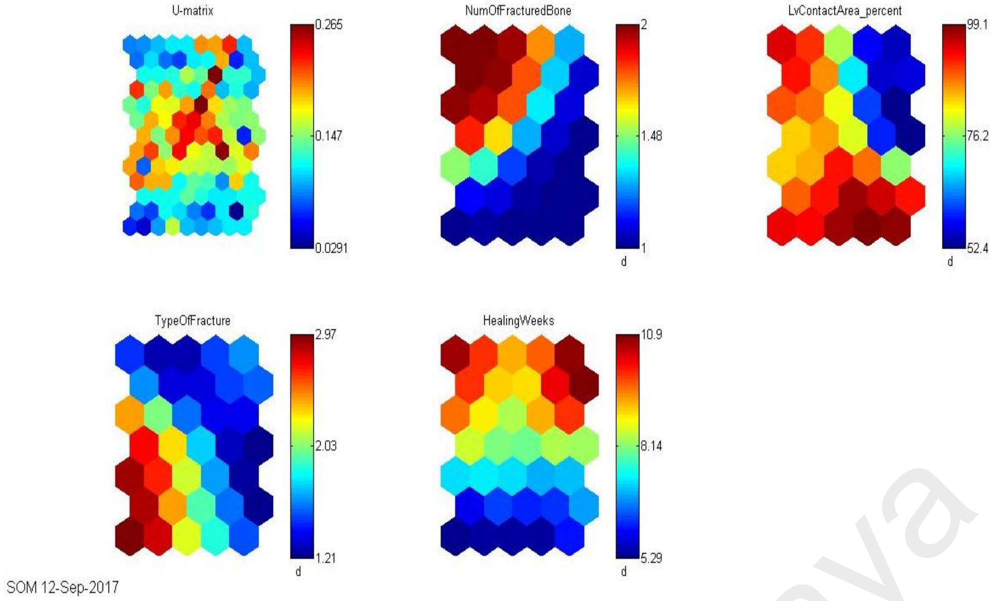

Figure 4.5: SOM map for selected variables against healing weeks

Figure 4.4 illustrates when the contact area percentage is large more than 87 % as discovered from the DT depending on the number of fractured bones and the type of fracture. SOM map as illustrated in Figure 4.5 includes age to explain healing time.

The blue colour represent low recovery which is 5 weeks while red colour represents longer time to fully recover and the colours in between indicate the range from 5 weeks to 10 weeks. In general, the blue colour in the Figures presented above indicate low value of the variable while red value indicates a higher presence of that particular variable. All the colours in between (shades of other colours indicate an average values for each individual variable) means the value is classified as average. Age is an important factor and it is identified as important variables from the variable selection method from RF however during the backward elimination process elimination of the age variable did not affect the RMSE therefore the variable age was not selected to developed the prediction model. The quantitation and topographic error is also higher when age is included in the

University of Malaya

SOM map. The quantization error is reported as 0.201 and the topographic error is reported as 0.065.

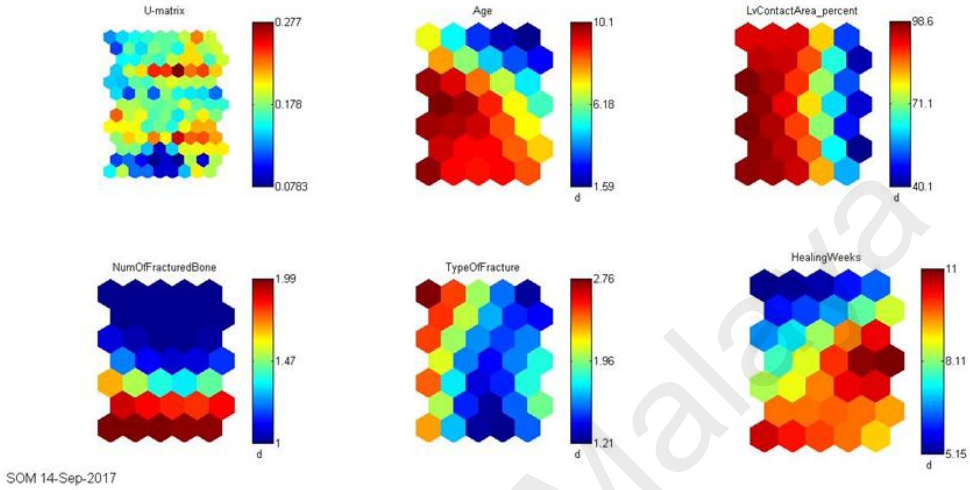

Figure 4.6: SOM map for selected variables including age against healing weeks

Figure 4.6 illustrates that younger children age heals faster compared to older children for all types of fractures.

University of Malaya

CHAPTER 5: DISCUSSION

In this study RF was used to determine the list of variable importance. The choice of variable importance measure plays an important role in determining model performance. The mean decrease in accuracy importance measure that is used in this study uses an unbiased splitting criterion and avoids both systematic bias and the increased variance as compared to the Gini measure of variable importance. The Gini measure of variable importance is used in RF for classification and is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees. RF for regression trees is applied in this study where, the node impurity is measured by the training residual sum of squares (RSS). The Gini measure is biased towards selection of predictors with many categories or larger continuous value as the predictors variables used in this study consist of categorical variables. The highest RF method accuracy in this study was achieve using the default value of number of variables for node splitting that is p/3 for regression trees, and to avoid over fitting, the number of trees should be sufficiently large as suggested by Breiman (2001). In this study, the number of tree parameter identified via trial and error is set to ntree 2000 to construct the RF model.

Variable importance evaluations based on the variable importance measures available from RF are widely used for data exploration and understanding. However, variable importance rankings, depends on the number of variables (mtyr value) used to split a node when designing RF, especially when the number of trees in RF is small. In this study default value of mtyr that is p/3 for regression trees is used. Brieman (2002) suggested that to start with the default mtry, half of the default and twice of the default. In our study, the default value gave the best results, however the result does not change extremely for different mtry value. The number of tree for development of RF model in

University of Malaya

this study is set to ntree 2000, as according to Brieman (2002) larger tree provides stable estimates of variable importance and proximity.

Backward elimination is performed in this study to identify variables that significantly affects healing week. Backward elimination algorithm relies only on significance as a sufficient condition to remove insignificant variables one-by-one from a model (Royston & Sauerbrei, 2008; Vittinghoff et al., 2011). In this study, backward elimination is performed based on results from RF importance method. The variables that when are removed causes significant increase in RMSE in the testing dataset of the RF model is deemed as important. This is based on the concept of 'purposeful selection algorithm' , where in this concept significance and change-in-estimate criteria (RMSE error rate on testing dataset that is used in this study) are combined for selecting significant variables for a final model construction (Bursac et al., 2008; Dunkler et al., 2014). The variable importance obtained by applying RF and backward elimination technique identified in this study are percentage of contact area, type of fracture and number of fractured bones. Based on AO guidelines, type of fracture and contact area between fragments have been showed to influence the fracture healing time. The number of affected bones involved in the fracture though has been shown to influence fracture healing time in the lower limb. This is because the single bone considered here was the femur, which has adequate soft tissue coverage, which promotes quicker healing. The tibia and fibula has poor soft tissue coverage, hence would be considered to take a longer time to achieve union (Audigé et al., 2005; Audigé et al., 2004; Slongo et al., 2006).

It has been well documented that age of the patient is an important criteria, where the younger children heal faster (Staheli, 2008). The variable age has been identified from the RF variable importance as one of the important variable that effects healing time.

University of Malaya

Previous study Malek et al. (2017) constructed the RF model for predicting paediatric fracture healing time based on only RF variable importance method without considering backward elimination. However, in this study using both RF and backward elimination method, the elimination of variable age does not significantly increase the testing RMSE. This variation could be a result of different combination of method used and due to the fact of limited number of dataset are available in this study. However, considering that variable age is ranked higher in RF method it has been included in the SOM map to explain its importance in fracture healing time.

Predictive models for ANN, SVM and RF was developed using all input variables and selected variables from RF variable selection method and backward elimination procedure. The values of performance metrics for each model are calculated considering all of the seventeen input variables and selected variables (i.e.: age, lateral contact area (measured in percentage), lateral diameter (measured of millimetre), number of fractured bone, anterior diameter (measured of millimetre), anterior contact area (measured of millimetre), lateral angulation (measured in degree), type of fracture, anterior angulation (measured in degree), anterior contact area (measured in percentage), bone involve, race, gender, bone part, bone segment, fracture segment).

RMSE, t-test and U-test have been calculated to evaluate the performance of the models as illustrated in Table 4.3 and 4.4. The leave on out cross validation (LOOCV) method is used in this study. The LOOCV method can be considered as an accurate indicator of performance of a classifier on unseen data and is a commonly used statistical technique when limited data is available. The LOOCV is used during the training to prevent over fitting when limited dataset is available such as in this study (Kim et al., 2006).

University of Malaya

Models developed using selected variable provided better results compared to models developed using all variables for RF (RMSE 2.28) and ANN (RMSE 1.99) model. SVM model using all variables (RMSE 1.82) performed slightly better than using selected variables (RMSE 1.96). The lower values of RMSE indicate better fit. T-test and U-test are used further to support model performance, since t-test assumes that the difference between the two distributions under comparison is normally distributed, Wilcoxon signed-rank test or U-test is used, which does not make the same assumption of normality and the test results in the same conclusions about significance. SVM model using all and selected variables reported a U-test of 0.95 and 0.91 higher than RF and ANN model respectively. The findings from this study indicated that there was a small difference between models considering all inputs and reduced variables with SVM models slightly outperformed ANN and RF. Similar studies done in orthopaedic domain comparing SVM , ANN and RF model for osteoporosis risk prediction resulted in SVM model outperformed the other machine learning method with higher accuracy. (Kim et al., 2013; Wu et al., 2016).

SVM model depends on mapping data to a higher dimensional space by using function of a kernel, the maximum-margin hyper-plane will be selected in order to separates training data. Hence SVM improves the accuracy via optimization the space separation. Thus, it produces a better result compared to the other models used on this study. SVM for regression is used in this study as it is a non-parametric technique and does not depend on distributions of the underlying dependent and independent variables. It depends on the kernel function. SVM for classification have been applied in orthopaedic field and the accuracy reported is around 80%. (Hayashi et al., 2015; Wang et al., 2016; Wu et al., 2016). However to our best knowledge SVM application especially SVM for regression have not been applied in the paediatric orthopaedic.

University of Malaya

ANN for selected variables performance is higher than ANN developed using all variables.

Sensitivity of ANN model in this study was studied for different number of hidden layer neurons. There is no rule for selecting the number of hidden layer neurons it is suggested that this number should be equal to one more than twice the number of input variables (Hecht-Nielsen, 1988) .Overfitting or underfitting of ANN are results of too many or two little number of neurons used in the hidden layer. In this study, we used trial and error method to identify the number of optimum neurons in the hidden layer. Number of neurons used is eight for ANN developed using all input variables and two for ANN developed for selected number of variables. It was identified that higher number of neurons were not making a significant difference in the accuracy of the ANN model. ANN model in this study was developed using only one hidden layer as it should be adequate for most of the applications (Principe et al., 2000).

SVM and ANN outperforms the RF model by a small margin as reported in this study. ANN and SVM for selected variables both managed to perform better in predicting the lower values and higher values compared to RF. ANN and SVM closely followed the healing week pattern between the actual and predicted week and therefore performed slightly better than RF.

This can be due to the fact that ANN and SVM models are considered good at fitting functions and recognizing patterns in various dataset. Both ANN and SVM can approximate practically all types of non-linear functions (Desai et al., 2008; Gulati et al., 2010). However, there are some limitations are such as standardized coefficients for variable may not be straightforwardly calculated and presented and it is considered as a “black box” approach. The complete insight into the internal workings of the model or information for evaluating the interaction of inputs is unknown (Dayhoff & DeLeo,

University of Malaya

2001). RF compared to ANN and SVM is relatively easier to use and RF provides insight to the variable importance which can provide valuable information to clinician. RF has been applied in the paediatric orthopaedic domain to predict outcomes of intramuscular psoas lengthening in patients with cerebral palsy with the accuracy of 78%. The usage of RF provides valuable information variable importance that affects psoas which age identified as the most important variable (Schwartz et al., 2013) .Other studies involving application of RF in paediatric orthopaedic is reported based on our previous work that combines both RF and SOM method to predict the lower limb healing time in children (Malek et al., 2017). The study conforms to literature based on RF variable selection capabilities that age is an important factor affecting healing time.

Results that are presented in this study were also visualized in a 2-dimensional representation using SOM technique. This allows the clinician, if there is confidence in the original training data, to place a new patient within the context of previous or similar cases. The results of SOM compliments the results obtained from a pruned DT. Pruning is applied in this study to reduce size of decision trees to increase the predictive accuracy and avoid overfitting. The lateral contact area percentage was selected by DT, as the lower limb fractures in children had only minimal displacement of the fragments. This is expected of fractures in children as they have a thick periosteal layer, preventing displacement (Ogden, 2000).

The results of SOM were displayed using variables obtained from RF variable importance method and both RF and backward elimination method. The final quantization and topographic errors obtained are 0.17 and 0.00 for map constructed using both RF selection method and backward elimination using lateral contact area percentage, number of fractured bones and type of fracture. The lower value of the error measures the SOM map accuracy (McKibbin & Ralis, 1978). The quantization and

University of Malaya

topographic error is 0.201 and 0.065 reported is higher when age have been include as identified from RF variable importance method. Both error is important to consider identifying the quality of the SOM map.

The finding from SOM map in this study identified that when the lateral contact area percentage is (90 %) or higher, the faster the fracture heals. The torus fractures healed the fastest followed by spiral fractures and finally transverse fractures. This is due to the fact that torus fractures are incomplete fractures with intact periosteum. Spiral fractures have a larger contact surface area between fragments as compared to transverse fractures and as such, heal faster (Staheli, 2008).

The fracture healing time in younger children (5 years and below) is reported within 6 weeks (Ryöppy, 1972). In older children (9 years and above) the healing time is reported to be within 8 weeks. The healing time in younger children (age below 5) is reported to be around 8 weeks when the contact area is less than 30%. This is can be explained as younger children have a thicker periosteal later, which is highly osteogenic. This facilitates faster healing of the bone. Contact area between the fragments play a role is that they allow callus to form more easily between the fracture fragments. The larger the contact area, the faster the fracture have been noted to heal (Schwartz et al., 2013).

Single bone fractures in this study, healed faster than fractures involving more than one bone. Most fractures analysed were femur fractures which have large soft tissue coverage, as opposed to fractures of the tibia and fibula. The importance of adequate soft tissue cover, to ensure healing thus comes into importance. The fracture of the femur in this study group was also among those from the younger age, and as such healed faster.

University of Malaya

A combination of applying RF, DT, SVM, ANN and SOM techniques prove its suitability to be an extremely powerful tool for predicting and selecting significant variables that affects fracture healing time. This can be an extremely powerful tool where the abundance of data obscures straightforward diagnostic reasoning. It is evident from this work that it is possible to create, a compressed data representation, using fracture healing time. A conclusion can be drawn that using such a map in conjunction with presentation with lower limb fracture may be a useful screening mechanism for detecting children with risk with obstructed longer healing time which may require special care. At this stage, it is not possible to claim the results here has universal application. Since it is based upon limited clinical data, and if used within a validation system and continually recreated as more data is collected, it can form a useful tool for placing a patient within a clinical context, allowing consensus to be achieved between clinicians and assessing the particular risk to a patient as well as charting their progress under treatment.

This study was an improvement of the previous study published (Malek et al., 2016) in terms of some limitations encountered. The limitation of this study, however, was the small sample size. The incidence of lower limb fractures in the paediatric population is less than that of the upper limb. A standard analysis requires at least 400 cases (Tseng et al., 2013).

Future enhancement should include more dataset and comparison with upper limb fracture dataset for the paediatric community.

University of Malaya

CHAPTER 6: CONCLUSION

In conclusion the lower limb fracture healing time could be assessed by using machine learning methods. However, application of machine learning methods has not been developed to its full potential especially in paediatric fracture healing. More studies are required to further improve its performance and models created in this study still need to be validated externally using more datasets. Based on the results obtained SVM produces the optimum results compared to other methods. However, it can be concluded a combination of applying RF, DT, SVM, ANN and SOM techniques prove its suitability to be an extremely powerful tool for selecting the most important variables, and for predicating fracture healing time.

The key finding of this study is that applications of machine learning provide an accurate and close to reality estimation for predicting the healing time rate in paediatrics orthopaedic fractures. Therefore this study can be extended to cover the upper limb bone fractures as well as any kind of bone fractures with larger datasets to provide a better diversity on the datasets and to ease perform the analysis

It was hard to refer to or to be guided to other research because there were no study done in paediatric orthopaedic field so this work is considered to be original.

University of Malaya

REFERENCES

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249-2260.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*(1), 69-80.

Audigé, L., Bhandari, M., Hanson, B., & Kellam, J. (2005). A concept for the validation of fracture classifications. *Journal of Orthopaedic Trauma, 19*(6), 404-409.

Audigé, L., Hunter, J., Weinberg, A. M., Magidson, J., & Slongo, T. (2004). Development and evaluation process of a pediatric long-bone fracture classification proposal. *European Journal of Trauma, 30*(4), 248-254.

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research, 9*(Sep), 2015-2033.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* Clinical Research Centre press.

Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry, 26*(1), 5-14.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121-167.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine, 3*(1), 17.

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning, 48*(1-3), 287-297.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge University Press.

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology, 309*(1), 229-240.

University of Malaya

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks. *Cancer, 91*(S8), 1615-1635.

Desai, K. M., Survase, S. A., Saudagar, P. S., Lele, S., & Singhal, R. S. (2008). Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: Case study of fermentative production of scleroglucan. *Biochemical Engineering Journal, 41*(3), 266-273.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*(1), 3.

Dunkler, D., Plischke, M., Leffondré, K., & Heinze, G. (2014). Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PloS one, 9*(11), e113677.

Fagbohungbe, B. O., Akinbode, G. A., & Ayodeji, F. (2012). Organizational determinants of workplace deviant behaviours: An empirical analysis in Nigeria. *International Journal of Business and Management, 7*(5), 207.

Geisser, S., & Johnson, W. (1993). Testing independence of fragment lengths within VNTR loci. *American Journal of Human Genetics, 53*(5), 1103-1106.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225-2236.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P. *European Journal of Epidemiology, 31*(4), 337-350.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician, 63*(4), 308-319.

Gulati, T., Chakrabarti, M., Sing, A., Duvuuri, M., & Banerjee, R. (2010). Comparative study of response surface methodology, artificial neural network and genetic algorithms for optimization of soybean hydration. *Food Technology and Biotechnology, 48*(1), 11-18.

Günther, F., & Fritsch, S. (2010). Neuralnet: Training of neural networks. *The R journal, 2*(1), 30-38.

Hayashi, H., Toribatake, Y., Murakami, H., Yoneyama, T., Watanabe, T., & Tsuchiya, H. (2015). Gait analysis using a support vector machine for lumbar spinal stenosis. *Orthopedics, 38*(11), e959-e964.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks, 1*(Supplement-1), 445-448.

Hekim, M. (2012). ANN-based classification of EEG signals using the average power based on rectangle approximation window. *Electrical Review, 88*(8).

University of Malaya

Hiramatsu, N., Kurosaki, M., Sakamoto, N., Iwasaki, M., Sakamoto, M., Suzuki, Y., . . . Matsuura, K. (2011). Pretreatment prediction of anemia progression by pegylated interferon alpha-2b plus ribavirin combination therapy in chronic hepatitis C infection: decision-tree analysis. *Journal of Gastroenterology, 46*(9), 1111-1119.

Hollmén, J. (1996). Process modeling using the self-organizing map.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). *An introduction to statistical learning* (Vol. 6). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). *An introduction to statistical learning* (Vol. 112). Springer.

Jiang, P., Missoum, S., & Chen, Z. (2014). Optimal SVM parameter selection for non-separable and unbalanced datasets. *Structural and Multidisciplinary Optimization, 50*(4), 523-535.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making, 11*(1), 51.

Kilmer, R. A., Smith, A. E., & Shuman, L. J. (1997). An emergency department simulation and a neural network metamodel. *Journal of the Society for Health Systems, 5*(3), 63-79.

Kim, H., Drake, B. L., & Park, H. (2006). Adaptive nonlinear discriminant analysis by regularized minimum squared errors. *IEEE Transactions on Knowledge and Data Engineering, 18*(5), 603-612.

Kim, S. K., Yoo, T. K., & Kim, D. W. (2013). *Osteoporosis risk prediction using machine learning and conventional methods.* Paper presented at the Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.

Kohonen, T. (1988). Self-Organization and Associative Memory. *Self-Organization and Associative Memory, 100 figs. XV, 312 pages. Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences*.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22.

Macready, W. G., & Wolpert, D. H. (1996). What makes an optimization problem hand? *Complexity, 1*(5), 40-46.

Malek, S., Gunalan, R., Kedija, S., Lau, C., Mosleh, M. A., Milow, P., . . . Saw, A. (2016). *A primary study on application of artificial neural network in classification of pediatric fracture healing time of the lower limb.* Paper presented at the 10th International Conference on Practical Applications of Computational Biology & Bioinformatics.

University of Malaya

Malek, S., Gunalan, R., Kedija, S., Lau, C., Mosleh, M. A., Milow, P., . . . Saw, A. (2017). Random forest and self organizing maps application for analysis of pediatric fracture healing time of the lower limb. *Neurocomputing*.

Malone, C. A., Sauer, N. J., & Fenton, T. W. (2011). A radiographic assessment of pediatric fracture healing and time since injury. *Journal of Forensic Sciences, 56*(5), 1123-1130.

Mantzaris, D. H., Anastassopoulos, G. C., & Lymberopoulos, D. K. (2008). *Medical disease prediction using artificial neural networks.* Paper presented at the BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on. (pp. 1-6). IEEE.

McKibbin, B., & Ralis, Z. A. (1978). The site dependence of the articular cartilage transplant reaction. *The Journal of Bone and Joint Surgery*. *60*(4), 561-566.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition* (Vol.544). John Wiley & Sons.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*: Massachusetts Institute of Technology Press.

Naghsh-Nilchi, A. R., & Aghashahi, M. (2010). Epilepsy seizure detection using eigen-system spectral estimation and multiple layer perceptron neural network. *Biomedical Signal Processing and Control, 5*(2), 147-157.

Navarro, H., & Bennun, L. (2014). Descriptive examples of the limitations of artificial neural networks applied to the analysis of independent stochastic data. *arXiv Preprint arXiv:1404.5598*.

Ogden, J. A. (2000). Injury to the immature skeleton. *Skeletal Injury in the Child*, 38-68.

Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition, 26*(9), 1277-1294.

Principe, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). *Neural and adaptive systems: fundamentals through simulations* (Vol. 672). New York: Wiley.

Provost, F., & Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine Learning, 30*(2), 127-132.

Quinlan, J. R. (1987). *Generating production rules from decision trees.* Paper presented at the ijcai (Vol. 87, pp. 304-307).

Rai, P. (2011). Hyperplane based Classification: Perceptron and (Intro to) Support Vector Machines. *CS5350/6350: Machine Learning September, 8*.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research, 3*(Mar), 1357-1370.

University of Malaya

Riedmiller, M., & Braun, H. (1993). *A direct adaptive method for faster backpropagation learning: The RPROP algorithm.* Paper presented at the Neural Networks, 1993. IEEE International Conference on (pp. 586-591). IEEE.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, *61*(4), 241-254.

Rojas, R. (1996). The backpropagation algorithm *Neural Networks* (pp. 149-182). Springer.

Royston, P., & Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables* (Vol. 777): John Wiley & Sons.

Sackinger, E., Boser, B. E., Bromley, J., LeCun, Y., & Jackel, L. D. (1992). Application of the ANNA neural network chip to high-speed character recognition. *IEEE Transactions on Neural Networks*, *3*(3), 498-505.

Sapthagirivasan, V., & Anburajan, M. (2013). Diagnosis of osteoporosis by extraction of trabecular features from hip radiographs using support vector machine: An investigation panorama with DXA. *Computers in Biology and Medicine, 43*(11), 1910-1919.

Schwartz, M. H., Rozumalski, A., Truong, W., & Novacheck, T. F. (2013). Predicting the outcome of intramuscular psoas lengthening in children with cerebral palsy using preoperative gait data and the random forest algorithm. *Gait & Posture, 37*(4), 473-479.

Shaikh, A. B., Sarim, M., Raffat, S. K., Ahsan, K., Nadeem, A., & Siddiq, M. (2014). Artificial neural network: A tool for diagnosing osteoporosis. *Research Journal of Recent Sciences.*

Simonis, R. B., Parnell, E. J., Ray, P. S., & Peacock, J. L. (2003). Electrical treatment of tibial non-union: a prospective, randomised, double-blind trial. *Injury*, *34*(5), 357-362.

Slongo, T., Audigé, L., Schlickewei, W., Clavert, J. M., & Hunter, J. (2006). Development and validation of the AO pediatric comprehensive classification of long bone fractures by the Pediatric Expert Group of the AO Foundation in collaboration with AO Clinical Investigation and Documentation and the International Association for Pediatric Traumatology. *Journal of Pediatric Orthopaedics*, *26*(1), 43-49.

Staheli, L. T. (2008). *Fundamentals of pediatric orthopaedics*. Lippincott Williams & Wilkins.

Stiglic, G., Kocbek, S., Pernek, I., & Kokol, P. (2012). Comprehensive decision tree models in bioinformatics. *PLOS ONE, 7*(3), e33812.

Subasi, A., & Ercelebi, E. (2005). Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine, 78*(2), 87-99.

University of Malaya

Suykens, J. A., Van Gestel, T., & De Brabanter, J. (2002). *Least squares support vector machines*. World Scientific.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences, 43*(6), 1947-1958.

Tseng, W. J., Hung, L. W., Shieh, J. S., Abbod, M. F., & Lin, J. (2013). Hip fracture risk assessment: artificial neural network outperforms conditional logistic regression in an age-and sex-matched case control study. *BMC Musculoskeletal Disorders*, *14*(1), 207.

Umadevi, N., & Geethalakshmi, S. (2012). *Multiple classification system for fracture detection in human bone x-ray images.* Paper presented at the Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on (pp. 1-8). IEEE.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988-999.

Ventura Dde, F., Ganato, L., Mitre, E. I., & Mor, R. (2009). Oculomotricity parameters in digital nystagmography among children with and without learning disorders. *Brazilian Journal of Otorhinolaryngol*, *75*(5), 733-737.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330-349.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2011). Linear, logistic, survival, and repeated measures models. Springer-Verlag. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models* (pp. 340): Springer Science & Business Media.

Wang, Z., Wen, X., Lu, Y., Yao, Y., & Zhao, H. (2016). Exploiting machine learning for predicting skeletal-related events in cancer patients with bone metastases. *Oncotarget*, *7*(11), 12612.

Wu, H. Y., Gong, C. S. A., Lin, S. P., Chang, K. Y., Tsou, M. Y., & Ting, C. K. (2016). Predicting postoperative vomiting among orthopedic patients receiving patient-controlled epidural analgesia using SVM and LR. *Scientific Reports, 6*, 27041.

Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*(Aug), 975-1005.

Ye, G. B., Chen, Y., & Xie, X. (2011). *Efficient variable selection in support vector machines via the alternating direction method of multipliers.* Paper presented at the Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics.

University of Malaya

Zhao, W., Adolph, A. L., Puyau, M. R., Vohra, F. A., Butte, N. F., & Zakeri, I. F. (2013). Support vector machines classifiers of physical activities in preschoolers. *Physiol Report, 1*(1).

Zorman, M., Eich, H.-P., Kokol, P., & Ohmann, C. (2001). Comparison of three databases with a decision tree approach in the medical field of acute appendicitis. *Studies in Health Technology and Informatics*(2), 1414-1418.

University of Malaya

LIST OF PUBLICATIONS AND PAPERS PRESENTED

1. Malek, S., Gunalan, R., Kedija, S., Lau, C., Mosleh, M. A., Milow, P., . . . Saw, A. (2016). *A primary study on application of artificial neural network in classification of pediatric fracture healing time of the lower limb.* Paper presented at the 10th International Conference on Practical Applications of Computational Biology & Bioinformatics.

2. Malek, S., Gunalan, R., Kedija, S., Lau, C., Mosleh, M. A., Milow, P., . . . Saw, A. (2017). Random forest and self organizing maps application for analysis of pediatric fracture healing time of the lower limb. *Neurocomputing*.

University of Malaya