MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

IMPROVING EFFICIENCY OF THE SOLUTIONS FOR CLASS IMBALANCE PROBLEMS USING DATA MINING TECHNIQUES

IZHAN FAKHRUZI

MASTER THESIS
Department of Computer Engineering

Thesis Supervisor
Dr. Öğr. Üyesi Betül Demiröz BOZ

ISTANBUL, 2018

MARMARA UNIVERSITY INSTITUTE FOR GRADUATE STUDIES IN PURE AND APPLIED SCIENCES

Izhan FAKHRUZI, a Master of Science student of Marmara University Institute for Graduate Studies in Pure and Applied Sciences, defended his thesis entitled "**Improving Efficiency of The Solutions for Class Imbalance Problems Using Data Mining Techniques**", on July 2, 2018 and has been found to be satisfactory by the jury members.

Jury Members

Dr. Öğr. Üyesi Betül Demiröz BOZ (Advisor)
Marmara University .....

Dr. Öğr. Üyesi Murat Can GANİZ (Jury Member)
Marmara Üniversitesi .....

Dr. Öğr. Üyesi Berna KİRAZ (Jury Member)
Fatih Sultan Mehmet Üniversitesi .....

APPROVAL

Marmara University Institute for Graduate Studies in Pure and Applied Sciences Executive Committee approves that Izhan FAKHRUZI be granted the degree of Master of Science in Department of Computer Engineering, Computer Engineering Program on 09.07.2018., (Resolution no: 2018/16-07).

Director of the Institute
Prof. Dr. Bülent EKİCİ

ACKNOWLEDGMENTS

In the name of Allah, the Most Gracious, the Most Merciful. This thesis completion would have been impossible without the strengths and the blessings He has given me.

I would like to express my special gratitude to my advisors Assistant Prof. Dr. Betül Demiröz BOZ for her supervision throughout the completion of the thesis and Assistant Prof. Dr. Mustafa Ağaoğlu for his guidance in the beginning of my thesis writing.

My deepest gratefulness to Prof. Dr. Haluk Rahmi Topçuoğlu, all teachers and the Institute of Pure and Applied Sciences who have given me invaluable knowledge and experiences during my master study in Marmara University.

I am highly indebted to my parents, wife and siblings for their endless love, prayers and understanding who have made a difficult time easier.

I owe a great thanks to all friends, Indonesian students in Turkey, and friends in Eyüp who kindly helped and encouraged me during my study here and to those who indirectly contributed in this research. Thank you very much.

CONTENTS

ÖZET

VERI MADENCILIĞI TEKNIKLERINI KULLANARAK SINIF DENGESIZLIK PROBLEMLERI İÇIN ÇÖZÜMLERIN VERIMLILIĞININ ARTIRILMASI

Klinik karar almada yanlış bir teşhis hastanın yaşamına zarar verebilir. Bu sebeple veri madenciliğinin sağlık sektörüne uygulanmasındaki kayda değer artış ölçüm doğruluğunu klinik teşhis öngörüsünde kritik performans ölçümlerinden biri haline getirmektedir. Bununla birlikte sınıf dengesizliği problemi yaygın olarak klinik veri kümelerini sıkıntıya sokmaktadır. Bu hal, veri kümelerindeki sınıflar eşitsiz biçimde ortaya konduğunda meydana gelmektedir. Bu durum algoritmaların verilerle overfitting uyuşmazlığına sebep olan ve klinik öngörüde zayıf doğruluk veren sinirsel ağ algoritmalarının işlerliğini azaltmaktadır. Torbalama metodu sınıf dengesizliği problemine yaklaşım becerisine sahip ve ölçme doğruluğunu artıran yaygın kümeleme metodlarından biridir. Bunun yanısıra torbalama metodu kararsız kümeleyicilerde olumlu biçimde işlemektedir. Kararsız kümeleyicilerden biri de sinirsel ağlardır. Bu sebeple bu çalışmada, yukarıdaki probleme yaklaşım konusunda torbalama tabanlı sinirsel ağ öne sürülmektedir. Deneysel sonuçlara göre bu yöntem doğru ölçmede konvansiyonel sinirsel ağdan daha iyi sonuç vermekte ve klinik teşhis öngörüsünde sınıf dengesizliği problemine başarılı bir yaklaşım sergileyebilmektedir.

Anahtar kelimeler: sınıf dengesizliği problemi; torbalama; sinirsel ağlar

ABSTRACT

IMPROVING EFFICIENCY OF THE SOLUTIONS FOR CLASS IMBALANCE PROBLEMS USING DATA MINING TECHNIQUES

In clinical decision making, an inaccurate diagnosis might harm patient's life. Therefore, the significant growth of data mining's implementation in healthcare industry takes the accuracy into one of the critical performance measures for clinical diagnosis prediction. However, clinical datasets commonly suffer from class imbalance problem. It occurs when the classes in the datasets are unequally presented. This situation degrades the performance of neural network algorithms which leads the algorithms to overfit the data and have poor accuracy in clinical prediction. Bagging method is one of the popular ensemble methods that is capable to address class imbalance problem and improve the accuracy. Furthermore, bagging method performs well with unstable classifiers. One of the unstable classifiers is neural networks. Therefore, bagging based neural network is proposed to address the above problem. From the experimental results, the proposed method achieves better accuracy than the conventional neural network and successfully addresses class imbalance problem on clinical diagnosis predictions.

Keywords: class imbalance problem; bagging; neural networks

SYMBOLS

v'	**:** New value of dataset in min-max normalization
v	**:** Original value of dataset in min-max normalization
min_A	**:** Minimum value of an attribute in dataset
max_A	**:** Maximum value of an attribute in dataset
new_{max_A}	**:** New maximum value, usually set as 1
new_{min_A}	**:** New minimum value, usually set as 0
$f(x)$	**:** The rectified linear unit function
k	**:** The number of folds in cross validation
TP	**:** True Positive, correctly identified result in classification
TN	**:** True Negative, correctly rejected result in classification
FP	**:** False positive, incorrectly identified result in classification
FN	**:** False negative, incorrectly rejected result in classification

ABBREVIATIONS

NN : Artificial Neural Network

BNN : Bagging Neural Network

UCI : University of California Irvine

AUC : Area Under ROC Curve

Bagging : Bootstrap Aggregating

CAD : Computer Aided Decision

PSO : Particle Swarm Optimization

MIoT : Medical Internet of Things

SMOTE : Synthetic Minority Over-Sampling Technique

ReLU : Rectified Linear Unit

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

In recent years, data mining techniques have been significantly applied in healthcare industry to assist a physician in giving a correct diagnosis to a patient (Jothi, Rashid, & Husain, 2015; Mazurowski et al., 2008). In making clinical decision for predicting diseases or diagnosing a patient, the accuracy is one of the critical performance measures since it leads to the treatment to the patient. Misdiagnosis might bring the loss of financial cost for the therapies and thread the patient's life (Han & Kamber, 2006; Zhou & Liu, 2006). Therefore, it is imperative that a correct prediction in making clinical decision has to be done to treat the patient properly and save their lives.

Many studies show that data mining techniques are able to improve the performance of clinical prediction models such as the study on resurgery prediction in intensive care using data mining approaches (Peixoto, Ribeiro, Portela, Filipe Santos, & Rua, 2017), Decision Tree to predict the type of birth through pregnancy characteristics (Pereira, Portela, Santos, Machado, & Abelha, 2015), Bagging C4.5 algorithm to support wise clinical decision-making in the healthcare industry (Lee, Xu, Li, & Yang, 2017). Specifically, neural network learning algorithms are popularly used for medical decision making and proved to have better accuracy (Han & Kamber, 2006; Mazurowski et al., 2008).

However, several weaknesses might encounter neural networks when the dataset with two class classification is not equally presented. One class has a large number of data than the others. This class imbalance problem is a common thing in medical records or clinical datasets, for example, the number of patients who has cancer is represented by only a few number while the other is represented by a large number (Jothi et al., 2015; Mazurowski et al., 2008; Zhou & Liu, 2006). This situation causes negative effects on the neural network's performance that can lead the algorithm to overfit the data and have poor accuracy (Fan, Wang, & Gao, 2016; Huang, Hung, & Jiau, 2006; Zhou & Liu, 2006).

There are a lot of approaches have been conducted by researchers to address class imbalance problem. One of the popular and effective approaches to handle the problem is ensemble method (Fan et al., 2016; Han & Kamber, 2006; Zhou & Liu, 2006). In

ensemble methods, boosting and bagging are two popular ways that have been proven to improve the accuracy of prediction models or learning algorithm (Han & Kamber, 2006; Kim & Kang, 2010; Mazurowski et al., 2008; Setiyorini & Wahono, 2014). However, bagging method gives better performance to class imbalance problem than boosting (Zhou & Liu, 2006). Moreover, bagging shows well performance with unstable classifier such as neural network (Collell, Prelec, & Patil, 2017). Therefore, in this study, bagging method is implemented in neural network to address class imbalance problem on clinical diagnosis predictions to gain more accurate results.

1.2.Statement of the Problem

Neural network learning algorithms show better prediction accuracy for medical decision making, however neural network learning algorithms have weaknesses when datasets with two class classification is not equally presented, popularly known as class imbalance problem, so that it leads the algorithm to overfit the data and to give poor accuracy.

1.3.Research Questions

How accurate is neural network if bagging method is implemented to handle class imbalance problem?

1.4.Purpose of the Study

- To develop a proposed method to address class imbalance problem on clinical prediction by applying ensemble strategy that is bagging based neural network.
- To develop a data mining technique of two class classification for predicting patients' health condition.
- To implement bagging method to address class imbalance problem on neural network to gain more accurate prediction.

1.5.Significance of the Study

- To prove the neural network ensemble produces more accurate predictions than the conventional neural network model particularly for clinical prediction.
- To give a contribution to the healthcare industry to get more accurate prediction for clinical decision making by using the proposed method.

1.6.Procedures

Materials

The experiments are conducted by using a computing platform based on Intel(R) Core(TM) i7-4510U CPU @ 2.60GHz, 8 GB RAM, and Microsoft Windows 8.1 Pro 64-bit operating system. The programming language used is Python 3.6.3.

Methodology

The research method used in this study is an experimental research method. As shown in Figure 1.1, the stages of this study are conducted as follows:

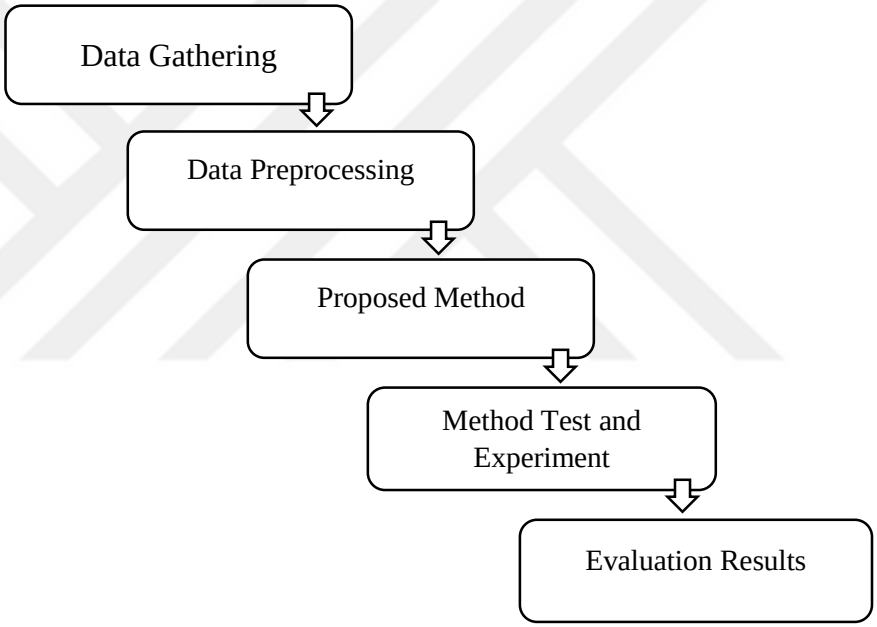

Figure 1.1. Stages of the Research Process

1.7.Limitations of the Study

This study focuses only on neural network with bagging method in clinical prediction using public data from University of California Irvine (UCI) Machine Learning Repository as testing and training data.

1.8.Organization of Study

This thesis is organized as follows: Chapter 2 provides related researches and literature reviews. Chapter 3 explains the methodology to conduct the study from the

data gathering stage to the evaluation result stage. Chapter 4 demonstrates the performance of the conventional neural networks and the proposed method, bagging neural networks, to predict clinical diagnosis as experimental results. Finally, conclusion and recommendation are provided in Chapter 5.

2. LITERATURE REVIEW

2.1.Related Works

Many studies show that data mining techniques are able to improve the performance of clinical prediction or assist clinical decision making. For example in maternity care, data mining techniques are applied to predict the type of delivery by Pereira et al. (2015). The purpose of the study is to find the most suitable delivery technique to the pregnant women whether to get normal delivery or caesarean section in advance by identifying the obstetric risk factors. The correct prediction assists the physicians in decision making process that is able to avoid misdiagnosis and wrong treatment to the pregnant women. Besides, the maternity care unit is able to give better services and safety to mother and child. In the study, four data mining techniques are implemented. From the study, data mining techniques give satisfactory results to recommend appropriate delivery type predictions, particularly Decision Tree which has higher accuracy than the others.

Nevertheless, accuracy in clinical prediction can be weaken by imbalanced dataset (Mazurowski et al., 2008). Class imbalance problem is typical characteristic among medical data where one class is underrepresented or has fewer number of samples than the other class. The study from Mazurowski et al. (2008) investigates the effect of class imbalance problem in clinical dataset towards neural network learning algorithms for computer aided decision (CAD) systems. In the study, CAD systems is the use of computer algorithm to assist a decision maker in giving recommendation to a physician for diagnosing a patient. Two sampling methods, oversampling and undersampling methods, are used to address imbalanced dataset. The purpose of the study is to investigate the effect of class imbalance problem on neural network learning algorithms. Therefore, two different neural network methods are implemented, backpropagation and particle swarm optimization (PSO), to see the effect of the class imbalance problem. Then, Area Under ROC Curve (AUC) is implemented to evaluate the classifier performances. It shows that the neural network learning algorithms' performances are degraded with even modest imbalanced dataset. However, backpropagation performance outperforms the PSO.

Other study proposes a novel bagging C4.5 algorithm based on wrapper feature selection to assist a physician on clinical decision making toward the high-dimensional

and high-uncertain data by Lee et al., (2017). Those data characteristics are typical data generated by Medical Internet of Things (MIoT). Therefore, Synthetic Minority Over-Sampling Technique (SMOTE) is applied to achieve better sampling in an attempt to reduce the size of dataset and data distortion. Then, the Wrapper method is applied to omit unnecessary features to achieve high-impact features. Afterwards the dataset is ready to be implemented into the learning algorithm. In the study, C4.5 algorithm is applied with an ensemble method, bagging, to improve the accuracy. From the study, the proposed method shows satisfactory performance compare to the other selected data mining techniques.

2.2.Literature Review

2.2.1. Healthcare Industry

The healthcare industry or medical industry consists of various industries, from profit to non-profit organizations that provides medical services and all related medical activities, such as (Jothi et al., 2015; Madadipouya, 2015):

1. Drugs
2. Medical Equipment
3. Medical Insurance
4. Healthcare Facilities

The healthcare industry is developing significantly at a rapid growth and considered as one of the largest industries in the world. The healthcare industry is a place with a huge amount of data that is collected from its operation such as administrative reports, electronic medical records, and so on (Jothi et al., 2015).

Clinical data is a collection of databases with a list of well-defined features or attributes that are related to clinical activities and operations. Clinical data is divided into six primary types (Health Sciences Library, 2018):

1. Patient or Disease registries
2. Electronic medical records
3. Claims data
4. Administrative data
5. Clinical trials data
6. Health surveys

The collected data has become priceless property for healthcare industry that serves as an input into important clinical decision making processes, such as diseases prediction, medical diagnosis and treatment of diseases. Therefore the implementation of data mining has been significantly increasing in an attempt to extract the data to find valuable information (Jothi et al., 2015; Madadipouya, 2015; Mazurowski et al., 2008; Ranjan & Kumar, 2016).

2.2.2. Neural Networks

Neural network learning algorithms are inspired by neural cells of human in processing information. Neural networks is one of the prominent data mining tools for classification and clustering. It is described as a structure of connected layers, comprises input, hidden, and output layers in which each layer has number of nodes and number of weight associated with every connection (Setiyorini & Wahono, 2014). In the learning phase, in order to be capable of predicting an accurate output, the neural network learns by modifying internal weights of each layers to produce an intended output (Han & Kamber, 2006). Figure 2.1. depicts the standard network architecture of neural network model .

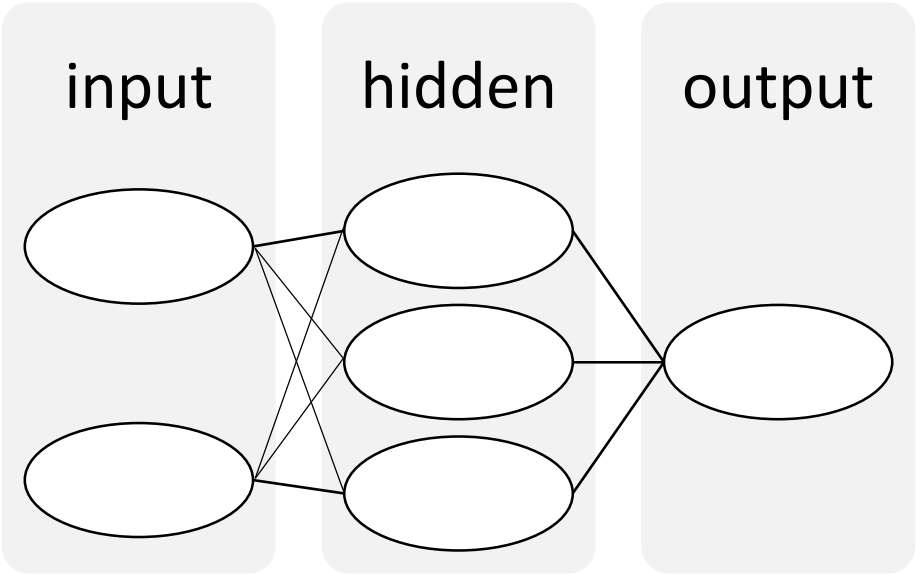

Figure 2.1. Neural Network's Architecture

2.2.3. Backpropagation Algorithm

The backpropagation algorithm is one of the artificial neural network algorithms. The algorithm is a supervised learning algorithm that learns by adjusting the weight in response to the calculated error. A standard network architecture is depicted in Figure 2.1., comprises three layers, one input layer, one hidden layer, and one output layer (Han & Kamber, 2006). The implementation of the backpropagation neural network consists of two phases (Cilimkovic, 2010):

1. Training or learning phase, in this phase dataset is processed into the algorithm in order the algorithm to learn and get the intended output.
2. Testing phase, this phase is done after the learning phase reach the intended output to test the model.

Principally, during the learning phase, the backpropagation neural network can be summarized into three main steps as follows (Cilimkovic, 2010):

1. Feedforward computation from input to output layer.
2. Backpropagation computation from output to input layer.
3. Weight adjustments or weight updates.

If the stop condition or the intended output is fulfilled, the algorithm is stopped.

2.2.4. Bagging Method

Bootstrap aggregating, commonly known as bagging, is a powerful ensemble method that is able to improve the accuracy of base classifiers and easy to be implemented. Bagging method is a technique that trains base classifiers in the ensemble using different sample data which is randomly drawn from the given dataset. Then, each classifier predicts its class prediction from each bootstrap of data and is combined together as a vote with equal weight and assigns the major vote as an accurate prediction. Bagging demonstrates well performance and accuracy with unstable base classifier like neural networks and is stronger towards the effects of noisy data and overfitting (Collell et al., 2017; Han & Kamber, 2006; Kim & Kang, 2010). From Figure 2.2. bagging method can be summarized as follows:

1. Sample data is randomly drawn from the training dataset and put into a set of bootstraps of data.

2. Each bootstrap of data is then proceeded by each classifier.
3. As a composite model, it returns all prediction of classifier as a majority vote.

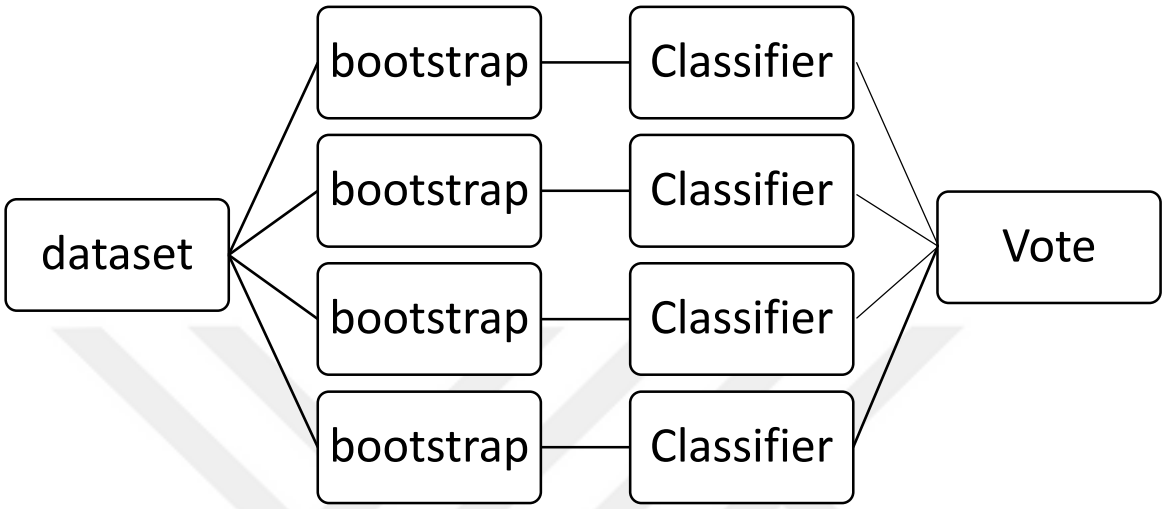

Figure 2.2. Bootstrap Aggregating (Bagging) Scheme

2.2.5. K-Fold Cross Validation

K-Fold Cross Validation is one of the common techniques for evaluating classifier accuracy by splitting the dataset into k equal size with randomly selected sample of data from the dataset as training and testing data (Han & Kamber, 2006). From Figure 2.3. k-fold cross validation can be summarized into five steps below:

1. The given dataset is partitioned into *k* number of folds with equal size of randomly selected sample data from the given dataset.
2. Proceed one fold as a testing dataset and the rest (*k*-1) as training dataset.
3. Apply the learning model to each fold of data
4. Repeat *k* times in which each data fold get its turn to be the testing data exactly once.
5. Finally the accuracy of *k* results from all folds is averaged as a composite result.

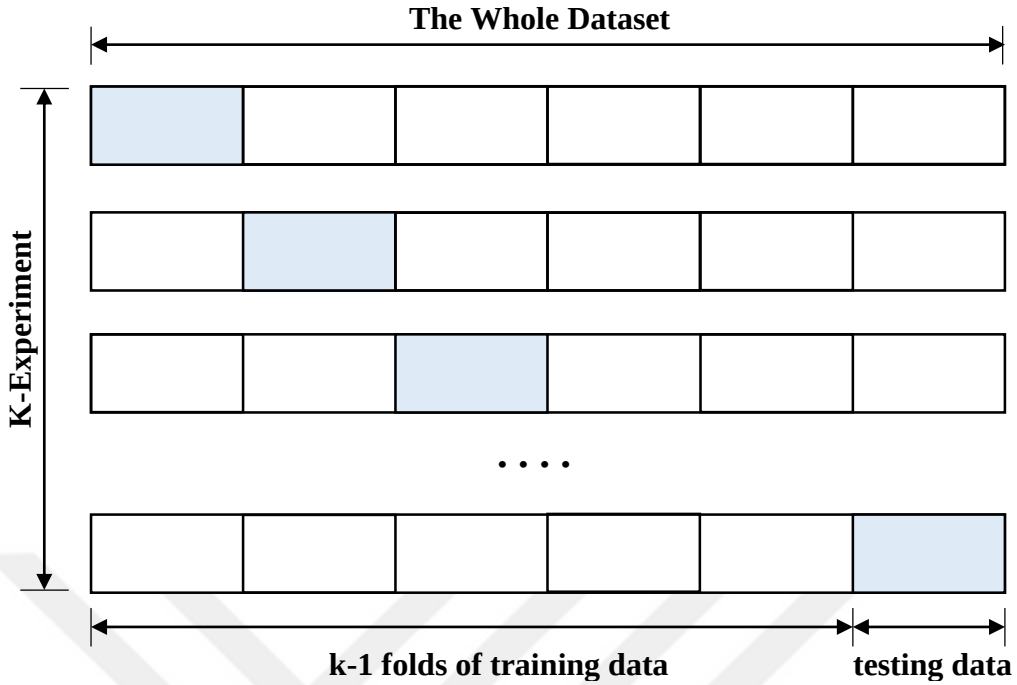

Figure 2.3. K-Fold Cross Validation Scheme

3. METHODOLOGY

3.1.Research Design

The research method used in this study is an experimental research method. The research stages are conducted as follows:

1. Data Gathering

   The datasets used in this study are collected from UCI Machine Learning Repository (Dua et al., 2017). They are Diabetes Dataset, Breast Cancer Dataset, and Patient Dataset. These datasets are clinical datasets with class imbalance problem.

2. Data Preprocessing

   In the preprocessing step, the selected datasets pass through the data cleaning and the data transformation processes before the datasets are ready to be proceeded in the learning algorithm models.

3. Proposed Method

   The proposed method in this study is neural networks with bagging method to address class imbalance problem on clinical diagnosis predictions in order to gain more accurate results. In particular, the neural network learning algorithm which is implemented in this study is backpropagation algorithm using Rectified Linear Unit (ReLU) activation function.

4. Method Test and Experiment

   In this step, each different topology of the neural networks and the proposed method are set and compared to measure the performance of both models. There are 10 different topologies are used in the experiment for each neural network model.

5. Evaluation Result

   After the experiments are conducted and compared for both models, then the performances are analyzed and evaluated based on the parameters and datasets used using area under ROC curve.

3.2.Data Gathering

The dataset source for this study is from UCI Machine Learning Repository. The repository is mainly used as a dataset source for machine learning algorithm by a large number of educators, students, and researchers. Besides, its collection of databases is

an open access (Dua, Karra Taniskidou, 2017). There are 3 datasets selected as representative of clinical datasets with class imbalance problem in this study.

a. Diabetes Dataset: This dataset is originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset has 768 patient records that consists of 500 non diabetes and 268 diabetes patient records. The objective is to predict the patient's diabetes condition based on diagnostic measurement.
b. Breast Cancer Dataset (Mangasarian, Wolberg, 1990). The dataset is from the University of Wisconsin Hospitals, Madison. This dataset has 699 patient records that consists of 458 non breast cancer and 241 breast cancer patient records. The objective is to detect breast cancer on patients based on diagnostic measurements.
c. Liver Dataset. The dataset was gathered from India, particularly at north east of Andhra Pradesh. This dataset has 583 patient records that contains 416 liver patient and 167 non liver patient records. The objective is to predict liver disease on patient based on diagnostic measurements.

These selected datasets are clinical datasets with uneven class distribution or called class imbalance problem. The detail descriptions of the datasets are summarized in Table 3.1.

Table 3.1. UCI datasets used in the experiment

Datasets	**Size**	**Attribute**	**Class**	**Class Distribution**
Diabetes	768	8	2	500/268
Breast Cancer	699	10	2	458/241
Liver	583	10	2	416/167

3.3.Data Preprocessing

In the preprocessing step, the selected datasets are not ready to be applied into the learning algorithm models due to its missing values, invaluable attributes, and uneven range of values for some attributes. Therefore, the raw datasets pass through the data cleaning and the data transformation processes steps before ready to be implemented. As it is shown in Figure 3.1, the processes as follows:

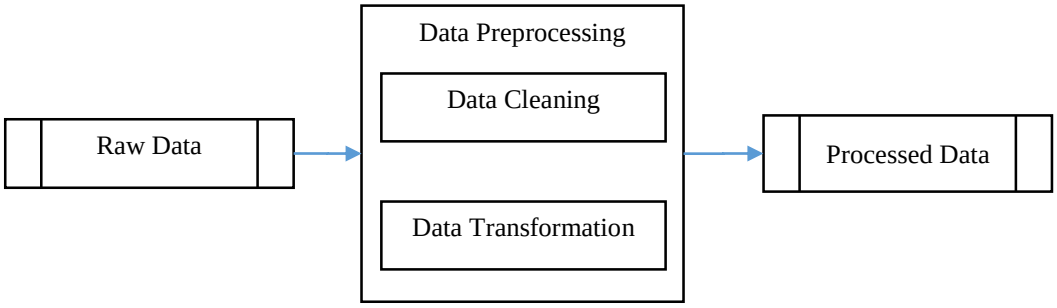

Figure 3.1. Data Preprocessing steps

a. In the data cleaning process, less-impact attributes are omitted and missing values are filled with appropriate values (Han & Kamber, 2006). In this study, the missing data in the selected dataset is filled with mean value and some attributes with a lot of missing or incomplete data are dropped.

   Summary statistics can help to identify missing or corrupt values in the dataset on each attribute. The summary statistics consists of the count, mean, the min and max values as shown on the figures below.

   - Diabetes Dataset

     Min attribute from statistics can be used as an indication if the attributes consist of missing values. As shown in Figure 3.2. there are minimum of zero values on most of the attributes. That indicates missing values or corrupt values in diabetes dataset.

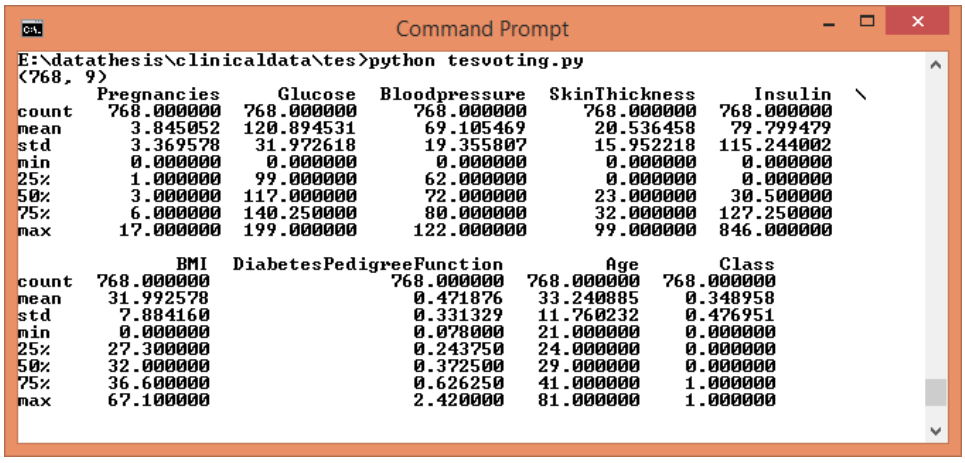

Figure 3.2. Statistical Summary in Diabetes Dataset

Figure 3.3 shows a clear statistics about the number of missing values on each attribute. It shows that there are a large number of missing values on each attribute on diabetes dataset. These missing values need to be proceeded in data cleaning process.

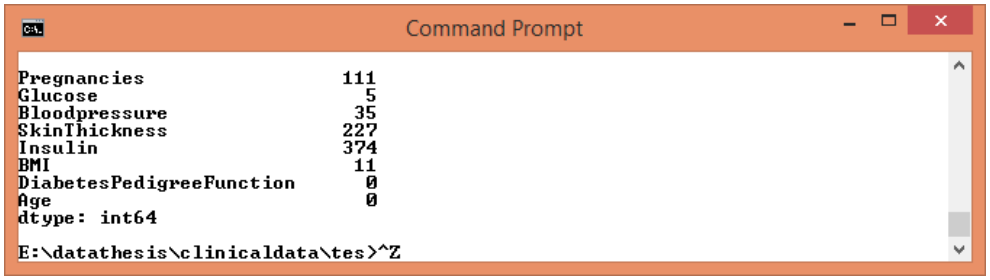

Figure 3.3. Missing Value Statistics on Each Diabetes Dataset Attributes

- Breast Cancer Dataset

  As shown in Figure 3.4. there is no minimum of zero values on the attributes. This indicates diabetes dataset has a good quality of data that does not suffer from missing values.

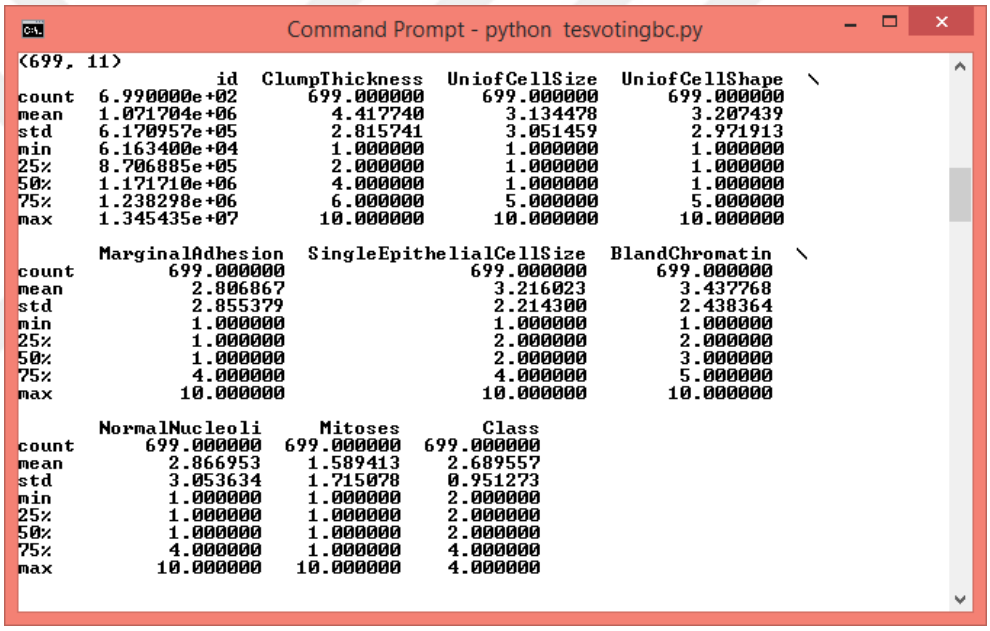

Figure 3.4. Statistical Summary in Breast Cancer Dataset

Figure 3.5 shows a clear statistics about the number of missing values on each attribute. It shows that each attribute on breast cancer dataset has complete data.

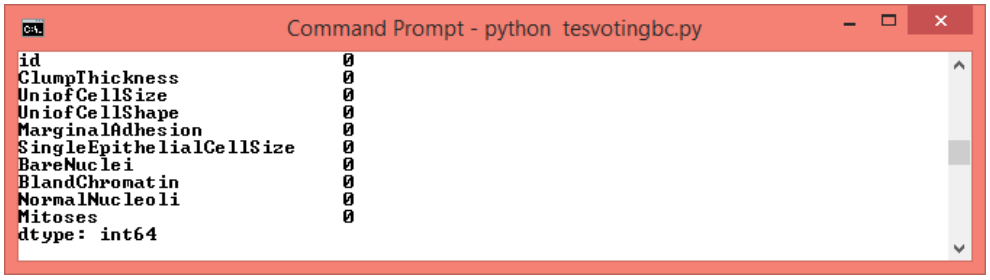

Figure 3.5. Missing Value Statistics on Each Breast Cancer Dataset Attributes

- Liver Dataset

  Similar to breast cancer dataset, as shown in Figure 3.6. there is no minimum of zero values on the attributes on liver dataset. This indicates diabetes does not suffer from missing values. However, the maximum values on some attributes show high values that indicates high variance in data distribution. This can be fixed in data transformation process.

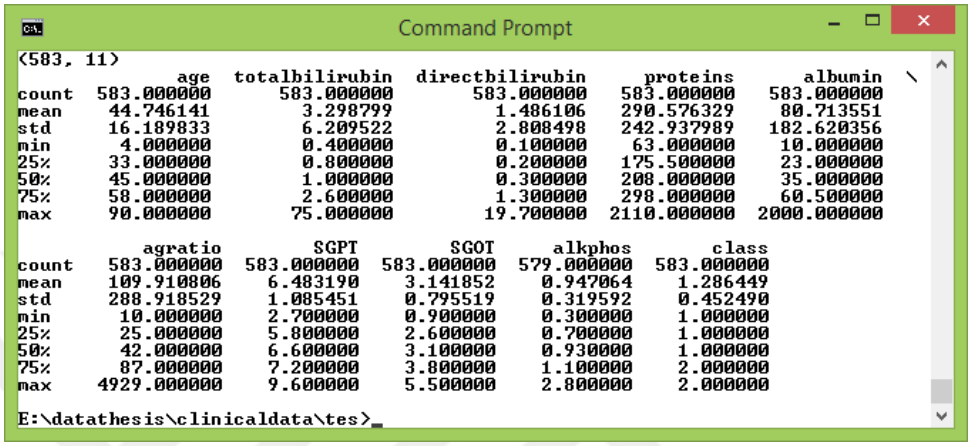

Figure 3.6. Statistical Summary in Liver Dataset

Figure 3.7 shows a clear statistics about the number of missing values on each attribute. It shows that each attribute on liver data set has complete data.

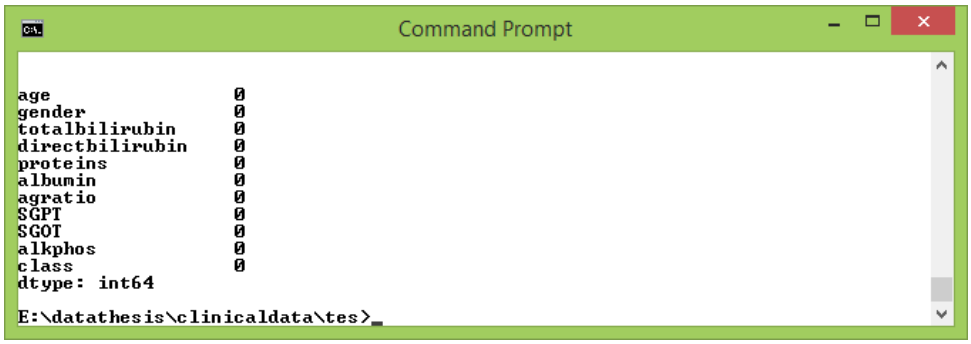

Figure 3.7. Missing Value Statistics on Each Liver Dataset Attributes

b. Afterwards, the dataset is normalized with min-max normalization with range from 0 to 1 as in formula 3.1. Data transformation process convert the original data into appropriate values to be proceeded in the data mining (Han & Kamber, 2006).

c. $$v' = \frac{v - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new_{min_A} \quad (3.1)$$

Before conducting data transformation process, boxplot chart can help to show the minimum and the maximum values of each attribute to see the high variance in the data distribution. The boxplot chart as shown on the figures below.

- Diabetes Dataset

  In Figure 3.8 Insulin attribute shows high variance with maximum of 800 value. However, the other attributes show relatively even distribution.

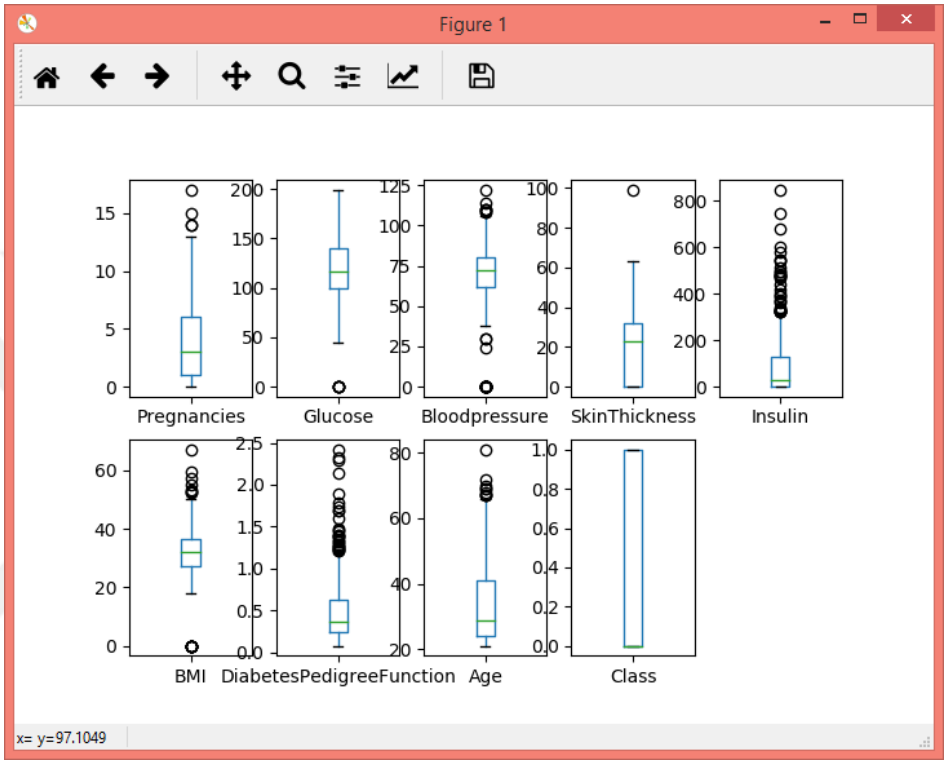

Figure 3.8. Boxplot Chart of Diabetes Dataset

Therefore, min-max normalization is done to have even distributions. After normalization, minimum and maximum values on each attribute has range from 0 to 1 as shown in Figure 3.9.

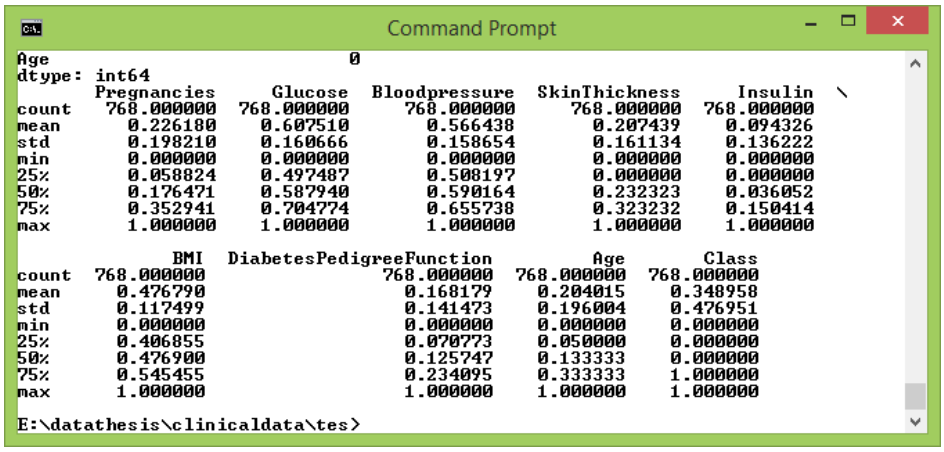

Figure 3.9. Statistical Summary in Diabetes Dataset After Transformation

- Breast Cancer Dataset

  From the boxplot char in Figure 3.10 shows that all attributes in breast cancer dataset has an even distribution with range 1 to 10.

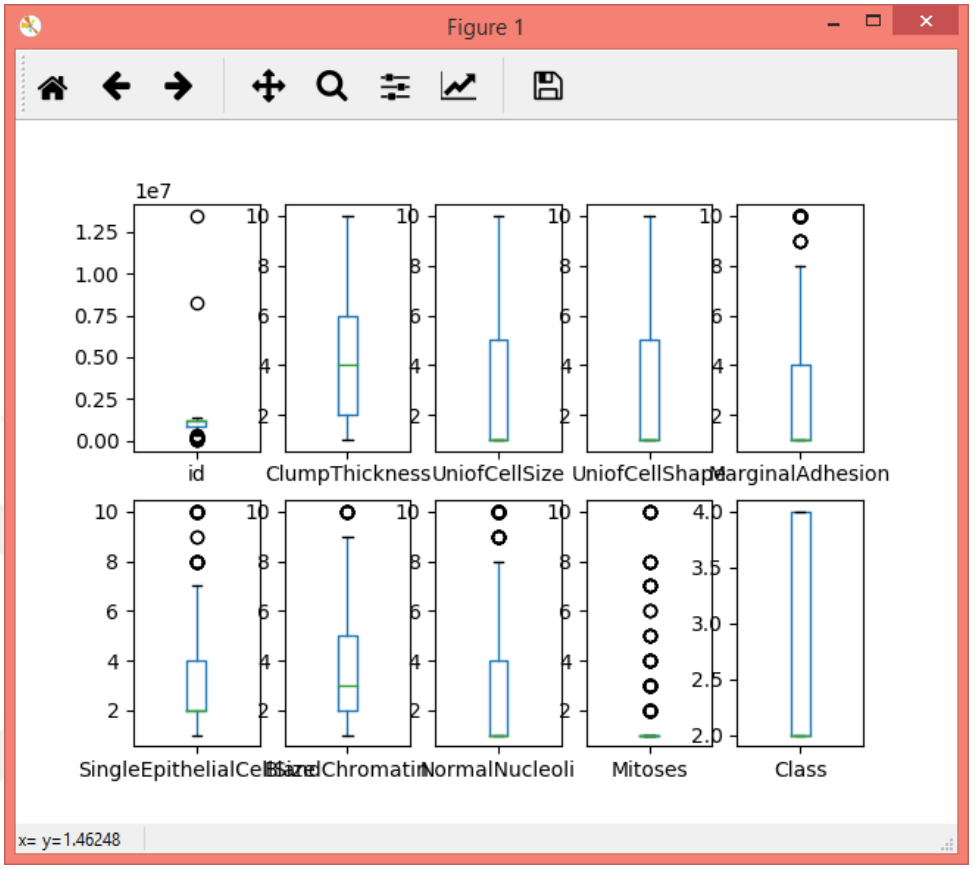

Figure 3.10. Boxplot Chart of Breast Cancer Dataset

However, this study uses min-max normalization with range from 0 to 1 to get optimized result from the learning method. After normalization, minimum and maximum values on each attribute has range from 0 to 1 as shown on the following figure.

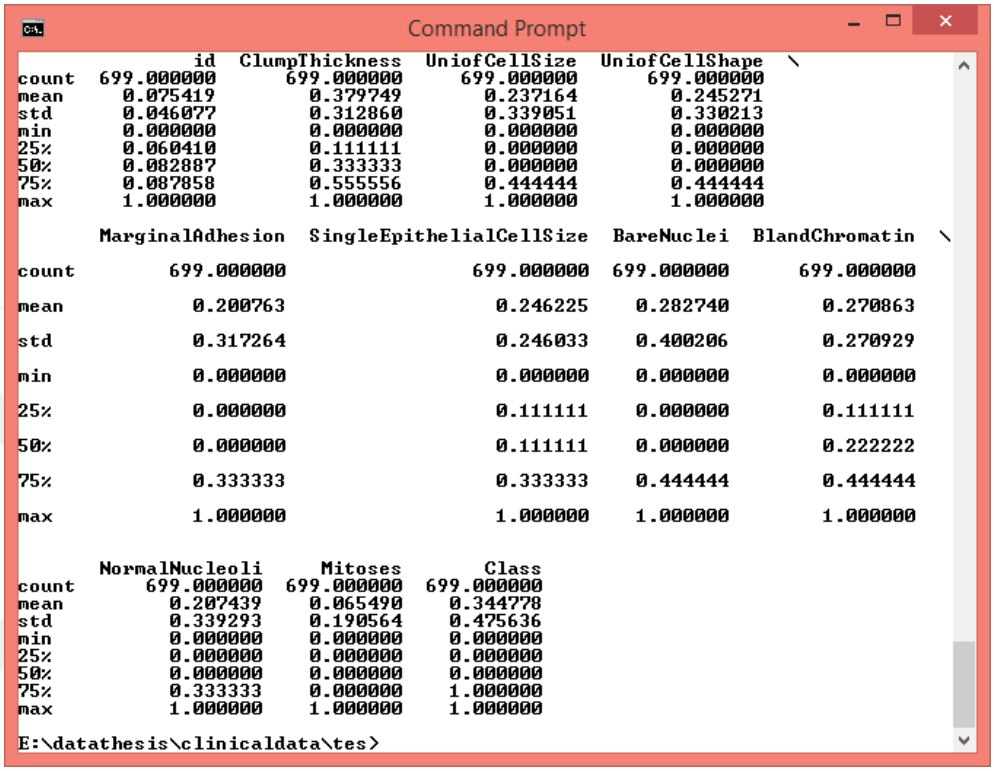

Figure 3.11. Statistical Summary in Breast Cancer Dataset After Transformation

- Liver Dataset

  In Figure 3.12 some attributes shows high variance that reach 5000 in maximum. Min-max normalization needs to be done in order the dataset to have an even distribution.

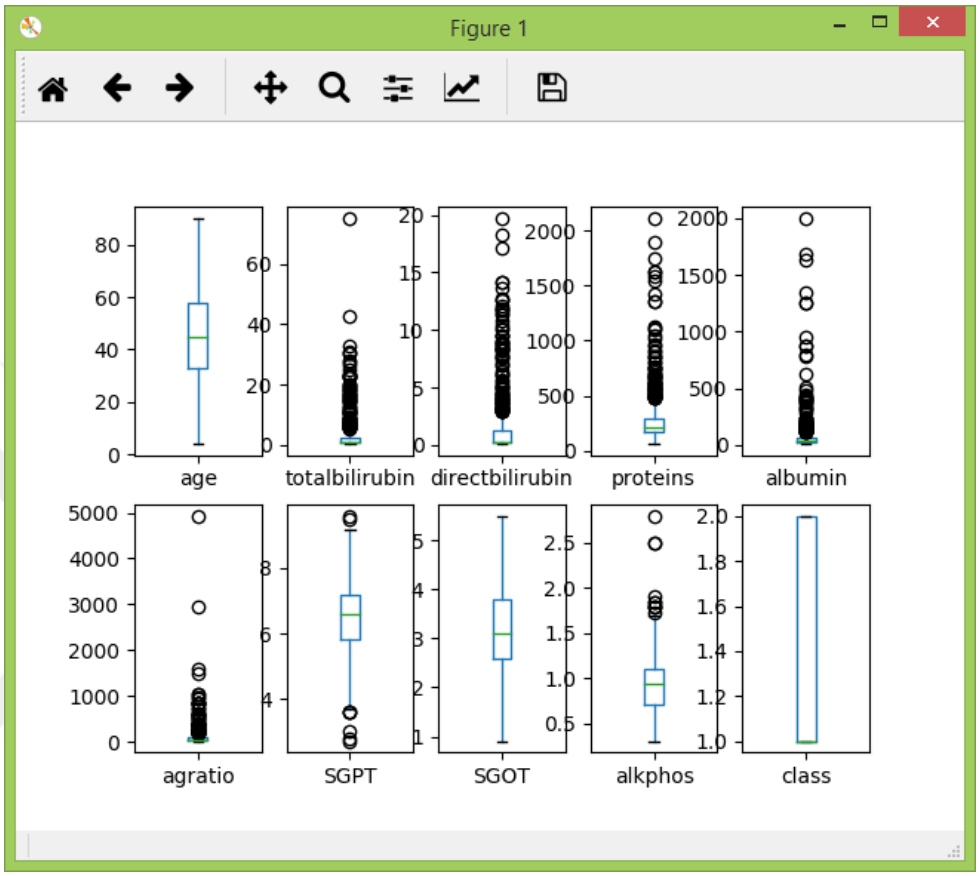

Figure 3.12. Boxplot Chart of Liver Dataset

After normalization, minimum and maximum values on each attribute has range from 0 to 1 as shown in Figure 3.13.

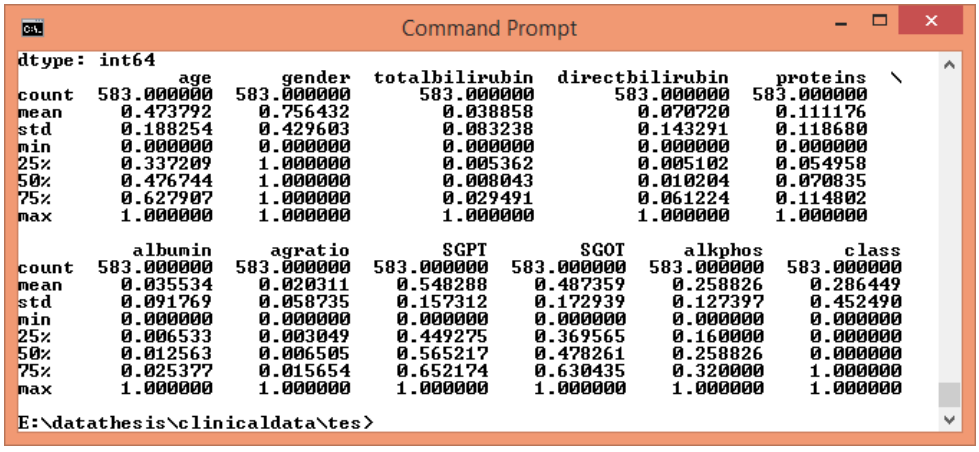

Figure 3.13. Statistical Summary in Liver Dataset After Transformation

3.4.Proposed Method

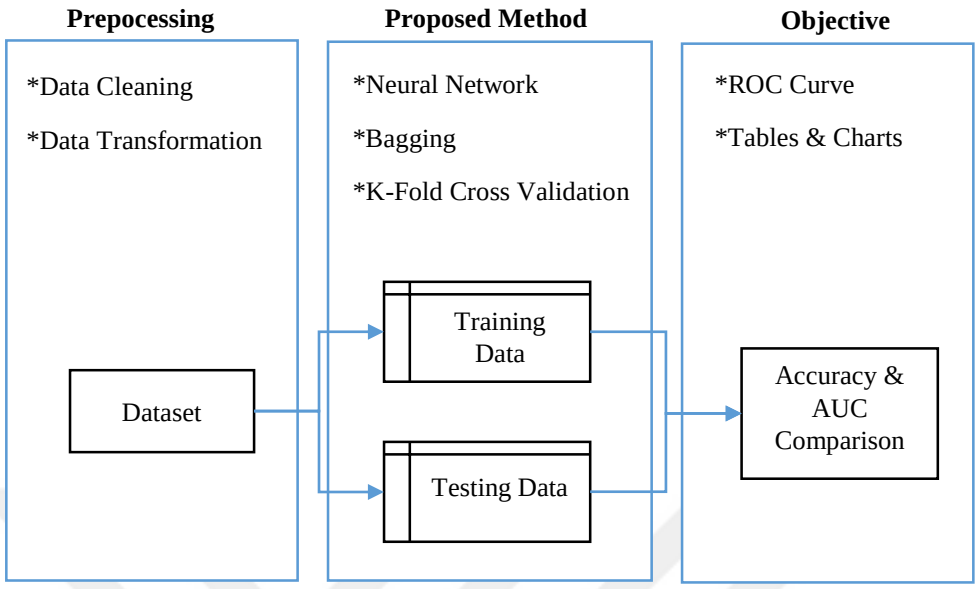

Figure 3.14. The Proposed Framework

Figure 3.2 depicts the whole process of the implementation of the bagging neural network to address class imbalance problem in clinical datasets. There are 3 major steps from the proposed framework, they are preprocessing, proposed method, and objective. In the preprocessing step the selected datasets are cleaned and transformed to be fitted into the learning algorithm models.

Then, the datasets are ready to be processed in the learning algorithm models. Next, the dataset is divided by k-fold cross validation method into training data and testing data with 10 folds. K-fold cross validation is one of evaluation methods for estimating a classifier's accuracy and is recommended due to its relatively low bias and variance (Peixoto et al., 2017). Afterwards the dataset is sent to be processed in the proposed method. In the proposed method, the dataset is divided into a set of bootstraps of sample data which is randomly selected by bagging method. Each bootstrap of data is processed by base classifiers. Each classifier predicts its class prediction and is combined together as a vote with equal weight and assigns the major vote as an accurate prediction.

In this study, the backpropagation neural network is implemented as a base classifier of bagging method using ReLU activation function. The algorithm is a supervised learning algorithm from neural network algorithm that learns by adjusting the weight in response to the calculated error. A standard network architecture comprises three layers, input, hidden and output layer (Han & Kamber, 2006). In the learning phase of the backprogation neural network, the dataset is forwarded and calculated in the network

until it reaches output layer, then the result from the output layer is back propagated and the weights are adjusted. The algorithm is repeated until it reaches the intended output.

ReLU activation function:

$$f(x) = \max(0, x) \tag{3.2}$$

3.5.Method Test and Experiment

In this section, the proposed method and the conventional neural network are tested to measure the performance. There are 10 different topologies are used in the experiment for each neural network model. 10 different topologies are implemented to prove that the bagging neural network generally works with better accuracy results for all different topologies compare to the conventional neural network. The parameters of neural network such as input layer, learning rate, number of iteration and so on, are set and adjusted according to each different clinical dataset in order to get maximized output or prediction results. For the detail example of the parameters and specifications of the conventional neural network and the bagging neural network for this study are shown in Table 3.2 below.

Table 3.2. Example of the Neural Network's parameter on diabetes dataset

Parameter	Specification
Architecture	1 hidden layer
Input layer	8 neurons
Hidden layer	from 3 to 13 neurons
Output layer	2 neurons
Learning rate	0.01
Iteration	700

3.6.Evaluation Result and Validation

For evaluating the experimental results, area under ROC curve (AUC) is used to measure the accuracy of the proposed method. AUC is performance metric for two class classification problems depicted in visual comparison. AUC demonstrates the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity). Sensitivity is the correct prediction from the positive class and specificity is the correct prediction from the negative class.

$$sensitivity = \frac{TP}{TP+FN} \quad (3.3)$$

$$specificity = \frac{TN}{TN+FP} \quad (3.4)$$

Furthermore, ROC curve is popularly used in the biomedical science (Mazurowski et al., 2008). The technique is a suitable approach for the proposed method since the model only has two prediction output, patient who is healthy or unhealthy. From the experiment results of this study, AUC is calculated for every selected clinical dataset to evaluate the prediction accuracy of the proposed method. The best classifier is with threshold value closer to 1.0. The calculation result of AUC can be depicted from the figure 3.3. It shows that the closer to the diagonal line or the area 0.5, the less accurate is the model. However, the closer to the area 1.0, the more accurate is the model.

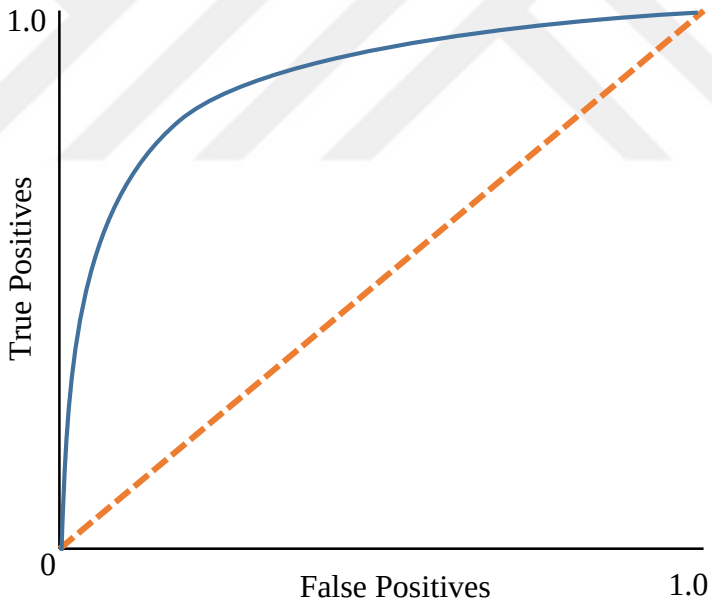

Figure 3.15. Example of ROC curves

4. RESULTS AND DISCUSSION

4.1.Results

The datasets used in this study are collected from UCIMLR. They are Pima Indians Diabetes Dataset, Wisconsin Breast Cancer Dataset, and Indian Liver Patient Dataset. These datasets are clinical datasets with class imbalance problem.

a. Diabetes Dataset

Backpropagation neural network is used in the experiment. There are 10 experiments with 10 different neural network topologies. Each topology used 1 hidden layer but each hidden layer for each topology used different number of neurons. ReLU activation was used for all neurons. The neural networks were trained for 700 iteration with 0.01 learning rate. The detail of the neural network's parameters on diabetes dataset can be seen in Table 4.1.

Table 4.1. Neural network's parameter on diabetes dataset

Parameter	Specification
Architecture	1 hidden layer
Input layer	8 neurons
Hidden layer	from 3 to 13 neurons
Output layer	2 neurons
Learning rate	0.01
Iteration	700

As summarized in Table 4.1., the parameter used for the conventional neural networks and the bagging neural network are the same. In Table 4.2. and Figure 4.1., the accuracy of the conventional neural networks show that 2 experiments out of 10 experiments give better performance than the bagging neural networks. However, 8 experiments out of 10 experiments of the bagging neural networks show better performance than the conventional neural network. This indicates that clinical prediction on diabetes dataset using the bagging neural networks outperform the accuracy of the conventional neural networks in general.

Table 4.2. Experimental results of diabetes dataset

Experiment	No of Neurons at Hidden layer	Accuracy		AUC	
		NN	BNN	NN	BNN
1	3	0.76046	0.76434	0.82763	0.82936
2	5	0.76306	0.76564	0.82492	0.83101
3	6	0.76695	0.76827	0.83308	0.82749
4	7	0.74593	0.77476	0.80812	0.83245
5	8	0.76049	0.76825	0.82501	0.82938
6	9	0.75784	0.77215	0.80486	0.83049
7	10	0.76567	0.77085	0.82887	0.83137
8	11	0.77806	0.76695	0.82784	0.83118
9	12	0.76953	0.76564	0.83271	0.82991
10	13	0.74356	0.77085	0.78950	0.83007
Average		0.761155	0.76877	0,820254	0,830271

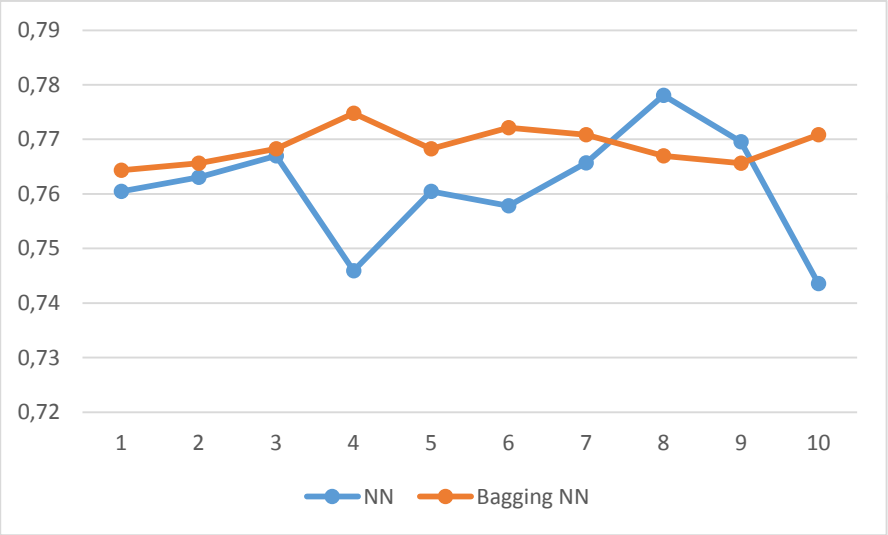

Figure 4.1. Experimental Results of Diabetes Dataset (Accuracy)

b. Breast Cancer Dataset

Similar to the diabetes dataset, backpropagation neural network is used in the experiment. There are 10 experiments with 10 different neural network topologies. Each topology used 1 hidden layer but each hidden layer for each topology used different number of neurons. ReLU activation is used for all neurons. In this case, the neural networks were trained for 1000 iteration with 0.1 learning rate. The detail of the neural network's parameters on breast cancer dataset can be seen in Table 4.3.

Table 4.3. Neural network's parameter on breast cancer dataset

Parameter	Specification
Architecture	1 hidden layer
Input layer	8 neurons
Hidden layer	from 3 to 13 neurons
Output layer	2 neurons
Learning rate	0.1
Iteration	1000

The parameter used for the conventional neural networks and the bagging neural networks on breast cancer dataset are the same as summarized in Table 4.3. In Table 4.4., the experimental results from both models, the conventional neural network and the bagging neural network, have good accuracies achieving over 0.96%. It can clearly be seen in Figure 4.2. that both models reach optimum results, however on average the bagging neural networks outperform the conventional neural networks.

This indicates that clinical prediction on breast cancer dataset using the proposed method gives better accuracy than the conventional neural network in general.

Table 4.4. Experimental results of breast cancer dataset

Experiment	No of Neurons at Hidden layer	Accuracy		AUC	
		NN	BNN	NN	BNN
1	3	0.92143	0.96571	0.94379	0.99234
2	5	0.96857	0.96571	0.99258	0.99287
3	6	0.96286	0.96571	0.99213	0.99244
4	7	0.96571	0.96571	0.99251	0.99265
5	8	0.96286	0.96714	0.99270	0.99245
6	9	0.96429	0.96429	0.99203	0.99267
7	10	0.96857	0.96714	0.99305	0.99252
8	11	0.95714	0.96571	0.99213	0.99258
9	12	0.96429	0.96714	0.99302	0.99269
10	13	0.96286	0.96286	0.99296	0.99286
Average		0.959858	0.965712	0,98769	0,992607

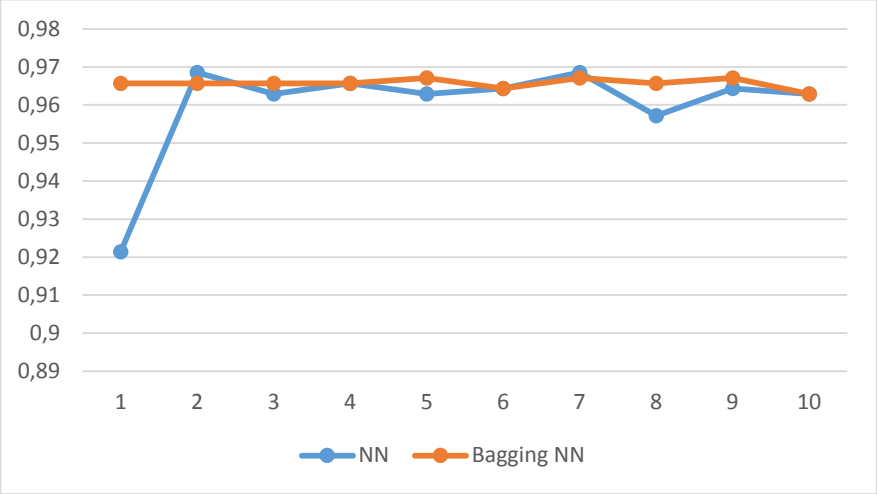

Figure 4.2. Experimental Results of Breast Cancer Dataset (Accuracy)

c. Liver Dataset

For liver dataset, backpropagation neural network is used in the experiment. There are 10 experiments with 10 different neural network topologies. Each topology used 1 hidden layer but each hidden layer for each topology used different number of neurons. ReLu activation is used for all neurons. In liver dataset experiment, the neural networks were trained for 1000 iteration with 0.01 learning rate. The detail of the neural network's parameters on liver dataset can be seen in Table 4.5.

Table 4.5. Neural network's parameter on liver dataset

Parameter	Specification
Architecture	1 hidden layer
Input layer	9 neurons
Hidden layer	from 3 to 13 neurons
Output layer	2 neurons
Learning rate	0.01
Iteration	1000

The parameter used for the conventional neural networks and the bagging neural network are the same as summarized in Table 4.5. From Table 4.2. and Figure 4.1., the experimental results of the conventional neural network show that only 1 experiment out of 10 experiments gives better performance than the bagging neural network. However, the experimental results of the bagging neural network show that 9 experiments out of 10 experiments have higher accuracy

than the conventional neural network. This indicates that the bagging neural networks significantly shows better performance than the conventional neural networks in general.

Table 4.6. Experimental results of liver dataset

Experiment	No of Neurons at Hidden layer	Accuracy		AUC	
		NN	BNN	NN	BNN
1	3	0.70824	0.71335	0.64862	0.71749
2	5	0.71856	0.71166	0.68885	0.71921
3	6	0.70655	0.71338	0.65774	0.72660
4	7	0.70652	0.71335	0.71185	0.72266
5	8	0.70991	0.71508	0.66205	0.72187
6	9	0.7083	0.71163	0.70325	0.72263
7	10	0.70999	0.71508	0.66518	0.71965
8	11	0.70313	0.71335	0.64594	0.72422
9	12	0.69968	0.71338	0.66383	0.72783
10	13	0.70663	0.71508	0.69730	0.72530
Average		0.707751	0.713534	0,674461	0,722746

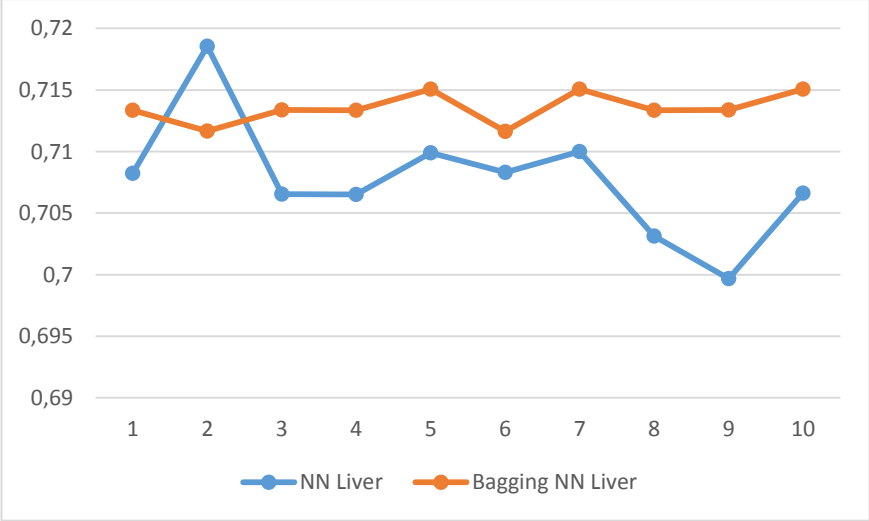

Figure 4.3. Experimental Results of Liver Dataset (Accuracy)

4.2.Discussion

In this study, AUC is used as an evaluation metric to measure the performance of the conventional neural network and the bagging neural network. The best classifier is the classifier with AUC threshold value closer to 1.0. From those tables of experimental results above, it shows that the conventional neural network on average has an AUC of 0.82 on diabetes dataset, AUC of 0.98 on breast cancer dataset, AUC of 0.67 on liver dataset, however the bagging neural network has an AUC of 0.83 on diabetes dataset, AUC of 0.99 on breast cancer dataset, AUC of 0.72 on liver dataset. Therefore, it is clearly seen from the results that the bagging neural network has better performance.

These experimental results show that bagging neural network is able to address class imbalance problem in clinical datasets so that it is able to give better performance and more accurate prediction in diagnosing patient's disease than the conventional neural networks. This ensemble method of bagging and neural network proposed in the study can be an effective model to address class imbalance problem in clinical datasets.

5. CONCLUSION AND FUTURE WORK

5.1.Conclusion

In this study, the effect of bagging in neural network with 10 different topologies for the class imbalance problem on clinical datasets were studied an evaluated empirically. An experimental study was conducted based on 3 selected clinical datasets of UCI Machine Learning Repository to predict whether the patient is healthy or unhealthy.

From the experiments of the diabetes dataset, on average the conventional neural network achieves 0.761% accuracy, while the bagging neural network achieves 0.768% accuracy. On the breast cancer dataset, on average the conventional neural network achieves 0.959% accuracy, whereas the bagging neural network achieves 0.965% accuracy. While, on the liver dataset, on average the conventional neural network achieves 0.707% accuracy, and the bagging neural network achieves 0.713% accuracy. From three different selected clinical datasets and ten different topologies the bagging neural networks shows better performance than the conventional neural network in general. In addition, data quality affects the learning algorithm to get optimized results than can be seen from the results given from breast cancer dataset that has a better data quality than the other datasets.

Finally, from the experimental results above, it can be concluded that bagging method is effectively able to address class imbalance problem for clinical diagnosis prediction with neural network algorithm so that it outperforms the performance of the conventional neural network.

5.2.Future Work

Several recommendations can be given for the future work from this study are as follows:

- Deeper topology of neural network or more hidden layers can be implemented to get maximized prediction results.
- To address class imbalance problem, another ensemble methods can be applied and compared to find the best results.
- The selected clinical datasets should have a good data quality because it effects the performance of the learning algorithm.

REFERENCES

Cilimkovic, M. (2010). Neural Networks and Back Propagation Algorithm. *Fett.Tu-Sofia.Bg*. Retrieved from http://fett.tu-sofia.bg/et/2006/ET2006 BOOK 1/Circuits and Systems/173 Paper-V_Skorpil.pdf

Collell, G., Prelec, D., & Patil, K. R. (2017). A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, *0*, 1–11. https://doi.org/10.1016/j.neucom.2017.08.035

Fan, Q., Wang, Z., & Gao, D. (2016). One-sided Dynamic Undersampling No-Propagation Neural Networks for imbalance problem. *Engineering Applications of Artificial Intelligence*, *53*, 62–73. https://doi.org/10.1016/j.engappai.2016.02.011

Han, J., & Kamber, M. (2006). Data Mining : Concepts and Techniques.

Huang, Y.-M., Hung, C.-M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, *7*(4), 720–747. https://doi.org/10.1016/j.nonrwa.2005.04.006

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia Computer Science*, *72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications*, *37*(4), 3373–3379. https://doi.org/10.1016/j.eswa.2009.10.012

Lee, S.-J., Xu, Z., Li, T., & Yang, Y. (2017). A Novel Bagging C4.5 Algorithm Based on Wrapper Feature Selection for Supporting Wise Clinical Decision Making. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2017.11.005

Madadipouya, K. (2015). a New Decision Tree Method for Data Mining, *2*(3), 31–37. https://doi.org/10.5121/acii.2015.2304

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, *21*(2–3), 427–436. https://doi.org/10.1016/j.neunet.2007.12.031

Peixoto, R., Ribeiro, L., Portela, F., Filipe Santos, M., & Rua, F. (2017). Predicting Resurgery in Intensive Care - A data Mining Approach. *Procedia Computer Science*,

113, 577–584. https://doi.org/10.1016/j.procs.2017.08.291

Pereira, S., Portela, F., Santos, M. F., Machado, J., & Abelha, A. (2015). Predicting Type of Delivery by Identification of Obstetric Risk Factors through Data Mining. *Procedia Computer Science*, *64*, 601–609. https://doi.org/10.1016/j.procs.2015.08.573

Ranjan, T., & Kumar, S. (2016). Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset. *Procedia - Procedia Computer Science*, *85*(Cms), 862–870. https://doi.org/10.1016/j.procs.2016.05.276

Setiyorini, T., & Wahono, R. S. (2014). Penerapan metode bagging untuk mengurangi data noise pada neural network untuk estimasi tuat tekan beton. *Journal of Intelligent Systems*, *1*(1), 36–41.

Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, *18*(1), 63–77. https://doi.org/10.1109/TKDE.2006.17

Health Sciences Library, University Libraries, University of Washington, https://guides.lib.uw.edu/hsl/data/findclin, 17 April 2018.

Dua, D. and Karra Taniskidou, E. (2017). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

CURRICULUM VITAE

Personal information	
First name(s) / Surname(s)	**Izhan Fakhruzi**
Permanent Address	9, Jl. KHW. Hasyim GG. Mutiara, 78118, Pontianak, Indonesia
Telephone	Mobile: +90 553 771 6981
E-mail	Izhan.fa@gmail.com
Nationality	Indonesia
Date of birth	30.03.1988
Gender	Male
Work experience	
Dates	from 2 January 2012 to 28 December 2012
Occupation or position held	Technology Intermediary – Innovation Manager
Name of employer	Business Technology Center (BTC) Network - Indonesia
Type of business or sector	Business, innovation and technology consulting
Dates	From 02.11.2011 to 25.11.2011
Occupation or position held	Apprentice
Name of employer	TUM-Tech GmbH - Germany
Type of business or sector	Business, innovation and technology consulting
Education and training	
Dates	From 2014 to Present (in progress)
Title of qualification awarded	Master of Science
Name and type of organisation providing education and training	Marmara University, Turkey
Level in international classification	Msc
Dates	From September 2005 to January 2010
Title of qualification awarded	Bachelor of Engineering
Name and type of organisation providing education and training	Tanjungpura University, Indonesia
Level in national classification	ST
Awards	- Awarded a scholarship from Ministry of Research and Technology of Indonesia (RISTEK) for a technology intermediary training, VDI|VDE|IT GmbH, Berlin, Germany (October 2011 – November 2011) - Awarded a scholarship from The. US department of State to study English and American culture in Ohio University (April 2009 – May 2009)