# A Comparison of Classification Methods in Vertebral Column Disorder with the Application of Genetic Algorithm and Bagging

Rizki Tri Prasetio[1] and Dwiza Riana[2]
[1,2]Magister of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia
(Tel : +62-21-741-0437; E-mail: rizki.rte@bsi.ac.id)
(Tel : +62-21-741-0437; E-mail: dwiza@bsi.ac.id)

**Abstract - Disorders of the spine are experienced by about two-thirds of adults and belong to the second most common disease after headache. Prediction of spinal disorders is difficult because it requires an experienced radiologist to analyze images of Magnetic Resonance Imaging (MRI). The use of Computer Aided Diagnosis (CAD) system can help the radiologist detect abnormalities in the spine more optimal. In the vertebral column data set which is available now has three classes that indicate the condition of the spine, that are herniated disk class, spondylolisthesis class and normal class. As well as on the data sets that has several classes, there are problems called the class imbalance which causes a lack of accuracy in the classification results. In this study, the combination of genetic algorithm and bagging technique are proposed to improve the accuracy of class classification on spinal disorders. Genetic algorithm is used for feature selection while bagging technique is used to solve the problem of class imbalance. The proposed method is applied to three classifier algorithms, namely naïve bayes, neural networks and k-nearest neighbor. The results showed that the proposed method makes a significant improvement in the classification of disorders of the spine for most classifier algorithms. The best algorithm after applied to genetic algorithms and bagging technique is k-nearest neighbor with an accuracy of 89.03%, 88.06% for the neural network and 86.13% for naïve bayes if validated using cross validation.**

**Keywords: Vertebral Column Disorder, Imbalanced Class, Genetic Algorithm, Feature Selection, Bagging Technique.**

## I. INTRODUCTION

Low-back pain is a disorder of the lower spine (lumbar) which is a medical disorder that significantly affects 50-80% of the population [1]. Approximately two-thirds of adults suffer from this disease [2] ,and it becomes the second most common disease after headache [3].

Radiographs which were initially often used to investigate spine disorders now become less attractive since Magnetic Resonance Imaging (MRI) was discovered where it is able to describe the spinal cord and other soft tissues better [4]. Although MRI is a very good solution, analyzing MRI images requires considerable experience because the error in the MRI image analysis can lead to the wrong treatment [5]. Use of Computer-Aided Diagnosis (CAD) system can help the radiologist detect abnormalities in the medical viewed by using pattern recognition and machine learning [4].

Previous research on the dataset vertebral column was done by some researchers in order to be able to classify disorders of the spine using various methods including Artificial Neural Network (ANN) [6], Naïve Bayes (NB) [7], backpropagation [4], Generalized Regression Neural network (GRNN) [8] as well as embedded reject option [9]. In previous research on the vertebral column dataset, no researcher applies k-nearest neighbor as classifier algorithm. According to [10], k-nearest neighbor is a conventional non-parametric classifiers algorithm which produces good performance and which is easy to implement at low dimensional datasets by a small scale.

Problems that arise in most other studies is that the classification of the vertebral column dataset suffer negative effects such as imbalanced class so that weighting features or pre-processing methods should be used to improve the performance of classification algorithms [4]. Class imbalance is the uneven distribution among classes where one class is more than the other classes [11]. Class imbalance problem is a challenge for machine learning and data mining, and it has attracted significant research in recent years. Classification algorithm that is affected by the class imbalance problem for a particular data set will see an overall good accuracy but a very bad accuracy in the minority [12]. Class imbalance problem is usually accompanied by the issue of high-dimensional data sets, and applying feature selection techniques is an action that needs to be done [13].

Various feature selection technique was widely recommended by the world researchers to overcome the problem of class imbalance, such as relief algorithm [12], Particle Swarm Optimization (PSO) [11], Genetic Algorithm (GA) [14]. Genetic Encryption (GA) which are able to effectively explore the large search space and usually required in the case of attribute selection [14].

In general, the class imbalance can be handled by two approaches, namely the level of the data and the algorithm level. Level algorithm approaches were done by improving the algorithm or combining (ensemble) single classifier in order to be better [15]. Ensemble method such as bagging and boosting

978-1-4673-7800-0/15/$31.00 ©2015 IEEE

is another method widely used to handle the issue of class inequality [16].

This research will be done by comparing the neural network classification methods, naïve Bayes and k-nearest neighbor by applying a genetic algorithm for selection feature combined with ensemble bagging method to solve the problem of class imbalances in the dataset vertebral column.

## II. VERTEBRAL COLUMN

Vertebral column is the central axis of the human skeleton, extending from the bottom of the skull up to the pelvic bone which consists of 26 irregular individua bonesl [17]. This complex system can get dysfunction that causes back pain with different intensities. Disk Hernia and Spondylolisthesis are examples of the pathology of the spine that causes pain [9].

Dataset obtained and used in this study is the Vertebral Column which is a collection of biomedical data developed by Dr. Henrique da Mota during the period of medical residency at the Group of Applied Research in Orthopaedics (GARO). This dataset is taken directly from the UCI Machine Learning Repository which can be downloaded via the website http://archive.ics.uci.edu/ml/datasets/Vertebral+Column.

This dataset has 310 records which classify 3 classes. Each record consists of 6 attributes and 1 label. All of the attributes are the numerical attribute. The dataset is organized into 2 datasets which are different but still relevant. The first dataset consists of a classification of patients classified into one of three categories: Normal (100 patients), Disk Hernia (60 patients) or Spondylolisthesis (150 patients). The second dataset combines Disk Hernia and Spondylolisthesis category into a single category labeled abnormal. The second dataset contains the classification of the patients classified into one of two categories: Normal (100 patients) or Abnormal (210 patients).

Every patient in this dataset is represented as a vector or a pattern with six biomechanical attributes, in accordance with the following parameters: angle of pelvic incidence, angle of pelvic tilt, lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis

## III. METHODS

### A. Genetic Algorithm

Genetic algorithms are evolutionary algorithms are the most popular algorithm which uses the basic principle of natural selection introduced by Charles Darwin. Genetic algorithms are applied as an approach for identifying value and solution search for a wide range of optimization problems [18]. There are more advantages of genetic algorithms than other traditional optimization algorithms, two of which are the ability to handle complex and parallel problems. Genetic algorithms can handle a wide range of optimization depending on the objective function (fitness) whether balanced or unbalanced, linear or non-linear, continuous or non-continuous, or with random noise.

Gist of the genetic algorithm includes encoding or encryption optimization functions as an array containing the bits or characters of a string to describe the chromosome, the operation of string manipulation by genetic operators, and selection in accordance with fittness, with the purpose to find a good and optimal solution for the problems being faced. Genetic algorithms have three main genetic operators:

1. Crossover is the process of exchanging part of the solution (chromosome) with another parent in order to produce different types of chromosomes that may be a new solution to solve the problems. Its main role is to provide mixing solutions and convergence in the sub-space (generating new solutions).
2. Mutations are the exchangingof one part of the solution chosen randomly, which increase the diversity of the population and generating mechanisms to avoid local minima.
3. Selection of the fittest or the elitism is the usage of solution with high fitness values to pass to the next generation, which is often done in terms of some form of selection of the best solution.

### B. Bagging Technique

Bagging stands for bootstrap aggregating, using subdataset (bootstrap) to produce a training set L (learning). L trains basic learning using the learning procedure which is not stable, and then, during the test, taking the mean [19]. Bagging is good to use for classification and regression. In the case of regression, to be stronger, one can take an average when combining predictions.

Bagging builds a set of classifier by sub-sample such as training to produce different hypotheses. After hypothesis differences are generated, they are combined with a voting mechanism for the classification and the average of estimation or prediction.

Bootstrap data created by the resampling uniform sample with replacement from the original training data. The classification can be trained in parallel ,and the final classification is produced by combining the ensemble classification. Bagging was considered to be a variance reduction technique for a given classifier. Bagging is known to be very effective when the classifier is not stable, that is when learning perturbing set can lead to significant changes in the behavior of the classification. Because bagging improves generalization performance, the reduction of variance (noise) is maintained or only slightly increasing the bias [20].

### C. Naïve Bayes Classifier

Naive Bayes is a method that does not have a rule [21]. Naive Bayes uses a branch of mathematics known as probability theory to find the greatest opportunities of the possibility of classification, by looking at the frequency of each classification in the training data. Naive Bayes is a popular classification method and in the top ten best algorithm in the data [22]. This algorithm is also known by the name of Idiot's Bayes, Simple Bayes, and Independence Bayes [23].

Naive Bayes classification is a statistical classification that can be used to predict the probability of membership of a class. Bayesian classification is based on the Bayes theorem, taken from the mathematician's name who is also England Prebysterian minister, Thomas Bayes (1702-1761) [23].

Bayesian classification has a similar classification capability with decision tree and neural network. Bayes rule is used to calculate the probability of a class. Naive Bayes algorithm provides a means of combining the preceding opportunities on condition to be a formula that can be used to calculate the odds of any possibilities that may occur [21].

The general form of Bayes's theorem can be seen in equation 1 below:

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \qquad (1)$$

Where:

X : Data with unknown class

H : Hypothesis of X data of specific class

P(H|X) : The probability of the hypothesis H based on the X condition (posterior probability)

P(H) : The probability of the hypothesis H (prior probability)

P(X|H) : X probability based on condition of hypothesis H

P(X) : X probability

## C. Neural Network Classifier

Neural Network (NN) or also called Artificial Neural Network (ANN) is an attempt to mimic the human brain functions. The human brain is believed to be made up of millions of small processing units that work in parallel, called neurons. Neurons are connected each other through the connection of neurons. Each individual neuron takes input from a set of neurons. It then processes these inputs and outputs passing to a set of neurons. Output is collected by other neurons for further processing [24]. Neural network does learning process of examples, such as the human brain which learns every instance or previous experience or past events.

Neural network is [25] a set of input or output units which are connected where each relationship has a weight. The network was intended to simulate the behavior of the biological system of the human nervous system, which consists of a large number of processing units, called neurons, which operate in parallel. Neuron is related to a synapse that surrounds other neurons. Nervous system were presented in a neural network in the form of a graph consisting of nodes (neurons) connected with a bow, which corresponds to the synapse [26] Since 1950s [27], neural network has been used for the purposes of prediction, not only for classification but also for regression with continuous target attribute

Neural Network consists of an input layer, one or more hidden layer and output layer. Here's an explanation of each layer [27]:

1. Input Layer

The input layer for receiving the input value of each record in the data. Input node number equals the number of predictor variables.

2. Hidden Layer

Hidden layer transforms the input value in the network. Each node in the hidden layer is connected to the nodes in the hidden layer of the previous or nodes in the input layer and to nodes in the next hidden layer or to nodes in the output layer. Number of hidden layer can be whatever.

3. Output Layer

Line connected to the output layer is derived from the hidden layer or an input layer and returns the output value corresponding to the variable input.

## D. k-Nearest Neighbour

K-NN algorithm is a method that uses a supervised algorithm [22]. Differences between supervised and unsupervised learning are on the purposes. Supervised algorithm is intended to discover new patterns in the data by connecting patterns existing data with new data while in unsupervised learning, the data do not yet have any pattern, and the purpose of unsupervised learning is to find patterns in a data [25]. The purpose of the k-NN algorithm is to classify new objects based on attributes and training samples [28] where the results of the new test samples were classified by the majority of categories on the k-NN.

Classification process on this algorithm does not use any model to be matched and only based on memory. K-NN algorithm uses adjacency classification as predictive value of the new test sample. Selection of the distance calculation techniques are other important things. Usually Euclidean Distance and Manhattan Distance are used to calculate the distance to the k-NN algorithm [25]. The formula for both of the distance calculation techniques (2), (3).

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(X_k - Y_k)^2} \qquad (2)$$

$$d(x,y) = \sqrt{\sum_{k=1}^{n}|X_k - Y_k|} \qquad (3)$$

## E. Proposed Methods

The proposed method which is suggested applies genetic algorithms for feature selection and applies bagging technique on naïve Bayes algorithms, neural networks, k-nearest neighbor to classify disorders of the spine from dividing the dataset into training data and data testing.

TABLE I
ACCURACY OF THREE CLASSIFIERS ON TWO VERTEBRAL COLUMN DATASETS (WITHOUT GENETIC ALGORITHM AND BAGGING)

| Validation | Vertebral Column 3 Classes | | | Vertebral Column 2 Classes | | |
|---|---|---|---|---|---|---|
| | Accuracy | | | Accuracy | | |
| | Naïve Bayes | Neural Network | k-Nearest Neighbor | Naïve Bayes | Neural Network | k-Nearest Neighbor |
| Cross Validation | 81.94% | 84.52% | 81.94% | 77.42% | 84.19% | 85.16% |
| 90% - 10% | 80.65% | 83.87% | 90.32% | 80.65% | 83.87% | 87.10% |
| 80% - 20% | 82.26% | 80.65% | 77.42% | 77.42% | 83.87% | 79.03% |
| 70% - 30% | 81.72% | 83.87% | 77.42% | 77.42% | 82.80% | 77.42% |
| 60% - 40% | 83.06% | 87.90% | 78.23% | 79.84% | 86.29% | 79.84% |

TABLE II
ACCURACY OF THREE CLASSIFIERS ON TWO VERTEBRAL COLUMN DATASETS (WITH GENETIC ALGORITHM AND BAGGING)

| Validation | Vertebral Column 3 Classes | | | Vertebral Column 2 Classes | | |
|---|---|---|---|---|---|---|
| | Accuracy | | | Accuracy | | |
| | Naïve Bayes | Neural Network | k-Nearest Neighbor | Naïve Bayes | Neural Network | k-Nearest Neighbor |
| Cross Validation | 86.13% | 88.06% | 89.03% | 82.90% | 87.74% | 88.71% |
| 90% - 10% | 90.32% | 90.32% | 96.77% | 83.87% | 90.32% | 96.77% |
| 80% - 20% | 90.32% | 88.71% | 90.32% | 83.87% | 90.32% | 91.94% |
| 70% - 30% | 87.10% | 88.17% | 87.10% | 87.10% | 89.25% | 90.32% |
| 60% - 40% | 87.10% | 88.71% | 87.90% | 86.29% | 88.71% | 88.71% |

An early data processing begins by dividing the dataset into training and testing data, applies a genetic algorithm for feature selection in the training data, and applies the bagging technique on the training dataset that has been its majority. After that, naïve Bayes algorithms, neural networks and k-nearest neighbor are applied in the dataset and then combining the results of all algorithms. The next stage is to validate the models which are produced after that, calculate how much accuracy generated by the model. If the desired accuracy has not been reached, so repeat the process of selection feature using genetic algorithms. This iteration will continue to run until the optimal feature is resulted. The proposed method can be seen in Fig. 1.

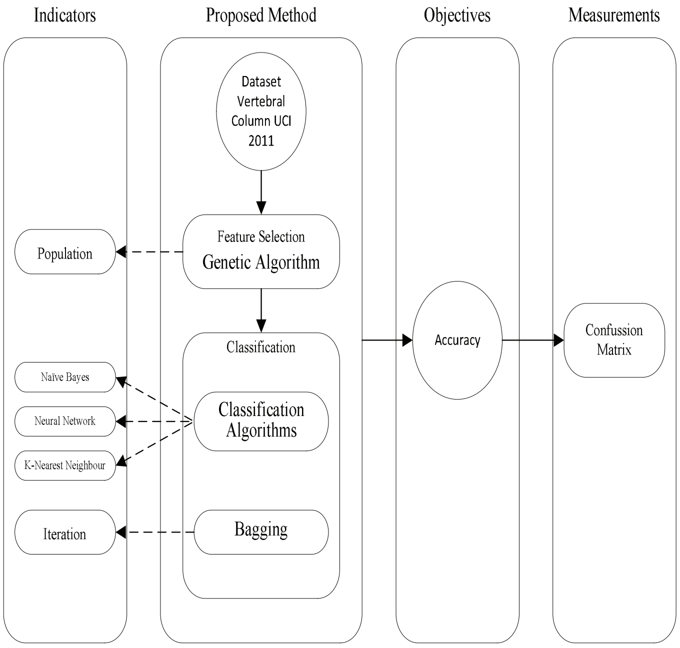

Fig. 1. Proposed Method

## IV. RESULT AND DISCUSSION

The study was conducted in two experiments which are experiments with the algorithm classifier (naïve Bayes, neural networks and k-nearest neighbor) without using feature selection based on genetic algorithms and techniques bagging (Table I) and experiments with the algorithm classifier using feature selection based on genetic algorithms and techniques bagging (Table II). Both are in the validation experiments using two types of validation, that are 10-fold cross validation and validation split. The experiment using standard parameter configuration for neural network (learning rate=0.3 and momentum=0.2), k-nearest neighbor (k=1), naïve bayes (laplace correction) and bagging (sample ratio=0.2).

The experiment results set forth in Table I declare that the neural network has the best accuracy average value among two other classification algorithms when applied to the dataset and the vertebral column of 3 and 2 classes while the experiment results set forth in Table II state that the k-nearest neighbor has the best accuracy value when applying genetic algorithms and bagging techniques on dataset vertebral column of 2 and 3 classes.

Based on the experiment results in this study, the difference in accuracy values were obtained on testing naïve Bayes algorithms, neural networks and k-nearest neighbor, combined with genetic algorithm and bagging the dataset vertebral column vertebral column 3C and 2C. To determine whether the proposed method can improve performance in the classification of disorders of the vertebral column significantly, testing using different test was done. t-Test Paired Two Sample for Means were used in any classification algorithm results between before and after using genetic algorithms and bagging techniques.

The test results of t-Test Paired Two Sample for Means generate that the application of genetic algorithms for feature selection combined with bagging technique can improve the performance of classification algorithms in terms of accuracy significantly in both datasets vertebral column, marked with P Value of t-Test <0.05. T-Test test results can be seen in Table III.

TABLE III
THE TEST RESULTS OF PAIRED TWO TAIL T-TEST

| Classifier | P value of t-Test | | Result | |
|---|---|---|---|---|
| | 2 Classes | 3 Classes | 2 Classes | 3 Classes |
| NB | 0.0038 | 0.0049 | **Sig. (P < 0.05)** | **Sig. (P < 0.05)** |
| NN | 0.0043 | 0.0204 | **Sig. (P < 0.05)** | **Sig. (P < 0.05)** |
| KNN. | 0.0051 | 0.0013 | **Sig. (P < 0.05)** | **Sig. (P < 0.05)** |

K-nearest neighbor algorithm is simple and high performance for a variety of applications because excess k-nearest neighbor is considered comparable to the much more complex algorithms such as neural networks and support vector machine (Lee et al., 2014). From the results of this study, it can be concluded that the KNN combined with GA and Bagging is superior when compared to other classifier algorithm on both the vertebral column datasets validated using either cross validation or split validation.

Comparison of the performance produced by the method proposed by previous studies can be seen in Table V.

TABLE IV
THE COMPARISON RESULT WITH THE PREVIOUS RESEARCH

| No. | Reseachers | Methode | Accuracy |
|---|---|---|---|
| 1 | Abdrabou [6] | Hybrid CBR and ANN | 85% |
| 2 | Reddy et al. [7] | Naïve Bayes | 83.74% |
| 3 | Unal et al. [4] | Multi Layer Perceptron | 85.48% |
| | | Naïve Bayes | 83.22% |
| 4 | Ansari et al. [8] | Feedforward backpropagation neural network (using half of dataset available) | 93.87% |
| 5 | Neto et al. [9] | SVM (Linear) | 84.30% |
| | | SVM (KMOD) | 85.90% |
| **6** | **This Reasearch** | **Combination of GA and bagging + naïve bayes** | **86.13%** |
| | | **Combination of GA and bagging + neural network** | **88.06%** |
| | | **Combination of GA and bagging + k-nearest neigbour** | **89.03%** |

## V. CONCLUSIONS

Genetic algorithms are applied to the selection of features and apply the bagging technique on several algorithm classifiers (naïve Bayes, neural networks and k-nearest neighbor) to solve the problem of class imbalances in the dataset vertebral column. Of the five experiments, all use two validation techniques that are cross validation and validation split. Genetic algorithms and bagging are proven effective to be able to improve the accuracy results of the dataset vertebral column classification, and furthermore the different test results among the three classifier algorithms combined with genetic algorithm and bagging with three classifier algorithms without genetic algorithm and bagging produce significant differences.

Comparison of the classification algorithms are proposed to compare the accuracy of the results among naïve Bayes, neural networks and k-nearest neighbor that have been combined with genetic algorithm and bagging techniques. Of 5 time-experiment, the k-nearest neigbour algorithm are proven to have the highest accuracy value compared with neural network and naïve Bayes.

In this study, in general, genetic algorithms applied to the selection of features and bagging techniques to overcome the problem of class imbalances can improve the accuracy in the dataset vertebral column classification, but some things can be applied to enhance the research, which uses another algorithm metaheuristic for selection feature and another ensemble technique application to overcome the problem of class imbalance and add another classification algorithm.

## ACKNOWLEDGMENT

This study is supported by Department Computer Science of STMIK Nusa Mandiri Jakarta.

## REFERENCES

[1] A. A. White and S. L. Gordon, "Synopsis: Workshop on Idiopathic Low-Back Pain," *Spine,* p. 141, 1982.

[2] R. A. Deyo and J. N. Weinstein, "Low Back Pain," *New England Journal of Medicine,* pp. 363-370, 2001.

[3] M. N. Brant-Zawadzki, C. S. Dennis, G. F. Gade and M. P. Weinstein, "Low Back Pain," *Radiology,* pp. 321-330, 2000.

[4] Y. Unal and E. Kocer, "Diagnosis of Pathology on the Vertebral Column with Backpropagation and Naive Bayes Classifier," *Technological Advances in Electrical, Electronics and Computer Engineering,* pp. 278-281, 2013.

[5] T. Videman, P. Nummi, M. Battie and K. Gill, "Digital assessment of MRI for lumbar disc desiccation. A comparison of digital versus subjective assessments and digital intensity profiles versus discogram and macroanatomic findings.," *Spine,* pp. 192-198, 1994.

[6] E. Abdrabou, "A Hybrid Intelligent Classifier for The Diagnosis of Pathology on the Vertebral Column," *Artificial Intelligence Methods and Techniques for Business and Engineering Applications,* pp. 297-309, 2012.

[7] S. K. Reddy, S. R. Kodali and J. L. Gundabathina, "Classification of Vertebral Column using Naive Bayes Technique," *International Journal of Computer Application,* pp. 38-42, 2012.

[8] S. Ansari, N. Naveed, F. Sajjad and I. Shafi, "Diagnosis of Vertebral Column Disorders Using Machine Learning Classifiers," *Information Science and Application,* 2013.

[9] A. R. Neto, A. G. Barreto, R. G. Sousa and J. S. Cardoso, "Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option," *Pattern Recognition and Image Analysis - 5th Iberian Conference,* 2011.

[10] A. Katari and M. D. Singh, "A Review of Data Classification Using K-Nearest Neighbour Algorithm," *International Journal of Emerging Technology and Advanced Engineering,* pp. 354-360, 2013.

[11] T. Deepa and M. Punithavalli, "An Innovative Optimization Algorithm for Feature Selection - A Comparative Study," *International Journal of Computer Science and Information Technology & Security,* pp. 20-24, 2013

[12] D. Tiwari, "Handling Class Imbalance Problem Using Feature Selection," *nternational Journal of Advanced Research in Computer Science & Technology* , pp. 516-520, 2014.

[13] N. V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations,* 2004.

[14] P. Villar, A. Fernandez and F. Herrera, "A Genetic Algorithm for Feature Selection and Granularity Learning in Fuzzy Rule-Based Classification Systems for Highly Imbalanced Data-Sets," *Information Processing and*

*Management of Uncertainty in Knowledge-Based Systems,* pp. 741-750, 2010.

[15] A. Saifudin and R. S. Wahono, "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *Journal of Software Engineering,* pp. 28-37, 2015.

[16] S. Zhongbin, S. Qinbao, Z. Xiaoyan, S. Heli, X. Baowen and Z. Yuming, "A novel ensemble method for classifying imbalanced data," *Elsevier Pattern Recognition,* pp. 1623-1637, 2015.

[17] P. Tate, Seeley's Principles of Anatomy and Physiology (2nd Edition), Ohio: McGraw-Hill Education, 2012.

[18] F. Gorunescu, Intelligent System Reference Library, Berlin Heidelberg: Springer-Verlag, 2011.

[19] L. Breiman, "Bagging Predictors," *Machine Learning,* pp. 123-140, 1996.

[20] M.-J. Kim and D.-K. Kang, "Classifier Selection in Ensembles using Genetic Algorithm for Bankruptcy Prediction," *Expert System with Application: An International Journal,* pp. 9308-9314, 2012.

[21] S. N. N. Alfisahrin, "Komparasi Algoritma C4.5, Naive Bayes dan Neural Network Untuk Memprediksi Penyakit Jantung," Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri, Jakarta, 2014.

[22] X. Wu and V. Kumar, The Top Ten Algorithms in Data Mining, New York: CRC Press, 2009.

[23] M. Bramer, Principle of Data Mining Second Edition, London: Springer, 2013.

[24] A. Shukla, R. Tiwari and R. Kala, Real Life Applications of Soft Computing, New York: CRC Press, 2010.

[25] J. Han and M. Kamber, Data Mining Concepts and Techniques Second Edition, San Francisco: Diane Cerra, 2006.

[26] E. Alpaydin, Introduction to Machine Learning, London: The MIT Press, 2010.

[27] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making, Cornwall: John Wiley & Sons, Ltd., 2009.

[28] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, New Jersey: John Wiley & Sons, Inc, 2005.