

A Hybrid Method to Predict Success of Dental Implants

Reyhaneh Sadat Moayeri*

Department of Computer and
Informatics Engineering,
Payame-Noor University, Tehran, Iran

Mehdi Khalili

Department of Computer and
Informatics Engineering,
Payame-Noor University, Tehran, Iran

Mahsa Nazari

Department of Computer and
Informatics Engineering,
Payame-Noor University, Tehran, Iran

Abstract—Background/Objectives: The market demand for dental implants is growing at a significant pace. Results obtained from real cases shows that some dental implants do not lead to success. Hence, the main problem is whether machine learning techniques can be successful in prediction of success of dental implants.

Methods/Statistical Analysis: This paper presents a combined predictive model to evaluate the success of dental implants. The classifiers used in this model are W-J48, SVM, Neural Network, K-NN and Naïve Bayes. All internal parameters of each classifier are optimized. These classifiers are combined in a way that results in the highest possible accuracies.

Results: The performance of the proposed method is compared with single classifiers. Results of our study show that the combinative approach can achieve higher performance than the best of the single classifiers. Using the combinative approach improves the sensitivity indicator by up to 13.3%.

Conclusion/Application: Since diagnosis of patients whose implant does not lead to success is very important in implant surgery, the presented model can help surgeons to make a more reliable decision on level of success of implant operation prior to surgery.

Keywords—Data Mining; Dental Implant; W-J48; Neural Network; K-NN; Naïve Bayes; SVM

I. INTRODUCTION

Dental plants are a sophisticated and unique technology with diverse applications which had a huge market in 2011 worth almost 7 billion USD¹. Although it has been over a decade since the successful use of single tooth implants started, many uses and conditions for implants remain conditional and little understood. Many conditions that dentists consider important include success rates and long-term survival affected by several factors including the location of substitution (i.e., denture replacement), implant, denture anchoring, bone density, tissue health, age of recipient, prosthetic complications, abutment and implant types and also materials and post-operative medicines². Therefore, this medical field of technology requires the combination of continuous clinical trials and technical innovation to improve implant reliability and survival rates and also reduce failure rates^{3,4}.

Data mining is a computational process used to discover patterns in large data sets via methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The purpose of this process is to extract

information from a data set and then transform it into a comprehensible structure for extra use. For data mining technologies, many data mining methods such as clustering, association, evolution, pattern matching, generalization, characterization, classification, data visualization and meta-rule guided mining have been developed⁵.

Ensemble methods which combine the output of individual classifiers have been very successful for the production of accurate prediction for many complicated classification tasks. These methods become successful when they have the appropriate ability to consolidate accurate predictions and also correct errors in many diver base classifiers⁶.

Two really good and well-known ensemble methods are as follows: a form of meta-learning which is called stacking and also ensemble selection. Stacking, makes a higher-level predictive model over the predictions of base classifiers, whereas, ensemble selection utilizes an incremental strategy to select base predictors for the ensemble while simultaneously balancing performance and diversity. These approaches have really superior performance in several areas because of their ability to utilize heterogeneous base classifiers^{7,8}.

Stacking does not really manipulate the training dataset. Instead, based on two levels, an ensemble of classifiers is generated. In the base level, different learning algorithms are used to train multiple classifiers. The diversity is provided because different learning algorithms make different errors in the same dataset. A meta-classifier is utilized to general the final prediction. The meta-classifier is trained using a learning algorithm via a meta-dataset that combines the outputs of base-level classifiers and the real class label. A problem that exists in stacking is how to acquire an "appropriate" configuration of the meta-classifier and base-level classifiers for each domain-specific dataset. The type of meta-classifier matters to the function of the base-level classifiers. To determine the configuration of stacking, some researchers have proposed different methods⁹.

Different approaches that combine many models, often called ensembles, have been explored. One of these approaches which is called "stacking", determines the optimally weighted average of many models through the minimization of predicted error. Wolpert introduced stacking in neural networks, whereas Breiman, extended the idea to uncensored regression models and then demonstrated that stacking can improve prediction error.

Breiman, discovered that combining completely different regression modes like ridge regression and subset regression, significantly reduced prediction error. LeBlanc and Tibshirani, discovered that stacking with a constraint of non-negative weights, is an efficient method to combine models. Van der Laan, *et al.* individually developed uncensored stacking as a "Super Learner" algorithm and offered results about the rate of convergence of the stacked estimator. Recently, Boonstra, *et al.* recently used stacking for the improvement of prediction when using different generation sequencing information in high-dimensional genome analysis^{10,11}.

A. Related Works

In Text¹², Oliviera, *et al.* have studied the performance of the four techniques of RBF DDA, SVM, K-NN and NNSRM on dental implants using 10-fold cross validation. Input variables used in this study were age, gender, position, type, smoking and illness. Results showed that the accuracy of the classifiers were 75.49%, 75.96%, 75.93%, and 72.18%, respectively.

In Text¹³, Braga, *et al.* proposed a decision model of prediction for dental implants by building a set of binary logistic models that could assess the probability of success or failure in oral rehabilitation process taking into account some genetic factors, individual habits, clinical and nonclinical factors. Results showed that they have obtained the AUC=0.789 (area under ROC curve).

II. BASIC THEORIES

A. Decision tree technique

Definition of a decision tree: it is a decision support system that utilizes tree-like graph decisions and their probable after-effect, including resource costs, chance event results and utility. A classification tree or decision tree, is utilized to learn a classification function which concludes the value of a dependent attribute (variable) considering the values of the independent (input) attributes (variables). This verifies a problem which is known as supervised classification, because the dependent attribute and the counting of classes (values) are provided. Decision trees are the most powerful approaches in data mining and knowledge discovery. It includes the technology to research large and complex data to discover useful patterns. It is very important because it provides the possible to model and extract knowledge from the available data¹⁴.

Specialists and theoreticians continually search for techniques to make the process more cost-effective, accurate and efficient. Decisions trees are very effective tools for many areas including information extraction, machine learning, data and text mining, and pattern recognition¹⁵.

B. Neural Network

An Artificial Neural Network (ANN) is an information processing paradigm inspired by information processing methods used by biological nervous systems like the brain. The key component of it is the novel structure of the information processing system. It is comprised of many highly interconnected processing elements (neurons) working

together to solve specific problems. Like people, ANNs learn by example. ANNs are typically comprised of hundreds of simple processing units wired together in a complicated communication network. Each unit or node is a simplified model of a real neuron which transmits a new signal or fires, in case it receives a sufficiently strong Input signal from the other nodes connected to it. A typical ANN is configured for a specific application like pattern recognition or data classification through a learning process. In biological systems, learning involves adjustments to the synaptic connections existing between the neurons¹⁶.

Neural networks are one of the methods of making classifiers in which learning model are shown by a collection of joined nodes besides with their weighted connections. Neural networks are widely used to design black box classifiers. Black box means in neural network based classifiers, there is no way to express the hidden knowledge of neural networks clearly. Exactly, unlike decision tree based classifiers which are completely interpretable¹⁷.

C. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyper plane. In other words, considering labeled training data (supervised learning), the algorithm gives an optimal hyper plane which classifies new examples¹⁸. Besides performing linear classification, SVMs can perform a non-linear classification efficiently using what is called the kernel trick, mapping their inputs into high-dimensional feature spaces implicitly¹⁹.

Given a dataset with n examples (x_i, y_i) , where each x_i is an input data and $y_i \in \{+1, -1\}$ corresponds to its bipolar label, $i=1, 2, \dots, n$. Using a nonlinear mapping $\phi(x)$, the input data is mapped into a high dimensional feature space F , in which the data are sparse and also possibly more separable. Then, the maximum margin which separates hyper-plane $w \cdot \phi(x) + b = 0$ is built in F , where w is a weight vector orthogonal to the hyper-plane, and b is an offset term. The margin is $1/\|w\|$. Maximizing the margin $1/\|w\|$ is equivalent to minimizing $\|w\|^2$, whose solution is found after solving the following quadratic optimization problem:

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n \xi_i \quad (1)$$

And

$$\text{Restrictions: } y_i(\omega \phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \text{ for } i = 1, \dots, n \quad (2)$$

Here, C is the penalty parameter which causes a trade-off between training error and generalization and ξ_i is a slack variable²⁰.

D. K-Nearest Neighbors (K-NN)

The KNN is the simplest classification technique for the times when there is almost no prior knowledge about the distribution of the data. It simply preserves the entire training set during learning and assigns a class represented by the majority label of its k -nearest neighbors in the training set to each query. The performance of a KNN classifier is determined primarily by the choice of K as well as the distance metric. The estimate is influenced by the sensitivity

of the selection of the neighborhood size K , and the reason for that is that the radius of the local region is determined by the distance of the K^{th} nearest neighbor to the query, and a different K yields different conditional class probabilities. If K is very small, the local estimate is usually going to be very poor due to the data sparseness and the noisy, ambiguous, or mislabeled points²¹.

E. Naïve Bayes

The Naïve Bayes classifier is a probabilistic classifier based on the Bayes theorem, regarding Naïve (Strong) independence assumption. Naïve Bayes classifiers assume that the effect of a variable value on a class is not related to the values of another variable. This assumption is referred to as class conditional independence. Naïve Bayes can usually perform more complex classification methods. It is especially suited when there is a high dimensionality of the inputs. When we want a more competent output, compared to other methods' output, we can utilize Naïve Bayes implementation. Naïve Bayesian is utilized to create models with predictive capabilities. An advantage of the naïve Bayes classifier is that it merely requires a small amount of training data for estimating the parameters required for classification^{22, 23}.

F. Hybrid Method

A combinative classifier is a method which combines several classifiers in order to promote the robustness and achieving higher performance. In fact, this method increases the accuracy of the classification via using the results of predictions of classifiers. One of the popular methods of combinations is stacking which is usually used to combine several different classifiers such as decision tree, neural network, etc.²⁴. In this method, a learning algorithm is trained to combine the predictions of many learning algorithms. First, all of the other algorithms are trained via using the available data, then a combiner algorithm is trained to make a final prediction with the use of all the predictions of the other algorithms as additional inputs. Usually, Stacking has a better performance than any trained models. It has been successfully used on both supervised learning tasks and unsupervised learning²⁵.

Stacking is related to combining multiple classifiers generated by using different learning algorithms L_1, \dots, L_N on a single dataset S , which is comprised of examples $s_i = (x_i, y_i)$, i.e., pairs of feature vectors (x_i) and their classifications (y_i) . In the first phase, a set of base-level classifiers C_1, C_2, \dots, C_N is generated, where $C_i = L_i(S)$. In the second phase, a meta-level classifier is learned that combines the outputs of the base-level classifiers²⁶.

These methods can further be used to evaluate the necessity of a dental implant and reduce the risks of using one by providing prosthodontists with predictions of the dental implant results based on a patient's physical condition and dental implant characteristics prior to performing surgery.

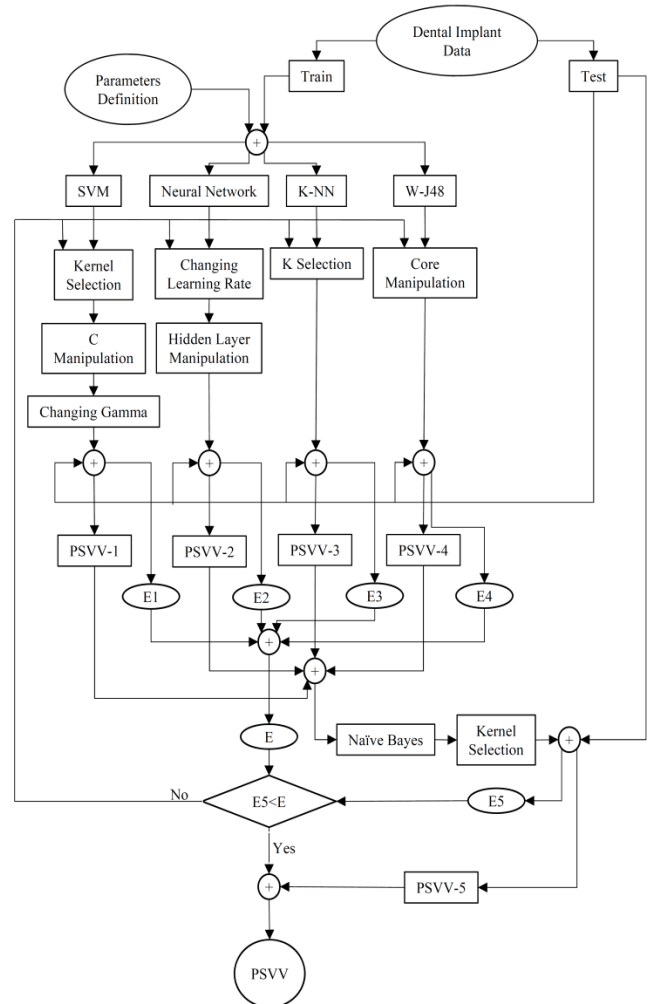
G. Cross Validation

Cross-validation (CV) has been widely used to facilitate model estimation and variable selection. In a typical K -fold CV process, the data set is randomly and evenly split into K parts (when possible). A candidate model is made based on

$K-1$ parts of the data set called a training set. Then the prediction accuracy of this candidate model is evaluated on a test set which contains the data in the hold-out part. By using each of the K parts as the test set and repeating the model building and evaluation process, we select the model with the smallest CV score as the 'optimal' model. In the K -fold CV procedure, each model is evaluated K times. The most common choice for evaluating a classification task is the accuracy. All other possible famous names of validation methods are seem to be as special cases of k -folds cross validation depending on the choosing value of k ^{27, 28}.

III. PROPOSED MODEL

The block diagram of the proposed combined predictive model is shown in **Error! Reference source not found.** and the steps involved are described as follows:



- 1) Block diagram of the proposed model to predict dental implant success
- 2) Divide the dental implant data with the related defined parameters in two sections: training (to design the related classifiers) and test (to calculate the minimum error of the classifiers)
- 3) Apply the training section parameters to each of the classifiers: SVM, Neural Network, K-NN and W-J48

- 4) Change the core and Gamma in SVM classifier, after selection of the suitable kernel, to achieve PSVV-1
- 5) Manipulate the learning rate and hidden layer in neural network classifier, select the suitable k in K-NN classifier and change the core in W-J48 classifier, respectively to obtain PSVV-2 to 4
- 6) Compare the results of the above classifiers with test section parameters to achieve related errors: E1 to E4
- 7) Combine four predictive success variable vectors: PSVV1 to PSVV4 and enter the result to stacking learner 1 i.e. Naïve Bayes algorithm
- 8) Apply Naïve Bayes classifier on the combined input, select the suitable kernel and finally compare the related output results with test section parameters, to achieve predictive success variable vectors PSVV-5 and error E5
- 9) Determine the minimum error value E from E1 to E4 to choose the final predictive success variable vector PSVV as follows:

$$\begin{cases} E5 < E: & PSVV = PSVV5 \\ E5 > E: & \text{Output of step 5} \end{cases}$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

According to the experts' opinion, the most important factors which influence the success or failure of dental implants are different. To evaluate the effectiveness of the proposed algorithm, we used 224 patient cases which had bone graft. This data set belongs to School of Dentistry of Tehran University and consists of 16 different dental parameters which are: gender, age, systemic, smoking, location, placement, loading, diameter, length, system, type, platform, connection, parallel taper, over-denture, and sinus lift.

Well-known performance indicators in medical problems are accuracy, sensitivity, and specificity²⁹.

Accuracy is the degree of how accurate is the prediction of implant success or failure.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Sensitivity is the degree of how accurate is the prediction of implant failure:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

Specificity is the degree of how accurate is the prediction of implant success:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

Here, TN is True Negative, TP is True Positive, FP is False Positive and FN is False Negative.

Rapid miner software was used as the data-classification tool to analyse the 224 implants. Comparison between proposed method and other methods of W-J48, Neural Network, SVM, K-NN and Naïve Bayes with three different cross validation techniques of 5, 7, and 10-fold are shown in the following figures.

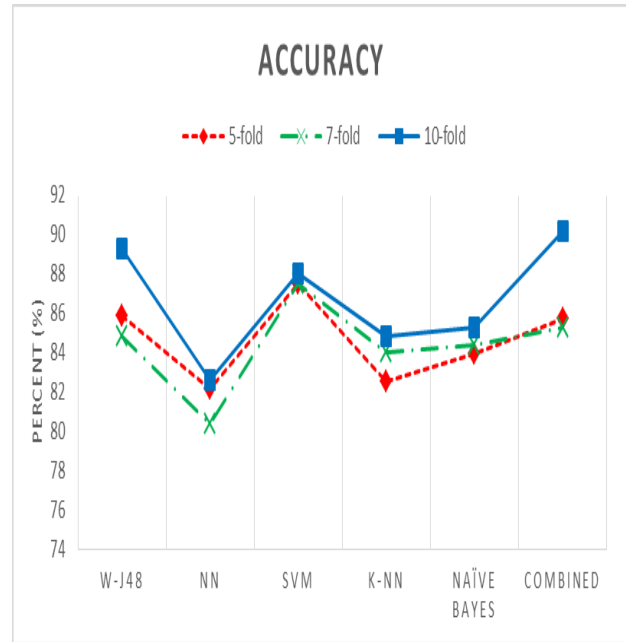


Fig. 1. Accuracy indicator for predictive models

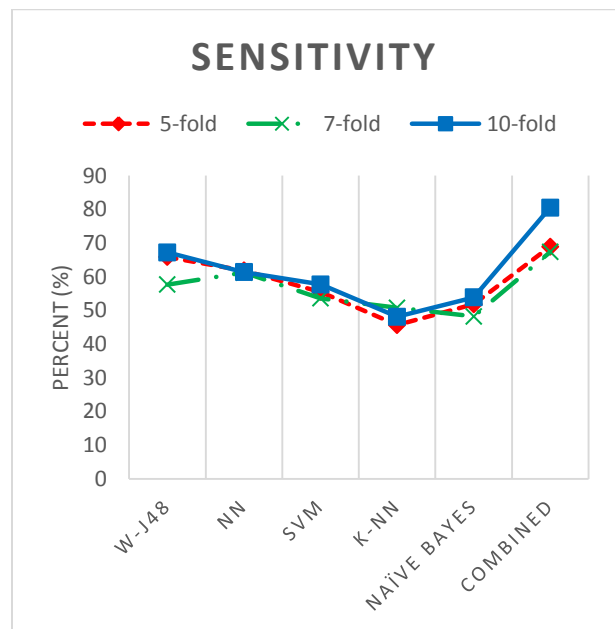


Fig. 2. Sensitivity indicator for predictive models

Comparison between different prediction methods with 5, 7, and 10-fold cross validation shows that the 10-fold cross validation gives higher performance than the others. Thus, the 10-fold cross validation has been selected for analysis. Results are summarized in **Error! Reference source not found.** as follows.

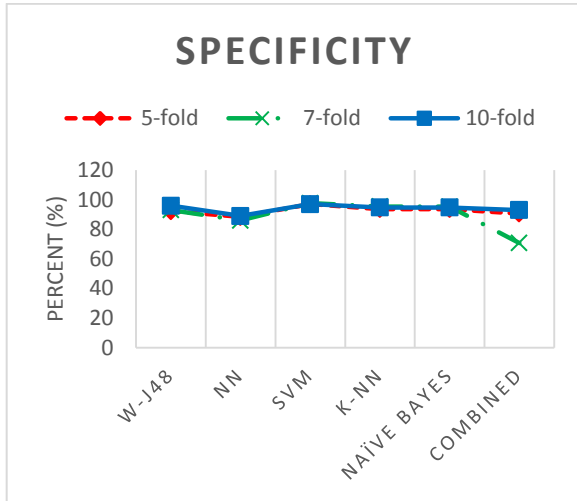


Fig. 3. Specificity indicator for predictive models

TABLE I. PERFORMANCE INDICATORS FOR DIFFERENT METHODS IN IMPLANTATION WITH 10-FOLD CROSS VALIDATION

Model	Accuracy	Sensitivity	Specificity
W-J48	89.31%	67.17%	95.92%
Neural Network	82.63%	61.33%	88.99%
SVM	88.00%	57.69%	97.09%
K-NN	84.05%	48.08%	94.77%
Naïve Bayes	85.30%	53.85%	94.77%
Proposed Method	90.22%	80.50%	93.01%

As shown above, the accuracy indicator of the proposed method with value of 90.22% is higher than prediction of W-J48 as most accurate single classifier with value of 89.31%.

From the viewpoint of sensitivity, the proposed method also gives better predictions than single component classifiers. The value of sensitivity indicator for the proposed model is 80.50% while it is 67.17% for W-J48 as the best single classifier.

However, from the viewpoint of specificity, the SVM model has highest value with prediction of 97.09% while the proposed model stands on the fifth rank before neural network model.

As another method for comparing models we can use ROC¹ diagram. ROC is graphical schematic which shows the performance of classifiers. It shows the False Positive Rate (FPR) versus True Positive Rate (TPR)³⁰.

The TPR defines the number of correct positive results that occur among all positive samples available during the test.

$$TPR = \frac{TP}{TP + FN} \quad (6)$$

On the other hand, FPR defines how many incorrect positive results occur among all negative samples which are available during the test.

$$FPR = \frac{FP}{FP + TN} \quad (7)$$

An ROC space is defined by FPR and TPR as x and y axes respectively. It shows relative trade-offs between false positive (costs) and true positive (benefits). TPR is equivalent to sensitivity and FPR is equal to 1-specificity^{31, 32}.

The more the surface under ROC curve, the more efficiency of the algorithm.

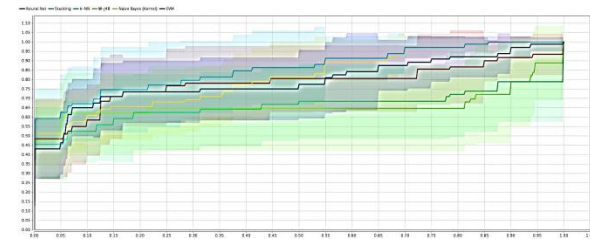


Fig. 4. ROC diagram for predictive models

In the above diagram, the six models of W-J48, Neural Network, SVM, K-NN, Naïve Bayes, and ensemble-based proposed model are shown with olive green, purple, red, green, yellow, and blue colors, respectively. From the above diagram, it is clear that the combined predictive model acts better than the other models.

V. CONCLUSION

In this study, we followed two important purposes. One is to show whether combination of algorithms has higher performance than the singular ones. The other purpose is to increase the prediction of implants which are not successful. This item is described by sensitivity indicator.

In order to substantiate the first purpose, we used five prediction classifiers using patients' data and then combined them in a way to achieve higher performance. The results of our study showed that the hybrid algorithm gains higher accuracy than using only one singular algorithm for classification of records. Also, it increased the sensitivity indicator significantly and since it is very important to identify the patients whose implant is unsuccessful, hence the second purpose of this study has been achieved.

ACKNOWLEDGEMENT

The authors are grateful to the School of Dentistry of Tehran University for Experimental Data Collection. Many people greatly contributed to this work. Special mention goes to Dr. Mohammad Javad Kharrazi Fard and Dr. Razieh Sadat Moayeri.

¹Receiver Operating Characteristic

REFERENCES

- [1] Ceramic Industry, Dental implants and prosthetics market continues growth, 2012, [Available at: <http://www.ceramicindustry.com/articles/92515-dental-implantsand-prosthetics-market-continues-growth>].
- [2] Mangano C, Piattelli A, Lezzi G, Mangano A, La Colla L. Prospective clinical evaluation of 307 single-tooth Morse taper-connection implants: a multicenter study. *The International Journal of Oral and Maxillofacial Implants*. 2010. 25 (2): 394-400
- [3] Wang T, Trappey C, Hoang S, Trappey A. Constructing a dental implant ontology for domain specific clustering and life span analysis. *Advanced Engineering Informatics*. 2013. 27(3): 346-357
- [4] Jung RE, Pjetursson BE, Glauser R, Zembic A, Zwahlen H, Lang NP. A systematic review of the 5-year survival and complication rates of implant supported single crowns. *Clinical Oral Implants Research*. 2008. 19(2): 119-130
- [5] Liao S, Chu P, Hsiao P. Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*. 2012. 39(12): 11303-11311
- [6] Yang P, Yang Y, Zhou B, Zomaya A. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics*. 2010. 5(4): 296-308
- [7] Whalen S, Pandey G. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics, Icahn School of Medicine at Mount Sinai. New York, USA. 2013
- [8] Altmann A et al. Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy. *PLOS ONE*. 2008. 3(10): 3470
- [9] Chen Y, Wong M, Li H. Applying Ant Colony Optimization to configuring stacking ensembles for data mining. *Expert Systems with Applications*. 2014. 41(6): 2688-2702
- [10] Boonstra P, Taylor J, Mukherjee B. Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics*. 2013. 14(2): 259-272
- [11] Wey A, Connett J, Rudser K. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*. 2015. 16(3): 537-549
- [12] Oliviera A, Baldisserotto C, Baldisserotto J. A Comparative Study on Machine Learning Techniques for Prediction of Success of Dental Implants. *MICAI 2005: Advances in Artificial Intelligence*. 2005. 3789: 939-948
- [13] Braga et al. Decision Model to Predict the Implant Success. *Computational Science and Its Applications*. 2012. 7333: 665-674
- [14] Korting T. C4.5 algorithm and Multivariate Decision Trees: IEEE Computer Society. 2013
- [15] Rokach L, Maimon O. *Data Mining with Decision Trees, Theory and Applications*. 2nd Edition: World Scientific Publishing Company. 2014
- [16] Sonali B, Wankar P. Research Paper on Basic of Artificial Neural Network. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2014. 2(1): 96-100
- [17] SanieiAbadeh M, Mahmoudi S, Taherparvar M. *Applied Data Mining*. 2nd Edition: NiazDanesh. 2012
- [18] Huang X. Support Vector Machine Classifier with Pinball Loss. *Pattern Analysis and Machine Intelligence*. 2013. 36(5): 984-997
- [19] Rahulamathavan Y, Phan R, Veluru S, Cumanan K. Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud. *Dependable and Secure Computing*. 2013. 11(5): 467-479
- [20] Zhang X, Qiu D, Chen F. Support vector machine with parameter optimization by a novel hybrid method and its application to fault diagnosis. *Neurocomputing*. 2015. 149(B): 641-651
- [21] Imandoust S, Bolandraftar M. Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *International Journal of Engineering Research and Applications*. 2013. 3(5): 605-610
- [22] Manjusha K, Sankaranarayanan K, Seena P. Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2014. 4(1): 864-868
- [23] Kaur G, Oberai N. Naïve Bayes Classifier with Modified Smoothing Techniques for Better Spam Classification. *International Journal of Computer Science and Mobile Computing*. 2014. 3(10): 869-878
- [24] Polikar R. Ensemble Based Systems in Decision Making. *Circuits and Systems Magazine*. 2006. 6(3): 21-45
- [25] Gehi B, Siddavatam I. Network Bandwidth Predictive analysis using Stacking. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2014. 2(8): 2166-2170
- [26] Dzeroski S, Zenko B. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*. 2004. 54(3): 255-273
- [27] Jung Y, Hu J. A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*. 2015. 27(2): 167-179
- [28] Abdalla M. Evaluation of Data Mining Classification Models. *IUG Journal of Natural and Engineering Studies*. 2014. 22(1): 151-165
- [29] Hauser R et al. Sensitivity, Specificity, Positive and Negative Predictive Values and Diagnostic Accuracy of DaTscan™ (Ioflupane I123 Injection): Predicting Clinical Diagnosis in Early Clinically Uncertain Parkinsonian Syndrome. *Journal of Neurology & Stroke*. 2014. 1(1): 1-13
- [30] Rehman A, Khanum A. Swarm Optimized Fuzzy Reasoning Model (SOFRM) for Diabetes Diagnosis. *Life Science Journal*. 2014. 11(3): 42-49
- [31] Hernandez J. ROC Curves for Regression. *Pattern Recognition*. 2013. 46(12): 3395-3411
- [32] Till R, Hand D. A Simple Generalization of the Area under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*. 2001. 45(2): 171-186