The Utilisiation of composite Machine Learning models for the Classification of Medical Datasets For Sickle Cell Disease

Mohammed Khalaf[1], Abir Jaafar Hussain[1], Robert Keight[1], Dhiya Al-Jumeily[1], Russell Keenan[2], Paul Fergus[1] and Ibrahim Olatunji Idowu[1]
[1]Faculty of Engineering and Technology, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK
[2]Liverpool Paediatric Haemophilia Centre, Haematology Treatment Centre, Alder Hey Children's Hospital, Eaton Road, West Derby, Liverpool L12 2AP, UK
M.I.Khalaf@2014.ljmu.ac.uk, {a.hussain, d.aljumeily, P.Fergus}@ljmu.ac.uk, R.Keight@2015.ljmu.ac.uk, Russell.keenan@alderhey.nhs.uk, I.O.Idowu@2009.ljmu.ac.uk.

***Abstract* —The increase growth of health information systems has provided a significant way to deliver great change in medical domains. Up to this date, the majority of medical centres and hospitals continue to use manual approaches for determining the correct medication dosage for sickle cell disease. Such methods depend completely on the experience of medical consultants to determine accurate medication dosages, which can be slow to analyse, time consuming and stressful. The aim of this paper is to provide a robust approach to various applications of machine learning in medical domain problems. The initial case study addressed in this paper considers the classification of medication dosage levels for the treatment of sickle cell disease. This study base on different architectures of machine learning in order to maximise accuracy and performance. The leading motivation for such automated dosage analysis is to enable healthcare organisations to provide accurate therapy recommendations based on previous data. The results obtained from a range of models during our experiments have shown that a composite model, comprising a Neural Network learner, trained using the Levenberg-Marquardt algorithm, combined with a Random Forest learner, produced the best results when compared to other models with an Area under the Curve of 0.995.**

Keywords —Machine Learning Algorithm; Sickle Cell Disease; Real-time data; Receiver Operating Characteristic Curve; The Area Under Curve; E-Health

I. INTRODUCTION

Sickle cell disease (SCD) is considered long-term disease that facing medical sector, affects human from early childhood. It has a high severe impact on the life expectancy and the patient's quality of life due to red blood cells (RBCs) abnormality. This kind of disease affects the vast majority of patients suffering from genetic blood disorders, where the haemoglobin is measured as abnormal behaviour in the blood. The major cause for the disease within affected populations lies with a group of ancestral disorders that have resulted in a protein mutation inside the RBC called haemoglobin. The common symptoms of SCD are caused by the abnormality of haemoglobin, the major protein inside RBCs which carries oxygen from the lungs and passes it to the human body. The symptoms potentially lead to serious infections, damage to some organs, and tissue death. Normally, the shape of RBCs are flexible and circular, which gives them with more freedom to move through blood vessels. However, the RBCs with those who suffer from SCD may look like a crescent (C-shaped). Moreover, Crescent RBCs usually break down or die more quickly than normal RBCs, resulting in anaemia.

With respect to sickle cell disease, the most significant symptoms that could show effects on patients are shortness of breath, fatigue, headaches, and dizziness[1]. There are three major types of sickle cell disease. The first common one is called Hb SS, when patients inherit sickle cell genes from both parents [2]. The second is a disease is called Hb SC, the patient usually inherits the sickle cell gene (S) and the second gene (C), produced from an abnormal kind of haemoglobin [3]. Finally, in S-beta thalassemia, the patient inherits one gene of sickle cell and beta thalassemia can be inherited from anaemia.

Recent research has demonstrated the positive effects of a drug called hydroxyurea/hydroxycarbamide in terms of modifying the disease phenotype [4]. The clinical procedure to manage SCD modifying therapy is significantly time consuming and difficult for medical staff. In order to curtail the significant medical variability presented by such difficult crisis, healthcare consultants need to improve adherence to therapy, which is regularly poor and subsequently results in fewer benefits and elevated risks to patients.

Up to the present, the new trend of machine learning is becoming essential for the analysis of data within the medical domain. This kind of techniques are being utilised in order to analysis the significance of healthcare factors in association with their integrations for prognosis, for instance, to overall patient management, for providing therapy and support, and the most important to predict the disease progression. [5].

The main objective of these enhancement is to develop the use of advanced technology in healthcare organisations [6]. Machine learning models have been developed and used to enhance medical decision support systems. Neural Network approaches is considerably used within classification techniques for medical fields. This is attributed to their features of parallel processing tools, self-organization, self-learning methods, and non-linearity [7].

The rest of this paper is structured as follows. Section 2 discusses the methodology of our experiments. Sections 3

illustrates the complete results. Finally, Section 4 shows the conclusion and future work.

II. METHODOLOGY

The most related work that shows in the field of machine learning architectures have been constrained to predict and check the severe crises of sickle cell disease, rather than using advance predication to provide accurate amounts of hydroxyurea/hydroxycarbamide in modifying the disease phenotype [8]. Presently, there is no standardisation of disease modifying therapy management. Through using the proposed computerised comprehensive management system, the aim is to develop a reproducible and optimised standard of care in various clinical settings across the UK, and indeed internationally. The main aim of this research is to utilise recent advances in machine learning approaches, in order to support the healthcare professionals in giving accurate amounts of medication for each individual patient according to their medical conditions. In this case, and due to the pattern of the SCD dataset, this research attempts to propose classification of the patient's data set records at an earlier stage, according to how much of a dose the patient will need to take. This can potentially avoid unnecessary admission to hospitals or special institutions, lower the costs significantly, mitigate patient illness before it gets worse over time, particularly with elderly people, unnecessary interventions, and to improve patient welfare.

There are different ways can obtain a better classification. Machine learning approaches can be applied to build strong integrated classifiers to use training and testing datasets, involving past patient observed cases that have been collected from the local hospital for Sickle Cell Disease in the city of Liverpool, UK, over the last ten years for 1168 patient records.

A. Data Collection

The original datasets that are used in our study for SCD were gathered within a ten-year period. Datasets were simulated using real-life data sets provided by local hospital when patient takes blood test sample, which contains a set of datasets that used for checking patients' condition. Each sample contains 12 features deemed important factors for predicting the SCD trait as showed in Table 1. These 12 features considered very important for SCD patients that clinicians mostly concentrate on to provide a suitable amount of medication dosage to tackle patient's condition. In order to work with a large amount of data, a local hospital has supported this research with a number of patient records for the purpose of obtaining better services and accuracy. The resulting dataset comprised 1168 sample points, with a single target variable describing the hydroxyurea/hydroxycarbamide medication dosage in milligrams. To facilitate our classification study, the target dosage was discretised into 6 bins, denoted classes 1 through 6. It formed through dividing the output range (in Milligrams) into membership intervals of equal size: Class 1: [$148 \leq Y < 427$mg], Class 2: [$427 \leq Y < 659$mg], Class 3: [$659 \leq Y \leq 937$mg], Class 4: [$937 \leq Y < 1201$mg], Class 5: [$1201 \leq Y < 1453$mg], Class 6: [$1453 \leq Y \leq 1700$mg].

TABLE I. CHARACTERISTICS OF SCD DATASETS

No	Types of Attributes
1	Weight
2	Haemoglobin(Hb)
3	Mean Corpuscular Volume (MCV)
4	Platelets(PLTS)
5	Neutrophils (white blood cell NEUT)
6	Reticulocyte Count (RETIC A)
7	Reticulocyte Count (RETIC %)
8	Alanine aminotransferase (ALT)
9	Body Bio Blood (BIO)
10	Hb F
11	Bilirubin (BILI)
12	Lactate dehydrogenase (LDH)

B. Experimental Setup

The models under study are composed of two types of integrated Machine learning algorithms: hybrid Neural Network with Levenberg-Marquardt learning algorithm [9], and Random Forest [9], combined using a further Levenberg neural network (H1), and hybrid Levenberg-Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis (H2). The competing model set is composed of a trained models using the Levenberg-Marquardt learning algorithm (LEVNN), a random forest classifier (RFC), Functional link neural network (FLNN) [10] and a Trainable decision tree Classifier (TREEC) [11]. Cross validation technique were used in these experiments for assessing how the outcomes of a statistical analysis could generalise to an independent datasets. The proportion of the datasets were divided into three part training, validation, and testing phase. This study used 10 fold-cross validation to find an average percentage of the correct classifications. The training set received 70 %, validation set 10%, while the testing set obtained 20% to estimate the accuracy and performance of the whole models. The main purpose of using cross validation is to avoid overfitting. In this context, this provides a great support, which indicate when more training is not producing optimal generalisation. However, the models can't be overfitting due to the large number of datasets.

These models are considered to be strong non-linear classifiers and are appropriate to act as comparators of high accuracy and performance. The linear model used includes a linear transformation function with a single layer neural network at each class output unit. To obtain performance estimates for the respective models, this study ran each simulation 50 times and calculated the mean of the responses. The full set of models used in the experiments are described in Table 2. The random oracle model (ROM) is used to establish random case performance through the assignment of random responses for each class [12]. This study is based on multi-classes, which use 6 classes to classify the amount of medication dosage as mention earlier in the data collection section.

TABLE II. CLASSIFICATION MODELS

Models	Description	Architecture	Training Algorithm	Role
ROM	Random Oracle Model	Pseudorandom number generator	N/A	Random Guessing Baseline
LEVNN	Levenberg-Marquardt learning algorithm	Context Units: One context unit for each output unit.	Levenberg-Marquardt	Non-linear Comparison Model
TREEC	Trainable decision tree Classifier	13 inputs, 3 outputs	Decision tree induction	Non-linear Comparison Model
RFC	Random Forest, Decision Tree Ensemble Classifier	Units: 13-30-3, tansig activations	Random feature bagging	Non-linear Comparison Model
H1	Levenberg-Marquardt learning algorithm and Random Forest, combined using Levenberg neural network	Hybrid	Gradient descent with momentum and adaptive learning rate backpropagation	Test model
H2	Levenberg-Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis	Hybrid	Gradient descent with momentum and adaptive learning rate backpropagation	Test model
FLNN	Functional link neural network	Units: 13-30-3, tansig activations.	Gradient descent with momentum and adaptive learning rate backpropagation	Test model
LNN	Linear Combiner Network	Units: 13-3, linear activations	Batch training with weight and bias learning rules	Linear Comparison Model

TABLE III. PERFORMANCE METRIC CALCULATIONS

Metric Name	Calculation
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Precision	TP/(TP+FP)
F1 Score	2 * (Precision*Recall)/(Precision + Recall)
Youden's J statistic (J Score)	Sensitivity + Specificity − 1
Accuracy	(TP+TN)/(TP+FN+TN+FP)
Area Under ROC Curve (AUC)	0 <= Area under the ROC Curve <= 1

Our classifier evaluation involves sensitivity, specificity, precision, the F1 score, Youden's J statistic, and whole classification performance and accuracy calculated as illustrated in Table 3. Additionally, the classifiers were characterised using the Area under the Curve (AUC) and Receiver Operating Characteristic (ROC) plots and, where the classification capability through all functional points was determined.

III. RESULTS

This section analyse the outcomes from the experiments as listed in Tables 4 and 5, showing results for training and testing of the classifiers, respectively. To clarify the results, this study provides further performance visualisations through the use of ROC plots (Figures 1 and 2) and the use of AUC comparison plots as illustrated in Figures 3 and 4. The results obtained from the experiments show that the chosen dataset exhibits significant non-linear relationships, presenting a challenge for the test models. Of the hybrid classifiers under study, the H2 outperformed the H1 as shown in Table 5, demonstrating capability both in fitting the training data and also in generalising to unseen examples. The calculated means of AUCs for the H2 model, obtained for six classes during training yielded an area of 1 (ideal), in comparison to 0.995 over the test sample. Classes 1 to 6 were found to show excellent performance and consistent generalisation from the training to the test sets for this model. It was found that the LEVNN model, a standalone neural network, was able to compete with the composite models H1 and H2, yielding an average AUC of 0.991, slightly outperforming the H1 model and showing an overall rank of second place. The H1 model yielded an average AUC of 0.988, ranking third overall, outperforming the lower ranking models by a reasonable margin. All three of the top performing models, the H2, LEVNN, and H1, obtained nearly ideal AUCs and represent viable candidates for future use.

This research found that the classifiers FLNN, RFC, TREEC, constituted the next performance level obtained, with test AUCs of 0.972, 0.961, and 0.95 respectively. Of these three models, the FLNN model achieved the largest AUC, with results generalising reasonably from the training to the test set. In contrast, the RFC and TREEC models, both founded on decision tree primitives, produced an excellent fit to the training data while failing to generalise to the same degree for the test set, indicating that a degree of overfitting may have occurred. In fact, the TREEC performed ideally and the RFC near-ideally in terms of fitting to the training set, in contrast with the FLNN, which stopped short of an ideal fit, yet maintained consistent generalisation. Additionally, the three aforementioned models can be contrasted with the H2, LEVNN, and H1 models, which produced exceptional results in terms of both training and generalisation.

The average test of AUCs for this LNN ranged 0.849, which is seen to demonstrate performance significantly below that of the other models. The ROM is indicated to follow the diagonal of the ROC plots for all classes (see Figures 1 and 2), demonstrating by compare the importance of the outcomes from the rest of trained classifiers. This process of guessing yields both train and test set AUCs lower than the LNN trained model baseline.

TABLE IV. THE MEAN PERFORMANCE FOR CLASSIFIER (TRAINING)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
ROM	0.493	0.542	0.179	0.253	0.0351	0.543	0.488
LEVNN	1	1	1	1	1	1	1
TREEC	1	1	1	1	1	1	1
RFC	0.999	0.999	0.994	0.996	0.997	0.999	1
H1	1	1	1	1	1	1	1
H2	1	1	1	1	1	1	1
FLNN	0.954	0.964	0.809	0.873	0.918	0.963	0.985
LNN	0.854	0.827	0.524	0.642	0.68	0.835	0.87

TABLE V. THE MEAN PERFORMANCE FOR CLASSIFIER (TESTING)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
ROM	0.548	0.547	0.197	0.268	0.0949	0.539	0.524
LEVNN	0.984	0.993	0.974	0.978	0.977	0.992	0.991
TREEC	0.913	0.926	0.684	0.737	0.839	0.921	0.95
RFC	0.892	0.918	0.648	0.741	0.81	0.915	0.961
H1	0.977	0.995	0.98	0.978	0.971	0.992	0.988
H2	**0.989**	**0.99**	**0.94**	**0.961**	**0.979**	**0.989**	**0.995**
FLNN	0.95	0.966	0.803	0.862	0.916	0.962	0.972
LNN	0.824	0.848	0.534	0.641	0.672	0.84	0.849

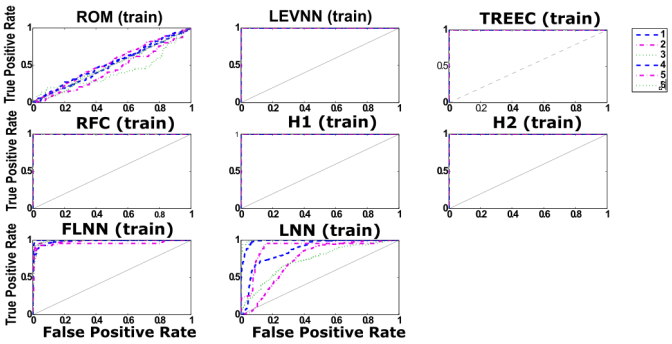

Fig. 1. ROC curve (Train) for classifiers

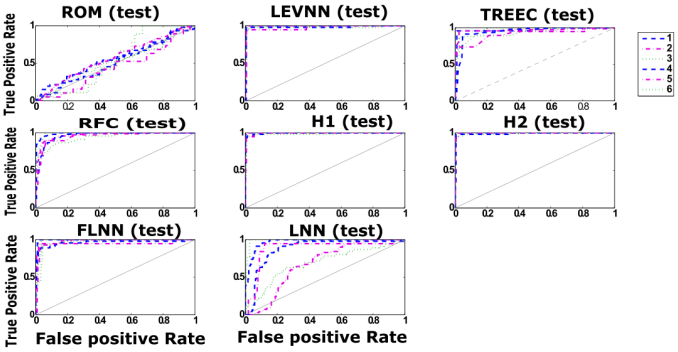

Fig. 2. ROC curve (Testing) for classifiers

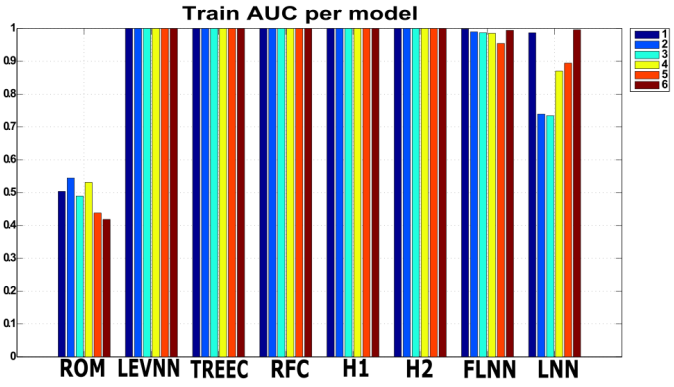

Fig. 3. Train AUC per model

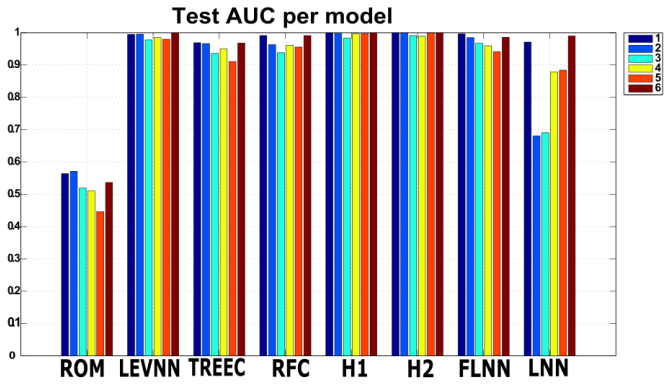

Fig. 4. Test AUC per model

Prior to proceeding to data modelling, the data representation was explored to investigate if any regularities could be uncovered within its structure. Additionally, the exploratory phase was used as a means of exposing any outliers and other questionable artefacts in the data if such defects were present, such that the results of later analysis would not be invalidated due to unsound input. Exploratory analysis is an important step in the machine learning approach, allowing the human advisor to gain an intuition of the data and also the potential learnability of such data. The results from data exploration can be used to guide the modelling phase, since a major component learnability is known to be a function of the correspondence between the learning algorithm and the type of representation it is supplied with. To undertake an exploration of the utilised data in these experiments, it is computed summary statistics, followed by visualisation methods including Stochastic Proximity Embedding (SPE), t-distributed Stochastic Neighbourhood Embedding (tSNE), and Principal Component Analysis (PCA) as reported in Figures 5, 6 and 7, respectively. Results from the visualisation procedures reveal that some discernible structure is present within the data. The PCA plot shows that there are possible clusters of values present within the data, a finding which is further elaborated through the tSNE plot, demonstrates that the data can be reasonably geometrically separated when considering various intervals of dosage level. Furthermore, the exploratory analysis shows no obvious defects that could call into question the results of subsequent analysis.

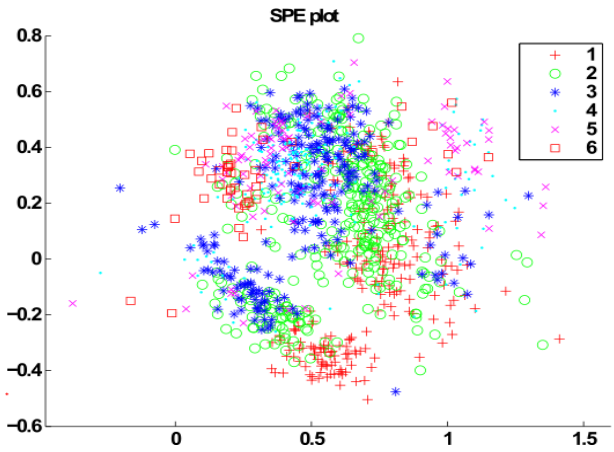

Fig. 5. SPE plot

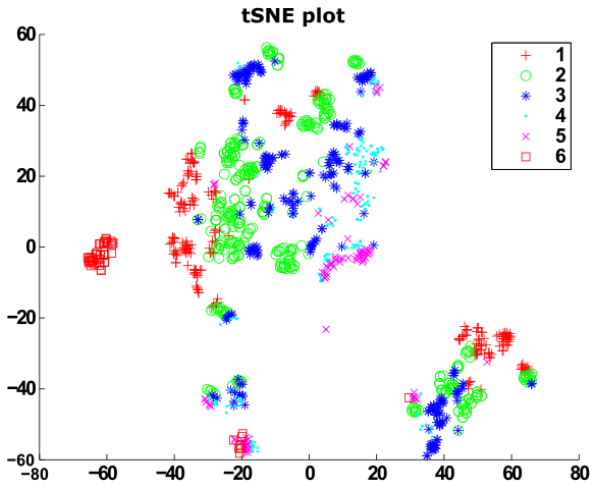

Fig. 6. tSNE plot

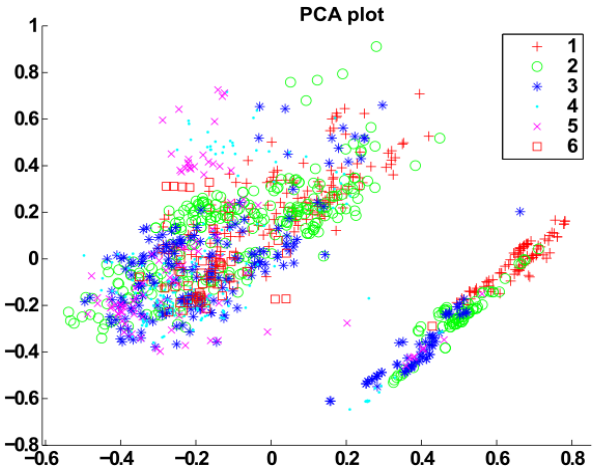

Fig. 7. PCA plot

IV. CONCLUSION

This study presents an experimental investigation into the use of different neural network algorithms to classify the level of dosage for SCD. In this paper, various model architectures are used for analysing the medical datasets obtained from SCD patients. The main purpose of this research is to examine the effectiveness of these models in terms of training and testing setting, investigating if such architectures could enhance classification results. It was discovered through experimental investigation, comprising the usage of patient sample data and approaches such as H1, H2, LEVNN and RFC, TREEC, and FLNN, that the analysis of medical data for the SCD objective is viable and yields precise results. The results obtained from a range of models during our experiments have shown that the proposed hybrid Levenberg-Marquardt learning neural network and Random Forest, combined using Fischer discriminate analysis (H2) generated significantly better outcomes over the other range of classifiers. The local hospital has supported this research with a more than 1100 sample examples for the purpose of obtaining better services and accuracy. Further research will be recommended to implement to make confirmation on our findings, where a large number of data could be utilised also to advance the performance of the results. This study has indicated only a certain number of neural networks architectures. In this circumstance, it is recommended however that a machine learning algorithms, for instance, ADABoost classifier and Automatic neural network classifier could be used to increase the scale and scope of this research.

REFERENCES

[1] D. J. Weatherall, "The role of the inherited disorders of hemoglobin, the first “molecular diseases,” in the future of human genetics," Annual review of genomics and human genetics, vol. 14, pp. 1-24, 2013.

[2] D. J. Weatherall, "The importance of micromapping the gene frequencies for the common inherited disorders of haemoglobin," British journal of haematology, vol. 149, pp. 635-637, 2010.

[3] V. Marsh, F. Kombe, R. Fitzpatrick, T. N. Williams, M. Parker, and S. Molyneux, "Consulting communities on feedback of genetic findings in international health research: sharing sickle cell disease and carrier information in coastal Kenya," BMC medical ethics, vol. 14, p. 41, 2013.

[4] M. Kosaryan, H. Karami, M. Zafari, and N. Yaghobi, "Report on patients with non transfusion-dependent β-thalassemia major being treated with hydroxyurea attending the Thalassemia Research Center, Sari, Mazandaran Province, Islamic Republic of Iran in 2013," Hemoglobin, vol. 38, pp. 115-118, 2014.

[5] G. D. Magoulas and A. Prentza, "Machine learning in medical applications," in Machine Learning and its applications, ed: Springer, 2001, pp. 300-307.

[6] H. Adams, "Medical Informatics: Computer Applications in Health Care," JAMA, vol. 265, pp. 522-522, 1991.

[7] B. Liu, M. Wang, L. Yu, Z. Liu, and H. Yu, "Study of Feature Classification Methods in BCI Based on Neural Networks," in Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, 2006, pp. 2932-2935.

[8] J. N. Milton, V. R. Gordeuk, J. G. Taylor, M. T. Gladwin, M. H. Steinberg, and P. Sebastiani, "Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models," Circulation: Cardiovascular Genetics, vol. 7, pp. 110-115, 2014.

[9] M. Panella, S. Marchisio, and F. Di Stanislao, "Reducing clinical variations with clinical pathways: do pathways work?," International Journal for Quality in Health Care, vol. 15, pp. 509-521, 2003.

[10] S. Dehuri, R. Roy, S.-B. Cho, and A. Ghosh, "An improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification," Journal of Systems and Software, vol. 85, pp. 1333-1345, 2012.

[11] D. Lavanya and K. U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," IJCA) International Journal of Computer Applications, vol. 26, 2011.

[12] X.-Y. Jia, B. Li, and Y.-M. Liu, "Random oracle model," Ruanjian Xuebao/Journal of Software, vol. 23, pp. 140-151, 2012.