# Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models

Mohammed Khalaf[a], Abir Jaafar Hussain[a,*], Robert Keight[a], Dhiya Al-Jumeily[a], Paul Fergus[a], Russell Keenan[b], Posco Tso[a]

[a] *Applied Computing Research Group, Faculty of Engineering and Technology, Liverpool John Moores University, Liverpool, L3 3AF, UK*
[b] *Liverpool Paediatric Haemophilia Centre, Haematology Treatment Centre, Alder Hey Children's Hospital, Eaton Road, West Derby, Liverpool L12 2AP, UK*

## ABSTRACT

This paper discusses the use of machine learning techniques for the classification of medical data, specifically for guiding disease modifying therapies for Sickle Cell. Extensive research has indicated that machine learning approaches generate significant improvements when used for the pre-processing of medical time-series data signals and have assisted in obtaining high accuracy in the classification of medical data. The aim of this paper is to present findings for several classes of learning algorithm for medically related problems. The initial case study addressed in this paper involves classifying the dosage of medication required for the treatment of patients with Sickle Cell Disease. We use different machine learning architectures in order to investigate the accuracy and performance within the case study. The main purpose of applying classification approach is to enable healthcare organisations to provide accurate amount of medication. The results obtained from a range of models during our experiments have shown that of the proposed models, recurrent networks produced inferior results in comparison to conventional feedforward neural networks and the Random Forest model. For our dataset, it was found that the Random Forest Classifier produced the highest levels of performance overall.

## 1. Introduction

Sickle cell disease (SCD) is a common serious genetic disease, which has a severe impact on the patient's quality of life and life expectancy due to red blood cell (RBCs) abnormality. SCD is phenotypically complex, with various medical outcomes ranging from early childhood mortality to a nearly unrecognised condition [1]. The main reason 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. According to the World health Organisation (WHO), 7 million new born babies each year suffer either from the congenital anomaly or from an inherited disease [2]. Furthermore, 5% of the population around the globe carries trait genes for the haemoglobin disorder, primarily, thalassemia and sickle cell disease [3]. In S-beta thalassemia, the patient inherits one gene of sickle cell and beta thalassemia can be inherited from anaemia.

SCD affects more than 1 million individuals in USA and there are over 75,000 hospitalisations costing approximately £300 million per year for treatment of SCD complications [4]. According to a National Health Services (NHS) investigation report, there are 250,000 people with sickle cell disease in the United Kingdom alone [5]. Moreover, the estimated cost in 2013 to admit patients to hospital reached more than £23.8 million per annum [6].

In the case of SCD, recent research has shown the beneficial effects of a drug called hydroxyurea/hydroxycarbamide in modifying the disease phenotype [7]. The clinical management of this disease modifying therapy is difficult and time consuming for clinical staff. In order to address the significant medical variability presented by such a crisis, healthcare professionals must improve adherence to therapy, which is frequently poor and subsequently results in elevated risks and less benefits to patients.

The development of medical information systems has played an important role in medical societies. The aim of these developments is to improve the utilisation of technology in medical applications [8]. Expert systems and various Artificial Intelligence methods and techniques have been used and developed to improve decision support tools for medical purposes. Machine Learning models (ML) is considered to be a powerful technique in the field of scientific research that enables computers to learn from data [9]. There are a number of machine learning technique for classification include the Artificial Neural Network, the Random Forest model, and the Support Vector Machine. In this paper, the application of machine learning approaches

for the problem of SCD medication dose management is considered.

The reminder of this paper is organized as follows. Section 2 will illustrate the related work, while Section 3 will discuss the classification of medical data, including model descriptions and recurrent neural network architectures used for classification. The methodology will then be introduced in Section 4, followed by the presentation of our results in Section 5. Finally, in Section 6 we discuss our conclusions and future works.

## 2. Related work

In recent years, healthcare organisations worldwide have faced many problems in meeting the demands of advanced medical sectors [10]. The main motivation for researchers is to produce a new system to support for health organisations and consequently to delivery benefits to patients. There are a number of research projects developed for healthcare environments based on machine learning approaches [11]. Several solutions have been proposed to provide support to physicians and medical professionals. Milton et al. [12] developed an ensemble model comprising 14 algorithms to predict genetic risk score (GRS) associated with different configurations of Single Nucleotide Polymorphisms (SNPs). The main purpose of using these 14 algorithms within their experiment is to predict Haemoglobin F (HbF) in sickle disease patients. The datasets collected were based on 814 patients. The datasets were trained and tested over three independent cohorts. The ensemble output of models described 23.4% of the variability in the discovery cohort, whereas the correlation between observed and predicted HbF in the three independent cohorts ranged between 0.28 and 0.44 [12].

Allayous et al. [13] demonstrated a new technique based on machine learning algorithms for quantifying the high risk of an acute splenic sequestration crisis, which is considered a serious symptom of (SCD). In their research, the main aim is to learn how to predict the level of severity depending on the training dataset. The dataset was gathered from "Centre Caribéen de ladrépanocytose" during 10 years for 42 children defined by 15 features. There are a numbers of machine learning methods used in their research that have the ability to evaluate the risk of acute splenic sequestration crisis in terms of classifying patient between sever and mild symptoms. The Area under Curve (AUC) and the Characteristics Receiver Operating Curve (ROC) were used to measure the accuracy of datasets. The highest numbers of accuracy were achieved with Adaboost algorithm with 92%, while the Ranktree algorithm achieved 90%, thus offering better models of diagnostic method.

Solanki [14], proposed machine learning approaches based on WEKA platforms. The research used two models comprising decision trees (J48) and Random tree in order to make a comparison for classifying specific blood groups. The outcome of the study indicated that the Random tree algorithm achieved better accuracy in comparison with other classifier. Our extensive researches indicated that there are no studies that have been applied for classifying SCD datasets for the provision of accurate medication dosage predictions. Currently, all hospitals and healthcare sectors are using manual approaches that depend completely on medical consultants, which can be slow to analyse, time consuming and stressful. However, this study provides a system that shifts from manual input to automated input approaches that can offer better outcomes with reduced error rates.

## 3. Classification

The importance of classification techniques in the medical community, especially for diagnosis purposes, has gradually increased [15,16]. The key reason for improving medical diagnosis is to enhance the humans ability to find better treatments, and to help with the prognoses of diseases to make the diagnoses more efficient [17], even with rare conditions [18]. The aim of the classifier is to learn how to extract useful information from the labelled data in order to classify unlabelled data. Various methods have been employed for the classification task [19]. They are categorised into two groups: linear and nonlinear classifiers. Linear classifiers are represented as a linear function (g) of input features x as illustrated in Eq. (1) [21].

$$g(x) = w^{T}x + b \tag{1}$$

Where w is a set of weighted values, b is a bias, and T refer to matrices transpose. For two classes, problem c1 and c2, the input vector x is assigned to class c1 if $g(x) > =0$ and to class c2, otherwise. The decision boundary between class c1 and c2 is simply linear. In the previous studies, several traditional linear classifiers were designed and applied to perform classification in different areas such as Linear Discriminant Analysis.

Nonlinear classifiers involve finding the class of a feature vector x using a nonlinear mapping function (f), where f is learnt from a training set T, from which the model builds the mapping in order to predict the correct class of the new data. Popular nonlinear classifier is the Artificial Neural Network (ANN) model. As a classifier, ANN has a number of output units, one for each class [20]. Nonlinear neural networks are able to create nonlinear decision boundaries between dissimilar classes using non-parametric approach [21]. Zhang [22] asserted that neural networks have the power to determine the posterior probabilities, which can be used as the basis for establishing the classification rule.

In this study, we consider the use of several classes of model for data classification, Random Forest, Support Vector Machines, and comprising ANN. Additionally, we trial a subclass of ANNs known as Recurrent Neural Networks, which offer greater representational capacity (and conversely increased complexity) in comparison to feedforward ANNs.

The knowledge representation encoded within ANN models is manifested in the form of directed connection weights, which collectively form the network's "program". In order to perform useful tasks, an appropriate configuration of weights must first be found through the use of a learning algorithm. Typically, during this learning procedure, the space of network weights is searched using an optimisation algorithm in search of solution that minimises an error defined according to an objective function of interest. Such an objective function is carefully chosen to facilitate generalisation. The dimensions of variation that contribute to the success of a neural network include the network connectivity pattern (architecture), the activation functions, determination of appropriate weights, and the training data presented to the network during learning. The computation at a single node of an ANN comprises a weighted sum of its inputs, in turn processed according to an activation function. Such a computation is demonstrated in Equation 2, where $y_j$ is the output from the $j$th unit in layer $y$, $w_{ji}$ represents the weight of the $i$th input, $x_i$ represents the value of the $i$th input, and $\sigma$ represents the activation function.

$$y_j = \sigma\left(\sum_{i=0}^{m} w_{ji}x_i\right)$$

ANN achievements have covered different types of medical applications [23]. The growth of medical information has played a significant role in healthcare organisations [24]. The target of these improvements is to develop the usage of technology in medical applications [25]. Various types of Artificial Intelligence (AI) techniques and expert systems have been applied in order to improve decision support tools for health purposes. In this research, we propose to use an advanced neural network algorithm for medical data classification based on machine learning for the purpose of utilising intelligent techniques for analysing the huge amount of SCD datasets. The proposed research aims to use machine learning to compare various algorithms and ensemble models within the medical data context. The main reason for using machine learning algorithms (neural networks) is to extract the