# AI MODEL TO PREDICT CHOLESTEROL AND HAEMOGLOBIN LEVELS IN BLOOD using NEAR INFRA-RED (NIR) SPECTROSCOPY

**Kosala**[1]

*Assistant professor,Department of Software Engineering,East china University Of Technology,China*

*Abstract:* ***Cardiovascular disease are leading cause of death every year,approximately 17.9 million people die due to CVD ,Cholesterol plays a vital role in these kinds of sufferings,this paper proposes AI model to predict the level of cholesterol and haemoglobin , Predicting the risk of cardiovascular disease is the key for primary prevention . Using patient data and other information, AI can help doctors and medical providers deliver more accurate diagnoses and treatment plans.***

***Traditionally, blood examination is done by gathering blood tests from patients and perform various tests on the samples. Our work gains ground towards non-invasive blood examinations by utilizing spectral data from the Near Infra-Red (NIR) frequencies instead of the latter.Most scientists are yet to adopt AI methods for spectral analysis on the account of a high risk of overfitting. One significant issue we faced was building a model that summed up well without overfitting. From our result, scientists will be able to use a provided application to input NIR spectroscopic data and receive relevant predictions.***

***We recommend an AI model to estimate blood and cholesterol levels based on bloodstream test measures. This non-invasive measurement approach is proposed by framing the task as a multi-class and multi-target classification problem. We focus on Machine learning and one Neural Networks techniques to model the absorption data in order to classify the level of cholesterol and haemoglobin as high, low or ok.In this study two models has been build, One using a classical machine learning and the other a neural network. After the data analysis, we created three classification models for predicting cholesterol low, cholesterol high and haemoglobin with average accuracy between 86 - 93%.***

***Keywords:*** **ArtificialIntelligence,spectralanalysis,AImodel,SVM,NLP,Classificationmodel,Neuralnetwork,ANN,Cardiovascular disease.**

## 1. INTRODUCTION

Data science can change the health care sector in countless ways. The right utilization of data science in healthcare, clinical associations can reduce costs and re-confirmations. Improve healthcare, medication conveyance and clinical decisions.

This makes data science medicine perhaps the main progression over the last decade. In this paper, we will respond to the greatest inquiry “A Machine learning model that detects the level of cholesterol and haemoglobin in blood samples”.

The aim is to give a version of predictive models from base to state-of-art, describing various types of predictive models, steps to develop a predictive model on Machine learning
and Neural Network and to reduce the loss of lives and to help the treatment of individuals even before they begin to suffer.

**Cardiovascular disease (CVD) is the leading cause of illness and death worldwide[1]**. Several risk assessment tools have been proposed to accurately predict the risk of CVD, among which the Framinghamriskscore(FRS),pooled cohort equation (PCE), systematic coronary risk evaluation (SCORE), and QRISK3 are widely used[2]– According to the WHO, an estimated 17.9 million people died from heart disease, an estimated 32% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries [5].

The National Academies of Sciences, Engineering, and Medicine estimates that some 12 million people receive incorrect diagnoses, sometimes with life-threatening consequences [4].

The Data scientist is a multi-abilities individual who can predict a phenomena dependent on data utilizing model. This word "Prediction" is exceptionally helpful for healthcare field, for instance dependent on data gathered from clients utilizing various sources, we can predict that this area of world will experience the ill effects of that infection, and for this situation we can move forward to treat it and this model offers the essential data with specialists so they can make a move before the circumstance gets critical.

Generally, we do blood examination by gathering blood tests from patients here we do gain ground towards painless blood examinations and perform various tests on the examples. For this purpose, we build AI models that can classify the degree of specific chemical compounds in sample from their spectroscopic information. Machine learning and Neural Network have become essential techniques in predicting medical results. Though our dataset has multi-class and multi-target classification problem. To work on healthcare dataset, we use SVM, time series analysis, NLP [3]. The frequent model utilised by the research papers we go through are SVM, KNN, Random Forest, Naïve Bays, Artificial Neural Network, etc.

The significant advantages of predictive analytics are to check the proficiently screens and investigates the demand for drug strategies. Predicts a patient's condition and proposes preventive measures. Gives quicker documentation of emergency clinic data. Helps in proficiently using specialists and different assets to support the most extreme number of patients. Predicts the future clinical emergencies of a patient. Medical data are sensitive so, our model is needed to properly extract the correct information from the data.

## 2. Objectives and Contributions

Our goal is to construct an AI model that distinguishes the level of cholesterol and haemoglobin in blood tests. To achieve this, we are going to apply different ML techniques to achieve a better accuracy in prediction.

We analysed existing invasive, insignificantly invasive, and non-invasive methodologies for blood haemoglobin level measurement determined to suggest data collection techniques, signal extraction processes, include computation systems, hypothetical establishment, and AI algorithms for developing a non-invasive haemoglobin level estimation point-of-care tool utilizing a smartphone [6].

We explored research papers connected with invasive, insignificant invasive, and non-invasive haemoglobin level measurement processes [7]. We researched the difficulties and chances of every strategy. We thought about the variety in data collection sites, bio signal handling procedures, hypothetical establishments, photoplethysmogram (PPG) signal and features extraction process, AI algorithms, and prediction models to work out haemoglobin levels [6]. This analysis was then used to prescribe practical ways to deal with a smartphone-based point-of-care tool for haemoglobin measurement in a non-invasive manner [6].

- The Data acquisition from **Bloods.ai**, gather the information on data origin and processed.
- EDA (Exploratory Data Analysis) adopting the Statistics and data Visualization.
- Feature engineering
- Model Training and deployment

# 3. State of the Art

### 3.1 Summary

Medical industry has a large availability of data. Therefore, Machine learning and Neural Networks have become essential techniques in medical decision-making. Their performance is improved according to the quality of the data available, enhancing disease prediction processes. In the medical field, these techniques are being widely used for modelling various human processes. Blood analysis is one of the essential disease detectors as it contains many parameters with different values that indicates the definite proof for a disease's existence in the human body [9].

As mentioned in the introduction, our dataset poses a multi-class and multi-target classification problem. There are many solutions available to resolve this problem. Some of the important multi-class and multi-target classification methods used in blood analysis by some researchers are SVM, KNN, Random Forest, Naïve Bays, and Artificial Neural Network.

The main objective of this research is to use machine-learning techniques for classifying the level of chemical compounds in blood samples; several techniques are performed for finding the most suitable algorithm that maximizes the prediction accuracy [8]. The following part of this section is organized as follows. Section 2.2 outlines the different classification methods on blood disease prediction, their advantage and disadvantage, undertaking multi-label and multi target classification problems. Section 2.3 represents Artificial Neural Network.

### 3.2 classification methods

Vector Machine (SVM), k-Nearest Neighbours (KNN), Random Forest, Naive Bayes, Artificial Neural Networks.

A decision tree classifier is a systematic approach for multiclass classification. SVM doesn't support multiclass classification natively. It supports binary classification and separating data points into two classes. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems [13]. SVM helps figure out if we have a bias or overfitting of the data by regularization parameter. SVM algorithm is not suitable for large data sets. Similarly, KNN classification method also degrades the performance of the algorithm with large datasets.

The computational complexity of Support Vector Machines (SVM) is much higher than for Random Forests (RF). This means that training a SVM takes longer to train than a RF when the size of the training data is higher. Our dataset has plentiful absorption, temperature and humidity data. Therefore, random forests should be preferred when the data set grows larger. It reduces overfitting problem and improves the accuracy.

In one of the articles based on machine learning method used to measure lipids in dried blood spots, the researchers applied Random forest machine learning to lipid profile data generated from human plasma samples to predict the circulating concentration of four clinically important lipoproteins, triglyceride, HDL, LDL and total cholesterol. Applying random forest machine learning to the lipid profile data, researchers were able to obtain a good estimate of TriG, HDL, LDL and total cholesterol concentrations.

Knowing their relative concentration is more important than knowing a patient's exact concentration of HDL, LDL or total i.e., if they fall outside of the healthy range [14]. Researcher estimated concentrations classified people into the correct clinical category with accuracies of 83.9%, 64.9%, 82.2%, and 64.9% for triglyceride, HDL, LDL and total cholesterol respectively [11].

Another effective and pragmatic machine learning algorithm used in Medical diagnosis is Naive bayes algorithm. In 2019 published research paper on Machine learning model for Haemoglobin estimation, researcher considered the execution of Tree Random Forest, J48 choice tree, Naïve Bayes and Lazy IBK. In this paper, they analysed the algorithms deepen on their accuracy, learning time and error rate. Through examination, they infer that Naïve Bayes algorithm has better classification accuracy over other algorithms [17]. Naïve bayes algorithm works on the principles of conditional probability as given by the Bayes' theorem. Naïve bayes used in real time prediction as it's fast and needs less training data. To work on multi class problem using Naïve bayes, one has to compute the probability of each class label in the usual way, then pick the class with the largest probability. Naïve bayes perform well in multi class prediction.

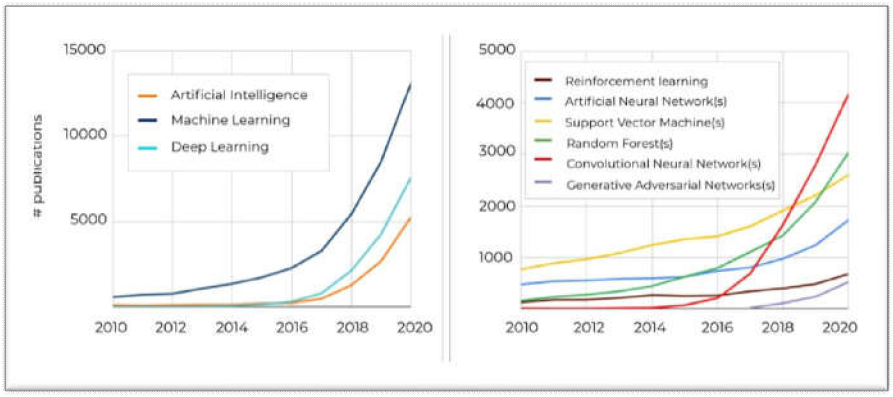


**Figure 1: Number of publications since 2010 - 2020 containing keywords related to AI/ML/DL methods [10]**

In the last decades, thousands of publications reporting the performance of new algorithms and/or original variants of the existing ones. The number of publications which are using some of the most popular ML/DL classification methods is presented in Fig. 1 [12]. We can see from the chart that since 2018, the use of Random Forest and Convolution Neural Network algorithms is rapidly increasing.

The research of these different approaches determines which model would provide the most accurate predictions on the current data. In our research, we are mainly focused on one Machine learning technique and one Neural Network technique to model the absorption data in order to classify the level of Hdl_cholesterol_human, Cholesterol_ldl_human and Haemoglobin(hgb)_human. In our classification methods, we separate classifiers for each target and train them using training data. These classifiers will be used to classify levels of targets (high, low, ok). As per published in the research paper from 2019, the group of researchers has tested several classifiers and calculated accuracies. The comparison of accuracies helps the physicians to select the best performing model to predict the blood diseases according to general blood test [9].

### 3.3 Artificial neural network

We know that in the medical sciences, it is very important to have a good interpretation of data and giving a correct, early diagnosis. Physicians usually suffer from an absence of good, accurate analysis of these laboratory data. They need a tool that would help them to make good decision [12]. The abundance of large datasets in the health-care sector and the interconnected complex relationships between each biological component has encouraged the scientific research community to incorporate Artificial Neural Network (ANN) models. As our dataset has multi-label classification, Neural network models can be configured to support multi-label classification and can perform well, depending on the specifics of the classification task. For multi-label classification, neural networks specify the number of target labels there is in the problem as the number of nodes in the output layer.

## 4. System's Architecture

### 4.1 Summary

The ultimate goal of this piece of work is not only to train models but also to design a tool that is functional and practical. The architecture of this system is depicted in the Fig. 2 as below. The data is collected using a scanner device. The data sets comprising of high quality, well-vetted bio donation scans which is provided to Web Application as an input data. The data is followed by pre-processing. In pre-processing, system performs Data Cleaning, Normalisation, Standardisation, handle Missing values, Encoding, etc. The processed data later feed to train a classification model. The classification model does predictions. The model will make instant prediction for future data without training the model again.

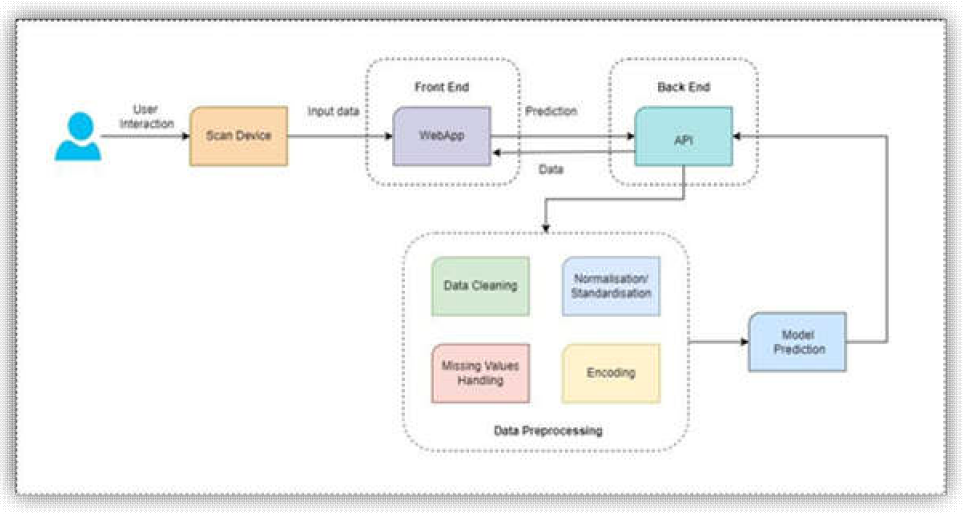


Figure 2: System Architecture

## 5. Methodology

### 5.1 General approach

To solve approach our problem of blood testing from a non-invasive blood analysis standpoint we decided to build two models. One using a classical machine learning approach and the other a neural network. This is because even though classical machine learning is most often the less powerful approach of the two, related work by other

studies in the field show acceptable performance [9]. More so, the latter is easier to explain to medical personnel and requires less complex infrastructure.

**5.2 Dataset description**

No Data, No Machine Learning. To build our model we made use of a carefully collected dataset by the Bloods AI institution containing **29,160** blood scan entries in the training dataset and **3660** entries in the test dataset. These samples were collected by passing Near Infrared light through the blood of donors and measuring the energy absorbed by 170 different wavelengths in the range 900 – 1700nm.
Here is a detailed explanation of the features presented in the table below.

| Parameters | Description |
|---|---|
| **Id** | Unique identifier assigned to each measurement |
| **Absorbance0, Absorbance1, …, Absorbance169** | Quantity of light reflected off of donators fingertip for 170 different wavelengths |
| **Std** | Standard deviation of scanning device |
| **Temperature** | Temperature at the time of measurement |
| **Humidity** | Humidity at the time of the measurement |
| **Donation Id** | Donor identifier |
| **Hdl_cholesterol_human** | Level of cholesterol high. Can be low, ok or high. |
| **Cholesterol_ldl_human** | Level of cholesterol low. Can be low, ok or high |
| **Haemoglobin(hgb)_human** | Level of haemoglobin: Can be low, ok or high |

**5.3Exploratory data analysis**

As displayed by the feature description table, a large part of the predictors consisted of absorbance readings. As a result, our exploratory data analysis focused on these.

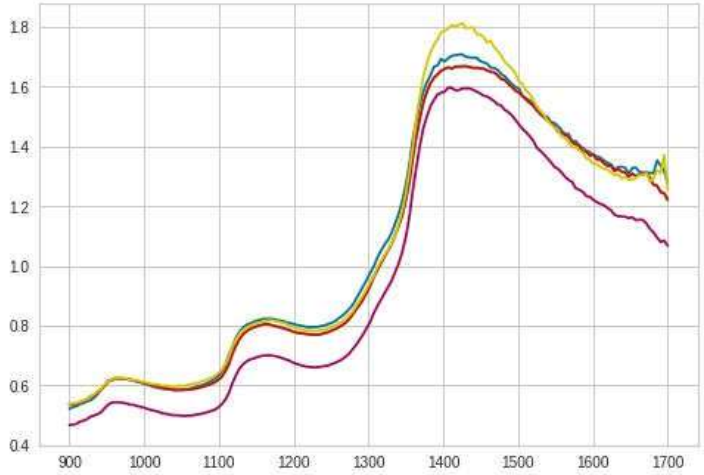


**Figure 3: Wavelength vs Absorbance readings plot for 4 samples**

Plotting the spectrum for four random samples, we observed that the readings got noisier at the edges. For example, in the figure 3 above we see irregularities in the ranges 900-1000nm and 1550-1700nm.This meant that for modelling, we had to leave out these edge wavelengths.

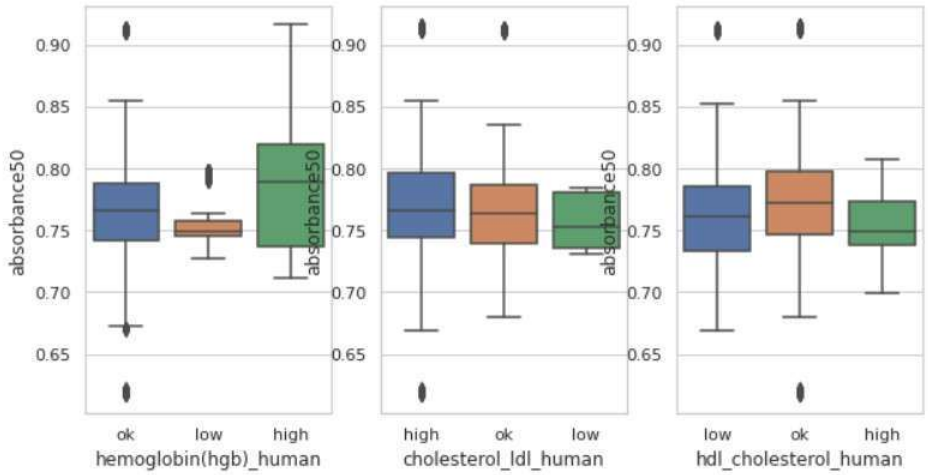


**Figure 4: Absorbance readings vs haemoglobin, cholesterol high, cholesterol low**

We then went ahead to analyse absorbance readings with respect to our three target variables. However, this did not provide any additional insight as readings did not show any clear disparity with respect to our target variables, as shown in the image above.

### 5.4 Modeling

For our classical machine-learning model, we trained three LightGBM models (one per target variable), which is a tree model like random forest but optimised for speed and performance. This decision was key because we anticipate future use cases for the models running on devices with limited computing power such as wearable devices.

Our work on ANNs involved a few steps. First, we trained a neural network **(Autoencoder)** to generate embeddings for each blood scan sample. This also served to reduce the dimensionality of our data. Next, we trained a clustering algorithm to classify these embeddings and hence our dataset. The combination of a Neural network plus Clustering algorithm provided robust a model. However, this performance did not match that of our LightGBMmodel.

## 6. Result

### 6.1 Current Results

At the moment, we have created our three Machine learning classification models (LightGBM) for predicting cholesterol low, cholesterol high and haemoglobin. These models are quick and have decent accuracies ranging from 86 - 93%.We have also trained an Autoencoder to reduce the dimensions of our predictors. And currently seek to determine an optimal embedding length that will give the best results in the clustering phase.

### 6.2 Development Anticipated

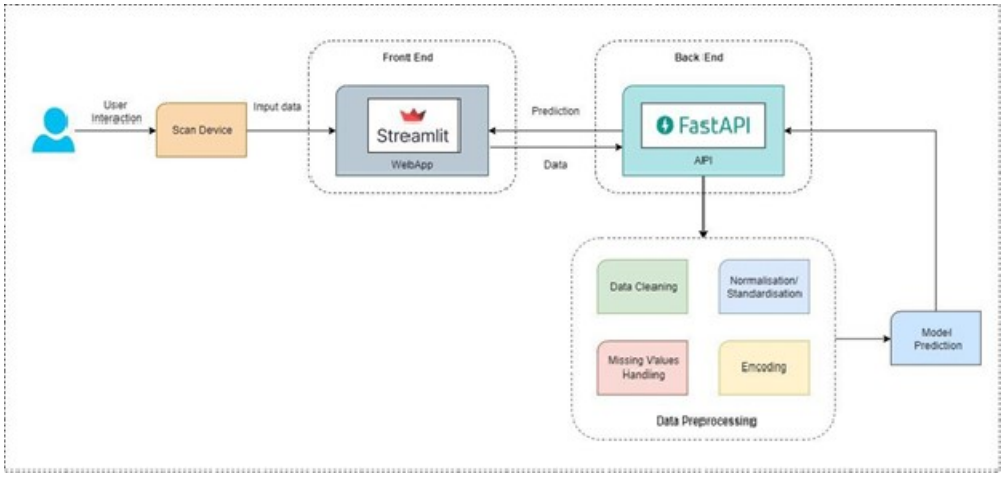


Figure 5: Web Application Design with technologies

As portrayed in the System Architecture section, we intend to build a sleek simple-to-use software to support our work. This will be a Web application at the Frontend powered by an API and our pre-trained models in the Backend. A scientist will be able to use the provided user interface to input **NIR** spectroscopic data and receive relevant predictions.

To build this solution, we will make use of a variety of tools and experiences of the team members. These can be elaborated as follows:

- Programming Language – Python
- Web Interface (Frontend) – Streamlit
- Application Programmable Interface (Backend) – Flask

## 7. Conclusions

This paper aimed to propose a non-invasive approach to blood analysis using artificial intelligence. By providing performant classification models that accomplish this task, we believe that our work takes a step forward in the right direction towards the democratisation of the latter.While the large dimensionality of spectroscopic data limits accuracy of our models, we recommend experimenting with scatter-correction methods and spectral derivatives pre- processing techniques that help with these.Our web application conceived to be used by scientists in labs is a first step that can be iterated upon to build smaller embedded systems. We are convinced blood analysis in the future will be commoditised just as the weighing scale is today.

## 8. References

*[1]Benjamin emelia ,j.et.al, 2019, heart disease and stroke statistics' Report from American heart association https://www.ahajournals. org/doi/10.1161 /CIR .0000000000000659*

*[2]D Agestino, General cardiovascular risk profile for use in primary care: The FraminghamHeartStudy.https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA .107.699579*

*[3] Zeeshan Mushtaq, 2018, When Data Science met Medicine! https://medium.com/the-research-nest/when-data-science-met-medicine-8d3971a0ade9*

*[4] Sharvari Santosh, 2021, Data Science In Healthcare https://www. analyticsvidhya. com/blog/2021/05/data-science-in-healthcare/*

*[5] WHO, Cardiovascular diseases https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1*

*[6] Bahareh Javid, 2018, Non-invasive Optical Diagnostic Techniques for Mobile Blood Glucose and Bilirubin Monitoring https://www.ncbi.nlm. nih.gov/pmc/articles /PMC6116315/*

*[7] Siva Rama Krishna Vanjari, 2017, Minimally invasive Mobile Healthcare Diagnostic platform for Glycated Hemoglobin Detection https://imprint-india.org/knowledge-portal-5699-minimally-invasive-mobile-healthcare-diagnostic-platform-for-glycated-hemoglobin- detection*

*[8] Md Hasanul Aziz, 2021, Automated Cardiac Pulse Cycle Analysis From Photoplethysmogram (PPG) Signals Generated From Fingertip Videos Captured Using a Smartphone to Measure Blood Hemoglobin Levels https://pubmed. ncbi.nlm.nih. gov/33760745/*

*[9] Fahad Kamal Alsheref, 2019, Blood Diseases Detection using Classical Machine Learning Algorithms https://pdfs.semanticscholar.org/ 94db/d6742b58220760 a3e11c 735bcde75a4a6c3b.pdf*

*[10] Darcy, Alison M., 2016, "Machine learning and the profession of medicine" https://pubmed. ncbi. nlm.nih.gov/26864406/*

*[11] Nantsupawat et al, 2019, Cholesterol levels https://www.nhs.uk/conditions/high-cholesterol/cholesterol-levels/*

*[12] AnaBarragán-Montero, 2021, Artificial intelligence and machine learning for medical imaging: A technology review https://www.sciencedirect. com/science /article/pii/S1120179721001733*

*[13] Simplilearn, 2021, Understanding Naive Bayes Classifier https://www. simplilearn.com/tutorials/machine-learning-tutorial/naive-bayes-classifier*

*[14] El-Sayed M., 2019, A Machine Learning Model for Hemoglobin Estimation and Anemia Classification https://www.academia.edu/38528570/ A_Machine_Learning_ Model_for_Hemoglobin_Esti mation_and_Anemia_Classification*