Mining Complete Blood Count Reports For Disease Discovery

Samiullah Jatoi[†] , M. Aamir Panhwar[††], M. Sulleman Memon [†††], Junaid Ahmed Baloch[††††] and Salahuddin Saddar[†††††]

[†]Software Engineer, Liaqat Medical University Hospital, Jamshoro
[††] PhD Research Fellow, Beijing University of Posts & Telecommunications, China
[†††] Department of Computer Systems Engineering Quaid e Awam UEST Nawabshah Pakistan
[††††, †††††]Department of Software Engineering , Mehran UET Jamshoro

Summary
Healthcare systems create a massive amount of data from medical tests. Data mining is the method to determine patterns in huge data sets such as medical examinations. Blood diseases are not the exception; there are many test data that can be collected from their patients. In this research, we have applied data mining technique to discover the core-relationship between Anemia and Thalassemia from Complete Blood Count (CBC)test. The relationship can be exploited to identify and predict the possibility of getting Thalassemia in the patients suffering from Anemia. We have performed experiments using blood test data set collected from Diagnostic and Research Laboratory of LUMHS in Pakistan. Naive Bayesian Network algorithm is used to analyze and evaluate the data set. The Final results show that Bayesian Network has the best capability to predict core-relate between diseases with an accuracy of 98%.

Key words:
CBC, KNN, Thalassemia, Anemia and Bayesian Network

1. Introduction

CBC Complete Blood Count is a crucial viewing blood scan which defines person's complete health state. CBC is the basic and the modest investigation test which can signify the diseases alike Thalassemia Primary (Thal-M) and Secondary (Thal-T), Iron deficiency patients, Cancer patients, Dengue, Anemia patients and the enduring who suffers from other blood diseases. A CBC blood test measures individual parameters and features of our blood, including:

- Red blood cells(RBC) carries oxygen
- White blood cells(WBC), flight with infections in body
- Hemoglobin(HB), oxygen-carries protein in RBC
- Hematocrit (HCT), the percentage of red blood cells to the fluid component, or plasma, in your blood Platelets, that helps in blood clotting. All these are also shown in Table 1. Defective rise in cell counts as exposed in a CBC may show that the patient has basic medical disorder that calls for additional decision [1].

A CBC test may go on well-ordered after a person has any amount of symbols and pointers that are connected to the conditions that may lead to any disorder in blood cells. When a person has tiredness, soreness, staining; a health professional might recommend for further CBC blood test to assist to notice the purpose and/or define its effect. In addition, less privileged people face difficulty to conduct expensive tests. For example tests and diseases related to bone -marrow conquest (a situation in which the bone-marrow does not create normal amounts of red blood cells, white blood cells, and platelets). However, CBC parameters indictor's reports can be analyzed with different data mining techniques to help predict the diseases with least cost.

CBC might remain well-ordered after a person has any amount of marks and indicators that can be linked to circumstances that affect blood cells. The year when an individual has lethargy or faintness gold has year infection, inflammation, bruising or bleeding; a health specialist may order a CBC to help detect the reason and/or determined its ruthlessness [2].

Research based on CBC report are to calculate and detect the accuracy level to Thalassemia defected patients (i.e. Major and Minor) interconnected to Anemia (i.e. Iron deficiency and Vitamin B12) will be detected. Classifying from CBC Report it might be useful for us to know or to predict disorder-ness in patients.

Main purpose of this research is to identify the accuracy level of complete blood count reports that has correlation between Thalassemia and Iron Deficiency Anemia.

Goals are enlisted below:

- Calculate the impact of other diseases with Thalassemia.
- Collected information is compared from CBC reports with other Anemia deficiency reports and HB Electrophoresis.

Hugin Expert Lite has occurred from the 1989 and is the directive strip now rising package for AI info & innovative

Manuscript received January 5, 2018
Manuscript revised January 20, 2018

decision occur supported on committed statistical models Theorem Nets.

Table 1: CBC parameters List

Parameters
RBC Count
RDW- Red Blood Cell Distribution Width
Hemoglobin
PCV –Packed Cell Volume
MCH–Mean Corpuscular Hemoglobin.
MCHC–Mean Corpuscular Hemoglobin Concentration.
MCV - Mean Corpuscular Volume
Total White Blood Cells
Lymphocytes
Neutrophils
Basophils
Monocytes
Platelet Count

2. Bayesian Network

A Bayesian network (Bayes network, a causal probabilistic network or Bayesian belief network, or simply belief network) is compacted pattern demonstration for thinking low doubt. Graphical models are a marriage stuck between the theory of probability and the theory of grapheme. They provide a natural tool for tackling the two problems that arise while applied mathematics and engineering - improbability and difficulty - and, above all, they are playing gradually vital role in the strategy and study of machine learning algorithms. The idea of modularity is major to the indication of a graphic model - a composite system is built by merging modest sections. Probability theory provides the glue to be combined through the parts to ensure that the whole system is consistent paths and provision of data interface models. The graphic theoretical side of the graphical models provides both intuitively attractive user interface, by which man can model human beings strong interaction with groups of variables, but also a data structure that suits algorithms. [3].

- P (EF) = 0.29 because 29% are normal patient
- P (NOR) = 0.71 because 71% are effected patient.
- P (EF I IDA) = 0.60 because 60% are iron deficiency anemia effected patient.
- P (EF I TH) = 0.40 because 40% are thalassemia effected patient.
- P (IDA I TH) = P (NOR). P (EF I IDA) / P (NOR). P (EF I IDA) + [P (EF). P (EF I TH)]
- P (TH I EF) = (0.29).(0.60) / [(0.71).(0.40)]
- P (TH I EF) = 0.37991

3. Related Work

The core objective is here to give an outline of most vital concepts that illustrat the main purpose of this research. As of the massive number of data in medical health fields, which are accessible today, numerous scholars hang upon data mining procedures to get new Awareness. A simple research on any data mining procedures gives a probable list of warning sign that contains both precise and imprecise therapies for diseases. Most of the scholars have done investigates on hematology section. Several journalists have practiced different approaches on the medical datasets that are related to heart diseases such as Neural Network, KNN, Bayesian Classifier and Classification based on Clustering Decision Tree [11]. The results reveal that Decision Tree is acceptable and occasionally decision tree and Bayesian classification have the same accuracy in predicting heart related diseases. However, other predicting methods such as Classification based on Clustering Decision Tree, KNN and Neural Networks do not perform well.

The authors have applied different methods like a Decision tree, Artificial Neural Networks and logistic regression [12]. The results show that the Decision Tree has the best accuracy of 93.6%. It provides the best predictor classifier as compared to Artificial Neural Network and logistic Regression model that gives 91.2%, 89.2% of accuracy related to Cancer diseases. In case of blood related diseases, there is 0.37991 probabilities that a patient may suffer from Iron Deficiency Anemia due to thalassemia. The authors in [10] have applied different techniques such as Decision Tree, Naïve Bayes, and Neural Network.

The Neural Network classifier has more significance accuracy that the results show the disease of Thalassemia, Iron deficiency patients. The authors in [4] have applied different techniques [4] such as Swarm Optimization Algorithms and Genetic Algorithm (GA) Technique for Predictive Disease. Diagnosing the acute diseases and predicting by identifying the recurrence of diseases in advance and providing health advice, Health care systems help people to lead a better life. Acute Diseases Like Cancer, stroke, lower respiratory infections, heart and chronic obstructive lung.

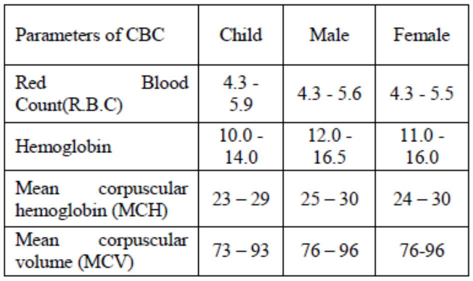
The authors [13] have applied Image processing based system that can automat

ically detect and count the number of RBCs and WBCs in the blood sample image results in Diseases from Blood Cell.

4. Diseases From CBC Report

Mostly, assumed parameters are RBC (Red Blood Cells), HB (Hemoglobin), MCH (Mean Corpuscular Hemoglobin) and MCV (Mean Corpuscular Volume) on these parametersseveral hemogolobin syndromes are seen by physicians to patients. These parameters promise two leading diseases Anemia and Thalassemia (Major or Minor). Furthermore, Anemia is partitioned into two channels either Iron Deficiency or Vitamin B12. These diseases are diagnosed when the values go beyond or below the reference ranges as shown in Table 2. Ranges are enlisted in three parts child, Male and Female.

Table 2::Enlisted of the three parts of Child, Male, and Female



Parameters of CBC	Child	Male	Female
Red Blood Count(R.B.C)	4.3 - 5.9	4.3 - 5.6	4.3 - 5.5
Hemoglobin	10.0 - 14.0	12.0 - 16.5	11.0 - 16.0
Mean corpuscular hemoglobin (MCH)	23 – 29	25 – 30	24 – 30
Mean corpuscular volume (MCV)	73 – 93	76 – 96	76-96

5. Anemia

Anemia is recorded most conversational murder place; it affects occurred in enforcement cells and haemo-protein. The accelerator is RBC that transaction oxygen from our lungs to the repose of our body. The condition of iron in our body is assignment to haemo-protein. Mostly age grouping who feature symptom human a deficiency of hamper. To investigate the symptom, weaken examines gore. This proves the sodding execution depend on (CBC). Financed by the results, they may need to do divers tests, such as testing become extinct delicate. [5].

5.1 Common Types of Anemia

5.1.1 Iron deficiency

Deficiency of Iron is the public form of anemia and its indication is greatest unrefined identify of anemia, this situation occurs after the individual sufficient press in the body. Iron deficiency is normally for the reason that of the poor/slow obsession of iron. Birth& growth is been the main reason of ID and its outcome in pregnancy related sign. Mild ID usually cause complications and create rick factor in life such as Heart problems, growth problems [6].

Anemia is unexceptionally the official with a CBC blood test. A specialist can instruct for extra pathology test to resolve the anemia and assist in deciding the treatment. These pathological blood tests deal offers combination of with:

i) *Serum Iron:* Serum Iron in our blood, RBC size and colour of RBC's cell stay light in color if they are lack.

j) *Ferritin:* Ferritin is the protein that assists by storing iron in our body. Small courses of ferritin show a small amount of iron level.

k) *Total Iron Binding Capacity (TIBC):* TIBC is a protein that carriages iron. The run is victimized that resolve the determination that carry's iron TIBC.[7]

5.1.2 Vitamin-deficiency Anemia

Vitamin-Deficiency consequences display that low phases of serum folate (folic acid) or vitamin-B12, frequently slowly eating ingestion. Anemia which is dangerous are based in which stomach passageway doesn't dissolve vitamin B12 [6].CBC also confirms or indicates that there is deficiency of Vitamin and for further confirmation there is also blood test for Serum Folate(Folic acid) and Vitamin B12.

6. Thalassemia

Thalassemia is heritable/genetic blood disorder. Persons suffer from thalassemia are not enough capable to get haemo-protein, which causes these symptom of anemia. Haemo-protein is opened in RBC and carryings oxygen to over all parts of the body. When sufficient haemo-protein is not present in RBC, oxygen cannot achievement to all parts of the body, and the body remain transforms starved for oxygen and is unable to the plastered. There are two assorted types of thalassemia disease; Beta (β) and Alpha (α) thalassemia. Alpha Thalassemia is a real ordinary disease. Indications seem the first two held of history and allowed boldness of the peel, penurious craving, petulance, and unfortunate to create. [14]. HB Electrophoresis blood test is done to insure that patient has thalassemia disorder or not.

7. Methodology

The related work planned, must be a method. These approaches and planes are the basic requirement or phases to be carry out in this research work.

7.1 Selection of Data

Data collection is one of the biggest jobs in the research. The benefit of procedures for mining production on statistical techniques is without imagination. Decision or records and databases that succeed these requirements are hard processes. Once in hand, the minor might proceed with activities to further themselves with the collection of data, determine the worth of group problems, attain the finest overview substance and notice interesting subsets to arrange assumptions for operational information. Collection was collected from Diagnostic and Research Lab. Number of patients are 400, and those where divided on the bases of age wise and by gender wise.

Male are 16 %.
Childs are 53%.
Female are 29%.

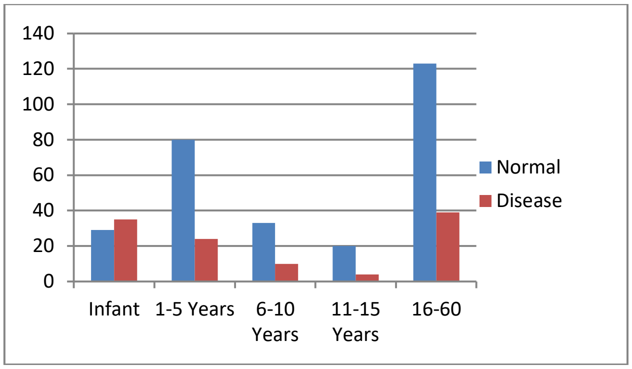


Fig. 1 Data Age wise

In Figure 1 we can see the age-wise discrimination that how many age wise patients are entered and on thisage, wise ranges are also deferred and also on gender wise.

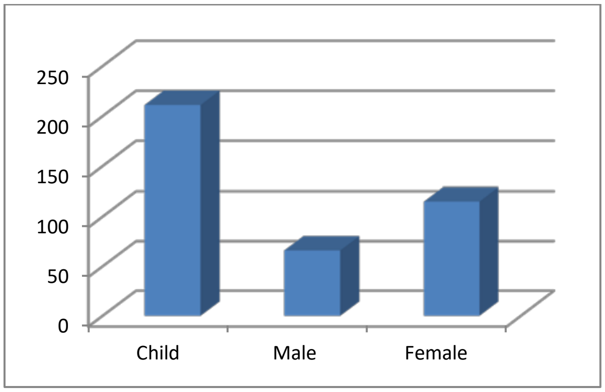


Fig. 2 Data Gender wise

Figure 1represents the disease w.r.t. age wise grouping. We can see that the most amount of disease occurs in the age group of 16-60. However, the Normal-Disease ratio in this group is least. The Normal-Disease ratio is very high in the infant group. Figure 2shows a total number of patients in a form of gender wise division. A total number of child ratios is high then that of female and male.

7.2 Data preprocessing

Data preprocessing practice is assessment of the growth of turning raw data into info service. A real gathering is unfinished, unpredictable, lack of convinced performances or tendencies and is probable to include various mistakes. Association of pretreatment is a proven method of resolution of the specified problems. Aggregation of pre-treatment prepares crude Assembly to encourage the transformation. The accumulation drives over a program of measures in pre-processing [8].

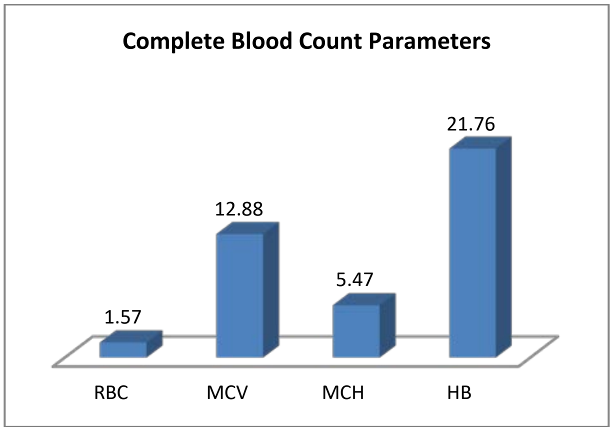


Fig. 3 CBC Report Parameters

Figure 3 depicts that these are the basic parameters of complete blood count on which research work is done and on these parameters prediction about of disease is also done by doctors.

- **Cleaning of data:** Data is cleansing data through processes for instance, missing values are filled, nosy data is smoothed and variation is resolved in data.
- **Integration of Data:** Different presentations are self-possessed and conflicts are resolved within data.
- **Reduction of Data:** This step is done to reduce the amount of data without disturbing the sanctity of data.
- **Discretization of Data:** In this step it checks the interchange values of a continuous attribute by dividing the range of concept of intervals

Fig 4 illustrates about the total number of patients affected by Thalassemia cases which are screened and shown that minor and major thalassemia and are counted by gender wise, and it shows that children have high ratio in Major.

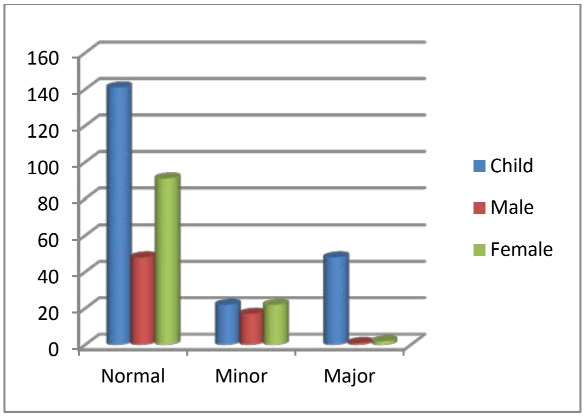


Fig. 4 Total no: of cases of Thalassemia

8. Results And Discussion

Data mining is an important element in Knowledge Discovery in Database (KDD) methods. Data mining contains algorithms and use of that algorithm to generate before unknown and theoretically valuable collection from the collection stored in the databases. They strength contains determinative which algorithm and parameters can check and equal a circumstantial data mining method with the indiscriminate standards of the KDD methods. Data mining techniques permit classification, regression, summarization, etc. Knowledge Discovery in databases is the practice of retrieving data aggregation from lower rise to higher rise knowledge [9].

Navies Bayes Network techniques of data mining is used to forecast the core-relationship between diseases. Fig shows the framework of Hugin lite software. The data collected for results are:

- A total number of Patients = 400.
- Patients Affected = 290
- Normal Patients = 110

The above figure shows the parameters of complete blood count and HB Electrophoresis with a standard deviation that are having core-relationship with each other and based on these parameters we have calculate the disease Anemia or Thalassemia. Figure 6 shows the interface of Hugin lite software to calculate the relationship between diseases. The result is 0.6% ratio of total number of affected patients those have Anemia but they have shown in CBC report that are thalassemia patients but they were not having any disease. Calculated results are shown from Hugin Lite software.

Fig 6 also shows the information that how many patients are affected from Anemia and Thalassemia from total number of patients that are counted from result. The word Anemia represents here patients which are affected from Vitamin B12 Deficiency and Iron Deficiency.

8.1 Efficiency Parameters

The framework is proposed to calculate the efficiency of parameters which are used and shows the domino effect.

Table 3: Parameters Comparison

Iron Deficiency	Low (MCV)	Low (MCH)	Low (HB)
Vitamin B12	High (MCV)	High (MCH)	Low (HB)
Thalassemia	Low (MCV)	Low (MCH)	High or Normal (R.B.C)

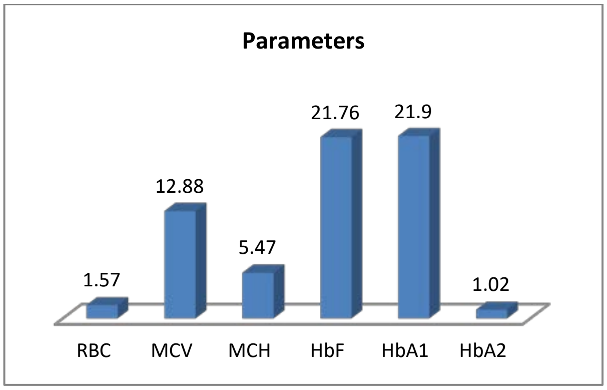


Fig. 5. Parameters

Formula:
Efficiency = No. of correct Predictions/ Total No. of Predictions
Efficiency = 0.98 which is calculated from data mining technique Navies Bayesian Network.

8.2 Dependences of Parameters

Table 03 indicates the core relationship among the complete blood count test parameters and their dependencies on parameters to predict diseases and after

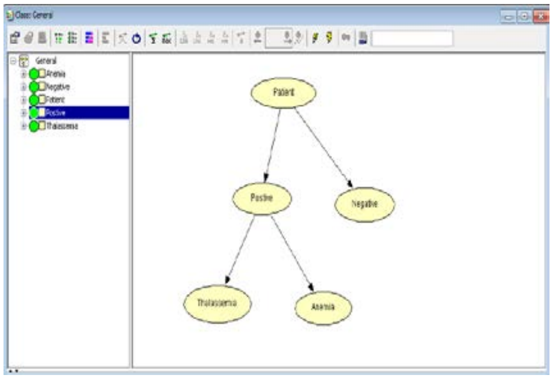


Fig. 6 Disease

that health consultant recommand for other blood tests to confirm and give final comments. Anemia's dependences on CBC factors are also shown same as with thalassemia.

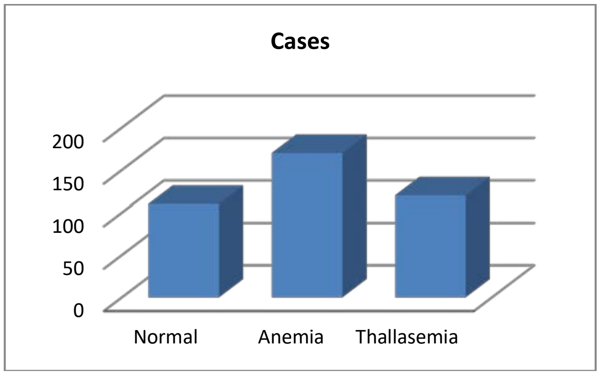


Fig.7 Results of all cases

9. Conclusion

By the modern summary growth in the number of bio-medical health data that is collected by automatic means in serious caution and the packed availability of cheap and reliable computing equipment, many investigators have started, however, are ready to start discovering data. The medical data mining produces business intelligence, which is useful for diagnosing the disease. Data mining methods are used medical health data for numerous numbers of diseases which are used to acknowledged and diagnosed for human health.

The conclusion can be started like this data mining techniques can provide an affordable solution to predict the acute diseases by analyzing its correlation with less serious diseases.

It is practiced from complete blood count reports that there are many patients how have a relationship between Hemoglobin disorders and Anima patients. Nonetheless, if patients have high or low MCH and MCV (they are directly proportional to one another) are counted as the main reasons for Anemia and Thalassemia.

10. Future Work Limitations and Challenges

Working in future, we can compare the other parameters of count formula blood and show that their core-relationship with the different corresponding and the anemia is caused, as we the know that the CBC is the first blood test primary for the discovery of the disease, with respect using data mining algorithms.

In medical is a considered as important although obscure tasks that are required to be carried out precisely and efficiently. More effective algorithms with very high accuracy are required to resolve the serious issues of human health. No matter how powerful these data mining techniques are, they should be used with great care in the biomedical applications. It would be virtually to discover the best mining algorithm of data for all medical areas.

References

[1] Khaki Jamei, Mehrzad, and Khadijeh Mirzaei Talarposhti. "Discrimination between Iron Deficiency Anemia (IDA) and β-Thalassemia Trait (β-TT) Based on Pattern-Based Input Selection Artificial Neural Network (PBIS-ANN)." Journal of Advances in Computer Research (2016).

[2] Eyad H. Elshami, Alaa M. Alhalees.; The International Conference on Informatics and Applications (ICIA2012) (ICIA2012) Malaysia: Automated Diagnosis of Thalassemia Based on Data Mining Classifiers; Jun – 2012 Page Numbers: 440-445.

[3] Rana Sabeeh Abbood Alsudani, Jicheng Liu; "The Use of Some of the Information Criterion in Determining the Best Model for Forecasting of Thalassemia Cases Depending on Iraqi Patient Data Using ARIMA Model": Journal of Applied Mathematics and Physics, 2017, 5, 667-679. https://www.techopedia.com/definition/14650/data-preprocessing; 14-JUL-2017

[4] M. Abdullah and S. Al-Asmari "Anemia types prediction based on data mining classification algorithms," Communication, Management and Information Technology – Sampaio de Alencar (Ed.) 2017.

[5] Sheenal Patel and Hardik Patel;" survey of data mining techniques used in healthcare domain" International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016.

[6] Shakil, Kashish Ara, Shadma Anis, and Mansaf Alam. "Dengue disease prediction using weka data mining tool." arXiv preprint arXiv: 1502.05167 (2015).

[7] Vanaja, S. and K. Ramesh kumar.; Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey: Journal of Computer Science, 2014.

[8] Vijaya shree, J., and N. Ch Sriman Narayana Iyengar. "Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review." International

Journal of Bio-Science and Bio-Technology 8.4 (2016): 139-148.

[9] Prof. Hina Malik, Roopali Randiwe, Jyotsna Patankar, Priya Bhure “Disease Diagnosis Using RBCs & WBCs Cell Structure by Image Processing” ;National Conference on Advances in Engineering and Applied Science (NCAEAS) :16th February 2017.

[10] Alaa M. El-Halees1, Asem H. Shurrab2; “Blood Tumor Prediction Using Data Mining Techniques”: Health Informatics - An International Journal (HIIJ) Vol.6, No.2, May 2017 DOI: 10.5121/hiij.2017.6202 23.

[11]National Institutes of Health Clinical Center :http://www.cc.nih.gov/comments.shtml 7/14 (DOA 10/05/2015).

[12] M. Abdullah and S. Al-Asmari “Anemia types prediction based on data mining classification algorithms,” Communication, Management and Information Technology – Sampaio de Alencar (Ed.) 2017.

[13] Khaki Jamei, Mehrzad, and Khadijeh Mirzaei Talarposhti. "Discrimination between Iron Deficiency Anemia (IDA) and β-Thalassemia Trait (β-TT) Based on Pattern-Based Input Selection Artificial Neural Network (PBIS-ANN)." Journal of Advances in Computer Research (2016).

[14] Miller, Jeffery L. "Iron deficiency anemia: a common and curable disease." Cold Spring Harbor perspectives in medicine 3.7 (2013): a011866.

[15] Elshami, Eyad H., and Alaa M. Alhalees. "Automated Diagnosis of Thalassemia Based on Data Mining Classifiers." The International Conference on Informatics and Applications (ICIA2012). The Society of Digital Information and Wireless Communication, 2012.

[16] S. Vijayarani and S. Sudha . "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples". Indian Journal of Science and Technology, Vol 8(17), August 2015.

[17] http://www.hematology.org/Patients/Anemia/; 16-JUL-17

Samiullah Jatoi received his B.E in Software Engineering in 2013 and currently doing his M.E in also Software Engineering from Mehran University of Engineering & Technology, Jamshoro. He is working as software engineer in the Laboratory in the Liquat University of Medical and Health Sciences, Jamshoro from last 3 years. His field of study is Net Work and Communication, Data Mining and programming languages. He also has CCN certification. He is member of Pakistan Engineering Council and IEEE.

Muhammad Aamir Panhwar received his B.E Biomedical Engineering degree(2006) from Mehran University of Engineering & Technology, Jamshoro, Pakistan and M.E Telemedicine & e-health system (2014) from Mehran University of Engineering & Technology, Jamshoro, Pakistan. He worked as a lecturer in Mehran University of Engineering & Technology, Jamshoro, Pakistan since for 8 years. Currently he is working toward his PhD from Beijing University of Posts and telecommunications, China. He is doing his research working on the from State Key Laboratory Intelligent communication, navigation and micro-Nano system, Beijing University of Posts and telecommunications, China. His research interests are heterogeneous, 5G networks, Wireless sensor networks, Femtocells, Microcells and e-health system.

M. Sulleman Memon received the B.E in Computer Engineering and M.E. in Software Engineering from Mehran University of Engineering & Technology, Jamshoroin 1990 and 2004, respectively. He is now a PhD scholar at Quaid e Awam University of Engineering and Technology, Nawabshah. He has submitted thesis. He is working as Assistant Professor in the Department of Computer Engineering. QUEST, Nawabshah. He is author of many International and national papers. He has presented his work at many countries of the word in International Conferences. His field of study is Wireless communications. He is Senior Member of IACSIT and member of Pakistan Engineering Council, ACM, and IEEE.

Junaid Ahmed Baloch received his B.E in Software Engineering and M.E in Software Engineering from Mehran University of Engineering & Technology, Jamshoro in 2012 and 2017 respectively. He is working as Lecturer in the Department of Software Engineer at Mehran University of Engineering & Technology, Jamshoro from last 4 years. His field of study is Software Requirment Engineering, Net Work and Communication and Data Mining. He is member of Pakistan Engineering Council and IEEE.

Salahuddin Saddar received his B.E in Computer Engineering and M.E in Software Engineering from Mehran University of Engineering & Technology, Jamshoro in 1994 and 2004 respectively. He is working as Assistant Professor in the department of Software Engineering from Mehran University of Engineering & Technology, Jamshoro from last 14 years. He is also author of some International and national papers. His field of study is Project Management, Net Work and Communication, Human Computer Interaction and Software Testing & Quality Assurance. He is member of Pakistan Engineering Council, ACM, and IEEE.