Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study

Basharat Naqvi[1], Arshad Ali[2], Muhammad Adnan Hashmi[3] and Muhammad Atif[4],

[1]Education Department, The Government of Punjab, Pakistan
[2,3,4]The Department of Computer Science & Information Technology, The University of Lahore, Lahore, 55150, Pakistan

Summary
The objective of supervised learning is to construct a distribution model by considering class values for the purpose of prediction. Nowadays data mining is playing an important role for prediction of diseases in the healthcare industry. Data mining process is used for feature selection, information extraction as well as to discover the unknown pattern and relations from the unstructured data. These patterns can be used to write a wise prescription for patients.

This paper introduces an expert system for the early prediction of a diabetic patient by using data mining classification techniques. This work is based on a dataset which comprises of 130-US Hospitals for the years 1999_2008 consisting of 50 attributes and more than 100,000 instances. The evaluation and comparison is performed by using RapidMiner, a software platform of data science, which supports various machine learning steps including results visualization. Random Forest, Decision Stump, Random Tree and ID3 were applied to mine the useful information from data. The extracted information will assist the practitioner to write the precise and wise prescription for diabetic patients.

This research work presents description of chosen classification models and dataset. Next, makes evaluation and performs a comparison of performance of 5 classification techniques on chosen dataset. Then, it provides results by considering evlaution metrics such as accuracy, precision and recall. This work finds that the decision tree is the best technique for prediction of disease in diabetic patients.

Key words:
Decision tree, ID3, diabetic patients, data mining

1. Introduction

It is an established fact that computer based information systems ate more effective as compared to traditional methods in disease diagnoses. In fact, computer-based methods not only facilitate the practitioners but also help in improving the quality in healthcare domain.

Data mining is the procedure of extracting useful information from a dataset [1]. It refers to extracting or mining knowledge from large amounts of data [2]. Many disciplines including biology, engineering, e-commerce, marketing, physics, communication networks, business, and health and so on can benefit from data mining solutions which can be applied to these areas effectively [3].

Machine learning (ML), a major discipline of data mining, methods have attracted the many researchers for diagnoses of diseases [4]. Previously, machine learning has already been applied in self-driving cars, detection of speech, well organized web search, and improving human perception. Presently, ML is found everywhere, and possibly being used by one without knowing it. ML techniques possess the ability to handle data of every scale, i.e., small, medium or large, and make a pool of data from dissimilar resources (combine) and to incorporate the contextual information in the learning [5].

It would be very difficult task to extract the hidden pattern and prognosis the disease effectively without adopting data mining approach and using machine learning methods. There exist many methods to extract patterns from data, which include Bayes’s theorem, regression analysis, decision trees, random forest, neural networks, cluster analysis, genetic algorithms, decision trees, random tree, decision stump, ID3, decision rules and support machine vectors.

In data mining approach, preprocessing is considered an important step which allows us to make data compatible with the algorithm by removing the missing and redundant values [3]. There is widespread use of data mining techniques in healthcare industry. In this paper, we have used the decision tree, ID3, Random Forest, decision stump and random tree algorithms to mine the useful information from the data. By using these algorithms, the diabetic patients have been classified into two categories (i) patients suffering from diabetics, and (ii) healthy patients indicated by 1 and 0 respectively. To make the data compatible with algorithms, we transformed data as detailed in the methodology section.

Focus: This work considers datasets related to diabetic patients and applies various techniques by considering evaluation metrics such as accuracy, recall, precision in order to carry out performance evaluation and comparison of chosen classifiers. The following research questions are focused in this work

- Weather different classification approaches are able to successfully predict outcome in diabetic patients?

Manuscript received August 5, 2018
Manuscript revised August 20, 2018

- Which prediction approach is better choice for outcome classification in diabetic patients?

Organization of the Paper: The remaining paper is organized as follows. Section 2 describes selected dataset related to diabetic, describes selected classification techniques. Section 3 presents an overview of existing related works. Section 4 describes the adopted methodology in detail. Evaluation metrics and results thereof are discussed in Section 5. Finally, Section 6 provides conclusion of the work.

2. Diabetic Dataset and Models

This section, presents description about three datasets which are selected for the purpose of evaluation and comparison of classification models. Then, a brief overview of various classification techniques used to analyze the selected datasets is provided.

Evaluation of classifiers is done on chosen dataset by using RapidMinor.

2.1 Dataset

This work considers dataset gathered from 130-US Hospitals concerning the period 1999-2008 and originally consists of 50 attributes and +100,000 instances. The class attribute is weather a patient has diabetic or not. Table 1 and Table 2 show the distribution of instances in terms of age and gender.

Table 1: Age-wise distribution

Age group	# of instances
0-10	161
11-20	691
21-30	1657
31-40	3775
41-50	9685
51-60	17256
61-70	22483
71-80	26068
81-90	17197
91-100	2793

Table 2: Gender-wise distribution

Gender	# of instances
Male	47055
Female	54708
Unknown	3

Table 3: Abbreviations

Description	Abbreviation
Iterative Dichotomiser 3	ID3
Rando forest	RF
Decision Tree	DT
Random Tree	RT
Decision Stump	DS

For this dataset, various classification techniques are applied (refer to Table 3), the details of which are provided in sub-section 1.2.

2.2 Classification Techniques

This sub-section describes classifications techniques in family of decision tree learning which were applied on dataset under consideration.

Decision Stump: This model actually consists of a one-level decision tree [6]. In this tree, one internal node is immediately connected to its leave nodes. It makes decision on the basis of only a single attribute given as input. There can be many possible variations which depend on the type of input attribute/feature. Authors of [7] coined the term “stump” in 1992.

ID3: Iterative Dichotomier 3 algorithm was invented by Quinlan [8]. It is used for the purpose of generation of a decision tree from dataset. It is mainly used in machine learning and natural language processing context. It performs the following steps: (1) iterates through every unused attribute and calculates the entropy of that attribute by using the original set S, (2) selects the attribute having minimum entropy and splits the set S into subsets by that attribute, (3) forms a decision tree node which contains that attribute, and (4) continues to recurse on each subset using leftover attributes

Random Tree: It is a tree which is formed by a stochastic process. Random trees normally refer to randomly built trees which have nothing to do with machine learning. However the popular machine learning framework Weka uses the term to refer to a decision tree built on a random subset of columns.

Random Forest: It is a method of combing multiple random trees into one huge classifier. It operates by building a multitude of decision trees at the time of training and providing a class as output that is mode of the classes. Brieman [9] introduced the random forest algorithm.

3. Related Works w.r.t. Classification

The methods widely used for classification are statistical, discriminant analysis, decision tree, Markov based, swarm intelligence, k-nearest neighbor, genetic classifiers, artificial neural network, support vector and association rule.

Breiman [10] used the random forest algorithm for the generation of multiple decisions. This approach is based on random selection of features for multiple decision generations.

Denzinger et al. [11] introduced the methodology for the detection of diabetics in initial stage by using multi-agent concept. In this work, the authors introduced the concept of mixture knowledge to extract the different aspects in diabetic patients.

The authors in [12] used the different classification mechanisms such as C4.5, LDA, and KNN in order to extract the useful pattern from the data. It was found that C4.5 performed better in term of accuracy and error rate as compared to other algorithms.

The authors in [13] applied artificial neural network (ANN) and Extended Classifier System (XCS) to extract the hidden layers from data with the help of perceptron to diagnose the diabetic patients.

Adidela et al. [14] used fuzzy ID3 approach to extract the useful pattern from data of diabetic patients. They divided the data into clusters by applying the Expectation – Maximization (EM) algorithm and classified each cluster into a tree like structure.

Patil BM and Joshi RC [15] used the Apriori algorithm to classify data into two categories: Patients having the diabetic disease were label as “yes”, healthy patients were labeled as “No”, the author has extracted association rules to classify the diabetic patients after applying some useful preprocessor steps.

Author in [16] used the J48 algorithm to diagnose the diabetic patients; Aljarullah AA has obtained the 78 % accuracy by using the J48 algorithm on diabetic patient dataset.

Jaya Rama [17] introduced a framework named as “duo mining” that is used for diagnosis the diabetic patients, the proposed framework based on many classification algorithms such as KNN, Naïve Bayes and SVM etc.

Mandal S and Dubey [18] used the hierarchal clustering to depict the useful trends in diabetic patients.

Kavitha K and Sarojamma [19] used the CART method for prognosis and monitoring the diabetic patients, they have divided the patients into two categories: low-risk patients and high-risk patients based on algorithm nature.

Many research works applied decision tree approach in order to analyze clinical data [20-22].

4. Methodology

In this paper, we used the Knowledge Discovery Process (KDD) model which consists of five steps as shown in Figure 1.

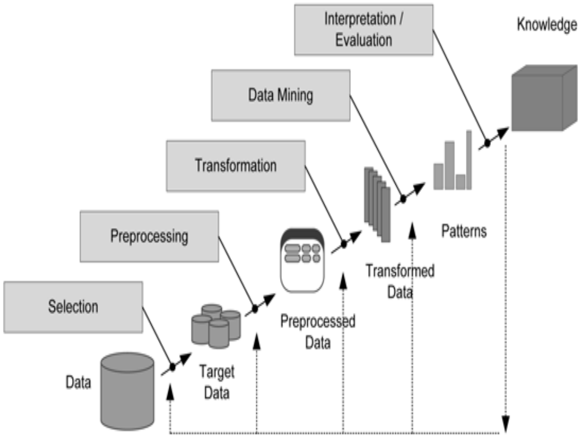


Fig. 1 KDD Process Model

4.1 Data Selection

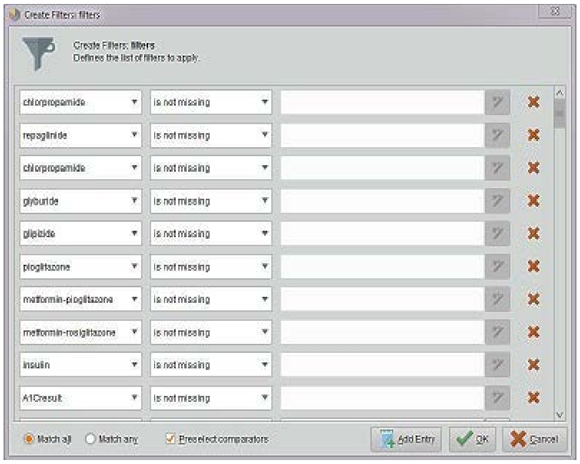
In this paper, we used the dataset gathered from US hospitals that consists of 50 attributes and 10 million instances [23]. These 50 attributes consist of different test results suggested by the practitioner, medication, patient encounter in the hospital as well as patient pre and post history during the stay in hospitals.

4.2 Preprocessing

It is an important step in data mining, in which we removed and discarded the missing or redundant values from the data. For this process, we applied the filter as shown in Figure 2.

4.3Transformation

In this step, all the transformations used in this paper to make the data compatible with certain algorithms are elaborated. Without applying this transformation step, we can’t extract useful information from the data. All the transformations are meaningful and sustain the information for which data is being transformed. These Transformations are shown in Table 4.



chlorpropamide	is not missing
repaglinide	is not missing
chlorpropamide	is not missing
glyburide	is not missing
glipizide	is not missing
pioglitazone	is not missing
metformin-pioglitazone	is not missing
metformin-rosiglitazone	is not missing
insulin	is not missing
A1Cresult	is not missing

Fig. 2 Preprocessing Step

Table 4: Data Transformation

Gender
Male=1,
Female=0
	Pioglitazone
	No=0
	Steady=1
		Tolazamide
		No=0, Steady=1

Examide
No=0,
Steady=1
	Glipizide
	No=0,
	Steady=1
		Tolbutamide
		No=0,
		Yes=1

Repaglinide
No=0,
Steady=1,
Up=2
	A1Cresult
	None=0,
	Norm=1
	>7=2, >8=3
		Metformin
		No=0,
		steady=1,
		Up=2,Down=3

Troglitazone
No=0,
Steady=1
	Acarbose
	No=0,
	Steady=1
		Rosiglitazone
		No=0, Steady=1

Gilmepiride
No=0,
Steady=1
	Miglitol
	No=0,
	Steady=1
		Citoglipton
		No=0, Steady=1

Nateglinide
No=0,
Steady=1
	Diabeties
	No=0, Yes=1
		Insulin
		no=0,
		steady=1,
		up=2, down=3

Gly_metformin No=0, Steady=1	Chlorpropamide No=0, Steady=1	
Age (0-10)=1, (10-20)=2, (20-30)=3, (30-40)=4, (40-50)=5, (50-60)=6, (60-70)=7, (70-80)=8, (80-90) = 9		

4.4 Data Mining

We applied Decision Tree, ID3, Random Forest, Random Tree and Decision Stump to mine the knowledge from the data.

4.5 Interpretation

Interpretation and evaluation of results obtained by algorithms as well as graphical description of results is detailed in result and discussion section. On the basis of results, we derived the rules as a prognosis of diabetic patients.

5. Evaluation Metrics and Results

A binary classifier provides outcome with two labels, for example, in terms of Yes/No and 1/0 against provided data as input. In order to evaluate performance after classification, observed class values of test dataset are compared with those predicted by classifier. Normally, one class is shown as positive (P) and the other one as negative (N). The dataset selected for this work consists of two observed labels i.e. “yes” or “no”. “Yes” indicates presence of diabetic while “no” represents absence of diabetic.

The confusion matrix is an important metric which helps figure out the accuracy and correctness of the model. For binary observed labels, the table of confusion matrix has two dimensions, namely observed and predicted. The columns show the observed/actual classification while rows provide predicted ones. It is worth mentioning that confusion matrix is not a performance measure in itself, however, it offers foundation for all of the performance measures. Four major items linked with confusion matrix are True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). True positives are data points classified as positive by the model that actually are positive (correct classification), TN are data points classified as negative by the model that actually are negative (correct classification), FP are data points classified as positive that actually are negative, and FN are data points the model identifies as negative that actually are positive (i.e. incorrect).

For a model to be 100% accurate, it must provide 0 FPs and 0 FNs, but this kind of scenario does not exist in real life. Every model being used for prediction of true class of the target attribute has some errors associated with it. The authors of [24] discussed various performance evaluation metrics.

5.1 Evaluation Metrics

Accuracy

The accuracy is obtained by dividing the number of accurate predictions over the number of total predictions. In other words, it is basically the number (or %) of correctly classified instances (CCI). In the selected dataset, target class is almost balanced; therefore, accuracy is a good measure to be used.

$$CCI = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \qquad (1)$$

Recall or sensitivity

Recall (REC) refers to the number of TPs divided by the number of TPs plus the number of FNs. It is basically the TP rate and also termed as sensitivity. Recall (REC) basically indicates proportion of instances identified as “yes” by the models which were “yes” actually. It is preferred to have recall as close to 100% as possible for the purpose of having more focus on minimizing FNs.

$$REC = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2}$$

Precision

Precision (PREC) refers to the number of TPs divided by the number of TPs plus the number of FPs. It is better to make precision as close to 100% as possible in order to be more focused towards minimizing FPs.

$$PREC = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{3}$$

5.2 Results and Discussion

Initially, the selected dataset consists of 50 attributes; then we reduced the attributes by applying the correlation matrix to get weight against each attribute and with the experts’ opinion. These selected attributes have the maximum value of information gain (IG) as compared to other attributes. Values 1 and 0 at leaf nodes in graph shown in Figure 3 represent the patient suffering from diabetic and healthy patients respectively. All the nodes in the graph expect leaf node represent the different test values recommended by domain experts. We selected the decision tree based on information gain instead of gain ratio. With the help of rule induction operator, we extracted the following rules:

- If insulin > 0.5 then patient was categorized as diabetic patient.
- If no_of_diagnosis > 0.5 and gender =1 and age > 6.5 then patient was suffering from diabetics.
- If metaformin <= 0.5 and glipizide <=0.5 then patient was categorized as healthy.
- If gilimepiride=0.5 and rosiglitazone <=0.5, patient was categorized as healthy.
- If insulin <=0.5 and metaformin >=0.5, then patient was categorized as diabetic patient
- .If value of gilipizide <=0.5 and glyburide >= 0.5, then patient was suffering from diabetics.

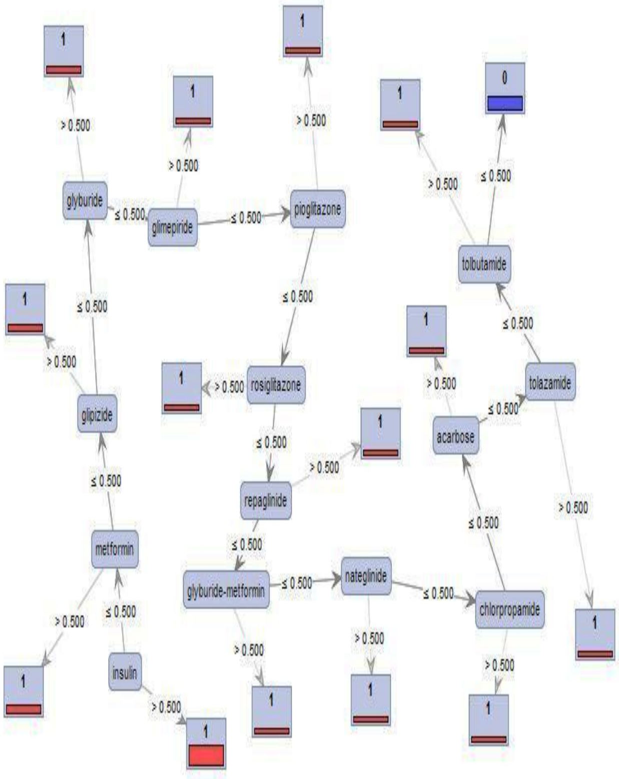


Fig. 3 Decision Tree (ID3) Results

- If value of gilimepiride <=0.5 and pioglitazone>=0.5 then patient was categorized as diabetic patient.
- If value of pioglitazone ≤ 0.500 and rosiglitazone > 0.500 and insulin <=0.5 then patient was categorized as diabetic patient.
- If rosiglitazone ≤ 0.500 and repaglinide > 0.500 and insulin<=0.5, then patient was suffering from diabetics.
- If nateglinide ≤ 0.500 and chlorpropamide > 0.500 and metformin ≤ 0.500 then patient was classified as diabetic patient.

5.3 Comparative Analysis

We applied Decision Tree, ID3, Random Forest, Random Tree and Decision Stump to mine the knowledge from the data. Comparative analysis of these algorithms is provided in Table 5 and Figure 4. According to comparative analysis, it was observed that ID3 attains the highest accuracy rate as compared to other algorithms as evident from Table 5. Moreover, ID3 performs better in terms of precision while decision stump has highest recall.

Table 5: Comparative Analysis

Algorithm	Accuracy	Precision	Recall
Random Forest	78.63	78.63	96.23
Random Tree	76.58	76.58	98.41
ID3	84.94	84.64	89.41
Decision Stump	74.76	74.76	100

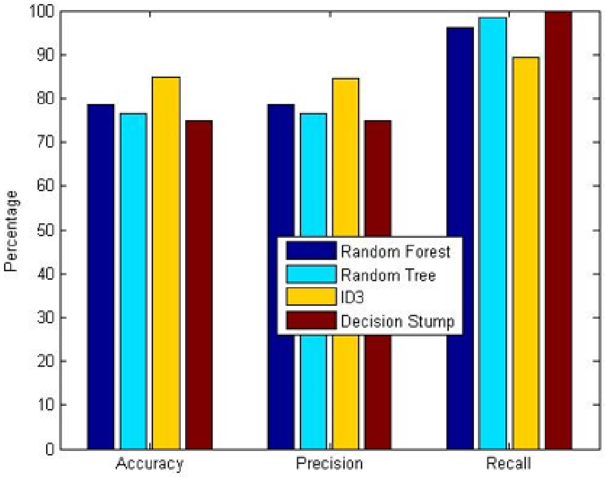


Fig. 4 Comparison of data mining techniques

6. Conclusions

In this paper, we used the data mining approach to prognosis the diabetics' patients by using different algorithms, comparative analysis of these algorithms are carried out. On the basis of accuracy, precision and recall, we identified Decision Tree approach along with rule induction operator as best to extract the useful information from the data. We extracted the rules which were cross verified from the domain expert. In this work, we also elaborated that how different values of tests affected the patients and attributes interdependency by using the rule induction operator. This work will help the practitioner as prognosis of diabetic patients as well as to write a wise decision for patients.

In future, we can automate the system that will assist the practitioner in a critical situation, when multiple factors involve in the formation of the patient state.

Acknowledgments

This work is mainly based on datasets taken from UCI machine learning repository [25].

References

[1] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[2] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[3] Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. Journal of Computer science, 2(2), 194-200.

[4] Elshazly, H. I., Elkorany, A. M., Hassanien, A. E., & Azar, A. T. (2013, November). Ensemble classifiers for biomedical data: performance evaluation. In Computer Engineering & Systems (ICCES), 2013 8th International Conference on (pp. 184-189). IEEE.

[5] Rambhajani, M., Deepanker, W. and Pathak, N. (2015) A Survey on Implementation of Machine Learning Techniques for Dermatology Diseases Classification. International Journal of Advances in Engineering & Technology , 8, 194-195.

[6] Iba, W., & Langley, P. (1992). Induction of one-level decision trees. In Machine Learning Proceedings 1992 (pp. 233-240).

[7] Oliver, J. J., & Hand, D. (1994, April). David,"Averaging over decision stumps, in machine learning" ECML-94. In European Conference on Machine Learning, Catania, Italy (pp. 231-241).

[8] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[9] Galván, I. M., Valls, J. M., García, M., & Isasi, P. (2011). A lazy learning approach for building classification models. International journal of intelligent systems, 26(8), 773-786.

[10] Breiman. L, "Random Forests", Machine Learning, Vol. 45 Issue 1, pp. 5-32, Springer, 2001.

[11] Gao, J., Denzinger, J., & James, R. C. (2005, November). CoLe: A cooperative data mining approach and its application to early diabetes detection. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE.

[12] Rajesh, K., & Sangeetha, V. (2012). Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT), 2(3).

[13] Afrand P, Yazdani NM, Moetamedzadeh H, Naderi F, Panahi MS. Design and implementation of an expert clinical system for diabetes diagnosis. Global Journal of Science, Engineering and Technology; 2012. p. 23–31. ISSN:2322-2441.

[14] Adidela DR, Lavanya DG, Jaya SG, Allam AR. Application of fuzzy ID3 to predict diabetes. Int J Adv Comput Math Sci. 2012; 3(4):541–5.

[15] Patil BM, Joshi RC, Toshniwal D. Association rule for classification of type-2 diabetic patients. 2nd International Conference of IEEE on Machine Learning and Computing; 2010. p. 67. DOI 10.1109/ICMLC.

[16] Aljarullah AA. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovative in Information Technology; 2011. p. 303–7.

[17] Jaya Rama Krishnaiah VV, Chandra Shekar DV, Satyab Prasad R, Rao KRH. An empirical study about type-2 diabetes suing duo mining approach. International Journal of Computational Engineering Research. 2012; 2(6):33–42.

[18] Mandal S, Dubey V. Implementation and evaluation of diabetes management system using clustering technique. Special Issue of International Journal of Computer Science and Informatics. 2(2):33–6.

[19] Kavitha K, Sarojamma RM. Monitoring of diabetes with data mining via CART Method. International Journal of

Emerging Technology and Advanced Engineering. 2012; 2(11):157–62.

[20] N. Sharma and H. Om, “Data mining models for predicting oral cancer survivability,” Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 2, no. 4, pp. 285–295, 2013.

[21] K.-J. Wang, B. Makond, and K.-M. Wang, “An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data.,” BMC Med. Inform. Decis. Mak., vol. 13, p. 124, 2013.

[22] H. M. Zolbanin, D. Delen, and A. Hassan Zadeh, “Predicting overall survivability in comorbidity of cancers: A data mining approach,” Decis. Support Syst., vol. 74, pp. 150–161,2015.

[23] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014.

[24] Altaher, A. (2017). Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting. International Journal of Advanced and Applied Sciences, 4(8), 43-49.

[25] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science

Basharat Naqvi received BSCS degree from Punjab University College of Information Technology (PUCIT), The Punjab University, Lahore, Pakistan in 2011. He obtained his MS (CS) degree from CS&IT Department of The University of Lahore (UoL), Lahore, Pakistan. He has been working as Secondary School Teacher at Schools Education Department, Govt. of Punjab, Pakistan since 2011. His research interests include social network analysis, data mining and machine learning.

Arshad Ali received the PhD degree in Computer Science and telecommunication jointly from the Institute of Telecom SudParis and UPMC (Paris VI) in 2013. He worked as a post-doctoral researcher at Orange Labs, Paris for 1 year. Currently, he is working as an assistant professor in the Computer Science & Information Technology Department at the University of Lahore, Pakistan. His research interests are in the areas of delay/disruption tolerant networks, wireless mobile ad hoc networks, network coding, Software metrics and supervised learning.

Muhammad Adnan Hasmi is an Assistant Professor in the Department of Computer Science & Information Technology, The University of Lahore. He received his Masters degree from University Rene Descartes, France and PhD degree from University Pierre and Marie Curie, France in 2012. His current research focuses on the development of agent oriented programming languages, proposing the mechanisms for the coordination of agents, development of chat bots and personal assistant agents.

Muhammad Atif got his PhD from Eindhoven University of Technology, Netherlands in 2011. His research interests include formal analysis of distributed algorithms and machine learning. Currently, he is associate professor at the University of Lahore, Computer Science and Information technology Department. Moreover, he is serving in Office of Research, Innovation and Commercialization ORIC as assistant director and supervising research and development projects.