International Journal of Research in Engineering, Science and Management
Volume-1, Issue-9, September-2018
www.ijresm.com
ISSN (Online): 2581-5782

An Improved Big Data Analysis of Diabetic Condition Based on Hemoglobin Protein

Mohammed Tanzim Shoaib[1], Manisha Joshi[2]

[1]*M.Tech. Student, Department of Medical Electronics, BMS College of Engineering, Bangalore, India*
[2]*Asst. Professor, Department of Medical Electronics, BMS College of Engineering, Bangalore, India*

***Abstract*: Machine learning has undergone significant development over the past decade and is being used successfully in many intelligent applications covering a wide array of data related problems. One of the most intriguing questions is whether machine learning can be successfully applied to the field of medical diagnostics. Moreover, there is a question as to what kind of data are needed. Several examples of successful applications of machine learning methods in specialized medical fields exist. Recently, a model capable of classifying skin cancers based on images of the skin was presented that achieves a level of competence comparable to that of a dermatologist7. There are however, no successful applications of machine learning that tackle broader and more complex fields in medical diagnosis, such as HbA1c level.**

Keywords: Big data, Hemoglobin Protein.

I. Introduction

It is increasingly recognized that the management of hyperglycaemia in the hospitalized patient has a significant bearing on outcome, in terms of both morbidity and mortality. This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions. However, the same cannot be said for most non-ICU inpatient admissions. Rather, anecdotal evidence suggests that inpatient management is arbitrary and often leads to either no treatment at all or wide fluctuations in glucose when traditional management strategies are employed. Although data are few, recent controlled trials have demonstrated that protocol driven inpatient strategies can be both effective and safe. As such, implementation of protocols in the hospital setting is now recommended. However, there are few national assessments of diabetes care in the hospitalized patient which could serve as a baseline for change. The present analysis of a large clinical database was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety. In particular, we examined the use of HbA1c as a marker of attention to diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus.

II. Aim and Scope

A. Methodology

This study used the health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States.

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.	0%

Group name	icd9 codes	Number of encounters	% of encounter	Description
Circulatory	390–459, 785	21,411	30.6%	Diseases of the circulatory system
Respiratory	460–519, 786	9,490	13.6%	Diseases of the respiratory system
Digestive	520–579, 787	6,485	9.3%	Diseases of the digestive system
Diabetes	250.xx	5,747	8.2%	Diabetes mellitus
Injury	800–999	4,697	6.7%	Injury and poisoning
Musculoskeletal	710–739	4,076	5.8%	Diseases of the musculoskeletal system and connective tissue
Genitourinary	580–629, 788	3,435	4.9%	Diseases of the genitourinary system
Neoplasms	140–239	2,536	3.6%	Neoplasms
Other (17.3%)	780, 781, 784, 790–799	2,136	3.1%	Other symptoms, signs, and ill-defined conditions
	240–279, without 250	1,851	2.6%	Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
	680–709, 782	1,846	2.6%	Diseases of the skin and subcutaneous tissue
	001–139	1,683	2.4%	Infectious and parasitic diseases
	290–319	1,544	2.2%	Mental disorders
	E–V	918	1.3%	External causes of injury and supplemental classification
	280–289	652	0.9%	Diseases of the blood and blood-forming organs
	320–359	634	0.9%	Diseases of the nervous system
	630–679	586	0.8%	Complications of pregnancy, childbirth, and the puerperium
	360–389	216	0.3%	Diseases of the sense organs
	740–759	41	0.1%	Congenital anomalies

Health Facts is a voluntary program offered to organizations which use the Cerner Electronic Health Record System .The database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-9, September-2018
www.ijresm.com ISSN (Online): 2581-5782

mortality, and hospital characteristics. All data were identified in compliance with the Health Insurance Portability and Accountability Act of 1996 before being provided to the investigators. Continuity of patient encounters within the same health system (HER system) is preserved.

B. Algorithm process

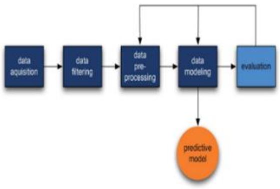

Fig. 1. Algorithmic process

- Data acquisition
- Data filtering
- Data pre-processing
- Data modelling
- Evaluation

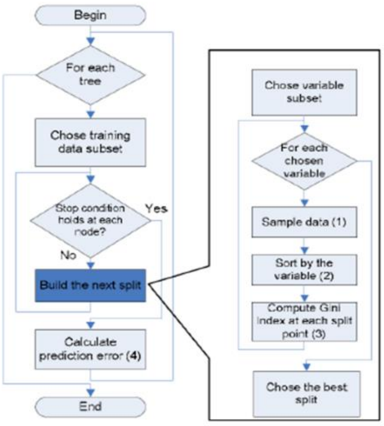

Fig. 2. Algorithm flowchart

III. Results and Discussion

In this study, we showed that a machine learning approach, using a random forest algorithm trained on large amounts of multianalyte sets of HbA1c level laboratory blood test results, is able to interpret the results and predict diseases with an accuracy on par with experienced diabetic specialists, while outperforming internal medicine specialists by a margin of more than two.

A. Random forest

Display: % of predicted classes

	Predicted			
Actual	None	>8	>7	Norm
None	92 %	67 %	74 %	79 %
>8	3 %	21 %	10 %	6 %
>7	2 %	6 %	8 %	5 %
Norm	3 %	5 %	8 %	10 %
	100 %	100 %	100 %	100 %

Fig. 3. Confusion matrix

HbA1C levels consists of none, norm, >7, >8
Random forests: None

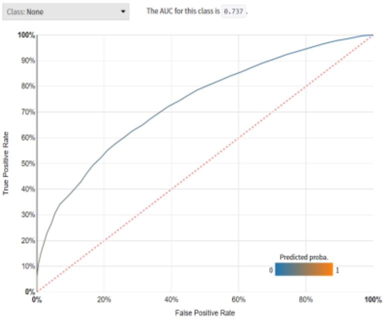

Fig. 4. Random forests: None

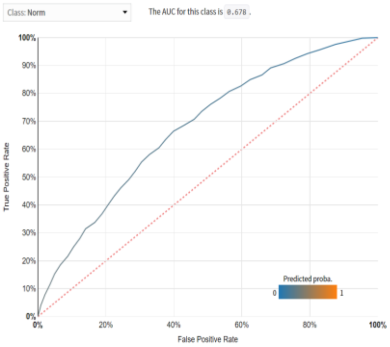

Fig. 5. Normal

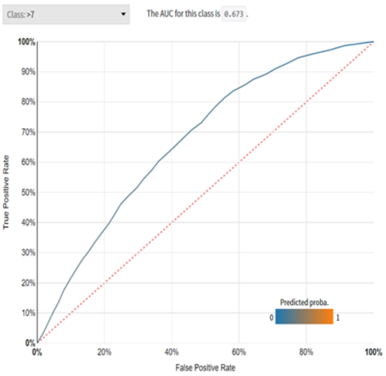

Fig. 6. >7

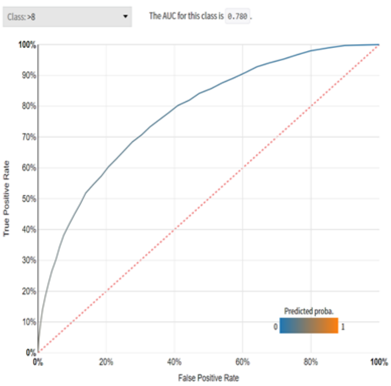

Fig. 7. >8

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-9, September-2018
www.ijresm.com ISSN (Online): 2581-5782

B. Logistic regression

Display: % of predicted classes

	Predicted			
Actual	None	>8	>7	Norm
None	84 %	37 %	-	50 %
>8	7 %	52 %	-	17 %
>7	4 %	5 %	-	17 %
Norm	5 %	6 %	-	17 %
	100 %	100 %	100 %	100 %

Fig. 8. Confusion matrix

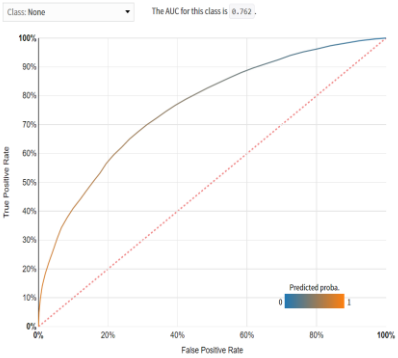

Fig. 9. Random forests: None

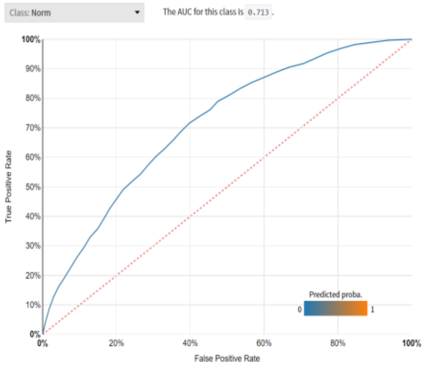

Fig. 10. Normal

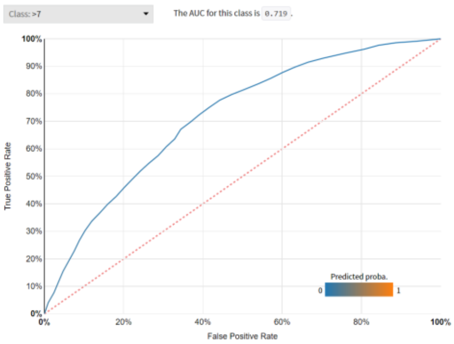

Fig. 11. >7

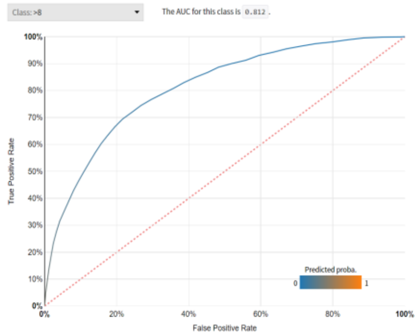

Fig. 12. >8

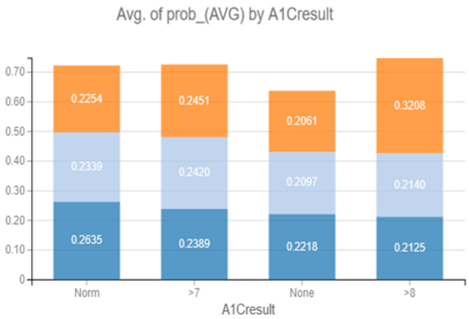

Fig. 12. Predictive analysis of A1C vs. HbA1C levels

IV. CONCLUSION

Machine learning models can recognize Hb1AC levels laboratory patterns that are beyond current medical knowledge, resulting in higher diagnostic accuracy compared to traditional quantitative interpretations based on reference ranges. These changes can be large, and physicians can observe them by checking for A1C level parameter values outside of normal ranges. Predictive models show great promise in medical laboratory diagnoses and could not only be of considerable value to both physicians and patients but also have widespread beneficial impacts on healthcare costs.

This study evaluated HbA1c by the of column chromatography with exchange resins in which patients with hemoglobin heterozygotes variants did not present a difference significant difference in relation to the control group.

REFERENCES

[1] Jordan, M. I. & Mitchell, T. M. “Machine learning: trends, perspectives, and prospects”. Science 349, 255–260, https://doi.Org/10.1126/science.Aaa8415 (2015).

[2] Van Ginneken, b. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. “Radiological physics and technology” 10, 23–32, https://doi.Org/10.1007/s12194-017-0394-5 (2017).

[3] K. Rajesh, V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis” in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.

[4] Sadhana, Savitha Shetty, “Analysis of Diabetic Data Set Using Hive and R”, International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.

[5] K. Rajesh, V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis” in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.

[6] Sadhana, Savitha Shetty, “Analysis of Diabetic Data Set Using Hive and R”, International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.
[7] Sabibullah M, Shanmugasundaram V, Raja Priya K, “Diabetes Patient’s Risk through Soft Computing Model”, International Journal of Emerging Trends & Technology in Computer Science, vol 2(6), 2013.
[8] V. H. Bhat, P. G. Rao, S. Krishna, and P. D. Shenoy, “An Efficient Framework for Prediction in Healthcare,” Most, Springer-Verlag Berlin Heidelberg , pp. 522-532, 2011.
[9] A. C. Tricco, N. M. Ivers, J. M. Grimshaw et al., “Effectiveness of quality improvement strategies on themanagement of diabetes: a systematic review andmeta-analysis,” The Lancet, vol. 379, no. 9833, pp. 2252–2261, 2012.
[10] M. C. Lansang and G. E. Umpierrez, “Management of inpatient hyperglycemia in noncritically ill patients,” Diabetes Spectrum, vol. 21, no. 4, pp. 248–255, 2008.
[11] R. Vinik and J. Clements, “Management of the hyperglycemic inpatient: tips, tools, and protocols for the clinician,” Hospital Practice, vol. 39, no. 2, pp. 40–46, 2011.
[12] K. J. Cios and G. W. Moore, “Uniqueness of medical data mining,” Artificial Intelligence in Medicine, vol. 26, no. 1-2, pp. 1–24, 2002.
[13] A. Frank and A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Science, 2010.
[14] R. M. Bergenstal, J. L. Fahrbach, S. R. Iorga, Y. Fan, and S. A. Foster, “Preadmission glycemic control and changes to diabetes mellitus treatment regimen after hospitalization,” Endocrine Practice, vol. 18, no. 3, pp. 371–375, 2012.