DIAGNOSING TYPE 1 DIABETICS SYMPTOMS USING BIG DATA ANALYTICS IN YOUTH

[1]Sultana Abdullah Albakry , [2]Bashayar Hijji Alshammari, [3]Kusum Yadav, [4]Afraa Sayah Alshammari
College of Computer Science and Engineering, Hail, Kingdom of Saudi Arabia
[1]s.albakry@uoh.edu.sa, 2 b.hijji@uoh.edu.sa, 3kusumasyadav0@gmail.com, [4]Afra.alshammari@uoh.edu.sa

ABSTRACT— *Nothing is more important than a healthy and peaceful life. A healthy life can be achieved with the alertness and awareness about the changes occurring in the human body system. An old saying that "prevention is better than cure" goes very well here. One can prevent the adverse conditions when they are aware before they occur and when the people are well equipped with the knowledge.*

Today, the world has been excited about machine learning and big data and the advancement in analytics. Big Data has proved its incredible influence from the global market to day-to-day life. The term Big Data is used to specify the huge volume of both structured and unstructured data. Big Data Analytics is the use of advanced analytical techniques against very large and diverse data sets that include different types and sizes. Big Data Analytics improves the health care system through the reduction run time and the optimal cost. It can assist the clinicians in decision making regarding any disease and improve patient compliance and satisfaction. This paper aims to diagnose the cause of sudden type 1 diabetics among young people (age ranging 28 years to 32 years) without any known symptoms. Unfortunately, medically the reason for the same has not been diagnosed.

Keywords—Big Data, Big Data Analytics, Weka Classification type 1 diabetics.

I. INTRODUCTION

Diabetes is an ancient disease that troubled humanity for centuries. Diabetes was first mentioned in a medical journal around the 15th century A.D. by Doctor Thomas Willis[4], who defined it as the disturbance of the human body's natural ability to produce insulin and manage blood sugar levels on daily basis. It is the main cause of many health problems like blindness, kidney failure, heart attacks, stroke and lower limb amputation due to complications. According to the World Health Organization, the number of people with diabetes rose from 108 million in 1980 to 422 million in 2014 [3]. It is the 7th leading cause of death hence, early identification is crucial to prevent severe outcomes of the disease. Prediabetes diagnostics are the process of noticing alarming signs that could develop to full case of diabetes and mostly affects younger age demographic. Saudi Arabia is one of the top 10 countries with a high prevalence of diabetes it is a scary territory especially in a society where more than 50% of its total population is younger than 20 years of age. Worldwide, over the past 20 years, the incidence of type 1 diabetes has almost doubled from 2.8% to 4.0% per year with more than 90% of the participants with diabetes being unaware of their disease.[2]. Diabetes not only is a medical problem, but it also has an economic impact on society estimated at $105 billion in 2016 alone.[5]

Out of other Preventive measures of developing type 1 or type 2 diabetes, noticing early signs and symptoms of prediabetes is one of the major concern in the modern medicine and its importance lies in the lifestyle consequences it imposes though relatively manageable. Utilizing machine learning technologies in order to further investigate a connection could prove useful as a measure to pre-diagnose individuals with Type 1 diabetes before more clinically significant symptoms emerges and manifests. Multiple successful studies which resulted in the real-world application in early detection of strokes that saved countless lives, another example of oncology research field is IBM's Watson system, which has a success rate of 99% of coherence with physician decisions of the treatment recommendations. [1] However, in most of the cases, diabetes and its consequences can be treated, avoided or delayed with diet, physical activity, medication and regular checkups and treatment for complications as they arise.[3].

Technology has helped to maintain diabetes to a great extent and today it's being used in helping patients maintain a healthy lifestyle and track blood level through a glucose monitoring system and nutrition coaching, which heavily use AI and machine learning to provide a significant insight to the patient and their physicians regarding their condition on daily basis. In recent years. OCR (Optical Character Recognition) and AR (Augmented reality) started to be utilized in informing the user of a meal's nutritional value and ingredients helping them make an informed decision about their dietary lifestyle. This has been very helpful for diabetics as they go on their daily life.

If we establish the role ML plays in assisting the process of eliciting key factors and recognizing early symptoms, we can design and apply tailored models and feed more data into the algorithms. This study aims to feed the available datasets into ML tailored models and analyze the output for viable patterns that would aid in recognizing prediabetes early signs. In the past, there were multiple attempts to execute a similar study manually through hundreds of personal visits and in-person surveys for family history inquiries. Technology can be applied here for a wider range and more condensed data repository for data mining and analysis. The purpose of this paper is to build intelligent diabetic classifier system which will detect diabetes using patient's data. To study the process of analyzing big data on diabetes symptoms in youth and the younger generation in order to better predict a preventative solution to diabetes complications that arise from ignorance to diagnose a case earlier.

II. RELATED WORK

Machine learning and big data usage in the biotechnology field had been significant and in an evolving way with different techniques as the literature review stated [1]. The machine learning techniques had been applied in diagnosing and prediction of verity diseases from strokes [2, 3], tumors [4, 5] and brain activities [6] in different parts of the body. For the scope of this paper, diabetes is the main concern and significant amount of research had been conducted in it. A recent paper by I. Kavakiotis et al were they provided a systematic review of the implementation of machine learning and data mining techniques in the field of diabetes research [7]. In their stud the focused on four aspects which were as follows:

1. Diagnosis and prediction.
2. Diabetic complications.
3. Genetic background and environment.
4. Health care and management.

Their research yields to find out that 85% of the machine learning techniques used were supervised and the remaining 15% were unsupervised. Moreover, they found that the Support Vector Machine (SVM) algorithm was the most extensive and successful used in this field. A research done by A. Rawshani et al studied the effect of ethnicity on glycaemic of patients with type two diabetes.nThey had used the Swedish National diabetes register (2002–2011) and choose patients who had recently diagnosed with type two diabetes, which they chose the period of diagnosis within twelve months [8]. Their findings show a vital relationship between ethnicity and type two diabetes, as non-Western origin have substantially higherHbA1c and therapy failure and higher risk to develop albuminuria, than native Swedes. The ethnicity impact was found more than the education and income impact, and at the same level as physical activities impact. Furthermore, J. Liu et al did research on predating type two diabetes from early changes in their metabolisms. In their study, they chose twenty-four markers of type two diabetes by using the Least Absolute Shrinkage and Selection operator regression [9]. In their novel prediction model, a relationship had been discovered and the specificity was improved for the young individuals. More validation for this model is required considering its novelty. Furthermore, this model has a high potential in resulting to a better understanding of the biological mechanisms responsible for the glycemic deterioration in prediabetes and diabetes. Another research by D. Lee et al did research investigating the association of low doses of some persistent organic pollutants (POPs) with type two diabetes, which they could not conclude but provided a further step in prediction diabetes using POPs as it is an important fact that affects obesity [10]. P. A. Modesti et al discovered new findings in type two diabetes in first-generation Chinese migrants who settled in Italy (Prato), by using a cross-sectional survey [11]. For analysis Logistic regression was used to evaluate the effect of the independent variables which are: age, sex, education categories, current smoking, alcohol use, total cholesterol, triglycerides, hypertension, BMI, and waist on the diagnosis of type two diabetes. As expected with age the disease correlated positively (55 – 59 years old) and is found in men more than women. Also, the same relationship is found with central obesity, waist-to-hip ratio and or waist-to-height. Additionally, M. Maniruzzaman et al had conducted a study on the role of outliers and missing values on the accuracy of machine learning in diabetes by replacing them with group median and median values [12]. For their study, ten different classifiers were used to investigate which are: linear discriminate analysis, quadratic discriminate analysis, naïve Bayes, Gaussian process classification, support vector machine, artificial neural network, Adaboost, logistic regression, decision tree, and random forest. The findings were a 10% improvement over developed techniques in literature and resulted in 100% validation by JK-based cross-validation protocol.

III. RESEARCH METHODOLOGY

A. Methodology

Nowadays, the healthcare institutions need to manage their data using information systems; as the system consists of huge data various data mining techniques are used to extract hidden information for developing an intelligent medical diagnosing information system. By using mining techniques we can predict the occurrence of diabetic most efficiently and also can provide the final prediction to support physicians' decisions for early, efficient patent's diagnosis.

To achieve the above-mentioned goals a survey has been conducted in Arabic as well as in English in the month of November and December 2018 consisting of the participants from KSA, India. Jordan, Tunisia Pakistan. The survey consists of mainly two sections i.e. the detailed information about the person who is answering the questionnaire, as he himself is diabetic or he is answering on behalf of a related diabetic. Then, the second part of the survey was on the medical history of the diabetic person. The second section was used to analyze with the help of algorithms implemented on the recorded dataset. Before analyzing the data the Arabic dataset was translated in English.

B. Participants

The participants in this study consisted of diabetics people of both types from all slice of society.

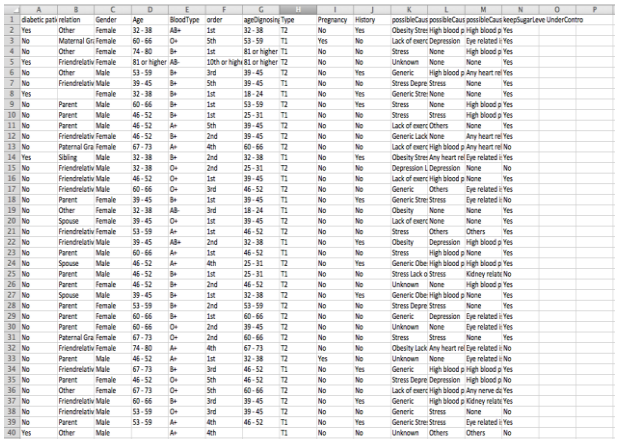
The main objective of this research is to build intelligent diabetic classifier System that gives a diagnosis of diabetic using diabetic patient's data. To develop this system, medical terms such as sex, Blood Type, and Patient's order among his/her siblings are used. In order to implement the models, Weka was used as the data-mining tool. Weka (Waikato Environment for Knowledge Analysis) is a data-mining tool written using java developed at Waikato. WEKA is a great data-mining tool to classify the accuracy on the basis of datasets by applying different types of algorithms and compared in the field of bioinformatics. Table 1 shows the WEKA data mining techniques that have been applied in this paper and other prerequisites like data set format.

Software	Dataset	Weka Data Mining Technique	Classification Algorithms	Operating System	Dataset File Format	Purpose
WEKA	Diabetic	Explorer	Naïve Bayes J48 ZeroR	Mac OS X	CSV	Classification

Fig. 1. Table 1.Weka data mining technique using different algorithms

C. Datasets

For analysis and diagnosing the reasons of two types of Diabetic, we used diabetic data set. The dataset has 12 attributes and 109 records, as shown in figure 1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	diabetic pati	relation	Gender	Age	BloodType	order	ageDiagnosis	Type	Pregnancy	History	possibleCaus	possibleCaus	possibleCaus	keepSugarLeve	UnderContro	
2	Yes	Other	Female	32 - 38	AB+	1st	32 - 38	T2	No	Yes	Obesity Stres	High blood p	High blood p	Yes		
3	No	Maternal Gra	Female	60 - 66	O+	5th	53 - 59	T1	Yes	No	Lack of exerc	Depression	Eye related i	Yes		
4	No	Other	Female	74 - 80	B+	1st	81 or higher	T1	No	No	Stress	None	High blood p	Yes		
5	Yes	Friendrelativ	Female	81 or higher	AB-	10th or high	81 or higher	T2	No	No	Unknown	None	None	Yes		
6	No	Other	Male	53 - 59	B+	3rd	39 - 45	T2	No	Yes	Generic	High blood p	Any heart rel	Yes		
7	No	Friendrelativ	Male	39 - 45	B+	5th	39 - 45	T1	No	No	Stress Depre	Stress	None	Yes		
8	Yes		Female	32 - 38	B+	1st	18 - 24	T1	No	Yes	Generic Stres	None	None	Yes		
9	No	Parent	Male	60 - 66	B+	1st	53 - 59	T2	No	Yes	Stress	None	High blood p	Yes		
10	No	Parent	Male	46 - 52	B+	1st	25 - 31	T1	No	No	Stress	Stress	High blood p	Yes		
11	No	Parent	Male	46 - 52	A+	5th	39 - 45	T2	No	No	Lack of exerc	Others	None	Yes		
12	No	Friendrelativ	Female	46 - 52	B+	2nd	39 - 45	T2	No	No	Generic Lack	None	Any heart rel	Yes		
13	No	Paternal Gra	Female	67 - 73	A+	4th	60 - 66	T2	No	No	Lack of exerc	High blood p	Any heart rel	No		
14	Yes	Sibling	Male	32 - 38	B+	2nd	32 - 38	T1	No	Yes	Obesity Stres	Any heart rel	Eye related i	Yes		
15	No	Friendrelativ	Male	32 - 38	O+	2nd	25 - 31	T2	No	No	Depression L	Depression	None	No		
16	No	Friendrelativ	Male	46 - 52	O+	1st	39 - 45	T1	No	No	Lack of exerc	High blood p	None	Yes		
17	No	Friendrelativ	Male	60 - 66	O+	3rd	46 - 52	T2	No	No	Generic	Others	Eye related i	Yes		
18	No	Parent	Female	39 - 45	B+	1st	39 - 45	T1	No	Yes	Generic Stres	Stress	Eye related i	No		
19	No	Other	Female	32 - 38	AB-	3rd	18 - 24	T1	No	No	Obesity	None	None	Yes		
20	No	Spouse	Female	39 - 45	O+	1st	39 - 45	T2	No	No	Lack of exerc	None	None	Yes		
21	No	Friendrelativ	Female	53 - 59	A+	1st	46 - 52	T2	No	No	Stress	Others	Others	Yes		
22	No	Friendrelativ	Male	39 - 45	AB+	2nd	32 - 38	T1	No	Yes	Obesity	Depression	High blood p	Yes		
23	No	Parent	Male	60 - 66	A+	1st	46 - 52	T1	No	No	Stress	High blood p	None	Yes		
24	No	Spouse	Male	46 - 52	A+	4th	25 - 31	T2	No	Yes	Generic Obe	High blood p	High blood p	Yes		
25	No	Parent	Male	46 - 52	B+	1st	25 - 31	T1	No	No	Stress Lack o	Stress	Kidney relate	No		
26	No	Parent	Female	46 - 52	B+	2nd	46 - 52	T2	No	No	Unknown	None	High blood p	Yes		
27	No	Spouse	Male	39 - 45	B+	1st	32 - 38	T2	No	Yes	Generic Obe	High blood p	None	Yes		
28	No	Parent	Female	53 - 59	B+	2nd	53 - 59	T2	No	No	Stress Depre	Stress	None	Yes		
29	No	Parent	Female	60 - 66	B+	1st	60 - 66	T2	No	No	Generic	Depression	Eye related i	Yes		
30	No	Parent	Female	60 - 66	O+	2nd	39 - 45	T2	No	No	Unknown	None	Eye related i	Yes		
31	No	Paternal Gra	Female	67 - 73	O+	2nd	60 - 66	T2	No	No	Stress	Stress	None	Yes		
32	No	Friendrelativ	Female	74 - 80	A+	4th	67 - 73	T2	No	No	Obesity Lack	Any heart rel	Eye related i	No		
33	No	Parent	Male	46 - 52	A+	1st	32 - 38	T2	Yes	No	Unknown	None	Eye related i	No		
34	No	Friendrelativ	Male	67 - 73	B+	3rd	46 - 52	T1	No	Yes	Generic	High blood p	High blood p	Yes		
35	No	Parent	Female	46 - 52	O+	5th	46 - 52	T2	No	No	Stress Depre	Depression	High blood p	No		
36	No	Other	Female	67 - 73	O+	5th	60 - 66	T2	No	No	Lack of exerc	High blood p	Any nerve da	Yes		
37	No	Friendrelativ	Male	60 - 66	B+	3rd	39 - 45	T2	No	Yes	Generic	High blood p	Kidney relate	Yes		
38	No	Friendrelativ	Male	53 - 59	O+	3rd	39 - 45	T2	No	No	Generic	Stress	None	No		
39	No	Parent	Male	53 - 59	A+	4th	46 - 52	T1	No	Yes	Generic Stres	Stress	Eye related i	Yes		
40	Yes	Other	Male		A+	4th		T1	No	No	Unknown	Others	Others	No		

Fig. 2. .Screeanshot of datasets

Table 2 shows the description of attributes in diabetes dataset used in this paper.

Attributes	Description
Gender	Male or Female
Age	Different Ages from 32 to 80
Blood Type	All types of blood
Patient's order	Patient's order among his/her siblings
Age First Diagnosed	Patient's age when diabetes was first diagnosed
Type	Type of diabetes, the patient is suffering from
Possible cause	
History	Is there any family history of diabetes
Medical condition BEFORE	Any medical condition the patient was suffering from, BEFORE getting diagnosed with diabetes
Medical condition AFTER	Any medical condition the patient is suffering from, AFTER getting diabetes
Sugar level control	Is patient able to keep the sugar level under control with the help of medication and change in the lifestyle

D. Data Mining Techniques

The analysis has been carried on using three data mining classification techniques, which are J48 and SVM and NaiveBayse. The accuracy of the classification was calculated by applying a confusion matrix. It shows how many instances have been assigned to each class. In this study, there are two classes in the experiment, and therefore it has a 2 x 2 confusion matrix. Ten-fold cross-validation technique was used to test the performance of the classification. In this method, the dataset is split into ten portions randomly and equally. Samples in each portion are left out in turn as the testing samples and samples in the rest portions are used to train the classifier. Therefore, each sample is tested exactly once in the original dataset.

J48—

J48 is a powerful algorithm for classifying data. It uses a pruning method to build a tree to reduce the size of the tree by removing overfitting data. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique provides the highest accuracy on training data. The overall concept is to make a tree that provides a balance of flexibility & accuracy.[16]

SVM—

Support vector machine(SVM) [13] is one of the powerful data mining algorithms for solving classifying problems. SVM is an algorithm used for learning classification and regression rules from data. SVM was initially proposed by Vapnik [14] in 1960 for classifying data. This machine classifies big data sets by categorizing a linear or non-linear separating surface in the input data. The separated surface depends on a subclass of the original data known as a set of support vectors. SVM makes a set of hyperplanes in a high dimensional space for classification. A good separation is done by the hyperplane that has the largest distance to the nearest training data points of any class, named functional margin. The generalization error of the classifier will be small If the functional margin is large.[13, 14, 15] Sequential Minimal Optimization (SMO) [12] is a simple algorithm that can quickly solve the SVM QP problem with no additional matrix storage and not using numerical QP optimization. The benefit of SMO is solving the Lagrange multipliers analytically.[13]

Naive Bayes—

Naive Bayes is a powerful algorithm for classification problems. It is based on Bayes' Theorem with an assumption of independence among predictors. This model is easy to build and particularly useful for very large data. Naïve and Bayes are two parts of this algorithm.The Naive Bayes algorithm assumes that the presence of an attribute value in a class is unrelated to any other values. [16]

Confusion matrix—

The accuracy of the solution to a classification problem can be illustrated by a confusion matrix technique. It contains information about actual and predicted classifications done by the selected classification algorithm.

Using data in the confusion matrix, the performance of the classification algorithm can be evaluated.[16]

IV. RESULTS OF ANALYSIS

The performance of three collected classification algorithms can be summarized using the confusion matrix which gives us a better idea of what our classification model is getting right and what types of errors it is making.

Classifying diabetes data using J48, SVM, and Naïve Bayes, the confusion matrix is generated to have two possible classes i.e. type1 or type2.

J48 Confusion Matrix—

Tab: 3

a	b	Classified as
74	**0**	**a = Type 2**
35	**0**	**b = Type 1**

SMO Confusion Matrix—

Tab: 4

a	b	Classified as
35	**2**	**a = Type 2**
26	**9**	**b = Type 1**

NB Confusion Matrix—

Table :5

a	b	Classified as
59	**15**	**a = Type 2**
26	**9**	**b = Type 1**

After running the classification, table 6 illustrates correctness results for the three selected al algorithms:

Table: 6

	Correctly Classified Instances	**Incorrectly Classified Instances**
J48	67.8899	32.1101
SMO	56.8807	43.1193
Naïve Bayes	62.3853	37.6147

From table 6, we can observe that J48 has the highest percentage of Correctly Classified Instances, whereas SMO has the minimum percentage.From these three data mining algorithms, we can find that J48 has the best performance to classify two types of diabetes dataset using Weka tool.

V. CONCLUSION

Big Data Analytics in using Weka analysis provides a systematic way for achieving better outcomes like reasons and causes of type 1 diabetics and help the user to identify medical historical data to predict and identify the symptoms of type 1 diabetics and so that the patient or the user can take the precaution before he suffers from the disease. Non-Communicable Diseases like diabetes is one of a major health hazard in the world. Early detection of diabetes can help the clinician and the patient to initiate the treatment on time, which can help to maintain the blood glucose level and prevent the complications, related to hyperglycemia and hypoglycemia. The prediction of this analysis will make the patient understand the complications which can occur. The goal of this research is to analyze medical historical data and predict type 1 diabetes using big data analytics. The design of predictive analysis system of diabetes treatment may give enhanced data and analytics yield the greatest results in healthcare. By employing location-aware healthcare services, anyone from an urban area to rural area can get proper treatment at low cost. This research mainly focused on the patients in the rural area. The patient or the clinician can use the data to find out diabetes. Treatment can be offered when the disease is confirmed. . Patients' awareness of diabetes and involvement can improve treatment compliance.

REFERENCES

[1] K. K. L. Wong, L. Wang and D. Wang, "Recent developments in machine learning for medical imaging applications," Computerized Medical Imaging and Graphics, vol. 57, pp. 1-3, 2017.

[2] G. N. Baksheev et al, "Validity of the 12-item General Health Questionnaire (GHQ-12) in detecting depressive and anxiety disorders among high school students," Psychiatry Research, vol. 187, (1), pp. 291-296, 2010;2011;.

[3] P. P. Sengupta et al, "Cognitive Machine-Learning Algorithm for Cardiac Imaging: A Pilot Study for Differentiating Constrictive Pericarditis From Restrictive Cardiomyopathy," Circulation. Cardiovascular Imaging, vol. 9, (6), 2016.

[4] M. Alilou et al, "Quantitative vessel tortuosity: A potential CT imaging biomarker for distinguishing lung granulomas from adenocarcinomas," Scientific Reports, vol. 8, (1), pp. 1-16, 2018.

[5] W. E. Muhlestein et al, "Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance," Neurosurgery, 2018.

[6] L. S. Prichep et al, "Classification algorithms for the identification of structural injury in TBI using brain electrical activity," Computers in Biology and Medicine, vol. 53, pp. 125-133, 2014.

[7] I. Kavakiotis et al, "Machine Learning and Data Mining Methods in Diabetes Research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104-116, 2017.

[8] A. Rawshani et al, "Impact of ethnicity on progress of glycaemic control in 131,935 newly diagnosed patients with type 2 diabetes: a nationwide observational study from the Swedish National Diabetes Register," BMJ Open, vol. 5, (6), pp. e007599-e007599, 2015.

[9] J. Liu et al, "Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study," Metabolomics, vol. 13, (9), pp. 1-11, 2017

[10] D. Lee et al, "Low Dose of Some Persistent Organic Pollutants Predicts Type 2 Diabetes: A Nested Case-Control Study," Environmental Health Perspectives, vol. 118, (9), pp. 1235-1242, 2010.

[11] P. A. Modesti et al, "New findings on type 2 diabetes in first- generation Chinese migrants settled in Italy: Chinese in Prato (CHIP) cross- sectional survey," Diabetes/Metabolism Research and Reviews, vol. 33, (2), pp. n/a, 2017.

[12] M. Maniruzzaman et al, "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers," Journal of Medical Systems, vol. 42, (5), pp. 1-17, 2018.

[13] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, (8), pp. 651-666, 2010.

[14] Joachims, T., "Making large-scale SVM learning practical", Advances in Kernel Methods –Support Vector Learning, B. Schokopf et al. (ed.), MIT Press, 1999.

[15] A. Ben-Hur et al, "Support vector machines and kernels for computational biology," PLoS Computational Biology, vol. 4, (10), pp. e1000173, 2008.

Tina R. Patil, Mrs. S. S. Sherekar, " Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," PLoS Computational Biology, vol. 4, (10), Apr 2013.