Comparing ML algo to predict Dibetes in women

Conference Paper · January 2021

CITATIONS
0

READS
65

1 author:

A. Saxena
Amity University
15 PUBLICATIONS **41** CITATIONS

SEE PROFILE

Comparing Machine Learning Algorithms to Predict Diabetes in Women and Visualize Factors Affecting It the Most—A Step Toward Better Health Care for Women

Arushi Agarwal and Ankur Saxena

Abstract Diabetes affects millions of people throughout the world, and more than half of the people suffering from it are women. Creating a better diagnosis and study tool will enable us to take a step forward in better healthcare. We use sklearn to create a model for the Pima Indians' Diabetes Dataset. The main goal is to compare the different algorithms to obtain the best accuracy. Prediction of diabetes in women is crucial as it not only ensures an early start of treatment, but also helps in prevention in cases of high probability of the disease occurring. We have not only focused on the detection part, but also tried to study and visualize the factors that were most correlated to a diabetic person. By studying the most common algorithms, we can figure out which area needs to be worked upon to develop better ways of healthcare. Machine learning has been actively used in health care and by implementing this in conditions like diabetes which affects a major population in the world, including almost 100 million Americans and more than 62 million Indians. The idea behind choosing the dataset was to get parameters and features, which are not determined by geography or region, but the overall physiology of women, so that most women throughout the world can be benefitted. The algorithms compared are decision trees, logistic regression, Naïve Bayes, SVM, and KNN. The final result got us an accuracy of 81.1% with the help of K-Fold and Cross-Validation.

Keywords Diabetes · Sklearn · Pima Indians · Decision trees · Logistic regression · KNN · Naïve Bayes · SVM · Diagnosis

A. Agarwal (✉) · A. Saxena
Amity Institute of Biotechnology, AUUP, Noida, India
e-mail: arushiagarwal14@gmail.com

A. Khanna et al. (eds.), *International Conference on Innovative Computing and Communications*, Advances in Intelligent Systems and Computing 1087,
https://doi.org/10.1007/978-981-15-1286-5_29

1 Introduction

Diabetes is one of the most common conditions affecting adults throughout the world, and while there are a lot of proposed lifestyle and medication changes advised for the treatment, very less-known and proven methods exist which ensure a reversal and betterment. Out of all the cases of diabetes, more than 50% of affected individuals are women. This is because the physiology of women causes a higher chance of diabetes compared to men and also makes it more difficult to control.

1.1 Type 1 Diabetes

It occurs when the immune system mistakenly kills the beta cells of pancreas responsible for making insulin; hence, insulin is not released in the blood, and sugar buildup takes place.

1.2 Type 2 Diabetes

It is the consequence of the body being unable to respond to insulin or when insulin produced is not enough. This can often be a result of insulin resistance or prolonged prediabetes. This constitutes almost 90% of all diabetes cases.

1.3 Gestational Diabetes

Sometimes women can be affected by diabetes only during the period of their pregnancy. This is usually due to the interference of pregnancy hormones with insulin. It is found in 9.2% of pregnancies (Fig. 1).

1.4 Detecting Diabetes

There are four methods of testing for diabetes, which are usually employed.

The A1C Test measures your average blood glucose for the period of 2–3 months. The advantages of being diagnosed this way are that the need of fasting is not there, and therefore, this is a comfortable option. Diabetes is diagnosed at an A1C of greater than or equal to 6.5%. However, the accuracy is usually lower in this method.

Fasting Plasma Glucose Test checks your fasting blood glucose levels. Fasting means after not having anything to eat or drink (except water) for at least 8 h before

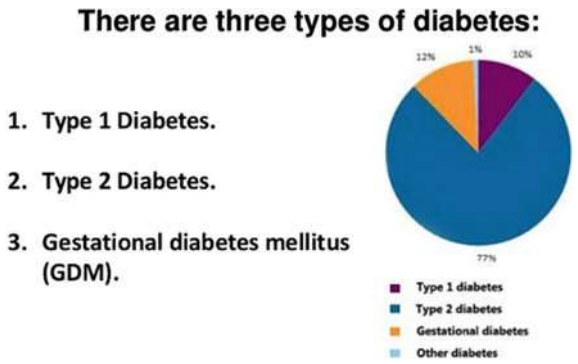


Fig. 1 Most common type of diabetes is type 2

the test. This test is usually done first thing in the morning, before breakfast. Diabetes is diagnosed at fasting blood glucose of greater than or equal to 126 mg/dl.

The Oral Glucose Tolerance Test is a two-hour test that checks your blood glucose levels before and 2 h after you drink a special sweet drink. It tells the doctor how your body processes glucose. Diabetes is diagnosed at 2-h blood glucose of greater than or equal to 200 mg/dl.

Random Plasma Glucose Test is a blood check at any time of the day when you have severe diabetes symptoms. Diabetes is diagnosed at blood glucose of greater than or equal to 200 mg/dl.

Most of the tests that are available, either take a long time, require fasting, have an insufficient accuracy or cause some other discomfort to the patient. Therefore, the need of better prediction and detection has been in place.

Moreover, none of these methods can tell the tendency of diabetes buildup in a person. It is only reasonable when it shows a prediabetes or a diabetes condition. This poses a requirement of a system which can accurately show how much tendency a woman has to build up this condition, so that it can be controlled before occurring. This is where machine learning plays a major role, as with sufficient data and training, it has the potential to detect and predict the onset of diabetes beforehand.

1.5 Machine Learning

Machine learning can help us build smart systems and machines, which learn by the examples and data fed into it. We do not need to program the machine for every instruction, which is very essential as every task cannot be manually programmed. Machine learning has been used in almost every part of our lives.

The task involves acquiring, analyzing, and wrangling our data, training our model to learn the data, testing the model to see how efficient it is, and then utilizing it while coming back to previous steps if needed. There are three ways, in which we can train our system:

Supervised learning is when we train the model with a set of example data, having the features, as well as the label or outcome. This involves learning through features and labels and then predicting based on the similarities and patterns to the training set.

Unsupervised learning involves training with only features, without any labels or outcomes. The model has to draw similarities and patterns in data and learn based on that. The most common method is clustering.

Reinforcement learning is based on learning through consequences and can be called similar to hit and trial method. This method is heavily used in gaming.

We will be comparing algorithms used heavily in supervised and unsupervised learning, for classification, regression, as well as clustering.

The algorithms we use and compare for this dataset are:

- Support vector machine,
- Decision tree,
- Logistic regression,
- K-nearest neighbor,
- Naïve Bayes.

1.6 About the Dataset

Pima Indians' Diabetes dataset is originally by the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of data from women of at least 21 years of age and takes into account the factors like BMI, number of pregnancies, blood pressure, insulin level, and so on. This is the best aspect of this dataset that it uses features that are physiological, and prevalent in women irrespective of their culture, location, etc.

This dataset is mainly focused on predicting whether a woman has diabetes or not, taking into account the different features provided to us, and the final outcome.

To try and predict diabetes using this dataset, we used Python's popular machine learning library called *scikit-learn*, or simply *sklearn*, with the intention to compare different types of machine learning algorithms to see which one can get us the maximum accuracy.

2 Methodology

2.1 Setting up the Environment

To get started with our model, we installed *Anaconda* to get our machine set up, as it makes it much easier to get all the libraries installed, along with the latest version

of Python (3.7 in our case). Another way to set up libraries is to use Pip to install libraries.

Jupyter notebook is the most suitable IDE for Python due to the user-friendly and simple interface. It is appropriate for machine learning as the graphs and data are displayed clearly in our kernel.

2.2 *Starting the Code*

We used sklearn, one of the most popular machine learning libraries of python. We imported our data in the form of a comma-separated values (CSV) format.

To start the code, we need to import a number of libraries and dependencies. The libraries we imported were inclusive of all, for data analysis, machine learning, and also visualization. Pandas, sklearn, seaborn, and matplotlib were some of the required ones.

Our dataset looks like this (first few rows) (Fig. 2):

The following are our parameters for the dataset and can be plotted individually like follows (Fig. 3):

We need to train our model to predict the outcome. Therefore, we need to divide our data into training and testing set. This can be done by using sklearn's *train_test_split*.

Training set (Fig. 4):

Test set (Fig. 5):

To better understand how our features impact diabetes in women, we obtained a correlation between the outcome and all the parameters considered (Fig. 6).

We further tried to study the relationship for all the parameters with each other using a representation of scatterplot matrix (Fig. 7).

```
In [7]: print (data.head())
```

```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

Fig. 2 Head() function of pandas displays the first five rows of the dataset

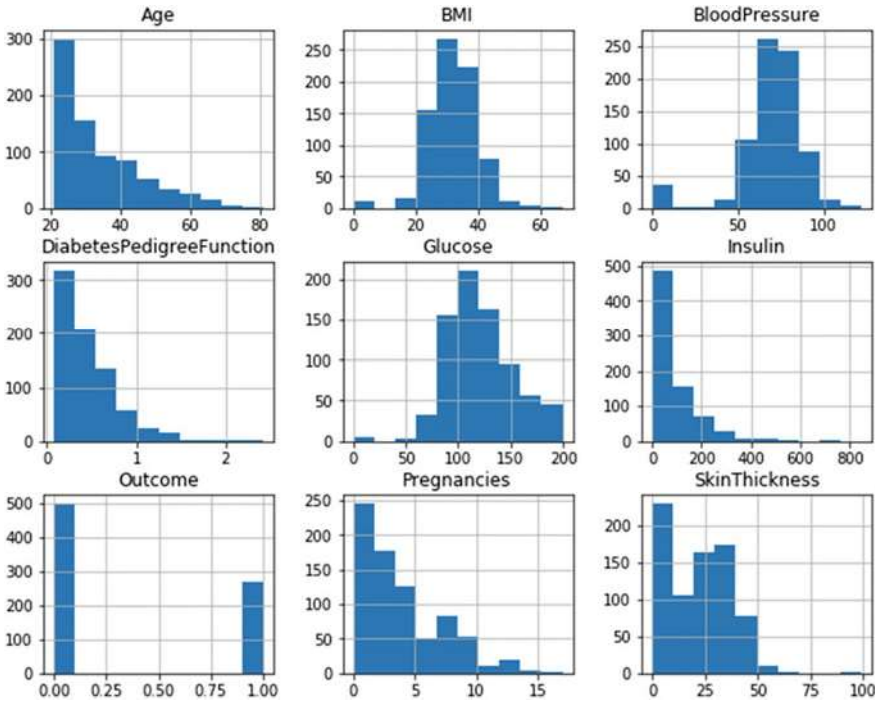


Fig. 3 Individual parameter can be studied through histograms. This is essential for data visualization and analysis

X_train:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
675	6	195	70	0	0	30.9
87	2	100	68	25	71	38.5
232	1	79	80	25	37	25.4
596	0	67	76	0	0	45.3
52	5	88	66	21	23	24.4

	DiabetesPedigreeFunction	Age
675	0.328	31
87	0.324	26
232	0.583	22
596	0.194	46
52	0.342	30

(614, 8)

Fig. 4 Training set for our model

X_test:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
476	2	105	80	45	191	33.7
322	0	124	70	20	0	27.4
76	7	62	78	0	0	32.6
94	2	142	82	18	64	24.7
215	12	151	70	40	271	41.8

	DiabetesPedigreeFunction	Age
476	0.711	29
322	0.254	36
76	0.391	41
94	0.761	21
215	0.742	38

(154, 8)

Fig. 5 Testing set for our model

```
In [20]: f,ax = plt.subplots(figsize=(10, 10))
         sns.heatmap(data.corr(), annot=True, linewidths=.5, fmt= '.1f',ax=ax)
         plt.show()
```

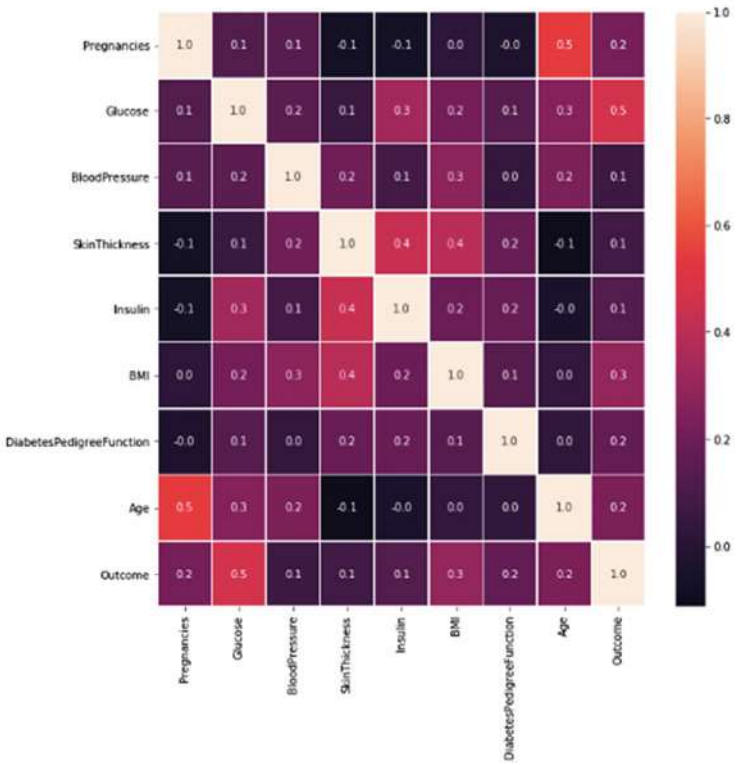


Fig. 6 This matrix shows us the inter-relationship of every parameter with each other by a float value

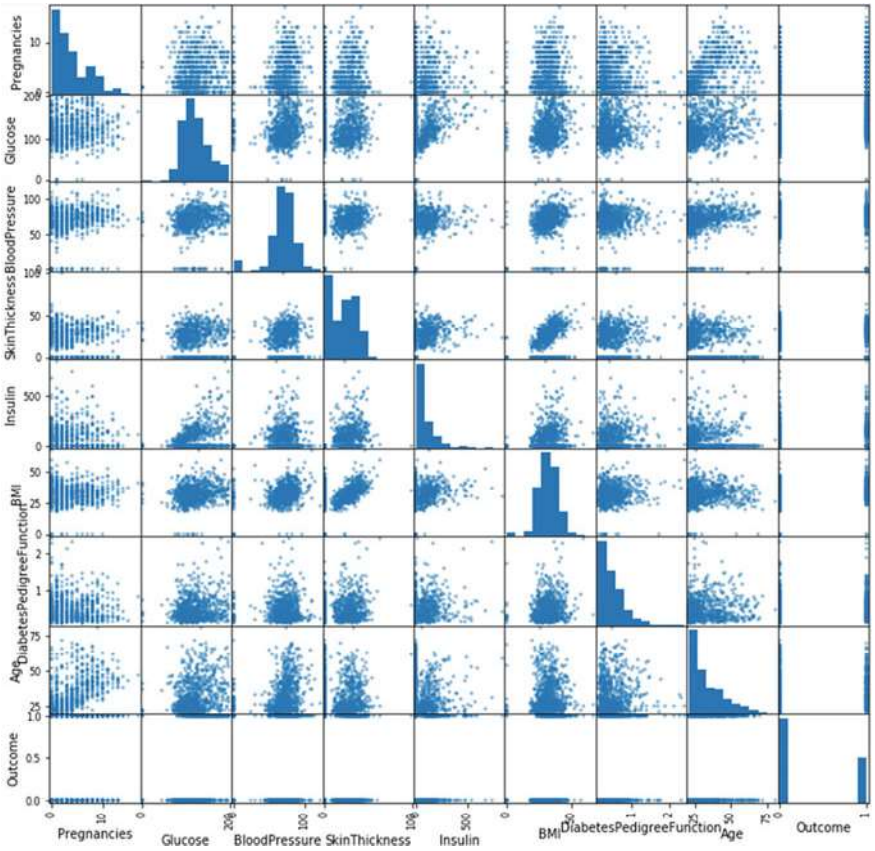
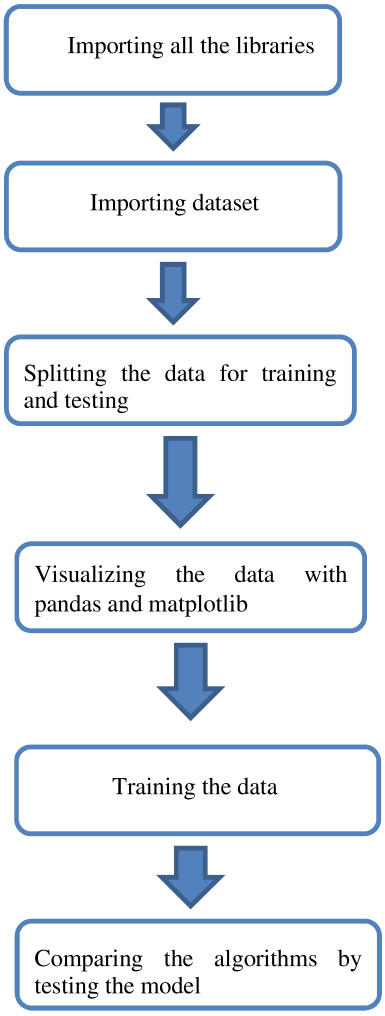


Fig. 7 Using a scatter-matrix, we can visualize the relationships between parameters in the form of scatter plots, which is easier to use for further analysis

The accuracies of all the algorithms were checked and compared. A function was used to do this to avoid repetition of code for each algorithm.

This will give us the accuracies, but to improve the results, we can implement K-Fold or Cross-Validation which can help us split our data into multiple parts (in our case, 10) so that we can use nine parts or "folds" to train our model and one part to test it. This will supposedly increase our accuracy due to more training.

The whole task can be summarized as follows:



3 Result

Our accuracy score that was obtained was (Fig. 8):

This gave support vector machine the highest accuracy score of **77.21%**

Our accuracy score after K-Fold or Cross-Validation was (Fig. 9):

In this case, the highest accuracy was by logistic regression giving a score of **81.16%** (Fig. 10).

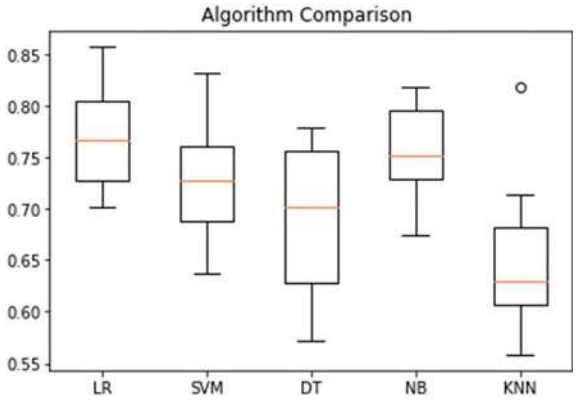
Fig. 8 Accuracies obtained originally

	Accuracy
Support Vector Machine	0.772129
Logistic Regression	0.769515
Naive Bayes	0.755178
K Nearest Neighbor	0.721275
Decision Tree	0.697847

Fig. 9 Accuracies obtained after K-Fold where folds = 10

	Accuracy
Logistic Regression	0.811688
Support Vector Machine	0.805195
Naive Bayes	0.785714
Decision Tree	0.753247
K Nearest Neighbor	0.720779

Fig. 10 Spread of accuracy score across each fold of each algorithm, through matplotlib



4 Conclusion

By trying different algorithms, we have reached the accuracy of 81.16%, through logistic regression. This accuracy is a positive indication, but not good enough to be used as a mainstream tool in the healthcare sector as of now. By visualizing and analyzing the data, we can conclude that some factors play much more significant role when it comes to diabetes. Factors such as glucose levels, BMI, age, and pregnancy

played more roles than skin thickness and blood pressure. We also concluded that while some factors may not be useful in indicating the future possibility of diabetes, they may have undergone notable changes because of already occurred diabetes, which in turn becomes essential for detection.

5 Future Scope

The model needs to be more accurate and specific. Also, it may be noted that because of several types of diabetes, it is not sufficient to just detect positive or negative, as the type needs to be indicated as well due to the difference in medication and lifestyle attributed to each type. The accuracy needs to be brought up to at least 98% for serious consideration over conventional methods. The dataset and features need advancement to cover every aspect of the condition and train the system most efficiently. Also, future improvements may be made through neural networks and more advanced algorithms after the dataset is improved.

Acknowledgements This paper would have not been accomplished without the support of Dr. Ankur Saxena, who was a guide and mentor throughout the process.

References

1. A. Rathore, S. Chauhan, S. Gujral, in *Detecting and predicting diabetes using supervised learning: an approach towards better healthcare for women*, IGDTUW Kashmiri Gate, Delhi, India
2. O. Chandrakar, J.R. Saini, *Development of Indian weighted diabetic risk score (IWDRS) using machine learning techniques for type-2 diabetes*. ACM COMPUTE'16, 21–23 Oct 2016
3. A.G. Karegowda, A.S. Manjunath, M.S. Jayaram, Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. Int. J. Soft Comput. (IJSC) **2**(2) (2011)
4. V.V. Vijayan, C. Anjali, Prediction and diagnosis of diabetes mellitus—a machine learning approach, in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Trivandrum, 10–12 Dec 2015
5. V.V. Kamadi, A.R. Allam, S.M. Thummala, A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. Appl. Soft Comput
6. Y. Hayashi, S. Yukita, Rule extraction using recursive-rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Inf. Med. Unlocked
7. A. Sarvwar, V. Sharma, Intelligent Naive Bayes approach to diagnose diabetes type-2. Special Issue Int. J. Comput. Appl. Issues Challeng. Netw. Intell. Comput. Technol. (2012)
8. R. Motka, V. Parmar, Diabetes mellitus forecast using different data mining techniques. IEEE Int. Conf. Comput. Commun. Technol. (ICCCT) (2013)
9. S. Sapna, A. Tamilarasi, M. Pravin, Implementation of genetic algorithm in predicting diabetes. Int. J. Comput. Sci. Issues **9**, 234–240 (2012)

10. S. Karatsiolis, C.N. Schizas, Region based support vector machine algorithm for medical diagnosis on pima Indian diabetes dataset, in *IEEE Conference on Bioinformatics and Bioengineering* (2012), pp. 139–144
11. A. AlJarullah Asma, Decision discovery for the diagnosis of type II diabetes, in *IEEE Conference on Innovations in Information Technology* (2011), pp. 303–307
12. D.M. Nirmala, S. Balamurugan, A. Appavu, U.V. Swathi, An amalgam KNN to predict diabetes mellitus, in *IEEE International Conference on Emerging Trends in Computing Communication and Nanotechnology (ICECCN)* (2013), pp. 691–695
13. P. Undre, H. Kaur, P. Patil, Improvement in prediction rate and accuracy of diabetic diagnosis system using fuzzy logic hybrid combination, in *International Conference on Pervasive Computing (ICPC)* (2015). pp. 1–4
14. S.S. Vinod Chandra, S. Anand Hareendran, in *Artificial Intelligence and Machine Learning* (PHI learning Private Limited, Delhi, 110092, 2014)
15. R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inf. **77**, 81–97 (2008)
16. A. Agarwal, A. Saxena, in *Machine Learning—A Simple and Modern Approach to Biometrics*. IndiaCom IEEE (2017)
17. A. Agarwal, A. Saxena Special Issue Malignant tumor detection using machine learning through scikit-learn: machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Int. J. Pure Appl. Mathem. **119**(15), 2863–2874 (2018) http://www.acadpubl.eu/hub/
18. J. Wiens, Clinical Infectious Diseases **66**(1), 153 (6 Jan 2018) https://doi.org/10.1093/cid/cix731
19. S. Saria, A.K. Rajani, J. Gould, D. Koller, A.A. Penn, Integration of early physiological responses predicts later illness severity in preterm infants. Sci. Transl. Med. 2:48ra65 (2010)
20. D.C. Kale, D. Gong, Z. Che et al. An examination of multivariate time series hashing with applications to health car, in *IEEE International Conference on Data Mining (ICDM)* (2014), pp. 260–69