JOURNAL OF
EMERGING INVESTIGATORS

Article

Entropy-based subset selection principal component analysis for diabetes risk factor identification

Valentino Pratama[1], Jimmy Tjen[2]

[1] Sekolah Menengah Atas (SMA) Santo Paulus Pontianak, Pontianak, West Kalimantan, Indonesia

[2] Informatics, Universitas Widya Dharma Pontianak, Pontianak, West Kalimantan, Indonesia

SUMMARY

Diabetes is one of the most common diseases, with an estimated 10% of the population suffering from this disease. Therefore, it is important to be able to identify the symptoms of this disease as early as possible before the patient's condition worsens and requires more expensive medical treatment. We aimed to study whether the entropy-based subset selection principal component analysis (E-ss PCA) can diagnose if a person is diabetic. The E-ss PCA is a novel machine learning method that can identify important parameters from a dataset. The E-ss PCA method was originally developed to fix the linearity problem that occurred when the principal component could not be written as the linear combination of the original parameters. Through the process, the E-ss PCA generates subsets of data that guarantee the linearity of variables of the subset. Via the E-ss PCA algorithm, we aim to verify which diabetic risk factors, such as pregnancy, triceps skinfold thickness, Body Mass Index (BMI), pedigree function, and age, are significant. Based on a dataset of diabetes patients from the United States National Institute of Diabetes and Digestive and Kidney Diseases, the E-ss PCA method was able to predict whether a person has diabetes or not with an average accuracy of 97.30%, which is higher than the classical PCA with an average accuracy of 94.45%. Furthermore, the proposed algorithm identified that the risk factors that accurately predict diabetes are BMI, triceps skin fold thickness, and blood pressure.

INTRODUCTION

Diabetes, one of the most common chronic diseases in humans, decreases the body's ability to regulate blood glucose levels which can result in a decrease in quality of life and life expectancy (1). Based on data from 2017, around 451 million people suffer from this disease, and it is estimated that this value will grow to 693 million by 2045 (2). Furthermore, at the end of 2021, it was estimated that about 10.5% of the global population has diabetes (3). Diabetes patients often experience other complications, such as heart disease, decreased eyesight/vision, amputation of body parts, and kidney disorders due to excess sugar levels in the bloodstream (4). There is no known method for curing diabetes, however, there are some preventive steps recommended, such as losing weight, consuming healthy foods rich in fiber, exercising regularly, and getting regular and adequate medical examinations. Early detection can allow patients to avoid severe complications, which may improve life quality and expectancy (5). Therefore, it is crucial to produce a precise predictive model that can diagnose whether a person has diabetes (6).

Principal component analysis (PCA) is an analytical method that extracts information from large data sets by maximizing the variance of the data through orthogonal projection. From this transformation, it is possible to shrink the dimensionality of the data but retain most of the information (in this case the variance) of the data (7). Several related works utilize this method (8–11).

Previous publications have reported the application of PCA to the field of diabetes care (8, 9). Choubey et al. used the PCA method to classify parameters that correlate with diabetes in diabetic patients (8). This method paired with the logistic regression and k-means was also used to predict diabetes in the Pima-Indian dataset, from which the authors concluded that the PCA method PCA enhanced the k-means clustering algorithm and logistic regression classifier in classifying diabetic and non-diabetic samples (9). Furthermore, recent publications proposed the idea using of the entropy-based subset selection (E-ss) PCA, to perform fault detection on structures, which has improved the performance of algorithms for other prediction tasks (10, 11). From the experiment, it was shown that the E-ss PCA improves the PCA algorithm in detecting faults, while potentially requiring a much smaller number of sensors (5 instead of 24).

The PCA method has also been used in various research topics related to health. While PCA is a versatile algorithm, it does have some flaws. In particular, PCA depends strongly on the linearity among variables. If this assumption is not satisfied, the PCA algorithm cannot perform optimally (10). The E-ss algorithm improves the existing method by specifically handling the linearity issue of the PCA. In particular, the E-ss algorithm fixes the problem of linearity in PCA by selecting a number of parameters that correlated strongly (in the entropy sense) with the target parameter. This step will efficiently remove the parameters that are not linearly correlated to the principal component. Thus, it helps to maintain the linearity between parameters and principal components.

In this research, we modified the E-ss method to determine risk factors that diagnose whether someone has diabetes or not. In particular, the E-ss PCA model is no longer viewed as a timed-based, auto-regressive function (i.e., a function that depends on the value of that function from the previous moment) as in previous studies, but as a time-invariant function, which is a function that does not depend to the time parameter (as patients can be considered as a discrete dataset) (10, 11). We hypothesize that this assumption will allow us first to verify whether the E-ss PCA algorithm can diagnose if someone has diabetes or not, better than the classical PCA and to study its accuracy in predicting the

https://doi.org/10.59720/22-284

presence of diabetes. Furthermore, we aim to study the risk factors that are significant in determining a person's potential for developing diabetes from the dataset obtained from the United States National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). In particular, we performed the experiment based on the E-ss PCA paired with a linear regression model algorithm, and to compare the result to the existing literature (12).

A patient shows signs of diabetes if the 2-hour post-load plasma glucose was measured to be at least 200 mg/dl during the examination or if detected during a medical checkup. We hypothesize that the factors that most strongly predict diabetes are BMI, and skinfold thickness. We believed these two factors were significant in determining whether a person has diabetes or not, as these metrics correlate with obesity. We showed that the E-ss PCA combined with a linear model is able to predict whether someone has diabetes, better than the classical PCA. With this research, an early diabetes prediction can be performed on diabetes-suspected patients, such that proper medical care can be applied to them in order to avoid more severe symptoms.

RESULTS

We compared our proposed algorithm for identifying diabetic samples to the classical PCA algorithm. In particular, the whole process consists of 2 steps: first, we generated a subset of parameters that can predict whether someone has diabetes, and then, we derived a mathematical model to predict the presence of diabetes based on the obtained dataset (12). We validated the algorithm against nondiabetic and diabetic samples. In this research, we considered 8 risk factors: pregnancy, glucose, blood pressure, triceps skinfold thickness, insulin, BMI, Diabetic pedigree function and age. To simplify the names of the risk factors, all of the risk factors are expressed as variables x_1 to x_8 (**Table 1**). For example, the model built by considering the *i*-th risk factor as the main risk factor is referred to as "x_i-model".

Model predictive accuracy

The model predictive accuracy was validated on various model, where, each model may have a different number of parameters on it, according to the other risk factors which strongly correlate (in the entropy sense) with the main risk factor. During our experiment setup, the classical PCA algorithm assigned many false negatives, decreasing its accuracy in some models to below 50%. Furthermore, we observed that the x_6-model generated from the subset of data containing the risk factors BMI, skinfold thickness, and blood pressure was the model which produced the highest accuracy for both cases (**Table 2**). Using this model, E-ss PCA predicted nondiabetic and diabetic samples with 96% and 99.25% accuracy, respectively, while classical PCA achieved 90.74% and 100% accuracy.

The visual representation of the E-ss PCA generated by the *x6*-model, which is the best model for the E-ss PCA, showed that there was a difference on the data pattern between diabetic and non-diabetic samples (**Figure 1**). As a comparison, the visual representation for the x_1-model, failed to detect any change of dynamics for diabetic samples with accuracy only 1.5% (**Figure 2**). In particular, for data that is non-diabetic, the new sample relatively followed the distribution of the model formed from the non-diabetic dataset (**Figure 1A**). However, for the diabetic samples, we found that the samples are distributed over a new kind of distribution such that they no longer occupied the upper and lower limits of the negative diabetes data (**Figure 1B**). We were able to validate that BMI and skinfold thickness are indeed the best parameters in deciding whether someone has diabetes or not according to the E-ss PCA algorithm.

Risk factor	Mathematical representation
Pregnancy	x_1
Glucose	x_2
Blood pressure	x_3
Tricep skinfold thickness	x_4
Insulin	x_5
Body Mass Index (BMI)	x_6
Diabetic pedigree function	x_7
Age	x_8

Table 1. Mathematical variable representation for each risk factor.

DISCUSSION

In this research, we have developed the E-ss PCA algorithm proposed by Smarra et. al. to predict whether someone is diabetic. Based on our research, the x_6-model version of E-ss PCA, which consists of BMI, triceps skin thickness, and blood pressure parameters is the best model for describing whether a person is diabetic or nondiabetic. In particular, this claim was made by considering the model predictive accuracy for both diabetic and nondiabetic was the highest (on average) than the classical PCA, while only requiring information from 3 (instead of 8 as in the classical method) risk factors. Conversely, model x_3 which consists of blood pressure and BMI information is the worst, as it categorized diabetic samples with an accuracy below 30%. These results show that, of the eight risk factors that we examined, BMI, triceps skin thickness, and blood pressure are the parameters that most influence the accuracy of the predictive model. Also, we found that the algorithm failed to derive the x_7-model. This is due to the E-ss PCA algorithm's failure to find any risk factor that correlates with the 7th risk factor, the pedigree function which is a function that calculates of diabetes likelihood of a person based on his/her age and diabetic family history. We believed that this parameter failed, as age (x_8) was also not considered to be a "good" parameter in predicting whether someone is diabetic. Thus, it is not possible to find any risk factors that correlate with this parameter. Intriguingly, E-ss PCA only uses at most five risk factors out of eight, yet is able to produce results close (or even better in several cases) with respect to the classical PCA. This result supports the capability of the E-ss PCA in minimizing the number of parameters as previously stated (11). In a real-life setup, this information is useful, since it allows early prediction given

Model	Subset (E-ss)	%A (non-diabetic)		%A (diabetic)	
		E-ss PCA	Classical PCA	E-ss PCA	Classical PCA
x_1	x_1, x_8	99 %	22.83 %	1.5 %	100 %
x_2	x_2, x_5	100 %	0 %	0 %	100 %
x_3	x_3, x_6	95 %	0 %	8.2 %	27.9 %
x_4	x_4, x_6, x_5	95 %	91.97%	11.5 %	100 %
x_5	x_5, x_4, x_2	96 %	97.53%	6.34 %	6.33 %
x_6	x_6, x_4, x_3	96 %	90.74%	99.25 %	100 %
x_7	-	-	48.76%	-	100 %
x_8	x_1, x_8	99 %	100%	1.49 %	52.5 %

Table 2. Model predictive accuracy comparison between E-ss and classical PCA algorithms.

https://doi.org/10.59720/22-284

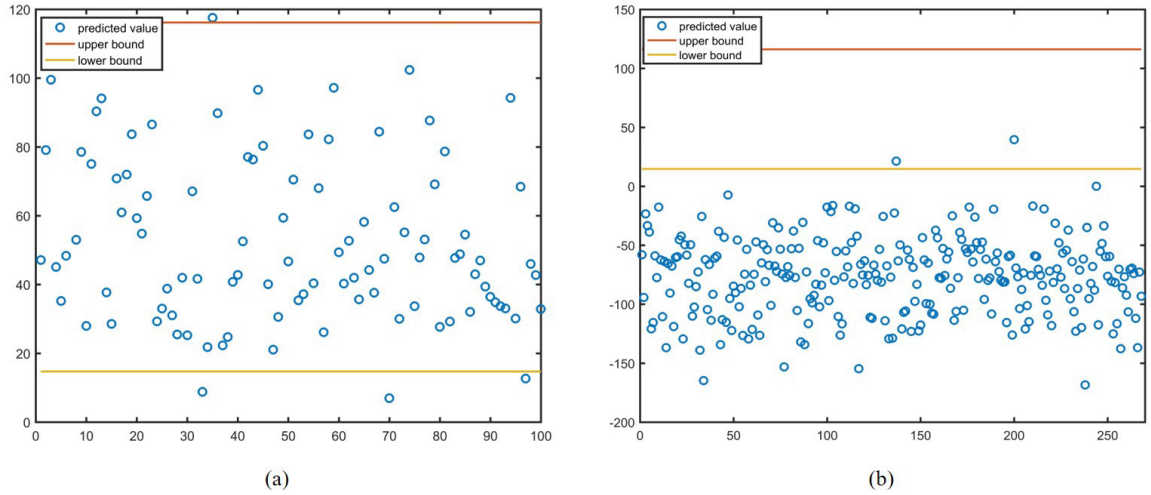

Figure 1. The visual representation of the E-ss PCA (x6-model) approach in diagnosing diabetes. Dynamics of the model on **A)** non-diabetic samples and **B)** diabetic samples. The algorithm can differentiate two kinds of datasets based on the change of the dynamics in the dataset.

some health parameters may take longer than the others in order to be measured precisely.

Furthermore, the risk factors that mostly influence the diagnosis results from samples collected from the NIDDK are BMI, triceps skin thickness, and blood pressure (12). This result supports with our hypothesis that skinfold thickness and BMI are the most significant factors in determining whether a person has diabetes or not. Moreover, these findings are consistent with several previous publications (13–17).

Reports have shown that BMI is a risk factor which strongly correlated with the likelihood of someone having diabetes from samples taken in the United States and Ghana (13, 14). Furthermore, other studies reported that skin thickness measured at the triceps affected the diagnosis of whether a person is diabetic (15, 16). There is also strong evidence of a correlation between high blood pressure and diabetes (17).

Based on our findings, the E-ss PCA method can accurately diagnose whether a person has diabetes or not with an average accuracy of 97.30% This accuracy is better than the classical PCA, which is 94.45%. Furthermore, the results of the risk factor analysis produced by E-ss PCA found that the most influencing risk factors, namely BMI, skin thickness, and blood pressure, were also supported by previous research. Thus, the E-ss PCA method can be a used as a framework for building machine learning software for diagnosing diabetes. In particular, the software can be used as an early indicator to determine whether someone is diabetic or not, such that a quick and precise diagnosis can be performed right on the spot when all necessary parameters related to the patient's health have been measured. For example, as in this research, BMI and skinfold thickness were identified as parameters that strongly correlated to diabetes. These parameters can be measured easily with respect to the conventional diabetes test which relies on the blood glucose measurement, where this measurement is usually considered to be relatively expensive in some developing countries.

In this paper, we considered only a single diabetic dataset from the US. In a future study, we would like to consider wider

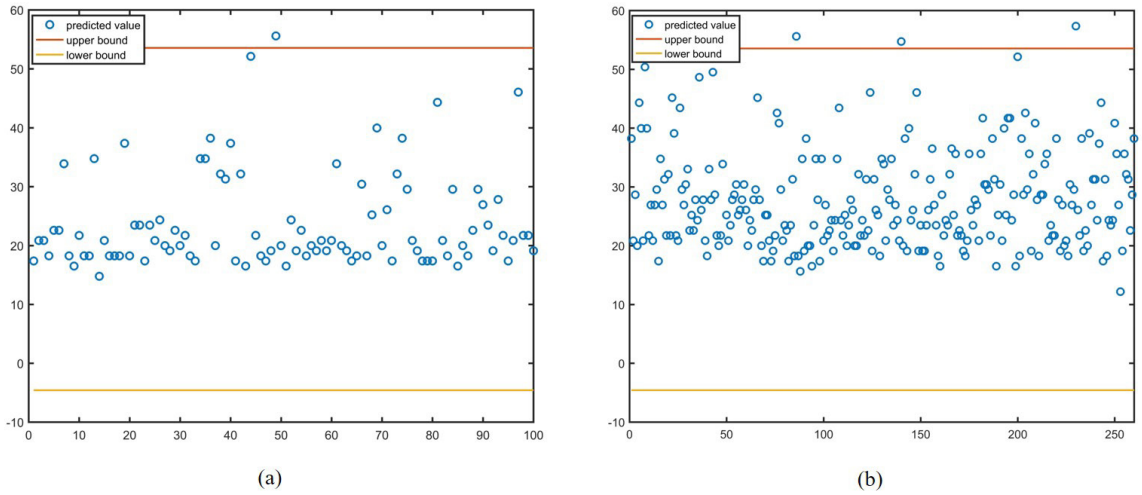

Figure 2 The visual representation of the E-ss PCA (x1-model) approach in diagnosing diabetes. Dynamics of the model on **A)** non-diabetic samples and **B)** diabetic samples. In this case, the model failed to differentiate the diabetic data.

https://doi.org/10.59720/22-284

data including the race of possibly diabetic patients in order to obtain a more accurate model.

MATERIALS AND METHODS

Case study: Diabetics data from the US NIDDK

Our methodology was validated on a dataset provided by the NIDDK (12). The dataset consists of 768 female patients (either nondiabetic and diabetic) as well as eight parameters: pregnancy, glucose, insulin, tricep skinfold thickness, BMI, pedigree function, and age. From the dataset, there are 500 non-diabetic samples while the rest, 268, are diabetic samples.

400 samples were used (which is 80% of the total available non-diabetic data) as the training dataset to build the E-ss PCA model. Then, the rest of the non-diabetic samples were used as a test dataset to validate the performance of the algorithm in predicting non-diabetic cases, and 268 samples from the diabetics' dataset were used to validate the algorithm in diagnosing diabetic cases. MatLab was used to perform all simulations in this study. Mathematical derivation of equations used in our model are given in the **Appendix**.

Received: January 14, 2023
Accepted: July 27, 2023
Published: November 18, 2023

REFERENCES

1. Egan, A. M. and S. F. Dinneen. "What is diabetes?" *Medicine*, vol. 47, no. 1, Jan. 2019, https://doi.org/10.1016/j.mpmed.2018.10.002.
2. Cho, N. H., et al. "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045." *Diabetes research and clinical practice,* vol. 138, Apr. 2018, https://doi.org/10.1016/j.diabres.2018.02.023.
3. "Percentage of diabetics in the global adult population in 2021, 2030, and 2045,", *Statista*, www.statista.com/statistics/271464/percentage-of-diabetics-worldwide. Accessed 12 December 2022.
4. Papatheodorou, K., et al. "Complications of diabetes 2017." *Journal of diabetes research*, Mar. 2018, https://doi.org/10.1155/2018/3086167.
5. Bilous, R., et al. "diabetes." *Handbook of diabetes,* John Wiley & Sons, 2021, pp. 7-10.
6. Lai, H., H. Huang, et al. "Predictive models for diabetes mellitus using machine learning techniques." *BMC Endocrine Disorders*, vol. 19, no. 1, Oct. 2019, https://doi.org/10.1186/s12902-019-0436-6.
7. Lever, J., et al. "Points of significance: Principal component analysis." *Nature methods*, vol. 14, no. 7, Jul. 2017.
8. Choubey, D. K., et al. "Performance evaluation of classification methods with PCA and PSO for diabetes." *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, Dec. 2020, https://doi.org/10.1007/s13721-019-0210-8.
9. Zhu, C., et al. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques." *Informatics in Medicine Unlocked*, vol. 17, Apr. 2019, https://doi.org/10.1016/j.imu.2019.100179.
10. Tjen, J., et al. "An entropy-based sensor selection algorithm for structural damage detection." *in 2020 IEEE 16th International Conference on Automation Science and Engineering*, Online virtual meeting, Aug. 2020, https://doi.org/10.1109/CASE48305.2020.9216828.
11. Smarra, F., et al. "Learning methods for structural damage detection via entropy-based sensors selection." *International Journal of Robust and Nonlinear Control*, vol. 32, no. 10, Mar. 2022, https://doi.org/10.1002/rnc.6124.
12. "Diabetes Dataset." *National Institute of Diabetes and Digestive and Kidney Diseases*, www.kaggle.com/datasets/mathchi/diabetes-data-set. Accessed 7 December 2022
13. Verma, S. and M. E. Hussain. "Obesity and diabetes: an update." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 11, no. 1, Feb. 2017, doi:10.1016/j.dsx.2016.06.017
14. NCD Risk Factor Collaboration. (NCD-RisC). "Trends in obesity and diabetes across Africa from 1980 to 2014: an analysis of pooled population-based studies." *International journal of epidemiology,* vol. 46, no. 5, Jun. 2017, https://doi.org/10.1093/ije/dyx078.
15. Alwash, S. M., et al. "Triceps skinfold thickness and body mass index and the risk of gestational diabetes mellitus: Evidence from a multigenerational cohort study." *Obesity Research & Clinical Practice*, vol. 16, no. 1, Feb. 2022, https://doi.org/10.1016/j.orcp.2021.12.005.
16. Liu, Hong y., et al. "Effects of waist-to-height ratio and skinfold thickness on impaired fasting glucose and diabetes in the elderly." *Chinese General Practice*, vol. 21, no. 25, Sep. 2018, https://doi.org/10.12114/j.issn.1007-9572.2018.25.013.
17. Chiriaco M., et al. "Association between blood pressure variability, cardiovascular disease, and mortality in type 2 diabetes: a systematic review and meta-analysis." *Diabetes, Obesity and Metabolism*, vol. 21, no. 12, Jul. 2019, https://doi.org/10.1111/dom.13828.
18. Wang, L. "Enhanced fault detection for nonlinear processes using modified kernel partial least squares and the statistical local approach." *The Canadian Journal of Chemical Engineering*, vol. 96, no. 5, Oct. 2018, https://doi.org/10.1002/cjce.23058.

Copyright: © 2023 Pratama and Tjen. All JEI articles are distributed under the attribution non-commercial, no derivative license (http://creativecommons.org/licenses/by-nc-nd/3.0/). This means that anyone is free to share, copy and distribute an unaltered article for non-commercial purposes provided the original author and source is credited.

APPENDIX

Mathematical derivation

Let $X = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \dots \boldsymbol{x_n}]; X \in \mathbb{R}^{m \times n}$ be the dataset related to the diabetic's risk factors, with $\boldsymbol{x_i} = [x_1(1)\ x_i(2) \dots x_i(m)]^{\intercal}; \boldsymbol{x_i} \in \mathbb{R}^m$ is the $i-th$ risk factors vector, where $i = 1,2,\dots,n$. As the first step, diabetic and nondiabetic data was split into different categories. Let, $\boldsymbol{x_j} = [x_j(1)\ x_j(2) \dots\ x_j(m)]^{\intercal}; \boldsymbol{x_j} \in \{0,1\}; j \in i$ be the categorical vector, which indicates whether someone is diabetic or not, i.e. 0 indicates the nondiabetic sample and 1 otherwise. Furthermore, let $X_d \in \mathbb{R}^{m^* \times n}$ be the subset of diabetic samples (that is: $x_j(k) = 1, \forall x_j(k) \in X_d; k = 1,2,\dots,m^*$) and $X_{td} \in \mathbb{R}^{m^{**} \times n}$ be the subset of nondiabetic samples (provided the condition: $x_j(k) = 0, \forall x_j(k) \in X_{td}; k = 1,2,\dots,m^{**}$) such that $X_d \cup X_{td} = X; X_d \cap X_{td} = \emptyset$.

In the next step, subsets of data for non-diabetic samples were constructed. For every $\boldsymbol{x_i} \in X_{td}$, let S_i be the subset of data generated from the E-ss algorithm as in [10] and [11] with the cardinality of the subset, $|S_i| = d$. For each subset S_i, the data which are contained in it will be transformed to their proper orthogonal subspace via the Eigen Value Decomposition (EVD). Let $V_i \in \mathbb{R}^{n^* \times n^*}$ be the eigen vector of the correlation matrix, R_i of the subset S_i. Consequently, for each subset, the matrix Ψ_i, which was the matrix of Principal Component (PC) was defined as follows:

$$\Psi_i = X_i V_i\,, \qquad (\mathbf{Eq.\ 1})$$

with $X_i \subset X_{td} = \{\boldsymbol{x_k} \in X_{td} | k \in S_i\}$ such that X_i was the dataset corresponding to the subset S_i. A mathematical equation to connect each PC was arranged in the last step. Let $\boldsymbol{\psi_{ik}}$ be the k^{th} PC from the orthogonal matrix Ψ_i, and $\boldsymbol{\psi_{i1}}$ denotes the first PC (in this case, the first column) of Ψ_i.

For each i, a simple regression model was constructed to relate $\boldsymbol{\psi_{i1}}$ with $\boldsymbol{\psi_{ik}}$ as follows:

$$\hat{\psi}_{i1}(t) = \sum_{k=2}^{n^*} \alpha_k \psi_{ik}(t)\,, \qquad (\mathbf{Eq.\ 2})$$

where α_k is the regression model parameter obtained from any regression model, e.g., the least square method (see: (18)) and $\hat{\psi}_{i1}(t)$ is the prediction of $\boldsymbol{\psi_{i1}}$ at instance t. Equation 2 produced a mathematical model on the orthogonal subspace, which was sensitive toward the nominal data (in this paper, it was the nondiabetic data). Thus, when exposed to diabetic data, the dynamic generated by the model would change abruptly. This idea then can be used in order to understand whether the test sample, in particular, is diabetic or not. Specifically, by defining an upper bound

$$u_b = \bar{\boldsymbol{\psi}}_{\boldsymbol{i1}} + M.\sigma(\boldsymbol{\psi_{i1}}) \qquad (\mathbf{Eq.\ 3})$$

and lower bound

$$l_b = \bar{\boldsymbol{\psi}}_{\boldsymbol{i1}} - M.\sigma(\boldsymbol{\psi}_{\boldsymbol{i1}}), \quad (\mathbf{Eq.\ 4})$$

where u_b and l_b are the bounds which can categorize whether a sample is diabetic or not, $M \in \mathbb{Z}^+$ is the multiplier which represents the number of standard deviations away from the mean, $\bar{\boldsymbol{\psi}}_{\boldsymbol{i1}}$ and $\sigma(\boldsymbol{\psi}_{\boldsymbol{i1}})$ respectively are the mean and standard deviation of $\boldsymbol{\psi}_{\boldsymbol{i1}}$. In general, M can take on any integer value. However, one needs to understand that choosing M which is sufficiently small will increase the chance of inducing false negatives, while a large M will increase the chance of false positives.

For any unknown sample, the diagnosis to identify whether a sample is diabetic or not can be done as follows: let $\boldsymbol{x_t} = [x_1(t)\ x_2(t)\ \dots x_n(t)]$ be the test sample. Then in this case, with a slight abuse of notation, the sample x_t is diagnosed as nondiabetic if it satisfies the following condition:

$$l_b \leq x_t V_i(1) \leq u_b \quad (\mathbf{Eq.\ 5})$$

and diabetic if otherwise, where $x_t V_i(1)$ denotes the first PC of the orthogonal transformation of the sample x_t.

Algorithm 1 shows the algorithm of diabetic identification based on the E-ss PCA algorithm (**Appendix 2**). Technically, the algorithm itself consists of three main parts: entropy-based subset selection, orthogonal projection by the PCA method, and model parameter identification with the least square algorithm. The E-ss PCA is known to have a time complexity of $O(mn^2)$ [11] where m is the number of samples and n is the number of parameters or features, while the least square algorithm, in this case, is in the $O(mn)$. Thus, the whole algorithm itself is running in the time complexity of $O(mn^2) + O(mn) = O(mn^2)$.

Furthermore, to represent the goodness of the model, the accuracy of the model is defined as follows:

$$\%A = \frac{T}{n} \times 100\% \quad (\mathbf{Eq.\ 6})$$

Where $\%A$ denotes the accuracy of the model, T is the number of true predictions and n is the total number of samples.

Algorithm 1: diabetes diagnosis based on the E-ss PCA algorithm

Input: $X_d, X_{td}; d; \theta; M; X_t$
Output: status
Process:
for $k = 1 : i$ do
S_k = e-ss(X_{td}, k, d, θ)
$d = |S_k|$
$X_k = \{\boldsymbol{x_j} \in X_{td} | j \in S_k\}$
$R_k = corr(X_k)$
$V_k = eig(R_i)$
$\Psi_k = X_k V_k$
$X = [\psi_{k2}\psi_{k2} \dots \psi_{kd}]$
$Y = \psi_{k1}$
$\beta = (YX^{\intercal})(X^{\intercal}X)^{-1}$
$b_a = \bar{Y} + M.\sigma(Y)$
$b_b = \bar{Y} - M.\sigma(Y)$
for $o = 1 : \text{length}(X_t)$
$\hat{\psi}_{i1}(o) = \beta X_t(o)$
if $\hat{\psi}_{i1}(o) > l_b$ && $\hat{\psi}_{i1}(o) < u_b$
status(o) = "non-diabetic"
else
status(o) = "diabetic"
end if
end for
end for

Figure S1. Algorithm for the diabetes prediction based on E-ss PCA