Use of Raman Spectroscopy to Diagnose Diabetes with SVM

Le Anh Duc and Nguyen Thanh Tung(✉)

International School, Vietnam National University, Ha Noi, Vietnam
me@anhducle.com, tungnt@isvnu.vn

Abstract. In this work, investigations were made for exploring the potential of machine learning in predicting Type 2 Diabetes Mellitus patients. In overall, 20 patients were assessed. The Raman spectrum was observed in four anatomical locations of the body: ear lobe, inner arm, thumb nail and cubital vein. The measurements were taken to examine the difference between Control and DM2 (9 well-controlled patients and 11 diabetic patients). To create effective diagnostic algorithms for categorization among these categories, multivariate approaches such as principal component analysis (PCA) paired with support vector machine (SVM) were applied. Based on the implemented classification systems, diabetic patients are classified using PCA-SVM shows the best potential of 80% in accuracy. Therefore, the taken approach successfully creates a classification and evaluation system. Overall, our findings show that the combination of Raman spectroscopy, PCA-SVM has several advantages in terms of preciseness and is suggested as a viable non-invasive diagnostic technique for diabetes.

Keywords: Diabetes · Raman spectroscopy · Machine learning · SVM · Diagnose

1 Introduction

Diabetes is a dangerous, chronic disease that arises when the pancreas fails to generate enough insulin (a hormone that regulates blood sugar, or glucose) or when the body fails to utilise the insulin that is produced adequately. Diabetes is a major public health issue, and it is one of four priority noncommunicable diseases (NCDs) that world leaders have identified for action. Diabetes has been progressively increasing in both the number of cases and the prevalence during the last few decades [1].

In Vietnam, diabetes cases have more than doubled in the previous ten years, with nearly twice as many patients diagnosed. According to the International Diabetes Federation, diabetes-related treatment expenditures in Vietnam averaged US$ 162.7 per patient per year [2], while average income was averaged at US$ 182.7 per month in 2020 [3]. Furthermore, the number of individuals with pre-diabetes is three times that of those with diabetes [4].

C. V. Phan and T. D. Nguyen (Eds.): ICTCC 2022, LNICST 473, pp. 79–87, 2023.
https://doi.org/10.1007/978-3-031-28790-9_6

Early detection for diabetes can be achieved with relatively inexpensive blood sugar monitoring. Nowadays, invasive blood glucose detection technology is now commonplace, convenient, and practical in both hospitals and home by glucometers that employ the procedure of blood sample first and then analyzing it in vitro for blood glucose measurement. Traditional glucose monitoring devices employ the electrochemical approach [5], which needs a certain amount of blood to be extracted out of the body through finger pricking or a thin lancelet placed subcutaneously. Unfortunately, most diabetics find it unpleasant to check their blood glucose levels on a regular basis since blood extraction is required many times each day for monitoring purposes.

2 Related Work

Recent studies show that blood glucose monitoring can be accomplished using non-invasive methods [6]. With the rise of worldwide diabetes in recent years, an increasing number of people have experienced discomfort and infections as a result of the intrusive nature of popular commercial glucose meters. Non-invasive blood glucose monitoring technology has become a global research focus as well as a novel way that might help a large number of patients. Based on the detection principle, researchers have been working on major problems of non-invasive blood glucose detection technology. Medical, materials, optics, electromagnetic waves, chemistry, biology, computational science, and other sectors are covered by this new method. The advantages and limitations of non-invasive and invasive technologies, as well as electrochemistry and optics in non-invasives were discussed in [7]. Therefore, non-invasive blood glucose monitoring will become more efficient, inexpensive, robust, and competitive on the market as wearable technology and transdermal biosensors advance.

In the last 20 years, a lot of researches has been conducted in non-invasive blood glucose testing. For non-invasive measurements, the researchers discovered a variety of optical techniques such as near-infrared (NIR) [8], photoacoustic spectroscopy, Raman spectroscopy [9], polarization techniques, and light scattering techniques [10]. A transilluminated laser beam is used to monitor glucose levels.

Non-invasive glucose monitoring is obviously the most attractive approach for diabetic patients, allowing for more frequent, if not continuous, assessments without discomfort or sensation. Until now, various approaches for non-invasive blood glucose testing have been proposed. However, the key challenge with non-invasive procedures is ensuring high accuracy of test result.

3 Machine Learning

There is no use in offering a definition for Machine Learning (ML) without first addressing the larger context in which it exists: the domain of Artificial Intelligence.

Artificial Intelligence (AI) is a notion that refers to any approach that allows computers to mimic human behavior in order to address and solve problems. Machine Learning, which uses statistical approaches to enable machines to recognize certain patterns by learning and improving through experience on a set of provided data, became more widespread in the 1980s.

In the decades afterwards, AI has been praised as the key to our civilization's brightest future, and derided as a harebrained idea of over-reaching propellerheads. Until 2012, there was a little bit of both. AI has surged in the last several years, particularly after 2015 [11]. Much of this is due to the widespread availability of GPUs, which makes parallel computing quicker, cheaper, and more powerful. It also has to do with the simultaneous one-two punch of almost endless storage and a deluge of data of all kinds (the whole Big Data movement)—photos, text, digital signal, transactions, mapping data, etc. (Fig. 1).

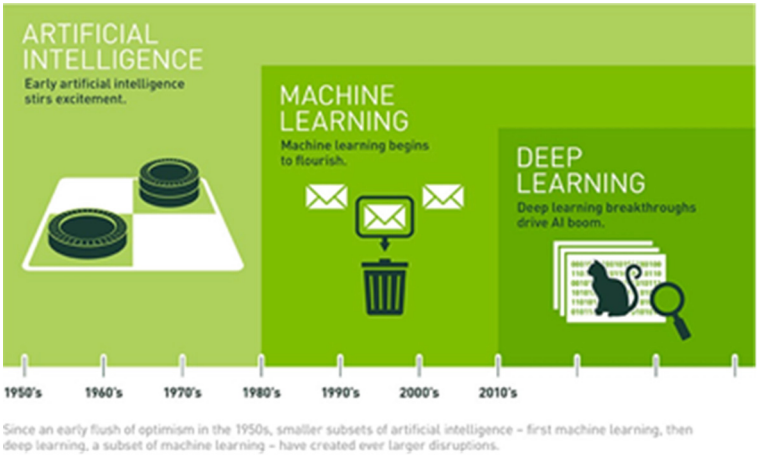

Fig. 1. Relationship between Artificial Intelligence, Machine Learning and Deep Learning

Machine Learning is a subset of AI. As defined by Wikipedia, Machine learning is the subfield of computer science that "gives computers the ability to learn without being explicitly programmed" [12]. Simply put, Machine Learning is a subfield of Computer Science that has the ability to learn on its own based on input without having to be specifically programmed.

Machine Learning has gone a long way and is a new subject in recent years, as computing power has been boosted to a new level and massive amounts of data have been collected by large technological businesses, Deep Learning (DL) was born. Deep Learning has enabled computers to perform tasks that were previously impossible: classifying hundreds of distinct objects in photographs, creating captions for images, imitating human voices and handwriting, communicating with humans, and even composing literature or music [13, 14].

4 Support Vector Machine

Support Vector Machine (SVM) is a linear model for classification and regression issues. It can handle linear and non-linear problems and is useful for a wide range of practical applications. The SVM concept is straightforward: The method draws a line or a hyperplane that divides the data into classes. The fundamental method to data classification

begins with attempting to develop a function that divides the data points into the relevant labels with (a) the fewest possible mistakes or (b) the greatest feasible margin. This is because larger vacant spaces around the splitting function result in less error since the labels are easier to differentiate from one another.

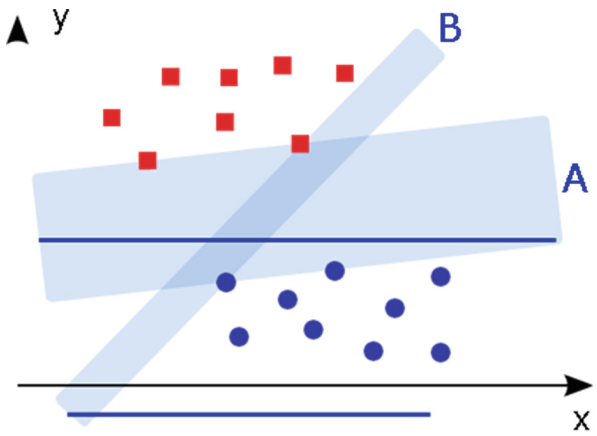

Fig. 2. Support Vector Machine separating a data set into two classes using two different linear separations, resulting in differing sized margins around the splitting functions.

Figure 2 shows that a data set may be separated by numerous functions with no errors. As a result, the margin surrounding a separating function is utilized as an extra parameter to assess separation quality. In this situation, separation A is preferable since it differentiates the two classes more precisely.

The objective of this algorithm is to achieve a hyperplane in an n-dimensional space that connects the data points to their possible classes. The hyperplane should be positioned as close to the data points as possible. The data points having the shortest distance to the hyperplane are referred to as Support Vectors [15]. Because of their near proximity, their effect on the exact position of the hyperplane is greater than that of other data points. The Support Vectors are the three points (2 blue, 1 green) laying on the lines in Fig. 3.

Any hyperplane can be written as the set of points $\mathbf{x}'$ satisfying

$$\mathrm{W}^T\mathrm{X} - b = 0$$

where $\mathbf{w}$ is the (not necessarily normalized) normal vector to the hyperplane. The input space consists of $\mathbf{x}$ and $\mathbf{x}'$. Therefore, $\Phi(\mathbf{x_i})$ represents the kernel function that turns the input space into a higher-dimensional space, so that not every data point is explicitly mapped. The kernel function can also be written as: $k(\mathbf{x}, \mathbf{x}')$.

5 Experimental Setup and Design

At the University of Guanajuato in Mexico, 11 patients with type 2 diabetes (DM2, 7 females, Age: 49.5 $\pm$ 6.7 years) and 9 healthy volunteers (Ctrl, 7 females, Age: 33.2

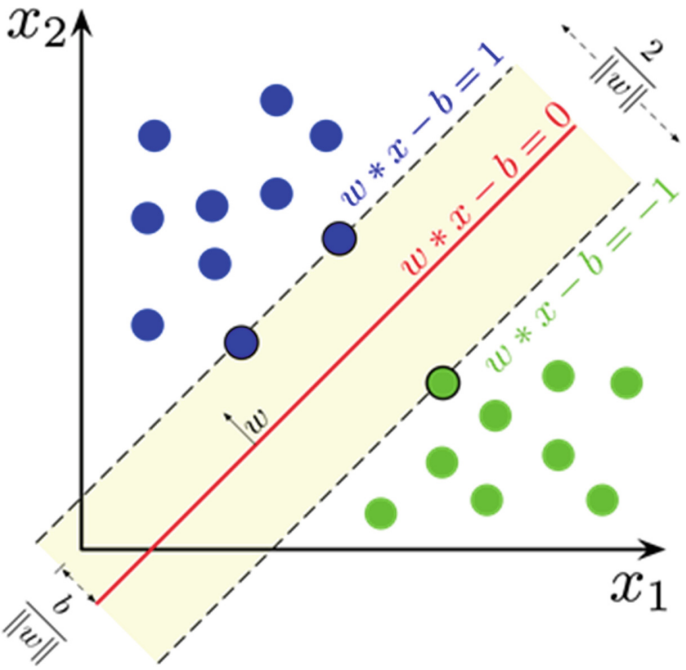

Fig. 3. Maximum-margin hyperplane and margins for an SVM trained on two classes of data. The support vectors are samples on the margin [15].

± 4.9 years) were investigated [16]. Before individuals were enrolled, the Institutional Review Board authorized the study and all subjects supplied informed consent. All DM2 patients in this research were previously diagnosed by their doctors using normal procedures such as a fasting plasma glucose test. Since the traditional method of monitoring HbA1C in human blood sample is regarded as the gold standard for long-term glycemic management in diabetic patients [17]. To evaluate HbA1C levels, a sample volume of 5 μl of blood was taken from each patient and analyzed using boronic acid affinity chromatography (LabonaCheck MH-200, Ceragem Medisys Inc.) (Fig. 4).

6 Implementation for PCA-SVM

Given 1000 dimension for each observation, it is hard to fit into an estimator with large number of features. The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

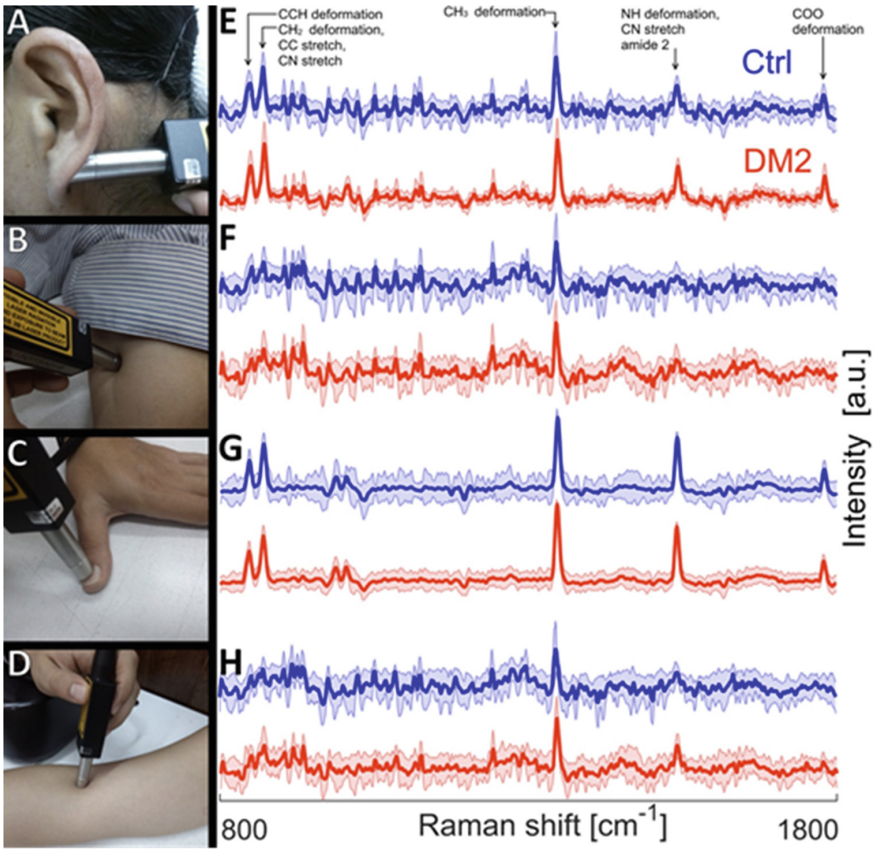

Fig. 4. Skin site images for in vivo Acquisition of Raman spectra: (A) ear lobe, (B) inner arm, (C) thumb nail, and (D) median cubital vein. The equivalent Raman observations (mean standard deviation) collected at an excitation wavelength of 785 nm (E-H) are also presented on the right side, with control spectra displayed in blue and DM2 spectra displayed in red [16].

As described previously, a PCA step is necessary to solve problems with high dimension data. Figure 5 shows an example of data acquired after being processed with Sci-kit Learn library in Python:

As can be seen in Fig. 6, particularly in ear lobe dataset, the initial number of components is chosen to match 99% explained variance ratio so that PCA compresses the main information in original dataset to produce new data that has 17 dimension. The explained variance ratio is the percentage of variation assigned to each of the chosen components. To minimize overfitting, the number of components should be selected to include in the ML model by adding the explained variance ratio of each component until it reaches a total of roughly 99%. After this processing step, the data is ready to be fed into a SVM model (Table 1).

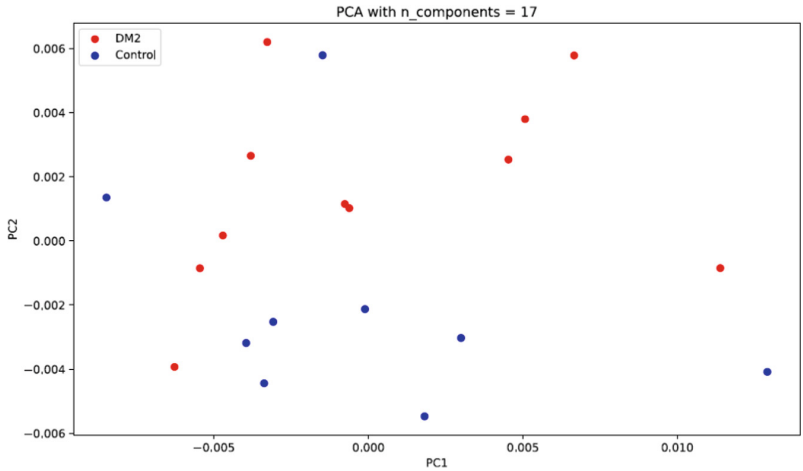

Fig. 5. Visualization of 2 first PCs obtained with total of 17 components in ear lobe dataset.

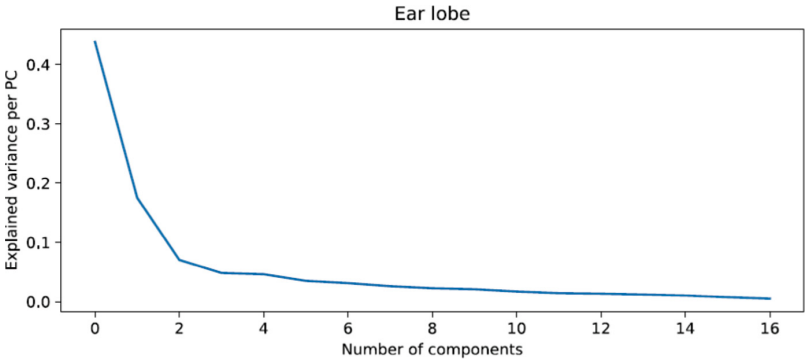

Fig. 6. Explained variance ratio for each PC.

Table 1. Experimental result of PCA-SVM.

	Mean accuracy (%)	
	5-fold mean accuracy	10-fold mean accuracy
Ear lobe	65	60
Inner arm	65	65
Thumb nail	70	70
Cubital vein	80	75

7 Conclusion

In this dissertation, I have several techniques for processing Raman spectroscopy, mainly 2 approaches for Diabetes Mellitus type 2 classification task, one is by using traditional Machine Learning method (combination of PCA and SVM). The investigation is extended from the dataset provided by [16]. Proposed accuracy for the finest classification model is 80%. Cubital vein is also suggested to be the most precise point of in vivo measurement.

This is a novel technique for fast non-invasive diabetes mellitus screening and can be practically utilized in clinical centers to assist traditional invasive testing procedure due to its considerable advantages in terms of handiness and preciseness. Raman spectroscopy and Machine Learning techniques, in general can be combined for detecting diabetic patients.

References

1. World Health Organization. Global Report on Diabetes (2016). ISBN 978 92 4 156525 7. http://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257eng.pdf
2. Triton Market Research. Vietnam Glucose Monitoring System Market 2019–2025 (2019). https://www.researchandmarkets.com/reports/4803838/vietnam-glucose-monitoring-system-market-2019-2025
3. VNA. "Vietnam's average monthly income in 2020 down 1 percent" (2021). https://en.vietnamplus.vn/vietnams-average-monthly-income-in-2020-down-1-percent/204206.vnp. Accessed 25 May 2022
4. Quinn Ryan Mattingly. The growing burden of diabetes in Viet Nam (2016). https://www.who.int/vietnam/news/feature-stories/detail/the-growing-burden-of-diabetes-in-viet-nam. Accessed 25 May 2022
5. Clark Jr, L.C., Lyons, C.: Electrode systems for continuous monitoring in cardiovascular surgery. Ann. New York Acad. Sci. **102** (1), 29–45 (1962). https://doi.org/10.1111/j.1749-6632.1962.tb13623.x, https://nyaspubs.onlinelibrary.wiley.com/
6. Caduff, A., Etienne Hirt, Y., Feldman, Z.A., Heinemann, L.: First human experiments with a novel non-invasive, non-optical continuous glucose monitoring system. Biosens. Bioelectron. **19**(3), 209–217 (2003)
7. Tang, L., Chang, S.J., Chen, C.J., Liu, J.T.: Non-invasive blood glucose monitoring technology: a review. Sensors, **20**(23), 6925 (2020). ISSN 1424–8220. https://www.mdpi.com/1424-8220/20/23/6925
8. Menon, K.U., Hemachandran, D., Abhishek, T.K.: A survey on non-invasive blood glucose monitoring using NIR (2013). https://doi.org/10.1109/iccsp.2013.6577220
9. Abdallah, O., Bolz, A., Hansmann, J., Walles, H., Hirth, T.: Design of a compact multi-sensor system for non-invasive glucose monitoring using optical spectroscopy (2012)
10. Anas, M.N., Nurun, N.K., Norali, A.N., Normahira, M.: Non-invasive blood glucose measurement. In: 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences, pp. 503–507 (2012)
11. Machine learning in healthcare – a brief introduction (2021). https://genomed4all.eu/2021/06/08/machine-learning-in-healthcare-a-brief-introduction/
12. Machine learning (2022). https://en.wikipedia.org/wiki/Machine_learning
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)

14. Brownlee, J.: Inspirational applications of deep learning (2019). https://machinelearningmastery.com/inspirational-applications-deep-learning/
15. Support-vector machine (2022). https://en.wikipedia.org/wiki/Support_vector_machine
16. Guevara, E., Torres-Galván, J.C., Ramírez-Elías, M.G., Luevano-Contreras, C., González, F.J.: Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools. Biomed. Opt. Express **9**(10), 4998–5010 (2018)
17. Jeppsson, J.O., et al.: Approved IFCC reference method for the measurement of hba1c in human blood (2002)