Spectral forecast: A general purpose prediction model as an alternative to classical neural networks

Paul A. Gagniuc, Constantin Ionescu-Tirgoviste, Elvira Gagniuc, et al.

COLLECTIONS

Paper published as part of the special topic on When Machine Learning Meets Complex Systems: Networks, Chaos and Nonlinear Dynamics

View Online   Export Citation   CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Combining machine learning with knowledge-based modeling for scalable forecasting and subgrid-scale closure of large, complex, spatiotemporal systems
Chaos: An Interdisciplinary Journal of Nonlinear Science **30**, 053111 (2020); https://doi.org/10.1063/5.0005541

Introduction to Focus Issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics
Chaos: An Interdisciplinary Journal of Nonlinear Science **30**, 063151 (2020); https://doi.org/10.1063/5.0016505

Supervised chaotic source separation by a tank of water
Chaos: An Interdisciplinary Journal of Nonlinear Science **30**, 021101 (2020); https://doi.org/10.1063/1.5142462

Spectral forecast: A general purpose prediction model as an alternative to classical neural networks

View Online   Export Citation   CrossMark

Paul A. Gagniuc,[1,2,a)] Constantin Ionescu-Tirgoviste,[2,3] Elvira Gagniuc,[4] Manuella Militaru,[4]
Lawrence Chukwudi Nwabudike,[3] Bujorel Ionel Pavaloiu,[1] Andrei Vasilățeanu,[1] Nicolae Goga,[1]
George Drăgoi,[1] Irinel Popescu,[2,5] and Simona Dima[2]

AFFILIATIONS

[1] Department of Engineering in Foreign Languages, Faculty of Engineering in Foreign Languages, University Politehnica of Bucharest, Splaiul Independentei 313, Bucharest 060042, Romania
[2] Center of Excellence in Translational Medicine, Fundeni Road 258, Bucharest 022328, Romania
[3] Institute of Diabetes, Nutrition and Metabolic Diseases "N.C. Paulescu," Ion Movila Street nr. 5-7, Bucharest 20475, Romania
[4] Department of Pathology, Faculty of Veterinary Medicine, University of Agronomic Sciences and Veterinary Medicine, Mărăşti Boulevard 59, Bucharest 011464, Romania
[5] Medical Academy of Sciences, Ion C. Brătianu Boulevard 1, Bucharest 030171, Romania

Note: This paper is part of the Focus Issue, "When Machine Learning Meets Complex Systems: Networks, Chaos and Nonlinear Dynamics."
[a)] Author to whom correspondence should be addressed: paul_gagniuc@acad.ro

ABSTRACT

Here, we describe a general-purpose prediction model. Our approach requires three matrices of equal size and uses two equations to determine the behavior against two possible outcomes. We use an example based on *photon-pixel coupling* data to show that in humans, this solution can indicate the predisposition to disease. An implementation of this model is made available in the supplementary material.

A novel prediction method is described, implemented, and tested. The model revolves around three known states: two extreme outcomes (*A* and *B*) and one measurement (*P*). These states are represented by matrices that include sets of homologous parameters. An information spectrum is described as a series of predicted states (M_1, M_2, M_3,..., M_d) generated between the two extreme outcomes (*A* and *B*). The predicted states are compared with the known state (*P*) from the measurements to generate a similarity index. The trend generated by the values of the similarity index indicates how a system may behave against these two extreme outcomes.

INTRODUCTION

Nonlinear behavior is part of all the phenomena we know in nature, from weather to living beings and beyond. In biological systems, nonlinear mechanisms are partially uncharted and a subject of general scientific interest for several decades.[1,2] In medicine, perhaps the most important research segment is represented by the prediction of different diseases. The accurate prediction of the occurrence of a disease has always been regarded as the main tendency of clinical practice. Genetic diseases, which can be triggered by environmental factors, are inherently nonlinear in nature and their onset is unpredictable. Such a disease, with a non-linear behavior, is diabetes. Neural Networks (NNs) were the hope for prediction in the medical field in the 1980s and this trend was reborn in recent years.[3–5] However, the medical context in which they are used and the partial subjectivity in the training of NNs slowly lose their much-anticipated value for certain specific tasks.[6,7] Essentially, NNs are classifiers of information with prior or "in-flight" adaptation to the environment. Any pattern recognition strategy uses this classification approach. Nevertheless, such a classification based on

prior adaptation to the environment may not imply a prediction. We are inclined to believe that NNs are misused in some cases and the classification process is often confused with the prediction process. For instance, in a previous study, we struggled with a fundamental problem in which we tried to use NNs for the prediction of diabetes onset.[8] There, we agreed that our NN correctly classified a new patient into one of the two classes, namely, type 1 diabetes (T1D) or type 2 diabetes (T2D). However, such a classification was a direct indication of the current state of the disease. In other words, it was merely a medical diagnosis. We then asked ourselves whether the NN can predict the evolution of a human subject over time in the hope that we can predict the onset of the disease. Our experimental data have indicated that our NN classified the set of data in a fair manner but failed to indicate any valuable information about the evolution of a subject over time.[9,10] In this respect, predictions that use Markov chains or classical statistical approaches showed more reliable results in the field of biology and medicine.[11–13] However, here we propose a novel method of analysis, with implementation (see the supplementary material), as an alternative to classical NNs. **Note**: the word *spectral* refers to a series of predicted states arranged linearly between two known states.
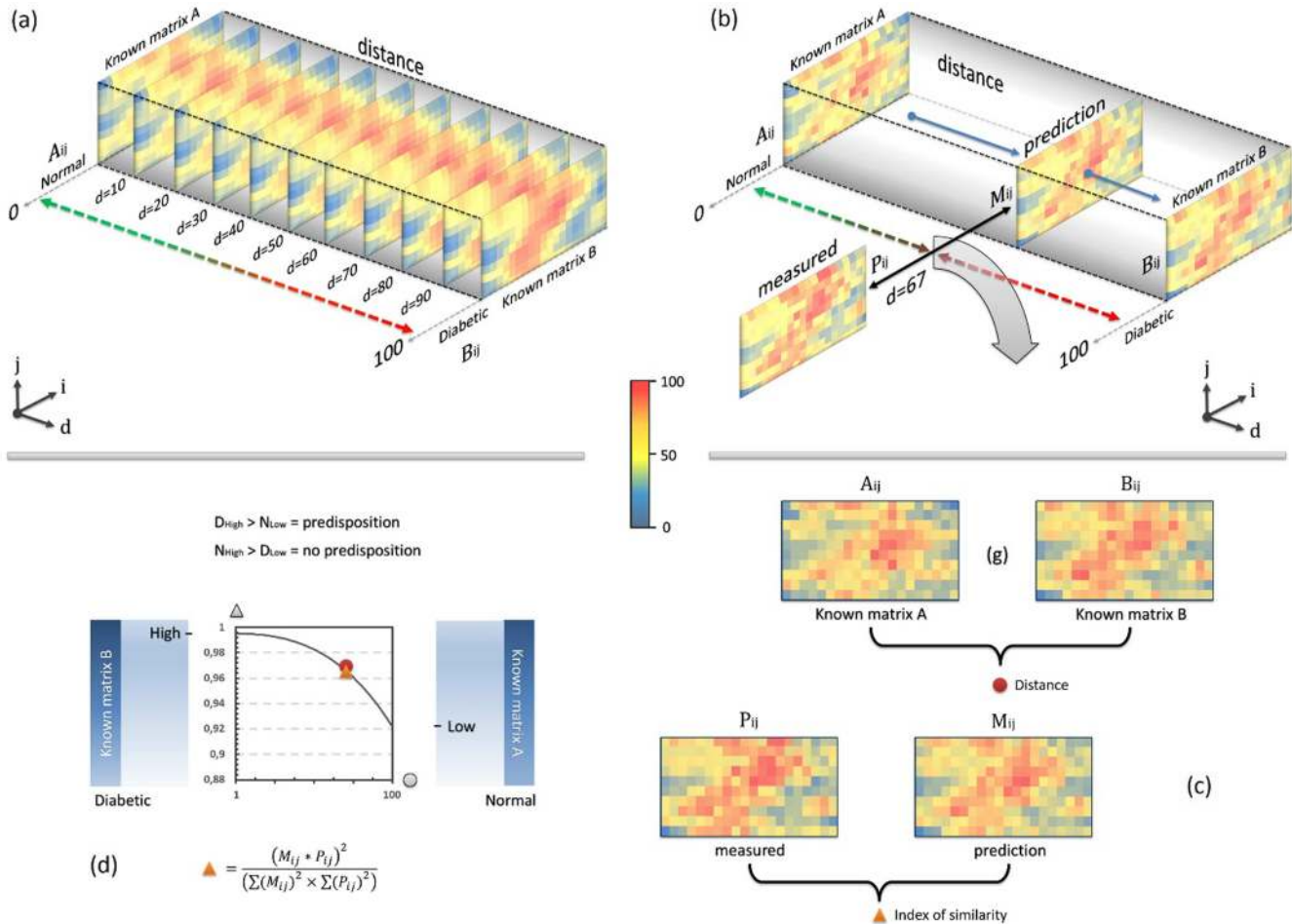

FIG. 1. The spectral forecast model. (a) The panel shows a spectrum of states. In each state, a matrix M is generated according to the data from matrix A, B, and distance d. The sequence of states represents a continuity between an initial state (matrix A) and a final state (matrix B). (b) The 3D approach to the method. The panel shows three known matrices, A, B, and P. Matrix A stores the representative data for one group, namely, the diabetic group, and matrix B stores data from the second group, the normal group. Matrix P contains new measured data that represent the current state of a system—a new subject, which is outside the two groups and whose evolution is of interest. A matrix M is created by using Eq. (1) and then compared with matrix P at each discrete step d by using Eq. (2). The panel shows the process frozen at $d = 67$ for exemplification. (c) It shows a 2D rationale to the method. The two matrices A and B are used to generate the M matrix. Matrix M is compared to matrix P in order to obtain a similarity value between 0 and 1, in which 1 means perfectly similar and 0 means totally dissimilar. (d) A trend example resulting from the prediction process. The value of the similarity index was plotted on the y axis at each step (x axis). The left side of the chart represents matrix B and the right side represents matrix A. The trend was built from small lines that made the connection between the similarity points of each discrete step. High to low or low to high features of the trends signify a predisposition or a protection for the disease. All matrices are shown in the form of heat maps, where dark red represents a maximum value of 100 and dark blue represents the minimum value of zero.

MATERIALS AND METHODS

In order to test our model, we collected and used the data related to the electrical activity signals of the human skin from our most recent experiment.[8-10]

Datasets and context

The electric activity on the skin surface of the trunk was measured by using 200 sensors in three groups: a control group A—18 normal subjects, a group B—18 diabetic subjects, and a test group C—20 normal subjects (ten subjects with confirmed family predisposition for T2D and ten subjects without family predisposition for T2D).[8] The electrical signals were collected using the photon-pixel coupling method and were stored as numerical values ($0 \dots 100$) in a 10×20 matrix for each subject.[8,9] An average was taken along the 18 subjects in group A and group B, yielding a 10×20 matrix for each, namely, matrix A and matrix B.[8] These average matrices represent the main characteristics of each group. A state space was considered between the two matrices of group A and group B. The number of states was established by a distance index (d) and each state in this spectrum was represented by a matrix M. Each subject in group C was then evaluated by a consecutive comparison of their matrix P with each matrix M in the spectrum.

The spectral forecast model

In our approach, we used three known matrices: A, B, and P. A matrix M was further used to formulate the entire spectrum of unknown information between matrix A and B [Fig. 1(a)]. For this calculation, we devised a novel equation shown in (1),

$$M_{ijd} = \left[\left(\frac{d}{\text{Max}(A_{ij})} \right) \times A_{ij} \right] + \left[\left(\frac{(\text{Max}(d) - d)}{\text{Max}(B_{ij})} \right) \times B_{ij} \right], \quad (1)$$

where M_{ij} represents the predicted matrix at every discrete step (d), A_{ij} represents the matrix of the normal group, and B_{ij} is the matrix associated with the diabetic group [Figs. 1(b) and 1(c)]. Also, d stands for distance and represents the total number of discrete steps taken from matrix A to matrix B. Thus, M_{ijd} can be considered a 3D tensor-like structure.

The evolution of P was predicted by a repeated comparison with matrix M at every discrete step (2). This comparison was made by using the similarity index,

$$S(d) = \frac{(M_{ijd} \times P_{ij})^2}{\left(\sum (M_{ijd})^2 \times \sum (P_{ij})^2 \right)}, \quad (2)$$

where S is the similarity index and represents the normalized dot-product of M_{ij} and P_{ij}. M_{ij} stands for the predicted matrix at every discrete step and P_{ij} is the matrix originated from a newly measured
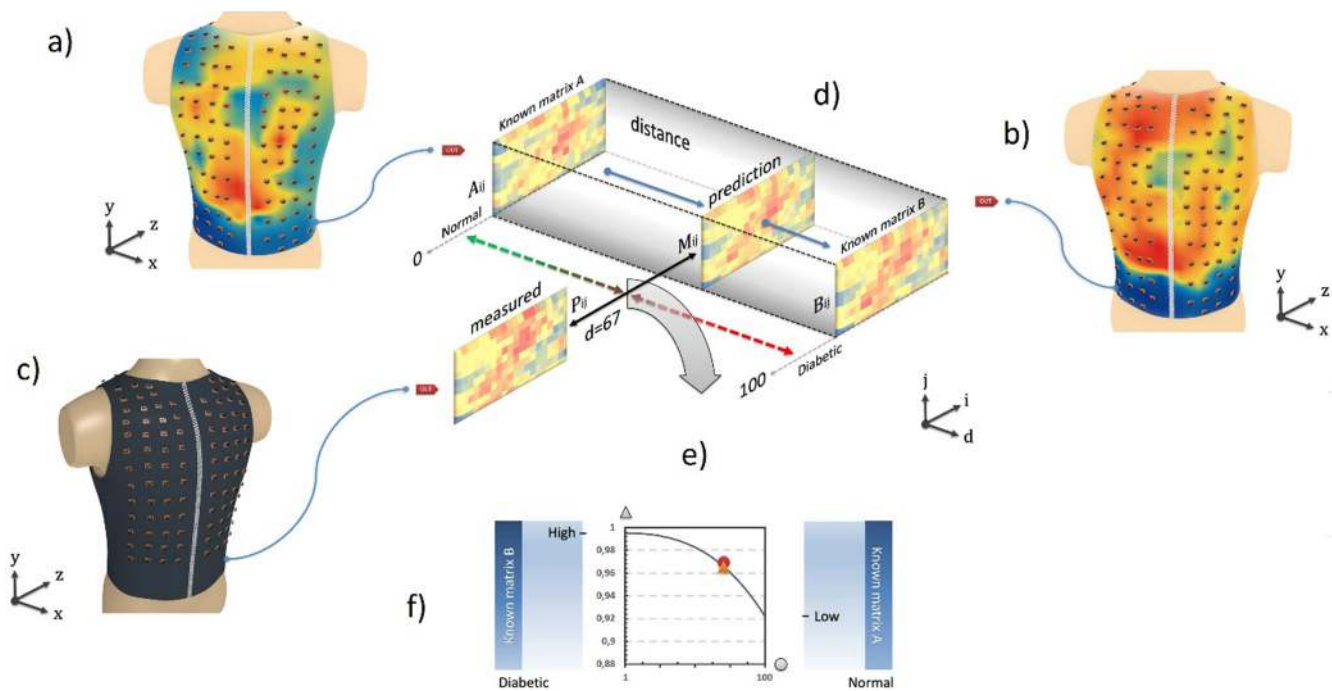

FIG. 2. The origin of the data. (a) The panel shows three known matrices, A, B, and P, and their provenience. (a) Data for the normal group represented by matrix A. (b) Matrix B stores data from the second group, the diabetic group. (c) The new measured data represented by matrix P. It represents the current state of a system; in our case, these measurements were made on a human subject. (d) shows a sweep made between matrix A and matrix B. (e) During the sweep, a matrix M was created at each discrete step d and then compared with matrix P, to generate the values of the similarity index. (f) The values of the similarity index were plotted on a graph to identify the future behavior of the newly measured system.

individual. The similarity index can take values between 0 and 1. As the similarity between the corresponding *i,j* elements of matrix *M* and *P* increases, the similarity index *S* tends to 1. In contrast, as the differences between the values of the corresponding *i,j* elements of matrix *M* and *P* are more frequent, the similarity index *S* tends to 0 [Fig. 1(d)] . The main result of the method is represented by a trend dictated by the values of the similarity index [Fig. 1(d)]. The trend was taken as the evolutionary route of the disease. In the supplementary material, we show a ready-to-use implementation of the method.

Note: The total number of discrete steps was arbitrarily chosen. In this specific case, the maximum value for distance (*d*) was set at 100 for ease. A higher number of discrete steps increased the resolution of the prediction, which was desirable in many situations.

For instance, in some cases, the trend developed both ascending and descending characteristics. At low resolutions (i.e., *d* < 10), many of these fluctuating features remained undetectable and the insight of the results significantly dropped.

RESULTS AND DISCUSSION

Discretization is a practical approach for many prediction algorithms and it is used for almost all computational solutions. Here, we used a discretization strategy to increase the resolution of the spectrum underlying two groups: a healthy group and a diabetic group [Figs. 2(a) and 2(b)]. The data from the healthy group were considered as the initial state (state 0) and the data from the diabetic group were used to formulate the final state (state 100). The
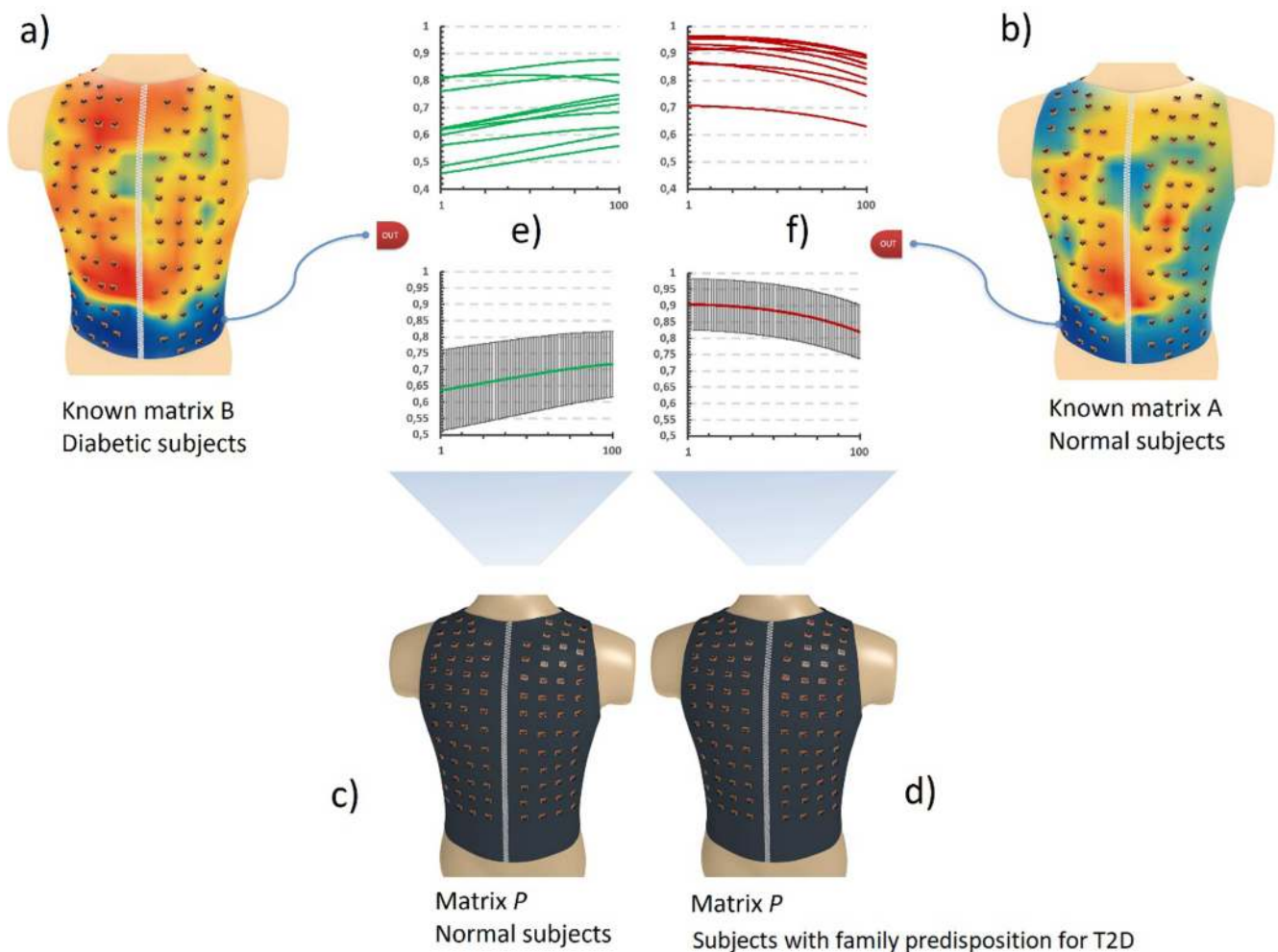

FIG. 3. The spectral forecast prediction—a diabetes case. (a) Average electrical signals from a diabetic group. (b) The average electrical signals from a normal group. (c) Electrical signals from normal subjects. (d) Electrical signals from subjects with family predisposition for T2D. (e) On top—the green lines represent the similarity index values per each normal subject and bottom—average and standard deviation of the normal group. (f) On top—the dark red lines represent the similarity index values per each individual with a predisposition to diabetes and bottom—average and standard deviation of the group with family predisposition for T2D.

number of intermediate states (state 1–state 99) was dictated by distance d, and the intermediate states properties were repeatedly formulated by matrix M [Figs. 2(c) and 2(d)]. To predict the evolution of a third group, a comparison was made along this spectrum [Figs. 2(c) and 2(d)]. Thus, data of matrix P from a new individual were compared to each matrix M in order to obtain the series of values for the similarity index [Figs. 2(e) and 2(f)]. To test the method, we decided to use our most recent data collected from a previous study.[8–10] The predisposition trend for individuals in group C has been correctly predicted 100% of the time (Fig. 3). The subjects with family predisposition for T2D have shown an average similarity index of 0.877 ± 0.024493, whereas normal individuals have shown an average similarity index of 0.68 ± 0.024499.

In the normal group, the mean of the similarity index showed a value of 0.68 ± 0.111 and a maximum value of 0.8784 and a minimum value of 0.4592 [Fig. 3(e)]. In the T2D predisposition group, the mean of the similarity index showed a value of 0.87741 ± 0.08 and a maximum value of 0.9626 and a minimum value of 0.63086 [Fig. 3(f)]. The method has been implemented and can be found in the supplementary material.

The meaning of the trend

Based on known clinical information and the observations made on each individual of the two groups, the trend of the similarity index values indicated whether a newly measured individual showed a predisposition or protection for T2D [Figs. 3(e) and 3(f)]. We speculate that the difference between the lower and the upper limit of the trend may represent a risk score for T2D. At this stage of the investigation, we can indicate if the newly measured subject tends toward the disease [Figs. 3(e) and 3(f)]. Experimentation has shown that the trend shaped by the similarity index does not exhibit only ascending or only descending features. Ascending and descending features of the similarity index can exist within the same plot. One example can be seen in Fig. 3(e), where one individual in the normal group shows both ascending and descending features. In the future, we will try to find the meaning of such a distribution because we speculate that it might be of particular importance for the prediction process.

A link between two unrelated data of the same dimension

The core of our method is represented by Eq. (1), which can have multiple uses on a wide range of values. One of these uses would be a normalization between two unrelated matrices with the same dimension. For instance, elements of matrix A may contain integers between 1 and 2×10^6 million and the elements of matrix B may contain probability values. In this case, Eq. (1) will mix the two matrices based on distance d. In the case of two probability matrices, Eq. (1) performs a normalization in favor of one of the matrices based on distance d. In other words, as matrix M is closer to matrix A, the homologous elements of matrix M will be more similar to matrix A than to those from matrix B. As matrix M will be closer to matrix B, matrix M will be more similar to matrix B. Consequently, if $d = 50$, matrix M will represent a mix equally similar to matrix A and matrix B.

Thoughts for the future

The important cases are those that show a maximum similarity index between the two groups. We suggest that these peak values may be a direct indication of the state of the subject before the onset of the disease. In order to predict the onset, we wish to establish a link between the temporal line of the disease and the states generated along the spectrum. Variations of the method may be constructed and we are eager to use other datasets in the same format (A, B, P). Future uses may include the field of meteorology, medical diagnostics, forensics, economic forecasts, or in the field of genetics for establishing the relationship between species. In biology, we also believe that Eq. (1) can be used for tissue structure prediction based on two groups of histological slides.

CONCLUSION

Here, we have shown the use of a novel prediction model. We proposed a simple method that provides an insight into the evolution of natural processes. To demonstrate the method, our current example considered the predisposition to disease in human subjects based on two known groups. In this approach, we correctly predicted the evolution of new subjects by using our previous data recorded from normal subjects and T2D subjects. Other applications of the method may further indicate the ideal conditions to which our method is appropriate or the limits of precision in the prediction of various metabolic diseases. In the future, we will try to make an association between a temporal line and the steps of the spectrum to indicate the time until the disease is triggered in days, months, or years. A ready-to-use implementation is present in the supplementary material, which can also be used for other types of data.

SUPPLEMENTARY MATERIAL

See the supplementary material for the implemented version of the spectral forecast model.

REFERENCES

[1] F. Mosconi *et al.*, "Some nonlinear challenges in biology," Nonlinearity **21**, T131 (2008).
[2] V. Rai, S. R. Nadar, and R. K. Upadhyay, "Nonlinear phenomena in biology and medicine," Comput. Math. Methods Med. **2012**, 183879.
[3] M. L. Astion and P. Wilding, "The application of backpropagation neural networks to problems in pathology and laboratory medicine," Arch. Pathol. Lab. Med. **116**(10), 995–1001 (1992).
[4] M. W. Kattan and J. R. Beck, "Artificial neural networks for medical classification decisions," Arch. Pathol. Lab. Med. **119**(8), 672–677 (1995).
[5] N. Shahid, T. Rappon, and W. Berta, "Applications of artificial neural networks in health care organizational decision-making: A scoping review," PLoS One **14**(2), e0212356 (2019).

[6] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," J. R. Soc. Interface **15**(141), 20170387 (2018).

[7] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—Beyond the peak of inflated expectations," N. Engl. J. Med. **376**(26), 2507–2509 (2017).

[8] C. Ionescu-Tirgovistea, P. A. Gagniuc, and E. Gagniuc, "The electrical activity map of the human skin indicates strong differences between normal and diabetic individuals: A gateway to onset prevention," Biosens. Bioelectron. **120**, 188–194 (2018).

[9] C. Ionescu-Tirgoviste, P. A. Gagniuc, and E. Gagniuc, "Maps of electrical activity in diabetic patients and normal individuals," Data Brief. **21**, 795–832 (2018).

[10] P. A. Gagniuc, C. Ionescu-Tirgoviste, and E. Gagniuc, "Photon-pixel coupling: A method for parallel acquisition of electrical signals for scientific investigations," Methods X **6**, 968–979 (2019).

[11] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2014), pp. 85–94.

[12] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM, 2018), pp. 43–51.

[13] P. A. Gagniuc, *Markov Chains From Theory to Implementation and Experimentation* (John Wiley & Sons, Hoboken, NJ, 2017), ISBN: 978-1-119-38755-8.