


Brief Communications

Can feature structure improve model's precision? A novel prediction method using artificial image and image identification

Yupeng He , MD, PhD^{*1}, Qiwen Sun, MD, PhD², Masaaki Matsunaga, MD, PhD¹,
Atsuhiko Ota, MD, PhD, MPH¹

¹Department of Public Health, Fujita Health University School of Medicine, Toyoake, Aichi 4701192, Japan, ²Independent scholar, Nagoya, Aichi 4640831, Japan

*Corresponding author: Yupeng He, MD, PhD, Department of Public Health, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan (yupeng.he@fujita-hu.ac.jp)

Abstract

Objectives: This study aimed to develop an approach to enhance the model precision by artificial images.

Materials and Methods: Given an epidemiological study designed to predict 1 response using f features with M samples, each feature was converted into a pixel with certain value. Permuted these pixels into F orders, resulting in F distinct artificial image sample sets. Based on the experience of image recognition techniques, appropriate training images results in higher precision model. In the preliminary experiment, a binary response was predicted by 76 features, the sample set included 223 patients and 1776 healthy controls.

Results: We randomly selected 10 000 artificial sample sets to train the model. Models' performance (area under the receiver operating characteristic curve values) depicted a bell-shaped distribution.

Conclusion: The model construction strategy developed in the research has potential to capture feature order related information and enhance model predictability.

Lay Summary

We aimed to demonstrate a novel method to investigate the effect of feature structure on model predictability with epidemiological data. The concept was inspired from image identification. Pixels in digital images are used as features when training the identification model. The quality of a given digital image will be damaged when pixels' position and their values changed arbitrarily, which obstructs the model training and model's precision. We assume the structure-related relationship exists in epidemiological data. Given a certain dataset, features are transformed to pixel values for generating artificial images. To explore the effect of feature structure, orders of pixels are randomly permuted and the model is trained using pixel-permuted artificial image sample sets. In the preliminary experiment, one binary response was designed to be predicted by 76 features. We randomly selected 10 000 artificial image sample sets to train the model. Models' performance (area under the receiver operating characteristic curve values) depicted a bell-shaped distribution. Namely, the performance of each model's predictability was studied and the feature structure information had a strong impact on model performance. Our novel model construction strategy has potential to capture feature order related information and enhance model predictability.

Key words: artificial image; image identification; prediction model; machine learning; neural network.

Introduction

Linear models are considered as cornerstone in epidemiological studies. They are widely used for estimating associations between factors and disease,¹ and for predicting the incidence or existence of disease.² These models encompass a range of techniques within generalized linear regression, such as linear regression, logistic regression, and Cox regression. Previous studies reported the limitations in linear models such as explaining nonlinear association, complexed interactions, etc. Nonlinear models were adopted to solve those problems.^{3,4} For example, nonlinear models represented by artificial neural networks have been introduced into epidemiological research.^{5,6}

To enhance the model accuracy, previous research implemented the following solutions: (1) Increasing the number of

features and expanding the sample size. (2) Optimizing the parameters in an attempt to achieve maximum accuracy.⁷ (3) Comparing various machine learning methods to identify the one that yields the highest performance. For instance, employing nonlinear models (eg, neural networks, support vector machines, etc.) to address the weaknesses in linear equations caused by multicollinearity among features,^{4,8} unsophisticated variable selection,⁹ and to extract more complex information.

Inspired from image recognition techniques, good feature structure is the key point for model training. Taking handwritten digit recognition as an example (Figures S1 and S2 in Appendix 1), the digital picture is composed of pixels (Figure S1-a), where the pixel values range from 0 to 255 (Figure S1-b),

Received: October 30, 2023; Revised: January 3, 2024; Editorial Decision: January 25, 2024; Accepted: February 1, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

corresponding to the color change from black to white (Figure S1-a). Each pixel is used as a feature for image recognition model training (Figure S1-b). If the order of the pixels is changed arbitrarily (Figure S1-c), handwritten digits cannot be correctly recognized.

This research investigated the effect of feature structure on model predictability with epidemiological data, which was less explored in this field. To introduce the feature structure, we generated artificial images with each pixel representing a feature. To explore the effect of feature structure, we randomly permuted the order of pixels and used pixel-permuted image sample sets to train the model. The model used in this study is neural networks which is an artificial intelligence-based computer analysis, designed to extract “certain information” from digitalized images. The performance of each model predictability was studied and the result showed the feature structure information had a strong impact on model performance.

On the other hand, previous studies on neural networks have demonstrated the exceptional accuracy achieved through image identification based on deep-learning techniques, often surpassing 98%.^{10,11} Therefore, the method discussed in this research has potential to develop a novel method for achieving higher precision predictions using artificial images and image identification technology.

Methods

Generating the artificial image

The process of generating artificial images involved 2 key aspects: variable pixelization and pixel sequencing.

Variable pixelization

In a grayscale image, pixels are represented conventionally by an 8-bit integer giving a range of possible values from 0 to 255. This is how an image stored in computer. While data collected from an epidemiology study is not conventionally arranged. Hence, to apply image recognition techniques, variables must be rearranged between 0 and 255, as the way of pixel is represented.

For each feature in a given epidemiological study, we applied the rescale function \mathbb{P} to normalize the feature’s value within the range of 0 to 255. This transformation ensured that each feature could be accurately represented in the artificial images.

$$\mathbb{P}(X_n) = \left\lfloor \left(\frac{X_n - \min(X_n)}{\max(X_n) - \min(X_n)} \right) \times 255 + 0.5 \right\rfloor,$$

$$n = 1, \dots, f$$

where X_n represents the value of feature n of a certain sample, $\min(X_n)$ corresponds to the minimum of X_n within the sample set, and $\max(X_n)$ corresponds the maximum of X_n within the sample set. f is the number of features. For instance, in an epidemiological study that included 5 samples; one of the features was their age: 20, 30, 40, 50, and 60 years. The age values were rescaled to 0, 64, 128, 191, and 255 gray-level. This pixelization process was crucial for ensuring that the artificial images accurately represented the underlying features.

Pixel sequencing

Given a certain epidemiology study with f features transformed to f pixels, there exist $f!$ possible pixel orders. To simplify our study design, we opted to organize pixels into a square array, without considering rotation (90° , 180° , or 270°) or flips (vertical, horizontal, or diagonal) of images. Consequently, the possibility of images being given f features is F , where F is equal to $1/8 f!$. Figure 1 shows an example of an artificial image of a certain sample with a certain pixel order.

Dataset expansion

To study the effect of feature structure on model predictability, we generated artificial images for each sample by using variable pixelization and randomly permuted the order of pixels to create distinct sample sets.

The original dataset, denoted as S_{original} , was obtained from an epidemiological study encompassing M samples and f features with a certain feature order. As previously described, f features can generate F different orders. Hence, the S_{original} dataset was expanded to F types of distinct datasets, represented by S_1, S_2, \dots, S_F (Figure 2). These expanded datasets, referred to as “candidate datasets”.

Data processing

Each of the F candidate datasets underwent an identical data processing procedure. Initially, we randomly divided each candidate dataset into training, validation, and test sets in a 70:10:20 ratio. To ensure balanced training, we applied the Synthetic Minority Oversampling Technique to the training set.¹² The model was trained using an image identification technique on the training and validation sets. Specifically, the model training process took place within the training set,

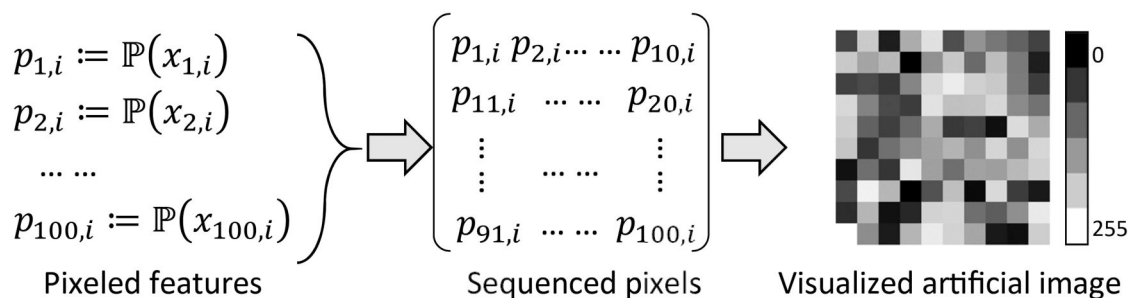


Figure 1. Given the feature information of a certain sample i , generating an artificial image by sequencing pixels in a square array—assuming the number of features is 100. $p_{n,i}$, $n = 1, \dots, 100$ represents the pixel value of feature n for sample i calculated by function \mathbb{P} .

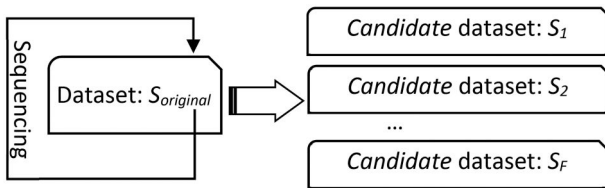


Figure 2. Dataset expansion. $S_{original}$: the original sample set with M samples and f features in a certain order. S_n , $n = 1, 2, \dots, F$: generated samples sets with distinct feature order.

with continuous validation to assess learning effectiveness. The training concluded once the loss function's value, evaluated on the validation set, ceased to increase, which indicated optimal model development. The model was then applied to the test set to assess its performance, and the results were recorded (Figure S2 in Appendix 1).

The model that achieved the highest performance was deemed the optimal prediction model. Consequently, this *candidate* dataset containing artificial images generated from a certain feature order, was considered the optimal dataset for model construction. These artificial images effectively captured the intricate relationship between the features and the response.

Preliminary experiment

In this section, we describe the process of model construction with the method introduced in previous sections. A preliminary experiment was carried out to explore the effectiveness of the method.

The model, namely the schizophrenia classifier was trained by using a sample set from online survey data collected by an internet research agency's pooled panel (Rakuten Insight, Inc., incorporated ~ 2.3 million panelists by 2022).¹³ The sample set comprised 223 patients with schizophrenia and 1776 healthy controls aged 20-75 years. For each sample, the following information was extracted from the survey: the existence of schizophrenia which corresponds to the response and 76 features, including demographic, health-related backgrounds, physical comorbidities, psychiatric comorbidities, and social comorbidities. The details of the study participants and variable definitions have been published elsewhere (Appendix 1).¹⁴

The models were trained using artificial neural network. We conducted another study using this machine learning technique based on the same dataset; therefore, we applied the same model structure to the current experiment.¹⁵ The model was structured with 5 hidden layers (neurons per each layer: 128-64-32-16-8), HeNormal weight initializer, ReLU activation function in the hidden layers, sigmoid activation in the output layer, a learning rate of 0.01, and early stopping when 5 consecutive updates were < 0.001 .¹⁵ Model performance was assessed using the area under the receiver operating characteristic curve (AUC).¹⁵ The following AUC thresholds were used to categorize model discrimination quality: 0.5 = no discrimination; 0.5-0.7 = poor discrimination; 0.7-0.8 = acceptable discrimination; 0.8-0.9 = excellent discrimination; and > 0.9 = outstanding discrimination. Given the extensive permutations of feature orders (namely, $1/8 \times 76!$), we experimented with 10 000 randomly selected *candidate* datasets to explore the impact of different feature orders on model performance. Statistical analyses were performed using Python 3.8 (Python Software Foundation, <http://www.python.org>), with the Jupyter Notebook (Jupyter, <http://www.jupyter.org/>) serving as the computational environment.

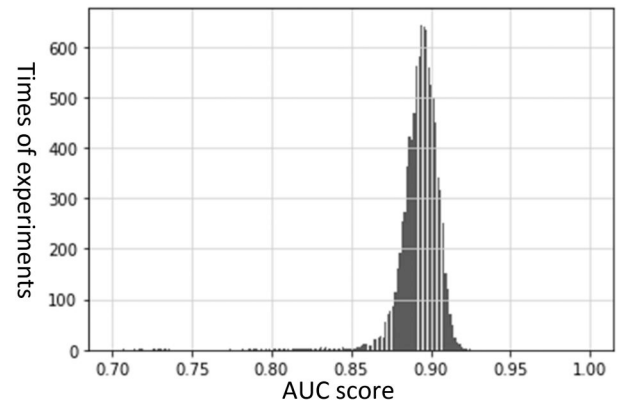


Figure 3. The distribution of AUC scores from the 10 000 experiments. Abbreviation: AUC, the area under the receiver operating characteristic curve.

Results

Based on our preliminary experiment, Figure 3 illustrates the distribution of AUC scores across 10 000 experiments. The majority of models yielded an AUC score of approximately 0.88, indicating excellent discrimination. However, some models achieved AUC scores between 0.5 and 0.7, and even fewer attained scores exceeding 0.93, demonstrating outstanding discrimination (Figure 3).

Discussion

This study introduces the novel concept of an artificial image, representing a departure from traditional epidemiological research methods. The novelty of this approach lies in its transformation of features into pixels to reconstruct images in reverse, enabling the application of techniques from diverse fields to classical epidemiological research.

In our preliminary experiment, we observed that the accuracy of the trained models varied depending on the positions of the features within the artificial image. Models exhibiting high accuracy correspond to a small number of datasets. This finding indirectly supports our hypothesis regarding the existence of optimal artificial images.

The core of our method revolves around increasing the number of dimensions. By doing so, our approach becomes a powerful tool for exploring and explaining complex non-linear information. We hypothesize that the linear equation structure commonly used in generalized linear models may lead to the loss of essential information among features.^{16,17} Moreover, linear equations often describe the relationship between the feature and the response as a monotonic increase or decrease which may oversimplify the intricate nature of the data. The construction of artificial images may provide a new perspective for model interpretation. For instance, we can potentially explain feature importance and feature interactions through their spatial locations within the artificial image: Features situated centrally may be considered more "important" than those on the periphery, and adjacent features may indicate closer interactions (Figure S3 in Appendix 1).

Despite the strengths of our innovative method, several challenges must be addressed for full implementation. First, the method demands significant computational power, which could potentially be mitigated with advancements in quantum computing. Second, in the preliminary experiment, the dataset lacked an adequate number of features. Therefore, it is

insufficient to rely on more mature image-recognition technologies such as convolutional neural networks for model training. We anticipate the introduction of at least 400 features in future experiments. Third, the relationship between features and response cannot be overlooked. Even the most advanced methods may struggle to construct a high-precision prediction model when faced with either no relation or a weak relation between the features and the response. Fourth, there is another limitation during the procedure of data splitting. In this work, data was randomly split to training, validation, and test set at the 70:10:20 ratio. This might cause bias in the results, especially when the response following a skewed distribution. These are important considerations for the continued development and application of our approach.

Conclusion

In this study, we introduced a novel concept and provided evidence of its potential for developing a novel predictive method using artificial images and image identification. The model construction strategy has potential to capture feature order related information and enhance model predictability.

Acknowledgments

The concept of the study has been submitted to Japan Patent Office for applying a patent.

Author contributions

Conceptualization and design: Y.H.; Data acquisition: A.O., M.M.; Data analysis and interpretation: Y.H.; Drafting of the manuscript: Y.H.; Theoretical support: Q.S.; Revising of the manuscript: Y.H., Q.S., M.M., A.O.

Supplementary material

[Supplementary material](#) is available at *JAMIA Open* online.

Funding

This work was supported by Fujita Health University (grant number 2126 to Y.H.), Japan Society for the Promotion of Science (grant number 22K21186 to Y.H.), and Ministry of Health, Labour and Welfare, Japan (grant number JPMH21GC1018 to A.O.).

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Bender R. Introduction to the use of regression models in epidemiology. In: Verma M, ed. *Cancer Epidemiology*. Vol. 471. Humana Press; 2009:179-195.
2. Lunt M. Introduction to statistical modelling: linear regression. *Rheumatology (Oxford)*. 2015;54(7):1137-1140.
3. Spuck N, Schmid M, Berger M. Detection of Nonlinearity, Discontinuity and Interactions in Generalized Regression Models. Accessed January 1, 2024. <https://arxiv.org/pdf/2310.20409.pdf>.
4. Masegosa AR, Cabañas R, Langseth H, Nielsen TD, Salmerón A. Probabilistic models with deep neural networks. *Entropy (Basel)*. 2021;23(1):117.
5. Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. *NPJ Digit Med*. 2020;3:10.
6. Mendez KM, Broadhurst DI, Reinke SN. The application of artificial neural networks in metabolomics: a historical perspective. *Metabolomics*. 2019;15(11):142.
7. Saitoh K, ed. Optimizing the hyperparameters. In: *Deep Learning Created from Scratch*. Tokyo: O'Reilly Japan; 2020:197-203.
8. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *Int J Epidemiol*. 2016;45(2):565-575.
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157-1182.
10. Rizvi M. Learn Image Classification on 3 Datasets Using Convolutional Neural Networks. 2020. Accessed January 1, 2024. <https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/>.
11. Kuo J. Cactus Image Classification Using Convolutional Neural Network that Reaches Over 98% Accuracy. 2020. Accessed January 1, 2024. <https://towardsdatascience.com/cactus-image-classification-using-convolutional-neural-network-cnn-that-reaches-98-accuracy-8432e068f1ea>.
12. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic minority over-sampling technique. *JAIR*. 2002;16:321-357.
13. Rakuten Insight, Inc. Accessed January 1 2024. <https://insight.rakuten.co.jp>.
14. Matsunaga M, Li Y, He Y, et al. Physical, psychiatric, and social comorbidities of individuals with schizophrenia living in the community in Japan. *Int J Environ Res Public Health*. 2023; 20(5):4336.
15. He Y, Matsunaga M, Li Y, et al. Classifying schizophrenia cases by artificial neural network using Japanese web-based survey data: case-control study. *JMIR Form Res*. 2023;7:e50193.
16. Huzaira K. Modeling Non-Linear Dynamic Systems with Neural Networks. 2020. Accessed January 1, 2024. <https://towardsdatascience.com/modeling-non-linear-dynamic-systems-with-neural-networks-f3761bc92649>.
17. Mathia K. Solutions of Linear Equations and a Class of Nonlinear Equations Using Recurrent Neural Networks. 1996. Dissertations and Theses. Paper 1355.