Int. j. inf. tecnol. (March 2024) 16(3):1279–1292
https://doi.org/10.1007/s41870-023-01635-7

ORIGINAL RESEARCH

Integrating prior knowledge to build transformer models

Pei Jiang[1] · **Takashi Obi**[2] · **Yoshikazu Nakajima**[1]

Received: 17 May 2023 / Accepted: 14 November 2023 / Published online: 2 January 2024

Abstract The big Artificial General Intelligence models inspire hot topics currently. The black box problems of Artificial Intelligence (AI) models still exist and need to be solved urgently, especially in the medical area. Therefore, transparent and reliable AI models with small data are also urgently necessary. To build a trustable AI model with small data, we proposed a prior knowledge-integrated transformer model. We first acquired prior knowledge using Shapley Additive exPlanations from various pre-trained machine learning models. Then, we used the prior knowledge to construct the transformer models and compared our proposed models with the Feature Tokenization Transformer model and other classification models. We tested our proposed model on three open datasets and one non-open public dataset in Japan to confirm the feasibility of our proposed methodology. Our results certified that knowledge-integrated transformer models perform better (1%) than general transformer models. Meanwhile, our proposed methodology identified that the self-attention of factors in our proposed transformer models is nearly the same, which needs to be explored in future work. Moreover, our research inspires future endeavors in exploring transparent small AI models.

Keywords AI · Knowledge · SHAP · Transformer · Reliable AI

✉ Takashi Obi
obi.t.aa@m.titech.ac.jp

[1] Institute of Biomaterials and Bioengineering, Tokyo Medical and Dental University, Kanda-Surugadai, Tokyo, Japan

[2] Institute of Innovative Research, Tokyo Institute of Technology, R2-60, Nagatsutachou 4259, Midori Ward, Yokohama, Kanagawa, Japan

1 Introduction

Artificial Intelligence (AI) technology has been used in various fields in our current society [1–5]. AI technology makes an innovative society possible and changes our lifestyles. For example, automatic car driving [6–9], face recognition systems [10–13], and computer aid detection in the medical area [14–17]. However, AI models are generally based on large data and huge parameters, called big AI models, especially in the computer version (diffusion model [18]) and the field of natural language processing. The robust Large Language Model (LLM): Generative Pre-trained Transformer (GPT) models [19] make our daily work more convenient and will even change our work life in the future. The GPT models have been used in various fields [20]. The transfer-based various models [20–27] indicate the possibility of Artificial General Intelligence (AGI) models. However, even with current AI technology, a prominent data-based AI model is impossible in some research fields. For example, in the medical area and biomedical, big data are not always available other than big AI models. Researcher Andrew Wu states the importance of "big AI in small data" [28] and also certificated the necessity of efficient AI models for small datasets. Moreover, a few million parameters in big AI models also cost colossal energy. Research about the energy saved by small AI models is urgently necessary. Therefore, we proposed to build AI models based on prior knowledge.

Besides the big AI models and huge parameter problems, some other limitations still exist in AI research. The black box problem is one of the most pressing issues in AI studies [29–33]. The black box problems lower the reliability of AI models. Meanwhile, current AI models are statistical-analysis-based models, not logic-theory-based models. This keeps the uncertainty of current AI models,

even though the big AI models are efficient. Therefore, understanding the AI models becomes necessary.

To clarify the AI models, Explainable AI (XAI) [34–38] has become one highlight topic in the AI research field. Currently, two kinds of XAI models exist: intrinsic (rule-based) and post hoc models [39]. The intrinsic models explain models by restricting the rules of machine learning models, e.g., linear regression, logistic analysis, and Grad-CAM [40]. In contrast, post hoc models interpret models after training, such as Local interpretable model-agnostic explanations (LIME) [41, 42] and Shapley Additive exPlanations (SHAP) [43]. The SHAP method is the most robust agent explanation model currently. SHAP method has been used in many fields [44–53] and was certificated robust [54–56]. The SHAP methods allow us to interpret the black box models and know the local and global reasons for one prediction or classification. There are also two kinds of SHAP methods: model agnostic (Kernel SHAP) and model specific (Tree SHAP, deep SHAP) [43, 57]. The model-specific SHAP methods are designed to explain the specific models to decrease the calculation or loss of the complex models. They can only be used for a particular situation. In contrast, the kernel SHAP can be used for any model type. However, the SHAP method is a causal-inference-based methodology. The logic among AI models still needs to be clarified. The SHAP methodology just increased the transparency of AI models in some aspects. Research on AI reliability and transparency is still urgently necessary. Are there also ways to explain AI models by instructing the rules of models? This still needs to be explored.

Even though the SHAP method explained AI models in some aspects, it already supplied some knowledge about AI models to us humans. Research by Feifei Li [58] certifies that human interaction will improve the performance of AI models, while the latest GPT4 models [19] also certify the necessary human insertion in large AI models. These situations show that human-knowledge-integrated AI models are one available research direction in AI studies. Currently, reinforcement models [59] give rewards in decision-making while knowledge distillation [60] models filter the knowledge (weights in layers) in AI models. Is there another efficient way to use knowledge in AI models? Can we make human knowledge-integrated AI models possible? Furthermore, how can we integrate knowledge into AI models efficiently? Our research makes one significant step to answering these questions. In this study, we proposed knowledge-integrated AI transformer models to improve the trust and efficiency of AI models. The main contribution of our study was summarized as follows:

- Prior knowledge-integrated transformer AI models were proposed in our study.
- Our proposed methodology paves the way to improve the transparency and reliability of AI models.
- Our study is one significant technical try for researching small and trustable AI models.
- Our proposed methodology certified the possibility of building knowledge-integrated neural network models.
- Our research helps us understand the logic of attention models.

The rest of this paper is organized as follows. We make a small literature review in Sect. 2. Our proposed methodology is introduced in Sect. 3. Section 4 describes the used datasets. We show the detailed results of our study in Sect. 5. Then, we discuss our effects in Sect. 6. Finally, we made one conclusion and discussed our future research direction in Sect. 7.

2 Literature review

2.1 Literature about prior knowledge

Some studies focused on building logic-based, trustable, explainable AI models [61–64]. Besides XAI to explore and explain the AI models to improve the reliability of AI models, some other studies try to build trustable AI models. Philip Slingerland et al. proposed adapting proposed trustable AI models to space mission autonomy [65], while Robin Cohen, Etc. [66] Sketched ways in which trust modeling may be leveraged towards trustable AI. Based on our current knowledge, few studies propose building knowledge-integrated AI models as to how to build trustable AI models. Meanwhile, some researchers state that AI models with human inserting can perform better [58].

Yann LeCun [67] proposed a word model that states we can build models like human learning progress. Our humans use our knowledge to make decisions and solve problems. Can the AI model also integrate knowledge to build more reliable models? Especially, do the AI models combine human knowledge to optimize themselves? Integrating knowledge to build AI models becomes one new research topic. However, there are few researchers focused on building knowledge-integrated AI models. Meanwhile, what is human knowledge, and how can human knowledge be integrated into AI models? There is no standardization currently. Therefore, we proposed using prior knowledge to build models. However, what can be treated as prior knowledge? While some studies use the pre-trained models as prior knowledge, we proposed using the XAI results to build models, especially building AI models based on small datasets, when most research focused on big data-based big AI models [19].

2.2 Literature about transformer models

At present, the transformer models [68, 69], which are the base model of generative AI models, become one highlight topic in AI. The attention model [70] is the primary structure of the transformer model. Using attention, we can check the connections among factors, like the research using attention to predict the connection among language tokens [23]. Even though the attention of the transformer model is also based on the Neural Network (NN) models, the attention models can help us understand the AI models in some aspects. The attention models in the LLM model can show the relationship among tokens. Especially after the attention model was used in the computer vision field, the vision transformer models can explain the images to let us know which areas are important [71]. Therefore, we also integrated prior knowledge to build transformer models for tabular data and compared our results with another tabular data transformer model: the Feature Tokenization Transformer (FTT) [72] model. Using self-attention models, we aim to clarify the relationship among the input features. Therefore, we can understand the AI models in some aspects.

3 Methodology

In this study, we proposed one knowledge-integrated self-attention transformer model. Unlike the attention mechanism using various methods to adjust the NN model weights, we proposed using ensemble SHAP values as knowledge to build transformer models. We first proposed ensemble SHAP value calculation methods to acquire more reliable knowledge. Then, we used the prior knowledge as the input of self-attention transformer models. The whole methodology structure is shown in Fig. 1.

Currently, the SHAP methodology is one of the most robust XAI methods and can be used to explain various models. Research also certificated the efficiency and robustness of SHAP methods [52, 54–56]. Because the SHAP value was calculated based on the casual inference theory, the SHAP value will be changed in different models. To balance the effect caused by various models, we proposed an ensemble SHAP value, which will consider all models' accuracy and kernel SHAP values. Therefore, we use the ensemble SHAP values as knowledge to build our models, not the hybrid SHAP value. The details are introduced in the following subsection.

3.1 Ensemble XAI to acquire knowledge

SHAP predicts an instance x by computing each feature's value's contribution to the prediction of one model. The SHAP explanation method computes Shapley's values from coalitional game theory. The feature values x of a data instance

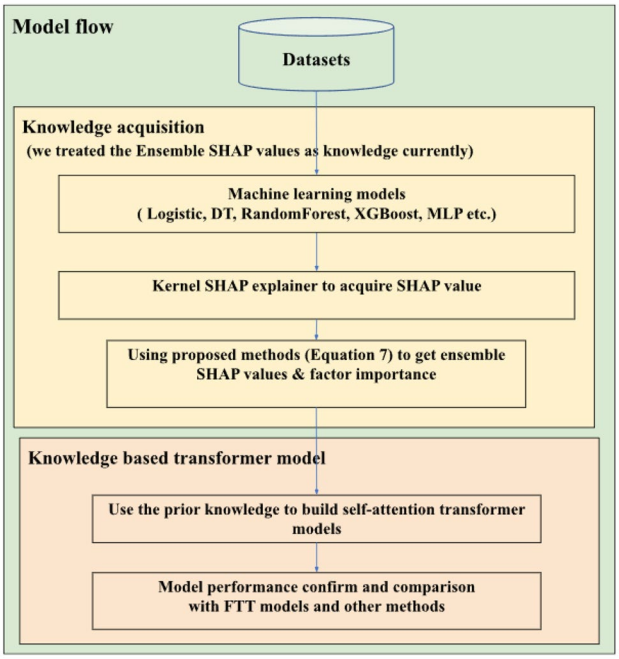

Fig. 1 The proposed methodology flowchart

act as players in a coalition. Shapley values tell us how to distribute the prediction among the features fairly. SHAP by appropriate the original model function to new function $f(x) = g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$. Where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$, which is treated as local factor importance. The z' represents the dataset of x, and M has the same feature space. In kernel SHAP, the $g(z')$ is linear model. The explanation of x is

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \pi_x(z')\left[f_x(z') - f_x(z' \setminus i)\right] \tag{1}$$

where the $f_x(z')$ is the function when the z_i' is 1, while the $f_x(z' \setminus i)$ is the original function when the z_i' is zero. The kernel of π_x is $\pi_x(Z') = \frac{(M-1)}{\binom{M}{|Z'|}|Z'|(M-|Z'|)}$ where the $|z'|$ is the number of nonzero entries in z' and $z' \subseteq x$ represents all z' vectors where the nonzero entries are a subset of the entries in x.

In kernel SHAP $\phi_0 = f(h_x(0))$ is set as 0, and the loss function of kernel SHAP becomes

$$L\left(\hat{f}, g, \pi_x\right) = \sum_{z' \in Z} \left[\hat{f}\left(h_x(z')\right) - g(z')\right]^2 \pi_x(z') \tag{2}$$

Kernel SHAP estimated the contribution of instance x by appropriate the x using a linear model and treating the weight of linear models as the local factor contribution ϕ_i. Although the kernel SHAP can help us understand the factor contribution in each model, the value of the factor ranking in each model is different. Because the kernel SHAP is based on the

appropriate calculation theory, the kernel SHAP values of different methods will differ [73]. When the SHAP method approximates a linear model, it just uses the predicted output of one model, which will affect the predicted outcome by how good the prediction model is. Meanwhile, because the model-agnostic explanation method only approximates the predicted outputs of models, the kernel SHAP values for all models have the same metric when we analyze one dataset. Therefore, our proposed ensemble SHAP method is available. Moreover, the goodness of one model also should be considered when calculating the factor's importance. Therefore, we also used the precision of the models to adjust the ranking of the factors. If one model has higher accuracy, it will be more critical in ensemble SHAP value calculation. The calculation is shown as Algorithm 1, Where the Acc_j is the accuracy of one classification or regression model, The N is the number of analytical approaches for one dataset, and the I_j is the factor of importance ranking in one analysis. Therefore, the single kernel SHAP value cannot stand the fundamental rank of factor importance. We proposed using ensemble SHAP values, shown in Algorithm 1. Even though we used local ensemble SHAP as input to build our proposed self-attention transformer model, we also checked the global ensemble SHAP value to confirm that our proposed ensemble SHAP method is efficient in our used datasets, which can be calculated as follows:

$$I = \sum_{j=1}^{N-1} W_j I_j = \sum_{j=1}^{N-1} \frac{\exp(Acc_j)}{\sum_{j=1}^{N-1} \exp(Acc_j)} \sum_{i=1}^{M} \phi_i \tag{3}$$

Although our previous study already certificated the efficiency of the ensemble SHAP methods [73].

Algorithm 1 Knowledge acquisition

i=0, j=0, N is all the number of analysis models,
for N analysis models, select N-1 analysis model **do**
  for $j <= N-1$: **do**
    $W_j = \frac{\exp(Acc_j)}{\sum_{i=1}^{N-1} \exp(Acc_j)}$ where Acc_j is the accuracy of jth analysis in the N-1 models
    for $i <= M$: **do**
      $\phi_j = \sum_{i=1}^{M} \phi_{ij}$
    end for
    $\phi_j = W_j * \phi_j$
    (# For global factor importance $I_j = W_j * I_j$)
    $\phi+ = \phi_j$
    (# For global factor importance $I+ = I_j$)
  end for
  $\phi = \frac{\phi}{N}$
  (# For global factor importance $I = \frac{I}{N}$)
end for
Final global factor importance is I

3.2 Prior knowledge to build transformer model

Our proposed whole methodology flow is shown in Fig. 1. We treated the proposed ensemble SHAP values as prior knowledge. Firstly, nine general and robust machine learning classification models: logistic analysis, Navie Bayside classification, quantitative discriminate analysis,k-nearest neighbors classification, AdaBoost, general Decision Tree, random forest classification, XGBoost, and Multi-Layer Perception classification, were used to make a classification in three classification-task datasets. Then, for one non-open dataset, the kernel SHAP was used to explain each classification model and got the contribution (local SHAP value) of factors for each model, while the importance ranking of factors was also reviewed. After we got the kernel SHAP value of factors, we used our proposed ensemble methodology to calculate the importance of the factors. Finally, we used the ensemble SHAP value as prior knowledge to build the self-attention transformer models. We compared our proposed knowledge-integrated self-attention transformer model with the FTT and other machine learning or NN models. Moreover, we also checked the self-attention of each transformer block in the FTT model and our proposed model. To confirm the efficiency of our models, we also tested various layers of self-attention transformer models in this study: 2 layers, 4 layers, 8 layers, and 12 layers. Reviewing the self-attention of each layer, we can understand the difference between the FTT model and our proposed model. Moreover, attention to transformer models also can help us understand the running rules of AI models. After checking the difference among various layer-deep transformer models, we also compared the average self-attention of our proposed transformer models with FTT models (Fig. 7) and the general coefficient among input factors (Fig. 8). The apparent difference is also shown in the results section and discussed.

4 Data source

To testify to the efficiency of our proposed models, three open data sets and one non-open data set were used to test our proposed methodology. The three open data sets are for

Table 1 Used datasets in this analysis

Datasets introduction					
	Type	Tasks	Samples	Pre-train	Test
PIDD	Open	Classify	768	614	614 & 154
Diabetes	Open	Classify	1000	800	800 & 200
Heart disease	Open	Classify	898	718	718 & 180
MHLW	Non-open	Classify	12,736	2548	10,188 & 2548

Table 2 Model performance comparison of classification models for the classification task datasets in our study

Methods	PIDD (%)	Diabetes (%)	Heart disease (%)	MHLW (%)
Logistic	77.92	94.00	75.38	60.36
K-nearest neighbors	72.73	92.00	64.58	51.53
Decision tree	67.53	97.00	65.37	55.38
Random forest	83.12	98.50	73.64	59.38
MLP	65.58	96.50	75.56	56.20
AdaBoost	79.22	98.50	75.63	61.22
Naive Bayes	77.27	90.50	70.19	58.36
QDA	74.03	93.50	70.44	58.36
XGBoost	75.32	98.50	76.02	58.59
FTT	64.65	89.75	49.85	50.87
Proposed	**65.30**	**89.75**	**50.94**	48.94

The bold results show the better performance of our proposed model than the FTT model

classification among the used open data sets. Pima Indians Diabetes Database (PIDD) [74], Mendeley open diabetes data set [75], and the heart disease dataset [76]. All open datasets can be downloaded from the Internet. The PIDD is a small diabetes dataset containing 768 diabetes samples and eight factors of diabetes: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pre-degree function, and age. Similarly, the US open diabetes dataset contains 11 risk factors for diabetes. BMI, HbA1c, age, etc., while the heart disease data sets have 17 factors. Moreover, The proposed method was also used to analyze the Ministry of Healthcare, Labor, and Welfare (MHLW) [77] census data. The MHLW dataset is non-objective-oriented; we used the newest MHLW (2018) data and deleted the null value samples. Finally, after pre-processing the datasets, 12,736 balanced samples were used to test our proposed methodology.

In our proposed methodology, samples of the datasets are divided into two parts: one part of the data was used to acquire prior knowledge, and the other part was used to train our proposed methodology. As shown in Table 1, for treating SHAP values as input models, we used 80 percentage data to obtain ensemble SHAP values and treated the ensemble SHAP values as a new input to self-attention transformer models and compared their performance with FTT and other machine learning models [logistic, K-nearest neighbors, decision tree, Multi-Layer Perception (MLP), AdaBoost, Naive Bayes classification, Quantum Discriminate Analysis (QDA) and XGBoost]. We used kernel SHAP to explain various classification models separately, and the factors' importance ranking to each model was reviewed. Then, we used the proposed ensemble SHAP value to build self-attention transformer models. Finally, we checked the performance of our proposed models. Details of the results are shown in the Results section.

5 Results

In this study, we proposed using ensemble SHAP value as knowledge to build self-attention transformer models. Then, we checked our proposed transformer models' performance and self-attention. We also compared the self-attention of our proposed models with FTT models and the general factor coefficients to confirm the efficiency of our proposed transformer models. All the results are shown as follows.

5.1 Model performance comparison of proposed transformer models

To confirm the efficiency of the proposed ensemble SHAP method, the final global factor importance (global ensemble SHAP value) is shown in Fig. 2. The ensemble global SHAP value can show the factor difference more clearly, which fits our general human common sense better. After we used the ensemble SHAP results as prior knowledge and used the knowledge to build self-attention transformer models, we compared our proposed models with FTT models and other classification models. The results (model accuracy: Acc) are shown in Table 2. In the MHLW dataset, our proposed models do not have the same level of performance as other classification methods. Because we only used 20% of the data to acquire knowledge. Then, we used the knowledge to build transformer models and acquired nearly the same level of performance (bold results in Table 2) as FTT

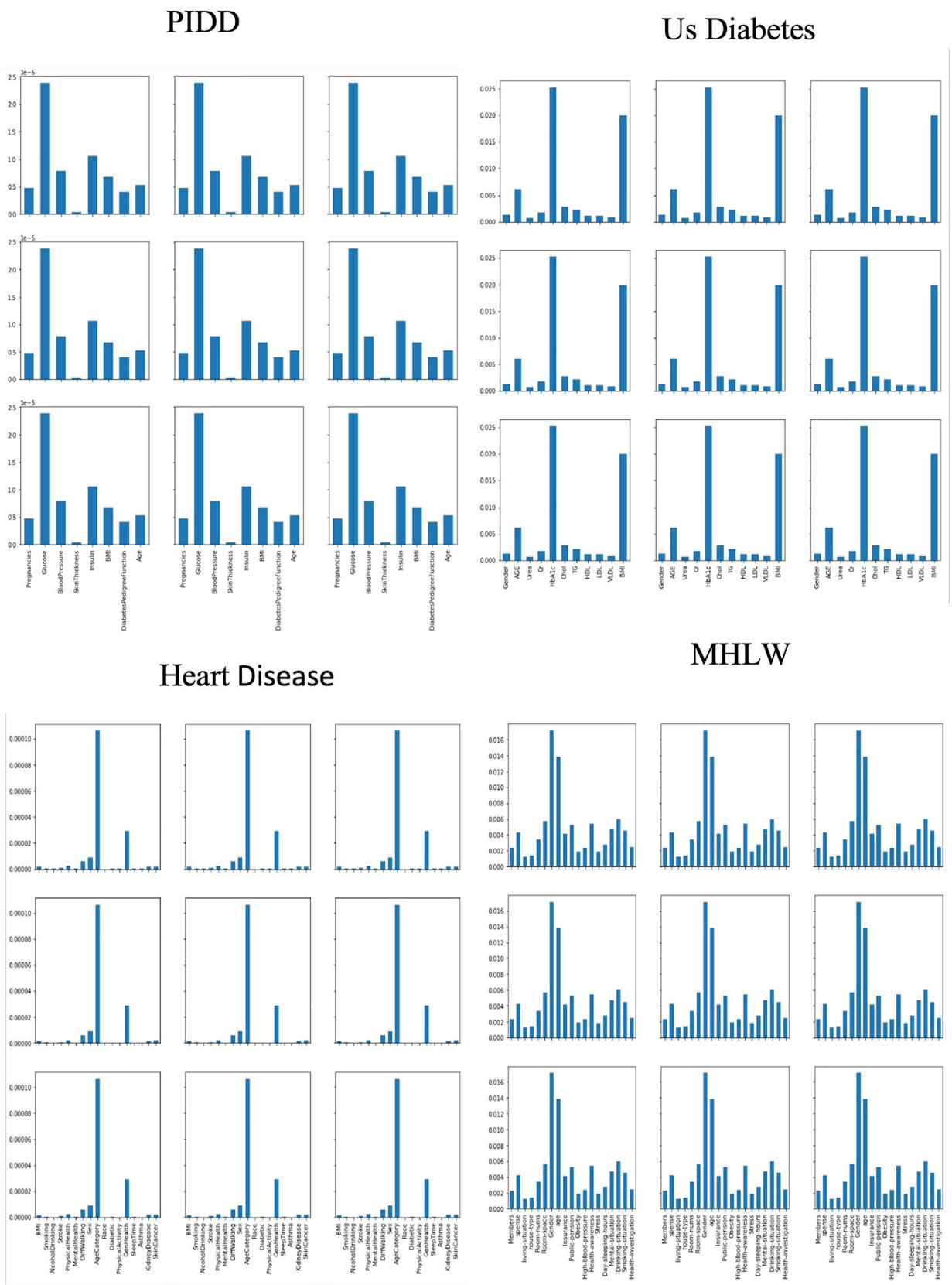

Fig. 2 Ensemble SHAP factor importance for four datasets

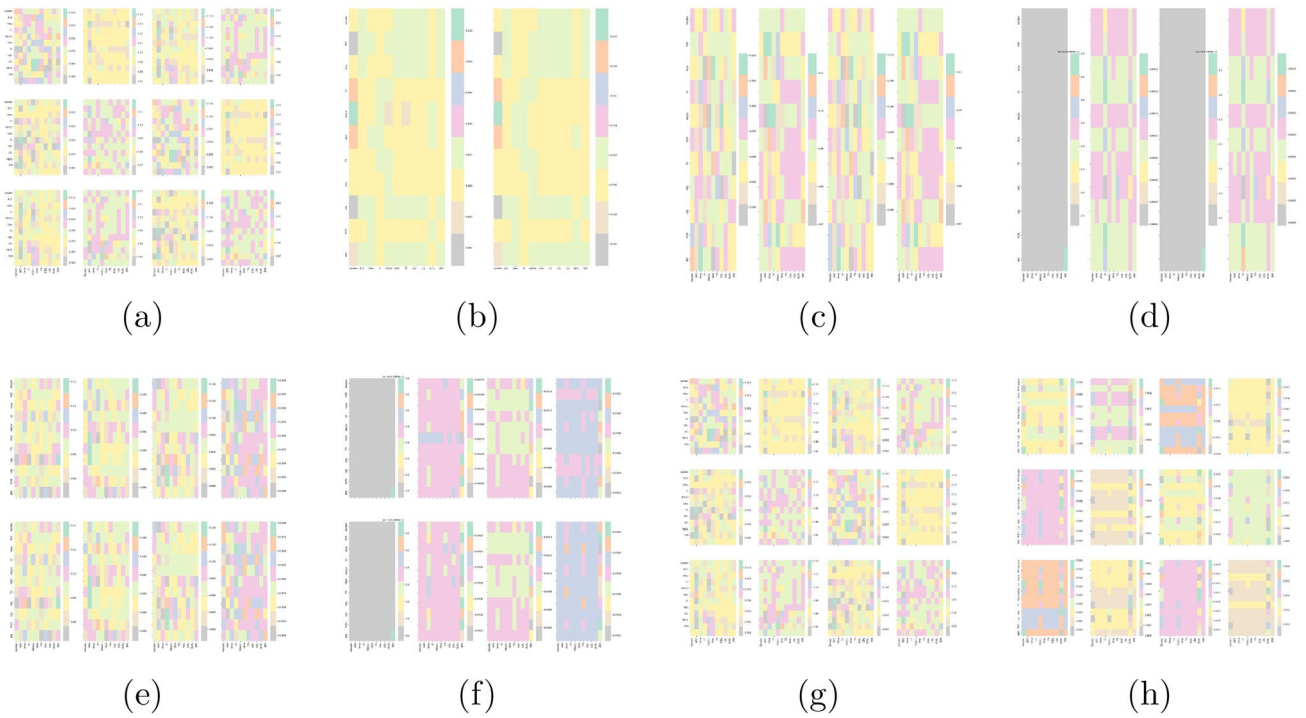

Fig. 3 The self-attention of proposed transformer models with general FTT models (Diabetes dataset) **a** self-attention in each layer for 2 layers of FTT models; **b** self-attention in each layer for 2 layers of proposed models; **c** self-attention in each layer for 4 layers of FTT models; **d** self-attention in each layer for 4 layers of proposed models; **e** self-attention in each layer for 8 layers of FTT models; **f** self-attention in each layer for 8 layers of proposed models; **g** self-attention in each layer for 12 layers of FTT models; **h** self-attention in each layer for 12 layers of proposed models

models. However, our proposed prior knowledge-integrated transformer model performs better (bold results in Table 2) than FTT models in the PIDD and heart disease datasets. Especially for the heart disease dataset, we used 20% of the data to acquire knowledge and build knowledge-integrated self-attention transformer models. Moreover, the attention of our proposed self-attention transformer models became more stable than general FTT models, as shown in the following subsection.

5.2 The self-attention comparison of transformer models

To understand the theory of the transformer models, we also checked the self-attention of each transformer block. We compared the FTT models and our proposed self-attention transformer models. The details are shown in Figs. 3, 4, 5 and 6.

When we check the self-attention among various transformer models, the attention in each transformer block changes randomly in FTT models. However, in our proposed self-attention transformer models, the attention of each transformer block becomes stable in all four datasets. To avoid possible randomness, we tested our proposed self-attention transformer models in 2 self-attention layers, 4 self-attention layers, 8 self-attention layers, and 12 self-attention layers transformer models. The results are shown in Figs. 3, 4, 5 and 6. The self-attention of our proposed knowledge-integrated transformer model becomes more stable than general FTT models, especially in the lower self-attention layer transformer models. In the 2 self-attention layer transformer models, the attention is the same. In the 4 self-attention layer and 8 self-attention layer transformer models, the attention is also nearly identical. Moreover, when we compare the self-attention of our proposed transformer model with the coefficients among features, we can

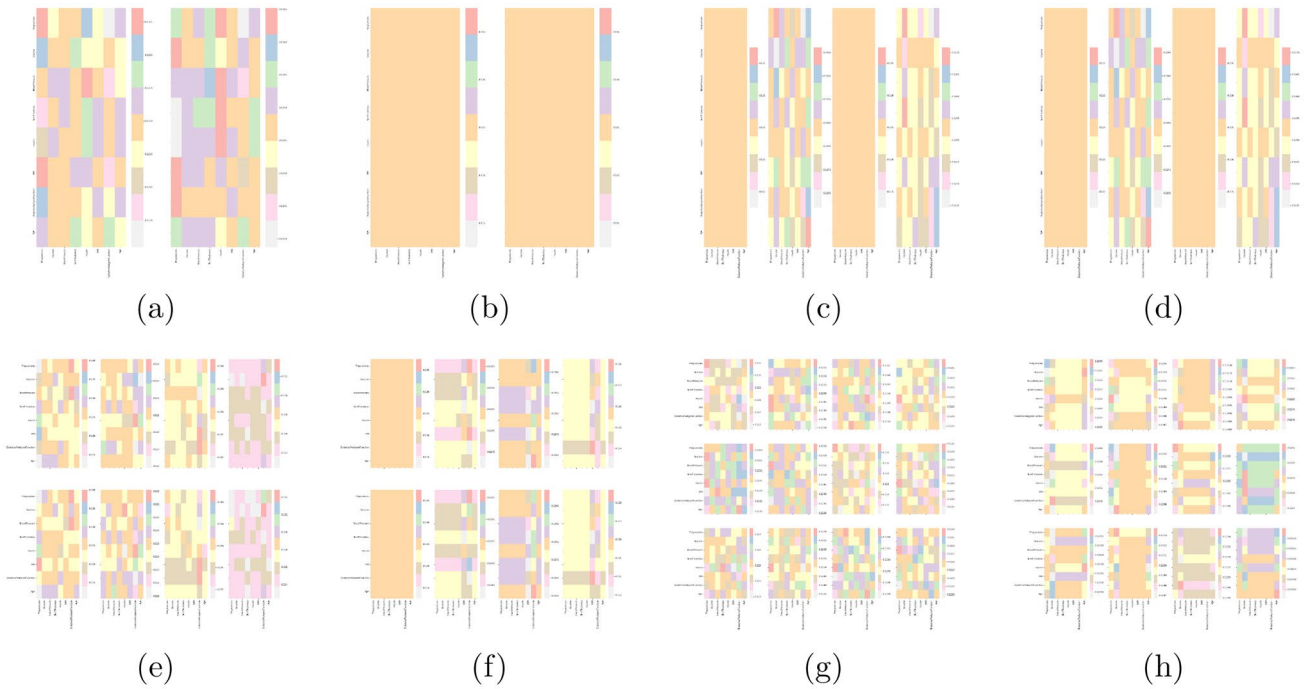

Fig. 4 The cooperation of proposed initial weight setting and general initial weight setting (PIDD dataset) **a** self-attention in each layer for 2 layers of FTT models; **b** self-attention in each layer for 2 layers of proposed models; **c** self-attention in each layer for 4 layers of FTT models; **d** self-attention in each layer for 4 layers of proposed models; **e** self-attention in each layer for 8 layers of FTT models; **f** self-attention in each layer for 8 layers of proposed models; **g** self-attention in each layer for 12 layers of FTT models; **h** self-attention in each layer for 12 layers of proposed models

find that the factors' self-attention (Fig. 7) of our proposed transformer models becomes similar to the factor coefficients in general machine learning models (Fig. 8). In contrast, FTT models' self-attention seems distributed randomly and has lower similarity with the factor coefficients (Fig. 8). When we use our proposed prior knowledge as input, the coefficients among factors seem to become similar (color in Fig. 8 becomes similar in each datasets)

6 Discussion

In this study, we used the proposed ensemble SHAP value as knowledge to build self-attention transformer models based on knowledge. The performance of our prior knowledge-integrated models has better performance than the non-knowledge-integrated FTT models. The better performance of our proposed models ensured that our proposed knowledge-integrated transformer model is an available research idea. Moreover, when we treated the ensemble SHAP value as knowledge and inserted the knowledge into transformer models, the self-attention of our knowledge-integrated transformer models became more stable than the general FTT model in all four tested datasets. Stable self-attention of each layer verified that the knowledge inserted in the transformer models influenced the transformer models. Meanwhile, stable self-attention of transformer models inspires us that we can interpret the AI models directly, rather than using agented methods [41–43] to explain the AI models. Moreover, our study certificated that knowledge-integrated AI methodology is achievable. Our results confirmed that a small AI model with knowledge is feasible for future research. Like the study of Feifei Li [58], our research also certifies that inserting knowledge in the AI model can help us improve the performance of traditional artificial intelligence methods. Moreover, our results inspire us to believe that a small AI model based on a small dataset [28] is possible.

While the reinforcement model rewards statement function and knowledge distillation filers the weights of NN

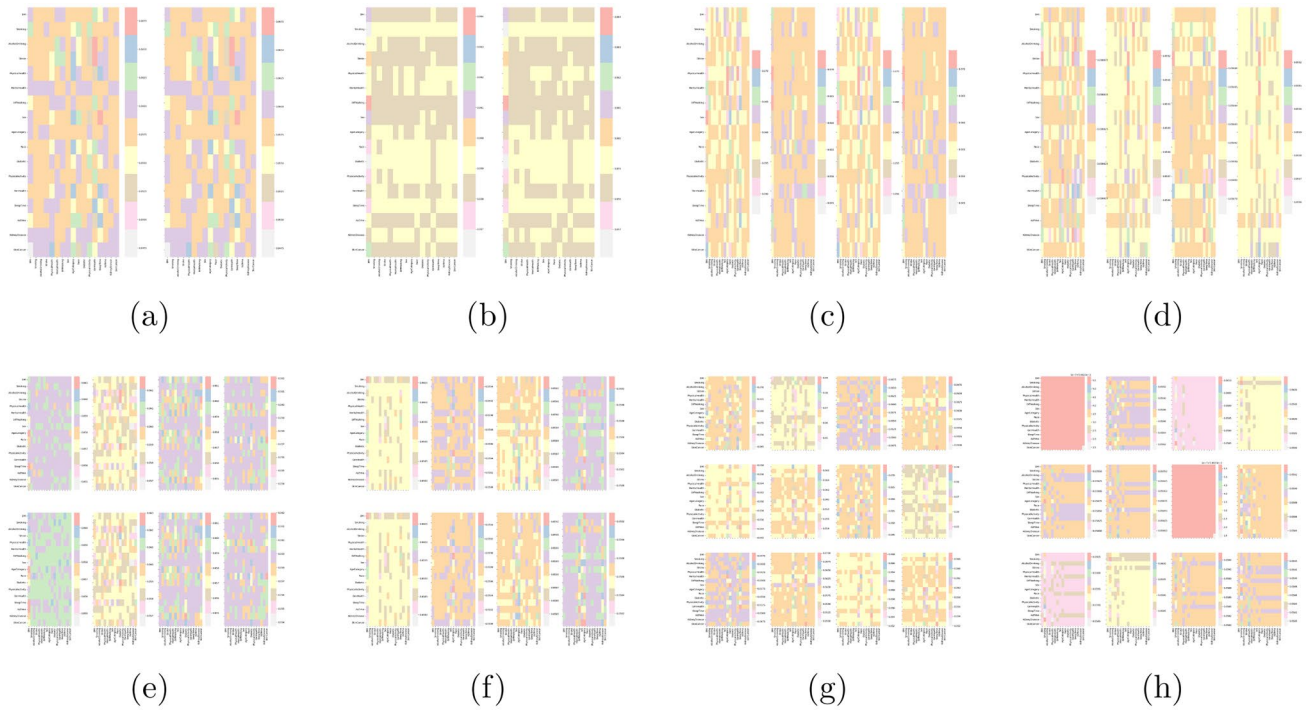

Fig. 5 The cooperation of proposed initial weight setting and general initial weight setting (Heart diseases) **a** self-attention in each layer for 2 layers of FTT models; **b** self-attention in each layer for 2 layers of proposed models; **c** self-attention in each layer for 4 layers of FTT models; **d** self-attention in each layer for 4 layers of proposed models; **e** self-attention in each layer for 8 layers of FTT models; **f** self-attention in each layer for 8 layers of proposed models; **g** self-attention in each layer for 12 layers of FTT models; **h** self-attention in each layer for 12 layers of proposed models

models, our proposed model used prior knowledge as the input of the transformer model, which can be transformed into other datasets and widely used in natural language processing, computer vision, and voice analysis areas. Our results also certified that our proposed model is available in classification models. Moreover, our study found that the attention of the transformer model becomes stable, which inspires us that we can probably understand the logic of NN models and make deep learning AI models transparent and reliable in the future.

Our proposed knowledge-integrated AI models used less data and performance than general AI models. Moreover, our results confirmed that our proposal is efficient. Our study certified that knowledge-integrated small AI models are available and efficient. Meanwhile, the attention results of our proposed transformer models show that the knowledge-integrated transformer models differ from the general transformer model. The self-attention of our proposed models becomes stable in each layer, unlike general transformer models (Figs. 3, 4, 5, 6). These inspire us that there must be logic and undefined rules in NN models. We can explore the real neural connection of NN in future studies and make AI models more transparent and reliable.

Certainly, there are also some limitations in our study. The prior knowledge used is different from the natural human experience. We will find quantified human knowledge to test our model in future work. However, our proposed methodology is one significant try for building a trustworthy small AI model, which will inspire more studies about reliable AI. Meanwhile, our study also confirmed that the small AI model with a small data set is feasible. At the same time, nearly all research efforts have focused on the large AI model, which is difficult in some research areas and wastes energy.

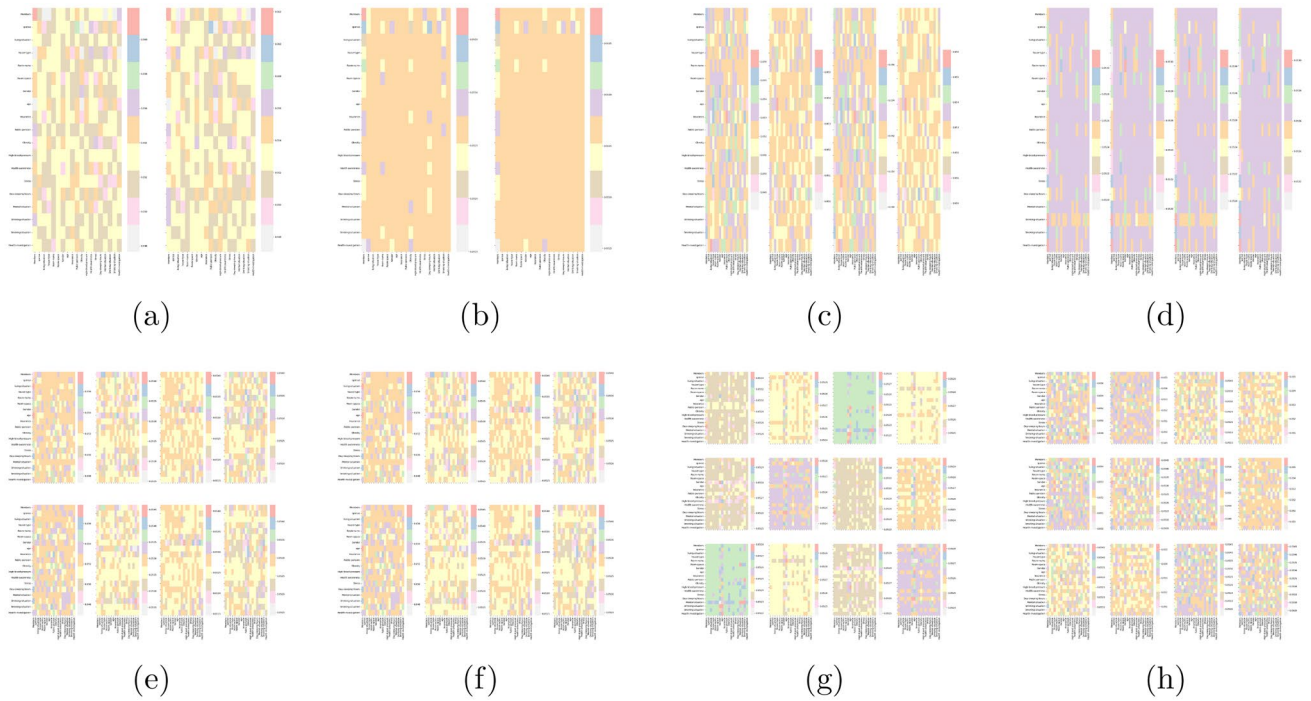

Fig. 6 The cooperation of proposed initial weight setting and general initial weight setting (MHLW dataset) **a** self-attention in each layer for 2 layers of FTT models; **b** self-attention in each layer for 2 layers of proposed models; **c** self-attention in each layer for 4 layers of FTT models; **d** self-attention in each layer for 4 layers of proposed models; **e** self-attention in each layer for 8 layers of FTT models; **f** self-attention in each layer for 8 layers of proposed models; **g** self-attention in each layer for 12 layers of FTT models; **h** self-attention in each layer for 12 layers of proposed models

7 Conclusion

In this study, we creatively designed knowledge-integrated AI models using prior knowledge to build transformer models. Our results confirmed the feasibility of our proposed methodology. Meanwhile, our research has certified that the research about trustable and logic-based AI models based on small data is feasible in the future. Indeed, there are some limitations to our study. More future work on trustable AI is still needed. However, our research inspires future studies about theory-based, trustable AI models in small-data-based. It paves the way for explaining and understanding the logic and theory of black-box AI models.

8 Future scope

Our future work will still explore the possibility of building transparent and reliable AI models, hoping to clarify the logic among NN models. Meanwhile, we will also consider using our proposed model in an actual life screen, especially in the medical and healthcare fields, which generally need more data to build big AI models.

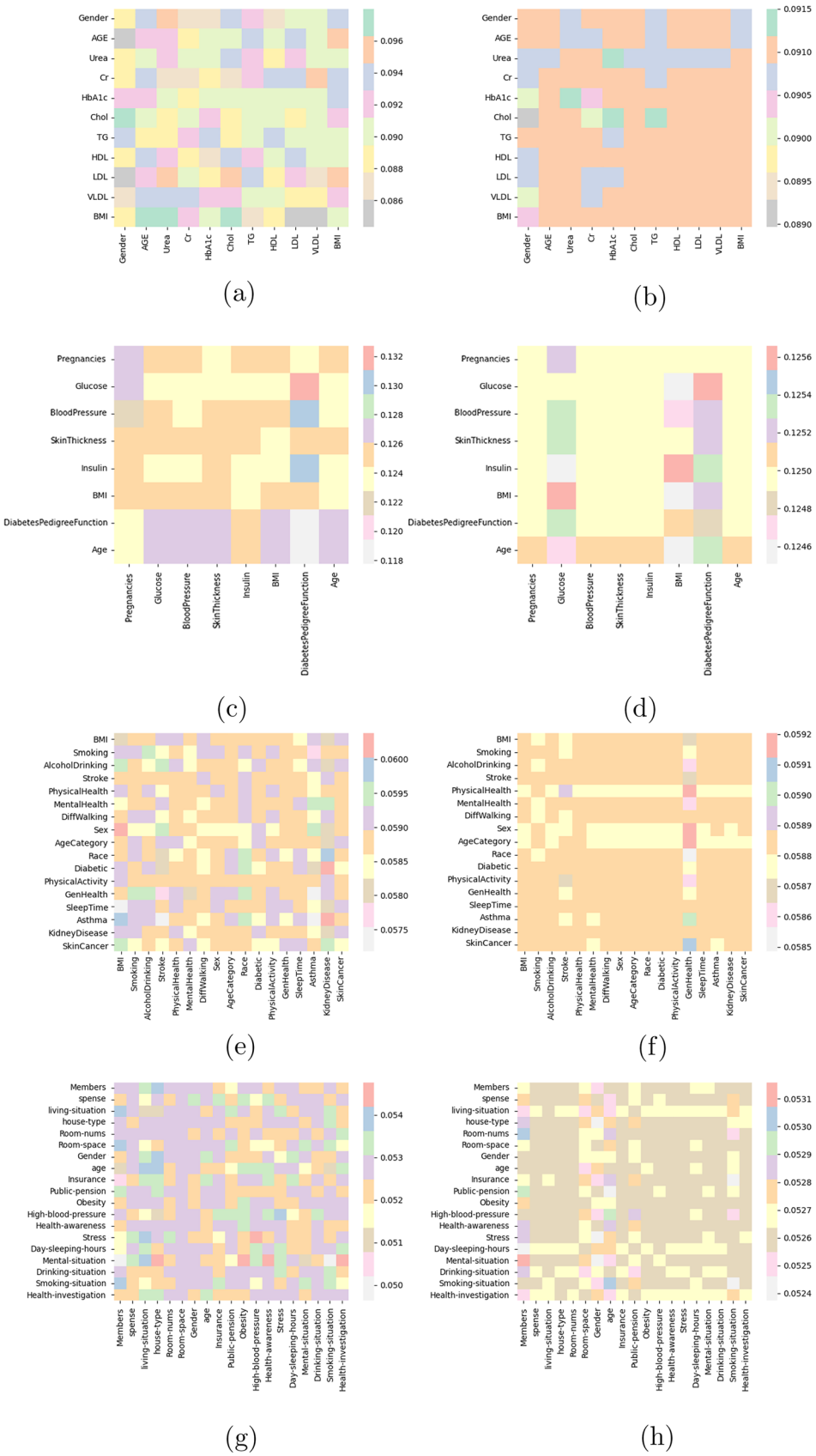

Fig. 7 The average of self-attention comparison among FTT models and proposed models (8 layers) **a** average self-attention of FTT models (US Diabetes); **b** average self-attention of proposed models (US Diabetes); **c** self-attention of FTT models (PIDD); **d** self-attention of proposed models (PIDD); **e** self-attention of FTT models (Heart disease); **f** self-attention of proposed models (Heart disease); **g** self-attention of FTT models (MHLW); **h** self-attention of proposed models (MHLW)

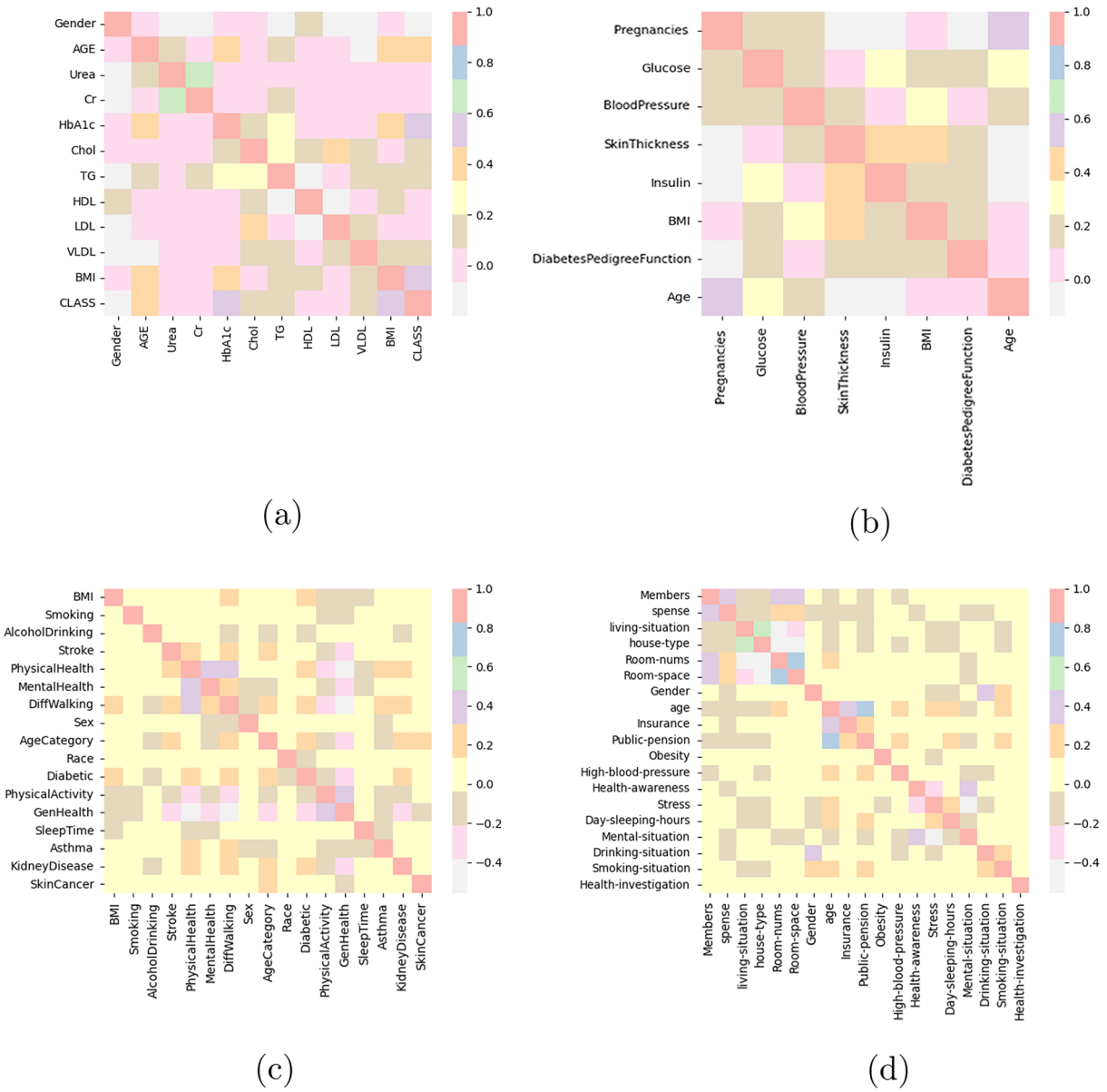

Fig. 8 The coefficients of prior knowledge integrated input **a** the factor coefficients of the US Diabetes dataset; **b** the factor coefficients of the PIDD dataset; **c** the factor coefficients of the Heart disease dataset; **d** the factor coefficients of the MHLW dataset

Funding The authors declare that no funds and grants were received during the preparation of this manuscript. However, we thank the Ministry of Healthcare, Labor, and Welfare of Japan for supplying the anonymous data for our study.

Data Availability The non-open dataset in this study needs usage approval from the Ministry of Healthcare, Labor, and Welfare of Japan.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval The ethics of the Ministry of Healthcare, Labor, and Welfare of Japan approved this study for using anonymous data.

References

1. Said Y, Alanazi A (2023) Ai-based solar energy forecasting for smart grid integration. Neural Comput Appl 35:8625–8635
2. Chang V, Bailey J, Xu QA, Sun Z (2023) Pima Indians diabetes mellitus classification based on machine learning (ml) algorithms. Neural Comput Appl 35:16157–16173
3. Sreekala K, Rajkumar N, Sugumar R, Sagar K, Shobarani R, Krishnamoorthy KP, Saini A, Palivela H, Yeshitla A (2022) Skin diseases classification using hybrid ai based localization approach. Comput Intell Neurosci 2022:7. https://doi.org/10.1155/2022/6138490
4. Wang C (2022) Ai-based heterogenous large-scale english translation strategy. Mob Inf Syst 2022:8344814
5. Du Y, Xu D (2022) Analysis of graphic design based on ai interaction technology. J Environ Public Health. 2022. https://doi.org/10.1155/2022/8493528
6. Alam A, Praveen S (2021) A review of automatic driving system by recognizing road signs using digital image processing. J Inform Electr Electron Eng (JIEEE) 2(2):1–9
7. Ma W, Zhao S, Xu S, Guo K, Qin K (2021) In: International conference on smart transportation and city engineering vol 12050. SPIE, pp 591–598
8. Yang M (2022) Research on vehicle automatic driving target perception technology based on improved msrpn algorithm. J Comput Cogn Eng 1(3):147–151
9. Du Y, Zhi Jy (2022) Impacts of attention level on manual takeover performance in automatic driving on high-speed railways. Int J Hum Comput Interact 1–10
10. Meng Q, Zhao S, Huang Z, Zhou F (2021) Magface: a universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14225–14234
11. Aggarwal D, Zhou J, Jain AK (2021) Fedface: Collaborative learning of face recognition model. In: 2021 IEEE international joint conference on biometrics (IJCB). IEEE, pp 1–8
12. Du H, Shi H, Zeng D, Zhang XP, Mei T (2022) The elements of end-to-end deep face recognition: a survey of recent advances. ACM Comput Surv (CSUR) 54(10s):1–42
13. Zeng J, Qiu X, Shi S (2021) Image processing effects on the deep face recognition system. Math Biosci Eng 18(2):1187–1200
14. Tian S, Wang M, Yuan F, Dai N, Sun Y, Xie W, Qin J (2021) Efficient computer-aided design of dental inlay restoration: a deep adversarial framework. IEEE Trans Med Imaging 40(9):2415–2427
15. Oza P, Sharma P, Patel S (2021) In: Proceedings of second international conference on computing, communications, and cyber-security. Springer, pp 377–392
16. Nazir A, Azhar A, Nazir U, Liu YF, Qureshi WS, Chen JE, Alanazi E (2021) The rise of 3d printing entangled with smart computer aided design during COVID-19 era. J Manuf Syst 60:774–786
17. Cohen MW, Gilo O, David L (2022) A computer aided medical classification system of COVID-19 ct lung scans using convolution neural networks. Comput Aided Des Appl 522–533
18. Croitoru FA, Hondru V, Ionescu RT, Shah M (2022) Diffusion models in vision: a survey. arXiv:2209.04747
19. Openai. https://openai.com/research/gpt-4. Accessed 01 May 2023
20. Bajaj D, Goel A, Gupta S, Batra H (2022) Muce: a multilingual use case model extractor using gpt-3. Int J Inf Technol 14(3):1543–1554
21. Mulla N, Gharpure P (2023) Leveraging well-formedness and cognitive level classifiers for automatic question generation on java technical passages using t5 transformer. Int J Inf Technol 15:1961–1973
22. Dowlagar S, Mamidi R (2021) Cmsaone@ dravidian-codemix-fire2020: A meta embedding and transformer model for code-mixed sentiment analysis on social media text. arXiv:2101.09004
23. Soni J, Mathur K (2022) Sentiment analysis based on aspect and context fusion using attention encoder with lstm. Int J Inf Technol 14(7):3611–3618
24. Sheik R, Parida SS, Nirmala SJ (2023) A hybrid model utilizing transfer learning for legal citation linking. Int J Inf Technol 15:2783–2792
25. Priya CSR (2023) Sentiment analysis from unstructured hotel reviews data in social network using deep learning techniques. Int J Inf Technol 15:3563–3574
26. George L, Sumathy P (2023) An integrated clustering and bert framework for improved topic modeling. Int J Inf Technol 15:2178–2195
27. Sengupta S, Mayya V, Kamath SS (2022) Detection of bradycardia from electrocardiogram signals using feature extraction and snapshot ensembling. Int J Inf Technol 14(6):3235–3244
28. Strickland E (2022) Andrew ng, ai minimalist: the machine-learning pioneer says small is the new big. IEEE Spectr 59(4):22–50
29. Wei CY, Luo H (2021) Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In: Conference on learning theory. PMLR, pp 4300–4354
30. Li Y, Shen W, Zhang Y, Chen H, Jiang M, Liu J, Jiang J, Gao W, Wu Z, Yang et al (2021) Openbox: a generalized black-box optimization service. In: Proceedings of the 27th ACMSIGKDD conference on knowledge discovery & data mining, pp 3209–3219
31. Wadden JJ (2022) Defining the undefinable: the black box problem in healthcare artificial intelligence. J Med Ethics 48(10):764–768
32. Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S (2021) Benchmarking and survey of explanation methods for black box models. arXiv:2102.13076
33. Durán JM, Jongsma KR (2021) Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. J Med Ethics 47(5):329–335
34. Knapič S, Malhi A, Saluja R, Främling K (2021) Explainable artificial intelligence for human decision support system in the medical domain. Mach Learn Knowl Extr 3(3):740–770
35. Sokolovsky A, Arnaboldi L, Bacardit J, Gross T (2021) Explainable machine learning-driven strategy for automated trading pattern extraction. arXiv:2103.12419
36. Covert I, Lundberg SM, Lee SI (2021) Explaining by removing: a unified framework for model explanation. J Mach Learn Res 22:209–211
37. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DKW, Newman SF, Kim J, Lee SI (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2(10):749–760. https://doi.org/10.1038/s41551-018-0304-0

38. Chen H, Lundberg S, Lee SI (2021) Explaining Models by Propagating Shapley Values of Local Components. Stud Comput Intell 914:261–270. https://doi.org/10.1007/978-3-030-53352-6_24
39. Molnar C (2022) Interpretable machine learning, 2nd edn. (add). https://christophm.github.io/interpretable-ml-book
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis **128**(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7. arXiv:1610.02391
41. Zhao X, Huang W, Huang X, Robu V, Flynn D (2021) In: Uncertainty in artificial intelligence. PMLR, pp. 887–896
42. Ribeiro MT, Singh S, Guestrin C (2016) In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, vol 13–17-August-2016. Association for Computing Machinery, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778
43. Meng Y, Yang N, Qian Z, Zhang G (2021) What makes an online review more helpful: an interpretation framework using xgboost and shap values. J Theor Appl Electron Commer Res 16(3):466–490. https://doi.org/10.3390/jtaer16030029
44. Feng DC, Wang WJ, Mangalathu S, Taciroglu E (2021) Interpretable xgboost-shap machine-learning model for shear strength prediction of squat rc walls. J Struct Eng 147(11):04021173
45. Wen X, Xie Y, Wu L, Jiang L (2021) Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with lightgbm and shap. Accid Anal Prev 159:106261
46. Li Z (2022) Extracting spatial effects from machine learning model using local interpretation method: an example of shap and xgboost. Comput Environ Urban Syst 96:101845
47. Chelgani SC, Nasiri H, Alidokht M (2021) Interpretable modeling of metallurgical responses for an industrial coal column flotation circuit by xgboost and shap—a "conscious-lab'' development. Int J Min Sci Technol 31(6):1135–1144
48. Yang C, Chen M, Yuan Q (2021) The application of xgboost and shap to examining the factors in freight truck-related crashes: an exploratory analysis. Accid Anal Prev 158:106153
49. Jabeur SB, Mefteh-Wali S, Viviani JL (2021) Forecasting gold price with the xgboost algorithm and shap interaction values. Ann Oper Res 1–21
50. Wang D, Thunéll S, Lindberg U, Jiang L, Trygg J, Tysklind M (2022) Towards better process management in wastewater treatment plants: process analytics based on shap values for tree-based machine learning methods. J Environ Manag 301:113941
51. Van den Broeck G, Lykov A, Schleich M, Suciu D (2022) On the tractability of shap explanations. J Artif Intell Res 74:851–886
52. Jiang P, Suzuki H, Obi T (2023) Interpretable machine learning analysis to identify risk factors for diabetes using the anonymous living census data of japan. Health Technol 13:1–13
53. Alwadi M, Chetty G, Yamin M (2023) A framework for vehicle quality evaluation based on interpretable machine learning. Int J Inf Technol 15(1):129–136
54. Mitrentsis G, Lens H (2022) An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. Appl Energy 309:118473
55. Zhao W, Joshi T, Nair VN, Sudjianto A (2020) Shap values for explaining cnn-based text classification models. arXiv:2008.11825
56. Wang J, Wiens J, Lundberg S (2021) In: International conference on artificial intelligence and statistics. PMLR, pp 721–729
57. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4768–4777
58. Krishna R, Lee D, Fei-Fei L, Bernstein MS (2022) Socially situated artificial intelligence enables learning from human interaction. Proc Natl Acad Sci 119(39):e2115730119
59. Lee D, Seo H, Jung MW (2012) Neural basis of reinforcement learning and decision making. Ann Rev Neurosci 35:287–308
60. Gou J, Yu B, Maybank SJ, Tao D (2021) Knowledge distillation: a survey. Int J Comput Vis 129:1789–1819
61. Ignatiev A, Narodytska N, Marques-Silva J (2019) On validating, repairing and refining heuristic ml explanations. arXiv:1907.02509
62. Tao G, Ma S, Liu Y, Zhang X (2018) Attacks meet interpretability: attribute-steered detection of adversarial samples. Adv Neural Inf Process Syst 31:7728–7739
63. Shih A, Choi A, Darwiche A (2018) A symbolic approach to explaining Bayesian network classifiers. arXiv:1805.03364
64. Narodytska N, Shrotri A, Meel KS, Ignatiev A, Marques-Silva J (2019) In: Theory and applications of satisfiability testing-SAT 2019: 22nd international conference, SAT 2019, Lisbon, Portugal, July 9–12, 2019, Proceedings 22. Springer, pp 267–278
65. Slingerland P, Perry L, Kaufman J, Bycroft B, Linstead E, Mandrake L, Doran G, Goel A, Feather MS, Fesq L et al (2022) In: 2022 IEEE aerospace conference (AERO). IEEE, pp 1–20
66. Cohen R, Schaekermann M, Liu S, Cormier M (2019) Trusted AI and the contribution of trust modeling in multiagent systems. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems. pp 1644–1648
67. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. https://openreview.net/pdf?id=BZ5a1r-kVsf. Accessed 24 Oct 2023
68. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2021) Transformers in vision: a survey. ACM Comput Surv (CSUR) 54:1–41
69. Lin T, Wang Y, Liu X, Qiu X (2021) A survey of transformers. arXiv:2106.04554
70. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
71. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16 x 16 words: transformers for image recognition at scale. arXiv:2010.11929
72. Sun L, Zhao G, Zheng Y, Wu Z (2022) Spectral-spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–14
73. Jiang P, Suzuki H, Obi T (2023) Xai-based cross-ensemble feature ranking methodology for machine learning models. Int J Inf Technol 15(4):1759–1768
74. kaggle. Pima Indians diabetes database (2006). https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed 15 May 2023
75. Rashid A (2020) Diabetes dataset. https://doi.org/10.17632/wj9rwkp9c2.1. Accessed 15 May 2023
76. C. for Disease Control, Prevention. Personal key indicators of heart disease (2020). https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
77. L. Ministry of Health, W. of Japan (2023). https://www.mhlw.go.jp/english/index.html. Accessed 15 May 2023