NutriVLM: Optimizing Multimodal Models for Comprehensive Nutritional Assessment

1st Jingxuan Zhang
Beihang University
Beijing, China
20241069@buaa.edu.cn

2nd Quan Duan
New York University
New York, United States
qd2045@nyu.edu

3rd Gong Huang
ShiFang Technology Inc.
Hangzhou, China
huanggong@buaa.edu.cn

4th Liu Liu
ShiFang Technology Inc.
Hangzhou, China
williuworld@163.com

5th Zhenbo Xu
Beijing Univ. of Posts and Telecommun.
Beijing, China
xuzhenbo@bupt.edu.cn

Qinghong Yang*
Hangzhou International Innovation Institute,Beihang University
Hangzhou, China
yangqh@buaa.edu.cn

***Abstract*—Nutritional analysis plays a vital role in promoting healthy eating and supporting personal health management and public health policies. Instead of in-efficient manual recording, the recent large vision-language model (LVLM) is more promising for automatic food recognition and nutritional analysis. In this paper, we introduce a high-quality nutrition multi-modal dataset and a novel evaluation framework named NutriVLM, which includes metrics such as food type recognition accuracy, weight recognition accuracy, nutritional recognition accuracy, and overall nutritional assessment. Through extensive evaluations using NutriVLM, we find that current popular LVLMs, such as GPT-4o and DALL-E, struggle to provide reliable results in nutritional assessment. To enhance their performance, we present a multi-round prompt optimization strategy focusing on food type identification, weight estimation, and nutritional information prediction. The proposed strategy brings significant performance improvements, especially when combining prompt optimization with optimal model selection. We believe our findings provide insights into the practical deployment of advanced LVLMs for food recognition and nutritional analysis.**

Keywords—multimodal models, nutritional analysis, food image recognition, evaluation metrics, machine learning

I. Introduction

In recent years, the increasing demand for healthy eating has underscored the significance of accurate nutritional analysis in both personal health management and the broader food industry [1, 2]. Nutritional data is essential for crafting personalized dietary recommendations and shaping public health policies [3]. However, existing nutritional analysis tools predominantly rely on manual data input, which is both labor-intensive and prone to errors, leading to inefficiencies in processing and inaccuracies in nutritional calculations [4]. To address these limitations, advancements in machine learning, computer vision, and natural language processing (NLP) have paved the way for the development of intelligent and automated nutritional analysis systems [5].

Despite the success of large vision-language models (LVLMs), such as GPT-4 and LLaVA, in multimodal content generation and understanding, their application to nutritional analysis remains challenging [6]. Experimental results demonstrate that while LVLMs can recognize food images and generate nutritional data, they often suffer from inaccuracies. These issues are exacerbated by hallucination phenomena, where the models produce contextually plausible but factually incorrect outputs [7]. In addition, variations in food types, image quality (e.g., lighting, angle), and dataset diversity contribute to nutritional assessment errors [8]. There are also instances where the models fail to output any nutritional information due to the complexity of the visual inputs, thereby limiting their practical utility in real-world nutritional analysis tasks [9, 10].

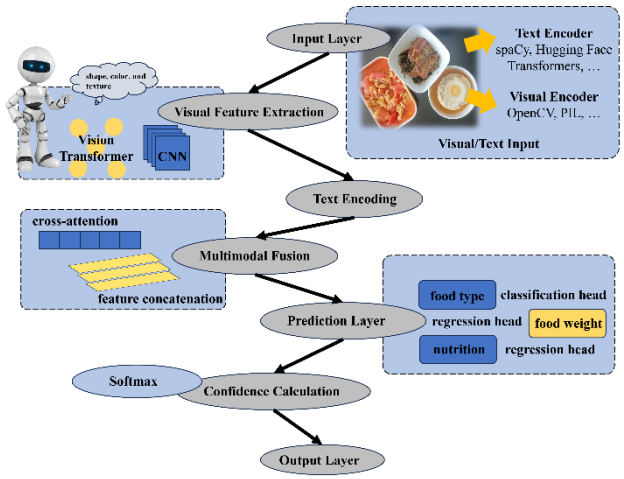

Fig. 1. Nutritionist Large Model: Multimodal Food Recognition and Nutritional Analysis Workflow

To address these challenges, this study introduces three key contributions aimed at enhancing the performance of LVLMs in food recognition and nutritional analysis:

1.Dataset Construction: The creation of a high-quality, annotated food image dataset encompassing diverse food types and detailed nutritional profiles. This dataset provides a solid foundation for developing and benchmarking multimodal models.

2.**NutriVLM** Evaluation Framework: The development of an evaluation framework that incorporates innovative metrics, such as Food Type Recognition Accuracy (FTRA), Weight Recognition Accuracy (WRA), Nutritional Value Assessment (NRA), and Overall Nutritional Assessment (ONA), to systematically assess and optimize model performance in nutritional analysis tasks.

3.Model Evaluation and Optimization: A comprehensive evaluation of state-of-the-art LVLMs, accompanied by targeted optimization strategies designed to improve their utility in food recognition and nutritional analysis across various real-world scenarios.

These contributions offer significant insights into the application of multimodal machine learning for nutritional analysis, setting the stage for more accurate and reliable AI-driven solutions in health and nutrition domains.

II. Related Work

A. Dataset Construction

To provide a robust foundation for evaluating the performance of multimodal models in food recognition and nutritional analysis, this study constructed a high-quality, diverse food image dataset, annotated with weight and nutritional information [11]. The dataset spans 10 distinct food categories: Fruits, Vegetables, Western Desserts, Western Dishes, Chinese Desserts, Chinese Dishes, Packaged Ready-to-Eat Food, Packaged Food Requiring Preparation, Beverages, and Bubble Tea, comprising more than 5,000 images. Each food sample in the dataset is meticulously labeled with detailed information, including food category, weight, and key nutritional components such as calories, protein, fat, and carbohydrates. All nutritional data were obtained directly from actual food packaging to ensure accuracy and consistency.

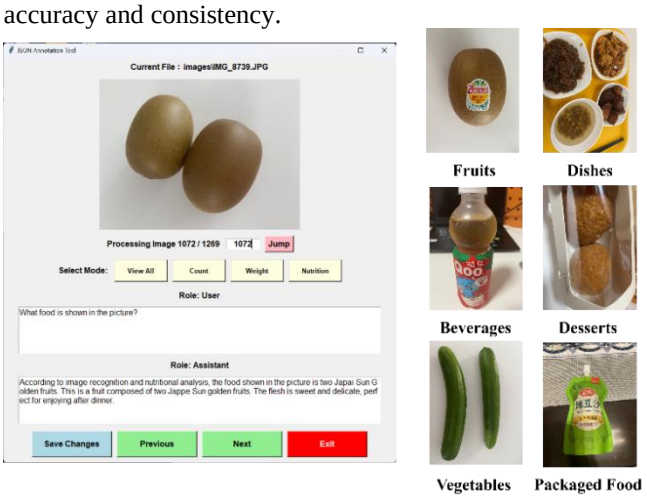

Fig. 2. Food Nutrition Information Q&A Annotation Tool with Sample Food Images

To further improve the dataset's quality and support model training on more complex food images, several image processing techniques were applied. These included Noise Reduction to eliminate unwanted artifacts, Color Correction to normalize and enhance the visual consistency of images, Edge Detection to sharpen object boundaries, and Contrast Enhancement and Sharpening to increase clarity and highlight key features [12]. These preprocessing steps enhanced the dataset's diversity and variability, ensuring that models could learn from a wide range of scenarios, thus improving their generalization capabilities.

To ensure the diversity and accuracy of the dataset, an automated Python-based annotation tool was developed to streamline the labeling process. This tool supports multi-angle image capture and enables multiple annotations for each image, allowing for comprehensive data coverage [13]. Once data collection and labeling were completed, preprocessing steps were applied to optimize the dataset. The data were then split into training, validation, and test sets to ensure that the models' performance was thoroughly evaluated across different stages of training, validation, and testing [14]. This structured approach ensures that the models are exposed to a wide variety of data conditions, enhancing their ability to perform effectively in real-world applications.

B. *NutriVLM Evaluation Framework Design*

The **NutriVLM** Evaluation Framework is meticulously crafted to provide a comprehensive, systematic assessment of multimodal models used in food recognition and nutritional analysis [15]. Considering the inherent complexity and diversity of food images and the challenges in estimating nutritional content, it is essential to have a robust evaluation framework that can assess various aspects of model performance in a unified manner. **NutriVLM** is designed to meet this need by combining multiple core metrics, allowing a holistic evaluation of how well models can classify food types, estimate their weight, and predict nutritional content.

Traditional evaluation methods often focus narrowly on classification accuracy, disregarding the challenges involved in estimating nutritional information or handling complex food-related datasets. **NutriVLM** goes beyond this by integrating a broader set of metrics to cover both recognition and prediction tasks. In doing so, it provides a detailed insight into the model's performance in food-related tasks, such as predicting nutritional values from images or estimating food weight from visual input. By focusing on these distinct yet interconnected tasks, **NutriVLM** provides a clearer understanding of how these models perform in real-world applications.

Moreover, the framework employs confidence-based weighting, a critical component that adjusts the influence of each prediction based on the model's confidence in its output. Predictions made with greater certainty are given more weight, ensuring that the evaluation reflects the model's reliability in making accurate predictions. This aspect of the framework is especially important in food recognition and nutritional analysis, where even small inaccuracies can lead to significant deviations in results.

NutriVLM comprises four essential metrics designed to capture the model's capabilities comprehensively:

- **Food Type Recognition Accuracy (FTRA)**: This metric evaluates the model's accuracy in classifying different food types. The metric is calculated as the proportion of correctly classified food types in the test set. The specific formula is as follows:

$$FTRA = \frac{\sum_{i=1}^{n}(\delta(\hat{y}_i, y_i) \cdot e^{confidence(\hat{y}_i)})}{\sum_{i=1}^{n} e^{confidence(\hat{y}_i)}} \quad (1)$$

Where $\delta(\hat{y}_i, y_i)$ is an indicator function equal to 1 if the predicted food type $\hat{y}_i$ matches the actual type y_i, and 0 otherwise [15]. The confidence term, $confidence(\hat{y}_i)$,

reflects the model's certainty in its prediction, computed using a softmax function applied to the model's logits [17].

- **Weight Recognition Accuracy (WRA):** This metric assesses the model's ability to predict the weight of the food. It is calculated by measuring the error between the predicted and actual weight, and evaluating accuracy based on a specified tolerance range. The formula is as follows:

$$WRA = 1 - \frac{\sum_{i=1}^{n} (\frac{|\widehat{w_i} - w_i|}{w_i} \cdot e^{confidence(\widehat{w_i})})}{\sum_{i=1}^{n} e^{confidence(\widehat{w_i})}} \quad (2)$$

Here, $\widehat{w_i}$ is the predicted weight, and w_i is the actual weight. Confidence again acts as a weighting factor, ensuring that predictions with higher certainty have more impact on the final accuracy.

- **Nutritional Recognition Accuracy (NRA):** This metric evaluates the model's ability to predict the nutritional information of foods (e.g., calories, protein, carbohydrates, dietary fiber, sodium). The metric compares the predicted nutritional values to the actual nutritional data, and the formula is as follows:

$$NRA = 1 - \frac{\sum_{j=1}^{5} \sum_{i=1}^{n} (\frac{|\widehat{n_{ij}} - n_{ij}|}{n_{ij}} \cdot e^{confidence(\widehat{n_{ij}})})}{\sum_{j=1}^{5} \sum_{i=1}^{n} e^{confidence(\widehat{n_{ij}})}} \quad (3)$$

Where $\widehat{n_{ij}}$ represents the predicted value for the *j*-th nutrient of the *i*-th food item, and n_{ij} is the actual value.

- **Overall Nutritional Assessment (ONA):** This metric combines the previous three metrics to comprehensively evaluate the model's performance in food type, weight, and nutritional recognition. This comprehensive metric provides a more holistic assessment of the model's real-world applicability. The formula is as follows:

$$ONA = \alpha \times FTRA + \beta \times WRA + \gamma \times NRA \quad (4)$$

Where α, β, and γ are coefficients that can be adjusted based on the importance of each metric in a specific application context.

Additionally, confidence-based evaluation plays a crucial role throughout the **NutriVLM** framework. Confidence is calculated using the softmax function, which converts the model's logits into probabilities, with the highest value representing the model's confidence in its prediction [18]. The formula for confidence is:

$$Confidence(\hat{y}) = \max\big(softmax(\boldsymbol{z})\big) \quad (5)$$

Where *z* represents the logits produced by the model, and the softmax function transforms these logits into probabilities [19]. Confidence is used to adjust the weighted contribution of the model in the evaluation, reflecting the model's confidence in its predictions.

The **NutriVLM** framework was applied to evaluate various multimodal models, including open-source models like Qwen and commercial models such as GPT-4o. By systematically comparing these models through the **NutriVLM** framework, this study identifies the strengths and weaknesses of each model, offering insights into how they can be optimized for food recognition and nutritional analysis tasks across different datasets and scenario.

C. Prompt Optimization and Model Combination Strategies

Prompt optimization and model combination were crucial in enhancing the performance of multimodal models for food recognition and nutritional analysis tasks [20]. The design of prompts directly impacted how models interpreted input data, processed signals, and generated output. Through multiple rounds of experimentation, prompt structures were refined to achieve the highest performance, particularly for Overall Nutritional Assessment (ONA). This iterative process enabled the models to better handle tasks like food classification, weight estimation, and nutritional prediction.

Various prompt configurations, such as "Identify the food type and weight" and "Provide nutritional information for the following food image," were tested. These optimized prompts significantly improved model accuracy in addressing common challenges like food type misclassification and incomplete nutritional data. The refined prompts resulted in a marked improvement across all metrics.

Alongside prompt optimization, combining different models further boosted performance. Comparative analysis showed that GPT-4o excelled in food type recognition while DALL-E performed well in nutritional value prediction. By leveraging their strengths, the combined models achieved higher scores across all evaluation metrics. This modular approach allowed each model to specialize, improving the accuracy of both food classification and nutritional data. As seen in Table III, the combined use of GPT-4o and DALL-E with optimized prompts outperformed the individual models, providing a comprehensive solution for complex tasks in real-world applications such as personalized nutrition and public health monitoring.

III. Experiments And Results

A. Experimental Setup

All experiments were conducted in a high-performance computing environment equipped with NVIDIA GPUs [21]. The experiments were implemented using Python, with model training and inference performed via the PyTorch framework. Data processing was handled using NumPy and Pandas, and Matplotlib was utilized for data visualization [22-24]. The experiments involved several multimodal models, including open-source models such as Qwen, CLIP, and ALIGN, as well as commercial models like Pangu, DALL-E, and GPT-4o.

The dataset, built by the research team, contained over 5,000 food images spanning 10 different categories, including fruits, vegetables, Western and Chinese dishes, and beverages. Each image was annotated with food category, weight, and nutritional data. The dataset was randomly split into training (70%), validation (15%), and testing sets (15%) to ensure robustness at every stage of the model's development.

To enhance the model's understanding of the food images, various prompt structures were created. For instance, prompts such as "Identify the food type and nutritional values for the image" and "Estimate the weight and nutrition for this food image" were used. Models' responses were recorded, analyzed, and evaluated for effectiveness. A k-fold cross-validation technique was also employed to improve model stability and generalization across different data split [25].

B. Experimental Results

The experimental results were evaluated using the four core metrics from the **NutriVLM** evaluation framework: Food Type Recognition Accuracy (FTRA), Weight

Recognition Accuracy (WRA), Nutritional Recognition Accuracy (NRA), and Overall Nutritional Assessment (ONA). As shown in Fig. 3, the performance of six models is compared based on these metrics. Each model is scored on FTRA, WRA, NRA, and ONA, and the results highlight the strengths and weaknesses of each model, offering insights into how the models performed across various evaluation metrics.

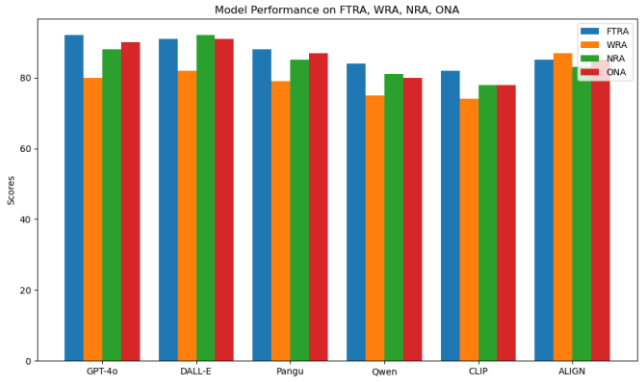

Fig. 3. Performance Comparison of Multimodal Models in Food Recognition and Nutritional Analysis

Additionally, the prompt optimization strategy introduced further improvements. The final prompt optimization strategy employed a multi-round approach where models were sequentially prompted to first identify the food type, then estimate the weight, and finally provide nutritional information. In the first round, the optimal prompt "Identify the food type in this image" allowed the models to classify the food type accurately. In the second round, the prompt "Estimate the weight of the food shown in this image" guided models to infer the weight based on the food's visual characteristics, such as size and density. Finally, the third-round prompt, "Provide the nutritional information (calories, protein, fat) for the food item in the image," enabled the models to predict nutritional data based on key metrics like calories and protein.

C. Result Analysis

The prompt optimization strategy significantly enhanced the performance of the models, particularly in the experiments involving the open-source models Qwen and CLIP. By refining the prompts, the models demonstrated notable improvements in the accuracy of food classification and nutritional information prediction. Furthermore, confidence analysis revealed that GPT-4o and DALL-E had higher confidence in their predictions, especially in complex scenarios, indicating strong stability and reliability.

TABLE I. MODEL PERFORMANCE ON DIFFERENT METRICS

Model	FTRA(%)	WRA(%)	NRA(%)	ONA(%)
GPT-4o	92	80	88	90
DALL-E	91	82	92	91
Pangu	88	79	85	87
Qwen	84	75	81	80
CLIP	82	74	78	78
ALIGN	85	87	83	85

After comparing the Overall Nutritional Assessment (ONA) scores of each model, it is evident that the prompt optimization strategy led to a marked improvement in model performance. Table 2 shows the ONA scores of each model before and after prompt optimization, along with the corresponding percentage improvements.

TABLE II. MODEL ONA SCORES BEFORE AND AFTER PROMPT OPTIMIZATION

Model	ONA (%) with Original Prompts	ONA (%) with Optimized Prompts	Improvement (%)
GPT-4o	85	90	5.88%
DALL-E	89	91	2.25%
Pangu	80	87	8.75%
Qwen	74	80	8.11%
CLIP	70	78	11.43%
ALIGN	82	85	3.66%

In the final analysis, the combination of GPT-4o and DALL-E proved to deliver the most effective results. This best model selection leveraged the strengths of GPT-4o in food type recognition and DALL-E in nutritional value prediction. The complementary abilities of the two models enabled more accurate and comprehensive analysis across all four key metrics: Food Type Recognition Accuracy (FTRA), Weight Recognition Accuracy (WRA), Nutritional Recognition Accuracy (NRA), and Overall Nutritional Assessment (ONA). The comparison of models before and after prompt optimization, along with the impact of combining GPT-4o and DALL-E, is presented in Table 3, showing significant improvements in model performance.

TABLE III. PERFORMANCE COMPARISON OF MODELS BEFORE AND AFTER PROMPT OPTIMIZATION

Model	FTRA(%)	WRA(%)	NRA(%)	ONA(%)	Improvement (%)
Original	92	80	88	90	-
Optimized Prompts	91	77	85	84	+2.4%
Optimized Prompts + Best Model	92	80	88	90	+9.7%

In summary, this experiment validated the performance of different models in food recognition and nutritional analysis tasks. By refining prompts and combining models, substantial improvements were achieved, providing valuable insights for future optimization and practical applications in the field of food recognition and nutritional analysis.

IV. CONCLUSIONS

This study developed a multimodal system for food recognition and nutritional analysis, evaluated through the NutriVLM framework. The results showed significant improvements in accuracy and stability through prompt optimization and model combination strategies. By refining prompts and leveraging the strengths of models like GPT-4o for food classification and DALL-E for nutritional analysis, the system achieved notable gains across key metrics. These findings provide a foundation for future research, particularly in optimizing AI models for real-world food recognition tasks, contributing to fields like personalized nutrition and public health monitoring.

V. REFERENCE

[1] Willett, W., et al., *Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems.* Lancet, 2019. **393**(10170): p. 447-492.

[2] Jiang, C. and J. Jing, *Recommending personalized dietary based on food nutrition and health knowledge base by using computer device, involves taking result data contained in target recommendation rule as recommended result*. Shenzhen Isoftstone Information Technolo.

[3] Castro-Vega, I., et al., *Validity, efficacy and reliability of 3 nutritional screening tools regarding the nutritional assessment in different social and health areas.* Medicina Clinica, 2018. **150**(5): p. 185-187.

[4] Kapse, V.M., Lovely, and R. Mishra, *Artificial intelligence-driven personalized nutrition assistant system for dietary health, has feedback module for monitoring user adherence to nutrition plans, and output module for presenting nutrition plans to user through interface module*. Noida Eng & Technology Inst.

[5] Zhang, D., W.-J. Li, and Aaai. *Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization*. in *28th AAAI Conference on Artificial Intelligence*. 2014. Quebec City, CANADA.

[6] Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks.* Communications of the Acm, 2017. **60**(6): p. 84-90.

[7] Li, F., et al., *Scales system of nutritional analysis and suggestions to establish a nutritional analysis and management system connected a database having plural nutrition information of subject matter*. Univ Chia Nan Pharmacy & Sci.

[8] Wang, F., S. Zhang, and H. Liu, *Open visual positioning method based on multimodal large model, involves obtaining image to be identified and target name, image and the target name are provided as input to a multi-mode large model, and visual target positioning description alignment is aligned with visual target*. Beijing Knowledge Atlas Technology Co.

[9] Kumar, K.G., et al., *Method for estimating food calorie using image artificial intelligence (AI) with RetinaNet feature extraction, involves comparing and calculating calories of food from identified ingredients along with approximate quantities using set of databases*. Chaitanya Bharathi Technology Inst.

[10] Lee, H.S., et al., *Image analysis-based nutrition information providing system using artificial intelligence, has service providing device that maps food information included in pre-stored food information and generates nutritional information using mapped food information*. Doinglab Inc; Doinglab Corp.

[11] Wang, W., *Data model conversion method, involves converting data of source target to final target by entity conversion tool of custom annotation data when code data is executed to self-defined annotation data, and performing annotation naming processing on code data*. Shanghai Tuhu Information Technology Co.

[12] Kokemohr, N., *Method for applying image enhancements involves hiding icons associated with image enhancements that are subsequent image enhancements relative to the selected image enhancement*. Kokemohr N; Google Inc.

[13] Jiang, H., et al., *Food category management method, involves combining multiple foods into package to obtain package information, where package information is provided with category information and item information for each item in package*. Shenzhen Xingfushangcheng Technology Co.

[14] Wang, J., et al., *RD-FGM: A novel model for high-quality and diverse food image generation and ingredient classification.* Expert Systems with Applications, 2024. **255**.

[15] Vaishnavi, S., et al., *Method for performing open-ended continual learning for real-world food recognition, involves using framework as ARCIKELM classifier for dynamically adjusting network architecture to reduce catastrophic forgetting*. St Martins Eng College.

[16] Bigaj, R., et al., *Method for detecting regression in relationship between performance indicator and AI metrics, involves identifying subset of AI metric outliers according to calculated baseline threshold and determined delta correction constant*. Int Business Machines Corp.

[17] Shao, H. and S. Wang, *Deep Classification with Linearity-Enhanced Logits to Softmax Function.* Entropy, 2023. **25**(5).

[18] Matsuo, T., *Confidence interval presentation program for machine learning model prediction, includes instructions for presenting combination of first output data and third output data, and reliability section*. Fujitsu Ltd.

[19] Cao, Y., R. Huang, and J. Wen, *Method for training performance of deep neural network, involves maintaining classification information from previous set of tasks by utilizing logits for matching during training on new set of tasks*. Royal Bank Canada.

[20] Chen, Z. and Y. Cai, *Method for optimizing large model prompt word, involves carrying out iterative optimization and testing on third prompt words, and determining sentence optimization prompt words according to test result and are output as optimal prompt words*. Guangzhou Pcitech Software Dev Co Ltd; Guangzhou Xinke Jiadu Technology Co Ltd; Pci Technology Group Co Ltd.

[21] El Zein, A., et al. *Performance evaluation of the NVIDIA GeForce 8800 GTX GPU for machine learning*. in *8th International Conference on Computational Science*. 2008. Cracow, POLAND.

[22] Zheng, H., G. Li, and K. Sun, *Method for deploying pyTorch model in program, involves providing compiled PyTorch library file to operating system in local storage path, and operating pytorch model based on PyTorCh model file after compiled PytorCh library file is loaded*. Shanghai Fengniao Jipei Information Tech.

[23] Hunter, J.D., *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 2007. **9**(3): p. 90-95.

[24] Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.

[25] Wong, T.-T. and P.-Y. Yeh, *Reliable Accuracy Estimates from k-Fold Cross Validation.* Ieee Transactions on Knowledge and Data Engineering, 2020. **32**(8): p. 1586-1594.