Devising a Set of Compact and Explainable Spoken Language Feature for Screening Alzheimer's Disease

Junan Li[1], *Yunxiang Li*[1], *Yuren Wang*[2], *Xixin Wu*[1], *Helen Meng*[1]

[1]Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong
[2]Dept. of Computer Science & Engineering, The Chinese University of Hong Kong
[1]{jli,yli,wuxx,hmmeng}@se.cuhk.edu.hk, [2]yrwang2@cse.cuhk.edu.hk

Abstract

Alzheimer's disease (AD) has become one of the most significant health challenges in an aging society. The use of spoken language-based AD detection methods has gained prevalence due to their scalability due to their scalability. Based on the Cookie Theft picture description task, we devised an explainable and effective feature set that leverages the visual capabilities of a large language model (LLM) and the Term Frequency-Inverse Document Frequency (TF-IDF) model . Our experimental results show that the newly proposed features consistently outperform traditional linguistic features across two different classifiers with high dimension efficiency. Our new features can be well explained and interpreted step by step which enhance the interpretability of automatic AD screening.

Index Terms: Alzheimer's disease detection, spoken language processing, large language models

1. Introduction

Alzheimer's disease (AD) detection presents a significant and growing challenge to healthcare and economic systems due to costly and complex diagnoses [1, 2, 3]. Current research underscores the importance of early intervention and the need for economically accessible, non-invasive and affordable alternatives for AD detection [4, 5, 6, 7]. Consequently, speech and language alternatives have emerged as early AD indicators, offering a promising, non-invasive diagnostic approach suitable for large-scale screening [8]. The Cookie Theft picture description task is one of the common cognitive assessment tasks that evaluates language and cognitive impairments through patients' descriptions of a complex scene.

For spoken language-based AD detection, two primary methods have recently been prevalent: linguistic feature extraction with classifier models and the use of pre-trained language models (PLMs) like BERT. Studies such as [9] focus solely on linguistic features, applying them across multiple languages. [10] combined linguistic features and classifiers for aphasia subtype classification, incorporating semantic coherence for robust results.

Another approach uses PLMs to capture semantic information and context. For instance, [11] utilized various acoustic and linguistic features, including BERT, to compare different PLMs. They found that BERT-based features significantly improved detection accuracy with both manual and ASR transcripts. Building on this, [12] advanced PLMs by incorporating prompt-based fine-tuning for AD detection, aligning training objectives with AD classification tasks for state-of-the-art results. An earlier study by [13] used PLMs like Whisper and BERT, integrating high-level acoustic and linguistic features along with task-related information to enhance accuracy. More recently, [14] investigated the impact of using ASR transcription on both linguistic feature-based methods and PLM-based methods.

More recently, Large Language Models (LLMs) such as GPT-4, which have shown remarkable capabilities in various tasks, are increasingly being explored for their potential in aiding AD detection. [15] explored the feasibility of using ChatGPT for primary screening of Mild Cognitive Impairment (MCI) based on text conversation analysis. Similarly, [16] assessed GPT-4's potential in dementia diagnosis, highlighting its strengths in zero-shot settings and interpretable explanations, but also its limitations, such as the inability to be fine-tuned and sensitivity to input quality. Moreover, prompt-based LLMs for AD detection face several challenges. Firstly, their outputs are not always controllable or traceable, as changes in the LLMs' versions may lead to shifts in their outputs. Secondly, substantial computational power is required due to the extremely large size of these models' parameters. Lastly, there are privacy concerns associated with uploading user data to the cloud.

Previous research has extensively utilized traditional linguistic features and language models. However, these studies did not explicitly consider task-specific features such as content coverage, which is critical for cognitive assessment. This work introduces a novel set of features, including those that leverage the Term Frequency-Inverse Document Frequency (TF-IDF) concept and features related to the Cookie Theft task. We utilize the advanced linguistic capabilities and visual processsing ability of LLMs to help the generation of our new features. The proposed new features are more interpretable for humans, thereby enhancing the explainability of AD detection. We compared our new features with 40 traditional linguistic features referenced in the literature [14]. The experimental results demonstrate that our new feature set, which is compact with only around 37.5% in dimensionality compared with the conventional feature set,consistently outperforms the traditional linguistic features. We achieved an competitive accuracy of 85.4% on the ADReSS test set using only a 15-dimensional feature set, highlighting the dimensional efficiency of our features.

To summarize, this work has three main contributions. First, we pioneered the breakdown of the Cookie Theft picture and leveraged the visual processing ability of LLMs to generate features. Our approach ensures that every step is clear and reasonable, leading to a traceable and explainable feature generation process. Secondly, we utilized tried and true technique from Information Retrieval (IR) to provide novel and grounded features from different perspectives. Lastly, we proposed a compact, effective, and explainable feature set that achieves competitive results compared to previous research.

This paper is organized as follows: Section 2 details the feature engineering process, including the definition and extraction

Table 1: *Fifteen proposed features description*

Feature Name	Description
Topic 1 Keywords Hit Rate	Hit rate of the keyword set 1 generated by LLM
Topic 2 Keywords Hit Rate	Hit rate of the keyword set 2 generated by LLM
Topic 3 Keywords Hit Rate	Hit rate of the keyword set 3 generated by LLM
BLEU-1	Averaged 1-gram BLEU score with 15 references generated by LLM
BLEU-2	Averaged 2-gram BLEU score with 15 references generated by LLM
BLEU-3	Averaged 3-gram BLEU score with 15 references generated by LLM
BLEU-4	Averaged 4-gram BLEU score with 15 references generated by LLM
METEOR	Averaged METEOR score with 15 references generated by LLM
TF-IDF similarity HC	Cosine similarity with the HC reference vector.
TF-IDF similarity AD	Cosine similarity with the AD reference vector.
TF-IDF Keywords Hit Rate	Hit rate of keywords selected by TF-IDF
avg_depth	Averaged parse tree depth
Filled Pauses	The number of filled pauses
Filled Pauses Ratio	The ratio between filled pauses and all tokens
WER	Word error rate

of the new features. Section 3 introduces the experimental settings and presents the results. Section 4 provides a discussion and an ablation study. Finally, Section 5 concludes the paper.

2. Method

2.1. Dataset

The dataset utilized in this study is derived from the ADReSS Challenge 2020 [17], which represents a curated subset of the Pitt Corpus within the DementiaBank database [18]. It comprises 156 speech samples and their corresponding transcripts from English-speaking participants engaged in the Cookie Theft picture description task. The participants are categorized into two groups: those without Alzheimer's disease (HC) and those with Alzheimer's disease (AD), with each group including 35 males and 43 females. The dataset is methodically divided into training and testing sets, featuring 108 participants in the training set and 48 participants in the testing set. Both sets are meticulously balanced for age, gender, and disease condition.

2.2. Feature Engineering

In our work, we propose 11 new features for this task. Table 1 summarizes these 11 features with their corresponding description. In this section, We will introduce the definition and extraction process for each feature in the following section.

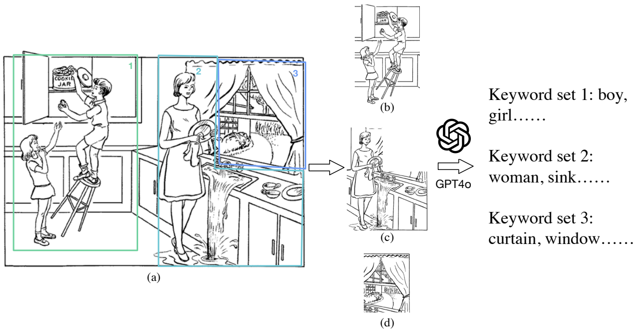
1
2
3
(a)
(b)
(c)
(d)
GPT4o
Keyword set 1: boy, girl……
Keyword set 2: woman, sink……
Keyword set 3: curtain, window……

Figure 1: *Schematic diagram of topic keyword generation. (a) shows the Cookie Theft picture and how we segment the picture. (b), (c) and (d) show the sub-pictures we crop. Three subpictures are sent to LLM with instructions for generating keywords.*

2.2.1. Topic Related Features

A critical aspect of the Cookie Theft picture description task is to evaluate the comprehensiveness of a subject's description in terms of picture content and topics. As illustrated in Fig. 1(a), the picture is segmented into three distinct topics: the boy and girl taking cookies, the mother and the water sink, and the window with curtain. We then segment the picture into three subimages based on the identified topics and send these cropped images to the multimodal LLM[1], leveraging its visual processing capabilities to generate relevant keywords, as illustrated in Fig. 1. For each sub-picture, we conduct 50 iterations of keyword generation and aggregate the results to ensure comprehensive content coverage. We manually check each iteration's output to prevent any potential hallucination and each step of the generation is trackable. With these three sets of keywords, we calculate the keyword hit rate within each topic to quantify the degree of detail in the descriptions.

Although the topic keyword method effectively evaluates how a subject describes local parts of the picture, it lacks information from the global picture, such as the connections between topics. To assess the description coverage of the entire picture, we can adopt metrics from the image captioning task (e.g., BLEU and METEOR scores). These metrics quantify the degree of match between the description and the 'golden standard,' making them suitable for our needs. To generate the 'golden standard,' we input the entire Cookie Theft picture into the multimodal LLM and leverage its visual analysis capability to produce detailed verbal descriptions of the picture. We performed 15 iterations of this generation, considering the 15 responses as the 'golden standards.' The averaged BLEU and METEOR scores over these 15 'golden standards' serve as the final score. Accordingly, we propose five new features: BLEU-1, BLEU-2, BLEU-3, and BLEU-4, calculated using different n-gram schemes, and the METEOR score. Table 2 shows the prompt template we used and example response from the LLM.

2.2.2. TF-IDF Related Features

Borrowing the idea of TF-IDF from IR [19], we propose a new feature called TF-IDF Score for this task. Let's consider each

[1] The LLM used in this work is the latest version of GPT-4o, 2024-05-13.

Table 2: *Prompt template and the response example for topic keywords and description generation.*

Instruction for topic keyword generation:
Imagine you are an expert on cognitive assessment using Cookie Theft picture description task. You have the knowledge of the Cookie Theft picture and the key point to assess the AD. Now I will provide you with a sub-picture of the Cookie Theft picture, please give me some key content words related to that part. These words should be helpful for people to distinguish AD patients that the missing of the words may indicate potential cognitive impairment. Please only give the keywords list separated by comma without any further explanation.

Response example:
boy, girl, cookie, jar, stool, reaching, cabinet...

Instruction for description generation:
This is the picture of the Cookie Theft description task which is widely used for cognitive assessment. Now imagine that you are an elderly people with healthy cognitive state. Please give me a verbal description of this picture to cover as much content as possible in the picture.

Response example:
"Well, the boy has climbed up on a stool to get to the cookie jar in the cupboard. He's giving a cookie to the girl who's eagerly waiting for it. The mother is busy washing dishes at the sink but hasn't noticed that the water is overflowing onto the floor. Outside the window, there are trees and another house, so it's probably a nice, sunny day."

subject's transcript as a document d and the training document set as D. Each document d has a corresponding label (HC or AD). We denote the HC document set as d_{HC}, AD document set as d_{AD}, where $d_{HC}, d_{AD} \in D$.

Then we obtain the TF of the term t in document d by calculating the the number of times t appears in d divided by the total number of terms in d i.e.

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

where: $f_{t,d}$ is the occurrences of term t in document d. $\sum_{t' \in d f_{t',d}}$ is the total number of terms in d.

The inverse document frequency (IDF) is a measure of how much information the word provides, that is, if it is common or rare across all subjects. It is defined as:

$$\text{IDF}(t) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{2}$$

where: $|D|$ is the total number of documents in the training set D. $|\{d \in D : t \in d\}|$ is the number of documents in which the term t appears (i.e., the document frequency of t).

Then the TF-IDF weight for a term t in d is the product of its TF and IDF:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{3}$$

Then we construct the TF-IDF vector for each document $d \in D$. Let T be the set of unique term from the document set. The i-th value of the vectors coresponds to the i-th term in T If the i-th term is in the document the value would be its TF-IDF, else 0, i.e.

$$\mathbf{v}_d = \left[\begin{cases} \text{TF-IDF}(t_i, d) & \text{if } t_i \in d \\ 0 & \text{if } t_i \notin d \end{cases}\right]_{i=1}^{|T|} \tag{4}$$

where $|T|$ is the total number of unique terms.

Then the HC reference vector is calculated by averaging the TF-IDF vectors of all documents in the document set d_{HC}:

$$\mathbf{v}_{\text{HC}} = \frac{1}{|d_{HC}|} \sum_{d \in d_{HC}} \mathbf{v}_d \tag{5}$$

Similarly, by replacing d_{HC} by d_{AD} and performing same calculation with Equation (5), we obtain the the AD reference vector $\mathbf{v}_{\text{AD}}$

Lastly, two similarity features of d are calculated as follows:

$$\text{TF-IDF similarity HC}(d) = \text{CosSimilarity}(\mathbf{v}_d, \mathbf{v}_{\text{HC}}) \tag{6}$$

$$\text{TF-IDF similarity AD}(d) = \text{CosSimilarity}(\mathbf{v}_d, \mathbf{v}_{\text{AD}}) \tag{7}$$

In our analysis, we observed that certain key terms, such as 'window' (objects) and 'overflow' (actions), may be overlooked by some AD subjects for various reasons. To quantify this observation, we propose using the keyword hit rate as a feature. To select appropriate keywords, we choose the top 30 terms that have the highest values in $\mathbf{v}_{\text{HC}}$ as keywords. The TF-IDF keyword hit rate is then determined by dividing the number of mentioned keywords by the total number of keywords (30).

We also add four linguistic features that are not included in the previous research into our feature set: averaged parse tree depth, filler pause number, filler pauses ratio and word error rate[2].

3. Experiment

3.1. Experiment settings

We constructed classifiers based on three widely recognized methods: Random Forest (RF) and XGBoost. To ensure optimal performance, we employed Bayesian Optimization [20] to determine the appropriate set of hyperparameters for each model. The hyperparameters identified through this process were kept fixed across all settings, ensuring consistency and robustness in our evaluation. In our work, we follow the standard train test split of ADReSS dataset.

We used three different feature sets in our experiment. The first set comprised 40 traditional linguistic features proposed by [14]. The second set included our 15 new features and the third set combined the linguistic features and the new features together.

3.2. Results

Table 3 presents the overall experimental results of this work. The bold numbers indicate the highest scores within the model and the red numbers represent the best score among all. It is evident that our new features consistently outperform traditional linguistic features. These results highlight the effectiveness of the new features. We achieve the best performance of 85.4% accuracy, this result is comparable to previous research which uses a fine-tuned BERT model and nearly ten times the number of feature dimensions. Furthermore, the new features are more intuitive for humans to understand and are closely related to the

[2] We use Whisper-large-v3 as the ASR system for WER evalaution

Model	Feature	ACC(%)	PRE(%)	REC(%)	F1(%)
RF	Linguistic Features (40)	75.0	87.5	58.3	70.0
	New Features (15)	**85.4**	**87.0**	**83.3**	**85.1**
	All Features (55)	80.6	89.3	69.6	78.2
XGBoost	Linguistic Features (40)	72.9	92.3	50.0	64.9
	New Features (15)	**83.3**	86.4	**79.2**	**82.6**
	All Features (55)	79.2	**93.8**	62.5	75.0

Table 3: *The overall results of 3 feature sets on Random Forest and XGBoost. ACC: accuracy, PRE: precision, REC: recall, F1: F1 score.*

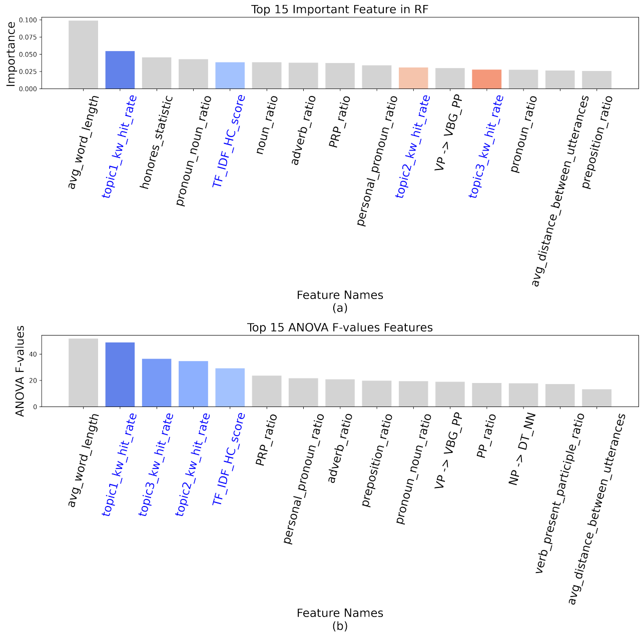

Figure 2: *Top 15 Important Features in Random Forest and ANOVA F-values. The charts display the top 15 features ranked by their importance in the Random Forest model (a) and by their ANOVA F-values (b). Newly proposed features are highlighted in blue names, with their corresponding bars in distinct colors.*

Cookie Theft picture description task, thereby enhancing the explainability of spoken language-based AD detection.

We found that combining the linguistic features with the new features may worsen performance compared to using only the new features which suggest the importance of applying feature selection to the linguistic features for filtering some noisy features.

4. Discussion

4.1. Feature Importance and ANOVA F-values

To further substantiate the effectiveness of our features, we extracted the feature importance from the RF model. Fig. 2(a) presents the top 15 important features in the RF. Notably, four of our new features ranking in the top fifteen. Additionally, we plotted the top 15 features with the highest ANOVA F-values. Fig. 2(b) indicates that our new features are highly relevant to AD detection as four of them ranking in the top five.

Among our proposed features, we identified that topic 1 keyword hit rate and the TF-IDF similarity HC are particularly effective, consistently ranking in the top five for both the importance of the RF feature and the ANOVA F values. Furthermore, other topic keyword features also demonstrated high effectiveness.

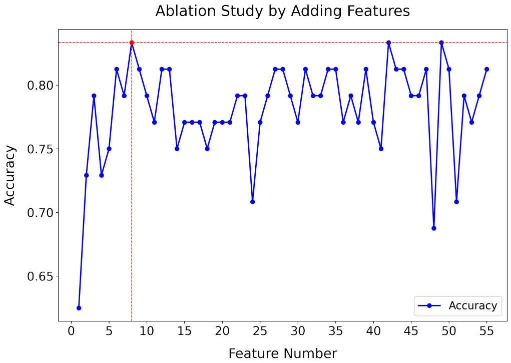

Figure 3: *Ablation study accuracy result. The x-axis indicates the number of feature we added into the experiment.*

4.2. Ablation Study

We also conducted an ablation study to dive deeper for the inverstigation. We incrementally added features based on their ANOVA F-values and assessed their impact on the accuracy of AD detection tasks. Fig. 3 illustrates the results of this study. A notable increase in accuracy is observed between feature numbers 1 and 8; however, accuracy declines and fluctuates as the number of features increases. The optimal result was achieved by incorporating four traditional linguistic features and four new features, however it does not outperform only using new features, hence the feature selection based on ANOVA F-values may not be suitable. Determining a more effective feature selection to better integrate traditional features with our new features will be a focal point for future research.

5. Conclusion

In conclusion, we have proposed a compact set of features that are both more explainable and more effective for AD detection. We introduced the concept of leveraging TF-IDF alongside advanced LLMs' visual processing ability to generate useful features. Our experiments demonstrate that our new features outperform the traditional features and achieve a competitive performance with high dimensional efficiency and interpretability.

6. References

[1] Z. Breijyeh and R. Karaman, "Comprehensive review on alzheimer's disease: causes and treatment," *Molecules*, vol. 25, no. 24, p. 5789, 2020.

[2] A. Wimo, B. Winblad, and L. Jönsson, "An estimate of the total worldwide societal costs of dementia in 2005," *Alzheimer's & Dementia*, vol. 3, no. 2, pp. 81–91, 2007.

[3] V. L, R. SH, R. M, P. M, L. J, C. M, and L. G., "Review of brief cognitive tests for patients with suspected dementia," *Int Psychogeriatr. 2014 Aug;26(8):1247-62. doi:*, vol. 10., 2014.

[4] Z. Arvanitakis, R. C. Shah, and D. A. Bennett, "Diagnosis and management of dementia," *Jama*, vol. 322, no. 16, pp. 1589–1599, 2019.

[5] L. C. Kourtis, O. B. Regele, J. M. Wright, and G. B. Jones, "Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity," *NPJ digital medicine*, vol. 2, no. 1, p. 9, 2019.

[6] S. G, H. I, V. V, K. J, and P. M., "Speaking in alzheimer's disease," *is That an Early Sign? Importance of Changes in Language Abilities in Alzheimer's Disease. Front Aging Neurosci. 2015 Oct 20;7:195. doi:*, vol. 10., 2015.

[7] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 674–681.

[8] L. Calzà, G. Gagliardi, R. R. Favretti, and F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, p. 101113, 2021.

[9] T. Melistas, L. Kapelonis, N. Antoniou, P. Mitseas, D. Sgouropoulos, T. Giannakopoulos, A. Katsamanis, S. Narayanan, and N. Demokritos, "Cross-lingual features for alzheimer's dementia detection from speech."

[10] L. Wagner, M. Zusag, and T. Bloder, "Careful whisper–leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification," *arXiv preprint arXiv:2308.01327*, 2023.

[11] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for alzheimer's disease detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6423–6427.

[12] Y. Wang, J. Deng, T. Wang, B. Zheng, S. Hu, X. Liu, and H. Meng, "Exploiting prompt learning with pre-trained language models for alzheimer's disease detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] J. Li, K. Song, J. Li, B. Zheng, D. Li, X. Wu, X. Liu, and H. Meng, "Leveraging pretrained representations with task-related keywords for alzheimer's disease detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] J. Heitz, G. Schneider, and N. Langer, "The influence of automatic speech recognition on linguistic features and automatic alzheimer's disease detection from spontaneous speech," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 15 955–15 969.

[15] C. Wang, S. Liu, A. Li, and J. Liu, "Text dialogue analysis for primary screening of mild cognitive impairment: Development and validation study," *Journal of Medical Internet Research*, vol. 25, p. e51501, 2023.

[16] Z. Wang, R. Li, B. Dong, J. Wang, X. Li, N. Liu, C. Mao, W. Zhang, L. Dong, J. Gao *et al.*, "Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today," *arXiv preprint arXiv:2306.01499*, 2023.

[17] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech," *Frontiers in computer science*, vol. 3, p. 780169, 2021.

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[20] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, pp. 455–492, 1998.