Health CLIP: Depression Rate Prediction Using Health Related Features in Satellite and Street View Images

Tianjian Ouyang
Department of Electronic Engineering
Tsinghua University
Beijing, China
oytj22@mails.tsinghua.edu.cn

Xin Zhang
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China

Zhenyu Han
Department of Electronic Engineering
Tsinghua University
Beijing, China

Yu Shang
Department of Electronic Engineering
Tsinghua University
Beijing, China

Yong Li*
Department of Electronic Engineering
Tsinghua University
Beijing, China

ABSTRACT

Mental health is a state of mental well-being that enables people to cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community. It has intrinsic and instrumental value and is integral to our well-being[14], and its correlation with environmental factors has been a subject of growing interest. As the pressure of society keeps growing, depression has become a severe problem in modern cities, and finding a way to estimate depression rate is of significance to relieve the problem. In this study, we introduce a Contrastive Language-Image Pretraining (CLIP) based novel approach to predict mental health indicators, especially depression rate, through satellite and street view images. Our methodology uses state-of-the-art Multimodal Large Language Model (MLLM), GPT4-vision, to generate health related captions for satellite and street view images, then we use the generated image-text pairs to fine-tune the CLIP model, making its image encoder extract health related features such as green spaces. The fine-tuning process is employed to bridge the semantic gap between textual descriptions and visual representations, enabling a comprehensive analysis of geo-tagged images. Consequently, our methodology achieves a notable R^2 value of 0.565 on prediction of depression rate in New York City with the combination of satellite and street view images. The successful deployment of Health CLIP in a real-world scenario underscores the practical applicability of our approach.

CCS CONCEPTS

• **Applied computing → Health informatics**; • **Computing methodologies → Machine learning**.

KEYWORDS

Satellite Imagery, Street View Imagery, Mental Health, Depression, Machine Learning, Contrastive Learning

ACM Reference Format:
Tianjian Ouyang, Xin Zhang, Zhenyu Han, Yu Shang, and Yong Li. 2024. Health CLIP: Depression Rate Prediction Using Health Related Features in Satellite and Street View Images. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651451

1 INTRODUCTION

Depression stands out as one of the most widespread and incapacitating mental conditions globally[1], and the escalating prevalence of depression has emerged as a profound and pressing concern[12], posing significant challenges to public health and well-being. Depression in the United States affects over 18 million adults annually, constituting approximately one in ten individuals, and stands as the primary cause of disability for those aged 15-44[7], contributing to over 41,000 suicides each year, a staggering figure that surpasses the 16,000 lives claimed by homicide according to 2013 CDC statistics[2]. Thus, how to accurately predict depression and identify the environmental factors that contribute to depression have become crucial problems in mental health research.

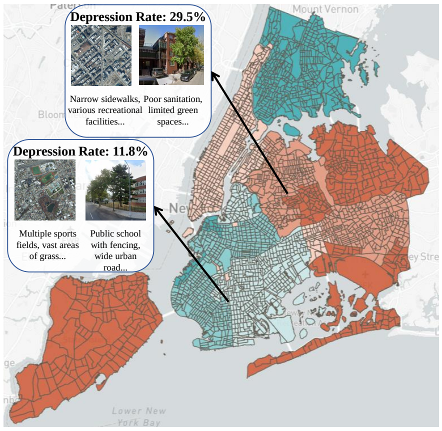

Depression Rate: 29.5%
Narrow sidewalks, various recreational facilities...
Poor sanitation, limited green spaces...
Depression Rate: 11.8%
Multiple sports fields, vast areas of grass...
Public school with fencing, wide urban road...
Mount Vernon
Lower New York Bay

Figure 1: Using satellite and street view image-text pairs to predict depression rate in New York City.

[*] Corresponding author (liyong07@tsinghua.edu.cn).

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05…$15.00
https://doi.org/10.1145/3589335.3651451

The sophisticated nature of this mental health issue brings great difficulties to its analysis and prediction. The fast developed machine learning methods like Random Forest(RF) and Deep Neural Networks(DNN) have been used to reveal the occurrence of mental illness[11][4]. The advent of large language models (LLMs) in recent years also opened new avenues for mental health problems[15], including the prediction and mitigation of depression[10]. However, these kinds of image-only or text-only methods have similar limitations: image or text, as a single modality, is hard to capture the complex health related features we want.

Thus, we integrate multi-source visual information including satellite and street view imagery, and coupled them with the capabilities of large language models, offering a novel approach to tackle the intricate landscape of mental health concerns. Specifically, we focus on predicting depression rates in the bustling urban environment of the New York City (Figure 1), leveraging a multimodal LLM framework that combines the power of imagery analysis and linguistic understanding.

2 BACKGROUND

The burgeoning field of health informatics has witnessed a paradigm shift with the integration of advanced technologies, notably satellite and street view imagery, into the exploration of mental health outcomes. The intricate relationship between environmental factors and mental well-being has spurred a quest for innovative methodologies capable of unraveling the complexities inherent in predicting depression rates. This paper introduces Health CLIP, a pioneering approach that combines multimodal analysis with machine learning techniques to predict depression rates in urban landscapes.

2.1 CLIP Model

Recent advancements in machine learning have ushered in transformative possibilities for mental health research, with models like CLIP (Contrastive Language-Image Pretraining) standing out as exemplars of multimodal capabilities. CLIP, developed by Radford [5], is a powerful model that combines vision and language understanding, enabling it to learn representations from diverse datasets. The model's ability to simultaneously process images and text has been applied to a myriad of tasks, from natural language understanding to image classification, setting a precedent for multimodal applications in various domains.

2.2 Multimodal Large Language Model

In parallel, the landscape of large language models (LLMs) has expanded, embracing multimodal architectures that fuse textual and visual information. Multimodal LLMs, such as the GPT-4 developed by OpenAI, integrate diverse data types, offering a holistic understanding of complex information. The synthesis of language and visual inputs enables these models to capture nuanced relationships within the data, making them particularly relevant for addressing intricate challenges like mental health prediction.

3 METHODS

3.1 Model Design

The escalation of mental health issues, particularly depression, on a global scale has underscored the urgency of devising comprehensive strategies for early diagnosis and intervention. Recognizing the spatial dimensions embedded in mental health disparities, Health CLIP seeks to exploit the correlations between machine learning algorithms and the wealth of information contained in satellite and street view imagery. The approach is encapsulated by the following equations:

$$R = MLP(Sat\_Feature, Str\_Feature)$$

where R stands for depression rate, the proportion of depressed people among adults over 18 years. $Sat\_Feature$ and $Str\_Feature$ are the feature embeddings extracted from satellite imagery and street view imagery through the image encoders of the fine-tuned CLIP model.

Health CLIP then uses fine-tuned image encoders to capture the spatial features needed for mental health prediction:

$$Sat\_Feature = Sat\_CLIP(Satellite\ Images)$$
$$Str\_Feature = Str\_CLIP(Street\ View\ Images)$$

where Sat_CLIP and Str_CLIP are the image encoders fine-tuned on the satellite and street view image-text pairs.

The main goal of the CLIP fine-tuning is to make the image features extracted from image encoder be more connected to the health related texts generated by GPT4-vision. The training process would make the confusion matrix of images and texts become diagonally larger, showing the increase in similarity.

The Health CLIP framework is depicted in Figure 2. The model comprises two components, one for handling satellite imagery and another for street view imagery. Initially, both satellite and street view images were input into GPT4-Vision to generate health-related captions. Then two CLIP models were separately fine-tuned using the image-text pairs generated in the previous step. Finally, the feature embeddings derived from the image encoders were integrated using a single MLP to predict depression rate.

3.2 Prompt Design

Prompts can largely determine the quality of the generated texts. The prompt design should be specific, descriptive and as detailed as possible, and articulate the desired output format through examples. Our prompt pushes GPT4-vision to capture fundamental facilities, green space, and landuse status in satellite images, and environmental cues that may impact mental health in street view images. The difference in prompt is because satellite image contains more high-level information, while street view image contains more low-level ones.

3.3 Training Loss

For the training process of CLIP model, contrastive loss[3] was adopted as criterion function:

$$L(x, y) = \frac{1}{2}\left(y \cdot D(W^T x) + (1 - y) \cdot \max(margin - D(W^T x), 0)\right)^2$$

where $L(x, y)$ is the contrastive loss, x is the input, y is the label (0 or 1), $D(\cdot)$ is the distance metric, W is the weight matrix, and $margin$ is a hyperparameter representing the minimum distance between positive and negative pairs. If the samples are similar ($y = 1$), it encourages the model to minimize the distance between the representations of similar examples. If the samples are dissimilar ($y = 0$),

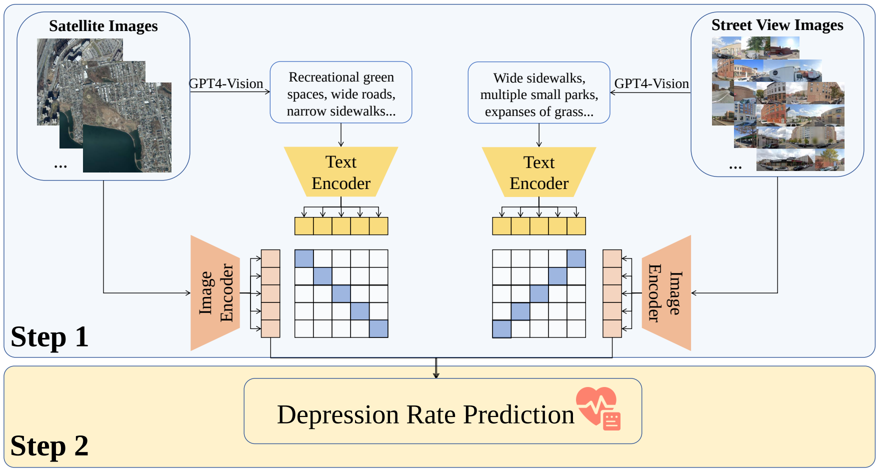

Satellite Images
...
GPT4-Vision
Recreational green spaces, wide roads, narrow sidewalks...
Text Encoder
Image Encoder
Wide sidewalks, multiple small parks, expanses of grass...
GPT4-Vision
Street View Images
...
Text Encoder
Image Encoder
Step 1
Depression Rate Prediction
Step 2

Figure 2: The Framework of Health CLIP. Satellite and street view images were fed into GPT4-Vision to generate health related captions. Then two CLIP models were fine-tuned separately using the image-text pairs generated above. At last, the feature embeddings come from image encoders were integrated using a single MLP to predict depression rate.

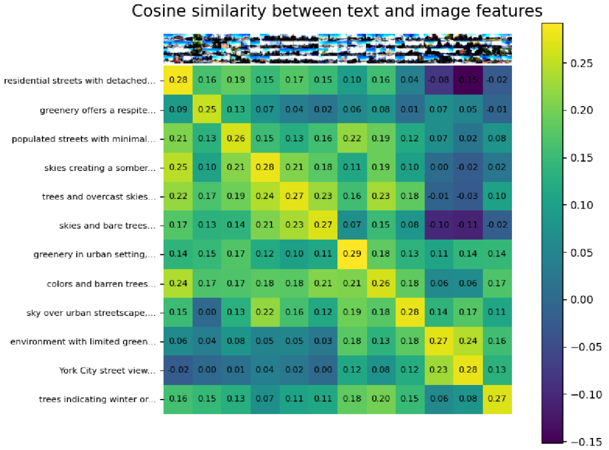


Figure 3: Confusion Matrix of Street View Images and Texts

it penalizes the model if the distance between the representations of dissimilar examples is less than the specified margin.

If the CLIP model is trained properly, the confusion matrix of image-text pairs should look like figure 3. The corresponding image and text would have a higher similarity score.

4 EXPERIMENTS

4.1 Datasets

We employed satellite and street view images of 1667 census tracts in New York City as our dataset, with the corresponding depression rate serving as ground truth labels. The 256×256 pixel satellite images with about 0.5m-resolution are obtained from ArcGIS, which are further merged along irregular region boundaries and concatenated into 512×512-pixel as the input satellite images. The 400×300-pixel street view images are obtained from Google API. In consideration of the limited information in single street view image, we combined every 12 street view images to a big image of size 1200 × 1200 to make it more representative. As for the depression rate data, we collect from U.S. goverment official health data website [6]. The dataset is split into training, validation, and test sets, with the proportions being 70%, 15%, and 15%, respectively.

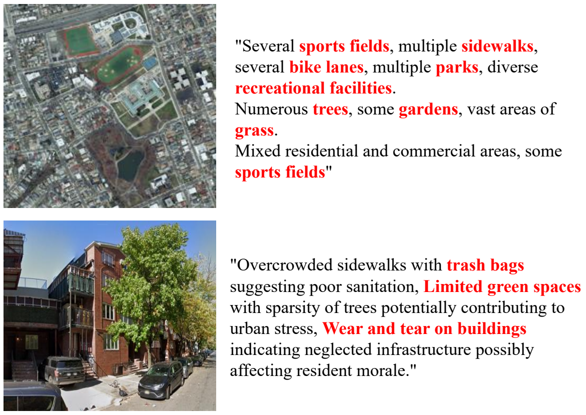
4.2 Setup and Baselines

We adopted the widely used rooted mean squared error(RMSE) and coefficient of determination(R^2) as evaluation metrics. In our experiment, we employed CLIP alongside different image encoders, specifically ResNet-50 and ViT-B/32. We select the AdamW optimizer,which integrate weight decay directly into the optimization process, and we also incorporate warm-up strategy for the first 10000 steps. The mini-batch size is fixed at 256, and the complete training process spans 50 epochs. In the depression prediction step, we conduct a grid search for optimal values of the learning rate, weight decay, and dropout.

In order to show the effectiveness of our Health CLIP, we use models below as comparision. The satellite imagery-based baseline includes: **RemoteCLIP**[9]: A vision-language foundation model for remote sensing, on the basis of CLIP model.
The street view imagery-based baselines includes: **Urban2Vec**[13]: An street view feature extraction model, assuming neighbourhoods have similar meanings. **SceneParse**[8]. A segmentation model trained on ADE20K dataset, calculate the percentage of each object in the image.

4.3 Results

We found that satellite imagery got the best performance using OpenCLIP model with ResNet-50 as its image encoder, while street view imagery reached its peak with ViT-B/32. For all models tested on satellite imagery, ResNet-50 fine-tuned with 20 epochs achieved the best R^2 value, while for street view imagery, ViT-B/32 fine-tuned with 10 epochs performed the best. For baselines, RemoteCLIP"ViT-B/32" reached the highest score on satellite part, and SceneParse performed the best on street view images. The best R^2 came from the combination of satellite and street view images, which is 0.565, meanwhile the RMSE is 1.558, indicating an accurate prediction



"Several **sports fields**, multiple **sidewalks**, several **bike lanes**, multiple **parks**, diverse **recreational facilities**.
Numerous **trees**, some **gardens**, vast areas of **grass**.
Mixed residential and commercial areas, some **sports fields**"

"Overcrowded sidewalks with **trash bags** suggesting poor sanitation, **Limited green spaces** with sparsity of trees potentially contributing to urban stress, **Wear and tear on buildings** indicating neglected infrastructure possibly affecting resident morale."

Figure 4: Alpha's Satellite Image-Text Pair, Beta's Street View Image-Text Pair

of depression rate. However, it is worth mentioning that, the fine-tuning process of CLIP model needs to be careful about the risk of overfitting, or the performance will drop dramatically.

Table 1: R^2 and RMSE of Satellite Images

Method Type	R^2	RMSE
OpenCLIP"RN50"	0.527	1.788
OpenCLIP"ViT-B/32"	0.393	2.025
RemoteCLIP"RN50"	0.043	2.545
RemoteCLIP"ViT-B/32"	0.392	2.027
OpenCLIP "Satellite + Street View"	**0.565**	**1.558**

Table 2: R^2 and RMSE of Street View Images

Method Type	R^2	RMSE
OpenCLIP"RN50"	0.248	2.140
OpenCLIP"ViT-B/32"	0.454	1.824
Urban2vec	0.027	2.331
SceneParse	0.365	1.884
OpenCLIP "Satellite + Street View"	**0.565**	**1.558**

4.4 Case Study

Census tract Alpha at Queens has a low depression rate of 11.8%, while census tract Beta at Brooklyn has a high depression rate of 29.5%. The satellite image-text pair of Alpha and street view image-text pair of Beta are shown in figure 4. Texts from Alpha contains lots of positive signs on mental wellness, including sports fields, gardens etc. These positive texts will help Health CLIP extract these positive features from its image encoder. On the contrary, the negative texts describing Beta's street view features such as trash bags, will make its image encoder extract depression related features, bringing out a prediction with higher depression rate.

5 CONCLUSION

In this paper, we introduced Health CLIP, a pioneering approach that integrates health-related features from satellite and street view imagery for predicting depression rates in urban environments, with a specific focus on New York City. Health CLIP positions itself at the intersection of machine learning and environmental determinants of health. We aim to capture the complex interplay between satellite and street view imagery features for enhanced predictive accuracy.

As we move forward, Health CLIP contributes to the evolving landscape of mental health prediction by presenting a comprehensive and innovative framework that addresses the intricate relationships between urban environments and mental health outcomes. By building upon the foundations laid by related works, we strive to advance the understanding of depression, offering valuable insights for public health interventions and urban planning strategies aimed at fostering mental well-being in city dwellers. The multidimensional and multimodal approach of Health CLIP represents a solid step toward a more nuanced and accurate understanding of mental health in the dynamic urban landscape.

ACKNOWLEDGMENTS

This work was supported National Key Research and Development Program of China under grant 2020YFA0711403, National Natural Science Foundation of China under grant U22B2057, National Natural Science Foundation of China under grant 62171260.

REFERENCES

[1] Amanda J Baxter, George Patton, Kate M Scott, Louisa Degenhardt, and Harvey A Whiteford. 2013. Global epidemiology of mental disorders: what are we missing? *PloS one* 8, 6 (2013), e65514.

[2] Centers for Disease Control and Prevention (CDC). 2013. *Web-based Injury Statistics Query and Reporting System (WISQARS).* National Center for Injury Prevention and Control, CDC (producer).

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

[4] Dominic B Dwyer, Peter Falkai, and Nikolaos Koutsouleris. 2018. Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology* 14 (2018), 91–118.

[5] Alec Radford et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020

[6] HealthData.gov. [n. d.]. PLACES: Local Data for Better Health - Census Tract Data. https://healthdata.gov/dataset/PLACES-Local-Data-for-Better-Health-Census-Tract-D/jpdw-4rwm/about_data.

[7] Ronald C Kessler, Wai Tat Chiu, Olga Demler, and Ellen E Walters. 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 617–627.

[8] Jihyeon Lee, Dylan Grosz, Burak Uzkent, Sicheng Zeng, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Predicting livelihood indicators from community-generated street-level imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 268–276.

[9] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. 2023. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. arXiv:2306.11029 [cs.CV]

[10] Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, Diagnose and Chat: Towards Explainable and Interactive LLMs-Augmented Depression Detection in Social Media. arXiv:2305.05138 [cs.CL]

[11] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. EPJ Data Science, 6 (15), 1-12.

[12] Theo et al. Vos. 2016. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The lancet* 388, 10053 (2016), 1545–1602.

[13] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (Apr. 2020), 1013–1020.

[14] World Health Organization. [n. d.]. *Mental health.* https://www.who.int/health-topics/mental-health#tab=tab_1

[15] Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385* (2023).