KDD 2023 International Workshop on Data Science for Social Good (DSSG-23)

Serina Chang, Ayan Mukhopadhyay, Anwar Said,
Aparna Taneja, Amulya Yadav

This workshop brought together researchers and practitioners across different strands of data science research and a wide range of important real-world application domains. The objective was to share the current state of research and practice, explore future work directions, and create collaboration opportunities. In addition, the workshop emphasized highlighting data science approaches for tackling the United Nations Sustainable Development Goals. We believe that data science research has an important role to play in providing unique insights about critical challenges faced by marginalized communities around the world. Workshop participants included both data science researchers and practitioners implementing machine learning and artificial intelligence for social impact, as well as domain experts interested in engaging with the SIGKDD community.

Mitigating Bias in Conversations: A Hate Speech Classifier and Debiaser with Prompts

Shaina Raza
Vector Institute of Artificial Intelligence
Toronto, ON, Canada
shaina.raza@vectorintitute.ai

Chen Ding
Toronto Metropolitan University
Toronto, ON, Canada
cding@torontomu.ca

Deval Pandya
Vector Institute of Artificial Intelligence
Toronto, ON, Canada
deval.pandya@vectorinstitute.ai

ABSTRACT

Discriminatory language and biases are often present in hate speech during conversations, which usually lead to negative impacts on targeted groups such as those based on race, gender, and religion. To tackle this issue, we propose an approach that involves a two-step process: first, detecting hate speech using a classifier, and then utilizing a debiasing component that generates less biased or unbiased alternatives through prompts. We evaluated our approach on a benchmark dataset and observed reduction in negativity due to hate speech comments. The proposed method contributes to the ongoing efforts to reduce biases in online discourse and promote a more inclusive and fair environment for communication.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Information retrieval**.

KEYWORDS

Language models, bias detection, fairness, language generation

ACM Reference Format:
Shaina Raza, Chen Ding, and Deval Pandya. 2023. Mitigating Bias in Conversations: A Hate Speech Classifier and Debiaser with Prompts. In *Proceedings of In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the era of social media and online platforms, communication and idea exchange have reached at its peak. Despite many benefits, these platforms also facilitate the spread of hate speech and offensive language. Hate speech often contains biases, perpetuating stereotypes and discriminatory language, which exacerbates the negative impact of such content on different targeted groups (based on race, gender, religion) [9]. Addressing these biases is a crucial step towards developing unbiased text processing systems and fostering healthy online interactions.

KDD '23, August 6-10, 2023, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

In this paper, we propose a debiasing technique that leverages language generation and in-context prompting [13] to minimize the influence of lexical biases. A prompt is an instruction usually consisting of a few words or sentences that provide context or constraints for the model to follow . The method works by first detecting hate speech using a classifier, then employing a debiasing component that generates less biased alternatives through incorporating context-aware prompts designed to reduce the presence of biased language patterns.

We evaluate our approach on benchmark dataset and demonstrate its effectiveness in debiasing hate speech texts. The results show a classifier accuracy of 95% and debiasing accuracy of 89%, along with a notable reduction in negative sentiment within hate speech comments. This method contributes to the ongoing efforts to reduce biases in online discourses.

2 RELATED WORK

Bias in language models and embeddings is a broad and subjective topic [19] . Research has identified gender bias in popular embeddings such as GloVe and Word2Vec [5] and quantified biases using the word embedding association test (WEAT) [6] [15]. Efforts have been made to reduce biases in Transformer-based language models like BERT and GPT-3 [4] and in conversational AI systems [3].

Hate speech refers to the use of derogatory, abusive or threatening language towards individuals or groups based on their race, ethnicity, gender, religion, sexual orientation or any other characteristic. Several studies have focused on developing machine learning models for detecting hate speech in text [7]. One work [20] presents a dataset, HOLISTIC BIAS, which consists of nearly 600 descriptor terms across 13 different demographic axes. The work demonstrates that this dataset is highly effective for measuring previously unmeasurable biases in token likelihoods and generations from language models, as well as in an offensiveness classifier.

Another work [12] quantifies sentiment bias through individual and group fairness metrics [19] and proposes embedding and sentiment prediction-derived regularization on the language model's latent representations. The role of individual neurons and attention heads in mediating gender bias across three datasets designed to gauge a model's sensitivity to gender bias are also studied [21].

A related paper [14] describes metrics for measuring political bias in language generation and proposes a reinforcement learning framework for mitigating political biases in generated text. StereoSet [16], a large-scale natural English dataset to measure stereotypical biases in four domains: gender, profession, race, and religion, is presented and it is shown that popular models like BERT, GPT-2, RoBERTa, and XLnet exhibit strong stereotypical biases.

A novel approach to mitigate gender disparity in text generation by learning a fair model during knowledge distillation is presented in a study [11] and two modifications based on counterfactual role reversal are proposed—modifying teacher probabilities and augmenting the training set. A related work [1] shows that GPT-3, a state-of-the-art contextual language model, captures persistent Muslim-violence bias, demonstrating that it appears consistently and creatively in different uses of the model. These biases become even severe compared to biases about other religious groups.

Prompt engineering has recently emerged as a promising approach to mitigate biases in language models [13] . In the context of hate speech detection, few-shot learning [2] can be useful in scenarios where there is limited labeled data available for a particular language or dialect. Recent studies [2] have shown promising results in using few-shot learning for hate speech detection, which is also a motivation. In this work, we use the OPT based model for debiasing but we introduce fairness aware prompts to achieve the goal of debiasing the texts.

3 METHODOLOGY

3.1 Hate speech classifier

We utilize the BERT [8] for building an efficient hate speech detection. The input to the BERT encoder consists of tokenized text sequences with special tokens [CLS] and [SEP] for classification and separation. The output from the BERT encoder is passed through a SoftMax layer to provide a probability distribution for the text being classified as hate speech or non-hate speech. The hate speech classifier is trained on a labeled dataset using binary cross-entropy loss as the objective function, which aims to minimize the difference between the predicted probability distribution and the true binary labels, encouraging the model to accurately classify hate speech and non-hate speech comments.

3.2 Debiasing model

We use the pre-trained OPT (OpenAI's Pre-trained Transformer) [18] model for debiasing hate speech classification. We employ the few-shot learning with prompts to further refine the debiasing process. Our approach to debiasing through prompts is inspired by similar works such as [22] , however, we make our own changes through task-specific examples.

To incorporate the debiasing prompts into the OPT model, we provide the prompt, the input text, resulting in a new input sequence. The OPT model processes this combined sequence and generates the debiased output based on the contextualized embeddings of both the prompt and the input text. We adjust the temperature of the OPT model during inference to control the level of randomness and diversity in the generated text samples. This helps to mitigate biases while preserving the overall meaning of the input text.

3.3 Pipeline

Our pipeline stacks both the hate speech classification and debiasing models as follows: (i) pass the input text through the hate speech classifier to obtain the predicted probability of it being hateful or non-hateful, and (ii) pass the hateful text through the debiasing model using few-shot learning with prompts to generate the debiased text during the language generation task. This two-stage approach first classifies the hate content in the input text, and if it is deemed hateful by the model, then it debiases it. We show our pipeline approach in Figure 1.

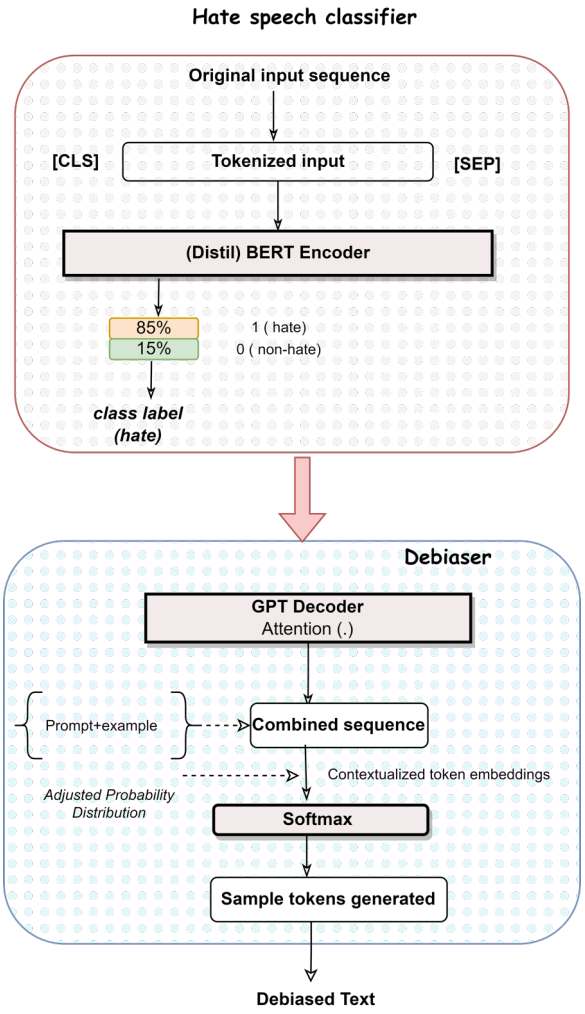


Figure 1: Proposed framework consisting of a classifier for hate speech and debiaser with prompt and examples.

4 EXPERIMENT AND RESULTS

4.1 Dataset

In this study, we utilize the Hate Speech Dataset from a white supremacy online community [10] . The dataset consists of a diverse range of text samples, including both overt and subtle instances of hate speech, posing significant challenges for automated classification and debiasing models. The dataset contains a total of 10,568 sentences, classified as conveying hate speech or not. The size of dataset, along with the average sentence length, word count, and vocabulary size for each class.

Table 1: Performance comparison of different methods on the hate speech classification and debiasing tasks.

Method	F1 Score ± SD
Classifier performance	
Rule-based	0.60 ± 0.03
SVM	0.70 ± 0.02
BERT-d	0.90 ± 0.04
RoBERTa-HS	0.92 ± 0.01
Proposed	0.95 ± 0.05
Debiasing performance	
Debiased with zero-shot learning	0.86 ± 0.02
Debiased with few-shot learning (5)	0.88 ± 0.03
Debiased with few-shot learning (10)	0.89 ± 0.02

- HATE: Sentences: 1,119, Sentence length: 20.39 ± 9.46, Word count: 24,867, Vocabulary: 4,148.
- NOHATE: Sentences: 8,537, Sentence length: 15.15 ± 9.16, - Word count: 144,353, - Vocabulary: 13,154.

In this study, we address the class imbalance problem in our dataset by using a hybrid approach combining under-sampling and over-sampling. We start by randomly removing instances from the 'NOHATE' class to balance the representation. After this, we augment the 'HATE' class byduplicating instances using a method like SMOTE. This ensures both classes are equally represented, improving our model's ability to learn from both hate speech and non-hate speech instances.

4.2 Hyperparameters and Evaluation

We fine-tuned the BERT model for hate speech classification to develop an efficient and accurate classifier. For the debiaser model, we use the GPT-2 -small model with 117M parameters. We tuned the model temperature (0.1 - 1.0) in our debiaser model to balance diversity in the generated text samples [17]. Additionally, we utilized few-shot learning with prompts to further refine the debiasing process. We used 5 and 10 examples per category in our few-shot learning experiments. For the classification task and to train the whole pipeline, we used 3 epochs, optimizing the parameters using the Adam optimizer with a learning rate of 5e-5, weight decay of 0.5 and an epsilon value of 1e-8. We searched a grid of hyperparameters, including batch sizes of 4, 8, and 16, 32, 64 . We limited the input sequences to 128 (subword) tokens and trained the model in batches of 16 (to avoid out-of-memory issues). We employ F1 score (by calculating the harmonic mean precision and recall) and accuracy. For training, we used Google Colab Pro, which provided access to an NVIDIA Tesla T4 GPU with 16 GB of memory.

4.3 Performance Evaluation

Table 1 presents the results of our proposed method for mitigating biases in hate speech classification. The classifier's performance is evaluated using F1-score metrics. Further, we introduce a novel measure - the bias score - to assess the effectiveness of our debiasing model.

In this study, our classifier generates a quantifiable bias score for each text, reflecting the degree of hate speech bias. Each text is scored both before and after the debiasing process. A decrease in this bias score post-debiasing is indicative of successful bias mitigation. This metric provides us a way to numerically gauge the effectiveness of our proposed debiasing model.

The results shows that the proposed method achieves the highest F1 score of about 95%, outperforming all other classification methods. Moreover, the debiasing performance of the proposed method with few-shot learning and prompts achieves an impressive F1 score of 89%, with a low standard deviation. This indicates that the proposed method can effectively mitigate biases in hate speech classification.

To evaluate the performance of the debiaser model, we conducted experiments with zero-shot and then 5 and 10 prompts with examples, respectively. The results show that using few-shot learning with prompts improves the F1 score compared to zero-shot learning. This indicates that the model is better able to mitigate biases in the input text with the guidance of a few examples. This technique has the potential to enhance the performance of the debiasing model further with more diverse and relevant prompt examples, which could be explored in future work.

Classification performance: Next, we show the performance of the classifier model on a sample of 100 examples and report the results in Figure 2. We observe in the Figure 2, the model has made

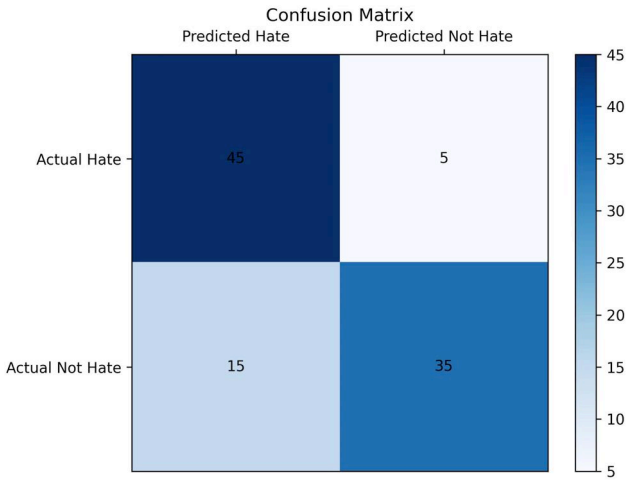


Figure 2: Confusion matrix for hate speech detection.

45 true positive predictions, 5 false positive predictions, and 35 true negative predictions. The false negative count is 15. The high true positive rate indicates that the model is correctly identifying a majority of the hate speech sentences, while the low false positive rate indicates that the model is not incorrectly labeling non-hate speech sentences as hate speech. However, the false negative count is 15, indicating that there is still room for improvement in correctly identifying all instances of hate speech. Since, these results are based on a sample of 100 instances and may not be representative of the model's performance on a larger dataset. Further evaluation and tuning of the model may be necessary to improve its overall performance.

Table 2: Performance comparison of original and debiased data on the hate speech classification task.

Metrics	Original Data	Debiased Data
Classifier performance		
Hate Speech Detection Accuracy	85%	83%
False Positives Rate	20%	12%
False Negatives Rate	15%	17%
Debiasing performance		
Bias Score*	65%	35%
Reduction in Bias Score	-	30%

*Bias score represents the degree of bias in the model, with a higher percentage indicating more bias. In this case, it's a measure of how much the model is biased towards incorrectly classifying certain non-hateful speech as hate speech.

Debiasing Performance: To further measure the effectiveness of our debiasing method, we measure and compare a range of performance metrics before and after debiasing. We first train our classifier on the original, biased dataset, and calculate the bias score. We then apply our debiasing process to the data, train the classifier on the debiased dataset, and again calculate the bias score. The change in the bias score serves as an indication of the effectiveness of our debiasing method.

As shown in Table 2, our debiasing method successfully reduced the bias score by 30%, from 65% to 35%. This indicates a significant reduction in the model's inclination to misclassify certain non-hateful speech as hateful. However, these improvements were accompanied by a slight decrease in hate speech detection accuracy, with the accuracy dropping from 85% to 83%. Additionally, there was a slight increase in the false negatives rate, which rose from 15% to 17%. These findings suggest that the debiasing caused slightly less accuracy in identifying the hateful speech. On the other hand, the false positives rate decreased from 20% to 12%, indicating an improvement in correctly identifying non-hateful speech .

5 CONCLUSION

We propose a hate speech classifier and debiaser that employs prompts to generate less biased alternatives. Our approach first detects hate speech using a classifier and then utilizes a debiasing component that generates less or no biased alternatives. Our proposed approach has some limitations. One limitation is the size and quality of the available training data, which may not be sufficient to cover all possible scenarios. Another limitation is the need for fine-tuning of the language model, which may require additional resources and expertise. Furthermore, the effectiveness of our approach may be affected by the quality of the underlying language model used for generating alternative text. Nonetheless, further research is needed to address the complex issue of online hate speech and biased language comprehensively.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[2] Aishwarya Anand, Murthy Devarakonda, Manik Gambhir, Arnav Wadhwa, and Mark J Carman. 2021. Few-Shot Learning for Hate Speech Detection on Social Media. In *Proceedings of the 1st Workshop on Online Abuse and Harms (WOAH 2021)*. 30–35.

[3] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521* (2021).

[4] Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. *arXiv preprint arXiv:2010.14534* (2020). arXiv:2010.14534 http://arxiv.org/abs/2010.14534

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2075–2086.

[7] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling Bias in Toxic Speech Detection: A Survey. *Comput. Surveys* (2023). https://doi.org/10.1145/3580494 arXiv:2202.00126

[10] Lara Grimminger and Roman Klinger. 2020. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 {US} Elections on the Basis of Offensive Speech and Stance Detection. *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 187–197. https://doi.org/10.18653/v1/2020.wassa-1.22

[11] Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574* (2022).

[12] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064* (2019).

[13] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9 (2023). https://doi.org/10.1145/3560815 arXiv:2107.13586

[14] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14857–14866.

[15] Emily May, Yao Wang, Shikha Bordia, Hannah Gao, Jilin Li, and Percy Liang. 2019. Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7479–7486.

[16] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[17] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. [n. d.]. INSTRUCTION TUNING WITH GPT-4. ([n. d.]). https://instruction-tuning-with-gpt-4.github.io/

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI* (2018).

[19] Shaina Raza, Deepak John Reji, and Chen Ding. 2022. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics* (2022), 1–21.

[20] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9180–9211. https://aclanthology.org/2022.emnlp-main.625

[21] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems* 33 (2020), 12388–12401.

[22] Heting Wu, Zhenxin Sun, Sheng Li, and Xiang Ren. 2022. Leveraging Prompt Engineering and Counterfactual Data Augmentation for Fairer Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

SchoolFinder: An AI-Assisted System for Localizing Schools from Satellite Imagery – Case Study of Sudan

Ferda Ofli
Qatar Computing Research Institute
HBKU
Doha, Qatar
fofli@hbku.edu.qa

Iyke Maduako
Office of Innovation
UNICEF
New York, NY, USA
imaduako@unicef.org

Masoomali Fatehkia
Qatar Computing Research Institute
HBKU
Doha, Qatar
mfatehkia@hbku.edu.qa

Ji Lucas
Qatar Computing Research Institute
HBKU
Doha, Qatar
jlucas@hbku.edu.qa

Aye Nyein Thaw
Office of Innovation
UNICEF
New York, NY, USA
athaw@unicef.org

Mohammad Amin Sadeghi
Qatar Computing Research Institute
HBKU
Doha, Qatar
msadeghi@hbku.edu.qa

Sanjay Chawla
Qatar Computing Research Institute
HBKU
Doha, Qatar
schawla@hbku.edu.qa

Do-Hyung Kim
Office of Innovation
UNICEF
New York, NY, USA
dokim@unicef.org

ABSTRACT

In 2019, UNICEF and ITU launched the Giga initiative to connect all unconnected schools in the world to the Internet in accordance with the sustainable development goal of providing quality education and lifelong learning. Surprisingly, the location of schools is not well documented in many parts of the developing world. In this work, we propose SchoolFinder, a computer-vision-based system that can identify schools from satellite imagery. The backbone of SchoolFinder is a data-centric adaptation of the state-of-the-art object detection algorithm for localizing schools from satellite imagery. Using an extensive ground-truth data set from Sudan that was hand-annotated via a field survey, we show that SchoolFinder can detect 90% of the school locations correctly. We have also integrated SchoolFinder with a well-known road-inference framework to identify the road network around school locations.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; • **Computer systems organization** → *Real-time system architecture*; • **Applied computing** → *Environmental sciences.*

KEYWORDS

satellite imagery, object detection, school mapping, data-centric AI

KDD'23, August 06–10, 2023, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Ferda Ofli, Iyke Maduako, Masoomali Fatehkia, Ji Lucas, Aye Nyein Thaw, Mohammad Amin Sadeghi, Sanjay Chawla, and Do-Hyung Kim. 2023. SchoolFinder: An AI-Assisted System for Localizing Schools from Satellite Imagery – Case Study of Sudan. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'23).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Quality Education is one of the 17 Sustainable Development Goals established by the United Nations [40] to *ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.*[1] According to the Sustainable Development Goals Report 2022 [41], during the COVID-19 pandemic, approximately 147 million children around the world missed over half of in-person instruction and 50% of primary schools globally lack access to the Internet. The report also highlights that recovering from the pandemic will require a significant investment in school infrastructure and services.

To tackle this challenge, United Nations Children's Fund (UNICEF) and the International Telecommunication Union (ITU) launched a joint initiative called Giga [7], which aims at connecting the unconnected schools in the world to the Internet to help bridge the digital divide in the world by 2030. One of the key components of this initiative is having access to accurate data about school locations. However, for many countries such as Sudan, school location records are often inaccurate, incomplete, or non-existent. Specifically, in Sudan, it is assumed that there are 20,000-25,000 schools across the country [29]. However, the majority of these school locations are not clearly documented. Traditional methods of the field visit and mapping of the school locations are not only heavily expensive but also do they prove impractical as some schools are located in remote,

[1] https://sdgs.un.org/goals/goal4

inaccessible, and insecurity-prone areas. Open data sources such as the OpenStreetMap (OSM)[2] could be very useful in some areas but it has little-to-no data in many developing countries. Alternatively, there has been significant progress in AI-based techniques including object and infrastructure detection and mapping from satellite imagery [11, 17]. However, applicability, generalizability, and scalability of such techniques on satellite imagery in the context of school detection and mapping at scale, especially in developing countries, have not yet been explored fully, which is indeed the gap that inspires this paper.

The aim of this study is to develop and deploy an AI model to accurately and comprehensively identify school locations from high resolution satellite imagery. To achieve this goal, we follow a data-centric modeling approach as described in [44]. Specifically, we iterate between data curation and model training steps to attain an optimal, unbiased, and fair school detection model. Our best model achieves a precision of 88.4% and a recall of 91.6% on the validation set. We then demonstrate the applicability of our model in a real-world deployment through a case study in Sudan where the developed system, SchoolFinder, is able to detect 90% of the school locations correctly.

We make the following contributions:

(1) We take a data-centric AI approach to iteratively improve the performance of a state-of-the-art object detection model, YOLO, for school location prediction.
(2) We have built a web-based system, SchoolFinder, that, in principle, can locate a school anywhere in the world provided we have a diverse training set.
(3) SchoolFinder can be used as prototype for other initiatives and tasks tied to SDGs that require the use of satellite imagery as a cost-efficient approach.

The rest of the paper is structured as follows. Section 2 reviews the relevant literature, Section 3 introduces the dataset, Section 4 explains the methodology, Section 5 presents the experimental results, and Section 6 concludes the paper.

2 RELATED WORK

State-of-the-art methods in computer vision domain leverage on large labeled image collections such as Places2 [48] or ImageNet [32] by using various deep learning architectures based on Convolutional Neural Networks (CNN) [8, 15, 20, 33, 34, 37]. One of the key strengths of systems based on deep learning is that they automatically infer a representation of the data suitable for the defined task. For example, the lower layers of deep learning correspond to a representation suitable for low-level vision tasks while the higher layers are more domain specific [10] and obviates the need for pre-defined feature engineering like SIFT [27]. Also, it is very common to use pre-trained weights of these large models for classification and detection purposes when there is not enough training examples or computational resources are insufficient.

Even though deep CNNs became popular initially for the image classification task, their capabilities were quickly transferred to the other tasks such as object detection. Early attempts in object detection used a set of bounding boxes and masked regions as input to the CNN architecture to incorporate shape information into the classification process to perform object localization [5, 8, 38, 39]. Later on, end-to-end techniques based on shared computation of convolutions for simultaneous detection and localization of the objects [3, 9, 13, 14, 24, 30, 31, 33].

Figure 1: Image grids for rural (left) and urban (right) areas

Remote sensing community has followed these trends closely and adopted the notion of large-scale satellite imagery datasets such as xView [21], DOTA [45], DIOR [22], and FAIR1M [36] as well as deep learning-based object detection models [1, 4, 12, 22, 25, 26, 47]. However, accurate building detection from satellite imagery *at scale* remains to be a challenging task, which makes detecting specific building types (i.e., schools) even more challenging. To this end, the closest works to our study have been presented in [19, 28] where the school mapping problem was formulated as a tile-level image classification task. While this formulation leads to a relatively easy solution, it ignores the most crucial component of the mapping problem: localization. In addition, classification-based solution suffers from too many false positives and negatives and requires substantial human effort for verification and localization of the results. Therefore, in this study, we approach the problem directly as an object detection task.

3 DATASET

This section elaborates on the satellite imagery acquisition and ground-truth school annotation tasks for our case study in Sudan.

3.1 Satellite Imagery

We systematically divide the entire Sudan area into 600m × 600m grids, which correspond to 1000 × 1000-pixel extent high-resolution satellite image tiles at 0.6m spatial resolution. This process yields a total of 5,694,227 grids for the entire country. However, majority of the grids correspond to desert areas with no human settlement data, let alone schools (see Figure 1 for rural vs. urban area comparison). Therefore, we collect human settlement data for Sudan from multiple sources including Google's Open Buildings[3] [35], UN OCHA country office in Sudan[4], Geo-Referenced Infrastructure and Demographic Data for Development[5], and Kontur, a geospatial data and real-time risk management solutions provider for business, government and humanitarian organizations.[6] In addition, we use land use and land cover maps from the European Space Agency (ESA)[7], ESRI[8], and Dynamic World dataset[9] [2] to delineate built-up

[2]https://www.openstreetmap.org/

[3]https://sites.research.google/open-buildings
[4]https://data.humdata.org/
[5]https://grid3.org/
[6]https://www.kontur.io/
[7]https://esa-worldcover.org/en
[8]https://livingatlas.arcgis.com/landcover/
[9]https://dynamicworld.app/

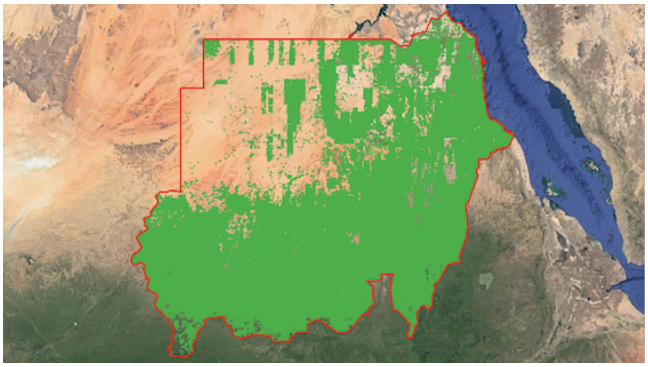


Figure 2: Selected grids (green areas) using human settlement and land cover/land use data from multiple sources

Figure 3: Example image tiles containing school buildings

areas from the rest of the images. Then, we collate all these geospatial data to filter only for the grids that contain human settlement areas, which results in a significantly reduced number of 1,282,586 grids as highlighted in Figure 2. Finally, using the selected grids, we download high-resolution satellite images with 0.6m spatial resolution from MAXAR's imagery archive under NextView license.[10] The imagery is collected from WorldView-3 sensor and composited with R, G, and B bands using natural composite method.

3.2 Ground-Truth School Locations

We obtained about 7,000 school locations (i.e., geo-coordinates) that were collected through a GPS field exercise in Sudan. We used these reference school locations to identify the relevant subset of satellite image tiles with schools and further inspected these images to annotate schools with bounding boxes as ground truth data. To this end, annotators were informed about the popular school building/compound design for primary and secondary schools in Sudan, which mostly follows the following principles: Three main buildings are usually noticed with the classroom building in the center and two other buildings set one on each side at 90° to the classroom building. The three main buildings have access verandas at the front. Additionally, there may be a playground, a kitchen, a separate store, a borehole, and a hand-pump inside the school area. Figure 3 depicts sample image tiles containing school locations.

Data labeling and label review process takes most of the data preparation time. We used data annotation toolkits such as Labelbox[11] and MakeSense[12] to systematically label schools with bounding box annotations. To this end, six different annotators

[10]https://hiu.state.gov/imagery/NextViewLicense.txt
[11]https://labelbox.com/
[12]https://www.makesense.ai/

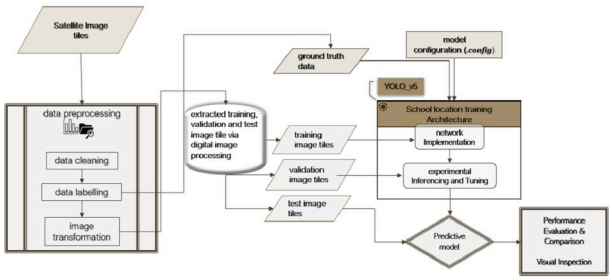


Figure 4: Methodology overview

annotated 6,704 image tiles with schools, taking a total of 48 hours. A review process was then carried out to verify and determine annotation agreement between the annotators. This took a total of 14 hours. Out of 6,704 image annotations, 6,554 were in agreement, yielding a percentage agreement score of 0.98 (i.e., 98%).

4 METHOD

This section outlines the model development workflow and deep learning framework adopted to detect school locations in RGB high-resolution satellite imagery, as depicted in Figure 4. Since our objective is to develop a model that can optimally account for the small-scale objects while also warranting run-time efficiency, we experimented with the popular YOLOv5 model architecture [18] to detect and localize schools.

Building successfully upon the previous versions of YOLO [30], YOLOv5 is considered an advanced object detector that achieves top performances on popular object detection benchmarks such as Pascal VOC [6] and Microsoft COCO [23] object detection challenges. Figure 5 illustrates the network architecture of YOLOv5. YOLOv5 is a useful choice of object detector because it incorporates the cross stage partial network (CSPNet) [42], which addresses the problem of repeated gradient information in large-scale backbones and integrates the gradient changes into the feature map, hence decreasing the parameters and FLOPS (floating-point operations per second) of the model.

YOLOv5 also boosts information flow by employing path aggregation network (PANet) [43], which adopts the feature pyramid network (FPN) structure with enhanced bottom-up path and improves the propagation of low-level features. The PANet utilizes the adaptive feature pooling to link feature grid and feature scale levels to allow valuable information in each feature level to propagate directly to the adjoining subnetwork. PANet also increases object localization accuracy using enhanced localization signals in lower layers. The head of YOLOv5 which is the YOLO layer, creates 3 distinct sizes of feature maps (18×18, 36×36, 72×72) to accomplish multi-scale [46] predictions, enabling the model to cope with varying object sizes of school location features.

These unique features of YOLOv5 do not only ensure high inference speed and accuracy, but also reduces the model architecture size, and hence, the memory footprint. Speed and accuracy are paramount for our case study as we aim to perform school location detection at a country-wide scale from millions of image tiles. For this reason, we opted for YOLOv5 model. We trained the

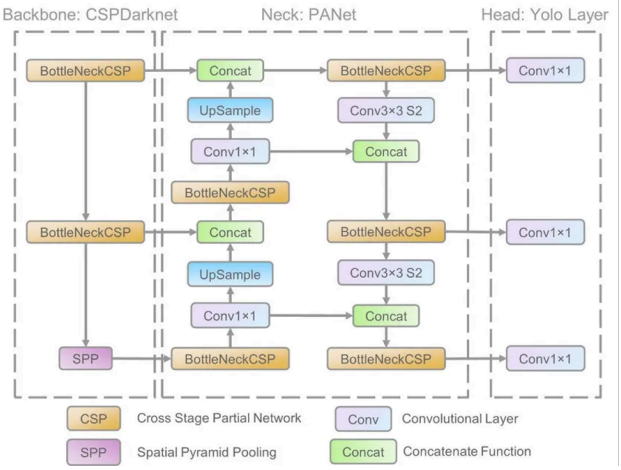


Figure 5: YOLOv5 architecture (image source: [46])

Table 1: Data split into training and validation sets

	Training	Validation	Total
School	4,178	1,045	5,223
Non-school	4,178	1,045	5,223
Total	8,356	2,090	10,446

model in an end-to-end fashion using pre-processed image tiles together with their corresponding ground-truth annotations to detect "school" locations. The output of the model comprises school location bounding boxes with associated loss values and prediction errors (probabilities).

It is notable that in future, YOLOv5 can be replaced with improved object detectors. Our overall school detection methodology is partly agnostic to the choice of object detection algorithm. Therefore, most of our focus is on making sure that the entire methodology is sound and rigorous.

5 EXPERIMENTS

5.1 Baseline Model

5.1.1 Dataset Split. We further examined the ground truth school image tiles collected in Section 3.2 and eliminated some uncertain images and poor quality annotations to reduce the noise, which left us with 5,223 school images. In order to create a more realistic dataset, we extended the initial set of school image tiles with additional image tiles that did not contain any schools. We then divided the resulting dataset into two splits with 80% for training and 20% for validation. Consequently, the extended dataset had a total of 8,356 images for training and 2,090 images for validation, evenly distributed across "school" and "non-school" classes, as summarized in the Table 1.

5.1.2 Model Training. As discussed in Section 4, we employed YOLOv5-large network architecture with an input size of 256×256. We used the SGD optimizer with an initial learning rate of 0.01.

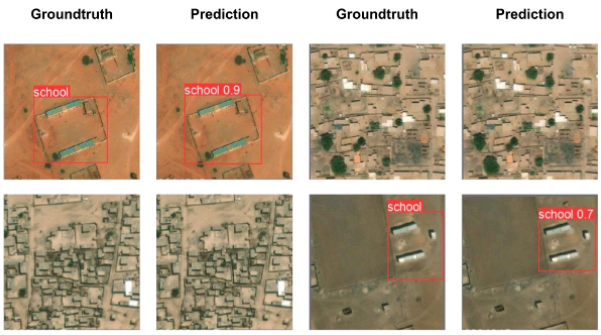


Figure 6: Example detection results on the validation set

The model was trained for 50 epochs with a batch size of 150. We performed data augmentations such as image flipping and rotations. The resulting model achieved precision, recall, and F1-scores of 0.810, 0.826, and 0.818, respectively, on the validation set (Table 3). Figure 6 presents some qualitative results from model predictions on the validation set.

5.1.3 Qualitative Analysis. The first version of the model made many false positive predictions when employed to map the entire Sudan. From our error analysis, we found that the model was making mistakes for the following types of images: (i) long rectangular buildings in industrial zones (Figure 7a), (ii) buildings with fences (Figure 7b), (iii) agricultural fields (Figure 7c), and (iv) desert areas and water bodies (Figure 7d). With these observations in mind, we move onto the data-centric model development.

5.2 Data-Centric Model Development

AI has traditionally focused on the development of new algorithms and methods for training better predictive models. Lately, the promise of deploying AI-based solutions in real-world applications such as self-driving cars has revealed that AI models are only as good as their data pipelines. That is, engineering, managing and understanding the data that feeds AI models can often matter much more in practice than modifying training algorithms and model architectures. Therefore, to produce higher performing AI systems, data-centric AI [44] suggests focusing on the data pipeline which typically involves (i) curating a dataset for labeling based on model performance after every iterative cycle to address the model's specific weaknesses and (ii) significantly increasing performance with a relatively small amount of training data.

5.2.1 Data Analysis and Refinement. As per data-centric AI, we re-investigated the training dataset to sort and categorize the images based on the structure and patterns of school buildings. The purpose was to pinpoint which type of school buildings was causing the false positives as well as to test whether removing a certain category from training would help reduce false positive predictions.

Through several modeling iterations, we found out that mainly the following two types of school buildings in the training set caused most of the false positive predictions: (i) single bar school buildings (Figure 8a) and (ii) schools that look like regular buildings

(a) Long rectangular buildings in industrial zones

(b) Non-school buildings with fences

(c) Agricultural fields

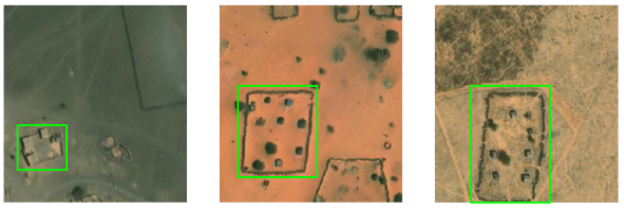
(d) Desert areas and water bodies

Figure 7: Most common mistakes

(Figure 8b). The reason is that there are a lot of non-school buildings in Sudan that look similar to those school buildings shown in Figure 3. So, we removed those types of images from the training dataset, added some non-school images from areas that were not present in the initial dataset, trained the models again on the new dataset, and ran prediction on the same sample region as before. The new model produced fewer false positives but at the same time, missed a lot of single bar school buildings. For the model to clearly distinguish school and non-school buildings that have similar structure and pattern, it would require a certain amount of such training samples.

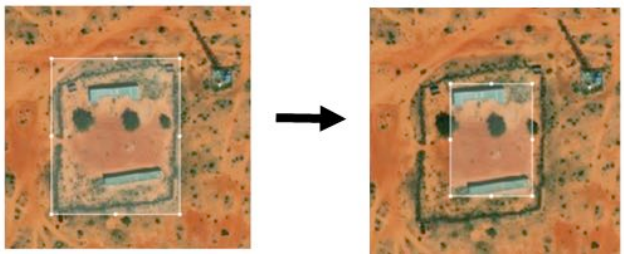
We also checked the labels and found that there were incorrect and inconsistent labels. So, we cleaned the labels and re-annotated a few to make them consistent (Figure 8c). We also changed the bounding box annotation scheme and drew boxes around school buildings rather than around school fences (Figure 8d). This was done because the model was producing false positive predictions on a lot of non-school buildings with fences (Figure 8e), sometimes even on fences without any buildings in them (Figure 8f).

(a) Single bar school building without (left) and with (right) fence



(b) Schools that look like regular buildings

(c) Correction of wrong labels



(d) Modification of existing bounding box annotation

(e) False positive prediction of non-school buildings with fence as school buildings

(f) False positive prediction of a fence with no building in it as a school building

Figure 8: Data quality improvement scenarios

Table 2: Training and validation set splits after data refinement

	Training	Validation	Total
School	1,935	491	2,426
Non-school	3,654	917	4,571
Total	5,589	1,408	6,997

Table 3: Comparison of baseline and data-centric models

	Precision	Recall	F1-score	mAP@0.5
Baseline	0.810	0.826	0.818	0.777
Data-centric	0.884	0.916	0.900	0.943

As a result of this iterative data refinement process, we were left with a total 5,589 images for training and 1,408 images for validation. As before, the data was split into 80% training and 20% validation sets. Training set contains 1,935 and 3,654 images for school and non-school images, respectively, while the validation set contains 491 and 917 images for school and non-school images, respectively (Table 2).

5.2.2 Model Training. The final model was trained on this new dataset following the same experimental setup explained before. The updated model achieved precision, recall, and F1-scores of 0.884, 0.916, and 0.900, respectively, on the validation set, outperforming the baseline model significantly as shown in Table 3.

5.3 Country-wide Evaluation

We test our school detection model's performance on all the grids selected and downloaded across Sudan except for the tiles used for model training purposes.

5.3.1 Inference Pipeline. As a pre-processing step, 1000×1000 images are sliced into smaller 256×256 tiles with 20% overlap between each tile. The reason for slicing the images into overlapping tiles is because schools will not always be located at the center of the 256×256 tile, and we do not want to miss schools that are situated at the intersection of multiple tiles.

After the pre-processing step, the model is run on the sliced tiles to produce output bounding boxes in pixel coordinates for detected schools. The predicted school locations are then geo-registered using the geo-coordinate grids for comparison with the ground truth school location data. Figure 9 shows the visualization of the model predictions. Blue-colored region is the 1,282,586 satellite image tiles for Sudan country filtered and downloaded using human settlement data from different sources and the red-colored dots are the detected schools while the green-colored dots are the ground truth school locations. Figure 10 provides a closer look at the predicted schools in rural and urban areas.

5.3.2 Qualitative Assessment & Field Validation. Initial quality assessment of the predicted results with field dataset has been carried out by comparing the predicted results with more than 7,000 school locations obtained from a field campaign. The assessment recorded more than 90% match with this field collected school locations as shown in Figure 11. A crowdsourced field validation programme is being planned for further validation of the results. For this purpose, we have developed an interactive online system.

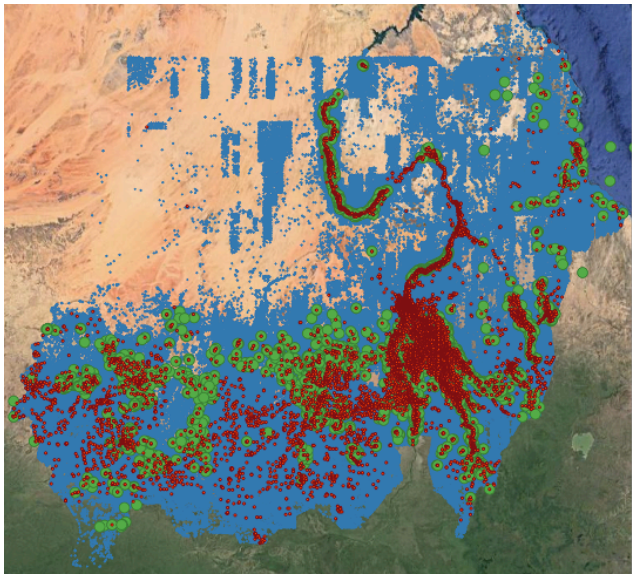


Figure 9: Country-wide model prediction results

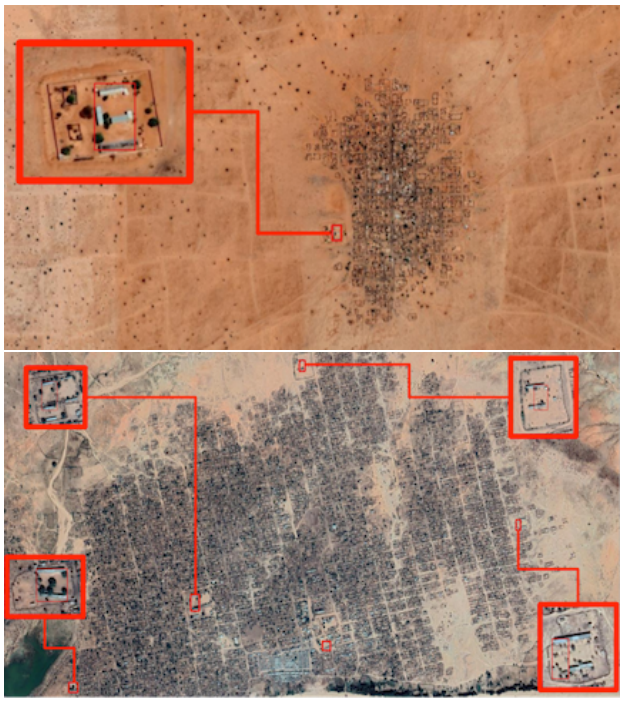


Figure 10: A closer look at the school detection results in rural (top) and urban (bottom) areas

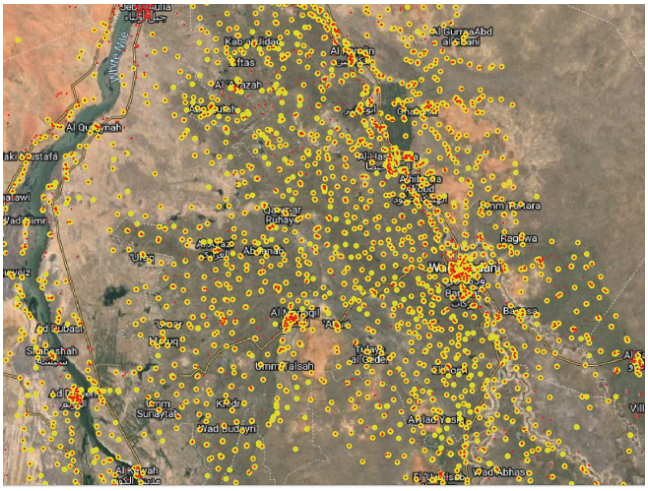


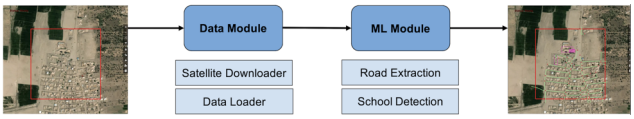
Figure 11: A region in Sudan illustrating the overlap between the predicted (red) and ground-truth (yellow) school locations

5.4 SchoolFinder Demo System

To complement the school mapping efforts, we developed an interactive demo system, SchoolFinder,[13] that can be used to explore the school detection model results for different parts of the globe. The front-end UI is built on an OSM open source project, iD Editor[14] and the back-end is based on Flask, a micro-framework.[15] On the demo interface, a user can scroll and zoom in to any part of the world map and choose an area of interest within the rectangular selection box that is overlaid on the map as shown in Figure 12. Once the area of interest has been chosen, the user can then activate the detection process which consists of first downloading satellite imagery for the selected area and then running the school detection model on the image (Figure 12a). Rectangular bounding boxes for the detected schools are then displayed on the map (pink boxes in Figure 12b). Using this demo can further facilitate exploration by enabling a user to visually inspect the model results in new locations and to validate the model by testing its results in school areas that are known to them.

Besides the school detection model, the system features a state-of-the-art road extraction module that automatically detects road networks [16] within the selected area of interest (green lines in Figure 12b). Combining the advantages of segmentation-based approaches with graph-based approaches, the road extraction module employs a novel encoding scheme called graph-tensor encoding, which encodes the road graph into a tensor representation. This encoding scheme allows training a model that can predict the positions of road network vertices and edges directly from a satellite image. We envision using the road network information to improve school mapping performance as future work.

[13]Available online at https://asm.qcri.org/.
[14]https://learnosm.org/en/beginner/id-editor/
[15]https://flask.palletsprojects.com/en/2.2.x/



(a) System workflow

(b) System output

Figure 12: SchoolFinder system

6 CONCLUSION

This study aimed to develop a scalable approach, called SchoolFinder, for locating schools by combining AI models with a web-based, high-resolution satellite imagery layer. In spite of the varying features of school locations in Sudan from rural to urban areas, this work proved that there are still yet identifiable overhead signatures common to school locations in Sudan that made it possible to detect schools from high-resolution satellite imagery with modern deep learning techniques. Our experimental evaluations also highlight the importance of data-centric AI for developing and deploying AI solutions to practical real-world problems.

REFERENCES

[1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. 2022. Slicing aided hyper inference and fine-tuning for small object detection. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, IEEE, Bordeaux, France, 966–970.

[2] Christopher F. Brown, Steven P. Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J. Pasquarella, Robert Haertel, Simon Ilyushchenko, Kurt Schwehr, Mikaela Weisse, Fred Stolle, Craig Hanson, Oliver Guinan, Rebecca Moore, and Alexander M. Tait. 2022. Dynamic World, Near real-time global 10 m land use land cover mapping. *Scientific Data* 9, 1 (2022), 251.

[3] J. Dai, K. He, and J. Sun. 2015. Convolutional feature masking for joint object and stuff segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 3992–4000. https://doi.org/10.1109/CVPR.2015.7299025

[4] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. 2021. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence* 44, 11 (2021), 7778–7796.

[5] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. 2014. Scalable Object Detection Using Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern*

Recognition. IEEE, Columbus, OH, USA, 2155–2162. https://doi.org/10.1109/CVPR.2014.276

[6] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111 (2015), 98–136.

[7] Giga. 2021. Connecting every school in the world to the internet. https://projectconnect.unicef.org/map. [Online; accessed: 2023-02-24].

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Columbus, OH, USA.

[9] Ross B Girshick. 2015. Fast R-CNN. In *International Conference on Computer Vision*. IEEE, Boston, MA, USA, 1440–1448.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press, Cambridge, MA, USA.

[11] Alex Goupilleau, Tugdual Ceillier, and Marie-Caroline Corbineau. 2021. Active learning for object detection in high-resolution satellite images. *arXiv preprint arXiv:2101.02480* (2021).

[12] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. 2021. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Online, 2786–2795.

[13] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2014. Simultaneous Detection and Segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, Zurich, Switzerland, 297–312.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. IEEE, Venice, Italy, 2961–2969.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, Las Vegas, NV, USA, 1–9.

[16] Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M Elshrif, Samuel Madden, and Mohammad Amin Sadeghi. 2020. Sat2graph: Road graph extraction through graph-tensor encoding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, Springer, Glasgow, UK, 51–67.

[17] Thorsten Hoeser, Felix Bachofer, and Claudia Kuenzer. 2020. Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications. *Remote Sensing* 12, 18 (2020), 3053.

[18] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. 2021. *ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations*. ultralytics. https://doi.org/10.5281/zenodo.4679653

[19] Do-Hyung Kim, Guzmán López, Diego Kiedanski, Iyke Maduako, Braulio Ríos, Alan Descoins, Naroa Zurutuza, Shilpa Arora, and Christopher Fabian. 2021. Bias in Deep Neural Networks in Land Use Characterization for International Development. *Remote Sensing* 13, 15 (2021), 2908.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (Eds.). Curran Associates, Inc., Lake Tahoe, CA, USA, 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[21] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. 2018. xView: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856* (2018).

[22] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing* 159 (2020), 296–307.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, Springer, Zurich, Switzerland, 740–755.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision (ECCV)*. Springer, Amsterdam, The Netherlands, 21–37.

[25] Wenchao Liu, Long Ma, Jue Wang, et al. 2018. Detection of multiclass objects in optical remote sensing images. *IEEE Geoscience and Remote Sensing Letters* 16, 5 (2018), 791–795.

[26] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. 2017. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* 55, 5 (2017), 2486–2498.

[27] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* 60, 2 (Nov. 2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[28] Iyke Maduako, Zhuangfang Yi, Naroa Zurutuza, Shilpa Arora, Christopher Fabian, and Do-Hyung Kim. 2022. Automated School Location Mapping at Scale from Satellite Imagery Based on Deep Learning. *Remote Sensing* 14, 4 (2022), 897.

[29] R. Obaid. 2021. Estimated numbers of schools in Sudan. Personal conversation.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Las Vegas, NV, USA, 779–788.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, and Jian Sun. 2016. Object detection networks on convolutional feature maps. *arXiv:1504.06066 (v2)* (2016).

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[33] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. 2014. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *International Conference on Learning Representations (ICLR 2014)*. CBLS, Banff, AB, Canada. http://openreview.net/document/d332e77d-459a-4af8-b3ed-55ba

[34] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

[35] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. 2021. Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283* (2021).

[36] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 184 (2022), 116–130.

[37] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, Boston, MA, USA, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[38] Christian Szegedy, Scott E. Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. 2015. Scalable, High-Quality Object Detection. *arXiv:1405.0312 (v3)* (2015).

[39] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. 2013. Deep Neural Networks for Object Detection. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Lake Tahoe, NV, USA, 2553–2561. http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf

[40] UN. 2015. The 17 Goals: Sustainable Development. https://sdgs.un.org/goals. [Online; accessed: 2023-02-24].

[41] UN-DESA. 2022. The Sustainable Development Goals Report 2022. https://unstats.un.org/sdgs/report/2022/. [Online; accessed: 2023-02-24].

[42] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. IEEE, Online, 390–391.

[43] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*. IEEE, Long Beach, CA, USA, 9197–9206.

[44] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal* (2023), 1–23.

[45] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 3974–3983.

[46] Renjie Xu, Haifeng Lin, Kangjie Lu, Lin Cao, and Yunfei Liu. 2021. A forest fire detection system based on ensemble learning. *Forests* 12, 2 (2021), 217.

[47] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. 2019. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Long Beach, CA, USA, 8232–8241.

[48] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

Received 24 May 2023; revised 24 May 2023; accepted 24 May 2023

Attention based Remote Photoplethysmography Estimation from Facial Video with Equilibrium in Time-Frequency Supervision

Sungpil Woo*[12], Muhammad Zubair*[1], Sunhwan Lim[1], and Daeyoung Kim[2]
[1]Autonomous IoT Research Section, Electronics and Telecommunications Research Institute (ETRI)
[2]School of Computing, Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
[1]woosungpil, zubair5608, shlim@etri.re.kr [2]woosungpil, kimd@kaist.ac.kr

ABSTRACT

In pre-clinical health monitoring, estimating physiological signals from video is a low-cost and convenient tool. Remote photoplethysmography (rPPG) involves placing a camera in a remote area to estimate a person's heart rate or Blood Volume Pulse (BVP). In this paper, we propose an attention based deep architecture for rPPG estimation that assimilate temporal relationship across a sequence of frames while focusing on the relevant features and regions by exploiting the inter-pixel relationship of feature maps. Also, we design a dynamic supervision strategy using frequency and time domain losses to mitigate overfitting for efficient estimation of rPPG signals. The proposed method was evaluated on two publicly available rPPG datasets (UBFC-rPPG and PURE). The findings of this study demonstrate that promising results can be achieved by enforcing an adequate balance between time-frequency supervision.

CCS CONCEPTS

• **Computing methodologies** → **Biometrics**.

KEYWORDS

Remote Photoplethysmography, Attention Mechanism, Dynamic Supervision, Health Monitoring, Synthetic Heart Rate

ACM Reference Format:
Sungpil Woo*[12], Muhammad Zubair*[1], Sunhwan Lim[1], and Daeyoung Kim[2]. 2023. Attention based Remote Photoplethysmography Estimation from Facial Video with Equilibrium in Time-Frequency Supervision. In *Proceedings of DSSG'23 (KDD 2023 Workshop on Data Science for Social Good).* , 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The ability to measure vital signs such as heart rate, oxygen saturation, and related physiological measures can be achieved using a technique called photoplethysmography (PPG). In this method, blood volume changes are monitored optically by monitoring light absorption in tissues. This is usually accomplished with a contact sensor attached to the skin and requires interaction with the subject limiting their usefulness and scalability [6]. In remote photoplethysmography (rPPG), on the other hand, the blood volume pulse is detected by tracking changes in the skin reflectance using a camera [3, 13] without any contact. This has great potential in many applications.

A significant improvement in the performance of rPPG techniques has been achieved through advancements in the fields of computer video, signal processing, and machine learning [6]. A number of early rPPG studies focused on finding signals within the image that was easier to access and perhaps more robust to nuisance variations, such as color over specific regions [10] or motion [1]. Most published rPPG methods work either by applying skin detection [10, 12] on a certain area in each frame or by selecting one or multiple regions of interest and track their averages over time to generate color signals[8, 9].

More advanced performance in rPPG has been achieved by deep learning to extract either heart rate or the BVP directly from camera images. They rely on the ability of deep networks to learn which areas in the image correspond to heart rate. This way, no prior domain knowledge is incorporated and the system learns rPPG concepts from scratch with capable of learning more robustly through noise[2, 8, 9, 11]. For example, HR-Net [11] uses a two-stage convolutional neural network (CNN) with a per-frame feature extractor and an estimator network with an attention mechanism. DeepPhys [2] is the first such end-to-end method to extract heart and breathing rates from videos with attention mechanism and time-frequency domain supervision. Similarly, [15] introduce a temporal difference transformer with multi-head self-attention mechanism to explore long-range spatio-temporal relationship for PPG measurement.

Moreover, temporal and frequency domain supervision have been employed in the literature [9]. The temporal domain supervision includes loss functions like negative Pearsons and mean square error that only focus on the trends of the signals. These loss functions are easy to converge. However, the models trained with time domain loss functions are prone to overfitting [15]. On the other hand, frequency domain supervision aims to control periodic features of the predicted signals within a target frequency band [15]. However, training based on frequency domain loss function causes deterioration in generalization and thus are considered inappropriate for real-world application. Therefore, an efficient strategy is needed to be adapted for an adequate balance between time-frequency supervision to alleviate the corresponding issues.

Motivated by the discussion above, we introduced an attention-based deep architecture for efficient estimation of heart rate and rPPG signals. In addition, we propose an intelligent training strategy by exploiting dynamic supervision to gradually change time and

* These authors contributed equally to this work

KDD 2023 Workshop on Data Science for Social Good, August 7, Long Beach, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

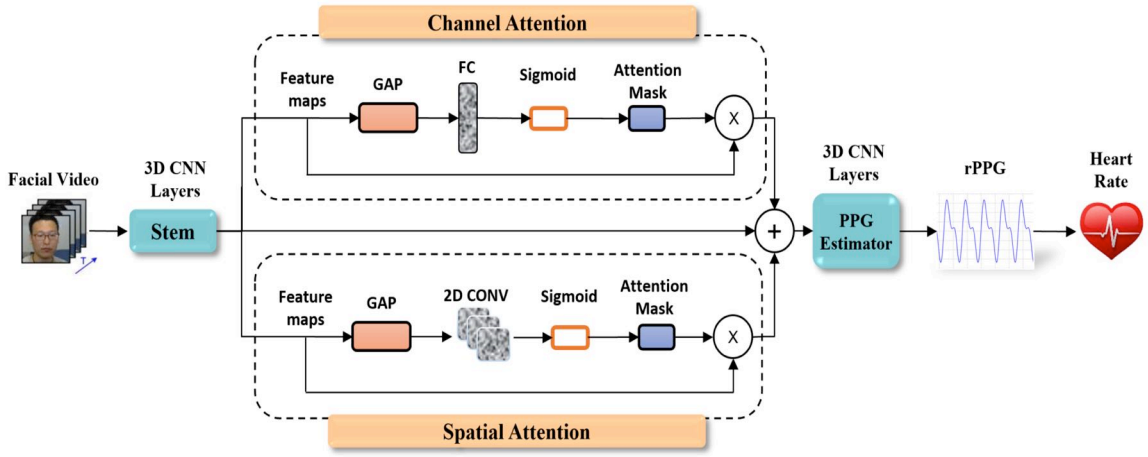


Figure 1: Overall workflow of proposed remote photoplethysmography estimation

frequency domain constraints for intrinsic feature learning. The contribution of this study is as follows:

- We propose an attention-based deep architecture that assimilates temporal relationships across a sequence of frames while focusing on the relevant features and regions by exploiting the inter-pixel relationship of feature maps.
- We also propose a dynamic supervision strategy using frequency and time domain losses to mitigate overfitting for efficient estimation of rPPG signals.

The remainder of the paper is organized as follows. In Section II, we describe the overall structure and workflow of our proposed model. In Section III, the experimental setting and experimental results are presented. We will conclude the paper in Section IV.

2 METHODOLOGY

An overview of the proposed approach for remote photoplethysmography estimation is shown in Fig.1. A detailed explanation of the proposed approach including model architecture and loss function is presented in this section.

2.1 Remote Photoplethysmography Estimation

The aim of the current study is to measure the heart rate efficiently using facial images by exploiting the most relevant information. Moreover, this study also explores the plausible potential of maintaining an adequate balance between time and frequency supervision during training.

The proposed architecture includes a stem comprised of 3D CNN layers that aim to learn a general representation F_s of a sequence of $T \in \{1, ...t\}$ input facial images. These feature maps (F_s) are further refined via attention modules to extract the most relevant information for the estimation of rPPG signals. Inspired by [7], we designed a spatial attention module and a channel attention module to extract the most relevant information from each frame in a sequence of T input images. The architecture of the proposed attention module is given in Fig.1. Afterward, an rPPG estimator incorporating 3D CNN, 1D CNN, spatial pooling, and normalization layers was used to generate 1D rPPG signal. Finally, the heart rate was computed using power spectral density (PSD) of the predicted rPPG signal [5].

Additionally, we used frequency domain supervision with complementary temporal domain supervision. We adopted an efficient strategy to dynamically scale frequency domain and temporal domain losses to mitigate the issues of convergence and overfitting. The proposed strategy allows the model to efficiently learn deep representation while taking trend-level and frequency-level constraints into account.

2.2 Attention Mechanism

The selection of relevant and most appropriate regions in facial frames is indispensable for the efficient estimation of rPPG signals. In this paper, we used the channel attention module and spatial attention module to boost the model performance. The channel attention module selects the most appropriate kernels and the spatial attention focuses on the salient regions of the feature maps extracted for each time-stamp of the sequence. The illustration of both modules is given below.

2.2.1 Channel Attention. The channel attention learns the weights for each channel based on its contribution to estimate the target rPPG signal. The channel attention mask exploits the inter-channel relationship of the input feature maps F_s to generation channel attention mask A_c. A global average pooling (GAP) operation is performed on the spatial axis followed by a fully connected layer and softmax operation to model the relationship between channels. The generated attention mask is then used to scale each channel yielding weighted feature maps that carry the most relevant information.

2.2.2 Spatial Attention. The spatial attention module model the inter-pixel relationship of the feature maps to generate a spatial attention mask A_s. Therefore, GAP is applied on the channel axis for dimensionality reduction of the input feature maps. The acquired

2D feature maps are passed through a convolution layer of a 3 x 3 filter and a sigmoid layer to obtain the spatial attention mask. The spatial attention mask is the most crucial part for the efficient estimation of rPPG signals as it carries pixel-level information of the most important regions in the frames.

2.3 Time-Frequency Supervision

In this study, we used time-domain supervision and frequency-domain supervision to assist the model in learning trends and periodic features of the target signals. It has been demonstrated that the gradual increase in frequency constraints mitigates the issue of overfitting by learning more general features for rPPG generation[15]. To explore further the impact of time-frequency supervision, we also adopted an effective strategy to control time and frequency domain constraints dynamically during training for intrinsic representation learning without overfitting. However, unlike [15], we explore both incremental and decremental strategies for time-frequency constraints for a smooth and fast convergence.

Additionally, we adopted label distribution learning strategy to predict the heart rate (HR) value within a specific range ([40, 180]). For this, Kullback-Leibler (KL) divergence loss [4] was computed between the power spectral density of the predicted signal and the corresponding sample from normally distributed ground truth HR values. In addition to KL loss, we also used frequency cross-entropy loss [9] for frequency-domain supervision. For time-domain supervision, we adopted Negative Pearson Loss [14]. The overall loss can be formulated as

$$\mathcal{L}_{\text{overall}} = \underbrace{\alpha \cdot \mathcal{L}_{\text{NP}}}_{\text{Temporal}} + \underbrace{\beta \cdot (\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{CE}})}_{\text{frequency}}$$
$$\alpha = \frac{1}{1 + \eta^{(\text{Epoch}_{\text{current}} - (\text{Epoch}_{\text{total}}/\rho))/\text{Epoch}_{\text{total}}}}, \quad (1)$$
$$\beta = \frac{1}{1 + \eta^{((\text{Epoch}_{\text{total}}/\rho) - \text{Epoch}_{\text{current}})/\text{Epoch}_{\text{total}}}}$$

where, η and ρ equal to 0.001 and 3.0, respectively. $\mathcal{L}_{\text{NP}}$ is the Negative Pearson loss, $\mathcal{L}_{\text{KL}}$ is the KL divergence loss and $\mathcal{L}_{\text{CE}}$ is the frequency cross entropy loss. The hyperparameter η is the most crucial parameter that controls the rate of change and strongly impacts the model performance.

3 EXPERIMENTAL SETTINGS

3.1 Datasets

We evaluated the proposed rPPG estimation method on two publicly available databases of physiological measurement. These databases include UBFC-rPPG, and PURE. **UBFC-rPPG** database includes uncompressed facial videos of 42 subjects with recorded PPG and heart rate data. For significant induction of variability in heart rate, a mathematical puzzle was carried out by subjects during recording. UBFC-rPPG does not include predefined folds for training and test sets. Therefore, we adopted five-fold cross-validation to match the evaluation protocols of [5]. **PURE** database comprises physiological data with facial images from 10 subjects. The data were acquired during 6 different activities (steady, talking, slow translation, fast translation, small rotation, and medium rotation). PURE also comes with predefined splits of training and test set.

3.2 Pre-rocessing

We first used a face detector on video frames to estimate the bounding box around the face [16]. The bounding box region for the face is estimated for each frame. A scale buffer of 10% was added to the estimated bounding box before resizing. We finally scale each frame to 128 x 128 resolution. For a single sample, 160 consecutive frames were stacked. During training, random RGB samples with sizes 160 x 128 x128 were used as model inputs. These simple preprocessing steps were adopted for all three databases.

4 RESULTS AND DISCUSSION

We conducted the experiments on UBFC-rPPG and PURE databases under different supervision conditions for model evaluation. We explore the impact of incrementing time constraints and decrementing frequency constraints on model performance for heartbeat estimation. On the contrary, we also evaluated the proposed model performance with increasing frequency domain constraints and decreasing time domain constraints.

Table 1 summarises the results of the proposed rPPG estimation model. The evaluation results were obtained for four different values of η. The η with a value less than 1 represents an incremental strategy of time domain loss and a decremental strategy for frequency domain loss. Similarly, η with a value greater than 1 represents an incremental strategy of time domain loss and a decremental strategy for frequency domain losses. The 5-fold cross-validation experiment on the UBFC database yielded an average mean absolute error (MAE) of 3.2, root mean square error (RMSE) of 4.8, and Pearson Correlation (PC) coefficient of 0.96 with a value of $\eta = 0.001$. Similarly, the experiments on the PURE database also demonstrate that the gradual decrease in frequency domain losses yields better performance as compared to incremental strategy.

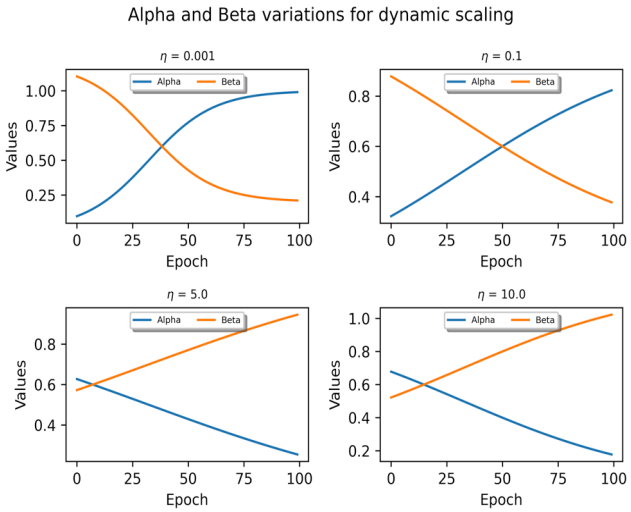


Figure 2: Visualization of alpha and beta during training

Table 1: Classification comparison with previously published results

Authors	Method	UBFC-rPPG			PURE		
		RMSE	MAE	PC	RMSE	MAE	PC
Wang et al.[3]	CHROM	20.3	10.6	–	17.18	6.23	30.71
De Haan et al.[13]	POS	8.3	3.5	0.90	3.14	10.7	0.95
Lee et al.[8]	Meta-rPPG	7.4	5.9	0.53	–	–	–
Gideon et al.[5]	Supervised with Saliency	4.6	3.6	0.95	2.6	2.1	0.99
Proposed	3dCNN + Attention [ρ=0.001]	**4.8**	**3.2**	**0.96**	**3.1**	**2.0**	**0.99**
	3dCNN + Attention [ρ=0.1]	5.1	3.4	0.93	5.7	2.7	0.97
	3dCNN + Attention [ρ=5.0]	4.9	3.3	0.93	6.5	2.4	0.97
	3dCNN + Attention [ρ=10.0]	5.5	3.8	0.92	6.8	3.2	0.96

The visualization of alpha and beta variations for dynamic scaling of time-frequency losses is given in Fig. 2. The value of η that controls the rate of change has illustrated a significant impact on heartbeat estimation. An important implication of the findings presented in Table 1 is that a gradual decrease in frequency constraints guarantees the extraction of more generalized features right from the start of training. Similarly, the gradual increase in time domain constraints eases the model convergence. As the model starts to learn the intrinsic periodic features of the target PPG signal from the start, therefore the risk of overfitting with increasing time domain constraints alleviates during training. In addition, the experimental findings also illustrate that the proposed attention mechanism also contributed efficiently to learning the most relevant features for rPPG estimation.

The current findings add substantially to our understanding of training rPPG estimation models with suitable scaling of time-frequency losses. Table 1 also presents the comparison of our proposed method with relevant studies. The study has confirmed that the proposed method outperforms the conventional rPPG estimation methods of CHROM[3] and POS[13]. The efficacy of the proposed scaling method is also justifiable by comparing the results with popular deep learning methods of rPPG measurement [5, 8]. Taken together, the findings of this study suggest that an adequate balance between time-frequency supervision is important to train a model for efficient estimation of rPPG signals.

5 CONCLUSION

In this paper, we propose an attention-based deep model with adequate equilibrium between time-frequency supervision for rPPG estimation. We demonstrated that the gradual decrease in frequency domain constraints enables the model to learn more generalized features for heartbeat prediction, improves model convergence speed, and mitigates the issue of overfitting. This research extends our knowledge of improving the generalizability of the rPPG estimation model by enforcing a sufficient balance between time-frequency supervision.

6 ACKNOWLEDGMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(2022-0-01032, Development of Collective Collaboration Intelligence Framework for Internet of Autonomous Things

REFERENCES

[1] Guha Balakrishnan, Fredo Durand, and John Guttag. 2013. Detecting Pulse from Head Motions in Video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 3430–3437. https://doi.org/10.1109/CVPR.2013.440

[2] Weixuan Chen and Daniel McDuff. 2018. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. *CoRR* abs/1805.07888 (2018). arXiv:1805.07888 http://arxiv.org/abs/1805.07888

[3] Gerard De Haan and Vincent Jeanne. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886.

[4] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26, 6 (2017), 2825–2838.

[5] John Gideon and Simon Stent. 2021. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3995–4004.

[6] M.A. Hassan, A.S. Malik, D. Fofi, N. Saad, B. Karasfi, Y.S. Ali, and F. Meriaudeau. 2017. Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control* 38 (2017), 346–360. https://doi.org/10.1016/j.bspc.2017.07.004

[7] Yangru Huang, Peixi Peng, Yi Jin, Junliang Xing, Congyan Lang, and Songhe Feng. 2019. Domain adaptive attention model for unsupervised cross-domain person re-identification. *arXiv preprint arXiv:1905.10529* (2019).

[8] Eugene Lee, Evan Chen, and Chen-Yi Lee. 2020. Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-Learner. *CoRR* abs/2007.06786 (2020). arXiv:2007.06786 https://arxiv.org/abs/2007.06786

[9] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. 2020. Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling. *CoRR* abs/2007.08213 (2020). arXiv:2007.08213 https://arxiv.org/abs/2007.08213

[10] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* 18, 10 (May 2010), 10762–10774. https://doi.org/10.1364/OE.18.010762

[11] Radim Spetlik, Jan Cech, Vojtěch Franc, and Jiri Matas. 2018. Visual Heart Rate Estimation with Convolutional Neural Network.

[12] Sergey et al Tulyakov. 2016. Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2396–2404. https://doi.org/10.1109/CVPR.2016.263

[13] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. 2016. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 7 (2016), 1479–1491.

[14] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. 2019. Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 151–160.

[15] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H. S. Torr, and Guoying Zhao. 2021. PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. *CoRR* abs/2111.12082 (2021). arXiv:2111.12082 https://arxiv.org/abs/2111.12082

[16] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*. 192–201.

Narrative to Trajectory (N2T$^+$): Extracting Routes of Life or Death from Human Trafficking Text Corpora

Saydeh N. Karabatis
Vandana P. Janeja
{saydeh1,vjaneja}@umbc.edu
University of Maryland, Baltimore County (UMBC)
Baltimore, Maryland, USA

ABSTRACT

Climate change and political unrest in certain regions of the world are imposing extreme hardship on many communities and are forcing millions of vulnerable populations to abandon their homelands and seek refuge in safer lands. As international laws are not fully set to deal with the migration crisis, people are relying on networks of exploiting smugglers to escape the devastation in order to live in stability. During the smuggling journey, migrants can become victims of human trafficking if they fail to pay the smuggler and may be forced into coerced labor. Government agencies and anti-trafficking organizations try to identify the trafficking routes based on stories of survivors in order to gain knowledge and help prevent such crimes. In this paper, we propose a system called Narrative to Trajectory (N2T$^+$), which extracts trajectories of trafficking routes. N2T$^+$ uses Data Science and Natural Language Processing techniques to analyze trafficking narratives, automatically extract relevant location names, disambiguate possible name ambiguities, and plot the trafficking route on a map. In a comparative evaluation we show that the proposed multi-dimensional approach offers significantly higher geolocation detection than other state of the art techniques.

CCS CONCEPTS

• **Information systems** → **Information integration**; • **Computing methodologies** → *Natural language processing*; • **Applied computing** → Anthropology.

KEYWORDS

Data Science, Natural Language Processing, Human Trafficking, Location Name Disambiguation

ACM Reference Format:
Saydeh N. Karabatis and Vandana P. Janeja. 2018. Narrative to Trajectory (N2T$^+$): Extracting Routes of Life or Death from Human Trafficking Text Corpora. In *Proceedings of Workshop name (KDD '23).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

KDD '23, August 06–10, 2023, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Both climate change and political unrest can lead to lack of resources, poverty, and instability that can force certain populations of affected areas to depart from their homeland and seek refuge in far away places. Since they lack proper documentation to legally enter non-neighboring countries, these migrants rely on a network of smugglers to facilitate their journey during which many human rights are violated. In addition, the smuggling process is a highly profitable money laundering business in which the migrant pays an exorbitant amount of money to the traffickers in order to be secretly moved from one location to another [12].

The Department of Economics and Social Affairs in the United Nations (UN) created a list of 17 Sustainable Development Goals (SDG) for peace and prosperity for all people. These goals provide an urgent call for action by all UN member countries that specifically target sustainability and development for underdeveloped countries [11]. Three out of the 17 SDGs target human trafficking (HT) thus making it easier for anti-trafficking organizations to address this "grave human rights violation" [11]. SDG 5 aims to "achieve gender equality and empower all women and girls" (forced sex exploitation, a form of HT violates this goal). SDG 8 aims to promote inclusive economic growth and "decent work for all" (unpaid/underpaid labor, a form of HT violates this goal). SDG 16 aims to "promote peaceful and inclusive societies" and "provide access to justice for all" (migrant smuggling across different countries violates this goal).

Climate change has been setting serious obstacles to various communities in Far East Asia and in the Pacific region. "This region includes 13 out of 30 countries most vulnerable to the impacts of climate change" [14]. The impacted communities have been seeing more people fall into poverty and seeking refuge outside their homeland. In addition, the unrest in some of the Near East countries, mainly Syria and Lebanon, has led to the displacement of several million people who have been looking for refuge in safer non-neighboring countries. In both cases, the migrants often rely on a network of criminal smugglers to facilitate their movement and to assist them during their journey. Those who succeed in reaching their destination must deal with the trauma experienced throughout their journey, however, not everyone is successful as there have been several reports of organ harvesting from migrants who fail to pay the expenses associated with the trip or migrants losing their lives along the journey before reaching safe haven [5].

It is commonly known that most migrants take the same paths that others have crossed earlier. These paths are often referred to as routes of life or death [13]. One way law enforcement agencies can disrupt migrant trafficking activities and save the lives of the

migrants is to predict the route by gaining knowledge about past trafficking routes. Non-profit organizations can reach out to victims during their trafficking routes so that they do not have to endure the abuse at the hands of their traffickers.

Many narratives posted on media and anti-trafficking organizations sites contain a varying number of details about the experience of survivors during their captivity. Those narratives are either written by journalists or by members of the humanitarian organizations following interviews with the survivors. In order to protect the identity of the victims, most media sites often provide vague location names of the route travelled by the victims. Very few narratives tell it all: place of origin, visited places along the transportation routes, and the destination. We were able to locate articles that narrate stories of victims and list the location names along the transportation route in chronological order without disclosing any personal information of the victims.

Given the overwhelming number of survivor narratives reported on the internet and since very few of these narratives contain particular names of places along the transporting routes, it is not feasible for a human to read each narrative, extract and identify the location names, and plot on a map the travelled trajectory. Therefore, it is necessary to create a tool that automatically parses the narratives, extracts location names, disambiguates these names if ambiguation is present, and plots the transportation route as narrated on a map. Such a tool helps the officials identify previous trajectories and gain insight of future routes in order to save the lives of the vulnerable trafficked victims.

The complexity of human trafficking activities requires both human and machine intelligence to untangle. Using web advertisement as input, Szekely et al. apply Artificial Intelligence (AI) methods to extract telephone numbers of sex laborers, Esfahani et al. propose a semi-automatic composite model to identify sex trafficking ads, Tong et al. describe the development of a multimodal deep learning model to identify sex trafficking ads, and Nagpal et al. propose an entity resolution pipeline to extract clusters of data with prior history of human trafficking activities [7–10]. None of these works mines location names from the advertisements. Extracting geographical tags from atypical language models posted on illicit activities websites is presented in [2] using DARPA MEMEX [4]. The authors describe an Integer Linear Programming (ILP) model that processes human trafficking advertisements and produces a set of geolocation names. This approach works only if the population of the extracted location names exceeds 15 thousand. It does not address ambiguity of location names. Molina-Villegas et al. describe a geographic name entity recognition and disambiguation model *GNER* in Mexican news articles using word embedding and semantics to develop a Mexican Geoparser[6]. The model achieves acceptable accuracy in recognizing geographic named entity, but fails to fully resolve location name ambiguity.

Most of the above referenced works do not mine geographic locations and the few that do have major limitations. We proposed a Narrative to Trajectory (N2T) prototype system that processes narratives to identify trajectories [3]. The initial version of N2T lacks disambiguating location names. In this paper we propose N2T$^+$, a Data Science (DS) and AI model for disambiguating location names and tracing precise human trafficking trajectories extracted from text corpora of victim narratives. The contributions of this paper are to:

- Identify location names from text corpora
- Disambiguate location names in the presence of ambiguities
- Allocate the precise coordinates to each location name

The remainder of the paper is as follows: In Section 2, we describe our novel methodology to identify the victim's trajectory. In Section 3, we evaluate our proposed prototype through experiments. In Section 4 we conclude our paper and propose future work of the N2T$^+$ system.

2 METHODOLOGY

N2T$^+$ accepts a narrative as input, preprocesses it, splits it into tokens, labels each token according to its semantics and its syntax in the sentence, disambiguates location tokens when ambiguity is present, assigns precise spatial coordinates to location tokens, and outputs the trajectory as narrated on a map. To do so, N2T$^+$ integrates structured (Gazetteer and Lexicon) with unstructured (text corpora) heterogeneous data, performs tokenization methods, and leverages contextual sliding widow and rules. A high level architecture of the N2T$^+$ system is illustrated in Fig. 1.

N2T$^+$ consists of five components. The early version of N2T$^+$ does not address location name disambiguation [3]. For this reason, we use a contextual sliding window, rules, and DB techniques to enhance the early version and resolve location name ambiguity.

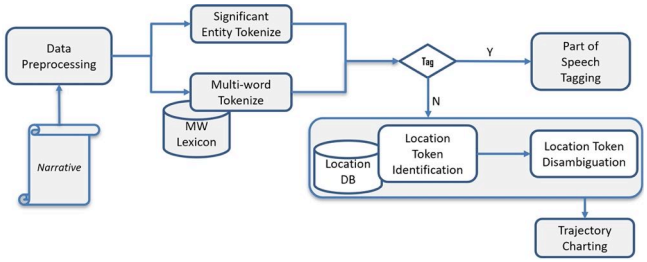


Figure 1: N2T$^+$ System Architecture

Data Preprocessing: Transforms the input text into a format applicable for analysis and prediction by performing data cleansing and replacing special and non-ASCII characters with NLP acceptable characters.

Tokenization: Splits text into meaningful tokens *To_V* by using NLP techniques that identify the divider between words in the text and generate tokens composed of either single or multi-words. However, the concept of using dividers (e.g. blank) by themselves for the purpose of generating tokens has its own limitations because some tokens are made out of multi-words where the designated divider can also be a unifier as in 'New Orleans'. For this purpose, we use:

- Significant Entity Tokenization (ST): It segments text into potential tokens of interest by applying lexical semantic rules specific to the language of the text. The output of this method is an ordered list of entities $To\_V_S = [(to\_v_i)]$ that are considered significant to the tokenizer.

- Multi-word Tokenization (MWT): It first applies sentence tokenization to the input. Each divided sentence is then tokenized into single tokens. Augmenting the generated list of tokens with the MW Lexicon helps generate tokens that also include multiple words. The output of this method is an ordered list of single word or multi-word tokens $To\_V_M = [(to\_v_j)]$.

At the end of the *Tokenization* step, $To\_V = To\_V_S \vee To\_V_M$.

Geospatial Token Identification and Disambiguation: *Tokenization* transforms the unstructured data into an ordered list of tokens *To_V* for each narrative. However, not all the identified tokens signify locations. In this step, we identify and disambiguate all the location tokens and assign the precise coordinates using a contextual sliding window, rules, and integration methods.

Location Disambiguation: It is commonly known that not all location names are unique. It is possible to encounter the same name that refers to different geographic places in the world: the city Tripoli is found in Greece, Lebanon, and Libya. We refer to this situation as *Homonym Ambiguity*. In addition, not all locations carry a distinct name. For example, Bombay is formally known as Mumbai. We refer to this situation as *Synonym Ambiguity*. To resolve the location ambiguity, we try to find the exact country/state where this location belongs to. Identifying the country/state helps in specifying the geospatial coordinates of the location token.

Our N2T$^+$ system includes a lookup Location dimension *Loc* that contains location names along with their country/state, longitude, and latitude values [*(to_v, to_cntry, to_long, to_lat)*] obtained from GeoNames [1]. We update *Loc* and add to it two fields [*(to_ha, to_sa)*] that signify if an entry is of *Homonym Ambiguity* or *Synonym Ambiguity* type respectively. The *Homonym Ambiguity* values contain the number of times the location name is found in *Loc* (*to_ha = 3* in case of *Tripoli*), whereas the *Synonym Ambiguity* values contain the formal name of the location if more than one name is given to the same location (*to_sa = Mumbai* in case of *Bombay*).

Augmentation: Using database join operation and while preserving the token order as they appear in the narrative, we augment the token list *To_V* with the *Loc* and generate a list of tokens *To_T* with values [*to_v, to_cntry, to_ha, to_sa*]. The values of [*to_v, to_ha, to_sa*] are the result of the join operation, while the value of the country [*to_cntry*] is determined based on the ambiguity of the token. If the token is not ambiguous, the value of the country is populated from *Loc*. Otherwise, it is initially set to *null* and will be resolved using in Algorithm 1. Resolving location ambiguity works as follows:

- In case of *Homonym Ambiguity*, we set the country name equal to the value of the country of the city that was last visited or will be visited next, using a contextual sliding window based on the principle of locality, which states that the current place is very likely adjacent to the place that was last visited or will be visited next. We apply this principle when the value of $to\_ha \geq 1$.
- In case of *Synonym Ambiguity* ($to\_sa \neq null$), we set the country of the location token to the country of the formal location.

Algorithm 1 identifies geolocation names, disambiguates the location tokens, assigns the correct country and coordinates to the disambiguated token, preserves the listing order of these geolocation names as they appear in the narrative, and detects whether a specific geolocation name is visited multiple times.

Algorithm 1 TokenDisambiguation

Require: *To_V(to_v) & Loc(to_v, to_cntry, to_long, to_lat, to_ha, to_sa)*
Ensure: *To_F(to_v, to_cntry, to_long, to_lat)*
Begin
$To\_T(to\_v, to\_cntry, to\_ha, to\_sa) \leftarrow To\_V \rtimes Loc$
// using *To_T*
while $i \leq n$ **do**
  if $to\_ha[i] \geq 1$ **then** // Homonym Ambiguity
    if $i = 1$ **then**
      $to\_cntry[i] \leftarrow to\_cntry[i+1]$ // first token in narrative
    else if $i = n$ **then**
      $to\_cntry[i] \leftarrow to\_cntry[i-1]$ // last token in narrative
    else if $to\_v[i] \in to\_cntry[i-1]$ **then**
      $to\_cntry[i] \leftarrow to\_cntry[i-1]$ // token exist in country of prior
    else if $to\_v[i] \in to\_cntry[i+1]$ **then**
      $to\_cntry[i] \leftarrow to\_cntry[i+1]$ // token exist in country of next
    end if
  else if $to\_sa[i] \neq null$ **then** // Synonym Ambiguity
    $to\_cntry[i] \leftarrow to\_cntry[to\_sa]$
  end if
  $To\_T \leftarrow To\_T + (to\_v[i], to\_cntry[i], to\_ha[i], to\_sa[i])$
end while
$To\_F(to\_v, to\_cntry, to\_long, to\_lat) \leftarrow To\_T \rtimes Loc$
End

N2T$^+$ contains a suite of four tokenization methods:

(1) ST: Significant Entity Tokenization
(2) MWT: Multi-Word Tokenization
(3) ST+Aug+DisAmbig: ST + Geospatial Augmentation + Disambiguation
(4) MWT+Aug+DisAmbig: MWT + Geospatial Augmentation + Disambiguation

Part of Speech Tagging: is an optional step that categorizes a token based on its semantics and position in a sentence. Applying NLP POS Tagging techniques to the tokens generated in the Tokenization step creates a list of 2-tuples (token, tag). This step is used to compare the results of only using AI techniques versus combining DS and AI techniques.

Trajectory Charting: plots the trajectory using the geospatial tokens generated in the tolenization step by depicting the travelled route over time as listed in the narrative.

3 EXPERIMENTAL RESULTS

3.1 Dataset and Ground Truth

Our text corpora is composed of several human trafficking narratives *(N_1 …N_n)*, a multi-word lexicon, and a location DB. The narratives are written by English speaking journalists and acquired from various news agencies and anti-trafficking organizations. They contain ambiguous location names. Their length varies between 605 and 4,212 words per narrative. To measure the performance of *N2T$^+$*, we conducted experiments and compared their results against the ground truth. To identify the ground truth, we manually read each of the narrative, extracted the location names, and saved them sequentially in a ground truth structure which was used later on to evaluate *N2T$^+$*.

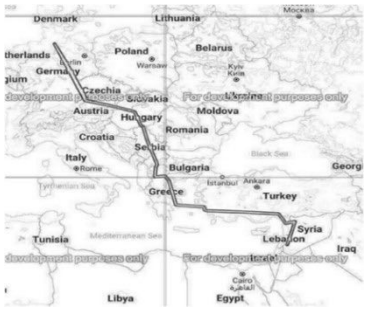


Figure 2: Trajectory of a Narrative

3.2 Results

We executed each of the four methods of N2T$^+$ using each narrative in the text corpus and generated a trajectory. Fig. 2 displays an example trajectory generated by N2T$^+$ using one migrant trafficking narrative. We compared every trajectory against the ground truth and calculated the performance measures for each method applied. Our goal is to extract as many geospatial tokens as possible from text corpora. *ST* and *MWT* failed to extract most geospatial tokens; therefore we decided to enhance them by augmenting each of these two methods with location dimension, using context, and applying the location disabmiguation approach resulting in methods *ST + Aug + DisAmbig* and *MWT + Aug + DisAmbig* respectively. The latter methods do not rely on the geospatial tagging process. They increase the true positive values and reduce false positive outcomes to a value close to zero resulting in higher F1-Score than *ST* and *MWT*. *MWT + Aug + DisAmbig* returns the highest F1-Score and Accuracy compared with all other methods because it uses sentence tokenization, Multi-word Lexicon, database joins, context, and the principle of locality to recognize and disambiguate multi-word location tokens as illustrated in Fig. 3.

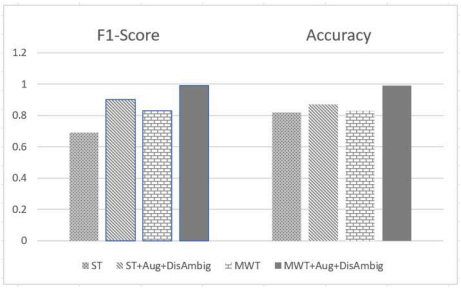


Figure 3: Performance Measures of N2T$^+$ Methods on Trafficking Narratives

We compared our N2T$^+$ *MWT + Aug + DisAmbig* to the proposed *ILP* [2] and *GNER* [6] and found out that our *MWT + Aug + DisAmbig* F1Score outperforms *ILP* by 21% and *GNER* by 37% as shown in Table 1. Such higher performance is attributed to the thorough data preprocessing, the utilization of NLP libraries, data augmentation, context, and principle of locality.

Table 1: Results of N2T$^+$ compared with other Methods

System	Precision	Recall	F1 Score
ILP	0.79	0.79	0.79
GNER	0.54	0.68	0.70
MWT + Aug + DisAmbig	0.98	0.91	0.96

4 CONCLUSION

We presented a system called Narrative to Trajectory (N2T$^+$), which extracts trajectories of trafficking routes. N2T$^+$ uses DS and AI techniques to analyze trafficking narratives, automatically extract relevant location names, disambiguate possible name ambiguities, and plot the trafficking route on a map. We first applied NLP libraries and found out that they lack retrieving some geospatial tokens. We then introduced geospatial dimension augmentation, context, and principle of locality concepts on top of the NLP libraries. In a comparative evaluation we show that our proposed multidimensional approach offers significantly higher geolocation detection and disambiguation than other techniques.

5 ACKNOWLEDGEMENTS

This work is funded in part by National Science Foundation (NSF) Award #2118285, "HDR Institute: HARP- Harnessing Data and Model Revolution in the Polar Regions"

REFERENCES

[1] GeoNames. 2021. GeoNames. https://www.geonames.org/
[2] R. Kapoor, M. Kejriwal, and P. Szekely. 2017. Using Contexts and Constraints for Improved Geotagging of Human Trafficking Webpages.
[3] S. Karabatis and V. Janenja. 2022. Creating Geospatial Trajectories from Text Corpora. In *Data-driven Humanitarian Mapping, 3rd KDD Workshop.* https://kdd-humanitarian-mapping.herokuapp.com/
[4] MEMEX. 2021. DARPA MEMEX. https://www.darpa.mil/program/memex
[5] Magdalena Mis. 2017. Organ trafficking booming in Lebanon as desperate Syrians sell kidneys, eyes. https://www.reuters.com/article/us-mideast-crisis-syria-trafficking/organ-trafficking-booming-in-lebanon-as-desperate-syrians-sell-kidneys-eyes-bbc-idUSKBN17S1V8
[6] Alejandro Molina-Villegas, Victor Muñiz Sanchez, Jean Arreola-Trapala, and Filomeno Alcántara. 2021. Geographic Named Entity Recognition and Disambiguation in Mexican News Using Word Embeddings. *Expert Syst. Appl.* 176, C (aug 2021), 8 pages. https://doi.org/10.1016/j.eswa.2021.114855
[7] C. Nagpal, K. Miller, B. Boecking, and A. Dubrawski. 2017. An Entity Resolution approach to isolate instances of Human Trafficking online.
[8] S. Shahrokh Esfahani, M. J. Cafarella, M. Baran Pouyan, G. DeAngelo, E. Eneva, and A. Fano. 2019. Context-specific language modeling for human trafficking detection from online advertisements.
[9] P. Szekely, C. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, and K. Knight. 2015. Building and Using a Knowledge Graph to Combat Human Trafficking.
[10] E. Tong, A. Zadeh, C. Jones, and L. Morency. 2017. Combating Human Trafficking with Deep Multimodal Models.
[11] UN United Nations. 2015. Transforming our World: The 2030 Agenda for Sustainable Development. https://sdgs.un.org/2030agenda
[12] UN United Nations. 2020. Protocol against the Smuggling of Migrants by Land, Sea and Air, supplementing the United Nations Convention against Transnational Organized Crime. https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-against-smuggling-migrants-land-sea-and-air
[13] Nick Walsh and et al. 2023. On one of the world's most dangerous migrant routes, a cartel makes millions off the American dream. https://www.cnn.com/2023/04/15/americas/darien-gap-migrants-colombia-panama-whole-story-cmd-intl/index.html
[14] WB World Bank. 2022. Climate and Development in East Asia and Pacific Region. https://www.worldbank.org/en/region/eap/brief/climate-and-development-in-east-asia-and-pacific-region

Received 5 June 2023

Monco: A Novel Deep Learning Pipeline to Predict Drug Candidates for the Inhibition of EZH2 Cofactors in Pediatric Neuroblastoma

Jaanak Prashar*

Jaagat Prashar*

jaanak007@gmail.com

jaagatp05@gmail.com

Jordan High School

Fulshear, Texas, USA

ABSTRACT

Pediatric Neuroblastoma (NB) is one of the most common cancers among infants, with 90 % of its cases occurring in children under the age of five and only 50 % of youth with high-risk NB surviving after five years. Currently, there are limited clinically approved drugs successful in treating high-risk NB due to cancerous side effects and undruggable MYC transcription factors. However, recent CRISPR-Cas9 loss-of-function screens have revealed the enhancer of zeste homolog 2 (EZH2) as a plausible therapeutic target for NB due to its proclivity for inhibiting tumor suppressor genes. Drug discovery is a time-consuming and costly process with the typical timeframe ranging from 10-15 years and developmental costs ranging from one to two billion dollars. Hence, computer-aided drug design (CADD) has become widely adopted; however, existing CADD remains largely incomprehensive due to its inability to account for dynamic protein structure, sufficient chemical descriptors for drug-like compounds, and computationally-efficient early screening of compounds. The purpose of this study was to develop a novel, robust, and production-ready *in silico* deep learning (DL) pipeline for *de novo* drug design using an ensemble of machine learning techniques. Monco, or machine learning for oncology, consists of three main functions: (1) identification of relevant genetic cofactors (2) development and optimization of drug-like compounds through deep learning (i.e., variational autoencoders, etc.) after cryptic pocket discovery, and (3) validation of compounds through molecular docking and Tanimoto similarity. To this end, this work serves as a potential efficient screening and discovery process for oncological therapeutics as a whole.

CCS CONCEPTS

• **Theory of computation** → **Design and Analysis of Algorithms**; • **Computing methodologies** → *Machine Learning*; • **Applied computing** → Life and medical sciences; • **Mathematics of computing** → mathematical software.

*Both authors contributed equally to this research.

KEYWORDS

pediatric neuroblastoma, drug discovery, generative models, computer-aided drug design, cryptic pocket discovery

1 INTRODUCTION

1.1 Pediatric Neuroblastoma

Pediatric Neuroblastoma (NB) is the most common extracranial cancer in children. NB carcinogenesis originates in sympathetic nerve tissue in general, typically in adrenal glands [1]. Existing treatments for pediatric neuroblastoma include chemotherapy, which increases the risk for bone and soft tissue tumors, surgical intervention, and anti-GD2 immunotherapy. Additionally, high-risk pediatric neuroblastoma, which accounts for roughly half of all neuroblastoma cases, has a long-term disease-free survival rate that is less than 50% [29]. Recently, a CRISPR-Cas9 screen revealed EZH2 as a strong genetic and pharmacological target to inhibit neuroblastoma tissue growth both *in vivo* and *in vitro* [9]. EZH2 performs its oncogenic function through trimethylation of Lysine 27 on Histone 3 (H3K27me3) relevant to tumor suppressor genes and its involvement in tumorigenic pathways [8, 32]. [36] further reports that EZH2 silences neuroblastoma suppressor genes CASZ1, CLU, RUNX3, and NGFR.

1.2 Computer-Aided Drug Design

In general, clinical drug discovery costs from 1 to 2 billion dollars and takes 10-15 years on average [33]. Thus, computer-aided drug design has emerged (CADD). CADD has historically contributed to the identification and optimization of lead and small molecule hit compounds [25]. Although several limitations exist in CADD that hinder the accuracy and efficiency of *in silico* drug development, emerging computational approaches have allowed for the analysis of the chemical space of small organic molecules, particularly in *de novo* drug design. [27] reports that CADD presents significant progress from high-throughput screening (HST) due to minimal previous knowledge for compound generation. Additionally, this study also shows that CADD has been employed in various drug discovery processes, including molecular design and drug validation. Quantitative structure-activity relationship (QSAR) has

been historically employed to show associations between molecular structure and activity [21]. Structural property information is necessary for successful small molecule inhibitor development. However, an overemphasis on structural information has been attributed to the toxicity of drug compounds [33]. In addition to patient comorbidities, an understanding of the drug delivery effects of small molecule inhibitors is necessary for iterative improvement of the prescribed drug. Currently, 97% of drug-indication pairs are unable to proceed through the last stage of the clinical drug discovery process for the Food and Drug Administration due to off-target effects [23]. Thus, designing small molecule inhibitors considering both QSAR and *in vivo* interactions is critical. The open-source machine learning toolkit AlphaFold is able to produce proteins with high accuracy through a deep learning transformer-based evolutionary algorithm, allowing for advanced drug design. Furthermore, while drug development processes have increasingly advanced, the availability of holistic production-ready platforms to synthesize drug-like molecules remains limited.

1.3 Insufficient Chemical Descriptors and Static Protein Structure

Existing frameworks for CADD do not comprehensively consider chemical descriptors required for effective drug delivery while emphasizing structural potency [33]. This work addresses the need for sufficient chemical descriptors through the adoption of predictive models for chemical space feature aggregation and the inclusion of chemical descriptors relevant to drug delivery. In spite of AlphaFold's protein structure prediction advancements, the adoption of static protein structures for drug identification hinders the accuracy of drug candidates. Roughly 50% of protein experimental domain structures lack apparent drug sites [26]. Static proteins are considered undruggable, but their pockets can be identified through the excited structural states they adopt in aqueous solutions. Limited accuracy of conformational protein changes was addressed through the implementation of AlphaFold's multiple sequence alignment (MSA) algorithm to detect cryptic pockets in the dynamic protein structure.

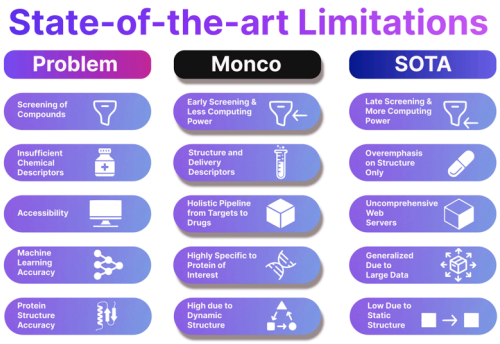


Figure 1: Comparison of SOTA software and Monco with respect to existing CADD limitations. Figure produced in Figma.

1.4 Early Screening and Holistic Pipeline

A primary limitation in CADD includes that existing web servers and toolkits generally process a stage of the *in silico* drug development process. Additionally, late screening in CADD is both computationally expensive and inefficient [19]. In this work, several self-organizing map reward functions including physicochemical information (i.e., bioactivity, drug-likeness, toxicity, and solubility) allow for the early detection of inactive compounds upon sampling of the latent chemical space. Through the development of early screening techniques, costs associated with CADD can be further reduced due to the limited amount of computational data needed at each step of the framework.

1.5 Chemical Space Identification and Data Aggregation

In general, drug chemical information is represented through the Simplified Molecular Input Line Entry System (SMILES) due to its linear nature compatible with text data for natural language processing methods [31]. Relevant SMILES sequences from the training data were tokenized into tensors before the production of a rich latent space for *de novo* molecular generation. Although CADD allows for the production of a chemical space for various SMILES sequences, the identification of chemical space regions for optimal drug candidate production remains as a limitation [25]. The implementation of self organizing maps (SOMs), a type of artificial neural network, in drug discovery has allowed for the definition of a reference set necessary for drug design [30]. Chemical space information was localized into regions of ideal drug molecules and ineffective drug molecules through the adoption of SOMs. Through the selection of relevant EZH2 cofactors from gene regulatory networks, the identification of drug sites, and the development of drug candidates, a holistic pipeline that allows for optimized drug discovery is produced. Prior sequence information of drugs is needed for accurate self-organizing maps and candidate generation as a whole. ChEMBL, a database consisting of binding, functional, and physicochemical molecular information, was employed for refined selection of the training data [13]. A primary challenge in deep learning molecular generation involves the limited amount of available biochemical data [3]. Furthermore, there is also a limited availability of diverse physicochemical information for SMILES candidates in a database and a need for increased integration of such data across databases. Thus, an ensemble of predictive models and mathematical calculations were implemented in conjunction with candidates obtained from ChEMBL to obtain the relevant descriptor information for drug generation and optimization.

1.6 Deep Learning Pipeline Framework

Deep learning has allowed for rapid advancements to be made in the drug discovery field [38]. Aforementioned limitations in *ab initio* protein design were addressed through the development of accessory models to integrate state-of-the-art software such as AlpahFold and GENTRL. In this work, Monco, a production-ready and novel deep learning pipeline is produced for highly accurate

and efficient identification of drug candidates and cofactors for NB. Monco's holistic drug discovery process is divided into five integral components: cofactor identification, pocket identification, *de novo* drug design, molecular docking, and further validation.

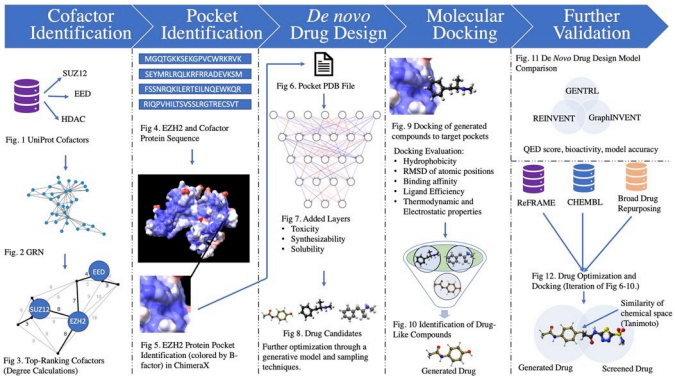


Figure 2: Novel Deep Learning Pipeline Architecture.

2 METHODS

2.1 Gene Regulatory Network (GRN) Analysis

UniProt is a database providing comprehensive and functionally annotated protein sequences, allowing for accessible protein information in stages of the *in silico* drug discovery process (i.e., molecular docking) [2]. Preliminary analysis of NB's genetic mechanisms involved identification of EZH2 and relevant cofactors from the public protein database. Additionally, Graphein, a geometric deep learning and network analysis library, was utilized to model the GRN for NB, particularly in adrenal sympathetic nervous tissue (although NB expression is not limited to a specific tissue).

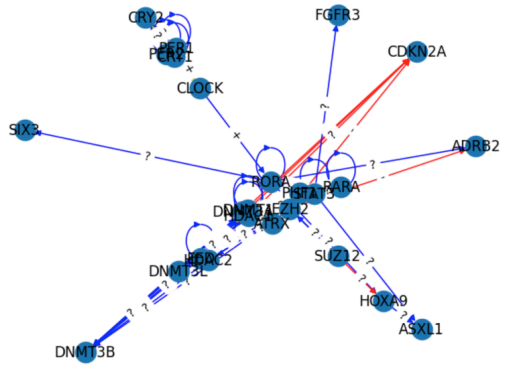


Figure 3: Cofactors and their respective cumulative edge values identified through indegree and outdegree calculations are as follows: HDCA1, 9; DNMT3B and DNMT1, 7; EED, 6; and DNMT3A, 5.

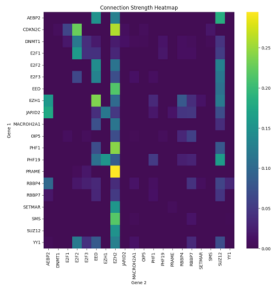


Figure 4: Gene-Gene interaction heatmap illustrating severity of connections between nodes.

Gene association relationships were analyzed through a gene regulatory network implementing genetic information from the Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRRUST) database of both human and mouse networks [18]. Further graph theory analyses including Dijkstra's algorithm and the minimum spanning tree were implemented to optimally evaluate relevant genes such as SUZ12, EED, and DNMT1. Furthermore, gene-gene interaction data was preprocessed and curated after conducting exploratory data analysis on the GeneMANIA database and BioGRID database. A graph neural network using PyTorch Geometric and GraphSAGE, a framework for inductive representation learning on graphs was employed to further substantiate gene-gene interaction analysis through a regression-based link prediction task [17]. GraphSAGE generates embeddings for unseen nodes through learned aggregation functions (expression A depicts a mean aggregator function). The neural network utilizes a mean squared error loss function as it evaluates a continuous value, the weight between two given nodes and thus illustrates the degree of connection between two given genes.

2.2 Multiple Sequence Alignment Cryptic Pocket Discovery

AlphaFold's MSA evolutionary algorithm was employed to produce 160 putative protein structures due to the static protein conformations generated by AlphaFold and other state-of-the-art models. Further analysis for cryptic pocket identification was conducted by superimposing the putative structures in ChimeraX to discover hidden drug sites that might emerge due to the protein's dynamic conformation in an aqueous environment. The default MSA technique from DeepMind, jackhammer, was adopted for cryptic pocket analysis due to its predictive capabilities and strong accuracy for hidden pocket detection. Five model parameters and 32 random seeds were attempted during the MSA run, resulting in the generation of 160 protein structures. Additionally, because mutations in the SET domain of the EZH2 protein are associated with increased trimethylation efficiency and cancer prognosis, subsequent cryptic pocket discovery analyses involved the development of a dynamic SET domain conformation through 160 additional putative structures [37]. Furthermore, in addition to increasing methylation efficiency, the SET domain is responsible for the primary activity of EZH2 in general [11].

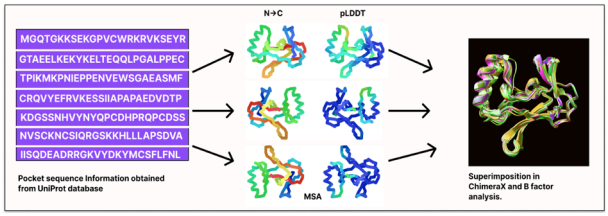
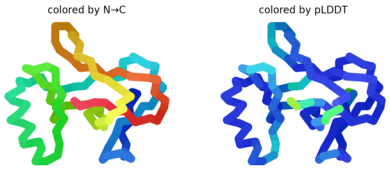
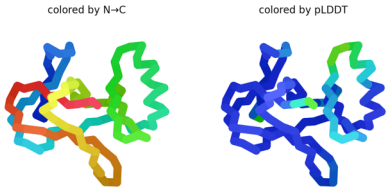


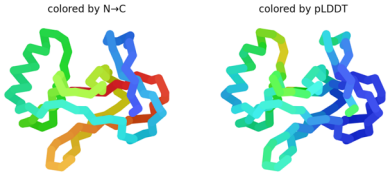
Figure 5: Cryptic Pocket Discovery MSA Workflow



(a) Example MSA Structure Generated in AlphaFold during initial Iterations.



(b) Example MSA Structure Generated in AlphaFold after several iterations.



(c) Example of MSA Structure Generated in AlphaFold toward the final iterations.

Figure 6: Multiple Sequence Alignment Protein Structures at Varying States.

2.3 Data Preprocessing

Traditionally, the Debye-Waller factor characterizes the attenuation of X-ray or neutron scattering in proteins [34]. Upon superimposition of the protein structures in the protein visualizer ChimeraX, the Debye-Waller factor, B factor, or atomic displacement factor was analyzed to determine putative pocket sites dependent on flexibility. Localized information characterizing protein rigidity and flexibility further allowed for accurate identification of drug sites due to high correlation with small molecule binding affinities [24].

Several preprocessing steps were taken to optimize model accuracy prior to the sampling and optimization of drug candidates. Aggregation of additional chemical descriptor information was conducted through predictive machine learning models. Upon normalization of chemical descriptor distributions obtained from parsing ChEMBL and implementing ensembled machine learning models, training SMILES sequences were tokenized into tensors. Furthermore, the Mann-Whitney U test was utilized to examine relative significance of chemical descriptors in the training data. Descriptors part of Lipinski's rule of Five, or Pfizer's rule of five, a traditional assessment of therapeutics, were further analyzed [14]. Additionally, an evaluation of the significance of drug candidate descriptors such as toxicity was conducted through comparison of drugs classified as active and inactive dependent on potency. The inclusion of IC_{50} values allowed for effective measures of drug efficacy, where lower values indicate the drug's potency at lower concentrations [6]. Small molecule inhibitors with an IC_{50} value of less than 1000nM were categorized as active, and those with an IC_{50} value greater than 10000nM were categorized as inactive.

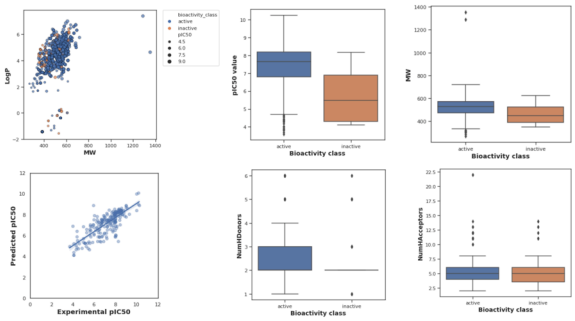


Figure 7: Exploratory Data Analysis of chemical descriptors and Mann-Whitney U Test conducted on various chemical descriptors to identify descriptors of interest.

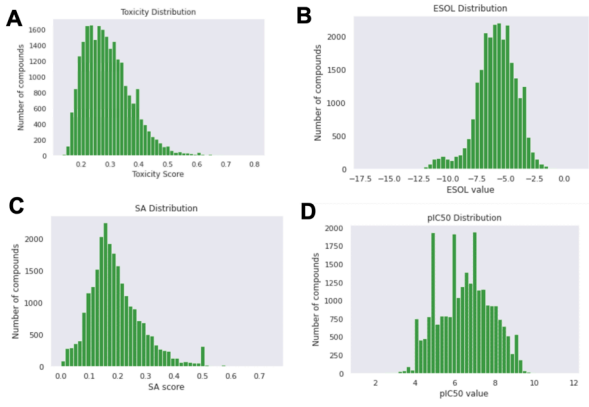


Figure 8: Distributions of chemical descriptors for SMILES generated from ChEMBL. A: Toxicity Distribution Frequency, B: Solubility Distribution Frequency, C: Synthetic Accessibility Distribution Frequency, D: Drug Potency Distribution Frequency.

2.4 Data Normalization

pIC_{50} values were obtained through computation of the negative logarithmic of IC_{50} values. Conversion to pIC_{50} allowed for accurate analysis of statistical significance of chemical descriptors analysis, higher reproducibility, and normalization [35]. Additional chemical descriptors significant for drug candidate optimization, including toxicity, solubility, and synthetic accessibility, were normalized through traditionally dividing all values by the maximum value in the training dataset for a relevant descriptor. More than 50,000 SMILES sequences were obtained by parsing the ChEMBL database for only EZH2-related known inhibitors. Initial data collection from ChEMBL including 10,000 SMILES sequences and subsequent analyses resulted in poor self-organizing map accuracies for each chemical descriptor. Thus, the ChEMBL parsing criteria was expanded to increase the scalability and accuracy of the pipeline. Inhibitors of histone methyltransferases (HMTs) as well as PcG proteins in general were incorporated to increase the training data size.

2.5 Synthetic Accessibility and Toxicity Prediction Models

Computational drug design requires synthesis of generated compounds for further validation because drug candidates with low synthetic accessibility scores are eliminated due to potential downstream complications [12]. Synthetic accessibility scores are further necessary for effective resource expenditure management. Compounds that are challenging to synthesize require a significant expenditure of costs and resources [12]. Toxicity of drug candidates significantly contributes to high drug development costs [15]. Thus, identification of toxicity in potential therapeutic compounds is necessary for early identification of adverse effects [16]. [28] reports that methodologies employed to analyze toxicity information of small molecule drugs aim to identify their effects on humans or the environment. This study also shows that existing machine learning predictive models such as ChemTox, which relies on quantitative-structure property relationship (QSPR), allow for collection of SMILES toxicity information. However, due to its generalized compatibility with highly varying datasets, the eToxPred deep learning predictive model was utilized to complement descriptor information obtained from ChEMBL through the addition of toxicity and synthetic accessibility scores [28]. The eToxPred model was also used due to its robust architecture, which includes a Deep Belief Network (DBN), a generative probabilistic architecture that allows for rapid training [28]. Furthermore, because eToxPred contains normalized toxicity values, toxicity values were renormalized for further exploratory data analysis of compounds.

2.6 Solubility Aggregation

Solubility is critical for the development of successful drug candidates [5, 10]. ESOL, or estimated solubility derived from compound structure, is important for *de novo* molecular generation as strong solubility properties allow for proper uptake of biologically active molecules and effective drug delivery [10]. The development of the ESOL calculation involves the following chemical properties: clogP, rotatable bonds, molecular weight, non-carbon proportion, aromatic proportion, hydrogen atom donor number, and the hydrogen atom acceptor number [10]. The ESOL mathematical framework was modified to allow for increased linear fit as part of the predictive computations on the training SMILES.

3 GENERATIVE TENSORIAL REINFORCEMENT LEARNING MODEL

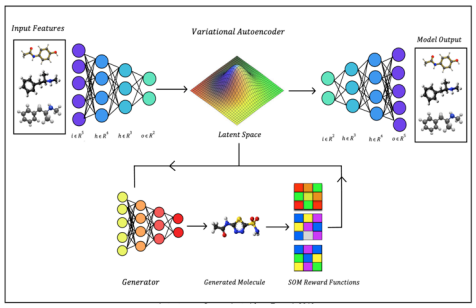
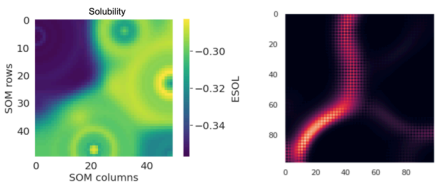


Figure 9: Generative Tensorial Reinforcement Model Architecture. Latent space image adapted from Ferreol, 2019. Figure was produced using the graphical software Figma.

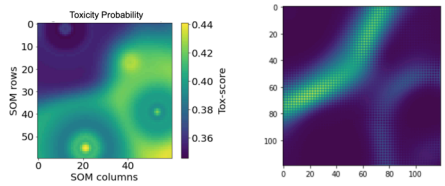
Tensor tokenized SMILES sequences were fed into a variational autoencoder. An important goal in drug discovery is to generate novel molecules with ideal properties. However, optimal candidate generation remains challenging due to the diversity of molecules that comprise the chemical space [22]. Variational autoencoders efficiently compress chemical space information to produce a rich latent space for novel candidate production. A generative tensorial reinforcement learning model (GENTRL) was implemented in relation to EZH2 cofactors for NB [39]. A reinforcement learning was integrated to train the generator to produce optimal drug candidates based on the provided set of chemical drug descriptors. Four SOMs trained on SMILES descriptor information from ChEMBL were used for accurate candidate generation and identification of the chemical space. SOM best-matching units were calculated using Euclidean distance due to simplicity in competitive learning and broadly in *in silico* drug discovery. Although an increase in training data size improved SOM accuracy, further hyperparameter optimization of SOMs was conducted through iterative modification of the number of rows, number of columns, starting learning rate, and ending learning rate. Preliminary analyses of the SOM with 50 rows, 50 columns, and 100,000 iterations allowed for optimal clustering of training molecules. In general, the assessment of molecular similarity of drug candidates produced is a critical stage in the drug discovery process [20]. Furthermore, a variety of methods have recently emerged for computing the structural similarity between molecules [20]. Tanimoto similarity, the ratio of the intersection and the union of two molecules, is widely used in cheminformatics applications [4]. Additionally, because diversity in drug candidates is necessary for reducing existing drug discovery failure rates, *de novo* drug candidates containing low shared similarity with training candidates were optimal for docking.

Table 1: QED values of top-ranking drug candidates.

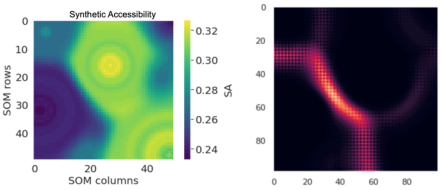
Generated Molecule ID	QED Score
1	0.93559
2	0.87985
3	0.89113
4	0.87176
5	0.82749
6	0.88119



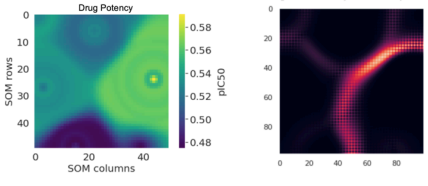
(a) Self Organizing Map and U Linear Matrix for Solubility



(b) Self Organizing Map and U Linear Matrix for Toxicity



(c) Self Organizing Map and U Linear Matrix for Synthetic Accessibility



(d) Self Organizing Map and U Linear Matrix for Drug Potency.

Figure 10: Self organizing maps for chemical descriptors for drug candidate optimization. Light regions of U Matrix represent neurons with short distances and thus clusters. Dark regions of the U Matrix represent cluster separators as they contain neurons that are separated by large distances.

$$d = min(\left\|\vec{x} - \vec{w}_{ij}\right\|) = min\sqrt{\sum_{t=0}^{n}\left[\vec{x}(t) - \vec{w}_{ij}(t)\right]^2}$$

$$W(t+1) = W(t) + \Theta(t)L(t)(V(t) - W(t))$$

$$L(t) - L_0 exp(\frac{-t}{\lambda}) \qquad \Theta(t) = exp\left(-\frac{dist^2}{2\sigma^2(t)}\right)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 11: First function depicts the Euclidean distance function used to update nodes and produce BMU (minimum of discriminant function values). Additional functions depict the adaptive process, which involves the learning decay rate and neighborhood size. Final equation depicts calculation of mean squared error.

4 MOLECULAR VALIDATION

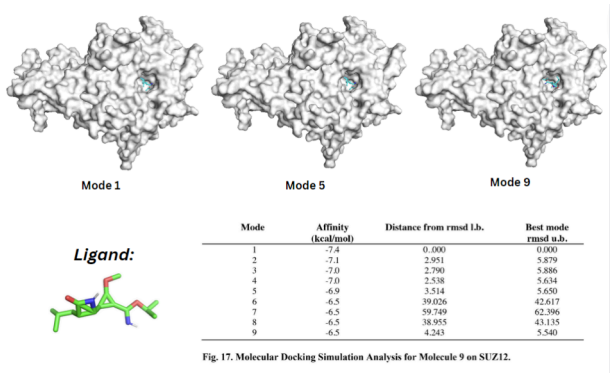


Figure 12: Produced Ligand Docked on SUZ12 Outperforms Tazmetostat's Binding Affinity of -7.1

Table 2: Molecular Docking Simulation Analyses for known Tazemetostat inhibitor. Analyses were used as a benchmark for analyzing the effect of existing yet largely ineffective inhibitors.

Mode	Affinity (kcal/mol)	Distance rmsd l.b.	Distance rmsd u.b.
1	-7.1	0.000	0.000
2	-7.0	20.020	25.009
3	-6.9	14.691	20.218
4	-6.9	7.496	16.079
5	-6.8	18.304	20.876
6	-6.8	18.559	21.217
7	-6.7	19.887	24.417
8	-6.7	14.226	19.256
9	-6.7	22.124	25.650

4.1 Molecular Docking

Upon the identification of relevant cofactors, pocket detection, and *de novo* drug design, small molecules are docked onto the relevant protein site for further assessment of their efficacy. AutoDock Vina, a traditional software for high-accuracy molecular docking of ligands to proteins, was utilized to assess the binding affinity of existing EZH2 inhibitors. To establish a reference for existing binding affinity values, existing drugs that target EZH2 such as Tazemetostat, 3-deazaneplanocin A (DZNep), and GSK126 were downloaded from the PubChem database

Ligands were converted to the PDBQT format through PyMol. Furthermore, Gasteiger charges were added, and non-polar hydrogens were merged using AutoDock Tools. Relevant pocket structures were prepared by the removal of all water molecules, addition of polar hydrogen molecules, and addition of Kollman charges, all of which are necessary for simulating accurate docking conditions. Results for molecular dynamic simulations were assessed over a range of nine binding sites in order to determine a relevant reference for inhibitors of EZH2. In order to evaluate the efficacy of the drugs produced by modifying GENTRL's reward functions, Tanimoto similarity was utilized, as minimal variance in drug structure is needed for diversity in testing. A wide array of structurally similar drug candidates in testing is conducive to a higher failure rate. Similarity was assessed by measuring overlapping chemical information over total chemical space information.

5 RESULTS AND DISCUSSION

Through its comprehensive chemical descriptor data aggregation and implementation of self-organizing map for candidate optimization, Monco expedites the *in silico* drug development process as a whole. The drug candidates in the training data parsed from the ChEMBL database contained an average Tanimoto similarity value of 0.1846. *De novo* drug candidates produced contained a Tanimoto similarity score of 0.1802, thereby allowing for similar diversity necessary for successful drug testing. [7] reports the assessment of quantitative estimate of drug-likeness scores as important in early drug development as it allows for compounds to be accurately ranked based on merit. This study also shows that the pharmaceutical strength of the drug candidate ranges from zero to one, indicating unfavorable and favorable properties respectively. The QED score of the highest-ranking drug candidate was 0.936. Furthermore, the fifth highest QED score measured among the produced drug candidates was 0.881, indicating that the drug-candidates produced from the deep learning framework are potent. Validation of generated drug compounds through AutoDock indicated comparable binding affinities in comparison to published inhibitors such as tazemetostat. Additionally, consideration of cryptic pocket sites and relative temperature factors allowed for a comprehensive understanding of protein and ligand interactions. Future work involves performing wet lab analyses for further validation of accuracy and reliability of drug candidates as well as iterative refinement of the pipeline using laboratory data. A primary limitation of this work involves the accessibility of computational power; subsequent research involves the analysis of ligand and protein interactions through robust molecular dynamics simulations such as Desmond from Schrödinger. Additionally, the role of off-target effects in the high drug discovery failure rate necessitates accounting for downstream effects in personalized patient data profiles. Furthermore, the implementation of more comprehensive chemical descriptors enables the development of novel drug molecules that are not only optimized, but also hold strong binding affinities to proteins such as SUZ12, a critical cofactor of EZH2.

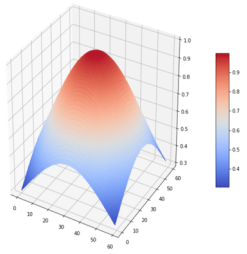


Figure 13: 3D Neighborhood Distance Weight Matrix of Potency Self Organizing Map.

Table 3: MSE results for self organizing map models. Because the values are significantly close to zero, the maps are highly accurate in organizing and predicting optimal drug properties.

Chemical Descriptor	Mean Squared Error
Potency	0.09502
Solubility	0.09612
Toxicity	0.07553
Synthetic Accessibility	0.09947

6 AUTHORS AND AFFILIATIONS

Jaagat Prashar and Jaanak Prashar are rising high school seniors and twin brothers passionate about drug discovery and health equity.

7 ACKNOWLEDGMENTS

We would like to acknowledge the industry professionals we presented our work to and received input from (no mentorship). Additionally, we would like to thank Regeneron as a category sponsor for the Regeneron ISEF (3rd Place Grand Award in Computational Biology and Bioinformatics), GENIUS Olympiad for naming us international Finalists, and the Conrad Foundation for naming us Conrad Innovators.

REFERENCES

[1] Neuroblastoma — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/books/NBK448111/. [Accessed 24-Jul-2023].

[2] and Alex Bateman, Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. D., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C. S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M. R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.-C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K. C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M. L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Mercier, P. L., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T. B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L.-S., Zhang, J., Ruch, P., and Teodoro, D. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research 49*, D1 (Nov. 2020), D480–D489.

[3] Bajorath, J. Deep machine learning for computer-aided drug design. *Frontiers in Drug Discovery 2* (Feb. 2022).

[4] Baldi, P., and Nasr, R. When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values. *Journal of Chemical Information and Modeling 50*, 7 (June 2010), 1205–1222.

[5] Barrett, J. A., Yang, W., Skolnik, S. M., Belliveau, L. M., and Patros, K. M. Discovery solubility measurement and assessment of small molecules with drug development in mind. *Drug Discovery Today 27*, 5 (2022), 1315–1325.

[6] Berrouet, C., Dorilas, N., Rejniak, K. A., and Tuncer, N. Comparison of drug inhibitory effects ($$\hbox {IC}_{50}$$) in monolayer and spheroid cultures. *Bulletin of Mathematical Biology 82*, 6 (June 2020).

[7] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry 4*, 2 (Jan. 2012), 90–98.

[8] Bownes, L. V., Williams, A. P., Marayati, R., Stafman, L. L., Markert, H., Quinn, C. H., Wadhwani, N., Aye, J. M., Stewart, J. E., Yoon, K. J., Mroczek-Musulman, E., and Beierle, E. A. EZH2 inhibition decreases neuroblastoma proliferation and in vivo tumor growth. *PLOS ONE 16*, 3 (Mar. 2021), e0246244.

[9] Chen, L., Alexe, G., Dharia, N. V., Ross, L., Iniguez, A. B., Conway, A. S., Wang, E. J., Veschi, V., Lam, N., Qi, J., Gustafson, W. C., Nasholm, N., Vazquez, F., Weir, B. A., Cowley, G. S., Ali, L. D., Pantel, S., Jiang, G., Harrington, W. F., Lee, Y., Goodale, A., Lubonja, R., Krill-Burger, J. M., Meyers, R. M., Tsherniak, A., Root, D. E., Bradner, J. E., Golub, T. R., Roberts, C. W., Hahn, W. C., Weiss, W. A., Thiele, C. J., and Stegmaier, K. CRISPR-cas9 screen reveals a MYCN-amplified neuroblastoma dependency on EZH2. *Journal of Clinical Investigation 128*, 1 (Dec. 2017), 446–462.

[10] Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences 44*, 3 (Mar. 2004), 1000–1005.

[11] Duan, R., Du, W., and Guo, W. EZH2: a novel target for cancer treatment. *Journal of Hematology &amp Oncology 13*, 1 (July 2020).

[12] Ertl, P., and Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics 1*, 1 (June 2009).

[13] Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research 40*, D1 (Sept. 2011), D1100–D1107.

[14] Goodwin, R., Bunch, J., and McGinnity, D. Chapter six - mass spectrometry imaging in oncology drug discovery. In *Applications of Mass Spectrometry Imaging to Cancer*, R. R. Drake and L. A. McDonnell, Eds., vol. 134 of *Advances in Cancer Research*. Academic Press, 2017, pp. 133–171.

[15] Guengerich, F. P. Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug Metabolism and Pharmacokinetics 26*, 1 (2011), 3–14.

[16] Gupta, R., Polaka, S., Rajpoot, K., Tekade, M., Sharma, M. C., and Tekade, R. K. Importance of toxicity testing in drug discovery and research. In *Pharmacokinetics and Toxicokinetic Considerations*. Elsevier, 2022, pp. 117–144.

[17] Hamilton, W. L., Ying, R., and Leskovec, J. Inductive representation learning on large graphs, 2017.

[18] Jamasb, A. R., Viñas, R., Ma, E. J., Harris, C., Huang, K., Hall, D., Lió, P., and Blundell, T. L. Graphein - a python library for geometric deep learning and network analysis on protein structures and interaction networks.

[19] Kiriiri, G. K., Njogu, P. M., and Mwangi, A. N. Exploring different approaches to improve the success of drug discovery and development projects: a review. *Future Journal of Pharmaceutical Sciences 6*, 1 (June 2020).

[20] Kumar, A., and Zhang, K. Y. J. Advances in the development of shape similarity methods and their application in drug discovery. *Frontiers in Chemistry 6* (July 2018).

[21] Kwon, S., Bae, H., Jo, J., and Yoon, S. Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics 20*, 1 (Oct. 2019).

[22] Lim, J., Ryu, S., Kim, J. W., and Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics 10*, 1 (July 2018).

[23] Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. L., Sausville, E. L., Lukow, D. A., Liu, L., Chait, A. R., Galluzzo, Z. C., Tucker, C., and Sheltzer, J. M. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine 11*, 509 (Sept. 2019).

[24] Liu, Q., Kwoh, C. K., and Li, J. Binding affinity prediction for protein–ligand complexes based on iβ/i contacts and b factor. *Journal of Chemical Information and Modeling 53*, 11 (Nov. 2013), 3076–3085.

[25] Medina-Franco, J. L. Grand challenges of computer-aided drug design: The road ahead. *Frontiers in Drug Discovery 1* (July 2021).

[26] Meller, A., Bhakat, S., Solieva, S., and Bowman, G. R. Accelerating cryptic pocket discovery using AlphaFold. *Journal of Chemical Theory and Computation* (Mar. 2023).

[27] Osakwe, O. The significance of discovery screening and structure optimization studies. In *Social Aspects of Drug Discovery, Development and Commercialization*. Elsevier, 2016, pp. 109–128.

[28] Pu, L., Naderi, M., Liu, T., Wu, H.-C., Mukhopadhyay, S., and Brylinski, M. eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacology and Toxicology 20*, 1 (Jan. 2019).

[29] Qi, Y., and Zhan, J. Roles of surgery in the treatment of patients with high-risk neuroblastoma in the children oncology group study: A systematic review and meta-analysis. *Frontiers in Pediatrics 9* (Oct. 2021).

[30] Schneider, P., Tanrikulu, Y., and Schneider, G. Self-organizing maps in drug discovery: Compound library design, scaffold-hopping, repurposing. *Current Medicinal Chemistry 16*, 3 (Jan. 2009), 258–266.

[31] Schoenmaker, L., Béquignon, O. J. M., Jespers, W., and van Westen, G. J. P. UnCorrupt SMILES: a novel approach to de novo design. *Journal of Cheminformatics 15*, 1 (Feb. 2023).

[32] Stairiker, C. J., Thomas, G. D., and Salek-Ardakani, S. EZH2 as a regulator of CD8 t cell fate and function. *Frontiers in Immunology 11* (Oct. 2020).

[33] Sun, D., Gao, W., Hu, H., and Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B 12*, 7 (July 2022), 3049–3062.

[34] Sun, Z., Liu, Q., Qu, G., Feng, Y., and Reetz, M. T. Utility of b-factors in protein science: Interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chemical Reviews 119*, 3 (Jan. 2019), 1626–1665.

[35] Thakur, A., Kumar, A., Sharma, V., and Mehta, V. PIC50: An open source tool for interconversion of PICsub50/subvalues and ICsub50/subfor efficient data representation and analysis.

[36] Wang, C., Liu, Z., Woo, C.-W., Li, Z., Wang, L., Wei, J. S., Marquez, V. E., Bates, S. E., Jin, Q., Khan, J., Ge, K., and Thiele, C. J. EZH2 mediates epigenetic silencing of neuroblastoma suppressor genes iCASZ1/i, iCLU/i, iRUNX3/i, and iNGFR/i. *Cancer Research 72*, 1 (Jan. 2012), 315–324.

[37] Wu, H., Zeng, H., Dong, A., Li, F., He, H., Senisterra, G., Seitova, A., Duan, S., Brown, P. J., Vedadi, M., Arrowsmith, C. H., and Schapira, M. Structure of the catalytic domain of EZH2 reveals conformational plasticity in cofactor and substrate binding sites and explains oncogenic mutations. *PLoS ONE 8*, 12 (Dec. 2013), e83737.

[38] Zhang, Y., Ye, T., Xi, H., Juhas, M., and Li, J. Deep learning driven drug discovery: Tackling severe acute respiratory syndrome coronavirus 2. *Frontiers in Microbiology 12* (Oct. 2021).

[39] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., Xing, L., Guo, T., and Aspuru-Guzik, A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology 37*, 9 (Sept. 2019), 1038–1040.

Imputation is challenging and ignoring can be worse: Learnings from activity trackers

Sharut Gupta*
MIT CSAIL

Narayan Hegde
Google Research

Srujana Merugu*
Amazon

Sriram Lakshminarasimhan
Google Research

ABSTRACT

Health studies based on activity trackers have to handle a key challenge in 'missing data'. They either discard missing data, drop participants from the study, or make use of simple interpolation techniques to fill in the missing values. The impact of missing data and the (in)accuracy of imputation can range from negligible to severe, depending on the nature of the study.

In this paper, we present the challenges faced during the imputation of missing data on multiple data types from trackers. An RNN-based approach augmented with time encodings proposed here shows strong performance overall in imputing missing heart rate values with ≈ 10-30% improvement on RMSE, MAPE, and MAE values. Unlike prior works that evaluate imputation algorithms just using RMSE and MAPE values, we demonstrate the efficacy of our imputation algorithm, not just using the standard error metrics, but also on the performance of a generic regression-based downstream application under different characteristics of missing data.

ACM Reference Format:
Sharut Gupta*, Narayan Hegde, Srujana Merugu*, and Sriram Lakshminarasimhan. 2023. Imputation is challenging and ignoring can be worse: Learnings from activity trackers. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Wearable activity trackers are being increasingly deployed to deliver digital health solutions due to their ability to detect and monitor vital physiological and activity signals such as step count, intensity, and heart rate in an affordable manner. Due to the high correlation between these activity signals and numerous preventable disease conditions such as obesity and hypertension, activity trackers have the potential to be an extremely cost-effective first-line of treatment for such diseases. Currently, thousands of studies are being performed every year to analyze the impact of mHealth (mobile health) solutions using trackers such as Garmin, Fitbit on lifestyle diseases and wellbeing.

Missing data, and more importantly, dealing with missing data, is a key challenge that emerges in these digital health studies. Data is often missing due to a variety of factors such as sensor malfunction, variation across user compliance rates and wearing behavior, poor network connectivity, and synchronization delays or failures. Since wearable compliance is critical for deriving inference and validating the accuracy of a hypothesis [11, 13], health studies address this issue through different approaches. Some studies end up removing the participants with missing data [18] (increasing recruitment and study costs), while some others interpolate missing data [8, 20] using techniques such as linear interpolation [3]. However, these approaches are often not accurate, since in most cases the data is *not missing at random*, especially in situations when participants intentionally remove the wearable during certain activities or to recharge.

The inference on the relationship between the variables of interest is heavily impacted by the approach taken for dealing with missing data. Consider the case of a downstream task that analyses the relationship between heart rate and step counts using a simple linear regression model. When the missing data is replaced by a commonly used imputation technique, the perturbation on the regression analysis becomes significant (Figure 1), even as the overall RMSE between the original and the imputed values are considered to be in an "acceptable" range. Often, it turns out that *even when the imputation results in low RMSE values, the underlying temporal correlation is wrongly inferred.*

Without high-fidelity imputations, the resulting conclusions regarding the healthcare hypothesis can be erroneous and have severe consequences downstream. Downstream tasks can be any clinical or user-facing application, which analyses data from wearables for making a lifestyle-related decision or health inference. This could finally be in a product, Health study, or in a critical, clinical setting.

Currently, there are multiple strategies used for time-series imputation: from simple fill mechanisms to averaging [10] and mining [24], to more data-driven approaches using machine learning and deep neural network models [1, 17, 25]. Most of these works however report the accuracy of imputation on metrics such as RMSE or MAPE [1, 25]. While pertinent, these metrics consider a *single marker in isolation.* None of these prior works investigate the impact on inferences related to complex downstream applications (such as healthcare) and the relationship between correlated variables (such as sleep, physical activities, stress, and heart rate). With time series data, there is an additional need to make judicious choices on a number of factors such as the temporal granularity of representation, prediction, and horizon for look-back. Often, these need to be chosen based on downstream application needs, data

*Work done while at Google Research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

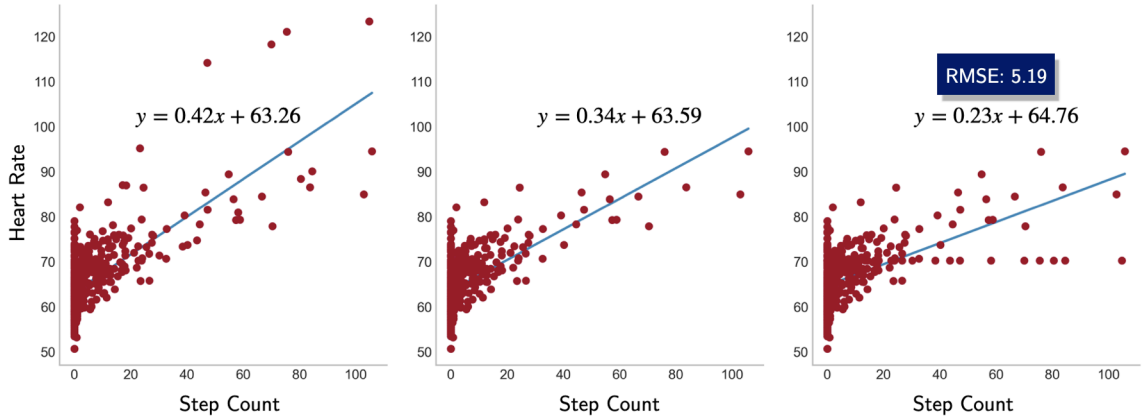


Figure 1: (Left) the regression line between between heart rate and step counts on the original data. (Middle) plot shows the same regression but with a 15% chunk of data removed. (Right) the missing values imputed using the standard backward fill algorithm, where missing values are replaced with the following non-missing value. The RMSE between the imputed and the original values is *only* 5, but more importantly, the regression line between heart rate and step count has been significantly altered.

availability, and the likely prediction performance. For instance, it is possible to achieve higher accuracy in predicting the average hourly heart rate compared to that for a 5-minute level granularity, but the latter is more useful for health alerts.

1.1 Contributions

(1) We present the varied challenges of imputing missing data on heart rate, sleep, and physical activity. We analyze practical imputation modeling challenges such as the granularity of the prediction, the amount of time to look-ahead and look-back for the imputation, etc.
(2) We propose and evaluate a generic RNN-based imputation framework that combines temporal encoding and achieves state-of-the-art results at a fine granularity of prediction with different data and missing data characteristics.
(3) We show the efficacy of our imputation algorithm using standard metrics like RMSE, MAPE, and MAE and also validate that the imputed values converge close to the original model when there is no missing data.

2 RELATED WORK

Tracker-based Health Studies: All mHealth studies have to deal with the practical realities of missing data. As stated earlier, a significant number of studies end up discarding both participants and missing records. However, some studies do incorporate imputation [16]. In one specific large-scale study on cognitive impairment [3], authors develop a *behaviorogram*, a time-aligned data channel representation of behavior, and create feature vectors and aggregations at multiple time scales (daily, minute-level, etc.). In another study of heart rate changes for pregnant women [1], authors employ a Multiple Imputation method using personalized models and medical history.

Time-series Analysis: Time-series forecasting techniques such as Vector Autoregression (VAR), ARIMA (Autoregressive integrated moving average), EWMA (Exponentially Weighted Moving Average [10]) have been around for several decades and have been analyzed in several domains such as finance, utilities, etc. for forecasting. In general, these statistical techniques have been shown to work well with univariate time series, with near-term forecasting. But on healthcare datasets with chunks of missing data used in this paper, these techniques did not outperform ML-based imputation models (refer Table 1).

Imputation with Neural Networks: Time-series imputation techniques have been explored with a variety of neural network architectures [2, 17, 26]. Some recent works have specifically worked on imputing missing heart rate values [1, 14, 25]. The HeartImp framework [14] makes use of deep learning GANs to build personalized heart rate models, and PTSI [25] makes use of both global and personal contextual information to impute heart rate values. The imputation depends on the features chosen, the granularity of imputation, and is data- and context-specific, and are not available for us to compare against.

IoT networks: Failure of sensors in IoT networks is a relatively common phenomenon and missing data is an expected occurrence. Several works have explored the usage of sliding windows and deep learning in settings of limited system resources [6, 27] to infer the missing data that would have been captured by the sensors. One work in the domain makes use of imputation using probabilistic models using spatial similarity in neighborhood sensor data [7]. The *missingness* characteristics of data in IoT environments tend to be quite different from that of personal health data with human behavior involved.

3 METHODOLOGY

To learn a user-level data imputation network, we use an encoder-decoder model [5]. This is motivated by the mere idea of encoding past and (or) future observations corresponding to a particular physiological feature, and then using this encoded context to predict

and fill missing values.

Sequence Building. Since the encoder-decoder network takes a sequence of values as input, we convert our time series data into small windows/chunks. At any given time, a window containing the historical values of the series is considered. We refer to the size of this window as Look-back (denoted as p). Specifically, at any given time t, the sequence corresponding to the k^{th} physiological marker,

$$x^k_{(t-p)}, x^k_{(t-p-1)}, ..., x^k_{(t-1)}, x^k_t$$

denotes the history or lag that the model wants to look at. Similarly, in order to leverage the observations post missing data, we also define a Look-ahead window (denoted as f). In particular, the series

$$x^k_{(t+d+1)}, x^k_{(t+d+2)}, ..., x^k_{(t+d+f-1)}, x^k_{(t+d+f)}$$

denotes the window corresponding to future observations. Here d denotes the window size corresponding to the sequence which is to be imputed. The look-back and the look-ahead windows are aggregated and used as input to the imputation network. The model is trained to predict the values across a window size d,

$$x^k_{(t+1)}, x^k_{(t+2)}, ..., x^k_{(t+d-1)}, x^k_{(t+d)}$$

which are as close as possible to the true observations. A pictorial representation of the constructed windows across the time series can be found in Figure 2.

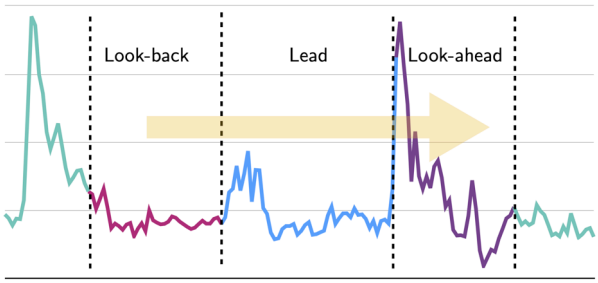


Figure 2: Windowing time series data into chunks that are to be used as input to the encoder-decoder model. Green corresponds to the original time series, pink denotes the past or look-back window, purple denotes the future or look-ahead and blue represents the window to be imputed or the lead.

Time Encodings. From a practical standpoint, different individuals may exhibit similar wearable sensory data distributions across a span of time, owing to their specific routines. For instance, an individual is more likely to go for a run/jog in the early morning or evening as compared to the afternoon. Such rhythmic variables which operate in a regular and periodic manner introduce temporal contextual signals which significantly affect the underlying sensory data distribution. To incorporate such temporal features, we encode each timestamp (second, minute, hour, week, month, and year) as a value ranging between -0.5 and 0.5. The encoded features include the minute of the hour, hour of the day, day of the week, day of the month, and week of the month.

The encoding function is given by:

$$f(x) = \frac{x}{\max(x)} - 0.5$$

, where x is the corresponding feature from the above list of features.

Encoder. The encoder of the imputation network consists of a recurrent neural network (RNN) [22] which encodes the input sequence into a representation that captures its characteristics and context. As shown in Figure 3, we use a Gated Recurrent Unit (GRU) [4] as the recurrent cell of this RNN. This cell adaptively captures dependencies of different time scales and modulates the flow of information inside the unit using two gates: the update gate (z_t) and the reset gate (r_t). Specifically, an input x_t to the unit is processed using the following mathematical operations:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\tilde{h}_t = g(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

where the gates are represented as,

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

We use a GRU over a Long Short-Term Memory (LSTM) as the basic cell of the RNN since the performance was stable and comparable across both networks, with the former being more efficient compared to the latter.

Decoder. The decoder network takes as input the representation vector encoded by the encoder network and predicts the values corresponding to the missing data. Our decoder module is symmetric in architecture to that of the encoder module. In particular, the base architecture of our decoder is an RNN with GRU as its recurrent cell. This RNN is followed by two fully connected layers with a ReLU activation function squeezed in the middle.

In addition to the encoded representation vector, we also use the encoded time embeddings corresponding to the timestamps of the prediction window of size d. The encoded representation vector along with the temporal features are sequentially passed through each GRU cell, followed by the two dense layers to output a sequence of imputed values.

In the above architecture, the decoder reads from the context vector and the time embeddings and generates multistep predictions by recursively feeding the model, predictions from the previous step in training time. Specifically, the i^{th} prediction estimate corresponding to the k^{th} variable (denoted as $\hat{x}^k_i$) is fed into the subsequent GRU unit to make the $(i+1)^{th}$ prediction estimate $\hat{x}^k_{i+1}$.

Teacher Forcing method uses the true value of data at the i^{th} step (denoted as x^k_i) to generate the next prediction $\hat{x}^k_{i+1}$. However, using this approach, errors made early in the sequence generation process are fed as input to the model and hence can be quickly amplified, because the model might be in a part of the state space it has never seen at training time. In order to smoothly bridge this gap, we use *Scheduled Sampling*, a sampling mechanism that randomly decides between $\hat{x}^k_i$ and x^k_i to use as the input to the following decoder unit.

We adapt this approach by using a custom sampling rate function which has a high probability of selecting the true values x_i^k during initial training steps (teacher forcing) and gently converges towards sampling purely from the estimates $\hat{x}_i^k$ (recursive). The teacher forcing during the initial steps helps in generating accurate short-horizon forecasts, during later steps, recursive training reduces the error accumulation across long-horizon predictions. At any step i, with a probability ϵ_i, the true value x_i^k is sampled, while with a probability $(1-\epsilon_i)^2$ the estimate comes from the previously made prediction. The probability ϵ_i is decayed with step i using the inverse sigmoid function $s_a^{-1}(x) = \frac{k}{k+\exp\frac{i}{k}}$. Here, $k \geq 1$ denotes the speed of convergence of the decay function.

4 DATA AND EXPERIMENTAL SETUP

4.1 Dataset

We use the Fitbit-Fitabase dataset [9], which contains data collected across 30 individuals from Fitbit wearable trackers over a span of two months in 2016. The data contains readings corresponding to various physiological features like physical activity, heart rate, and sleep monitoring. Physical activity is reported using the number of steps, metabolic equivalents (METs), intensity and calories burned recorded in one-minute intervals. Intensity is categorically encoded, with categories as sedentary, light activity, moderate activity, and very active state.

4.2 Training Details

Our proposed encoder-decoder module is trained end to end by minimizing the average mean squared error (L2 norm) between the true observations and the predicted imputed values. We also test the DILATE (DIstortion Loss including shApe and TimE) loss criterion [12], which aims to accurately detect sudden abrupt changes in the data. Specifically, DILATE is a weighted linear combination of shape (represented using Dynamic Time Warping - DTW) and temporal localization error (represented using the Temporal Distortion Index - TDI), which makes it suitable for improving the trajectory of forecasting. The encoder's and decoder's RNN cell is a unidirectional single-layered GRU with 128 features in the hidden layer. The decoder's RNN is followed by a dense layer with a hidden dimension of 16. The final dense layer operated on the neurons in the hidden layer to produce an output vector of length $d = 96$. We use a look-back window size of $p = 96$, a look-ahead size of $f = 48$, and impute the data for a prediction window size of $d = 96$. We use Adam optimizer with its default parameters ($\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-7}$) and a learning rate of 10^{-3} for the joint training of the encoder and the decoder network. The models are implemented in PyTorch [19]. Training is done on mini-batches of size 16, for up to 500 epochs, with an early stopping criterion on the validation error. The training stops when the validation error stops to improve for over 10 epochs.

5 RESULTS AND DISCUSSIONS

Heart Rate (HR) is a highly sensitive physiological marker that can vary significantly with slight changes in the time series data. For instance, monitoring peaks in HR serves as the primary clinical indicator for cardiovascular dysfunction and resting heart rate (RHR) has been shown to have a significant association with cardiovascular and non-cardiovascular outcomes. Even for a single person, the heart rate variation is quite high when the person is involved in any intense activity, with a long tail for the most intense activity. Considering the sensitivity and abruptness of heart rate, it is imperative to preserve any variations with maximum precision. To handle the missing values present across this physiological marker, simply removing them would certainly result in the loss of valuable information and user-specific characteristics.

Missing heart rate values could be imputed at the granularity of the tracker device (once every 15 seconds) or at an aggregate level of a day (aligning to the circadian rhythm), or at a larger timeframe in the order of days or weeks that potentially reflect people's habits and patterns. For the rest of the paper, we report the performance of our model at a 15-minute granularity.

5.1 Time Encoding for Imputation

As discussed in Section 3, time encodings (TE) construct signals which account for user-specific characteristics like routines and periodic behavior. To understand their impact on heart rate imputation, we train an Encoder-decoder framework with and without time embeddings.

With the addition of time encodings the model is able to learn the finer variations (at the minute level) in heart rate, while inducing faster convergence. In some cases, the model trained using time embeddings is able to learn the peculiarities of the data at merely the 200th round of training, while a model trained without time embeddings fails to do so even at the 500th round of training, as shown in Figure 4. In our experiments, we find that our method (either with or without TE) outperforms other models on all the users with heart rate data.

Table 1 also depicts that the encoder-decoder model with time encodings exhibits higher performance compared to the one without time encodings. This superior performance is observed across all performance metrics including MAE, RMSE, and MAPE, against two popular statistical methods: VAR and Prophet [23].
VAR is a statistical technique that models each variable as a linear combination of past values of itself, the past values of other markers in the system, and an error term.
Prophet is a univariate time series forecasting tool that is based on a decomposable additive model where non-linear trends are fit with seasonality.

Model	MAE	RMSE	MAPE
Seq2Seq with TE	**10.13**	**12.83**	**0.12**
VAR	11.58	14.71	0.14
Prophet	13.43	16.15	0.25

Table 1: Accuracy of imputation on the heart for various methods on look-back (past context) of 24 hours, prediction (output) of 24 hours at a granularity of 15 minutes on users with heart rate data.

5.2 Window Size for Imputation

We try to understand the impact of varying window sizes on the accuracy of heart rate imputation. In particular, we vary the past

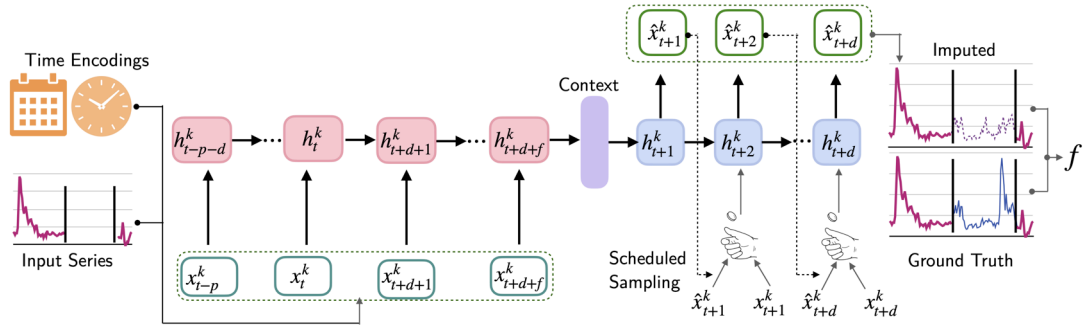


Figure 3: A schematic representing the Encoder-decoder architecture used to time series imputation. Besides the conventional training paradigm, our framework incorporates additional temporal features and scheduled sampling.

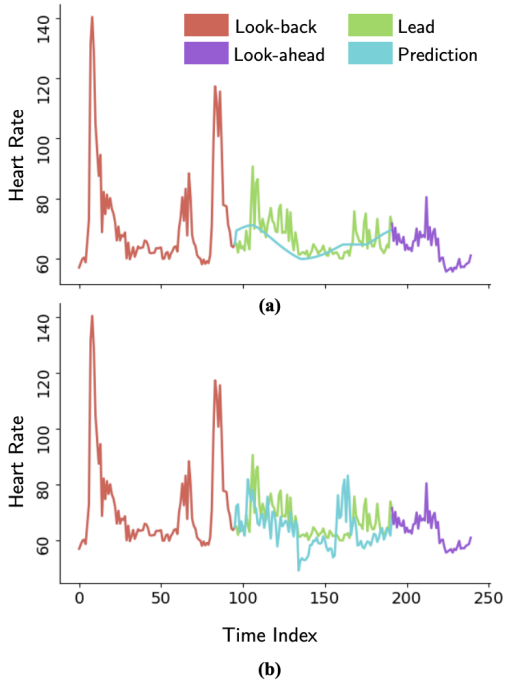


Figure 4: Performance of imputation framework (a) without temporal encoding at epoch 500 (b) with temporal encoding at epoch 200

window size also referred to as look-back (p), future window size also referred to as look-ahead (f) and prediction window size also referred to as the lead. Specifically, we vary each of these variables across long horizons (12 hours, 1 day, 2 days) and short horizons (2 hours, 3 hours, 4 hours).

Table 2 shows the performance metrics obtained from the experiments conducted on a subset of user data. Clearly, very large horizon window sizes (e.g., $p = f = 192$) lead to high RMSE, DTW and TDI metrics, indicating the inefficacy of large windows to capture the relevant spatiotemporal context. Similarly, for small window sizes ($p = 16, f = 8$), although the model achieves a decent RMSE, it fails to account for the long-term dependencies and regresses to the underlying trend in the data. Further, forecasting for such a short time span can only cater to specific types of missing data. Between the two extremes, there does exist one (or more) appropriate windows (e.g., $p = 96, f = 48, d = 96$) which not only performs well on the standard error metrics but also has learned the shape (DTW) and temporal features (TDI) in the data.

p, f, d	MAE	RMSE	MAPE	DTW	TDI
16, 8, 12	4.4	5.1	.07	0.2	1.78
16, 8, 8	4.2	5.4	.06	**.17**	0.81
48, 48, 12	5.7	7.1	.08	.27	0.86
48, 48, 24	5.9	7.4	.09	.34	4.48
48, 48, 48	5.8	6.9	.09	.46	6.15
96, 24, 24	4.3	5.6	.06	.28	2.26
96, 24, 48	**3.5**	4.88	**.05**	.37	2.77
96, 48, 8	5.2	6.4	.08	.2	0.82
96, 48, 48	4.0	5.1	.06	.38	4.8
96, 96, 24	5.0	6.3	.06	.33	2.64
96, 96, 48	5.4	6.7	.08	.55	1.17
96, 48, 96	3.6	**4.80**	**.05**	.50	**0.64**

Table 2: Performance of data imputation framework on heart rate for various past window sizes (look-back = p), prediction window sizes (lead = d), and future window sizes (look-ahead = f)

5.3 Impact on Downstream task

To simulate the impact of missing data and imputation, we construct a multiple linear regression problem whereby we use data corresponding to very high activity and step count to predict the heart rate at a 15-minute granularity. As shown in Figure 5(a, i), the hyper-plane which best fits this data can be mathematically represented as $Y = 9.03X_1 + 0.36X_2 + 63.42$ where X_1 denotes high intensity activity, X_2 denotes the step count, and Y corresponds to the heart rate. Here $\theta_1 = 9.03$ represents the regression coefficient of Y with respect to X_1 when X_2 is considered constant and $\theta_2 = 0.36$ represents the regression coefficient of Y with respect to X_1 when

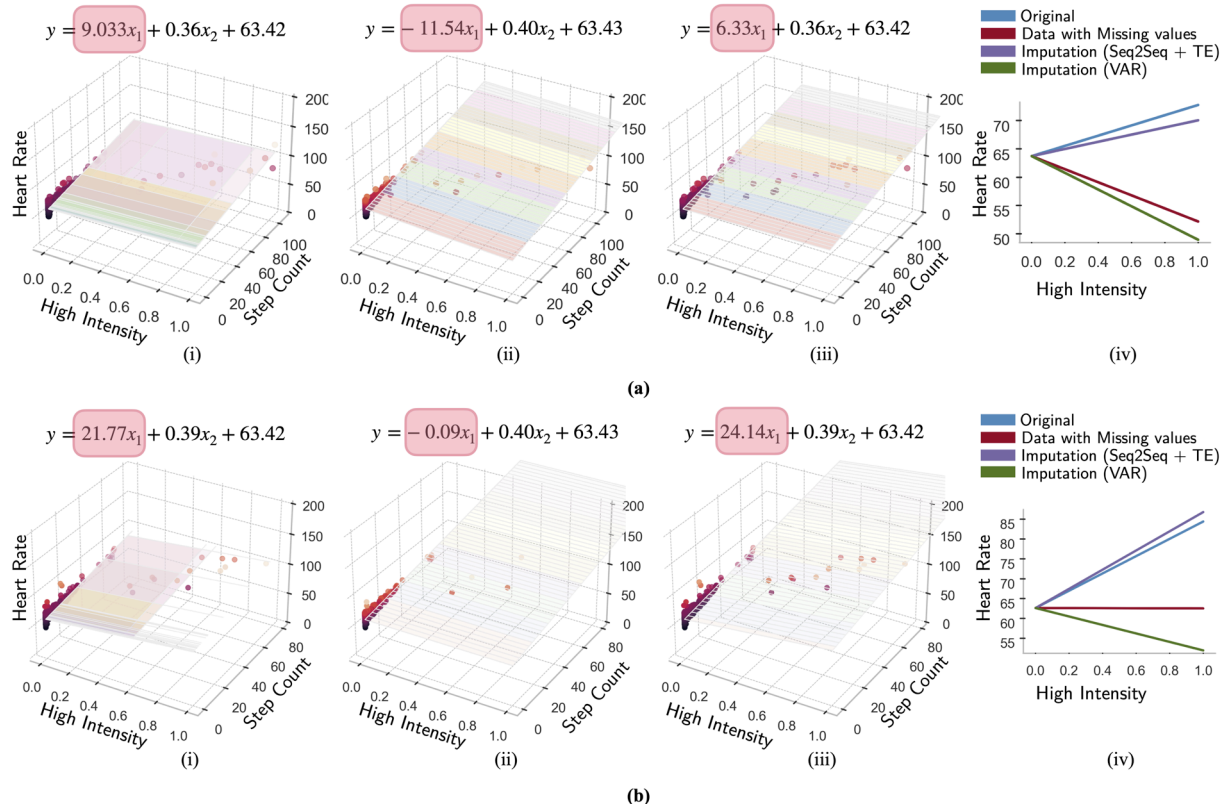


Figure 5: Analysis of the downstream task of estimating the heart rate of a random user using the step count and fraction of time (in minutes) spent in high-intensity activity. Each plot depicts a hyper-plane that best fits the (i) original data; (ii) (Top) data obtained by synthetically removing data corresponding to an entire day; (Bottom) data obtained by synthetically removing data between 4:00 PM to 7:00 PM for a week; (iii) data obtained post imputation using our imputation framework; (iv) depicts the projection of hyper-plane onto the axis corresponding to high intensity.

X_2 is considered constant. Similarly, the hyper-plane corresponding to data in Figure 5(b, i) is $Y = 21.77X_1 + 0.39X_2 + 62.62$.

Now, we synthetically induce missing information based on user compliance characteristics, as described by Rubin et al. [15, 21].

Missing Completely At Random (MCAR): While different individuals exhibit different compliance behaviors, it is possible that a highly compliant user forgets to wear his/her tracker. This would result in a large continuous chunk of missing values corresponding to that user's data spanning hours or even days. We synthetically design this type of missing information by removing data for a randomly selected user across a span of one day.

Not Missing At Random (NMAR): Some users prefer to remove their wearables when indulging in high-intensity tasks or physical activities. Such behavior can certainly result in missing information of various lengths across a day. Since it is preferred to work out during the evening or early morning, depending on the personal characteristics of the user, this missing information might be periodic and seasonal in nature. We simulate this scenario by discarding data from 4:00 PM to 7:00 PM across a week for a randomly selected user.

We analyze the shift in the fitted hyper-plane after introducing missing values based on MCAR and NMAR. As shown in Figure 5(ii), θ_1 changes from a value of 9.03 to -11.54 in MCAR and from 21.77 to -0.09 in NMAR, violating the original hypothesis that the heart rate varies positively with the variations in high-intensity activity. There is not much variation in the values of θ_2 and intercept c.

However, after imputing the missing values using our algorithm (Figure 5(iii)), we observe that θ_1 changes to 6.33 in MCAR and 24.14 corresponding to NMAR, which is very close to our original hypothesis. Figure 5(iv) shows the relationship between Intensity and heart rate after imputation using VAR. θ_1 changes from of 9.03 to -14.76 (MCAR) and from 21.77 to -10.67 (NMAR), violating our original hypothesis and underscoring the importance of an accurate imputation framework that looks beyond the standard error metrics. This analysis supports our claim to optimize imputation for downstream tasks.

6 CONCLUSION AND FUTURE WORK

Activity trackers provide the platform to conduct behavior and health studies on a large scale with open recruitment and low study cost. The data collected from these hardware devices can come

up with many problems that may lead to wrong conclusions. Our work explores the implications of missing data, which is one of the most common problems. The proposed RNN-based algorithm with temporal encoding features to impute data is accurate with standard metrics while having limited impact on a downstream task that correlates physiological markers. Future works could include jointly optimizing standard metrics and downstream tasks to develop custom imputation methods. We believe there are still a significant number of open challenges in dealing with missing values in healthcare data, its effect on downstream analysis, modeling user behavior, and dealing with class imbalances.

REFERENCES

[1] Iman Azimi, Tapio Pahikkala, Amir M. Rahmani, Hannakaisa Niela-Vilén, Anna Axelin, and Pasi Liljeberg. Missing data resilient decision-making for healthcare iot through personalization: A case study on maternal health. *Future Generation Computer Systems*, 96:297–308, 2019.

[2] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8, 04 2018.

[3] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, Marc Sunga, Han Hee Song, Hyun Joon Jung, Belle Tseng, and Andrew Trister. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD 2019, page 2145–2155, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[6] Monidipa Das and Soumya K. Ghosh. A deep-learning-based forecasting ensemble to predict missing data for remote sensing analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5228–5236, 2017.

[7] Berihun Fekade, Taras Maksymyuk, Maryan Kyryk, and Minho Jo. Probabilistic recovery of incomplete sensed data in iot. *IEEE Internet of Things Journal*, 5(4):2282–2292, 2018.

[8] Daniel Fuller, Javad Rahimipour Anaraki, Bongai Simango, Machel Rayner, Faramarz Dorani, Arastoo Bozorgi, Hui Luan, and Fabien A Basset. Predicting lying, sitting, walking and running using apple watch and fitbit data. *BMJ Open Sport & Exercise Medicine*, 7(1), 2021.

[9] Robert Furberg, Julia Brinton, Michael Keating, and Alexa Ortiz. Crowd-sourced fitbit datasets 03.12.2016-05.12.2016, May 2016.

[10] Charles C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

[11] Artie Konrad, Victoria Bellotti, Nicole Crenshaw, Simon Tucker, Les Nelson, Honglu Du, Peter Pirolli, and Steve Whittaker. Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3829–3838, New York, NY, USA, 2015. Association for Computing Machinery.

[12] Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. *Advances in neural information processing systems*, 32, 2019.

[13] I-Min Lee, Eric J. Shiroma, Masamitsu Kamada, David R. Bassett, Charles E. Matthews, and Julie E. Buring. Association of Step Volume and Intensity With All-Cause Mortality in Older Women. *JAMA Internal Medicine*, 179(8):1105–1112, 08 2019.

[14] Suwen Lin, Xian Wu, Gonzalo Martinez, and Nitesh V. Chawla. *Filling Missing Values on Wearable-Sensory Time Series Data*, pages 46–54.

[15] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[16] Todd D. Little, Terrence D. Jorgensen, Kyle M. Lang, and E. Whitney G. Moore. On the Joys of Missing Data. *Journal of Pediatric Psychology*, 39(2):151–162, 07 2013.

[17] Yonghong Luo, Xiangrui Cai, Ying ZHANG, Jun Xu, and Yuan xiaojie. Multivariate time series imputation with generative adversarial networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[18] Gonzalo J. Martinez, Stephen M. Mattingly, Shayan Mirjafari, Subigya K. Nepal, Andrew T. Campbell, Anind K. Dey, and Aaron D. Striegel. On the quality of real-world wearable data in a longitudinal study of information workers. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–6, 2020.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[20] Eujessika Rodrigues, Daniella Lima, Paulo Barbosa, Karoline Gonzaga, Ricardo Oliveira Guerra, Marcela Pimentel, Humberto Barbosa, and Álvaro Maciel. Hrv monitoring using commercial wearable devices as a health indicator for older persons during the pandemic. *Sensors*, 22(5), 2022.

[21] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[22] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536, 1986.

[23] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[24] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525, 06 2001.

[25] Xian Wu, Stephen Mattingly, Shayan Mirjafari, Chao Huang, and Nitesh V. Chawla. Personalized imputation on wearable-sensory time series via knowledge transfer. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1625–1634, New York, NY, USA, 2020. Association for Computing Machinery.

[26] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

[27] Yi-Fan Zhang, Peter J. Thorburn, Wei Xiang, and Peter Fitch. Ssim—a deep learning approach for recovering missing time series sensor data. *IEEE Internet of Things Journal*, 6(4):6618–6628, 2019.

Analyzing and Predicting Low-Listenership Trends in a Large-Scale Mobile Health Program: A Preliminary Investigation

Arshika Lalan
Google Research
arshikal@google.com

Shresth Verma
Google Research
vermashresth@google.com

Kumar Madhu Sudan
ARMMAN
madhu@armman.org

Amrita Mahale
ARMMAN
amrita@armman.org

Aparna Hegde
ARMMAN
aparnahegde@armman.org

Milind Tambe
Google Research
milindtambe@google.com

Aparna Taneja
Google Research
aparnataneja@google.com

ABSTRACT

Mobile health programs are becoming an increasingly popular medium for dissemination of health information among beneficiaries in less privileged communities. Kilkari is one of the world's largest mobile health programs which delivers time sensitive audio-messages to pregnant women and new mothers. We have been collaborating with ARMMAN, a non-profit in India which operates the Kilkari program, to identify bottlenecks to improve the efficiency of the program. In particular, we provide an initial analysis of the trajectories of beneficiaries' interaction with the mHealth program and examine elements of the program that can be potentially enhanced to boost its success. We cluster the cohort into different buckets based on listenership so as to analyze listenership patterns for each group that could help boost program success . We also demonstrate preliminary results on using historical data in a time-series prediction to identify beneficiary dropouts and enable NGOs in devising timely interventions to strengthen beneficiary retention.

KEYWORDS

Time-Series Prediction, Mobile Health

1 INTRODUCTION

Mobile Health (mHealth) programs make use of mobile phone devices to deliver healthcare services while raising awareness about critical information and knowledge that contributes to optimal health outcomes. They play a key role in making healthcare more accessible for the less privileged [5, 9, 18].

In this paper, we focus on our collaboration with ARMMAN [3], an India-based non-profit organization which conducts mHealth programs in India to increase awareness during antenatal and postpartum care for pregnant women and mothers. It has adopted mHealth interventions to assist in bringing down maternal and child mortality rates. mMitra, an initiative by ARMMAN, is a free mobile call service that delivers organized preventive care information to enrolled women throughout their pregnancy and child infancy, on a weekly/bi-weekly basis. mMitra has successfully deployed SAHELI [17], a system to efficiently utilize the limited capacity of healthcare resources to boost engagement with the program.

In 2016, the Ministry of Health & Family Welfare (MoHFW) launched the Kilkari program, a free mobile health (mHealth) education service that sends women preventive care information during pregnancy and child infancy. Kilkari is an IVR service designed to deliver weekly pre-recorded, stage-specific audio messages to pregnant women and mothers with children under the age of 1 year. Currently operational in 18 states and Union Terriroties, Kilkari has reached over 30 million women and their children to date, and has 3 million active subscribers. ARMMAN is a technical, content & creative production, and implementation partner to the MoHFW in making Kilkari available pan-India.

Like most mHealth programs [1, 7], Kilkari continues to evolve with improvements in technology and infrastructure. We discuss some unexplored questions based on these new developments in Kilkari. Our key contributions are :

- **Analyzing listenership patterns of different beneficiary buckets:** We segment our cohort of beneficiaries into four buckets based on listenership (High Pickup Rate - High Engagement Rate, High Pickup Rate - Low Engagement Rate, Low Pickup Rate - High Engagement Rate, Low Pickup Rate - Low Engagement Rate) and analyse each bucket individually. This segmentation helps us analyze the problem of listenership using two metrics - pickup and engagement - which we show are equally important. While previous works have only used pickup rates as an indication of listenership, we show that high pickup rates could be coupled with low engagement rates and thus should be taken into account to bolster program success.

KDD '23, August 06–10, 2023, Longbeach, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Analyzing the impact of time slot on call pickup and call engagement:** For our analysis, we construct seven time slots in the interval from 8 AM to 10 PM, divided into two-hour intervals. All calls, including reattempts, are put into one of these time slots. Our findings show that the beneficiaries in different buckets tend to exhibit a preference for some time slots over the others, which can be useful in boosting listenership
- **Predicting low-listeners:** Our final contribution is predicting engagement rates (the rate at which beneficiary listens to a call for more than 30 seconds) along with pickup rates (the rate at which beneficiary picks up calls) of the entire cohort using preliminary time-series modelling. We showcase how to make such predictions using only listenership trajectories and no demographic features, which can be useful in programs with sensitive and limited beneficiary information, a key characteristic of most mHealth programs. Additionally, predicting low-listenership of beneficiaries from historical data can assist NGOs in planning timely interventions to improve beneficiary retention.

Through the study of Kilkari as a real-world use case, our analysis offers insights that can be extended to other large-scale mHealth programs.

2 RELATED WORKS AND BACKGROUND

Adherence monitoring in healthcare is an extensively studied problem for diseases like HIV [16], cardiac problems [15], Tuberculosis [6, 13], etc. In mobile health programs as well, techniques such as sequential modelling [12] and markov decision process models [6, 8] have been used to improve beneficiaries adherence to audio messages.

Kilkari is the largest maternal mHealth messaging program in the world [2]. Given the scale of Kilkari's outreach, it has been a subject of multiple research studies in the past [4, 10]. Infrastructural and technological investments have led to the successful resolution of some of the key challenges discussed in previous works, while others have become less significant in the course of time.

[4] address the process of scaling Kilkari successfully across geographies. At scale, Kilkari was redesigned to make calls throughout the day as opposed to based on subscriber preference to increase cost-efficiency. An unintentional consequence was that sometimes calls were made at times that could prove unsuitable to beneficiaries. [4] also highlight certain considerations that need to be taken into account in achieving success at scale among the rural population, such as sim churn and gender gap in mobile phone access and digital literacy. [11] investigate the impact of variations in phone ownership levels among different social strata, such as income and education. The paper highlights that a successfully picked call may not guarantee that it was answered by the intended recipient, as it may have been answered by another unaware family member. Consequently, it would be beneficial to analyse call engagement (duration of call listened to) in conjunction with pickup rate to measure the outcome of the program, as opposed to analysing pickup rates exclusively. Our work represents the first attempt in predicting both pickup rates and engagement rates for program beneficiaries.

The key contribution of this paper centers around the challenges that Kilkari faces presently, in particular emphasizing predicting and contacting beneficiaries in their preferred time slots, as well as prioritizing program engagement rates along with pickup rates. Finally, we use predictive time-series modelling to predict engagement rates and pickup rates of the cohort. [14] also uses a predictive model for low listenership prediction. However, our work uses both engagement and call pickup as metrics for low listenership, which our secondary analysis from bucketing beneficiaries shows are equally important metrics. We thus learn separate binary classification models for both these target variables. Moreover, our secondary analysis also underscores factors such as time slot of placing calls and technical success rate of calls which can be important for boosting pickup and engagement rates for the low listener groups, which has not been highlighted by [14].

3 PROBLEM FORMULATION

3.1 Data Description

The Kilkari mobile health program is an outbound service that makes periodic automated voice calls to pregnant women and new mothers, starting from the second trimester of pregnancy until the time the child is one year old. One voice message is scheduled for every beneficiary in each of these 72 weeks, which contains information on topics such as maternal and child health, immunization and family planning.

Each beneficiary receives a maximum of 9 call attempts over 4 days, until they pick up the phone to increase the probability of them picking up the call. For every call attempt, the time and day of the attempt as well as the technical status is noted in the call records database. The technical status gives information on whether the call was picked up or the line was busy, switched off or out of network coverage.

In our analysis, we consider beneficiaries enrolled in the Kilkari program in the Indian state of Orissa between the year 2020-2022. Particularly, we take the set of 240K beneficiaries registered in Orissa and who received a call in the first week of 2022. We then create a dataset of their call history trajectory containing information on call attempts number, date and time of call, gestational age of beneficiary at the time of call, technical status of call and duration of the call.

3.2 Time Slot Analysis

Every week, a beneficiary receives the first call attempt on the same day as the day of their Last Menstrual Period date. If the beneficiary doesn't pick up the call, the call is attempted again.

However, the time of receiving a call in a day is currently randomized. In our analysis, we demonstrate the value in predicting and utilizing a favourable time slot for beneficiaries.

3.3 Low-Listenership Prediction

Low-listenership of beneficiaries can be characterised using various metrics. Based on discussions with domain experts, we consider two definitions of low-listenership:

(1) low-pickup rate: beneficiaries who have picked up less than 3 calls within a 6 week time window

(2) low-engagement rate: beneficiaries who have engaged with a call less than 3 times within a 6 week time window. Engagement is defined as listening to a call for more than 30 seconds (average message length is typically around 90 seconds).

Based on these definitions, we consider the following problem: Can a prediction be made in advance on when a beneficiary will become a low-listener. In such a case, early intervention by the NGO can help keep beneficiaries engaged with the health-information calls in the long run.

This low-listenership prediction problem can be formulated as a time-series prediction task. Starting at some point in time, we use history of beneficiaries' listenership trajectory for $N_{features}$ weeks to predict low-listenership N_{offset} weeks in the future. To convert the raw dataset described previously for the time series prediction task, we split listenership trajectory of all beneficiaries into multiple rolling windows. Specifically, each window is of length $N_{features} + N_{offset} + 6$ weeks where the last 6 weeks are used for defining the low-listenership binary flag. For each beneficiary, we create time series of duration of calls every week, number of attempts, status of calls, and the date and time of calls. Given these temporal features, we try to predict whether a beneficiary would have low-listenership. Note that beneficiaries enter and exit the Kilkari program at different points in time. Thus, we would get different number of time-windows for different beneficiaries.

Finally, we split the beneficiaries into train (80%) and test (20%) sets, and thus create the train and test time series datasets. Splitting on beneficiaries rather than temporally allows us to learn a prediction model that can predict low-listenership for new beneficiaries which enter the system and have at least $N_{features}$ weeks of call history.

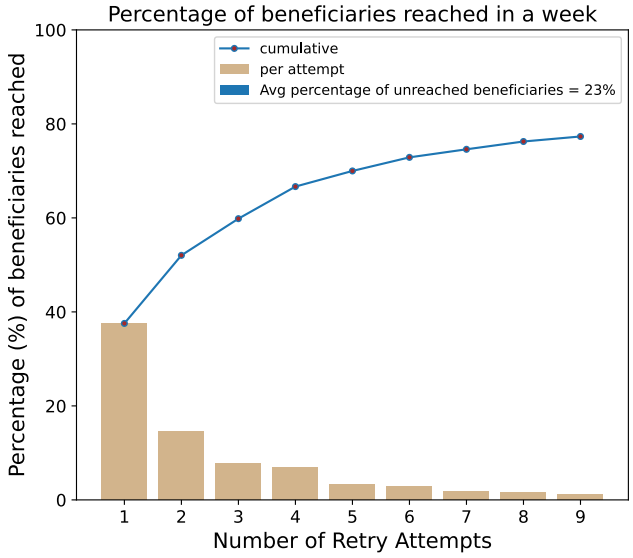


Figure 1: Efficacy of each successive attempt

Pickup \ Listenership	High	Low
High	1212	2651
Low	1421	6087

Table 1: Distribution of beneficiary in each bucket

4 EXPERIMENTS

4.1 Measuring efficacy of each successive attempt

The IVR System makes a maximum of 9 attempts to a beneficiary in a week. Fig.1 shows the percentage of beneficiaries reached with each successive attempt. The magnitude of the bars indicates the effectiveness of an attempt in reaching beneficiaries. We see a steady decrease in the percentage of beneficiaries that are reached with each successive attempt. The blue line indicates the cumulative percentage of beneficiaries reached. It can be seen that on average, 23% beneficiaries are not reached despite 9 attempts. Given the scale of the program, this is a significant number.

4.2 Analysing Listenership Patterns of Beneficiary buckets

Based on call pickup and call engagement behaviour, the beneficiaries were divided into four buckets: High Pickup Rate - High Engagement Rate, High Pickup Rate - Low Engagement Rate, Low Pickup Rate - High Engagement Rate, Low Pickup Rate - Low Engagement Rate. To avoid noisy patterns caused due to ad hoc listenership, two extreme set of beneficiaries were picked from the original data - those that have a high listening trajectory of greater than 50 weeks and those who have a lower listening trajectory of less than 20 weeks. These sets were constructed for the purpose of preliminary analysis, and we plan to expand them further in the future to include more beneficiaries. The previously mentioned four buckets were then constructed from the union of these two sets of beneficiaries. The distribution of beneficiaries per bucket is as per Table 1.

Three major conclusions can be drawn from the analysis:

- High pickup rate need not translate to high engagement.
- Contacting beneficiaries in their preferred time slots could increase listenership.
- Technical failures reduce pickup rates and could potentially lead to dropouts.

Some external factors could also affect beneficiary pick up rates and engagement. For example, there was an observed decrease in listenership of specific beneficiaries during the New Year period, which could be a possible phenomenon during most major festivities. Fig.2 illustrates this for one beneficiary. The first graph indicates which attempt day the call was picked on. The red hashed bars indicate weeks when the call was not picked up. The second plot shows whether the beneficiary engaged with the call (green bar) or not (red bar). Finally, the last graph indicates the week of the year at each Kilkari message index. We see the period of new year's starts around message index 45, around which the beneficiary doesn't pick up calls.

For figures 3, 4, 5 and 6, the six graphs display various characteristics of these aggregate groups. The first left graph shows the average pickup rate. The first right graph shows on average, what attempt day did the beneficiaries pick up the call in ('blue' if attempt day 1, 'green' if attempt day 2, 'brown' if attempt day 3, 'red' if

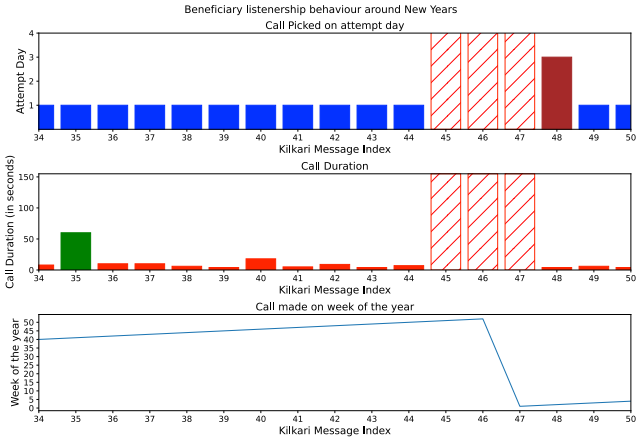


Figure 2: Listenership pattern of a beneficiary around new year - the red hashed bars indicate that the beneficiary did not pick up calls around new years

attempt day 4). The second left graph indicates the average technical success ratio (a green dot indicates a technical success ratio greater than 0.5, a red dot will indicate otherwise). The second right graph indicates which time slot did most of the beneficiaries pick up in out of the 7 time slots . Time Slot 0 is the earliest time slot (8AM-10PM) and Time Slot 6 is the latest (8PM-10PM) The third left graph indicates the call duration on average for every message index ('red' i.e non-engaging if call duration <= 30.0 seconds else 'green' i.e engaging). The third right graph indicates which day of the week did most of the beneficiaries pick up in. Day 0 indicates a Monday and Day 6 a Sunday. The color-coded scatter plots of time slot and day of the week are for easy visualisation of which points are in abundance.

4.2.1 Bucket 1: High Pickup Rate - High Engagement Rate beneficiaries. On average, this group of beneficiaries (Fig. 3) picked up the call on the first attempt day. The average technical success rate for these beneficiaries was extremely high, around 90%. Thus, high pickup rates seem to be indicative of a good technical success ratio. Most of these beneficiaries seem to pick up across diverse call time slots, with a slight preference for time slot 5 (6 PM to 8 PM).

4.2.2 Bucket 2: High Pickup Rate - Low Engagement Rate beneficiaries. The second bucket (Fig.4) comprises of beneficiaries that highlight one of the key focus points of this paper - **engagement in the program is a better metric than pickup rates to measure program success**. Despite a high average technical success rate, the engagement in the program remains low. While a predictive model for pickup rates will prefer this set of beneficiaries over beneficiaries with lower pick up rates, a predictive model for call engagement will ensure that beneficiaries more receptive to the program are given their due attention.

4.2.3 Bucket 3: Low Pickup Rate - High Engagement Rate beneficiaries. The third bucket (Fig.5) of beneficiaries tends to highly engage with the program, despite lower pickup rates. A possible reason for lower pick up rates could be the low average technical

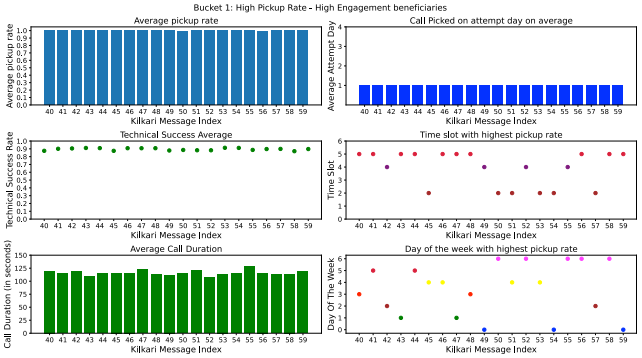


Figure 3: Listenership patterns of beneficiaries with high pick up rates and high listenership

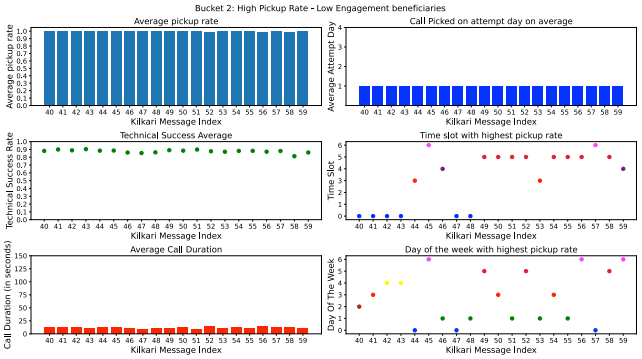


Figure 4: Listenership patterns of beneficiaries with high pick up rates and low listenership

success rates of calls. These beneficiaries also display a strong preference for the morning and evening slots, with most calls being picked up during the first two and last two slots of the day. Thus, providing a higher number of attempts to address technical failures and rearranging calls for beneficiaries in this bucket for early or late in the day can potentially improve their performance in the program .

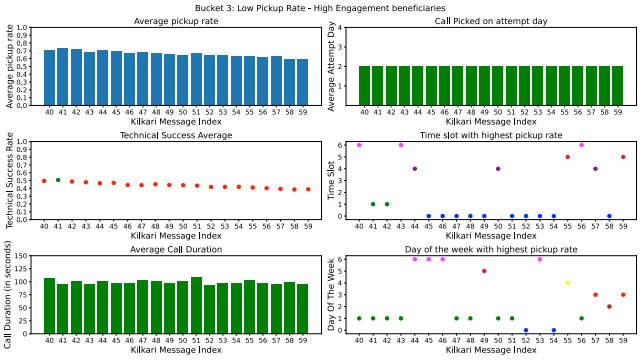


Figure 5: Listenership patterns of beneficiaries with low pick up rates and high listenership

4.2.4 Bucket 4: Low Pickup Rate - Low Engagement Rate beneficiaries. The average technical success rate for these beneficiaries (Fig. 6) is the lowest out of all the buckets, and is a meager 20%. This indicates that most calls made to these beneficiaries do not go through. This could be contributory to their low pickup rates. Finally, the beneficiaries also show a preference for the first and last time slots of the day. Addressing the correct time slots to place calls, driving engagement for calls that do connect despite lower pick up rates, and increasing the number of attempts to counter the effect of technical failures can potentially boost the performance of these beneficiaries.

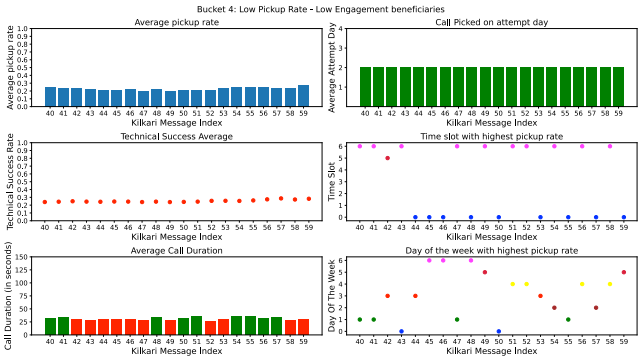


Figure 6: Listenership patterns of beneficiaries with low pick up rates and low listenership

4.3 Technical Failures and pick up rates

Certain beneficiaries are more susceptible to technical failures than others, which can affect their pickup rates and might eventually lead to them getting dropped out of the program. Dropouts are crucial in Kilkari as they save resources for the program by avoiding unnecessary attempts to phone numbers that are no longer in service, or to people who were incorrectly registered in the program. However, an unintended effect may be the possibility of dropouts in poor network coverage areas. While such issues can be addressed in due course with improvements in network coverage, it is critical to make multiple attempts to ensure maximum beneficiaries can receive technically successful calls.

Fig.7 shows the listening trajectory of a beneficiary who picks up all the calls. The first graph shows the attempt day on which the call was picked ('blue' if attempt day = 1, 'green' if attempt day = 2, 'maroon' if attempt day = 3, 'red' if attempt day = 4). The second graph shows the engagement of the beneficiary every week (green bar implies engaging and red bar implies non-engaging). The red hashed bars in both the graphs indicate calls that haven't been picked up. Finally, the final graph plots Boolean values of whether the beneficiary received technically successful (green) and failed (red) calls. Weeks with only red dots indicate that no technically successful calls were made that week.

We see that the beneficiary loses up to 5 weeks of messages due to a series of technical failures. It is interesting to note that after 6 weeks of no listenership, beneficiaries are dropped from the program.Kilkari has the flexibility of increasing the number of attempt days that beneficiaries are contacted over, hence identifying the segment of beneficiaries prone to technical failures and reaching them over a greater number of attempts can help further boost pickup rates in the program.

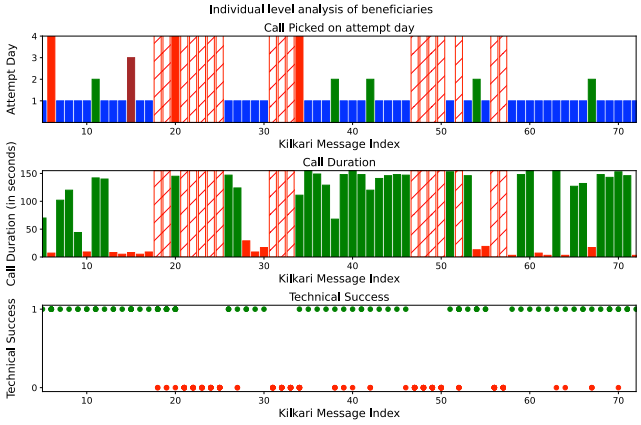


Figure 7: A beneficiary who picks up all calls except when they are complete technical failures - the beneficiary loses upto 5 weeks of messages due to technical failures

4.4 Analysis of time slot on listener pickup and engagement rates

Currently, calls in Kilkari are made at random times during the day. For analysis, we have constructed seven time slots in the interval from 8 AM to 10 PM, divided into two-hour intervals. Calling beneficiaries in their preferred time slot can help in increasing pickup and engagement rates. As seen in Fig.8, there is a clear preference in the total cohort for morning time slots, which boasts of the highest pickup rates. However, for beneficiaries with low pickup rates and low engagement, all slots perform equally poorly with the exception of the first and last time slot (Fig.9). As discussed previously, one possible reason is the issue of phone ownership, where the beneficiary has access to the phone in only the earlier or later parts of the day.

Currently, there are plans to update the IVR infrastructure to allow for maximum outreach. While bandwidth will no longer be an issue in the newer version of the IVR system, concurrency of calls still acts as a bottleneck. Thus, it might not be possible to call all beneficiaries in the morning or evening slots and leave the remaining slots with no bandwidth utilization. We aim to build a predictive model to calculate the top-k preferred time slots of beneficiaries, and optimize calling times in a way that can bring up the overall pickup and engagement rates of the cohort while working under the constraint of the number of concurrent calls that can be made at a particular time.

4.5 Predicting low-listeners

To segment beneficiaries into different clusters and to help NGOs plan for timely interventions that can increase program retention,

Model	Features	Balanced Accuracy	Precision@5	AUC
Logistic Regression	duration, attempt, status, date	0.756635	0.871566	0.828824
Logistic Regression	duration, attempt, status	0.756471	0.869925	0.828644
Feedforward NN	duration, attempt, status	0.756136	0.872381	0.828275
Feedforward NN	duration, attempt	0.756120	0.872837	0.828449
LSTM	duration, attempt	0.756097	0.876322	0.828866
Feedforward NN	duration, attempt, status, date	0.755964	0.871730	0.828090
Logistic Regression	duration, attempt	0.755885	0.874805	0.828331
LSTM	duration, attempt, status, date	0.755800	0.866317	0.827406
LSTM	duration, attempt, status	0.754932	0.865422	0.826992
Random	duration, attempt, status, date	0.501689	0.410071	0.501652
Random	duration, attempt	0.501505	0.409128	0.501464
Random	duration, attempt, status	0.500297	0.408431	0.498771

Table 2: Overall Results in predictive modeling for low-engagement prediction

Model	Features	Balanced Accuracy	Precision@5	AUC
Feedforward NN	duration, attempt, status, date	0.797638	0.430001	0.872797
Logistic Regression	duration, attempt, status, date	0.796492	0.424793	0.870911
Feedforward NN	duration, attempt, status	0.796021	0.425285	0.870253
Logistic Regression	duration, attempt, status	0.795685	0.422127	0.870115
Feedforward NN	duration, attempt	0.795241	0.420856	0.869499
Logistic Regression	duration, attempt	0.794780	0.416838	0.868858
LSTM	duration, attempt	0.794487	0.421922	0.869370
Random	duration, attempt	0.503113	0.055565	0.503159
Random	duration, attempt, status, date	0.502418	0.053555	0.500930
Random	duration, attempt, status	0.501334	0.055196	0.499087

Table 3: Overall Results in predictive modeling for low-call pickup prediction

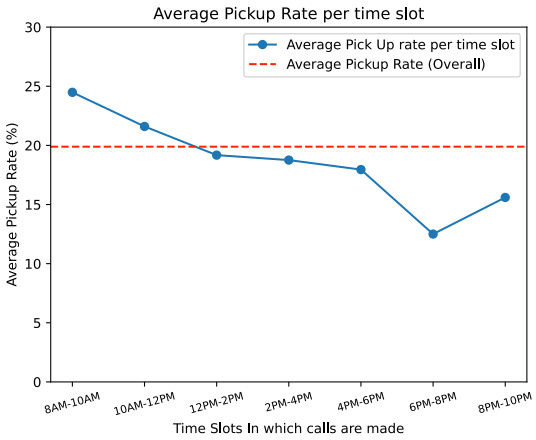


Figure 8: Pickup rates per time slot for the full cohort

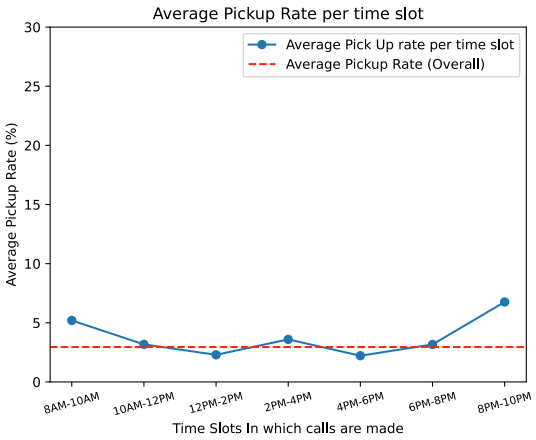


Figure 9: Pickup rates per time slot for bucket 4

we make use of beneficiaries' historical trajectories in a time-series model to predict low pickup rates and low engagement rates. Our model does not make use of any beneficiary demographic features, and is thus especially relevant for mHealth programs, where beneficary information is often sensitive and limited.

We use the dataset described in Section 3.3 to predict whether beneficiaries would showcase low-listenership behaviour in the future. Specifically, we consider two definitions of low-listenership, low-pickup and low engagement. We thus learn separate binary classification models for these two target variables. As input, we

consider $N_{train} = 6$ weeks of listenership indicators. This includes time series of duration of calls listened to every week, number of call attempts every week and number of calls with every status codes every week. We consider the following status codes that is logged in the Kilkari system: call picked up, phone busy, phone switched off, phone out of network, any other reason of not reaching. This results in a total of 7 features every week. Crucially, we develop predictive models which only use listenership behaviour data and no beneficiary specific information.

We consider the following models in our experiments:

- **Random**: This is the baseline model which predicts low listenership by sampling from uniform random distribution
- **Logistic Regression**: This uses a logistic regression model to predict the target variable. Time series features are flattened and given as input to the model
- **Feedforward NN**: This uses a dense feedforward neural network. We use 3 layers with 128 hidden units each and finally a sigmoid acitvation for output. Binary crossentropy loss is used to optimise the weights of the neural network. This model also uses flattened time series as input.
- **LSTM**: We use Keras implementation of Long Short Term Memory model for encoding sequential information from beneficiaires listenership. Specifically, we use LSTM cell with 128 hidden units followed by three layers of 128 hidden units each. Finally, we use sigmoid activation for output. and Binary crossentropy loss as optimization objective.

We evaluate all models on three metrics:

- **Precision@K%:** This is the precision in predicting low-listenership flag if we set threshold to $(100 - K)^{\text{th}}$ percentile of values. This measures the fraction of low-listeners that we will find if top $K\%$ of beneficiaires are chosen according to model predictions. Since health resources are limited, we use $(K = 5\%)$ in our experiments.
- **Balanced Accuracy:** Since we have imbalanced target distribution, we extend the notion of accuracy to imbalanced classes. Balanced accuracy is the arithmetic mean of sensitivity and specificity.
- **AUC**: We also report the area under the ROC curve (AUC) metric to compare all the models.

In table 2 and table 3, we compare different models on multiple evaluation metrics for low-engagement and low-pickup targets respectively. We notice for both the targets, ML models can perform much better than random. This showcases that input features are predictive of low listenership. However, we do not see any added value from time series model like LSTM, as opposed to logistic regression or feedforward neural network. Lastly, we find that adding additional features, such as status of calls and attempts results in slight increase in predictive performance.

5 ETHICS AND DATA USAGE

Acknowledging the responsibility associated with real-world AI systems for undeserved communities, we have closely coordinated with domain experts from the NGO throughout our analysis. This study falls into the category of secondary analysis of the aforementioned dataset. We use the previously collected engagement trajectories of different beneficiaries participating in the service call program to train the predictive model and evaluate the performance. All the data collected through the program is owned by the NGO and only the NGO is allowed to share data while the research group only accesses an anonymized version of the data.

6 CONCLUSION

In this preliminary work, we perform secondary data analysis for Kilkari, the largest maternal mobile health program in the world. We characterize beneficiaries' engagement and pickup behaviour with the program through their temporal patterns in voice message listenership. We showcase that apart from beneficiaries' call pickup rates, which has been the primary focus of several previous works, call engagement rates, preferred time slots and good call technical success ratios are also critical for successful outreach of the program. Lastly, we demonstrate that historical data can be used to predict low-listenership of beneficiaries to provide tailored approaches for beneficiary clusters and help NGOs perform timely intervention to increase beneficiary retention. Our proposed modelling approach relies only on past listenership trajectories, thus removing dependence on any sensitive or limited beneficiary information. These results also open up new line of ML research in areas such as feature engineering, intervention optimization and measuring behaviour change with better access to information such as voice messages. Our analysis and techniques can be extended to other large-scale mHealth programs.

REFERENCES

[1] Clara B Aranda-Jan, Neo Mohutsiwa-Dibe, and Svetla Loukanova. 2014. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. *BMC public health* 14, 1 (2014), 1–15.

[2] ARMMAN. [n. d.]. Kilkari. https://armman.org/kilkari/ Accessed on April 5, 2023.

[3] ARMMAN. 2019. Assessing the Impact of Mobile-based Intervention on Health Literacy among Pregnant Women in Urban India. https://armman.org/wp-content/uploads/2019/09/Sion-Study-Abstract.pdf. Accessed: 2022-08-12.

[4] Sara Chamberlain, Priyanka Dutt, Anna Godfrey, Radharani Mitra, Amnesty Elizabeth LeFevre, Kerry Scott, Jai Mendiratta, Vinod Chauhan, and Salil Arora. 2021. Ten lessons learnt: scaling and transitioning one of the largest mobile health communication programmes in the world to a national government. *BMJ Global Health* 6, Suppl 5 (2021), e005341.

[5] Yutao Guo, Deirdre A Lane, Limin Wang, Hui Zhang, Hao Wang, Wei Zhang, Jing Wen, Yunli Xing, Fang Wu, Yunlong Xia, et al. 2020. Mobile health technology to improve care for patients with atrial fibrillation. *Journal of the American College of Cardiology* 75, 13 (2020), 1523–1534.

[6] Jackson A. Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, Bistra Dilkina, and Milind Tambe. 2019. Learning to Prescribe Interventions for Tuberculosis Patients Using Digital Adherence Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Jul 2019).

[7] Samaneh Madanian, Dave T Parry, David Airehrour, and Marianne Cherrington. 2019. mHealth and big-data integration: promises for healthcare system in India. *BMJ health & care informatics* 26, 1 (2019).

[8] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12017–12025.

[9] Judith McCool, Rosie Dobson, Robyn Whittaker, and Chris Paton. 2022. Mobile health (mHealth) in low-and middle-income countries. *Annual Review of Public Health* 43 (2022), 525–539.

[10] Diwakar Mohan, Jean Juste Harrisson Bashingwa, Kerry Scott, Salil Arora, Sai Rahul, Nicola Mulder, Sara Chamberlain, and Amnesty Elizabeth LeFevre. 2022. Optimising the reach of mobile health messaging programmes: an analysis of system generated data for the Kilkari programme across 13 states in India. *BMJ Global Health* 6, Suppl 5 (2022), e009395.

[11] Diwakar Mohan, Kerry Scott, Neha Shah, Jean Juste Harrisson Bashingwa, Arpita Chakraborty, Osama Ummer, Anna Godfrey, Priyanka Dutt, Sara Chamberlain,

and Amnesty Elizabeth LeFevre. 2021. Can health information through mobile phones close the divide in health behaviours among the marginalised? An equity analysis of Kilkari in Madhya Pradesh, India. *BMJ Global Health* 6, Suppl 5 (2021), e005512.

[12] Siddharth Nishtala, Lovish Madaan, Aditya Mate, Harshavardhan Kamarthi, Anirudh Grama, Divy Thakkar, Dhyanesh Narayanan, Suresh Chaudhary, Neha Madhiwalla, Ramesh Padmanabhan, et al. 2021. Selective Intervention Planning using Restless Multi-Armed Bandits to Improve Maternal and Child Health Outcomes. *arXiv preprint arXiv:2103.09052* (2021).

[13] Louise Pilote, Jacqueline P. Tulsky, Andrew R. Zolopa, Judith A. Hahn, Gisela F. Schecter, and Andrew R. Moss. 1996. Tuberculosis Prophylaxis in the Homeless: A Trial to Improve Adherence to Referral. *Archives of Internal Medicine* 156, 2 (01 1996), 161–165.

[14] Sanket Shah, Shresth Verma, Amrita Mahale, Kumar Madhu Sudan, Aparna Hegde, Aparna Taneja, and Milind Tambe. [n. d.]. Preliminary Results in Low-Listenership Prediction in One of the Largest Mobile Health Programs in the World. ([n. d.]).

[15] Youn-Jung Son, Hong-Gee Kim, Eung-Hee Kim, Sangsup Choi, and Soo-Kyoung Lee. 2010. Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research* 16, 4 (2010), 253–259.

[16] Albert Tuldrà, Ma José Ferrer, Carmina R. Fumaz, Ramon Bayés, Roger Paredes, David M. Burger, and Bonaventura Clotet. 1999. Monitoring Adherence to HIV Therapy. *Archives of Internal Medicine* 159, 12 (06 1999), 1376–1377.

[17] Shresth Verma, Gargi Singh, Aditya S. Mate, Paritosh Verma, Sruthi Gorantala, Neha Madhiwalla, Aparna Hegde, Divy Hasmukhbhai Thakkar, Manish Jain, Milind Shashikant Tambe, and Aparna Taneja. 2023. Deployed SAHELI: Field Optimization of Intelligent RMAB for Maternal and Child Care. In *Innovative Applications of Artificial Intelligence (IAAI)*.

[18] Christopher S Wood, Michael R Thomas, Jobie Budd, Tivani P Mashamba-Thompson, Kobus Herbst, Deenan Pillay, Rosanna W Peeling, Anne M Johnson, Rachel A McKendry, and Molly M Stevens. 2019. Taking connected mobile-health diagnostics of infectious diseases to the field. *Nature* 566, 7745 (2019), 467–474.

Accurate Measures of Vaccination and Concerns of Vaccine Holdouts from Web Search Logs

Serina Chang[†]
Stanford University
serinac@cs.stanford.edu

Adam Fourney
Microsoft
adam.fourney@microsoft.com

Eric Horvitz
Microsoft
horvitz@microsoft.com

ABSTRACT

To design effective vaccine policies, policymakers need detailed data about who has been vaccinated, who is holding out, and why. However, existing data in the US are insufficient: reported vaccination rates are often delayed or missing, and surveys of vaccine hesitancy are limited by high-level questions and self-report biases. Here, we show how large-scale search engine logs and machine learning can be leveraged to fill these gaps and provide novel insights about vaccine intentions and behaviors. First, we develop a *vaccine intent classifier* that can accurately detect when a user is seeking the COVID-19 vaccine on search. Our classifier demonstrates strong agreement with CDC vaccination rates, with correlations above 0.86, and estimates vaccine intent rates to the level of ZIP codes in real time, allowing us to pinpoint more granular trends in vaccine seeking across regions, demographics, and time. To investigate vaccine hesitancy, we use our classifier to identify two groups, *vaccine early adopters* and *vaccine holdouts*. We find that holdouts, compared to early adopters matched on covariates, are 69% more likely to click on untrusted news sites. Furthermore, we organize 25,000 vaccine-related URLs into a hierarchical ontology of vaccine concerns, and we find that holdouts are far more concerned about vaccine requirements, vaccine development and approval, and vaccine myths, and even within holdouts, concerns vary significantly across demographic groups. Finally, we explore the temporal dynamics of vaccine concerns and vaccine seeking, and find that key indicators emerge when individuals convert from holding out to preparing to accept the vaccine.

KEYWORDS

COVID-19, vaccination, search logs, graph machine learning

1 INTRODUCTION

COVID-19 vaccines provide significant protection against severe cases of SARS-CoV-2 [46, 59], yet a large portion of the United States remains unvaccinated. Effective vaccine policies—for example, where to place vaccine sites [49, 74], how to communicate about the vaccine [18, 72], and how to design campaigns to reach unvaccinated populations [5, 22, 60]—rely on detailed data about who is seeking vaccination, who is holding out, and why. However, existing data are insufficient [43]. Reported vaccination rates are frequently delayed [2], missing at the county-level and below [70], and missing essential demographic data [33, 42]. Surveys provide a starting point for understanding vaccine hesitancy but are often limited by high-level questions [16], small or biased samples [13, 71], and self-reporting biases (e.g., recall or social desirability bias) [3, 66] especially in sensitive contexts such as vaccination [36].

Here, we demonstrate how large-scale search logs from Bing and machine learning (ML) can be leveraged to fill these gaps, enabling fine-grained estimation of vaccine rates and discovering the concerns of vaccine holdouts from their search interests. While search logs are powerful, with widespread coverage, real-time signals, and access to personal interests, the vast amounts of data they provide are unlabeled and unstructured, consisting of billions of natural language queries and clicks on search results. To derive meaning from these queries and clicks, we first impose structure by constructing *query-click graphs*, which encode aggregated query-click patterns as bipartite networks. Second, using a combination of semi-supervised graph ML techniques and manual annotation, we develop two computational resources that enable us to extract vaccine behaviors from large unlabeled search logs.

First, we develop a *vaccine intent classifier* that can accurately detect when a user is seeking the COVID-19 vaccine on search. Our classifier achieves areas under the receiver operating characteristic curve (AUCs) above 0.90 on held-out vaccine intent labels in all states, and demonstrates strong agreement with CDC vaccination rates across states ($r = 0.86$) and over time ($r = 0.89$). Using our classifier, we can estimate vaccine intent rates to the level of ZIP code tabulation areas (ZCTAs), approximately 10x the granularity of counties and preceding lags in reporting. We carefully correct for bias in our estimates from non-uniform Bing coverage, and demonstrate minimal additional bias from our classifier, as it achieves equivalent true and false positive rates across regions.

Second, we construct a novel *ontology of COVID-19 vaccine concerns* on search. Our ontology consists of 25,000 vaccine-related URLs, clicked on by Bing users, that we organize into a hierarchy of vaccine concerns from eight top categories to 36 subcategories to 156 low-level URL clusters. Unlike surveys, our ontology discovers these concerns directly from users' expressed interests and explores them at multiple scales. Furthermore, by measuring individuals' interest in each concern from their clicks, we capture revealed preferences, side-stepping potential biases in self-reporting [24, 66].

Combining our ontology with the vaccine intent classifier allows us to conduct a thorough analysis of how individuals' vaccine concerns relate to whether they decide to seek the vaccine. We use our classifier to identify two groups of users—vaccine early adopters and vaccine holdouts—and compare their search behaviors. We identify significant differences in their vaccine concerns and news consumption; for example, compared to early adopters matched on covariates, vaccine holdouts are 69% more likely to click on untrusted news sites. We find that vaccine concerns also differ significantly even within holdouts, varying across demographic groups. Finally, we analyze the temporal dynamics of vaccine concerns and vaccine seeking, and discover that individuals exhibit

[†] Research performed during an internship at Microsoft.

telltale shifts in vaccine concerns when they eventually convert from holding out to preparing to accept the vaccine.

Our contributions can be summarized as follows:

(1) A novel vaccine intent classifier, developed with graph ML and human annotation, that achieves AUCs above 0.9 on all states and strong agreement with CDC vaccination rates;
(2) Bias-corrected estimates of vaccine intent rates from our classifier, including estimates for over 20,000 ZCTAs;
(3) A hierarchical ontology of COVID-19 vaccine concerns, including 25,000 URLs clicked on by Bing users, 156 URL clusters, 36 subcategories, and eight top categories;
(4) Analyses of vaccine holdouts' search concerns and news consumption, comparing to early adopters and studying dynamics over time.

We are publicly releasing our code, vaccine estimates, and ontology.[1] We hope that our resources, methods, and analyses can provide researchers and public health agencies with valuable insights about vaccine behaviors, helping to guide more effective, data-driven interventions.

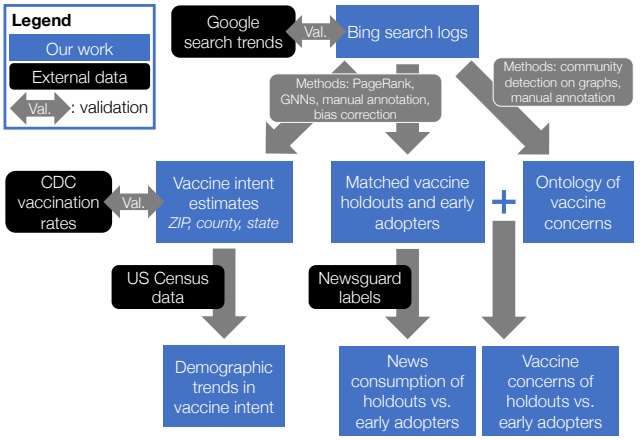


Figure 1: Our work integrates a variety of datasets and methods to analyze vaccine behaviors from search logs.

2 DATA

Our work uses a variety of datasets, including Bing search logs, CDC vaccination rates, US Census data, and Newsguard labels (Figure 1). Bing is the second largest search engine worldwide and in the US, with a US market share of around 6% on all platforms and around 11% on desktop [65]. Despite having non-uniform coverage across the US, Bing has enough penetration in the US that we can estimate representative samples after applying inverse proportional weighting (Section 4). The Bing data we use consist of individual queries made by users, where for each query, we have information including the text of the query, an anonymized ID of the user, the timestamp, the estimated geolocation (ZIP code, county, and state), and the set of URLs clicked on, if any. Since our work is motivated by insufficient vaccine data and vaccine concerns in the US, we limit our study to search logs in the US market. However, the methods we introduce could be extended to study vaccination rates and vaccine concerns in other languages and countries. We apply our vaccine intent classifier (Section 3) to all Bing search logs in the US from February 1 to August 31, 2021.[2]

To evaluate our vaccine intent classifier, we compare it to vaccination rates reported by the CDC (Section 4). The CDC provides daily vaccination rates at the levels of states [27] and counties [26]. CDC data are essential but limited, with a substantial portion of county-level data missing. These limitations serve as one of the motivations of our work, since we hope that our vaccine intent classifier can serve as a complementary resource to monitor vaccination rates, especially in smaller regions. To characterize demographic trends in vaccine intent, we use data from the US Census' 2020 5-year American Community Survey [15]. To capture political lean, we use county-level data from the 2020 US presidential election [53]. To quantify the trustworthiness of different news sites, we use labels from Newsguard [52]. Finally, to evaluate the representativeness of Bing search trends, we compare them to Google search trends, which are publicly available online [34].

Data ethics. Our work was approved by the Microsoft IRB office and by an internal privacy review process which included officers from both Microsoft Research and the Bing product team. When we use search logs, we are mindful of the need to balance privacy and social benefits when using potentially sensitive user data. While we study individual search logs, since we need to be able to link individual vaccine outcomes (as predicted by our classifier) to search interests, those sessions are assembled using only anonymous user identifiers, which are disassociated from any specific user accounts or user profiles, and cannot be linked to any other Microsoft products. Likewise, in this anonymous view of the logs, location and demographic data were limited to ZIP code-level accuracy. Finally, we are careful to only report results aggregated over thousands of individuals. Aside from Bing search logs, all of the data sources we use are publicly available and aggregated over many individuals.

3 VACCINE INTENT CLASSIFIER

Our first goal is to develop a classifier that can accurately detect when a search user is expressing vaccine intent, i.e., trying to get the COVID-19 vaccine (e.g., book an appointment or find a location). Detecting vaccine intent requires precision: for example, if a user issues the query [covid vaccine], they may be trying to get the vaccine, but they could also be generally curious about vaccine information or eligibility. Thus, we begin by defining a set of regular expressions that allow us to identify vaccine intent queries, i.e., queries that *unambiguously* express vaccine intent. To be included, the query must include both a COVID-19 term ("covid" or "coronavirus") and a vaccine term ("vaccin", "vax", "johnson", etc.). In addition, the query must satisfy at least one of the following criteria: (1) matching some variant of "find me a COVID-19 vaccine", (2) containing appointment-related words or location-seeking words, (3) containing a pharmacy name.

[1] https://github.com/microsoft/vaccine_search_study.
[2] February 2021 was the earliest that we could study following data protection guidelines, which allow us to store and analyze search logs up to 18 months in the past. We end in August 2021, since the FDA approved booster shots in September and our method is not designed to disambiguate between vaccine seeking for the primary series versus boosters.

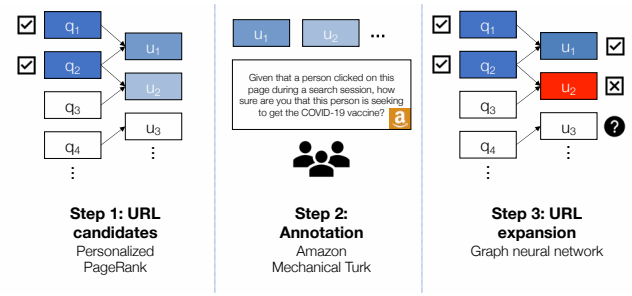


Figure 2: Our pipeline of methods to identify a large, high-precision set of vaccine intent URLs.

However, in addition to maintaining high precision, we seek to detect as many users as possible who have expressed vaccine intent, so that we have sufficient statistical power for our downstream analyses. Since our search logs contain both queries and clicks, we lose the opportunity to detect many more users if we only detect vaccine intent based on queries. For example, a user may issue the ambiguous query [covid vaccine], but then click on the URL for the CVS COVID-19 vaccine registration page, thus clarifying their intent through their clicks [61]. The challenge with URLs is that they are less formulaic than queries, so we cannot easily define regular expressions to identify URLs expressing vaccine intent.

Our key insight is that, while we cannot use regular expressions to identify URLs, we can use them to identify vaccine intent queries and then use those queries to identify URLs, based on common query-click patterns. For example, vaccine intent queries such as [cvs covid vaccine] or [covid vaccine near me] may result in clicks on the CVS COVID-19 vaccine registration page. To capture these patterns, we construct *query-click graphs* [20, 45], which are bipartite networks between queries and URLs where an edge from a query to a URL indicates how often this query is followed by a click on this URL. Specifically, we construct a query-click graph per US state, aggregating over queries and clicks from two representative months in our study period (April and August 2021). Then, our pipeline proceeds in three steps (Figure 2): first, we use personalized PageRank to propagate labels from queries to URLs, so that we can generate a set of URL candidates (Section 3.1); next, we present the URL candidates to annotators on Amazon Mechanical Turk to label as vaccine intent or not (Section 3.2); finally, we use those labels to train graph neural networks (GNNs) so that we can further expand our set of vaccine intent URLs (Section 3.3).

3.1 Personalized PageRank for URL candidates

Personalized PageRank [14] is a common technique for seed expansion, where a set of seed nodes in a graph are identified as members of a community, and one wishes to expand from that set to identify more community members [40]. In our case, the vaccine intent queries act as our seed set, and our goal is to spread the influence from the seed set over the rest of the query-click graph. Given a seed set S, personalized PageRank derives a score for each node in the graph that represents the probability of landing on that node when running random walks from S.

State	URL
CA	https://myturn.ca.gov/ https://www.cvs.com/immunizations/covid-19-vaccine https://www.goodrx.com/covid-19/walgreens https://www.costco.com/covid-vaccine.html https://www.walgreens.com/topic/promotion/covid-vaccine.jsp
NY	https://covid19vaccine.health.ny.gov/ https://www.cvs.com/immunizations/covid-19-vaccine https://www.walgreens.com/topic/promotion/covid-vaccine.jsp https://vaccinefinder.nyc.gov/ https://www.goodrx.com/covid-19/walgreens
TX	https://www.cvs.com/immunizations/covid-19-vaccine https://vaccine.heb.com/ https://www.walgreens.com/topic/promotion/covid-vaccine.jsp https://corporate.walmart.com/covid-vaccine https://dshs.texas.gov/covidvaccine/
FL	https://www.publix.com/covid-vaccine https://www.cvs.com/immunizations/covid-19-vaccine https://www.walgreens.com/topic/promotion/covid-vaccine.jsp https://floridahealthcovid19.gov/vaccines/ https://www.goodrx.com/covid-19/walgreens

Table 1: Top 5 URLs from Personalized PageRank (S-PPR) for the four largest states in the US.

We run personalized PageRank from the seed set of vaccine intent queries (S-PRR) to derive scores for all URLs in each query-click graph. Then, we order the URLs from each state according to their S-PPR ranking and keep the *union* over states of their top 100 URLs as our set of URL candidates, resulting in 2,483 candidates. The number of URLs we have in the union is much lower than the number of states multiplied by 100, since there is overlap between states. However, there is also substantial heterogeneity in top URLs across states, reflecting state-specific vaccine programs and policies (Table 1). By constructing separate graphs and running S-PPR per state, our approach is uniquely able to capture this state-specific heterogeneity. In supplementary experiments, we show that an alternative approach that uses a combined graph over states severely hurts performance for small states (Section A2.2).

S-PPR also provides scores for all queries in the graph, but we found that the seed set was comprehensive in identifying vaccine intent queries. The top-ranked queries that were not in the seed set tended to be location-specific, such as [covid vaccine new york], which is suggestive of vaccine intent but not unambiguous enough. Thus, in the subsequent steps of annotation and GNN expansion, we only seek to add URLs, and consider regular expressions sufficient for identifying queries. However, we also selected a sample of regular expression-detected queries to present to annotators, to validate whether they were truly vaccine intent. To capture a diverse sample, we use the union over the top 5 and bottom 5 queries per state (ranked by S-PPR), after filtering out queries that were issued by fewer than 50 users, resulting in 227 queries to label.

3.2 Annotation on Amazon Mechanical Turk

In this step, we present our URL candidates (and sampled queries) to annotators on AMT. The first question we ask is, "Given that a person clicked on this page during a search session, how sure

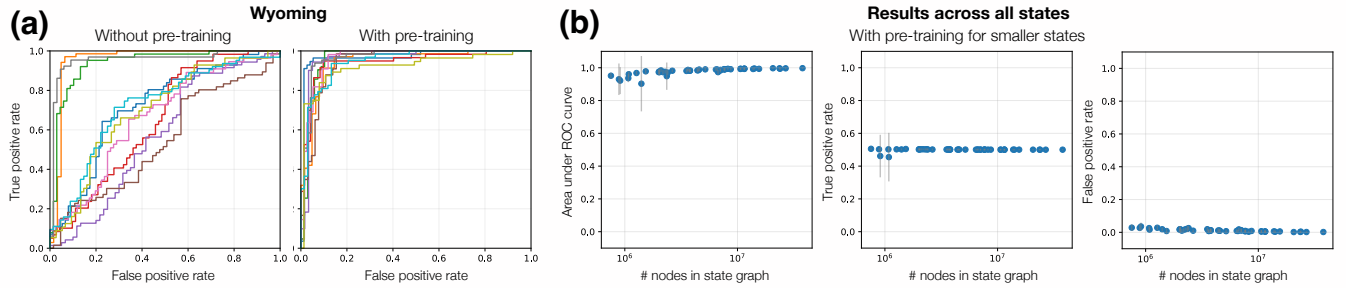


Figure 3: (a) GNN results with and without pre-training for Wyoming, one of the smallest states. Each line represents one of 10 random trials. (b) Final GNN results for all 50 states, with pre-training for smaller states. Each dot represents a state, with its y-coordinate representing the mean metric over 10 trials and grey bars indicating standard deviation.

are you that this person is seeking to get the COVID-19 vaccine?", with answer options from Highly Likely to Unlikely. Second, we ask them to indicate what other intention(s) the person might have, such as seeking information about vaccine safety. While we use the answers to the first question to construct our vaccine intent labels, we include the second question for precision, since annotators are encouraged to think of alternative intents and should only label positively for vaccine intent if the URL seems unambiguous. In the query version of our task, we ask very similar questions but replace language about clicking on the page with issuing the query.

For each URL, we first present it to three annotators. If all three give it a positive label (i.e., Highly Likely or Likely), then we label this URL as vaccine intent. If two give it a positive label and one does not, we assign it to one more annotator, and label it as vaccine intent if that annotator gives a positive label. In other words, we require vaccine intent URLs to receive three positive annotations. With this relatively strict bar, we still find that a large majority (86%) of our URL candidates are labeled as vaccine intent. Furthermore, we observe a clear relationship between S-PPR rank and the percentage labeled as vaccine intent: for example, around 90% of URLs from ranks 0 to 20, around 81% of URLs from ranks 40-60, and around 71% of URLs from ranks 80 to 100 (Figure A2). We also find a very high positive rate (96%) among the queries that we tested, thus validating our regular expressions.

3.3 Graph neural networks for expansion

Since manual annotation is expensive, we wish to augment our efforts by training ML models on the AMT labels, then use the models to expand our set of vaccine intent URLs. We formulate this problem as semi-supervised node classification on a graph, since the URLs are nodes in the query-click graph and we are trying to predict whether a URL indicates vaccine intent or not, given labels for a subset of URLs. In this section, we provide an overview of our modeling procedure, with details in Section A1.

GNN architecture and training. To solve this problem, we design a GNN [39] that consists of character-level convolutions (CNN) and graph convolutions. We use the CNNs to capture textual information in the queries and URLs, since text can be informative for this problem (e.g., the appearance of "vaccine"). The graph convolutions allow us to learn representations of URLs that draw from the representations of their neighboring queries, which draw from the representations of their neighboring URLs, and so on. In this way, we can capture "similar" URLs in embedding space (similar in terms of both text and graph structure).

To train and test our model, we randomly split the URL labels into a train set (60%), validation set (15%), and test set (25%). However, some states have much smaller graphs, and therefore, fewer positive and negative labels. For example, for Wyoming, we only have 245 positive and 276 negative URLs. We find that with such few labels, the model cannot adequately learn how to predict vaccine intent, with AUCs far below those of large states (Table A1). To address this issue, we *pre-train* the model on S-PPR rankings, which requires no additional supervision. Our intuition is that S-PPR already performed remarkably well at predicting vaccine intent, as we discussed in the prior section. Furthermore, S-PPR rankings do not require any manual labels; we derive them entirely from our initial vaccine intent queries, which were automatically labeled using regular expressions. This pre-training encourages the model to learn URL representations that are predictive of S-PPR rankings, which we find help substantially with predicting vaccine intent.

Evaluating GNN performance. We evaluate model performance by computing its AUC on the held-out test set. Furthermore, to account for randomness from model training and data splitting, we run 10 random trials for every model/state, where in each trial, we re-split the URL labels, retrain the model on the train set, and re-evaluate the model's performance on the test set. First, we find that pre-training significantly improves performance for the smaller states; for example, the mean AUC for Wyoming increases from 0.74 to 0.95 (Figure 3a, Table A1). We find that pre-training seems unnecessary for the larger states, such as Connecticut and Tennesssee, where we are already achieving high AUCs above 0.98. After incorporating pre-training for smaller states (fewer than 5,000,000 nodes), we are able to achieve AUCs above 0.90 for all 50 states and above 0.95 for 45 states (Figure 3b).

Discovering new vaccine intent URLs. Finally, we use our trained GNNs to identify new vaccine intent URLs. In order to decide which new URLs to include, we need a score threshold. Our goal is to set the threshold such that any URL that scores above it is very likely to truly be vaccine intent (i.e., we want to maintain high precision). Borrowing the idea of "spies" from positive-unlabeled learning [8], our idea is to use the held-out positive URLs in the test set to

determine where to set the threshold. We consider two thresholds: (1) t_{med}, the median score of the held-out positive URLs, and (2) t_{prec}, the minimum threshold required to achieve precision of at least 0.9 on the held-out test set. Then, we only include URLs that pass both thresholds in at least 6 out of the 10 random trials. Even with this strict threshold, we discover around 11,400 new URLs (Table A2), increasing our number of vaccine intent URLs by 10x. In the following section, we also evaluate the impact of adding these URLs on our ability to estimate regional vaccine intent rates. We find that the new URLs not only increase our coverage of vaccine intent users by 1.5x but also further improve our agreement with reported vaccination rates from the CDC (Table 2).

4 ESTIMATING VACCINE INTENT RATES

Using our classifier, we can estimate regional rates of vaccine intent. In this section, we discuss how we correct for bias in our estimates, validate against CDC vaccination rates, and use our estimates to derive insights about fine-grained vaccination trends.

Bias evaluation. In Section A2, we decompose potential bias in our approach into two key sources: first, bias from non-uniform Bing coverage, and second, bias from non-uniform true positive rates (TPR) and false positive rates (FPR) of our classifier. We show that, if we can correct for non-uniform Bing coverage and show that our classifier's TPRs and FPRs do not significantly differ across regions, our vaccine intent estimates should, theoretically, form unbiased estimates of true vaccination rates. We evaluate our classifier's TPRs and FPRs on held-out vaccine intent labels, using the same score threshold we used for discovering new vaccine intent URLs. We find that our classifier does indeed achieve statistically equivalent TPRs and FPRs across states (Figure 3b), suggesting that our classifier contributes minimal additional bias. We discuss below how we correct for non-uniform Bing coverage. Additionally, to evaluate the representativeness of Bing data, we compare search trends for vaccine intent queries between Google and Bing and find that, even before applying corrections to Bing data, the trends are highly correlated (Figure A4).

Estimating coverage-corrected rates. When we apply our classifier to Bing search logs from Feburary 1 to August 31, 2021, we find 7.45 million "active" Bing users who expressed vaccine intent through their queries or clicks. We focus on active Bing users, i.e., those who issued at least 30 queries in a month, since we can reliably assign them to a location based on their mode ZIP code (or county or state) from those queries. Given a ZCTA z, we compute $N(\hat{v}, z)$, the number of active Bing users from z for whom we detect vaccine intent. Furthermore, we estimate the ZCTA's Bing coverage as $\frac{N(b,z)}{N(z)}$, where $N(b, z)$ is its average number of active Bing users over the months in our study period and $N(z)$ is its population size from the 2020 5-year American Community Survey [15]. Then, our coverage-corrected vaccine intent estimate $\tilde{p}(v, z)$ for ZCTA z is

$$\tilde{p}(v,z) = \frac{\frac{N(\hat{v},z)}{N(z)}}{\frac{N(b,z)}{N(z)}} = \frac{N(\hat{v},z)}{N(b,z)}.$$

To estimate the vaccine intent rate for a set Z of ZCTAs, e.g., a state or county, we simply take the population-weighted average.

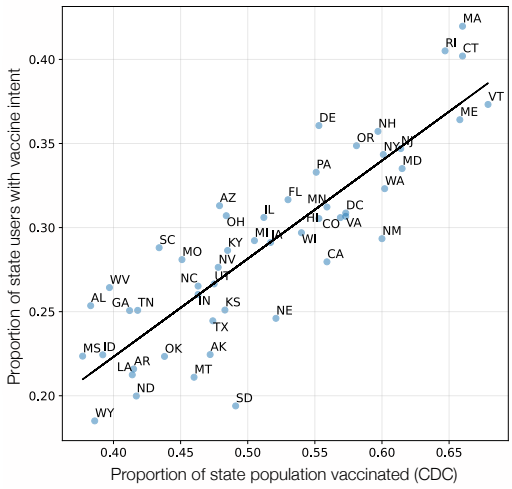


Figure 4: Comparing CDC state vaccination rates vs. estimated vaccine intent rates from Bing search logs.

Pipeline step	CDC corr.	# vaccine intent users
Only queries	0.62	3.18M
+manual URLs	0.80	4.95M
+manual and GNN URLs	0.86	7.45M

Table 2: Each step of our classification pipeline (Section 3) improves both our correlation with CDC vaccination rates and our coverage of vaccine intent users.

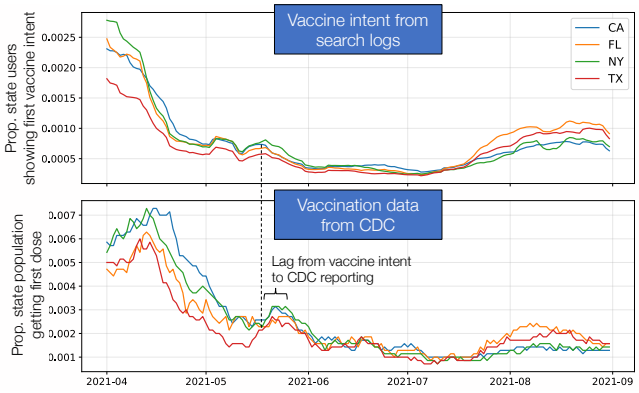


Figure 5: Rates over time of first vaccine intent (top) vs. first dose from CDC (bottom) for the four largest states in the US.

Comparison to CDC vaccination data. When we compare our vaccine intent estimates to state-level vaccination rates from the CDC, we observe strong correlation ($r = 0.86$) on cumulative rates at the end of August 2021 (Figure 4). Notably, we find that the correlation drops to $r = 0.79$ if we do not correct for Bing coverage in our estimates. Furthermore, we find that each step of our classification pipeline—only using queries from regular expressions, incorporating manually annotated URLs from personalized PageRank and AMT, incorporating URLs found by GNNs—improves both

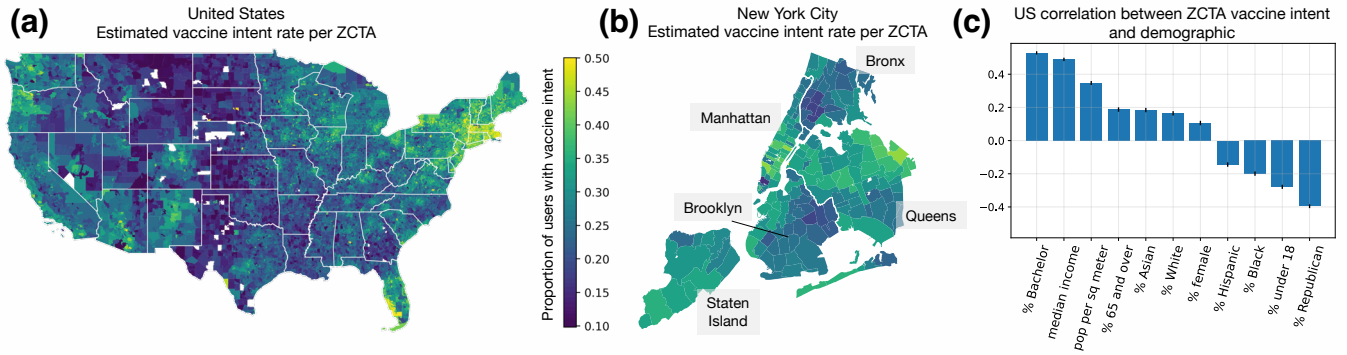


Figure 6: (a) Using our classifier, we can estimate vaccine intent rates per ZCTA, approximately 10x the granularity of counties. (b) Zooming in on New York City shows that estimated vaccine intent rates vary substantially across ZCTAs, even within the same city or county. (c) Correlations between ZCTA vaccine intent rates and demographic variables.

our correlation with CDC rates and the number of users we are able to identify (Table 2). Notably, if we only use queries, the correlation drops to $r = 0.62$ and we lose 57% of the users we identified with our full classifier, demonstrating the value of adding vaccine intent URLs through our graph ML framework.

Additionally, we compare our vaccine intent estimates to the CDC's vaccination rates over time. We observe strong correlations here as well, especially if we allow the CDC time series to lag behind the vaccine intent time series (Figure 5). With lags of 7-15 days (IQR), the median correlation over states reaches $r = 0.89$; without a lag, the median correlation drops to $r = 0.78$. The CDC's lag demonstrates an advantage of our classifier, as it can detect vaccine seeking in real time without delays from reporting.

Granular trends in vaccine seeking. Our vaccine intent classifier allows us to pinpoint who was seeking the COVID-19 vaccine, where, and when. We estimate cumulative vaccine intent rates up to the end of August 2021 at the level of ZCTAs (Figure 6a), approximately 10x the granularity of counties, which is the finest-grained vaccination data the CDC provides and, still, with many counties missing or having incomplete data [70]. We observe substantial heterogeneity in vaccine intent at the ZCTA-level, even within the same states and counties. For example, when we focus on New York City, we see that Manhattan and Queens have higher vaccine intent rates, and within Queens, ZCTAs in the northern half have higher rates (Figure 6b), aligning with reported local vaccination rates in New York City [11].

We can also use our estimates to characterize demographic trends in vaccination. When we measure correlations between ZCTA vaccine intent rate and different demographic variables, we find that overall demographic trends from our estimates align closely with prior literature [37, 41, 71, 76]. For example, we observe strong positive correlations with education, income, and population density, and a strong negative correlation with percent Republican (Figure 6c). However, we discover more nuanced trends when we look closer. Demographic trends vary significantly across states (Figure A5), especially for race and ethnicity, and trends change over time. For example, we estimate that older ZCTAs were much likelier to seek the vaccine early in 2021 but this trend fell over time (Figure A6a), reflecting how the US vaccine rollout initially prioritized seniors [38], and we see an increase in vaccine intent from more Republican ZCTAs in summer 2021 (Figure A6b). Thus, our classifier both confirms existing findings and enables new analyses with finer granularity across regions, demographics, and time.

5 SEARCH CONCERNS OF HOLDOUTS

We use our vaccine intent classifier to identify two groups: *vaccine early adopters*, who expressed their first vaccine intent before May 2021, and *vaccine holdouts*, who waited until July 2021 to show their first vaccine intent, despite becoming eligible by April.[3] Comparing the search interests of these two groups allows us to discover relationships between expressed vaccine concerns, news consumption, and vaccine decision-making. To reduce potential confounding, we match each holdout with a unique early adopter from the same county and with a similar average query count, since we know that the populations seeking vaccination changed over time and we do not want our comparisons to be overpowered by regional or demographic differences. In our following analyses, we compare the search interests of the matched sets, with over 200,000 pairs.

Vaccine holdouts are more likely to consume untrusted news. First, we analyze the trustworthiness of news sites clicked on by vaccine holdouts versus early adopters. We use ratings from Newsguard, which assigns trust scores to news sites based on criteria such as how often the site publishes false content and how it handles the difference between news and opinion [52]. We find that, in the period while vaccine holdouts were eligible but still holding out (April to June 2021), holdouts were 69% (95% CI, 67%-70%) likelier than their matched early adopters to click on untrusted news, defined by Newsguard as domains with trust scores below 60. Furthermore, we see that as the trust score from Newsguard degrades, the likelier it was that holdouts clicked on the site, relative to early adopters (Figure 7a). For example, sites that are known for spreading COVID-19 misinformation, such as Infowars [25], RT [6], and Mercola [31], were much likelier to be clicked on by holdouts.

[3]We did not consider as holdouts those who never showed vaccine intent during our study period, since those users may have gotten their vaccine in ways that are not visible via search data. In comparison, individuals who did not show their first vaccine intent until July 2021 likely did not receive the vaccine before.

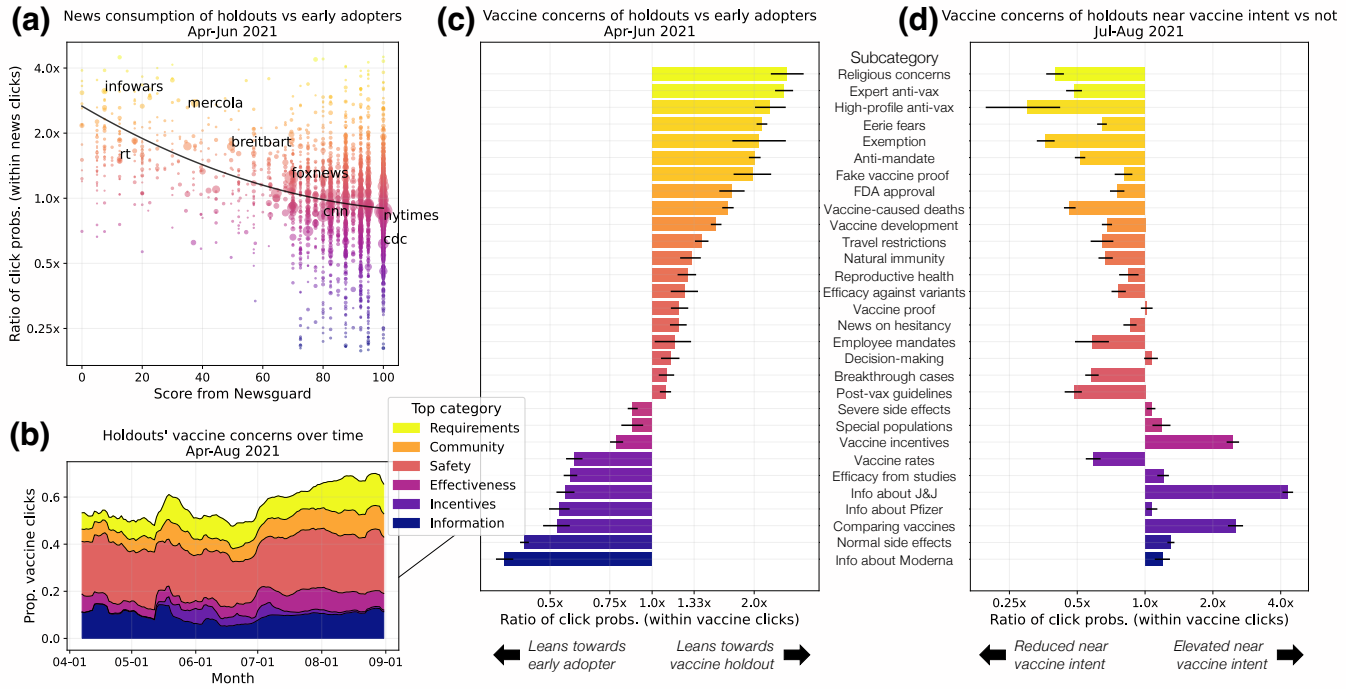


Figure 7: In all subfigures, news/categories are colored from yellow to dark purple to represent most holdout-leaning to most early adopter-leaning. (a) The lower the trust rating from Newsguard, the likelier it is that vaccine holdouts click on the news site, relative to early adopters. (b) Holdouts' top category concerns include Vaccine Safety, Requirements, and Information, with varying proportions over time. (c) Comparing holdouts vs. early adopters' relative probabilities of clicking on each subcategory (from April to June 2021) reveals each group's distinctive concerns. (d) Near when holdouts express vaccine intent (±3 days) in July and August 2021, their concerns become much more like the concerns of early adopters, with a few important differences.

Ontology of vaccine concerns on search. To characterize vaccine-related search interests in far more detail, we construct a hierarchical ontology of vaccine concerns, defined in terms of 25,000 vaccine-related URLs that were clicked on by early adopters or holdouts. We construct our ontology from the bottom-up: first, we seek to automatically partition the URLs into clusters. Leveraging graph ML again, we formulate this as a community detection problem on graphs, and apply the Louvain algorithm [12] to the collapsed URL-URL graph (collapsing the bipartite query-click graph over queries). We find that this approach results in remarkably coherent clusters (Table A3), due to the strength of the signal contained in query-click graphs, and outperforms standard topic modeling approaches such as LDA [10]. Based on these clusters, we design a comprehensive set of subcategories and top categories, and sort the clusters accordingly. For example, we identify one cluster of news stories announcing vaccine passport requirements in cities, which we sort under the proof of vaccination subcategory and Vaccine Requirements top category. This bottom-up approach allows us to discover and measure vaccine concerns directly from users' search interests and analyze them at multiple scales, providing complementary insights to more traditional surveys.

In Figure A1, we summarize our resulting ontology, which consists of 8 top categories and 36 subcategories. Some top categories encompass a number of distinct subcategories: for example, under Vaccine Safety, we include normal side effects, severe side effects, concerns about reproductive health, vaccine history and development, FDA approval, fear of vaccine-caused deaths, and "eerie" fears (e.g., myths about vaccine shedding or becoming magnetic [28]). At the top category-level, we find that vaccine holdouts are, by far, the most concerned about Vaccine Safety, which accounts for 23% of their vaccine-related clicks, followed by Vaccine Information (10%) and Vaccine Requirements (9%). We also observe changes in interests over time (Figure 7b): for example, interest in Vaccine Incentives increased in May 2021, and interest in Vaccine Effectiveness grew in June 2021, following the spread of the Delta variant.

Distinctive concerns of holdouts vs. early adopters. Our ontology allows us to compare the vaccine concerns of holdouts and their matched early adopters. First, during the period from April to June 2021, we find that holdouts were 48% less likely than early adopters to click on any vaccine-related URL. Furthermore, their distribution of concerns within their vaccine-related clicks differed significantly (Figure 7c). Using the subcategories from our ontology, we find that holdouts were far more interested in religious concerns about the vaccine; anti-vaccine messages from experts and high-profile figures; avoiding vaccine requirements by seeking exemptions, banning mandates, or obtaining fake proof of vaccination; eerie fears and vaccine-caused deaths; and FDA approval and vaccine development. In comparison, early adopters were much more concerned

about normal side effects, vaccine efficacy, comparing different types of vaccines, and information about each vaccine (Moderna, Pfizer, and Johnson & Johnson). These differences reveal the importance of a fine-grained ontology; for example, at the top category level, we would see that both groups were interested in Vaccine Safety but miss that early adopters were more concerned about normal and severe side effects, while holdouts were more concerned about eerie fears and vaccine-caused deaths. Our approach also allows us to study *who* is expressing these concerns in greater granularity. Even within holdouts, we observe significant variability in concerns across demographic groups (Figure A7). For example, holdouts from more Democrat-leaning ZCTAs were particularly concerned about FDA approval and vaccine requirements, while holdouts from more Republican-leaning ZCTAs were more concerned about eerie fears and vaccine incentives.

Holdouts appear like early adopters when seeking the vaccine. In our final analysis, we exploit the fact that all of our vaccine holdouts eventually expressed vaccine intent to explore how vaccine concerns change as an individual converts from holdout to adopter. From July to August 2021, we analyze how holdouts' vaccine concerns change in the small window (±3 days) surrounding their expressed vaccine intent, compared to their typical concerns outside of that window. We find that in those windows, holdouts' vaccine concerns nearly reverse, such that they look much more like early adopters than their typical selves (Figure 7d nearly reverses 7c). During this time, holdouts become far more interested in the Johnson & Johnson vaccine, comparing different vaccines, and vaccine incentives, and less interested in anti-vaccine messages and vaccine fears. Notably, not all early adopter-leaning concerns reverse as dramatically; for example, even while expressing vaccine intent, holdouts remain less interested in the Pfizer and Moderna vaccines, which may reflect how vaccine hesitant individuals were quicker to accept the one-shot Johnson & Johnson vaccine, instead of the two-shot mRNA vaccines [21, 73]. Furthermore, there are some early adopter-leaning concerns that holdouts do not pick up on during this time, such as interest in vaccine rates. We hypothesize that these concerns are more reflective of an early adopter "persona" rather than of concerns that would become relevant when seeking the vaccine, such as comparing different vaccines.

6 RELATED WORK

Our work centers Bing search logs, which have been used to study other health issues such as shifts in needs and disparities in information access during the pandemic [67, 68], health information needs in developing nations [1], experiences around cancer diagnoses [55, 56], concerns rising during pregnancy [29], and medical anxieties associated with online search [75]. Our efforts build on prior work that extracts insights about the COVID-19 vaccine from digital traces, such as social media [50, 57, 58] and aggregated search trends [7, 23, 48]. Our work is also related to other efforts to detect health conditions online, such as predicting depression from social media [19] and monitoring influenza from search queries [32].

Our work seeks to address the challenges of working with digital traces [24, 54] and limitations of prior work [32, 44] by developing ML and human-in-the-loop methods to precisely label search logs and evaluate bias. Furthermore, as one of the first works to use *individual* search logs to study the COVID-19 vaccine, we have the rare opportunity to link vaccine outcomes (predicted by our classifier) to the same individual's search interests. Our graph ML pipeline is also similar to other "big data" approaches that, due to the scale of unlabeled data, manually annotate a subset of data, train machine learning models to accurately predict those labels, then use those models to label the rest of the data [17, 30, 35, 47]. We extend this approach in several ways, such as by using personalized PageRank to select URLs for more efficient annotation and by setting a strict classification threshold based on "spies" to ensure high precision.

7 DISCUSSION

We have demonstrated how large-scale search logs and machine learning can be leveraged for fine-grained, real-time monitoring of vaccine intent rates and identification of individuals' concerns about vaccines. There are limitations to our approach: for example, while we can achieve finer granularity than existing data, we still miss within-ZCTA heterogeneity in vaccine intent. Furthermore, our efforts to minimize bias in our estimates are substantial but imperfect (e.g., we can only approximate TPRs and FPRs of our classifier). We also assume in this work that vaccine intent can be detected through single queries or clicks, but more sophisticated models could incorporate entire search sessions or browsing data beyond search. However, in favor of simplicity and considerations of privacy, we label vaccine intent at the query and click-level.

Despite these limitations, our resources demonstrate strong agreement with existing data and enable analyses that have not been available before. For example, our fine-grained vaccine intent estimates can help public health officials to identify under-vaccinated communities, informing where to place vaccine sites or whom to prioritize in online or real-world outreach programs. Furthermore, our novel ontology and analyses of individuals' vaccine concerns inform how to intervene, guiding messaging strategies for different holdout populations. Lastly, our observation that holdouts resemble early adopters when they eventually seek vaccination indicates that individuals might follow similar paths towards vaccine acceptance. Future work could model these trajectories, try to identify key influences (e.g., vaccine mandates), and use these models to ideally allocate limited resources for interventions.

To facilitate policy impact and future research, we are releasing our vaccine intent estimates and our ontology of vaccine concerns. We hope that these resources will be useful for conducting detailed analyses of COVID-19 vaccine behaviors and vaccination rates. The ontology can also be employed widely in web and social media research; for example, to study how certain classes of URLs (e.g., eerie fears) are disseminated on social media or surfaced by search engines. Finally, we note that our graph ML techniques for intent detection are applicable beyond vaccines, and could be applied to precisely detect other intents of interest, such as seeking stimulus checks or COVID-19 tests. More broadly, we hope that our work can serve as a roadmap for researchers of how to derive rigorous behavioral and health insights from search logs, including how to precisely detect user intents and interests, evaluate and correct for bias, validate against external data, and release resources to promote reproducibility, transparency, and future work.

REFERENCES

[1] Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M. Small, and H. Andrew Schwartz. 2019. Using Search Queries to Understand Health Information Needs in Africa. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM '19)*.

[2] Yasmeen Abutaleb and Lena H. Sun. 2021. How CDC data problems put the U.S. behind on the delta variant. *The Washington Post* (2021). https://www.washingtonpost.com/health/2021/08/18/cdc-data-delay-delta-variant/.

[3] Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare* 9 (2016), 211–217.

[4] Emily Anthes, Madeleine Ngo, and Eileen Sullivan. 2021. Adults in all U.S. states are now eligible for vaccination, hitting Biden's target. Half have had at least one dose. *The New York Times* (2021). https://www.nytimes.com/2021/04/19/world/adults-eligible-covid-vaccine.html.

[5] Susan Athey, Kristen Grabarz, Michael Luca, and Nils Wernerfelt. 2023. Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines. *Proceedings of the National Academy of Science (PNAS)* 120, 5 (2023).

[6] Julian E. Barnes. 2021. Russian Disinformation Targets Vaccines and the Biden Administration. *The New York Times* (2021). https://www.nytimes.com/2021/08/05/us/politics/covid-vaccines-russian-disinformation.html.

[7] Shailesh Bavadekar, Adam Boulanger, John Davis, Damien Desfontaines, Evgeniy Gabrilovich, Krishna Gadepalli, Badih Ghazi, Tague Griffith, Jai Gupta, Chaitanya Kamath, et al. 2021. Google COVID-19 Vaccination Search Insights: Anonymization Process Description. *arXiv* (2021).

[8] Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: a survey. *Machine Learning* 109 (2020), 719–760.

[9] Alexis Benveniste. 2021. New York City will require vaccines for entry to restaurants and gyms. *CNN Business* (2021). https://www.cnn.com/2021/08/03/business/new-york-city-vaccine-requirements/index.html.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* (2003), 993–1022.

[11] Matthew Bloch, Larry Buchanan, and Josh Holder. 2021. See Who Has Been Vaccinated So Far in New York City. *The New York Times* (2021). https://www.nytimes.com/interactive/2021/03/26/nyregion/nyc-vaccination-rates-map.html.

[12] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (2008).

[13] Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. 2021. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* 600 (2021), 695–700.

[14] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* (1998).

[15] United States Census Bureau. 2020. American Community Survey Data. https://www.census.gov/programs-surveys/acs/data.html.

[16] United States Census Bureau. 2021. Household Pulse Survey COVID-19 Vaccination Tracker. https://www.census.gov/library/visualizations/interactive/household-pulse-survey-covid-19-vaccination-tracker.html.

[17] Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Science (PNAS)* 119, 31 (2022).

[18] Wen-Ying Sylvia Chou and Alexandra Budenz. 2020. Considering Emotion in COVID-19 Vaccine Communication: Addressing Vaccine Hesitancy and Fostering Vaccine Confidence. *Health Communication* 35, 14 (2020), 1718–1722.

[19] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM'13)*.

[20] Nick Craswell and Martin Szummer. 2007. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*.

[21] Bob Curley. 2021. Why Some People Still Prefer the Johnson & Johnson COVID-19 Vaccine. *Healthline* (2021). https://www.healthline.com/health-news/why-some-people-still-prefer-the-johnson-johnson-covid-19-vaccine.

[22] Hengchen Dai, Silvia Saccardo, Maria A. Han, Lily Roh, Naveen Raja, Sitaram Vangala, Hardikkumar Modi, Shital Pandya, Michael Sloyan, and Daniel M. Croymans. 2021. Behavioural nudges increase COVID-19 vaccinations. *Nature* 597 (2021), 404–409.

[23] Parris Diaz, Pritika Reddy, Reshna Ramasahayam, Manish Kuchakulla, and Ranjith Ramasamy. 2021. COVID-19 vaccine hesitancy linked to increased internet search queries for side effects on fertility potential in the initial rollout phase following Emergency Use Authorization. *Andrologia* 53, 9 (2021).

[24] Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, and Jaime Teevan. 2014. Understanding User Behavior Through Log Data and Analysis. In *Ways of Knowing in HCI*. Springer New York, New York, NY, 349–372.

[25] Luis Ferré-Sadurní and Jesse McKinley. 2020. Alex Jones Is Told to Stop Selling Sham Anti-Coronavirus Toothpaste. *The New York Times* (2020). https://www.nytimes.com/2020/03/13/nyregion/alex-jones-coronavirus-cure.html.

[26] Centers for Disease Control and Prevention. 2023. COVID-19 Vaccinations in the United States,County. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh.

[27] Centers for Disease Control and Prevention. 2023. COVID-19 Vaccinations in the United States,Jurisdiction. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc.

[28] Centers for Disease Control and Prevention. 2023. Myths and Facts about COVID-19 Vaccines. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/facts.html.

[29] Adam Fourney, Ryen W. White, and Eric Horvitz. 2015. Exploring Time-Dependent Concerns about Pregnancy and Childbirth from Search Logs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. 737–746.

[30] Matt Franchi, J.D. Zamfirescu-Pereira, Wendy Ju, and Emma Pierson. 2023. Detecting disparities in police deployments using dashcam data. In *Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency 2023 (FAccT'23)*.

[31] Sheera Frenkel. 2021. The Most Influential Spreader of Coronavirus Misinformation Online. *The New York Times* (2021). https://www.nytimes.com/2021/07/24/technology/joseph-mercola-coronavirus-misinformation-online.html.

[32] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457 (2009), 1012–1014.

[33] Alice Goldfarb and Kara W. Schechtman. 2021. State-Level Vaccine Demographic Data is Messy and Incomplete—We Need Federal Data, Now. *The COVID Tracking Project* (2021). https://covidtracking.com/analysis-updates/state-level-vaccine-demographic-data-is-messy-and-incomplete.

[34] Google. 2023. Google Trends. https://trends.google.com/trends/?geo=US.

[35] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2021. Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science* 24 (2021), 395–419.

[36] Rodrigo Jiménez-García, Valentín Hernandez-Barrera, Cristina Rodríguez-Rieiro, Pilar Carrasco Garrido, Ana López de Andres, Isabel Jimenez-Trujillo, María D Esteban-Vasallo, Maria Felicitas Domínguez-Berjón, Javier de Miguel-Diez, and Jenaro Astray-Mochales. 2014. Comparison of self-report influenza vaccination coverage with data from a population based computerized vaccination registry and factors associated with discordance. *Vaccine* 32, 35 (2014), 4386–4392.

[37] Ashish Joshi, Mahima Kaur, Ritika Kaur, Ashoo Grover, Denis Nash, and Ayman El-Mohandes. 2021. Predictors of COVID-19 Vaccine Acceptance, Intention, and Hesitancy: A Scoping Review. *Frontiers in Public Health* 9 (2021).

[38] Berkeley Lovelace Jr. 2021. CDC expands Covid vaccination guidelines to everyone 65 and older. *CNBC* (2021). https://www.cnbc.com/2021/01/12/covid-vaccine-trump-administration-to-expand-eligibility-to-everyone-65-and-older.html.

[39] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR '17)*.

[40] Isabel M. Kloumann and Jon M. Kleinberg. 2014. Community Membership Identification from Small Seed Sets. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 1366–1375.

[41] Sarah Kreps, Sandip Prasad, John S. Brownstein, Yulin Hswen, Brian T. Garibaldi, Baobao Zhang, and Douglas L. Kriner. 2020. Factors Associated With US Adults' Likelihood of Accepting COVID-19 Vaccination. *JAMA Network Open* 3, 10 (2020), e2025594–e2025594.

[42] Nancy Krieger, Pamela D Waterman, Jarvis T Chen, Christian Testa, and William P Hanage. 2021. Missing again: US racial and ethnic data for COVID-19 vaccination. *The Lancet* 397, 10281 (2021), 1259–1260.

[43] Sharon LaFraniere. 2022. 'Very Harmful' Lack of Data Blunts U.S. Response to Outbreaks. *The New York Times* (2022). https://www.nytimes.com/2022/09/20/us/politics/covid-data-outbreaks.html.

[44] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343 (2014), 1203–1205.

[45] Xiao Li, Ye-Yi Wang, and Alex Acero. 2008. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*.

[46] Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, Natalie Groves, Gavin Dabrera, et al. 2021. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New England Journal of Medicine* 385, 7 (2021), 585–594.

[47] Ian Lundberg, Jennie E. Brand, and Nanum Jeon. 2022. Researcher reasoning meets computational capacity: Machine learning for social science. *Social Science Research* 108 (2022), 102807.

[48] Sean Malahy, Mimi Sun, Keith Spangler, Jessica Leibler, Kevin Lane, Shailesh Bavadekar, Chaitanya Kamath, Akim Kumok, Yuantong Sun, Jai Gupta, et al. 2021. Vaccine Search Patterns Provide Insights into Vaccination Intent. *arXiv* (2021).

[49] Zakaria Mehrab, Mandy L. Wilson, Serina Chang, Galen Harrison, Bryan Lewis, Alex Telionis, Justin Crow, Dennis Kim, Scott Spillmann, Kate Peters, Jure Leskovec, and Madhav Marathe. 2022. Data-Driven Real-Time Strategic Placement of Mobile Vaccine Distribution Sites. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (IAAI'22)*.

[50] Goran Muric, Yusong Wu, and Emilio Ferrara. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health and Surveillance* 7, 11 (2021).

[51] Nambi Ndugga, Latoya Hill, Samantha Artiga, and Sweta Haldar. 2021. Latest data on COVID-19 vaccinations by race/ethnicity. *Kaiser Family Found (KFF)* (2021). https://covid-19archive.org/files/original/f90f767bdd1cd10911587853d70a6320f29bf9b7.pdf.

[52] Newsguard. 2022. Rating Process and Criteria. https://www.newsguardtech.com/ratings/rating-process-criteria/.

[53] Dave Leip's Atlas of U.S. Elections. 2022. Store - Election Data. https://uselectionatlas.org/BOTTOM/store_data.php.

[54] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019).

[55] Michael J. Paul, Ryen W. White, and Eric Horvitz. 2015. Diagnoses, decisions, and outcomes: Web search as decision support for cancer. In *Proceedings of the 24th international conference on World Wide Web (WWW'15)*.

[56] Michael J. Paul, Ryen W. White, and Eric Horvitz. 2016. Search and Breast Cancer: On Episodic Shifts of Attention over Life Histories of an Illness. *ACM Transactions on the Web* 10, 2 (2016).

[57] Francesco Pierri, Brea L. Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports* 12, 5955 (2022).

[58] Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of Change: Impact of COVID-19 on Vaccine-Related Opinions of Twitter Users. In *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM'22)*.

[59] Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, et al. 2020. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine* 383, 27 (2020), 2603–2615.

[60] Nathaniel Rabb, Megan Swindal, David Glick, Jake Bowers, Anna Tomasulo, Zayid Oyelami, Kevin H. Wilson, and David Yokum. 2022. Evidence from a statewide vaccination RCT shows the limits of nudges. *Nature* 604 (2022), E1–E7.

[61] Filip Radlinski, Martin Szummer, and Nick Craswell. 2010. Inferring Query Intent from Reformulations and Clicks. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*.

[62] Lydia Saad. 2021. More in U.S. Vaccinated After Delta Surge, FDA Decision. *Gallup* (2021). https://news.gallup.com/poll/355073/vaccinated-delta-surge-fda-decision.aspx.

[63] Michael Siegel, Isabella Critchfield-Jain, Matthew Boykin, Alicia Owens, Rebeckah Muratore, Taiylor Nunn, and Joanne Oh. 2022. Racial/Ethnic Disparities in State-Level COVID-19 Vaccination Rates and Their Association with Structural Racism. *Journal of Racial and Ethnic Health Disparities* 9, 6 (2022), 2361–2374.

[64] Marianna Sotomayor, Jacqueline Alemany, and Mike DeBonis. 2021. Growing number of Republicans urge vaccinations amid delta surge. *The New York Times* (2021). https://www.washingtonpost.com/politics/growing-number-of-republicans-urge-vaccinations-amid-delta-surge/2021/07/20/52a06e9c-e999-11eb-8950-d73b3e93ff7f_story.html.

[65] StatCounter. 2023. Desktop Search Engine Market Share United States Of America, Jan - Dec 2021. https://gs.statcounter.com/search-engine-market-share/desktop/united-states-of-america/2021.

[66] Seth Stephens-Davidowitz. 2014. The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics* 118 (2014), 26–40.

[67] Jina Suh, Eric Horvitz, Ryen W. White, and Tim Althoff. 2021. Population-Scale Study of Human Needs During the COVID-19 Pandemic: Analysis and Implications. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM'21)*. 4–12.

[68] Jina Suh, Eric Horvitz, Ryen W. White, and Tim Althoff. 2022. Disparate impacts on online information access during the Covid-19 pandemic. *Nature Communications* 13, 7094 (2022).

[69] Tom Tapp. 2021. Los Angeles City Council Votes 13-0 To Create Vaccination Requirement For Indoor Public Spaces Such As Restaurants, Movie Theaters, Concert Venues. *Deadline* (2021). https://deadline.com/2021/08/los-angeles-city-requires-vaccination-vaccine-indoors-1234813086/.

[70] Jennifer Tolbert, Kendal Orgera, Rachel Garfield, Jennifer Kates, , and Samantha Artiga. 2021. Vaccination is Local: COVID-19 Vaccination Rates Vary by County and Key Characteristics. *Kaiser Family Foundation (KFF)* (2021). https://www.kff.org/coronavirus-covid-19/issue-brief/vaccination-is-local-covid-19-vaccination-rates-vary-by-county-and-key-characteristics/.

[71] Gianmarco Troiano and Alessandra Nardi. 2021. Vaccine hesitancy in the era of COVID-19. *Public Health* 194 (2021), 245–251.

[72] Raymond John D Vergara, Philip Joseph D Sarmiento, and James Darwin N Lagman. 2021. Building public trust: a response to COVID-19 vaccine hesitancy predicament. *Journal of Public Health* 43, 2 (2021), e291–e292.

[73] Noah Weiland. 2021. One and Done: Why People Are Eager for Johnson & Johnson's Vaccine. *The New York Times* (2021). https://www.nytimes.com/2021/03/04/health/covid-vaccine-johnson-and-johnson-rollout.html.

[74] Rebecca L. Weintraub, Kate Miller, Benjamin Rader, Julie Rosenberg, Shreyas Srinath, Samuel R. Woodbury, Marinanicole D. Schultheiss, Mansi Kansal, Swapnil Vispute, Stylianos Serghiou, et al. 2023. Identifying COVID-19 Vaccine Deserts and Ways to Reduce Them: A Digital Tool to Support Public Health Decision-Making. *American Journal of Public Health* 113, 4 (2023), 363–367.

[75] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Transactions on Information Systems* 27, 4 (2009).

[76] Farah Yasmin, Hala Najeeb, Abdul Moeed, Unaiza Naeem, Muhammad Sohaib Asghar, Najeeb Ullah Chughtai, Zohaib Yousaf, Binyam Tariku Seboka, Irfan Ullah, Chung-Ying Lin, and Amir H. Pakpour. 2021. COVID-19 Vaccine Hesitancy in the United States: A Systematic Review. *Frontiers in Public Health* 9 (2021).

APPENDIX

The Appendix provides additional results and experiments, including detailed descriptions of our ontology (Figure A1), results from developing our vaccine intent classifier (Section A1), our decomposition and evaluations of bias (Section A2), and additional analyses of vaccine intent trends and vaccine concerns (Section A3).

A1 VACCINE INTENT CLASSIFIER: ADDITIONAL RESULTS

Annotation results. As discussed in the main text, in the second step of our classification pipeline, we present URLs to annotators on Amazon Mechanical Turk. We find that a large majority (86%) of our URL candidates are labeled as vaccine intent, when we require at least three positive annotations to qualify a URL as vaccine intent. Furthermore, we observe a clear relationship between S-PPR rank and the percentage labeled as vaccine intent, whether we set the threshold at two or three annotations (Figure A2). For example, when we require three positive annotations, around 90% of URLs from ranks 0 to 20 qualify, around 81% of URLs from ranks 40-60 qualify, and around 71% of URLs from ranks 80 to 100 qualify. Thus, we find that S-PPR predicts vaccine intent remarkably well, with a high rate among its top URLs and agreement with a decreasing rate as the ranking drops.

Details from GNN experiments. In the final step of our classification pipeline, we train GNNs to learn vaccine intent labels and discover new URLs. Since there are not enough URL labels from AMT for smaller states, we experiment with *pre-training* the GNN on S-PPR rankings. In practice, before training the model on the URL labels from AMT, we train the model to predict the URLs' S-PPR rankings that we derived in the first step of our pipeline. Since S-PPR rankings become less meaningful in the long tail of URLs, we focus on predicting the top $K = \max(1000, q_{\max})$ S-PPR rankings, where $q_{\max}$ is the maximum rank (where lower rank corresponds to higher S-PPR score) of the last seed set query.

To test the effect of pre-training on S-PPR rankings, we select six representative states that vary in graph size and US region. We find that pre-training significantly improves performance for the smaller states. For example, the mean AUC for Wyoming increases from 0.74 to 0.95 (Table A1). Specifically, due to the low number of URL labels for smaller states, we observe great variance in the model's performance if we do not pre-train the model, leading to some trials that perform well and some that perform poorly (Figure 3a). Performance becomes far more stable for smaller states after we incorporate the pre-training objective. We find that pre-training seems unnecessary for the larger states, such as Connecticut and Tennesssee, where we are already achieving high AUCs above 0.98. So, we set a generous cutoff of 5,000,000 nodes (still larger than the graph size for Connecticut) and we pre-train all states with fewer than 5,000,000 nodes in our data, of which there are 26. After incorporating pre-training for these smaller states, we are able to achieve AUCs above 0.90 for all 50 states and above 0.95 for 45 states (Figure 3b).

As a supplementary analysis, we can also use AUC to evaluate the predictive performance of S-PPR alone and GNN-PPR, i.e., the GNN pre-trained on S-PPR rankings *before* it is also trained on AMT labels. Here, we evaluate on *all* AMT labels, since none of them were used in constructing S-PPR or GNN-PPR scores. In fact, evaluating on AMT labels is particularly challenging, since we chose to label only the top-ranked URLs according to S-PPR, so we are asking S-PPR to distinguish between URLs that it already considers similar. We conduct this experiment on the 26 smaller states for which we pre-trained our GNNs.

State	# nodes	AUC w/o pre-train	AUC w/ pre-train
WY	752865	0.741 (0.146)	0.951 (0.014)
AK	909357	0.796 (0.187)	0.921 (0.074)
DE	1269327	0.864 (0.134)	0.968 (0.007)
MT	1533071	0.857 (0.139)	0.978 (0.011)
CT	4407722	0.987 (0.005)	0.984 (0.008)
TN	7712443	0.991 (0.003)	0.990 (0.003)

Table A1: Effects of pre-training on S-PPR rankings for six selected states. We report the mean and standard deviation of AUC on the test set over 10 random trials.

First, we find across these states that S-PPR still performs better than random, with a mean AUC of 0.569, which complements our annotation results showing that even within its top-ranked URLs, S-PPR rankings still correlate with true rates of vaccine intent labels (Figure A2). Second, we find that GNN-PPR consistently *outperforms* S-PPR by 10-15 points, with a mean AUC of 0.675. This is somewhat surprising, since GNN-PPR was only trained to predict S-PPR rankings, without any additional labels. We hypothesize that GNN-PPR outperforms S-PPR because, unlike S-PPR, the GNN can incorporate textual information from URLs and queries, in addition to graph structure. So, while S-PPR incorrectly upweights high-traffic URLs such as facebook.com that are often reached on random walks starting from the vaccine intent queries, GNN-PPR recognizes that these URLs do not look like the rest of high-ranking URLs and correctly excludes them. However, in order to achieve this difference between S-PPR and GNN-PPR, it is important not to overfit on S-PPR. So, we employ early stopping during pre-training; that is, we train the GNN on S-PPR rankings until they achieve a correlation of 0.8 and then we stop pre-training.

Our evaluation results demonstrate that our GNNs are able to accurately predict vaccine intent labels in all 50 states, which is essential as we use our GNNs to discover new vaccine intent URLs. In Table A2, we provide a uniform random sample of the URLs that our GNNs discovered. The majority of them seem to express vaccine intent, with several news stories about new vaccine clinics and information about vaccine appointments. Furthermore, the supplemental analysis of S-PPR and GNN-PPR shows that due to the expressive power of the GNN (with character-level CNN) and the predictive power of S-PPR from a well-designed seed set, we can achieve decent performance without *any* labels at all. These methods, which should be explored more deeply in future work, may be useful in a zero-shot context, allowing lightweight, effective prediction before acquiring any labels.

Top category	Subcategory	Description
Safety	Normal side effects	Expected side effects: sore arm, shoulder, fever, etc
	Severe side effects	Rare but plausible side effects, severe, potentially long-term: blood clots, myocarditis, etc
	Reproductive health	Concerns about fertility, breast feeding, menstruation
	Vaccine-caused deaths	Fear of deaths *caused* by COVID vaccine
	Eerie fears	Eerie and debunked fears: shedding, magnets, microchips, etc
	Vaccine development	History of vaccine development, fear of mRNA technology, ingredients in COVID vaccine
	FDA approval	FDA approval of COVID vaccines
Effectiveness	Efficacy from studies	How effective the vaccine is, how long immunity lasts, how long for vaccine to take effect
	Efficacy against variants	How well does vaccine work against variants (mostly Delta)
	Breakthrough cases	Breakthrough COVID cases, symptoms when vaccinated
	Natural immunity	Is natural immunity better than vaccine, do I still need vaccine
Community	Vaccine rates	Vaccine trackers, rates of vaccination over time: by state, by country, etc
	News on hesitancy	Reporting on vaccine hesitancy and anti-vaxxers, how to talk to vaccine hesitant
	Expert anti-vax	Anti-vaccine messages from scientists and doctors
	High-profile anti-vax	Anti-vaccine messages from high-profile figures: politicians, celebrities, etc
	Religious concerns	Religious concerns about the vaccine, seeking advice from religious leaders
Information	Decision-making	Pros and cons of COVID vaccine, should I get the vaccine?
	Comparison	Comparing Moderna vs Pfizer vs J&J, side effects, efficacy
	Moderna	General news on Moderna vaccine, rollout, side effects, efficacy
	Pfizer	General news on Pfizer vaccine, rollout, side effects, efficacy
	Johnson & Johnson	General news on J&J vaccine, emphasis on blood clots and efficacy
	Special populations	COVID-19 vaccine for special populations: autoimmune disease, rheumatoid arthritis, etc
	Post-vax guidelines	Guidelines after vaccination: masking, testing, quarantine
Requirements	Travel	Vaccine requirements to travel: for cruises, other countries, etc
	Employment	Employer vaccine mandates: healthcare, government, educators, etc
	Vaccine proof	Required proof of vaccination to enter places: restaurants, gyms, concert venues, etc
	Exemption	Seeking exemption on vaccine requirements, religious or medical
	Fake vaccine proof	Seeking fake proof of vaccination
	Anti-mandate	States banning mandates, lawsuits against employer mandates
Incentives	Vaccine incentives	Vaccine incentives: lotteries, gift cards, free groceries, giveaways, etc
Availability	Locations	Where to get COVID vaccine (some missed vaccine intent URLs): CVS, Walgreens, etc
	Children	Are COVID vaccines for children available / recommended
	Boosters	Are boosters available / recommended
Other	New / non-US vaccines	Other COVID vaccines: Novavax, Astrazeneca, Sinovax
	Non-COVID vaccines	Non-COVID vaccines: flu, MMR, varicella, meningitis, etc
	Pet vaccines	Vaccines for pets, mostly dogs and cats

Figure A1: Our ontology of vaccine concerns consists of 8 top categories and 36 subcategories.

URL	t_{med}	t_{prec}
https://www.chesco.org/4836/61876/COVID-Authorized-Vax	7	10
https://patch.com/new-jersey/princeton/all-information-princeton-area-covid-vaccine-sites	9	10
https://dph.georgia.gov/locations/spalding-county-health-department-covid-vaccine	9	10
https://www.abc12.com/2021/04/22/whitmer-says-covid-19-vaccine-clinics-like-flint-church-are-key-to-meeting-goals/	7	10
https://www.delta.edu/coronavirus/covid-vaccine.html	10	10
https://www.lewistownsentinel.com/news/local-news/2021/01/scheduling-a-virus-vaccine-appointment/	9	10
https://www.laconiadailysun.com/news/local/covid-vaccine-clinics-at-lrgh-franklin-now-open-to-public/article_aa4b67e0-601a-11eb-a889-1bd4e6c83de1.html	6	10
https://www.insidenova.com/headlines/inside-woodbridges-new-mass-covid-19-vaccination-site-the-lines-keep-moving/article_eca45b88-8db0-11eb-a649-4bbeccd82cc3.html	9	10
https://www.keloland.com/news/healthbeat/coronavirus/avera-opens-covid-19-vaccine-clinic/	10	9
https://bangordailynews.com/2021/04/06/news/maine-to-kick-off-statewide-mobile-covid-19-vaccine-clinics-in-oxford-next-week-sk6sr8zcdk/	8	9
https://morgancounty.in.gov/covid-19-vaccinations/	9	10
https://www.firsthealth.org/specialties/more-services/covid-19-vaccine	10	10
https://healthonecares.com/covid-19/physician-practices/covid-19-vaccine-information.dot	9	10
https://patch.com/florida/stpete/drive-thru-covid-19-vaccine-sites-open-florida	9	10
https://vaccinate.iowa.gov/eligibility/	7	10
https://www.baynews9.com/fl/tampa/news/2021/03/17/new-walk-in-vaccine-site-at-tpepin-hospitality-centre-opens-today	10	10
https://www.doh.wa.gov/Emergencies/COVID19/VaccineInformation/FrequentlyAskedQuestions	10	10
https://www.emissourian.com/covid19/vaccine-registration-open-for-franklin-county/article_3638f7a0-5769-11eb-9bba-3f2611173784.html	10	10
https://www.fema.gov/press-release/20210223/maryland-open-covid-19-vaccination-center-waldorf-fema-support	10	10
https://kingcounty.gov/depts/health/covid-19/vaccine/forms.aspx	10	10

Table A2: A random sample (`random_state=0`) of 20 URLs from GNN. t_{med} and t_{prec} indicate how often the URL passed the median cutoff and precision cutoff, respectively, out of the 10 trials.

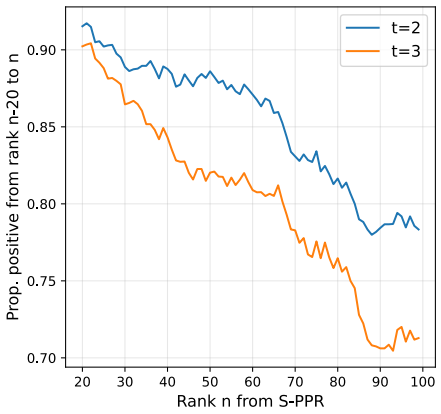


Figure A2: Comparison of S-PPR rank vs. proportion of URLs around that rank that are labeled as vaccine intent. $t = 3$ and $t = 2$ indicate how many positive annotations were required to qualify for vaccine intent.

A2 BIAS DECOMPOSITION AND EVALUATIONS

A2.1 Decomposition of bias

For a given individual, let $v \in \{0, 1\}$ indicate whether they actually had vaccine intent (up to a certain time) and $\hat{v} \in \{0, 1\}$ indicate whether our classifier labels them as having vaccine intent. Furthermore, let r represent the individual's home region, such as their state or county. We would like to estimate the regional vaccine intent rate, $\Pr(v|r)$, but we do not have access to v, only to $\hat{v}$. To understand how using $\hat{v}$ in place of v may bias our estimates, let us relate $\Pr(\hat{v}|r)$ to $\Pr(v|r)$. First, we introduce another variable b, which represents whether the individual is a Bing user. Note that $\hat{v} = 1$ implies that $b = 1$, since our classifier can only identify vaccine intent from users who appear in Bing search logs.

With these variables, we have

$$\Pr(\hat{v}=1|r) = \underbrace{\Pr(b=1|r)}_{\text{Bing coverage of } r} \Big[\Pr(v=1|r) \underbrace{\Pr(\hat{v}=1|b=1,v=1,r)}_{\text{Classifier TPR for } r} + \Pr(v=0|r) \underbrace{\Pr(\hat{v}=1|b=1,v=0,r)}_{\text{Classifier FPR for } r} \Big]. \tag{1}$$

$\Pr(b = 1|r)$ represents the probability that an individual from region r is a Bing user, i.e., the Bing coverage of r. Incorporating b, v, and r into $\Pr(\hat{v}|b, v, r)$ reflects all of the factors that affect whether the classifier predicts vaccine intent. As discussed, if the user is not a Bing user ($b = 0$), then the probability is 0, so we only consider the $b = 1$ case. If $v = 1$, predicting $\hat{v} = 1$ would be a true positive; if $v = 0$, it would be a false positive. Conditioning $\hat{v}$ on region r reflects the possibility that individuals from different regions may express vaccine intent differently and the classifier may be more prone to true or false positives for different regions. Finally, we make the assumption here that $b \perp v|r$; that is, conditioned on the individual's region, being a Bing user and having vaccine intent are independent. This misses potential within-region heterogeneity,

but to mitigate this in practice, we use ZCTAs as our regions, which are relatively fine-grained.

Based on this decomposition, we can see that if Bing coverage, TPR, and FPR are uniform across regions, then $\Pr(\hat{v}|r)$ will simply be a linear function of $\Pr(v|r)$. Unfortunately, we know that Bing coverage is not uniform. However, we observe $b = 1$ and can assign users to regions, so we can estimate Bing coverage per region and correct by inverse coverage. Thus, our estimate corresponds to a coverage-corrected predicted vaccine intent rate, $\tilde{p}(v, r) = \frac{\Pr(\hat{v}=1|r)}{\Pr(b=1|r)}$. If we refer to the true vaccine intent rate as $p(v, r)$, then we can see that $\tilde{p}(v, r)$ is a linear function of $p(v, r)$ when TPR and FPR are uniform:

$$\frac{\Pr(\hat{v} = 1|r)}{\Pr(b = 1|r)} = \Pr(v = 1|r)\text{TPR} + (1 - \Pr(v = 1|r))\text{FPR} \quad (2)$$

$$\tilde{p}(v, r) = \text{FPR} + (\text{TPR} - \text{FPR})p(v, r).$$

Furthermore, if FPR is low, then $\tilde{p}(v, r)$ is approximately proportional to $p(v, r)$. Thus, our first two strategies for addressing bias in our estimates are:

(1) Estimate Bing coverage per region and weight by inverse coverage, which we discussed in Section 4,
(2) Evaluate whether our classifier has similar TPRs and FPRs across regions and whether FPRs are close to 0, which we discuss below.

These efforts are our first two lines of defense against bias. After this, we furthermore compare our results to established data sources, such as the CDC's reported vaccination rates and Google search trends, where we find strong correlations for both.

A2.2 Evaluating bias in vaccine intent classifier

Our primary source of bias is uneven Bing coverage, which we found can vary by more than 2x across ZCTAs. However, after correcting for Bing coverage, we also want to know that our classifier does not significantly contribute to additional bias. To do this, we must establish that our classifier's TPRs and FPRs do not vary significantly or systematically across regions. The challenge is that we cannot perfectly evaluate these rates, because we do not know all true positives or true negatives. However, we can approximate these metrics based on the labeled URLs that we do have and furthermore make methodological decisions that encourage similar performance across groups.

Evaluating bias in generating URL candidates. Recall that in the first step of our pipeline, we generate URL candidates for annotation by propagating labels from vaccine intent queries to unlabeled URLs via personalized PageRank on query-click graphs. Since all URL candidates then go through manual inspection in the second step, we do not have to worry about the false positive rate at this stage. However, we do need to worry about the true positive rate (i.e., recall). For example, if we only kept COVID-19 vaccine registration pages for pharmacies that are predominantly in certain regions, then we could be significantly likelier to detect true vaccine intent for certain states over others. So, through the design and evaluation of our label propagation techniques, we aim to ensure representativeness in vaccine intent across the US.

The most important design decision is that we construct query-click graphs *per state*, then we run S-PPR per graph and take the union over states of top URLs as our set of URL candidates. Running this process separately for each state allows us to capture how vaccine intent varies regionally, with state-specific programs and websites for scheduling the vaccine (Table 1). To demonstrate the risks of not using a state-specific approach, we try an alternative approach where we construct a joint graph that combines the queries and clicks for 6 states (the same 6 states as those used in the pre-training experiments of Table A1).

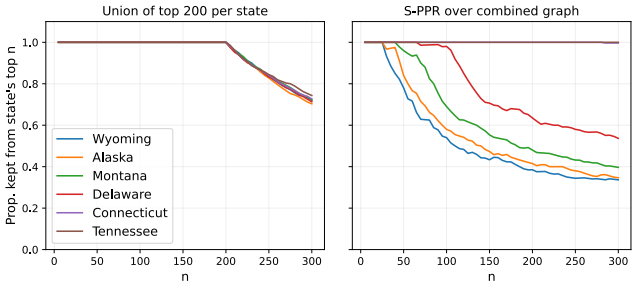


Figure A3: Comparing our union-over-states (left) to a combined graph approach (right) for generating URL candidates.

To represent our union approach, we take the union over these 6 states of the top 200 URLs per state, which results in 935 URLs. We compare this to a joint approach, where we take the top 935 URLs from running S-PPR on the joint graph. To evaluate each approach, we compute the proportion of each state's top N URLs that are kept across different values of N. While we cannot be sure that every URL in the state's top N is truly vaccine intent, from our annotation results, we saw high positive rates for top-ranking URLs (Figure A2), so we would like to see similar recall at these ranks.

By design, our union-over-states approach ensures equivalent, 100% recall up to $N = 200$ for all states (Figure A3, left). In comparison, we find that the joint approach yields different recalls as early as $N = 30$, with much higher recall for large states than small states (Figure A3, right). For example, it keeps less than 80% of Wyoming's URLs around rank 50 and less than 60% around rank 100, while keeping 100% of Tennessee's throughout. Furthermore, even past $N = 200$, where our union-over-states approach no longer has guarantees, we find that it still achieves far more similar recalls between states than the joint approach. Thus, our design decisions enable similar recalls between states, which helps to reduce downstream model bias. We also cast a wide net when constructing query-click graphs (taking all queries and clicks that co-occur in a session with any query that includes a COVID-19 or vaccine-related word), which may also improve recall and reduce bias, in case our choice of initial keywords was not representative of all vaccine intent searches across the US.

Evaluating bias in URL expansion from GNN.. In the third step of our pipeline, we use GNNs to expand our set of vaccine intent URLs beyond the manually labeled ones. We would like to see that the performance of GNNs is similarly strong across states, to ensure that the GNN is not creating additional bias when expanding the URL set. We discussed in Section A1 that, after incorporating pre-training on S-PPR rankings for smaller states, GNNs could achieve AUCs above 0.90 for all 50 states. The main metrics of interest

when considering bias, however, are TPRs and FPRs. Unlike AUC, which is evaluated across decision thresholds, TPR and FPR depend on the chosen threshold t above which data points are predicted to be positive. In our setting, we set $t = \max(t_{\text{med}}, t_{\text{prec}})$, since we required new vaccine intent URLs to score above these two thresholds (in at least 6 out of 10 trials): (1) t_{med}, the median score of positive URLs in the test set and (2) t_{prec}, the minimum threshold required to achieve precision of at least 0.9 on the test set. Then, we estimate TPR as the proportion of positive URLs in the test set that score above t and FPR as the proportion of negative URLs in the test set that score above t.

We find that TPR is highly similar across states and hovers around 0.5 for all states (Figure 3b, middle). This is because in almost all cases, t_{med} is the higher of the two thresholds and thus the value of t, so the true positive rate lands around 0.5 since t_{med} is the median score of the true positives. FPR is also highly similar across states and very low (around 0.01; Figure 3b, right), which suggests that the quantity we estimate, $\tilde{p}(v, r)$, is not only a linear function of the true vaccine intent rate, $p(v, r)$, but also approximately proportional to it (Eq. 2). The low FPR is encouraged but not guaranteed by our second threshold, t_{prec}. This threshold ensures that precision is over 0.9, which is equivalent to the false positive rate *among the predicted positives* being below 0.1, which typically corresponds to low false positive rates over all true negatives (which is what FPR measures). The GNN's similar AUCs, TPRs, and FPRs across states, as well as the equivalent recalls in our label propagation stage, increase confidence that our classifier is not adding significant bias to our estimates.

A2.3 Comparison to Google search trends

Following prior work using Bing data [68], we compare Bing and Google queries to evaluate the representativeness of Bing data.

Search trends over time. First, we compare daily search interest in the US over our studied time period from February 1 to August 31, 2021. Google Trends provides normalized search interest over time on Google, such that 100 represents the peak popularity for that time period, 50 means the term is half as popular, and 0 means "there was not enough data for this term." To match this, for a given query, we compute the total number of times it was searched on Bing in the US per day, then we divide by the maximum number and multiply by 100. Again, we apply 1-week smoothing to both the Bing and Google time series. We do not correct the Bing time series with Bing coverage here, since we cannot correct the Google time series with Google coverage, and we want the time series to be constructed as similarly as possible.

We evaluate 30 of the most common vaccine intent queries, including [cvs covid vaccine] and [covid vaccine finder].[4] We observe strong Pearson correlations, with a median correlation of $r = 0.95$ (90% CI, 0.88-0.99) (Figure A4a). These correlations are similar to those reported by Suh et al. [68], who conduct an analogous longitudinal analysis comparing Bing and Google search trends on COVID-related queries and report correlations from $r = 0.86$ to 0.98. Remaining discrepancies between Bing and Google are likely due to differences in the populations using these search engines, as well as potential unreported details on how Google normalizes their search interest trends (e.g., Google may be normalizing differently for [covid vaccine near me], which shows unusual peaks in Google trends and is the the only query for which we do not observe a strong correlation).

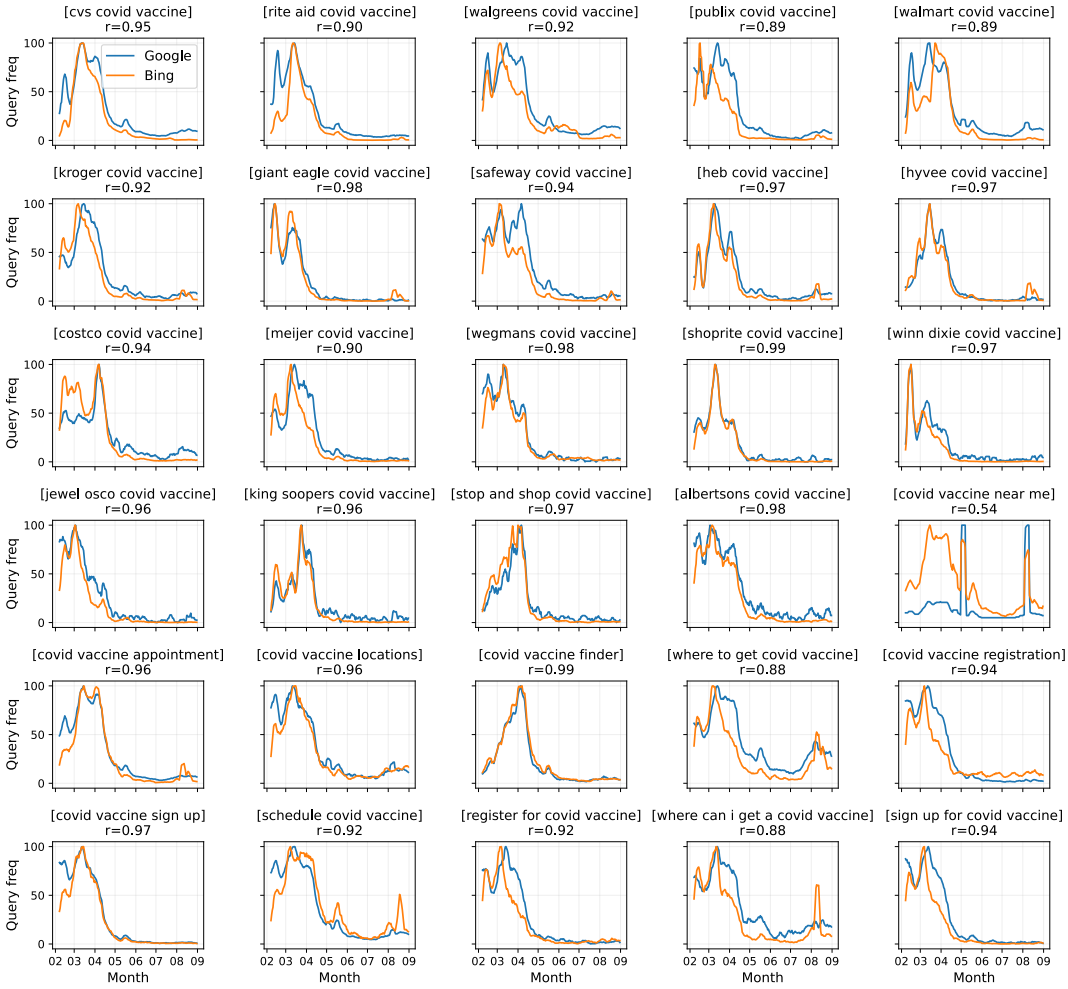
Search trends across states. Google also provides normalized search interest across US states, where search interest is defined as the fraction of searches from that state that match the query and search interest is normalized across regions such that 100 represents maximum popularity. To imitate this process, we first assign each vaccine intent query to a state based on where the query originated. Then, we approximate the total number of queries (all queries, not just vaccine intent) from each state by summing over the query counts of the active users assigned to each state. We compute the fraction of queries from each state that match the query, then we divide by the maximum fraction and multiply by 100 to normalize across states.

We observe strong Pearson correlations in this analysis too, with a median correlation of $r = 0.95$ (90% CI, 0.57-0.99) across the same 30 vaccine intent queries (Figure A4b). The correlations tend to be stronger on the pharmacy-specific queries, where certain regions dominate, compared to general location-seeking queries such as [covid vaccine near me], which are trickier since they follow less obvious geographical patterns. For the pharmacy-specific queries, we also observe substantial heterogeneity in terms of which region dominates. For example, [publix covid vaccine] is more popular in southern states, with Florida exhibiting the maximum normalized search interest on Google (100), followed by Georgia (26) and South Carolina (20). Meanwhile, [cvs covid vaccine] is more popular in the Northeast, with the top states being Massachusetts (100), New Jersey (96), Rhode Island (90), and Connecticut (65). These differences, reflected in the Bing search trends too, once again highlight the need for regional awareness and representativeness when developing our vaccine intent classifier.

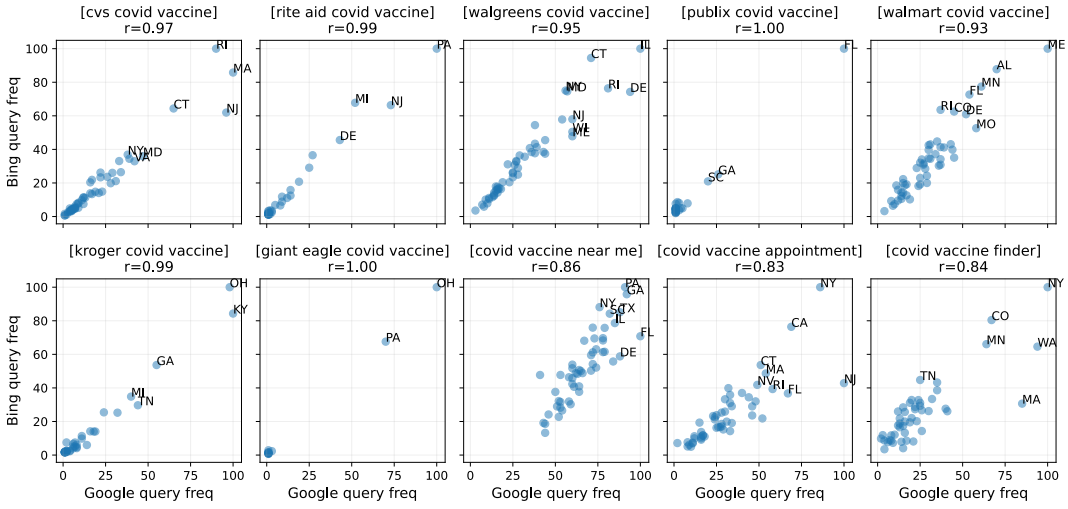
A3 ADDITIONAL ANALYSES

State-level demographic trends in vaccine intent. To investigate more granular demographic trends, we measure correlations per state (only including the ZCTAs in the state) for the 10 largest states in the US. For this finer-grained analysis, we drop percent Republican, since we only have vote share at the county-level, but we keep all other demographic variables, which we have per ZCTA. We find that correlations are mostly consistent in sign across states, but the magnitude differs significantly (Figure A5). For example, the positive correlation with percent 65 and over is around 2x as high in Florida as it is in the second highest states, reflecting the large senior population in Florida and the push for seniors to get vaccinated. In most states, we also see positive correlations for percent Asian and percent White, and negative correlations for percent Black and percent Hispanic, aligning with prior research on racial and ethnic disparities in COVID-19 vaccination rates [51, 63]. Positive and negative correlations for race are particularly strong in certain states, including New York and Florida for percent White/Black, and California and New York for percent Hispanic.

[4]We identify 30 representative vaccine intent queries from the top 100 vaccine intent queries, where we choose one standard query for each pharmacy that appears (e.g., [cvs covid vaccine]) and one for each location-seeking query (e.g., [covid vaccine near me]), and drop variants such as [cvs covid vaccines] and [covid 19 vaccine near me].



(a) US search trends over time.



(b) Search trends across states.

Figure A4: Comparing search trends on Google vs. Bing for 30 of the most common vaccine intent queries.

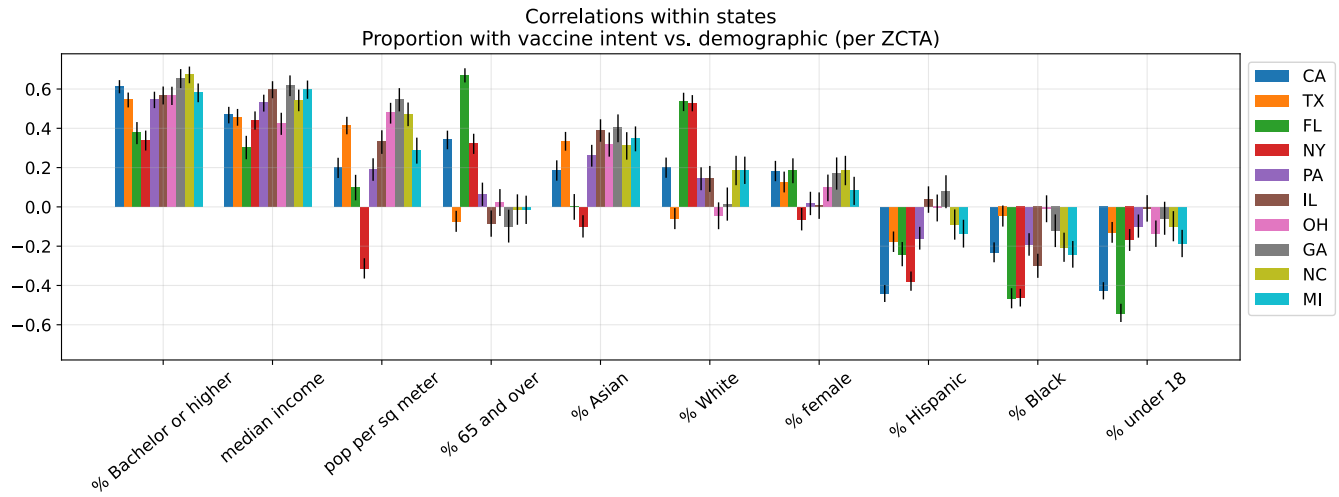
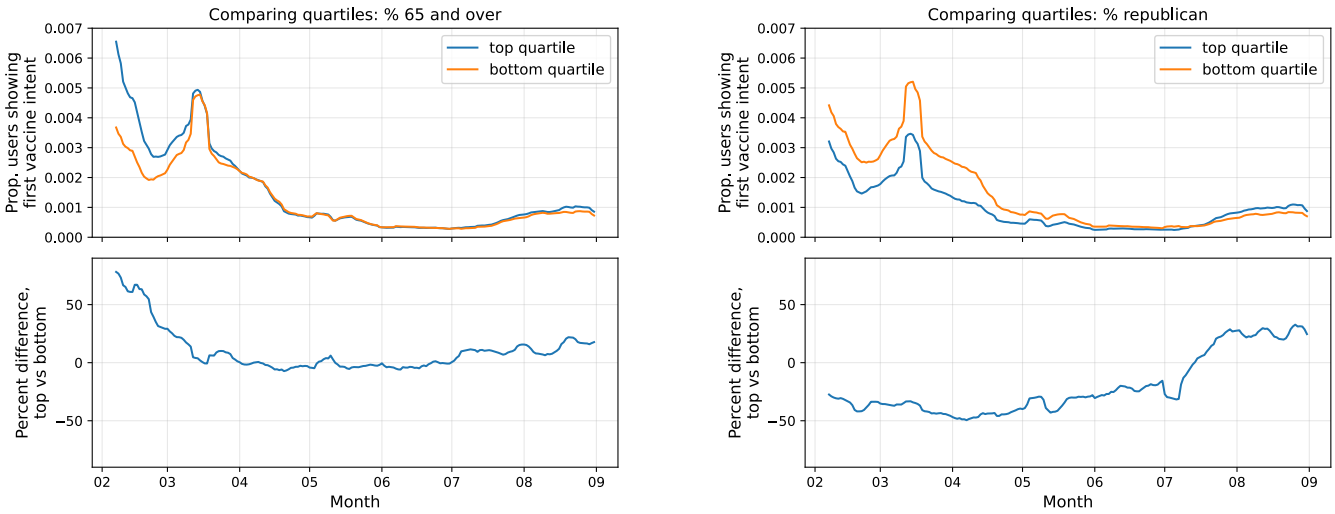


Figure A5: Correlations between ZCTA vaccine intent rate and demographic variables, for the 10 largest US states. Error bars indicate 95% CIs.

Changes in demographic trends over time. To evaluate changes in demographic trends over time, we separate ZCTAs into top and bottom quartiles, e.g., based on ZCTA median income, and compute each quartile's daily proportion of users showing their *first* vaccine intent. Then, computing the ratio of the top quartile's over bottom quartile's time series reveals changes in demographic trends over time. For example, we estimate that older ZCTAs were much likelier to seek the vaccine early in 2021 but this trend fell over time (Figure A6a), reflecting how the US vaccine rollout first prioritized seniors then expanded to general eligibility [4, 38]. We also see an increase in vaccine intent from more Republican ZCTAs in summer 2021 (Figure A6b), reflecting new calls from Republican leaders to get vaccinated [64] and a self-reported uptick in vaccinations among Republicans [62].

Examples of URL clusters. To construct our ontology of vaccine concerns, we begin by automatically partitioning URLs into clusters, using the Louvain community detection algorithm [12] on the collapsed URL-URL graph. We find that our automatic approach produces remarkably coherent clusters, with each cluster covering a distinct topic. The cluster annotations are provided in the ontology that we release, with URLs mapped to 156 unique clusters. We provide a sample of the clusters in Table A3, listing each cluster's most frequently clicked URLs and top query, which we obtain by summing over all queries that led to clicks on URLs in the cluster. From the top query and URLs, we observe distinct topics covered in each cluster: one on CDC masking guidelines after vaccination, one on the Vaccine Adverse Event Reporting System (VAERS), one about religious exemptions for COVID-19 vaccine requirements, and one about side effects of the Johnson & Johnson vaccine.

Holdout concerns across demographic groups. We conduct an additional analysis to analyze variation in holdout concerns across demographic groups. For a given demographic variable, we compute its median value across all ZCTAs, split holdouts into those from ZCTAs above the median versus those from ZCTAs below the median, then compare the vaccine concerns of those two groups of holdouts (by measuring their click ratios). We find significant variability across demographic groups in terms of holdout concerns (Figure A7). Compared to holdouts from more Republican-leaning ZCTAs, holdouts from more Democrat-leaning ZCTAs were far more interested in requirements around employee mandates and vaccine proof, which may be because jurisdictions run by Democrats were likelier to have vaccine requirements [9, 69] while several Republican governors in fact banned such requirements. Meanwhile, holdouts from more Republican-leaning ZCTAs were more interested in eerie vaccine fears, fears of vaccine-caused deaths, and vaccine incentives. We also find that, compared to holdouts from lower-income ZCTAs, holdouts from higher-income ZCTAs were significantly more interested in vaccine requirements, vaccine rates, and anti-vaccine messages from experts and high-profile figures, while holdouts from lower-income ZCTAs were more interested in vaccine incentives and religious concerns about the vaccine.



(a) Top and bottom quartiles for percent 65 and over.

(b) Top and bottom quartiles of percent Republican.

Figure A6: We quantify changes over time in demographic trends by estimating average vaccine intent rates per quartile over time (top) and computing their percent difference (bottom).

# URLs	Top query	Top URLs	% Clicks
206	[cdc mask guidelines]	https://www.cbsnews.com/news/cdc-mask-guidelines-covid-vaccine	8.0
		https://www.cdc.gov/media/releases/2021/p0308-vaccinated-guidelines.html	6.9
		https://www.usatoday.com/story/news/health/2021/05/13/covid-vaccine-cdc-variant-fda-clots-world-health-organization/5066504001	4.5
		https://www.nytimes.com/2021/05/13/us/cdc-mask-guidelines-vaccinated.html	4.4
139	[vaers database covid-19]	https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vaers/index.html	17.0
		https://rightsfreedoms.wordpress.com/2021/07/22/vaers-whistleblower-45000-dead-from-covid-19-vaccines-within-3-days-of-vaccination-sparks-lawsuit-against-federal-government	6.8
		https://www.theburningplatform.com/2021/07/03/latest-cdc-vaers-data-show-reported-injuries-surpass-400000-following-covid-vaccines	5.7
		https://vaersanalysis.info/2021/08/20/vaers-summary-for-covid-19-vaccines-through-8-13-2021	4.9
137	[religious exemption for covid-19 vaccination]	https://www.verywellfamily.com/religious-exemptions-to-vaccines-2633702	16.5
		https://www.fisherphillips.com/news-insights/religious-objections-to-mandated-covid-19-vaccines-considerations-for-employers.html	5.1
		https://www.law360.com/articles/1312230/employers-should-plan-for-vaccine-religious-exemptions	3.9
		https://www.kxly.com/who-qualifies-for-a-religious-exemption-from-the-covid-19-vaccine	3.3
113	[johnson and johnson side effects]	https://www.openaccessgovernment.org/side-effects-johnson-johnson-vaccine/109505	20.3
		https://www.healthline.com/health/vaccinations/immunization-complications	8.1
		https://www.msn.com/en-us/health/medical/these-are-the-side-effects-from-the-johnson-and-johnson-covid-19-vaccine/ar-bb1f03fq	4.3
		https://www.healthline.com/health-news/mild-vs-severe-side-effects-from-the-johnson-and-johnson-covid-19-vaccine-what-to-know	4.3

Table A3: The 4 highest-modularity clusters with at least 100 URLs. For each cluster, we provide its number of URLs, its most frequent query, its top 4 URLs (by click frequency), and percentage of clicks over all clicks on URLs in the cluster that the URL accounts for.

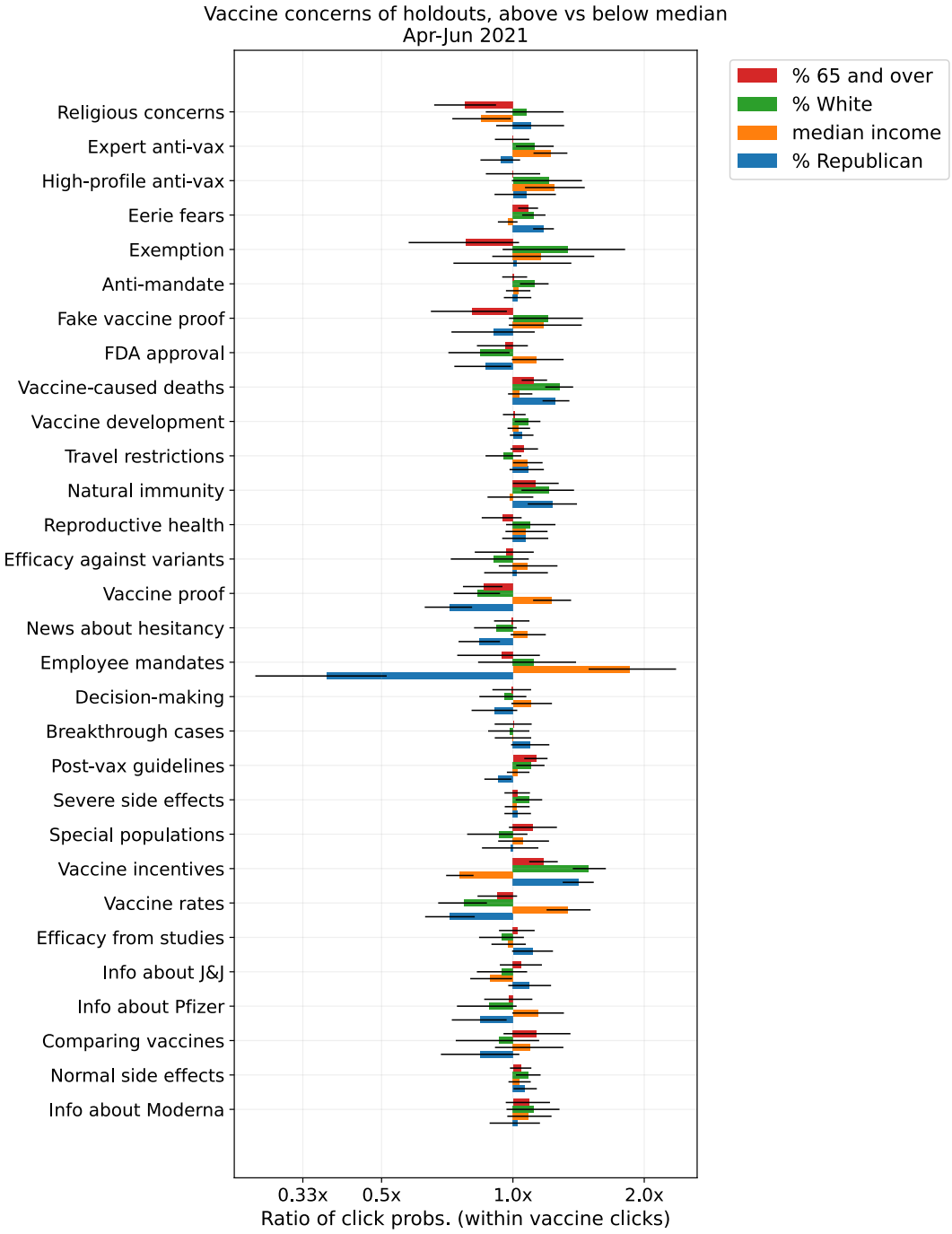


Figure A7: Variability in holdout concerns across demographic groups. For each demographic variable (e.g., percent Republican), we compare the concerns of holdouts from ZCTAs above the variable's median versus holdouts from ZCTAs below the median. Subcategories are ordered from most holdout-leaning to most early adopter-leaning, following Figure 7c. Error bars indicate bootstrapped 95% CIs.