Wearable Health Condition Detection with Vector Database

Malinda Huang, Yucan Wu

COS597A

Why **Wearable** Health Condition Detection?

Why **Wearable** Health Condition Detection?

- **Limitations** of traditional health monitoring and disease detection:
  - Diagnosis is often delayed until symptoms become apparent
  - Subtle variations and early signs of conditions often go unnoticed
  - Require frequent visits to healthcare facilities
  - Limited personalization in one-size-fits-all approaches

What about Neural Network? 

- **CovidDeep** framework [1] utilized deep neural networks with Wearable Medical Sensors and achieved great accuracy

Limitations

- **Model Adaptability** 
- **Resource Intensity** 

[1] S. Hassantabar, N. Stefano, V. Ghanakota, A. Ferrari, G. N. Nicola, R. Bruno, I. R. Marino, K. Hamidouche, and N. K. Jha, "CovidDeep: SARS-CoV-2/COVID-19 Test Based on Wearable Medical Sensors and Efficient Neural Networks," arXiv preprint arXiv:2007.10497, 2020.

Introducing Our Solution:
Vector-based Wearable Health Condition Detection

- A medical monitoring system that
  - **collects** physiological data from wearable devices
  - **monitors** health condition real-time through vector search using *Pinecone*
- Vector database advantages:
  - ○ **Dynamic Learning:** New symptom patterns *without exhaustive retraining*
  - ○ **Personalized Care:** Utilize user's history data
  - ○ **Proactive Detection:** Early detection and continuous monitoring
  - ○ **Edge Optimization:** Low cost and fast
  - ○ **Explainable Results**
- **App Demo** at the end

Methodology

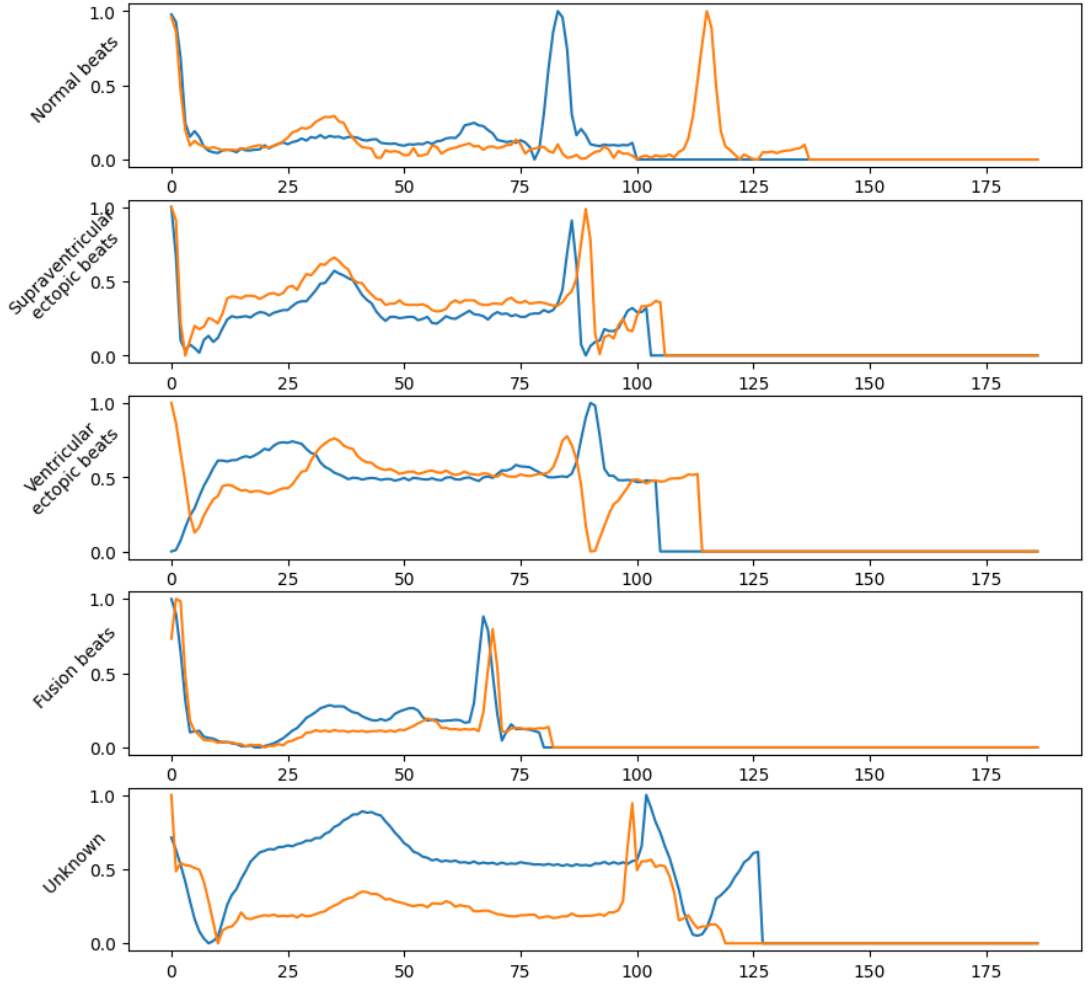
1. **Data**
   - Publicly available electrocardiogram (ECG) datasets
   - Pre-processing
2. **Vector Database** construction using training sets
3. **Evaluation** using testing sets
   - Prediction accuracy: Top k similarity search majority vote (k = 1, 5, 10)
4. **App**
   - Choose database(s) with satisfactory accuracy
   - Build an app that takes in ECG data collected from Apple Watch for health condition detection

Experimented with three ECG datasets

- MIT-BIH Arrhythmia Dataset [1]
- PTB Diagnostic Database [2]
- PhysioNet/CinC Challenge 2017 [3]

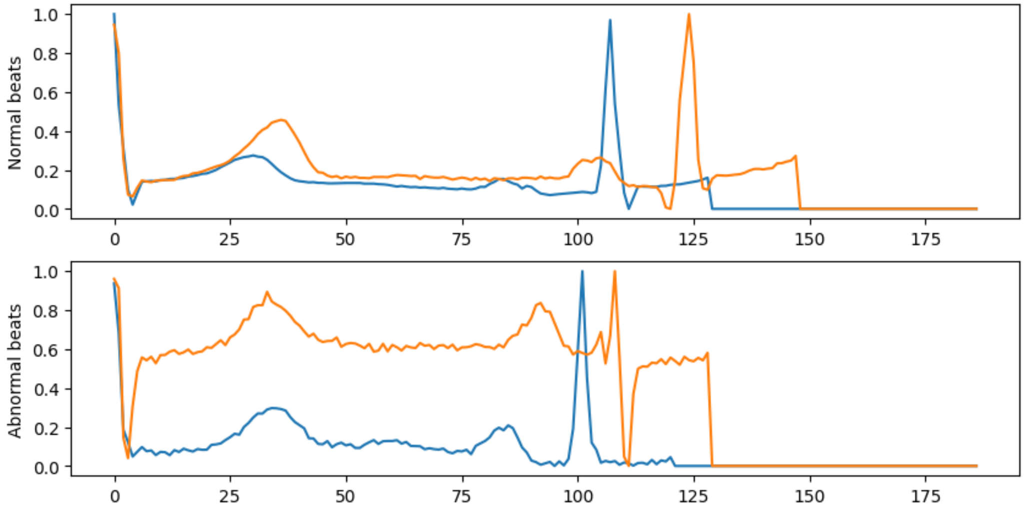
[1] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," in IEEE Engineering in Medicine and Biology Magazine, vol. 20, no. 3, pp. 45-50, May-June 2001, doi: 10.1109/51.932724.
[2] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," Biomedical Engineering / Biomedizinische Technik, vol. 40, no. s1, pp. 317–318, 1995, doi:10.1515/bmte.1995.40.s1.317.
[3] G. Clifford et al., "AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017," 2017 Computing in Cardiology Conference (CinC), Sep. 2017, doi: https://doi.org/10.22489/cinc.2017.065-469.

MIT-BIH & PTB Datasets: Beat type classification



MIT-BIH

- ECGs are **cropped** into single beats, **padded** to the same length, and **normalized**



PTB

MIT-BIH Dataset: Results

- Test accuracy:

Distance Metric	Top 1	Top 10
Cosine	**97.69%**	97.30%
Dot product	71.66%	77.85%
Euclidean	97.62%	97.17%

- Compare with EfficientNet accuracy: 96.62%

PTB Dataset: Results

- Test accuracy:

Distance Metric	Top 1	Top 10
Cosine	95.81%	93.34%
Dot product	71.56%	71.56%
Euclidean	**99.21%**	95.26%

- Compare with CNN Autoencoder accuracy: 99.9%

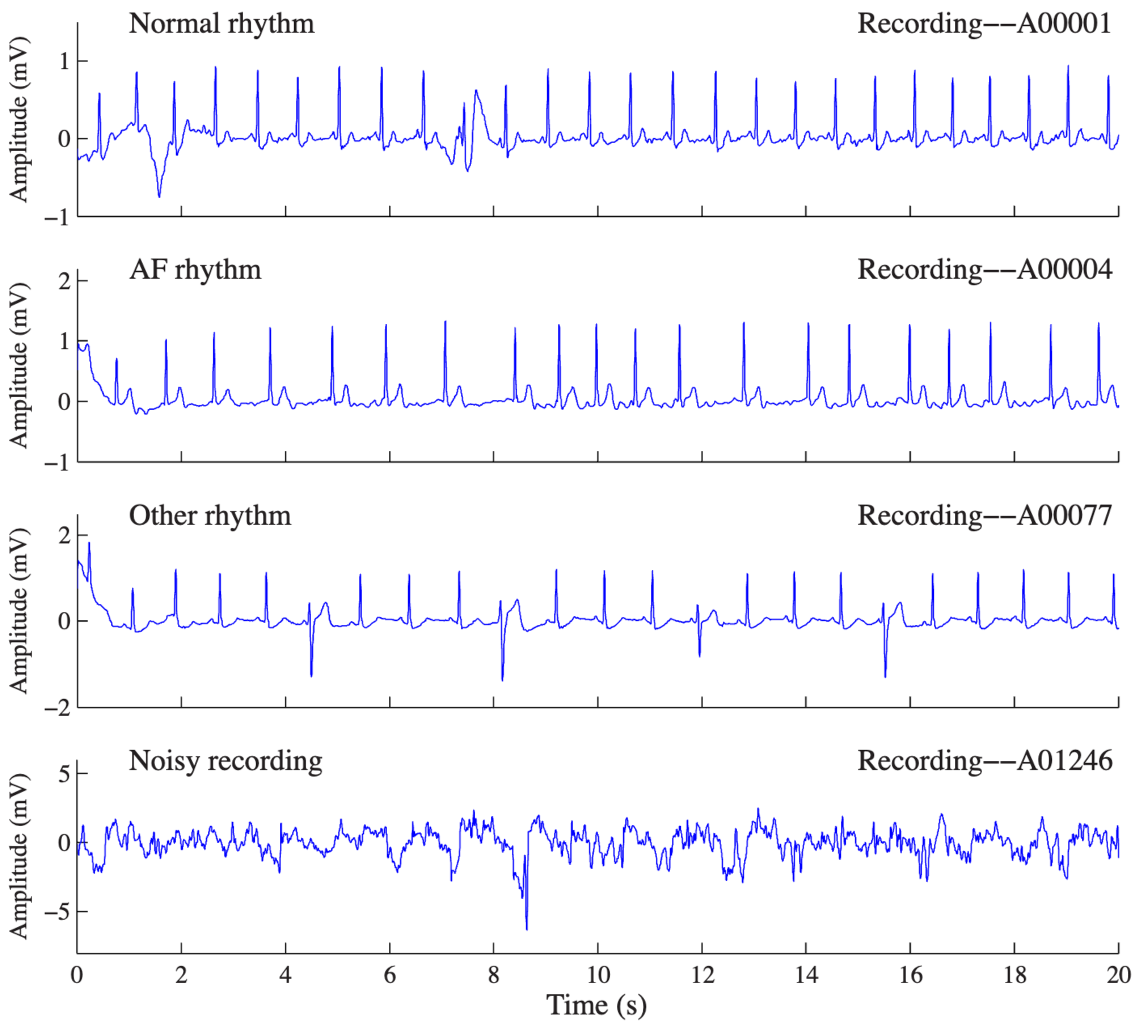
Insights from MIT-BIH & PTB datasets

- Vector search methods can achieve **comparable or higher accuracy** than neural network models without the need of training

- Different datasets have different optimal **distance metrics**

Atrial Fibrillation Detection

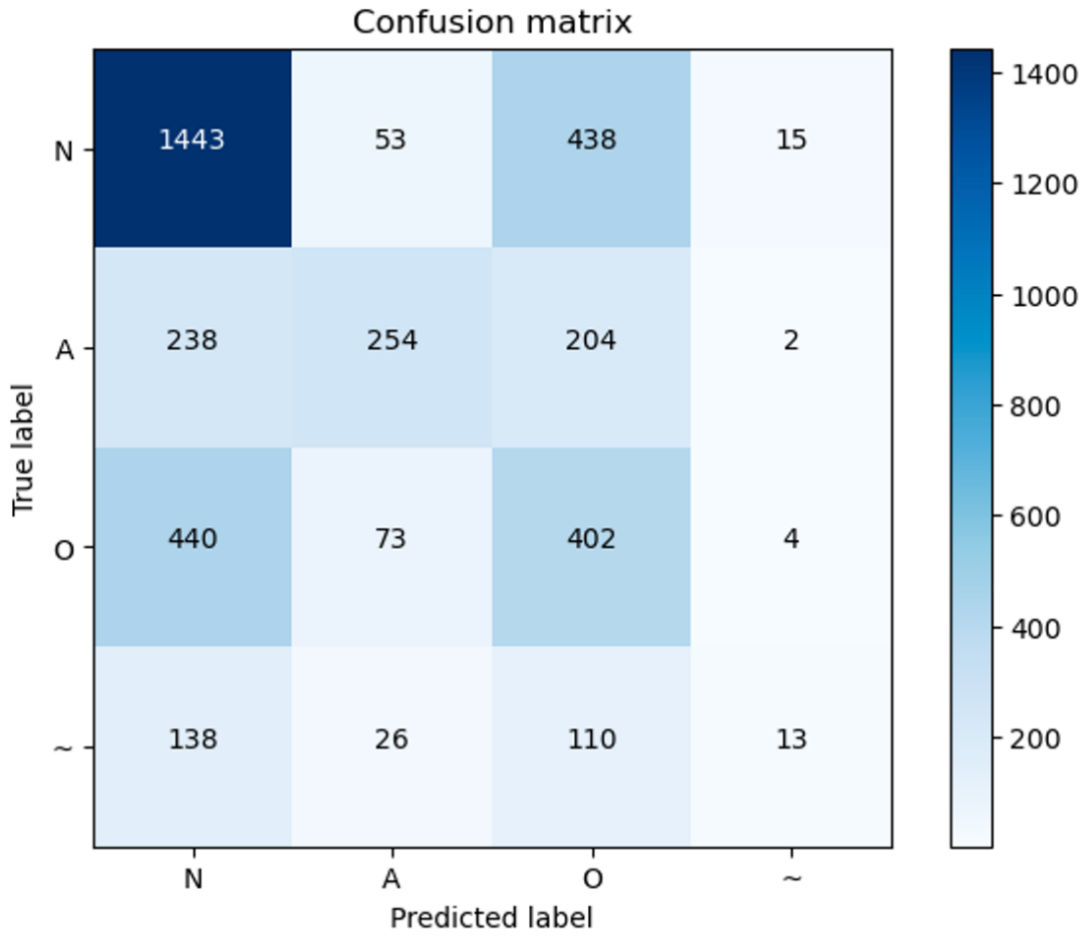
PhysioNet/CinC Challenge 2017 Dataset



PhysioNet/CinC Challenge 2017 Dataset: Results

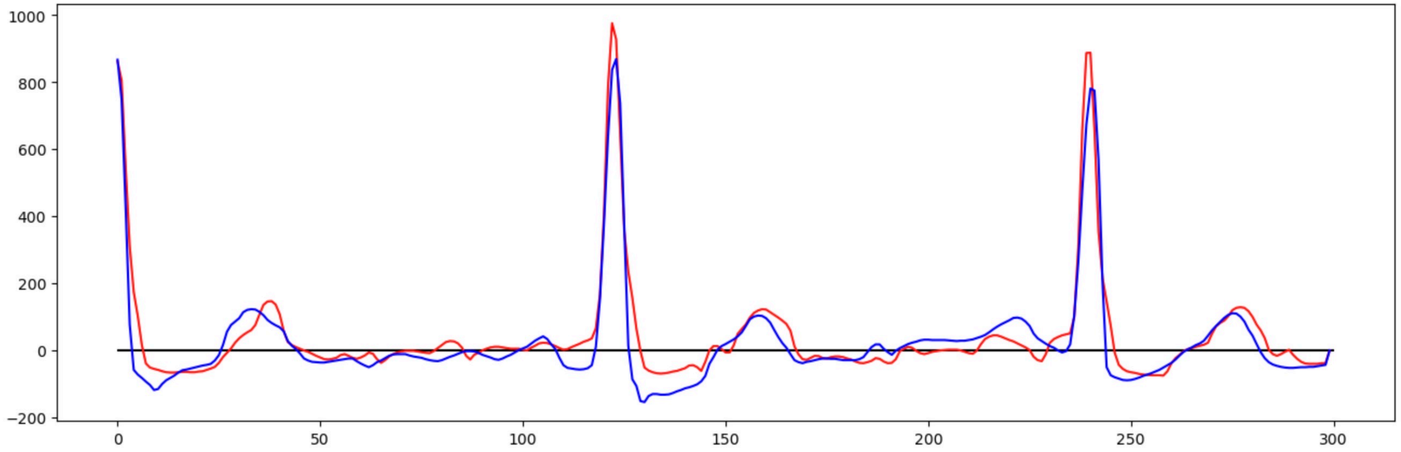
- Top 1 Accuracy: 54.81%
- Top 10 Accuracy: 58.29%

- Confusion among classes 1,2,3



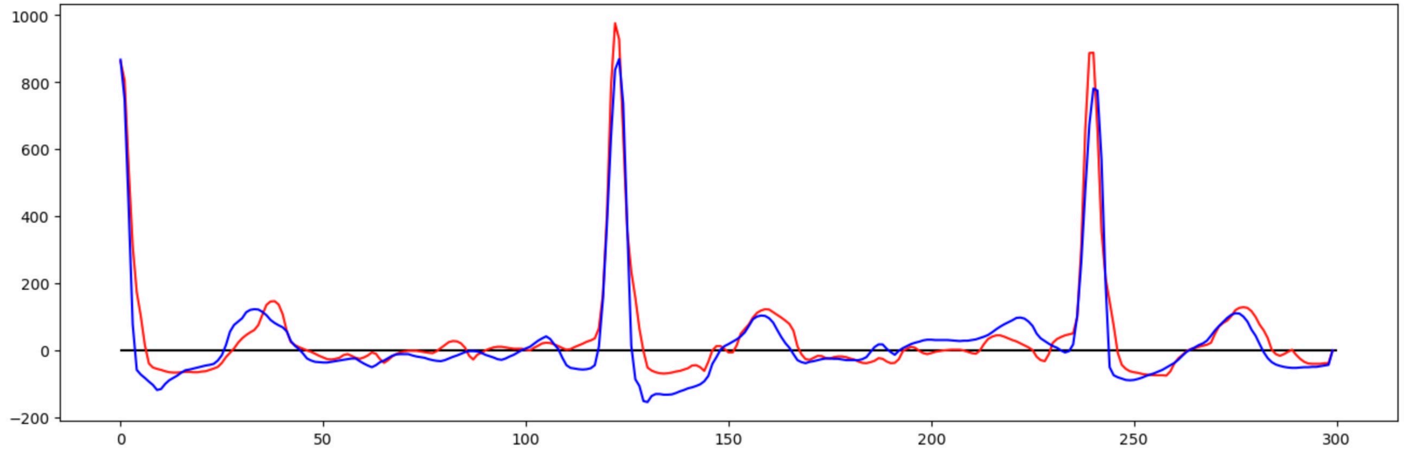
PhysioNet/CinC Challenge 2017 Dataset: Results

- Unable to achieve high accuracy due to the **difficulty level of the data & task**



PhysioNet/CinC Challenge 2017 Dataset: Results

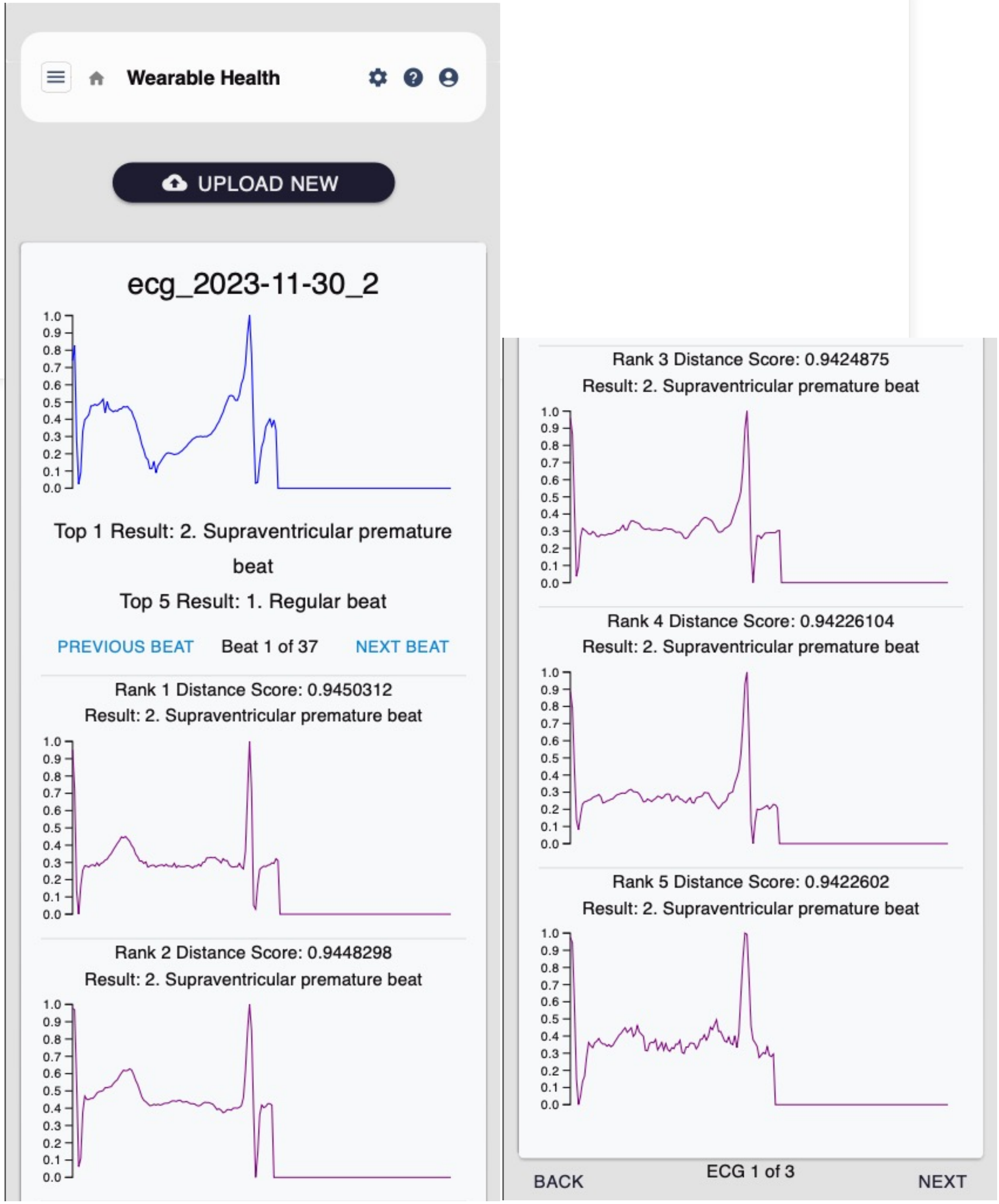
- Unable to achieve high accuracy due to the **difficulty level of the data & task**



- Blue = normal, red = other rhythm!

App demo

- Use ECG data collected from Apple Watch
  - In Apple Health app, click profile icon and scroll down to the end
  - Click "Export All Health Data" button, save the export.zip, and upload it to app
- Pinecone with MIT-BIH dataset

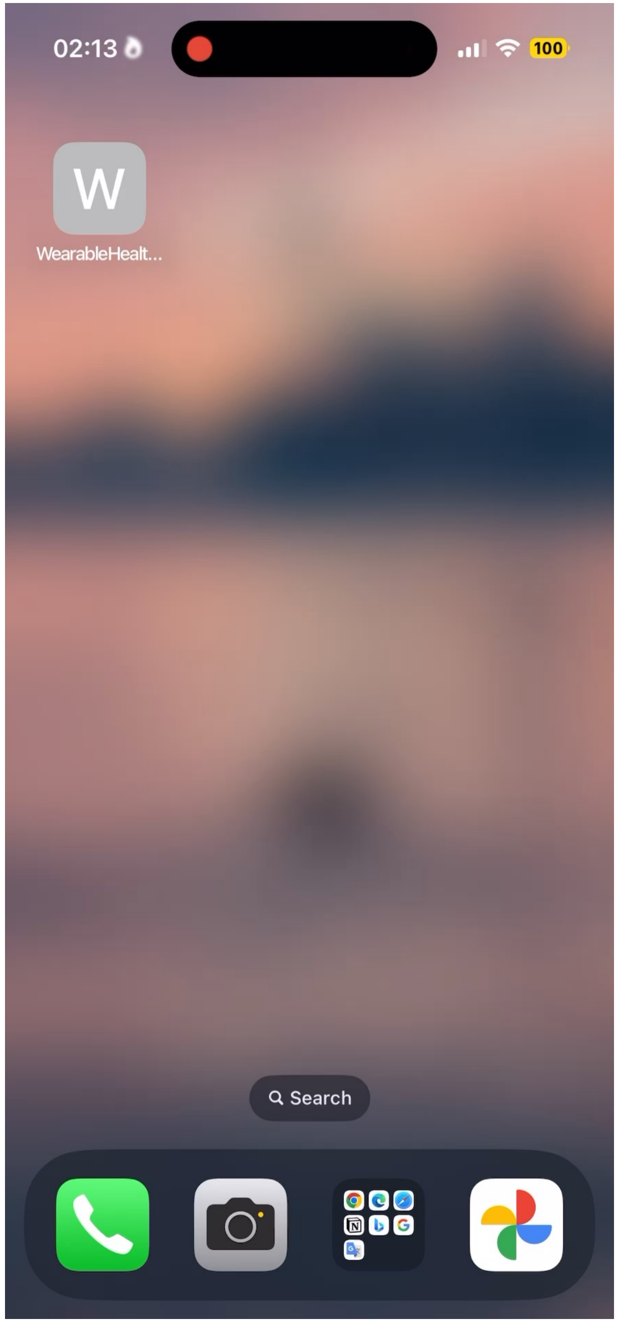


Video Demo

http://cos597a.yucanwu.com/
(available until Dec 9, 2023)

http://cos597a.yucanwu.com/demo

(available until Dec 31, 2023)



Video Link: https://drive.google.com/file/d/1_WHb4aabwujUbzh40ARQyKZ5RU7dtKKe/view?usp=sharing

Findings:
Pros and Cons of Vector-Search-Based Approach

- On certain tasks, vector database can perform comparably **well** or **better** than other machine learning models **without the need of training**
- Can add newly collected data, thereby **"fine-tune"** without re-training (and **personalization**)

- Direct vector database is only suitable for **certain easy datasets** & classification tasks due to the inability of **feature selection/extraction**
- Need to **tune hyperparameters** (distance metric, top k nearest neighbors, padding, etc.)
- Doesn't work well on class-imbalanced datasets

Conclusion & Future Work

- We showed that **real-time ECG health condition detection** is possible with high accuracy through vector database & similarity search

Future work:

- Incorporate **multiple** vector databases in the app => detect **multiple** health conditions
- Incorporate **multi-modal data** collected from smart watches
  - WESAD Dataset [1]: ECG, temperature, acceleration, electrodermal activity, etc.

Challenges:

- **Feature extraction** to make more complicated tasks possible

[4] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection,” Proceedings of the 20th ACM International Conference on Multimodal Interaction, Oct. 2018, doi: https://doi.org/10.1145/3242969.3242985.