DEPARTMENT OF
Computer Science
TCU COLLEGE OF SCIENCE & ENGINEERING

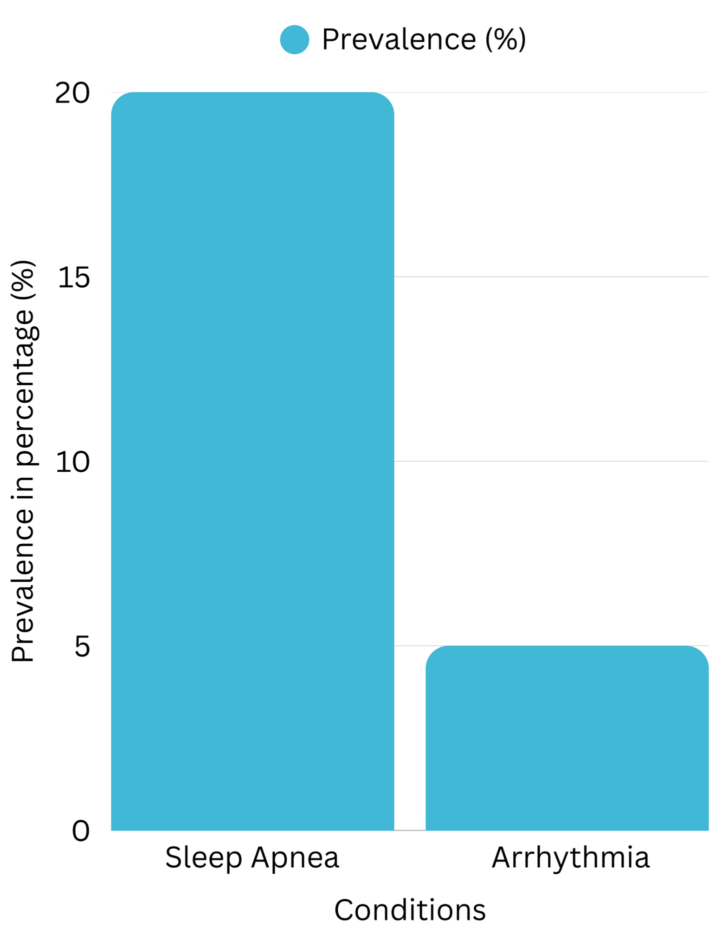
AI-Enhanced Early Detection of Disease Using Hybrid Real and LLM-Generated Wearable Data

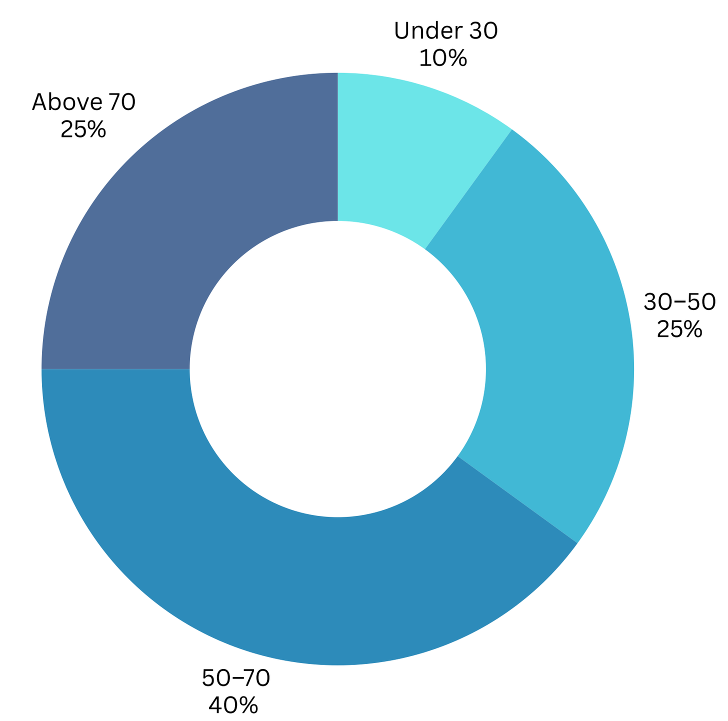
Author: Sujit Bhandari Advisor : Dr. Robin Chataut

SciCom Let's Talk Science

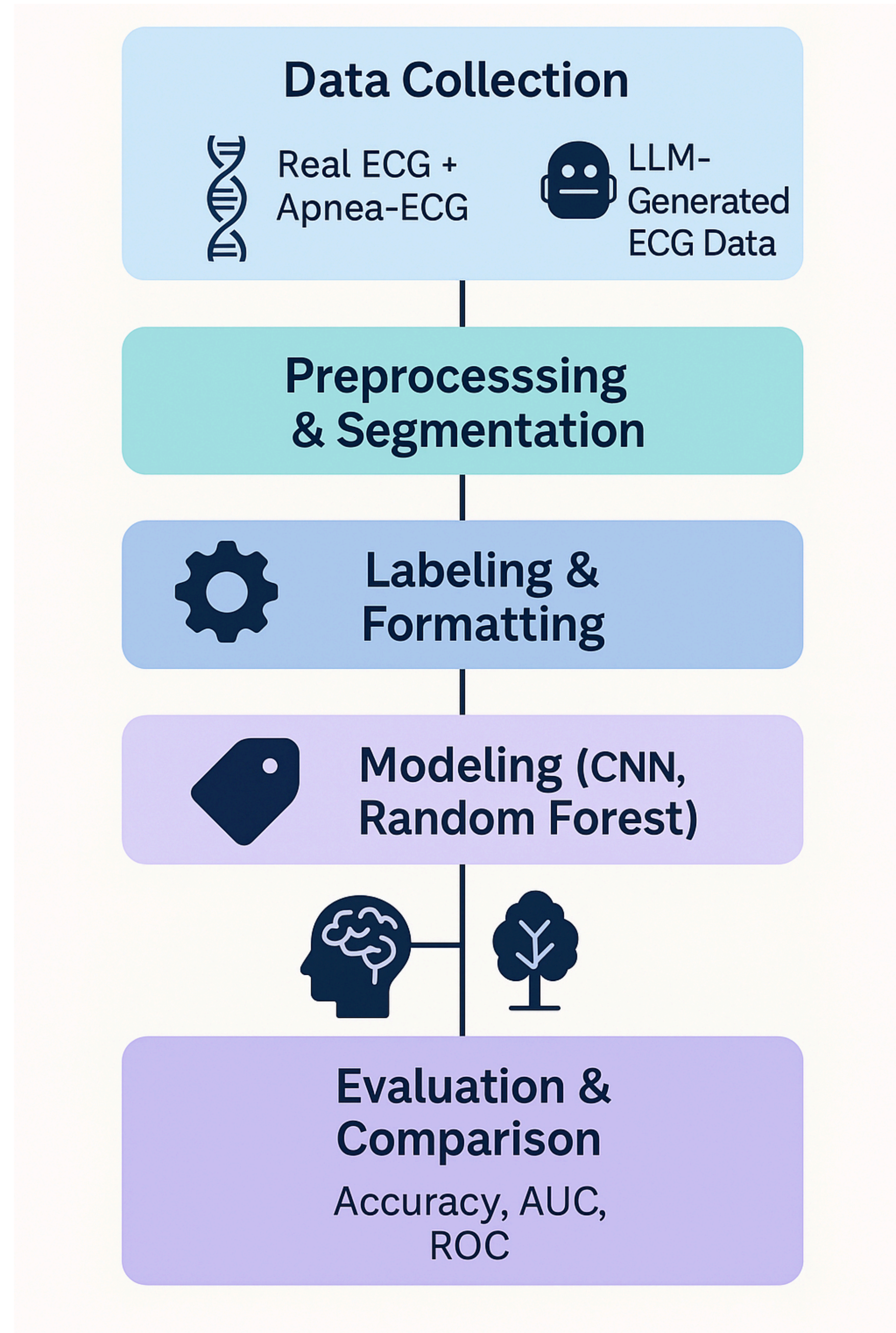
Introduction

- Early detection of arrhythmia and sleep apnea significantly improves health outcomes but is limited by privacy concerns and data scarcity.
- 20% of adults globally suffer from sleep apnea, increasing risks of hypertension and diabetes.
- 5% of adults experience arrhythmia, elevating the risk of stroke and heart failure, yet large-scale detection is limited due to privacy issues and lack of sufficient data.





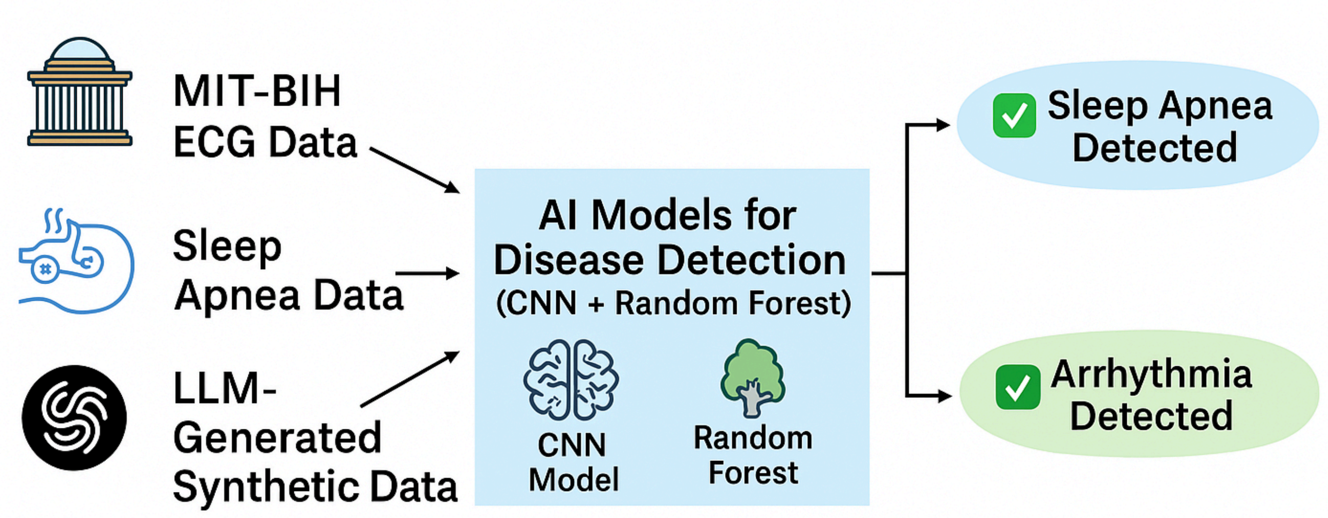
Methodology



Research Objective

- Build AI-driven diagnostic tools to detect early signs of arrhythmia and sleep apnea from physiological data.
- Develop CNN and Random Forest models to accurately identify these conditions from ECG signals.
- Create and utilize a hybrid dataset (real-world + synthetic data) to overcome data scarcity and privacy limitations.
- Evaluate and compare AI models for accuracy, interpretability, and performance using real-world physiological signals.

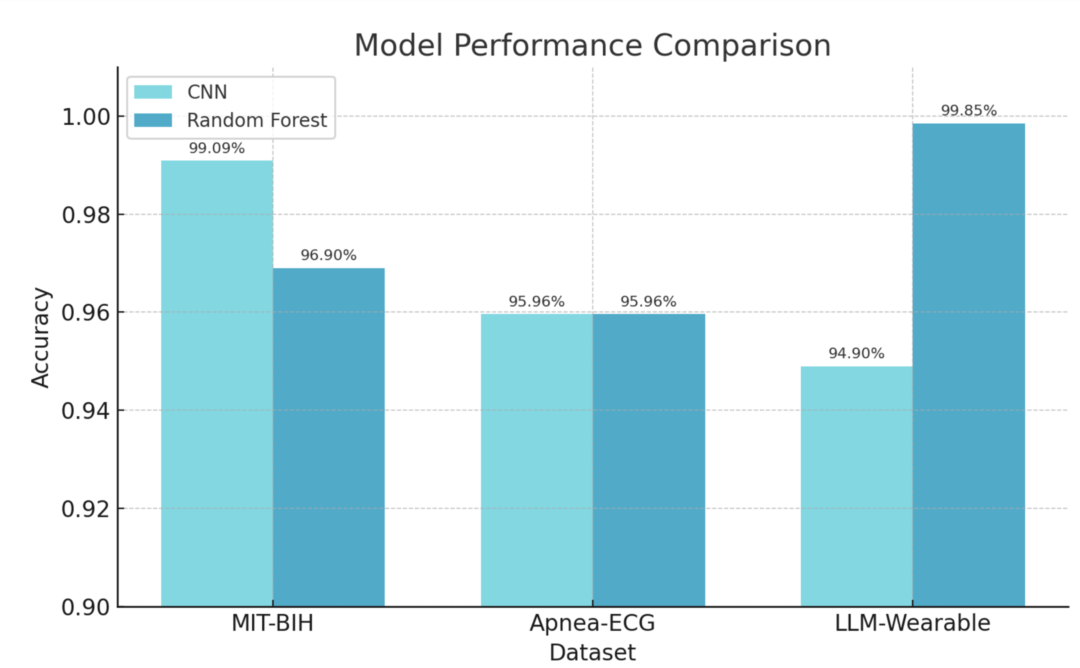
AI-Based Disease Detection Pipeline

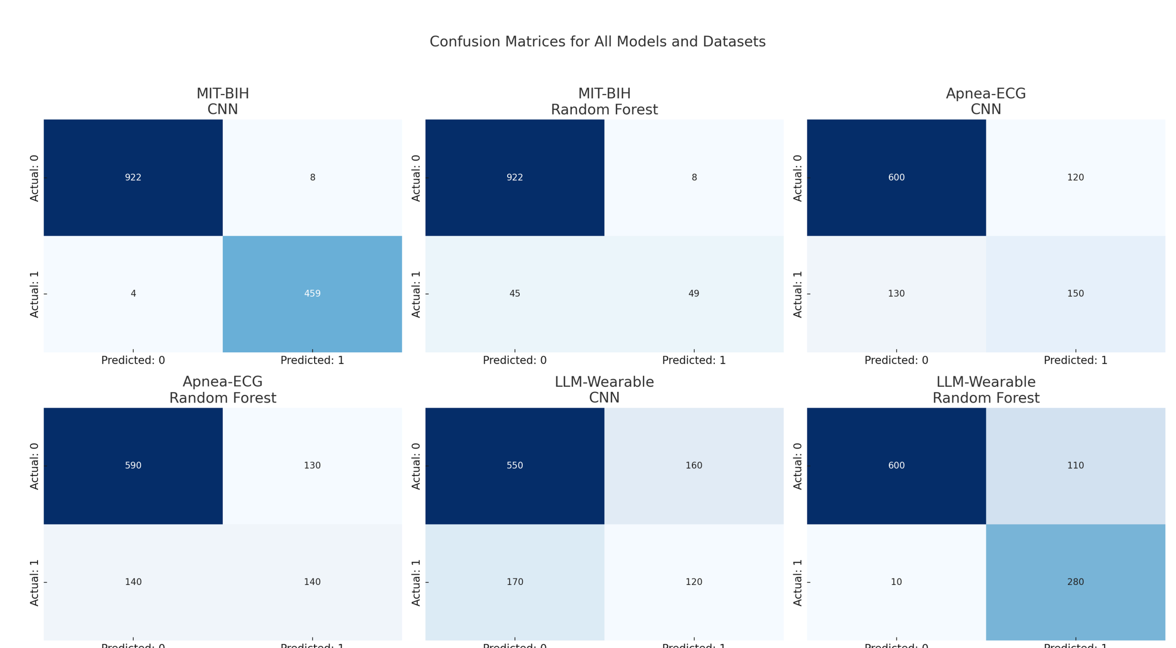


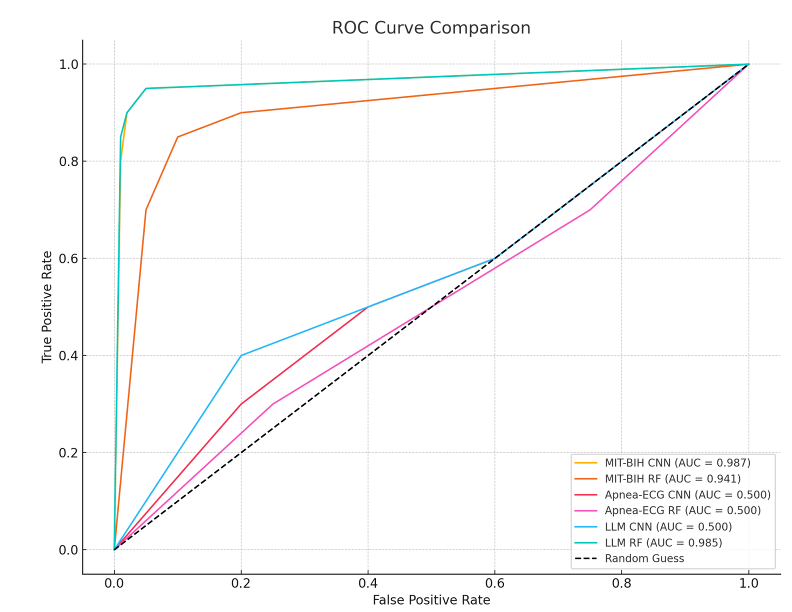
Hybrid training using real and synthetic physiological data

Output used to enable accurate and scalable diagnosis

Results

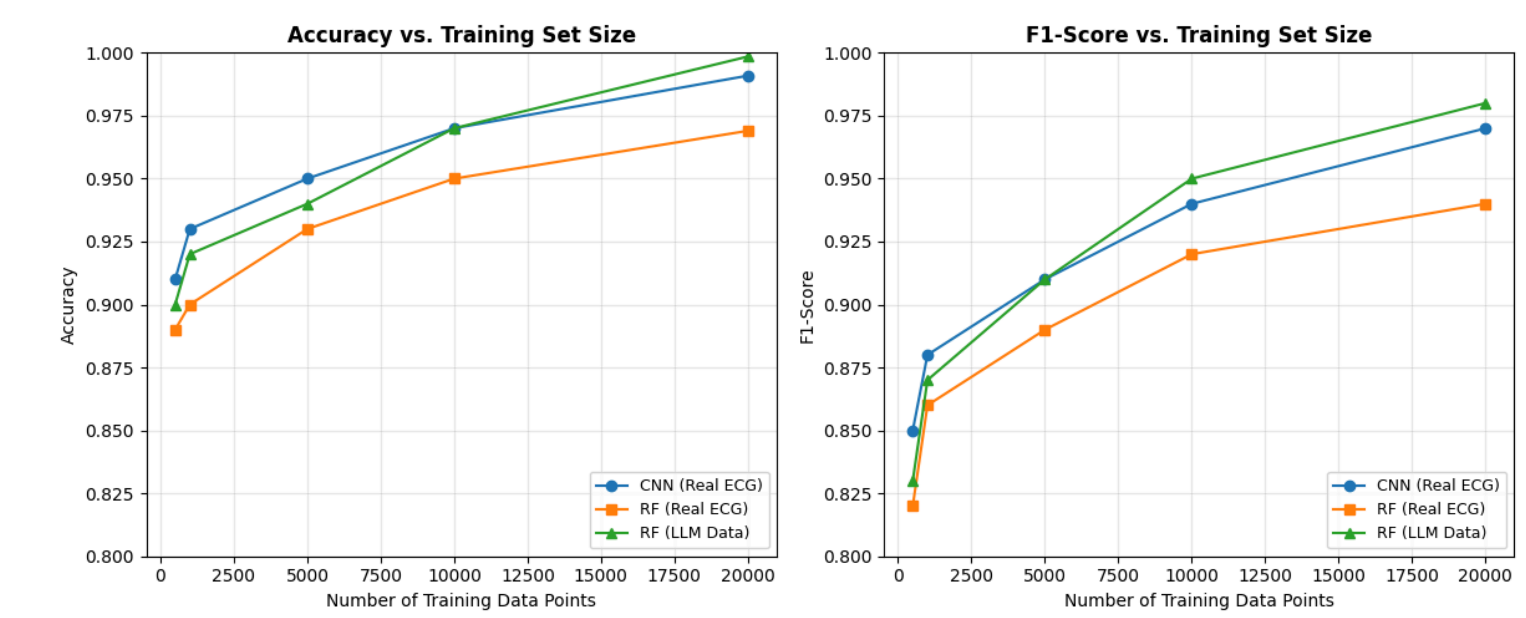






- CNN on MIT-BIH achieved the highest performance with 99.09% accuracy and AUC of 0.987, demonstrating strong reliability on real ECG signals.
- LLM-generated wearable data enabled realistic model behavior with moderate AUC, offering promise for privacy-preserving diagnosis.
- Random Forest underperformed on temporal ECG data but excelled with synthetic signals, achieving 99.85% accuracy on LLM data.

Machine Learning Analysis



CNN and Random Forest models showed improved accuracy and F1-score with larger training sets, with LLM-trained RF outperforming others at scale.

Conclusion and future Works

Our study shows that CNN outperforms Random Forest for arrhythmia and sleep apnea detection, benefiting from both real and LLM-generated physiological data. This hybrid approach ensures robust, accurate, and privacy-preserving disease detection. Looking ahead, we plan to expand detection to more conditions, refine LLM-generated signals for increased realism, and explore real-time wearable monitoring as well as Transformer-based architectures to further advance AI-driven healthcare diagnostics. Moreover, we aim to release open-source tools for reproducibility and encourage broader adoption. Finally, we intend to collaborate with healthcare providers to pilot these models in clinical settings, driving real-world impact.

References

MIT-BIH Arrhythmia Dataset
https://physionet.org/content/mitdb/

Apnea-ECG Dataset
https://physionet.org/content/apnea-ecg/