tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Unleashing The Power of Tiny Neural Network Models in Medical Devices"

Zhaojing (Jim) Huang – PhD Candidate at the School of Biomedical Engineering, University of Sydney

Leping (Steve) Yu – MPhil Candidate at the School of Biomedical Engineering, University of Sydney

May 14, 2024

TINY ML

www.tinyML.org

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

Executive Strategic Partners

Qualcomm AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency
Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
Continuous learning, contextual, always-on, privacy-preserved, distributed learning

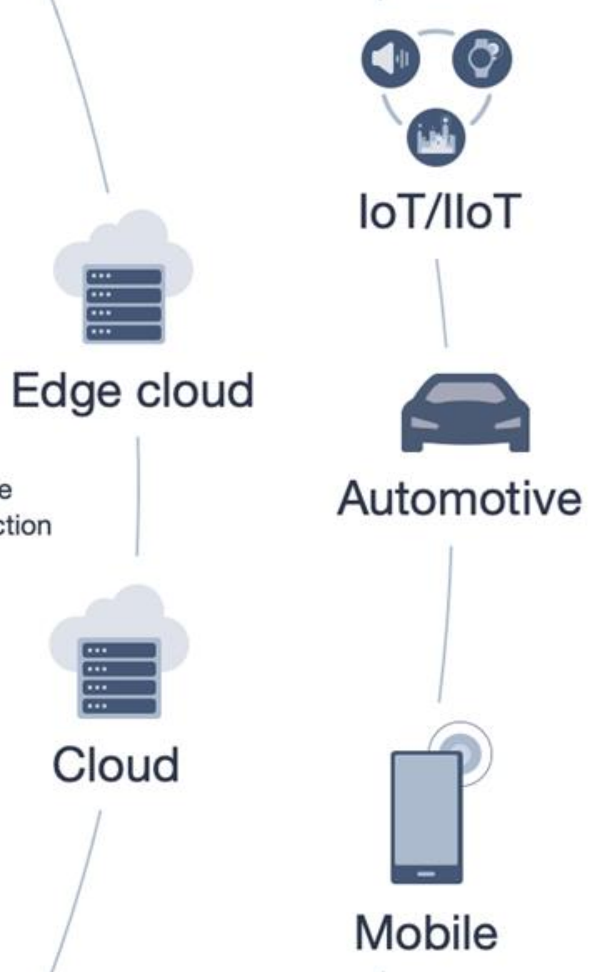
Efficient learning
Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception
Object detection, speech recognition, contextual fusion

Reasoning
Scene understanding, language understanding, behavior prediction

Action
Reinforcement learning for decision making



Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

embedUR

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners

Build the Future of tinyML

on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.

renesas.com

RENESAS
BIG IDEAS FOR EVERY SPACE

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

synaptics®

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

Silver Strategic Partners

Join Growing tinyML Communities:

meetup

20k members in
50 Groups in 42 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

4k members
&
16k followers

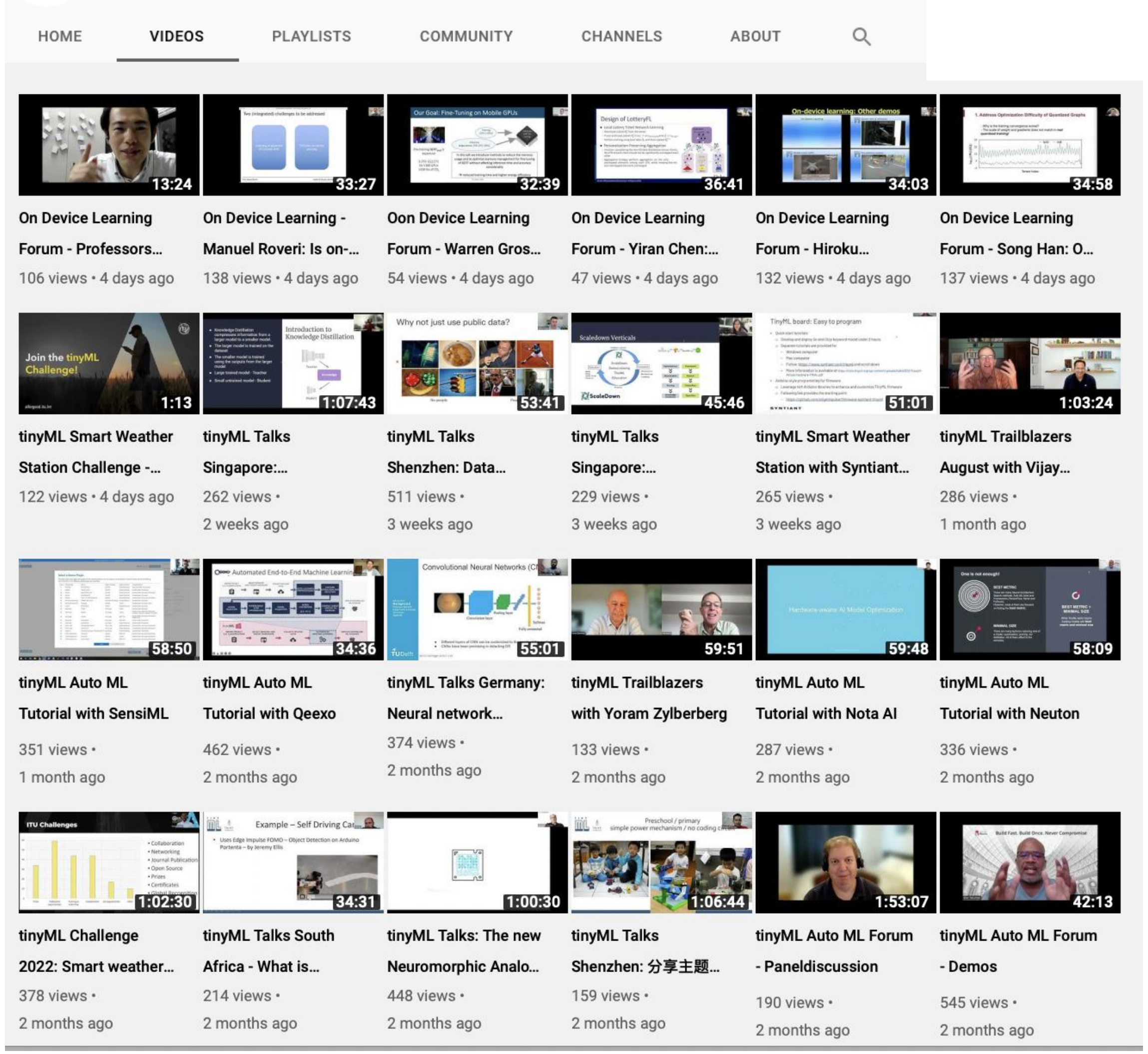
The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML

TINY ML TALKS webcast

12.6k subscribers, 685 videos with 461k views



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

TINY ML TALKS webcast

tinyML EMEA 2024

Amplifying Impact – Unleashing the Potential of TinyML

tinyML EMEA
June 24 -26, 2024 in Milan, Italy

REGISTER NOW

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums youtube.com/tinyml

Please use the Q&A window for your questions

Zhaojing (Jim) Huang

Zhaojing (Jim) Huang is a second-year PhD student in the School of Biomedical Engineering at the University of Sydney. His research focuses on the application of tinyML in the analysis of bio-signal data, particularly in the realm of medical diagnostics. With a keen interest in leveraging cutting-edge technology for healthcare advancements, he is committed to exploring the potential of machine learning in addressing critical challenges in biomedical engineering.

Leping (Steve) Yu

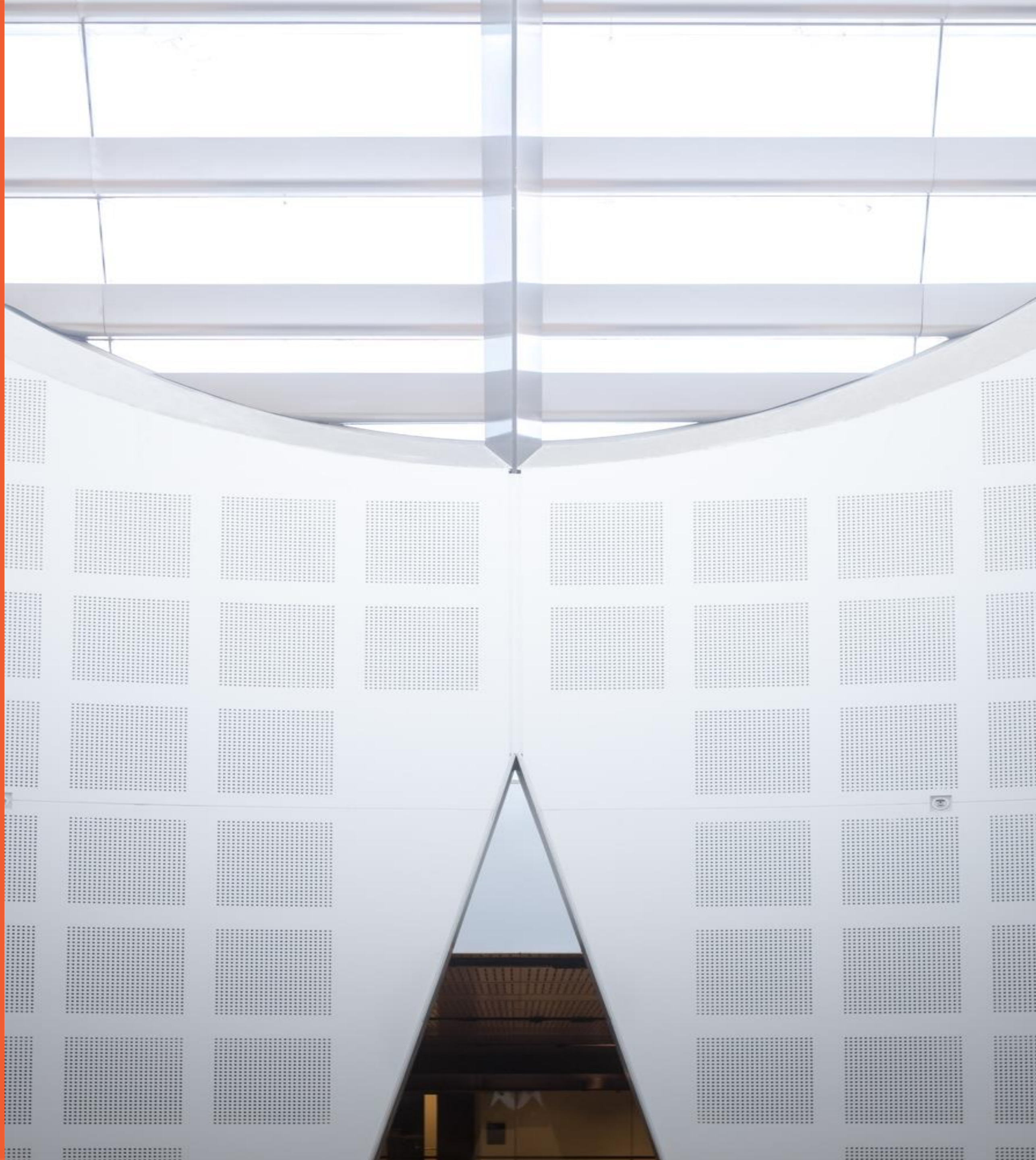
Leping Yu, a second-year Master's student at the University of Sydney's School of Biomedical Engineering, is dedicated to researching circuit design, signal processing, and system development, particularly in the realm of biosignal hardware, showcasing a strong interest in exploring diverse devices for biosignal measurements.

Unleashing The Power of Tiny Neural Network Models in Medical Devices

Presented by

Zhaojing (Jim) Huang

Leping (Steve) Yu



THE UNIVERSITY OF SYDNEY

Format

1. Introduction
2. Challenges
3. Strategies & Goals
4. Current Research
5. Future Directions

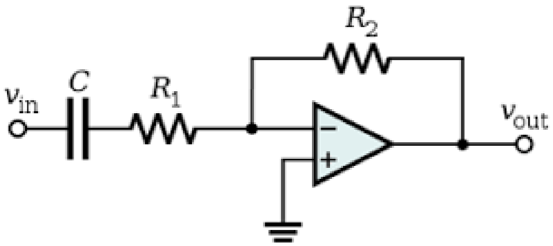
1. Introduction

Contextualizing and Providing an Overview of the Research Subject

THE UNIVERSITY OF SYDNEY

Wearable Biosignal Devices

- **Wearable Devices:**
  - Body-worn devices gather heart rate, ECG, EEG data
- **Biosignal Processing:**
  - Analyzes wearable biosignals with filtering, noise reduction, and feature extraction
- **Applications:**
  - Used in healthcare for continuous monitoring, early detection, personalized medicine, fitness tracking, stress management, and well-being



Topic Introduction

- **Machine Learning (ML):**
  - *Data-Driven Autonomy:* ML enables autonomous decision-making through data learning.
  - *Broad Influence:* It transforms healthcare, finance, autonomous vehicles, and language processing.
- **Medical/Bio-Signal ML:**
  - *Enhancing Diagnostics:* ML applies to complex medical bio-signals, improving patient care.
  - *Healthcare Transformation:* It revolutionizes healthcare through bio-signal analysis, like ECG and EEG, from disease detection to treatment optimization.

Vision and Goals

Vision:

- Achieving Precise AI Processing on Resource-Limited Medical Devices

Goals:

- *Optimize Bio-Signal Models:* Develop generalizable bio-signal processing models with low latency and minimal power consumption
- *Achieve AI Precision:* Enable accurate AI-based processing on resource-constrained medical devices

2. Challenges

Challenges hindering the pursuit of the vision and objectives

THE UNIVERSITY OF SYDNEY

Main Challenges

- Model Generalizability
  - Model performance deteriorates when deployed on external datasets
- Personalization
  - Customizing algorithms for specific user preferences to offer personalized experiences
- Model Size
  - Deploying large model architectures on devices is challenging
- Model Performance
  - The model's performance may be suboptimal with a smaller architecture
- Power Consumption
  - Power efficiency is crucial for battery-powered wearables

Model Generalizability

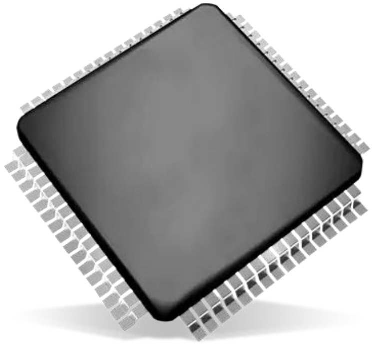
- Population Diversity
  - Model generalizability across diverse demographics (age, gender, ethnicity) is crucial
- Environmental Factors
  - Noise, movement artifacts, and different conditions, impact data quality and model performance
- Long-Term Adaptability
  - Maintaining model accuracy over time with continuous adaptation to changing signals and user behaviors

Personalization

- Dynamic User Preferences
  - Adapting to evolving user preferences for effective personalized recommendations over time
- Individual Variability
  - Personalized models account for individual differences in physiology, health, and lifestyle for tailored recommendations
- Long-term Engagement
  - Maintaining engagement with personalized insights and recommendations for consistent usage and adherence

Model Size

- Resource Constraints
  - Creating compact, efficient models for resource-constrained wearables
- Model Complexity
  - Balancing model complexity and size for accuracy and computational efficiency on wearable devices



Model Performance

- Accuracy
  - Aiming for high prediction accuracy in healthcare decisions for reliability and trustworthiness
- Low Latency
  - Reducing model processing time for quicker responses and feedback with minimized latency
- Resource Efficiency
  - Optimizing model resource usage for effective performance on resource-constrained wearables

Power Consumption

- Battery Life
  - Optimizing model design to extend wearable device battery life for continuous operation by reducing power consumption
- Efficient Algorithms
  - Energy-efficient models for data processing, inference, and communication to reduce computational load and power usage
- Low-Power Components
  - Utilizing low-power components in wearables to save power while maintaining performance

3. Strategies & Goals

Addressing the Challenges

THE UNIVERSITY OF SYDNEY

Strategies & Goals

- Utilizing ECG as the Primary Bio-Signal in the Study
- Study Techniques to Enhance Model Generalizability
- Create Models with Shallow Network Architecture
- Develop Power-Efficient Models
- Enhance Model Accuracy
- Hardware Compatibility

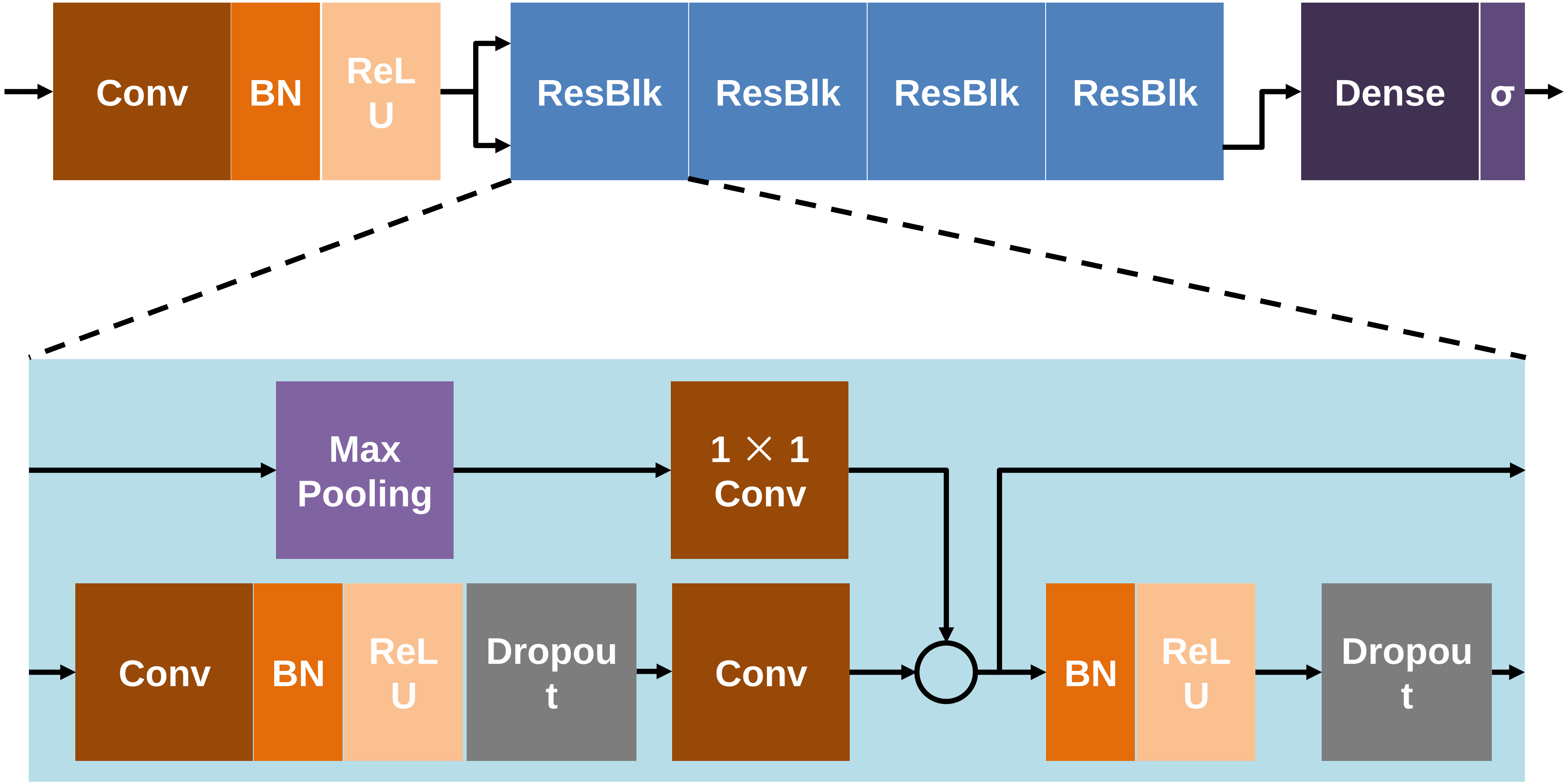
4. Current Research

Work Achieved Thus Far

THE UNIVERSITY OF SYDNEY

Generalizability

- Employing a model trained on the world's largest known 12-lead ECG abnormality dataset

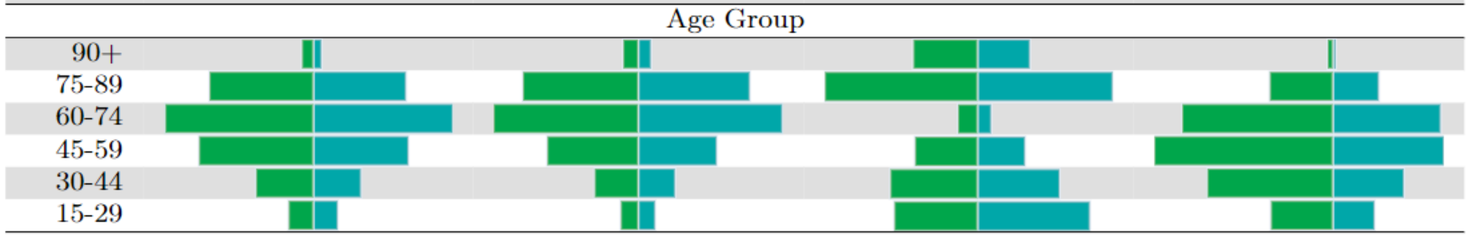


Ribeiro, Antônio H., et al. "Automatic diagnosis of the 12-lead ECG using a deep neural network." Nature communications 11.1 (2020): 1760.

Generalizability

- Selecting a subset from the dataset with a range of different characteristics
  - Balance of classes, Multi-abnormality patients, Age, Random

Subset	I	II	III	IV
1dAVb	3,234	6,677	3,341	333
RBBB	4,320	7,020	3,819	589
LBBB	3,443	4,935	3,411	340
SB	3,367	3,624	3,324	357
AF	3,561	4,043	3,846	377
ST	3,299	3,219	3,395	455
Normal	3,000	3,000	3,000	18,793



Huang, Zhaojing, et al. "Generalization Challenges in ECG Deep Learning: Insights from Dataset Characteristics and Attention Mechanism." Future Cardiology (Accepted).

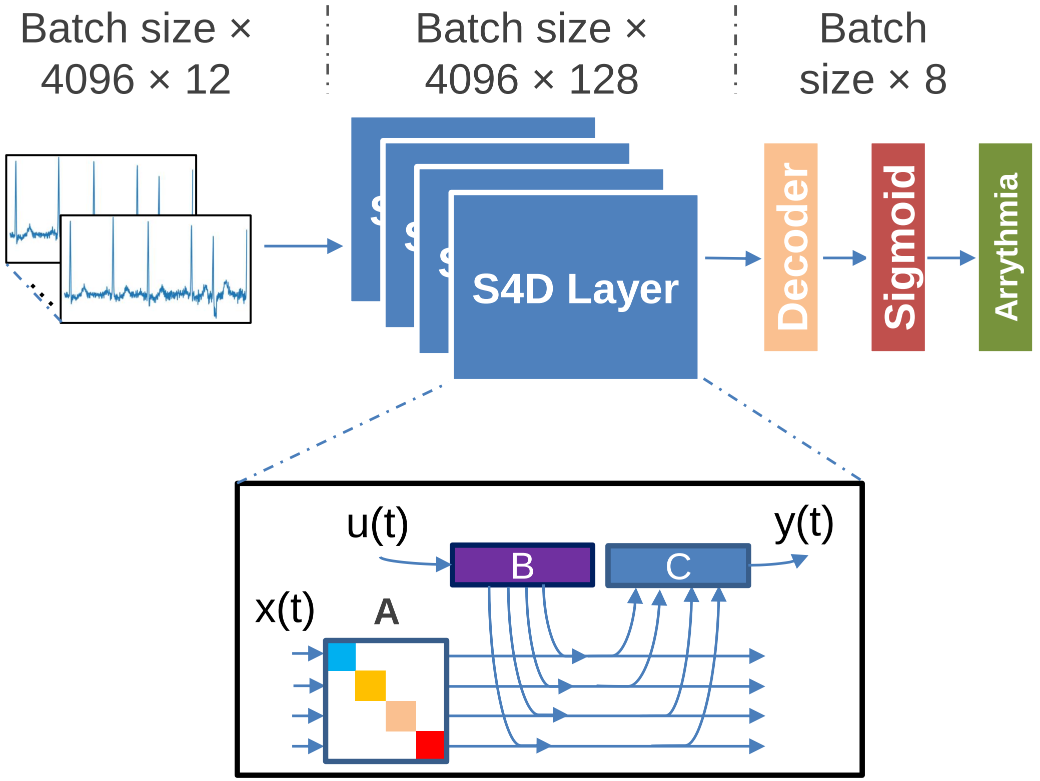
Generalizability

- Dataset characteristics play a vital role in model generalization
- A balanced dataset, even at just 1% of a larger set, can outperform larger set in generalization
- Self-attention mechanisms improve model generalization

Huang, Zhaojing, et al. "Generalization Challenges in ECG Deep Learning: Insights from Dataset Characteristics and Attention Mechanism." Future Cardiology (Accepted).

Shallow Architecture

- Diagonal State Space Sequence (S4D) model
  - Share a similar foundation with the Mamba model
  - Faster due to parallel computations for state variable updates, ideal for long sequences
  - Simpler implementation with fewer parameters and calculations compared to complex models like LSTMs
  - Advantageous for limited computational resources



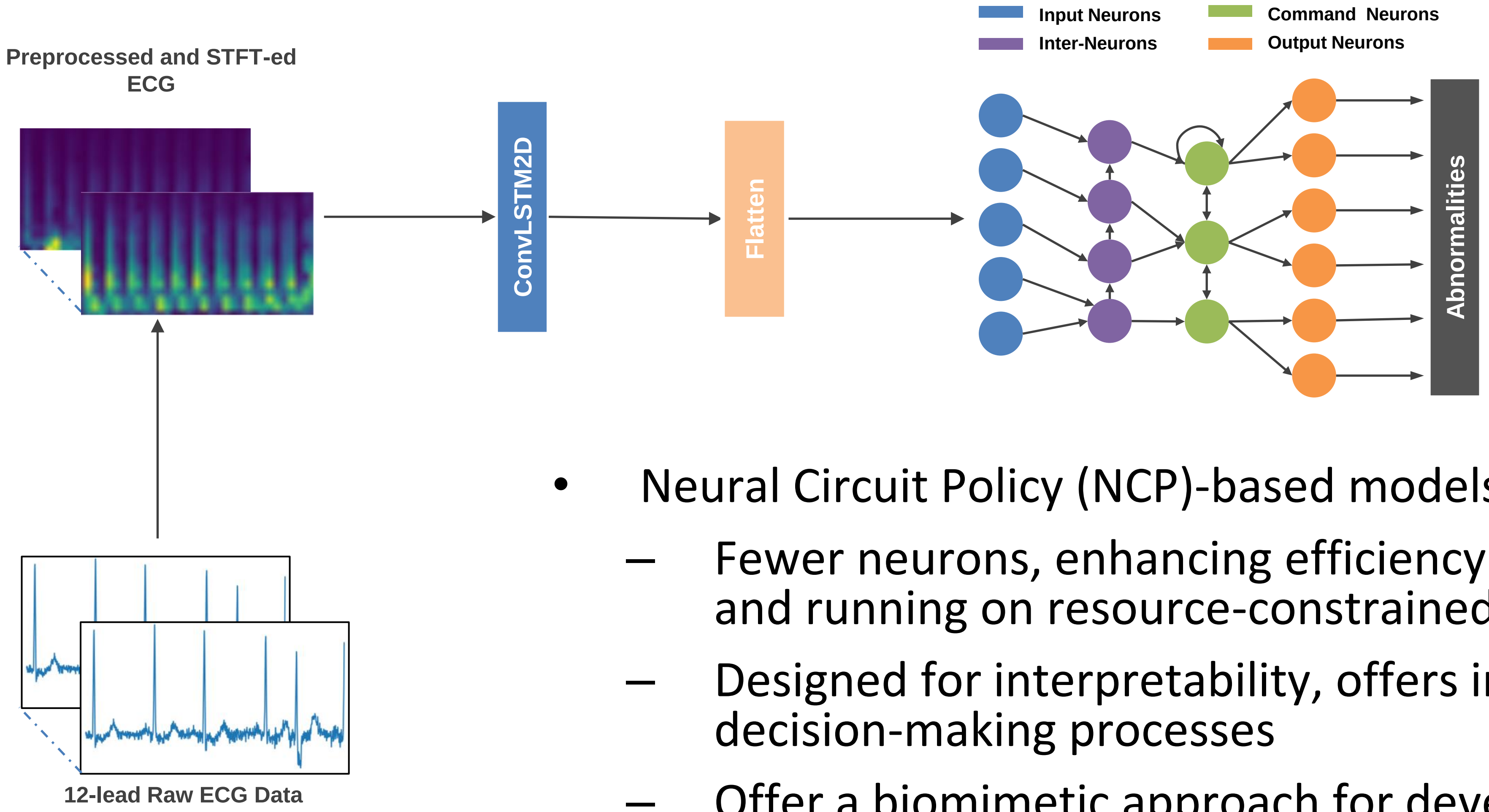
Huang, Zhaojing, et al. "S4D-ECG: A shallow state-of-the-art model for cardiac abnormality classification." Cardiovascular Engineering and Technology (2024): 1-12.

Shallow Architecture

- Diagonal State Space Sequence (S4D) model
  - Employed a stacked model with 4 S4D layers
  - Showcasing good performance and strong generalization capabilities
  - Demonstrated effective handling of moderate noise levels in the signal
  - Achieved excellent detection performance using only 1-lead data (Potential for edge device implementation)

Huang, Zhaojing, et al. "S4D-ECG: A shallow state-of-the-art model for cardiac abnormality classification." Cardiovascular Engineering and Technology (2024): 1-12.

Shallow Architecture



- Neural Circuit Policy (NCP)-based models
  - Fewer neurons, enhancing efficiency for training and running on resource-constrained devices
  - Designed for interpretability, offers insight into decision-making processes
  - Offer a biomimetic approach for developing efficient and robust AI systems.

Huang, Zhaojing, et al. "Efficient edge-AI models for robust ECG abnormality detection on resource-constrained hardware." Journal of Cardiovascular Translational Research (2024): 1-14.

Shallow Architecture

- Neural Circuit Policy (NCP)-based models
  - Introduces models: ConvLSTM2D-liquid time-constant network (CLTC) and ConvLSTM2D-closed-form continuous-time neural network (CCfC)
  - Both models perform comparably on TNMG data, with CCfC slightly more accurate and CLTC better at handling empty channels
  - Successfully deployed on a resource-constrained microcontroller, confirming generalization on the CPSC dataset
  - Efficient resource use: 70.6% memory and 9.4% flash memory of a STM32F746G microcontroller

Huang, Zhaojing, et al. "Efficient edge-AI models for robust ECG abnormality detection on resource-constrained hardware." *Journal of* Cardiovascular Translational Research (2024): 1-14.

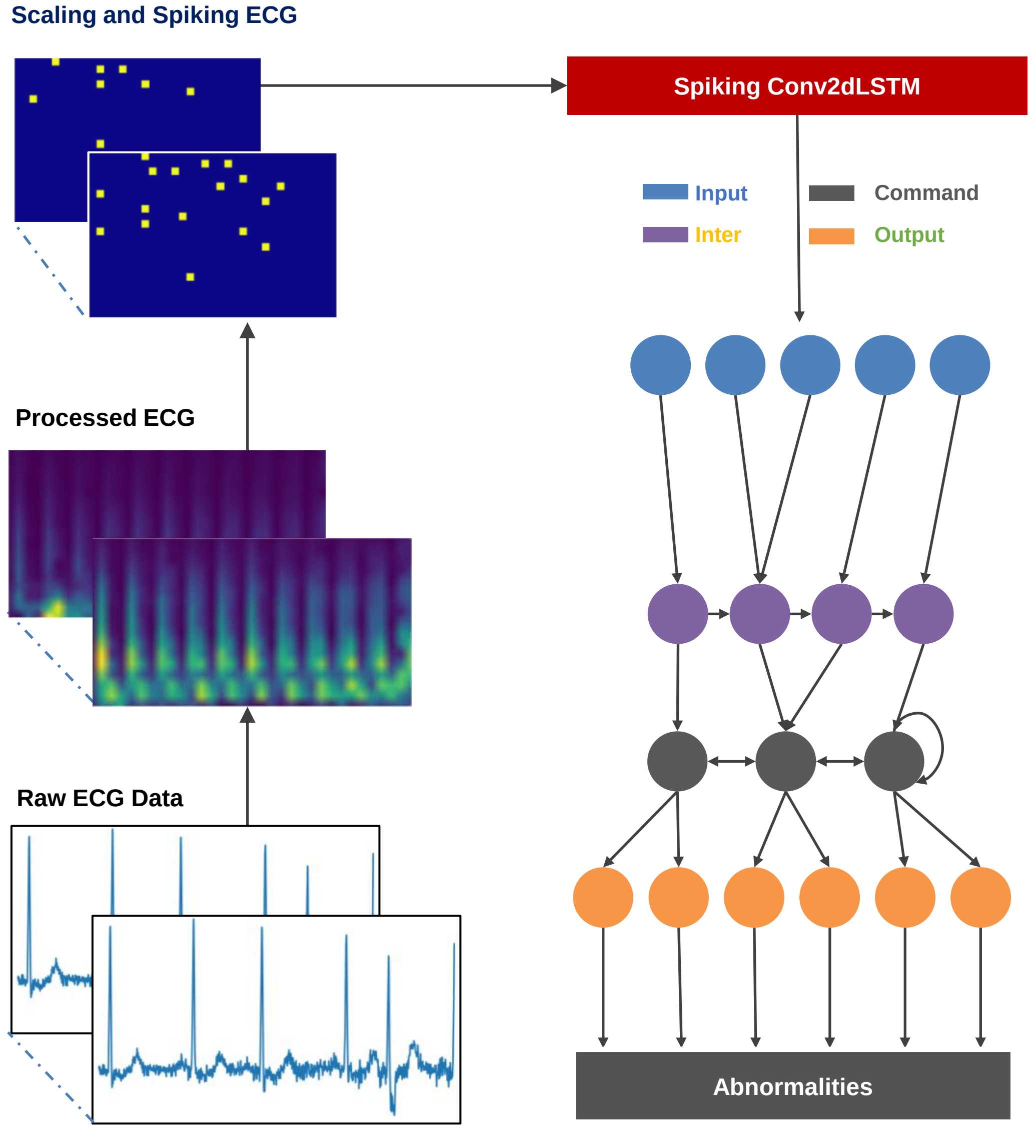
Power-Efficient Model

- Spiking Neural Networks (SNN)
  - Efficiently model temporal dynamics and event-based processing
  - Mimic biological neural networks closely, enhancing realism in data processing
  - Potentially achieve higher computational efficiency and reduced energy consumption compared to traditional artificial neural networks
  - Particularly effective for tasks involving temporal information, such as time-series data or event-based recognition

Huang, Zhaojing, et al. "On-device edge-learning for cardiac abnormality detection using a bio-inspired and spiking shallow network." APL Machine Learning 2.2 (2024).

Power-Efficient Model

- Fusion of spiking 2D ConvLSTM2D with bio-inspired CfC
- Comparable performance to non-spiking ConvCfC model
- Power Efficient: 4.68 µJ/Inf on a neuromorphic chip vs 450 µJ/Inf on a conventional processor



Huang, Zhaojing, et al. "On-device edge-learning for cardiac abnormality detection using a bio-inspired and spiking shallow network." APL Machine Learning 2.2 (2024).

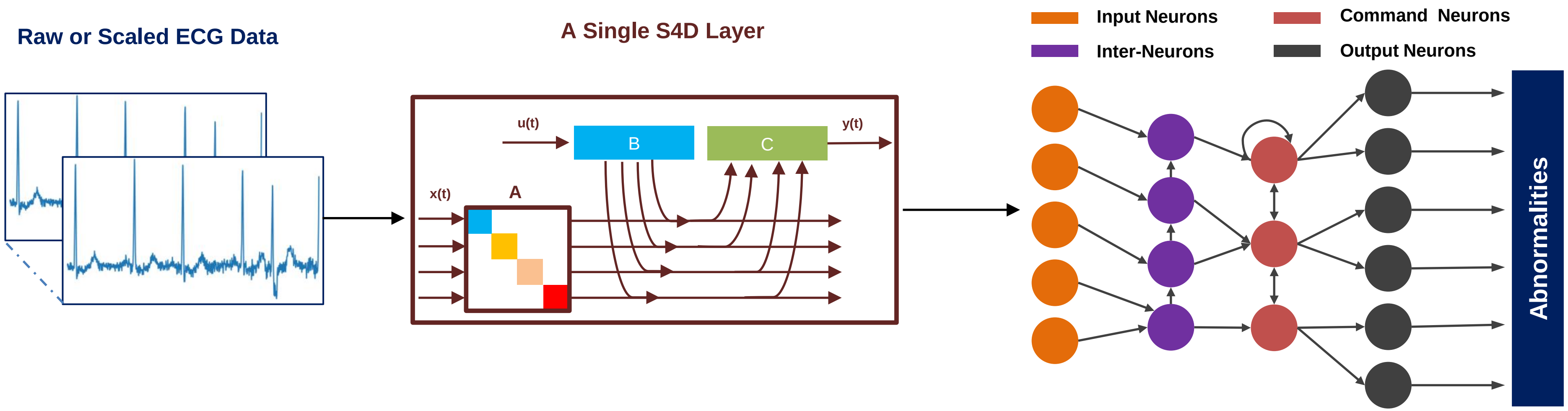
Power-Efficient Model

- Proof of concept demonstration for on-device training on the resource-constrained Radxa Zero microprocessor
- Superior robustness in handling missing ECG channels during inference compared to non-spiking ConvCfC model
- Effective single-lead ECG analysis with reasonable accuracy, despite focus on computational complexities

Huang, Zhaojing, et al. "On-device edge-learning for cardiac abnormality detection using a bio-inspired and spiking shallow network." APL Machine Learning 2.2 (2024).

Tiny Efficient Model

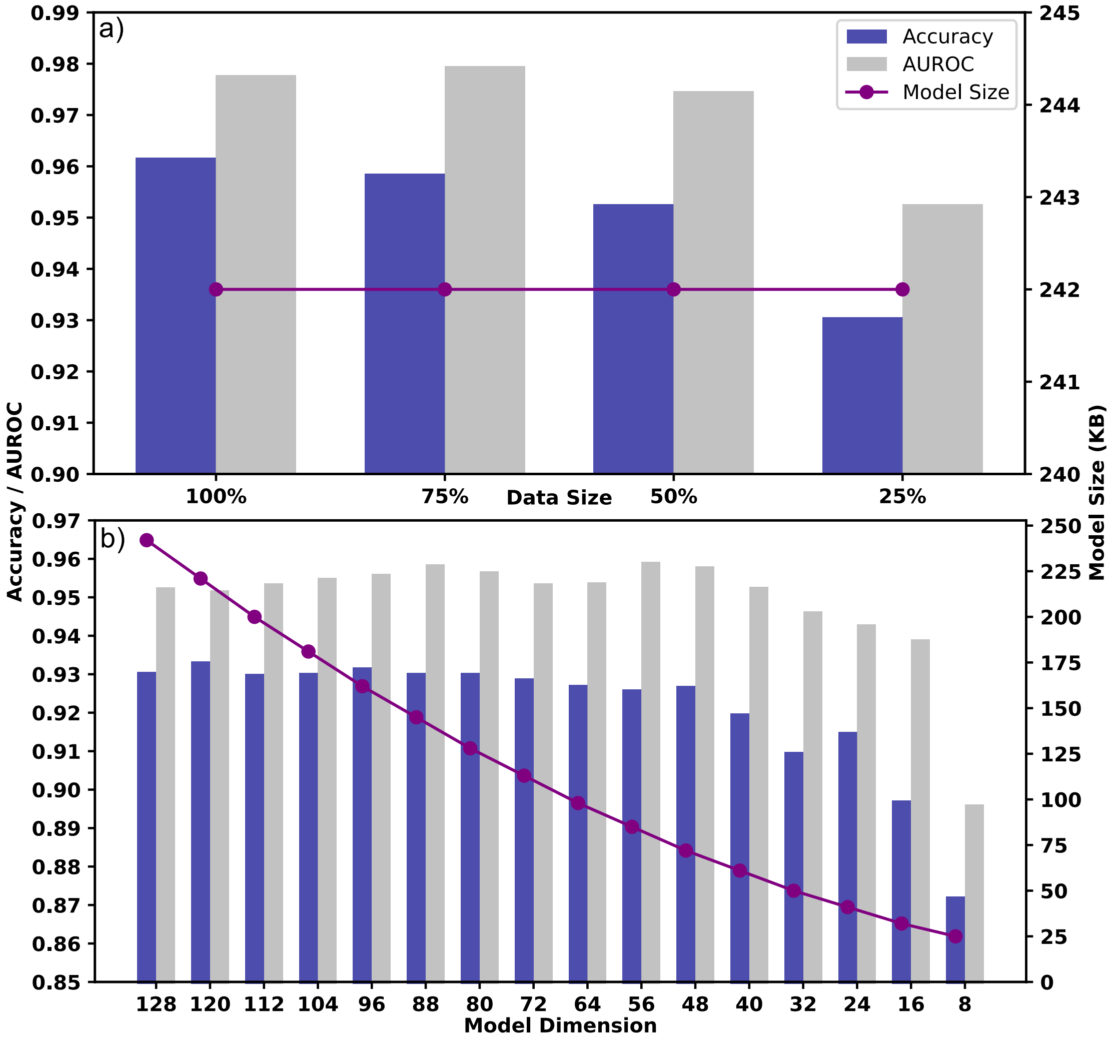
- Combined the S4D and NCP models to create a hybrid model, leveraging the strengths of both
- Achieved high performance and maintained a compact 242kB architecture
- Deployed this model on the Radxa Zero microprocessor for on-device training demonstrations



Huang, Zhaojing, et al. "Cardiac abnormality detection with a tiny diagonal state space model based on sequential liquid neural processing unit." APL Machine Learning 2.2 (2024).

Tiny Efficient Model

- Achieves precise detection with minimal data input, significantly lowering latency
- The model can be reduced to just 25KB, meeting even more rigorous resource constraints on devices
- Its compact size streamlines on-device fine-tuning, enhancing personalization capabilities



Huang, Zhaojing, et al. "Cardiac abnormality detection with a tiny diagonal state space model based on sequential liquid neural processing unit." APL Machine Learning 2.2 (2024).

5. Future Directions

Advancing Further

Future Direction

- Develop novel methods to enhance on-device inference accuracy
- Implement advanced techniques to preserve personalized fine-tuning capabilities
- Employ strategies to effectively address privacy concerns

Thank you!

Questions?

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org