In-Network Intelligence for Low-Latency Privacy-Preserving Health Monitoring

Aristide Tanyi-Jong Akem†‡, Huiqi Yvonne Lu†, and Noa Zilberman†
†University of Oxford, ‡University of Southampton
†{aristide.akem, yvonne.lu, noa.zilberman}@eng.ox.ac.uk, ‡a.tj.akem@soton.ac.uk

Introduction The Internet of Medical Things (IoMT) is transforming healthcare from reactive, hospital-based care toward continuous remote monitoring with wearable and ambient sensors. Most patient-monitoring deployments rely on the cloud for data aggregation and inference, which introduces latency, bandwidth, and connectivity dependencies and exposes sensitive physiological data to external networks. We propose **VISTA** [1], an in-network computing framework that embeds asynchronous sensor-stream aggregation and machine learning (ML) inference directly within a P4-programmable data plane on an IoMT edge gateway, enabling low-latency, privacy-preserving patient monitoring without cloud offload or reliance on personal mobile devices (Fig. 1).

Motivating Use Case: Consider a care home where dozens of residents are monitored using wearable sensors that stream heart rate, oxygen saturation, temperature, respiratory rate, blood pressure, and consciousness. Each resident may have ten or more devices, producing thousands of concurrent streams. Smartphones cannot serve multiple residents and typically depend on the cloud, whereas detecting deterioration or early signs of infection requires correlating signals across patients. The local IoMT gateway, which already collects all on-site sensor traffic, is the natural point for such analysis.

Why In-Network ML? Three challenges motivate moving inference into the network: *(i) data privacy and locality* — physiological data do not leave the gateway; *(ii) scalability* —in-network ML fit for national and populational scale of patient monitoring, especially when a large number of patients generate heterogeneous, asynchronous streams that exceed clinical capacity; and *(iii) resilience* — monitoring must continue under network or cloud disruption. Prior in-network ML targets traffic classification or intrusion detection on switches and SmartNICs [2]. VISTA is the first to bring in-network ML to clinical sensing on an IoMT gateway.

Implementation VISTA is prototyped in P4 on the bmv2 software switch, deployed in a Docker container on a Dell IoT Edge Gateway 5200. The pipeline in Fig. 1 maps four stages end-to-end. ① Incoming sensor packets carry patient and sensor IDs, timestamp and feature value, parsed at ingress and dispatched into a stateful aggregation module that opens per-patient time windows and indexes per-feature registers by patient ID. ② When a window is complete**,** two parallel match-action chains run in lock-step: one encodes the NEWS2 early-warning chart as range-match tables to compute the deterioration score, the other a compact XGBoost ensemble mapped to match-action tables to ③ predict health risks and generate alerts. Finally, ④ an alert packet carrying the NEWS2 level (Red/Amber/Green) and the ML risk predictions is sent to the clinician dashboard.

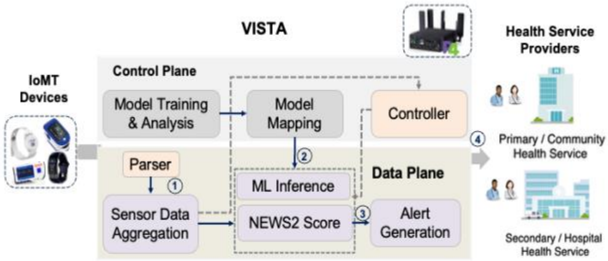

Fig. 1. Overview of patient monitoring with VISTA

Evaluation and Results: We evaluate VISTA on two datasets: a synthetic sepsis and heart-failure dataset of 2,000 patients (192,000 samples over 24-hour traces) and a maternal health risk clinical dataset. VISTA's inference performance closely matches that of the offline server-based XGBoost models. Accuracy differs by less than 0.001 for both prediction tasks: 0.9303 *vs* 0.9300 for sepsis and 0.9586 *vs* 0.9579 for heart failure. Similarly, the macro-F1 scores differ by less than 0.002, with 0.6815 *vs* 0.6809 for sepsis and 0.8223 *vs* 0.8204 for heart failure, and the weighted F1 scores follow the same trend. For example, the in-network ML with XGBoost matches its offline counterpart to within 0.001 accuracy (0.958 vs. 0.959 for heart failure; 0.930 for sepsis). Average packet processing time within the gateway is 2.4 ms (75% under 2 ms; 99th percentile 19.4 ms), roughly two orders of magnitude lower than reported for fog/cloud deployments (21–363 ms [3,4]). The system sustains a 100% alert success rate when scaling to 50,000 concurrent patients, with a mean CPU utilisation of 9.2% and a stable 580 MB resident set. Replacing raw-data uplinks with on-gateway inference cuts wire-line overhead by **90%**.

Summary: VISTA demonstrates that asynchronous multi-sensor aggregation and clinical ML inference can run end-to-end within a programmable IoMT gateway data plane, delivering millisecond-scale latency, privacy via data locality, and operations that survive cloud outages on commodity edge hardware. Future work will extend the prototype to live wireless sensor deployments, hardware P4 targets, and federated model updates across gateways.

Acknowledgement: This work has been published in the *IEEE Internet of Things Journal* (DOI: 10.1109/JIOT.2026.3688022). It was partly funded by the Leverhulme Trust and EU Horizon SmartEdge (101092908, Innovate UK 10056403).

References

[1] A. T.-J. Akem, H. Y. Lu, and N. Zilberman. In-network machine learning for real-time patient monitoring on IoMT edge gateways. IEEE Internet of Things Journal, 2026.
[2] C. Zheng et al..Planter: rapid prototyping of in-network machine learning inference. ACM SIGCOMM Computer Communication Review, 54,(1): 2–20, 2024.
[3] A. M. Rahmani et al. Exploiting smart e-health gateways at the edge of healthcare IoT: A fog computing approach. Future Generation Computer Systems, 78:641–658, 2018.
[4] T. Alsahfi et al. Optimizing healthcare big data performance through regional computing. Scientific Reports, 15(1), 2025.