International Journal of Communication and Information Technology

**E-ISSN:** 2707-6628
**P-ISSN:** 2707-661X
www.computersciencejournals.com/ijcit
IJCIT 2025; 6(1): 24-28
Received: 10-11-2024
Accepted: 14-12-2024

**E Edith Esther**
Assistant Professor, Department of Computer Science and Engineering, GRT Institute of Engineering & Technology, Tiruttani, Tiruvallur, Tamil Nadu, India

**P Shobha Rani**
Professor, Department of Artificial Intelligence and Data Science, R.M.K. Engineering College, Kavaraipettai, Tiruvallur, Tamil Nadu, India

**N Kamal**
Professor & Head, Department of Computer Science and Engineering, GRT Institute of Engineering & Technology, Tiruttani, Tiruvallur, Tamil Nadu, India

**Corresponding Author:**
**E Edith Esther**
Assistant Professor, Department of Computer Science and Engineering, GRT Institute of Engineering & Technology, Tiruttani, Tiruvallur, Tamil Nadu, India

# TINYML: A cutting-edge technology revolutionizing machine learning in healthcare

**E Edith Esther, P Shobha Rani and N Kamal**

**DOI:** https://doi.org/10.33545/2707661X.2025.v6.i1a.107

**Abstract**
In the recent technological revolution, Machine learning (ML) models are hugely successful due to the availability of the resources like powerful servers with large storage capacities and multiple graphical processing units (GPUs). Eventually, this paved way for large neural networks running virtually on unlimited cloud resources. However, there has been a shift towards the need for creating machine learning models tailored for edge devices that facilitate data transmission between local networks and the cloud. The aim is to eliminate delays caused by sending data to remote servers for processing, which can degrade performance and hinder the user experience. Instead, machine learning models must operate locally on these edge devices. This demand has led to the development of Tiny Machine Learning (TinyML), a technology designed for devices with constraints in data, memory, processing power and connectivity. Thus, this transformative technology promises to revolutionize healthcare landscape, paving the way for a healthier and more affordable future for everyone. This paper explores the underlying principles, key features and potential applications of TinyML, as well as its future prospects in the healthcare sector.

**Keywords:** TinyML, machine learning, limited memory, micro-controllers, healthcare, internet of things

## 1. Introduction

Tiny Machine Learning (TinyML) is an emerging, cutting-edge technology that enables the deployment of machine learning (ML) models on highly resource-constrained devices. It optimizes deep learning networks to run on tiny hardware, such as microcontrollers, sensors and embedded systems [1]. Typically, ML algorithms rely on previously collected data for input, which is then analysed to predict new outputs. Traditionally, ML and hardware are primarily linked to the cloud, where factors like latency, power consumption and connection speeds must be considered. To overcome these challenges, there is a need to reduce the computational demands of neural networks and algorithms by moving them closer to the network edge [2]. TinyML facilitates the execution of ML algorithms at the edge, enabling local data processing instead of relying on cloud infrastructure. This approach enhances data privacy by minimizing the need to transmit sensitive information to cloud servers, addressing security concerns while enabling more intelligent decision-making and real-time predictions, particularly in healthcare applications [3, 4]. Furthermore, TinyML's ability to significantly reduce energy consumption is crucial for many edge devices. These high energy-efficient edge devices can operate on battery power for extended time periods without frequent recharging [4].

Despite its increasing potential, TinyML continues to encounter various technical challenges, such as (i) the limited computational capacity of devices, (ii) the need for lightweight models that can operate on small-scale hardware and (iii) the development of efficient training methods tailored for these resource-constrained environments. However, TinyML is advancing rapidly and is poised to play a key role in the future of edge computing, enabling a broad array of applications across different industries. This technology holds particular promise in healthcare, where real-time data processing, low power consumption and portability are essential [4, 5]. In this paper, we examine the prowess of TinyML and its revolutionary potential in healthcare applications.

## 2. Principles of TinyML

The working principle of TinyML involves compressing complex ML models so they can operate efficiently on devices with constraints in memory, processing power and energy [6]. This is achieved through techniques like model quantization, pruning and efficient architecture design, which reduce the size and computational load of the models. Operationally, TinyML works by using sensors to collect data, which is then processed locally by the embedded ML model to make predictions or decisions without relying on cloud-based servers. This enables real-time processing, low latency and reduced power consumption, making TinyML ideal for Internet of Things (IoT) applications like smart homes, digital health, wearables and industrial automation [7, 8].

## 3. Features of TinyML

Some of the key features of TinyML are as follows;

**Low Latency:** TinyML processes data directly on the device (edge), reducing the need for constant communication with cloud servers. This enables faster responses and reduced latency.

**Low Power Consumption:** TinyML models are designed to operate on devices that have very low power consumption (<1 mW). This is critical for battery-powered or always-on devices, such as sensors and wearables. These microcontroller devices consume very little power, hence doesn't have to charge the power often.

**Low Bandwidth:** Processing on the edge eliminates the need for continuous data transfer, which helps in low-bandwidth environments.

**Privacy & Data Security:** By keeping data processing on-device, TinyML can enhance privacy and security. As sensitive data does not need to be transmitted to the cloud and data is not stored in any servers.

**Low-cost Hardware:** TinyML leverages inexpensive, off-the-shelf hardware, such as microcontrollers and specialized ML chips (e.g., Google's Edge TPU, ARM Cortex-M), making it a cost-effective solution for mass deployment.

**Energy Savings:** TinyML often utilizes specialized hardware accelerators, such as Tensor Processing Units (TPUs) or Neural Processing Units (NPUs), designed to run ML models efficiently with minimal energy usage.

**Internet Free Operation:** TinyML systems are designed to function without the need for constant internet connectivity. This is important for remote environments with intermittent connectivity or in cases where reducing dependency on cloud services is desired.

**Model Optimization:** TinyML involves optimizing machine learning models so that they can fit into devices with limited memory and processing power. Techniques like quantization, pruning, and knowledge distillation are often used to reduce the size of the models.
The optimization techniques, low-power design and integration with IoT make TinyML a promising technology for a wide range of applications in industrial automation, healthcare, agriculture and beyond.

## 4. Operational design of TinyML

The operational design of TinyML typically involves a combination of lightweight algorithms, energy-efficient hardware and optimized software frameworks to ensure that even with limited resources, the device can run ML models with high accuracy and responsiveness.

### 4.1 How to use it?

**Arduino Nano 33 BLE Sense**

TinyML along with Edge Impulse, they are turning the pervasive Arduino board into a powerful embedded ML platform. The Arduino Nano 33 BLE Sense is the preferred hardware for deploying ML models on edge. It as a 32-bit ARM Cortex-M4F microcontroller running at 64MHz with 1MB of program memory and 256 KB RAM. This microcontroller provides enough power to run TinyML models without frequent charging. The Arduino Nano 33 BLE Sense comprises of colour, brightness, proximity, gesture, motion, vibration, orientation, temperature, humidity, and pressure sensors. It has a digital microphone and a Bluetooth low energy (BLE) module.

**Raspberry Pi**

Raspberry Pi is also used to run the TinyML model, which requires much less power and space.

**Machine Learning Framework**

**There are various frameworks are available for TinyML. Some of them are**

1. TensorFlow Lite
2. PyTorch Mobile
3. Edge Impulse. Out of that

TensorFlow Lite (Product of Google) is the most preferred framework to deploy the TinyML models on microcontrollers.

### 4.2 What is TensorFlow Lite?

TensorFlow Lite (TFL) is a lightweight version of TensorFlow, which is a free and open-source library designed to enable machine learning on embedded devices with limited computational resources. TFL makes the inference at the edge by deploying ML models on linux based embedded devices such as *Raspberry Pi* or Microcontrollers and edge devices that uses Android or iOS. Figure 1 shows various process carried out by TFL. This is used in edge devices that run with less latency and minimum load. It converts the existing models into an optimized version and saves them in *tflite* file format.

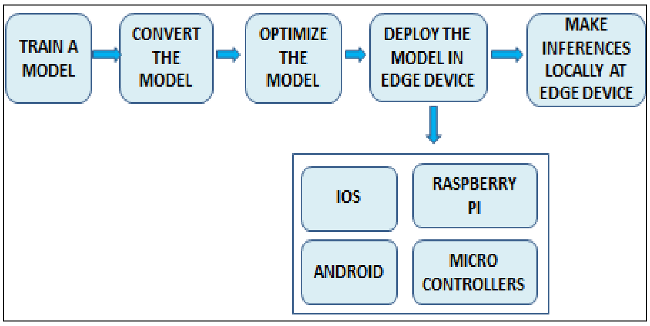


**Fig 1:** Diagrammatic representation showing the TensorFlow lite

### 4.3 How does TF Lite work?
It works by converting pre-trained TensorFlow models into a more compact format optimized for lower latency and smaller file size, suitable for on-device inference.

#### A. Selection and Training of model
This process involves two main steps: (a) model conversion and (b) optimization. First, a model trained in TensorFlow is converted into the TensorFlow Lite format (.tflite file) using the TFL Converter. Optimization of models were achieved by *Quantization* (Reducing the precision of the weights and activations) and *Weight Pruning* (process of trimming parameters within a model to improve performance).

#### B. Deployment and Inference Tasks
Once converted, the model is deployed onto a device, where TensorFlow Lite’s runtime library handles the execution of inference tasks. It ensures that the model runs efficiently, even on resource-constrained devices, by utilizing hardware accelerators (Like GPUs or specialized chips).

#### C. Inference and Mapping from edge devices
Making an inference on edge devices involves the following steps. Initialize the interpreter and load the model onto the device. Then, allocate the model and retrieve the input and output tensors needed for processing. Next, pre-process the data to ensure it's in the right format for the model. Perform the inference on the input tensor, allowing the edge device to analyze the data. Finally, map and obtain the result from the inference, providing the desired output based on the processed input.

## 5. Applications of TinyML
TinyML has a wide range of applications across industries where low-power, low-cost and real-time processing are crucial. Some of the notable applications of TinyML include:

### 5.1 Predictive maintenance
Using TinyML on low powered devices, it is possible to monitor the machine and predict faults ahead of time constantly. TinyML is used in industrial applications for anomaly detection and predictive maintenance by analysing sensor data from machinery. TinyML algorithms can monitor vibration, temperature, or sound data from industrial machines, detecting early signs of wear and tear. This allows companies to schedule maintenance before major failures occur.

### 5.2 Precision Agriculture
TinyML is being used to enhance smart agriculture, where it helps farmers make data-driven decisions with limited resources like internet. TinyML can analyse data from soil moisture, temperature sensors, or drone imagery to predict crop health and optimize irrigation. TinyML models can help detect pests or diseases in crops early on, leading to timely interventions and reducing the need for pesticides.

### 5.3 Smart Home Devices
TinyML enables smart home systems (Like thermostats, lights, and security cameras) to process sensor data locally. For example, motion detection or activity recognition can be done directly on a device, leading to faster response times.

### 5.4 Speech and Voice Recognition
TinyML plays a crucial role in enabling speech and voice recognition on low-power devices without needing an internet connection. Devices like smart speakers or portable voice assistants can use TinyML to process voice commands locally for improved speed and privacy. TinyML enables speech-to-text functionality on devices, allowing users to dictate messages or notes without relying on cloud services. This has applications in transcription, personal assistants and accessibility features.

### 5.5 Environmental Monitoring
TinyML can assist in monitoring environmental conditions in remote or challenging locations. TinyML can process data from sensors or cameras deployed in natural habitats, helping researchers to detect changes in ecosystems, like changes in air quality or temperature. TinyML can be applied in air quality monitoring systems, to detect air pollutants or particulate matter. These systems can be placed in various locations to provide continuous environmental monitoring.

## 6. Revolution of TinyML in Healthcare
The healthcare field generates vast amounts of data, particularly from portable and wearable devices. This raw data is typically transmitted to the cloud for processing, as machine learning algorithms like Neural Networks require significant computational power. However, the emerging technology of TinyML offers a transformative solution by enabling autonomous, secure devices that can collect, process, and send alerts without the need to transmit data to external systems. Some of the unique applications of TinyML in healthcare are detailed below;

### 6.1 Personalized Healthcare
TinyML allows for more personalized healthcare by adapting treatments based on continuous monitoring. Using data from wearable sensors or other personal devices, TinyML models can recognize an individual’s health patterns and recommend personalized interventions, such as adjusting medication dosages or physical activity levels. This creates more efficient, patient-specific care while reducing unnecessary treatments.

### 6.2 Early Disease Detection
TinyML models can be trained to detect early signs of diseases like diabetes, Parkinson’s disease, or heart conditions by analysing physiological signals such as movement patterns, skin temperature, or changes in voice. For example, wearable sensors can detect subtle motor changes in patients with Parkinson’s, while wearable glucose monitors can predict spikes in blood sugar levels in diabetic patients. By processing data locally, the devices can deliver early warnings to patients or caregivers without the need for cloud-based processing.

### 6.3 Remote Patient Monitoring
TinyML enables remote monitoring of patients in their homes, a crucial feature for individuals with chronic conditions or those who live in remote areas. Devices such

as wearable ECG monitors or pulse oximeters can run TinyML models to analyse patients' data and send alerts to healthcare providers if any anomalies are detected. This reduces the need for frequent hospital visits and helps ensure timely intervention in case of emergencies.

### 6.4 Wearable Health Monitoring Devices
TinyML enables the use of small wearable devices, like smartwatches, fitness trackers, or medical patches, to monitor vital health parameters such as heart rate, blood pressure, glucose levels, and respiration. These devices can process sensor data locally, without needing to send information to a cloud server, allowing for faster, real-time analysis. For example, a TinyML model could continuously monitor a patient's ECG (electrocardiogram) and alert them to potential arrhythmias.

### 6.5 Medical Imaging
TinyML is increasingly being applied in medical imaging, where lightweight models run on edge devices to analyse images such as X-rays, ultrasound scans, or CT scans. By performing basic image analysis directly on the device, TinyML can assist healthcare professionals in identifying abnormalities like tumours or fractures. This can speed up diagnoses, especially in remote or underserved areas without access to high-powered computational resources.

### 6.6 Drug Delivery and Management Systems
TinyML can be used in automated drug delivery systems, where real-time data is analysed to adjust medication dosage. For example, an insulin pump for diabetic patients can use TinyML to process glucose levels and adjust insulin delivery based on real-time readings. This ensures that patients receive personalized, dynamic treatment, which can improve health outcomes and reduce complications.

## 7. Challenges and Limitations of TinyML
TinyML faces several challenges, primarily the limited computational power of devices. This requires developing lightweight ML models that can run efficiently on microcontrollers with memory and processing constraints [5, 18]. Another challenge is ensuring data privacy and security, particularly in sensitive applications like healthcare or personal data [19]. Additionally, training ML models typically requires large datasets, which may be difficult to collect on small devices. Collaborative training methods, such as federated learning, are being explored to address this by enabling decentralized model training without transferring raw data [20].

## 8. Summary and Conclusion
In summary, TFL makes ML feasible on mobile and embedded devices by converting models into efficient formats, optimizing their execution and supporting hardware accelerators. This allows developers to deploy sophisticated AI applications in resource-constrained environments, offering faster inference times, lower power consumption, and improved privacy.

In conclusion, the post-COVID pandemic has highlighted the urgent need for innovative healthcare solutions. Technologies like TinyML, which offer low power consumption, minimal memory usage and no network dependency, have the potential to save lives at an affordable cost. With TinyML's ability to enable real-time, personalized health monitoring, the healthcare sector is poised for transformative growth. As this technology continues to advance, it promises to enhance patient care, reduce healthcare costs, and improve quality of life, making healthcare more accessible and effective for all.

**Conflict of Interest:** The authors declare that there is no conflict of interest.

**Authors' Contributions:** EE conceptualized, conducted the literature review, designed the outlay of this paper and wrote the draft manuscript. SR and KN corrected the draft manuscript. All the authors reviewed and approved the final version of the manuscript.

**Acknowledgements:** The authors acknowledge Dr. S. Arumugam, Principal and Committee Members of R&D Cell, GRT Institute of Engineering & Technology for providing the technical support and infrastructural facilities to carry out this work.

## References
1. Abadade Y, Temouden A, Bamoumen H, Benamar N, Chtouki Y, Hafid AS. A comprehensive survey on TinyML. IEEE Access. 2023;11:96892-96922. DOI:10.1109/ACCESS.2023.3294111.
2. Ray PP. A review on TinyML: State-of-the-art and prospects. J King Saud Univ Comput Inf Sci. 2022;34(4):1595-1623. DOI:10.1016/j.jksuci.2021.11.019.
3. Doyu H, Morabito R, Höller J. Bringing machine learning to the deepest IoT edge with TinyML as-a-service. IEEE IoT Newsletter. 2020.
4. Zhen Y, Guo Y, Chen M. The role of machine learning in enabling IoT-based healthcare applications. IEEE Access. 2021;9:38859-38890.
5. Zhao X, Li S, Wu X. Tiny machine learning for edge computing: Challenges and opportunities. Future Gener Comput Syst. 2020;108:103-115.
6. Sistla SMK, Konidena BK. IoT-edge healthcare solutions empowered by machine learning. J Knowl Learn Sci Technol. 2023;2(2):126-135. DOI:10.60087/jklst.vol2.n2.p135.
7. Tsoukas V, Boumpa E, Giannakas G, Kakarountas A. A review of machine learning and TinyML in healthcare. ACM Digital Library. Association for Computing Machinery. ISBN: 978145039555. 2021.
8. Shafique M, Theocharides T, Reddy VJ, Murmann B. TinyML: Current progress, research challenges, and future roadmap. In: 58th ACM/IEEE Design Automation Conference (DAC). 2021. p. 1303-1306. DOI:10.1109/DAC18074.2021.9586232.
9. Rancea A, Anghel I, Cioara T. Edge computing in healthcare: Innovations, opportunities, and challenges. Future Internet. 2024;16(3):329. DOI:10.3390/fi16090329.
10. Rajapakse V, Karunanayake I, Ahmed N. Intelligence at the extreme edge: A survey on reformable TinyML. ArXiv. 2022. DOI:10.48550/arXiv.2204.00827.
11. Soro S. TinyML for ubiquitous edge AI. ArXiv. 2020.

Available from: https://arxiv.org/ftp/arxiv/papers/2102/2102.01255.pdf
12. Dutta L, Bharali S. TinyML meets IoT: A comprehensive survey. Internet Things. 2021;16:100461. DOI:10.1016/j.iot.2021.100461.
13. Fedorov I, Stamenovic M, Jensen C, Yang LC, Mandell A, Gan Y, *et al*. TinyLSTMs: Efficient neural speech enhancement for hearing aids. ArXiv Preprint. 2020. Available from: https://arxiv.org/abs/2005.11138
14. Doyu H, Morabito R, Höller J. Bringing machine learning to the deepest IoT edge with TinyML as-a-service. IEEE IoT Newsletter. 2020.
15. Banbury C, Reddi JA, Lam *et al*. Benchmarking TinyML systems: Challenges and direction. ArXiv. 2020. DOI:10.48550/arXiv.2003.04821.
16. Beniczky S, Karoly P, Nurse E, Ryvlin P, Cook M. Machine learning and wearable devices of the future. Epilepsia. 2021;62:S116-S124.
17. Bharadwaj HK, Agarwal A, Chamola V, Lakkaniga NR, Hassija V, Guizani M, *et al*. A review on the role of machine learning in enabling IoT-based healthcare applications. IEEE Access. 2021;9:38859-38890.
18. Sze V, Chen Y, Yang T. Efficient processing of deep neural networks: A tutorial and survey. Proc IEEE. 2017;105(12):2295-2329.
19. Chen Z, Guo D, Chen C. Data privacy in TinyML applications. IEEE Access. 2020;8:196138-196151.
20. McMahan HB, Moore E, Ramage D, Band S. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017.
21. Gupta P, Chouhan AV, Wajeed MA, Tiwari S, Bist AS, Puri SC. Prediction of health monitoring with deep learning using edge computing. Meas Sens. 2023;25:100604. doi:10.1016/j.measen.2022.100604.
22. Ani R, Krishna S, Anju N, Aslam MS, Deepa OS. IoT-based patient monitoring and diagnostic prediction tool using ensemble classifier. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2017. p. 1588-1593.
23. Vincent ACSR, Sengan S. Edge computing-based ensemble learning model for healthcare decision systems. Sci Rep. 2024;14(1):26997. DOI:10.1038/s41598-024-78225-5.
24. Sanchez-Iborra R, Skarmeta AF. TinyML-enabled frugal smart objects: Challenges and opportunities. IEEE Circuits Syst Mag. 2020;20(3):4-18. DOI:10.1109/MCAS.2020.3005467.
25. Lane ND, Georgiev P, Punter J. TinyML: Machine learning on the edge. IEEE Commun Mag. 2019;57(6):96-102.