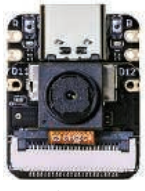
CUTTING AI DOWN TO SIZE



A $14 chip incorporating tinyML AI models, actual size shown.

Many artificial intelligence models are power hungry and expensive. Researchers in the Global South are increasingly embracing low-cost, low-power alternatives

By **Sandeep Ravindran**

The small drone circles the cashew tree, its rotor arms splayed out from its compact body like a water strider's. As it rises, its camera captures a bird's-eye view of the foliage, which shades from a dark glossy green at the tree's bottom to a purplish green at the top. Meanwhile an artificial intelligence (AI) model built into the drone determines whether the leaves are diseased—based on telltale black or brown splotches—or healthy. If needed, the system can one day direct other drones toward individual sick plants to treat their disease with a spritz of pesticide.

This system is the handiwork of Bala Murugan, a computer scientist at the Vellore Institute of Technology in Chennai, India. Murugan comes from a family of cashew farmers, including his father and uncle. "They apply a lot of pesticides on the cashew," he says. "I wanted to build a so-

PHOTO: SEEED STUDIO

lution to minimize the application of pesticides." But he needed a solution that didn't require internet connectivity, which is often hard to come by in rural India.

Murugan drew on his technical expertise. During his Ph.D., he had worked on small, cheap processors. Now, he realized AI models running on such small devices could help farmers like those in his family quickly identify and treat cashew disease. "That is when I ended up using tinyML," he says.

TinyML (the ML stands for machine learning) is a low-cost, low-power implementation of AI that is being increasingly adopted in resource-poor regions, especially in the Global South. In contrast to the large language models (LLMs) that have dominated the news with their versatility and uncanny knack for humanlike expression, tinyML devices currently have modest, specialized capabilities. Yet they can be transformative. Murugan's tinyML-equipped drones, for example, have been able to identify cashew leaves with the fungal disease Anthracnose with 95% to 99% accuracy. They should save farmers time they would otherwise spend looking for signs of disease themselves. And their ability to target treatments to diseased plants removes the need to indiscriminately spray pesticides on all the plants, which is both expensive and damaging to health and the environment.

Murugan is one of many researchers in the Global South finding uses for tinyML. The devices can serve as low-cost aids for teaching AI skills, but they are also providing homegrown solutions to problems that are not being sufficiently addressed by tech companies in the Global North, from detecting plant diseases to tracking wildlife. About 15 million tinyML devices were shipped in 2020, and that number, according to one estimate, could grow to 2.5 billion by 2030.

Part of the appeal for Murugan and others is that once the AI model is trained on a personal computer, it can often run for weeks on low-power tinyML devices powered by everyday batteries, sipping as little electricity as a typical laser pointer. The devices don't need internet connectivity, which can be scarce in resource-poor regions of the world looking to embrace AI solutions. Despite its limited capabilities, "I think [tinyML] is the future," says Marcelo Jose Rovai, a data scientist at the Federal University of Itajubá (UNIFEI). "It's fantastic for developing countries."

IN AI, "THE TREND HAS BEEN THE BIGGER, the better," says Thomas Basikolo, program officer in the Telecommunication Standardization Policy Department of the United Nations's International Telecommunication Union. The trend has culminated in current state-of-the-art generative AI models—those, including LLMs such as ChatGPT, that create new content. But the rise of these models "has so many issues," Basikolo says.

As LLMs are fed more and more training data and parameters to fine-tune and expand their ability to answer queries, the models have also grown dramatically in cost. The latest AI hardware chips can cost about $40,000 each, and ChatGPT alone uses tens of thousands of such chips.

Small but mighty

Tiny machine learning (tinyML) devices are orders of magnitude cheaper and less power hungry than the chips used to run artificial intelligence (AI) like large language models.

TinyML

$2–$60
Cost per device (including sensors)

≤1–100 milliwatts
Average power consumption per device

LLMs

$25K–$70K
Average cost per AI chip, requires tens of thousands of chips

700–1200 watts
Average power consumption per chip

Generative AI models are also hungry for power, which means more water usage and greenhouse gas emissions. Some estimates have ChatGPT sucking up almost 600 megawatt-hours of energy every day—more than 50 times what an average U.S. household consumes in a year. (And it took considerably more energy than that to initially train ChatGPT.)

These models typically run remotely on huge data centers accessed through the internet—or "in the cloud"—which means users need internet connectivity. The estimated tens of billions of liters of potable water needed every year to cool data centers leave a profound environmental impact, especially in areas where drinking water is already scarce. "Every time that you ask [ChatGPT] a question several liters of water are used cooling the big machines that are running big data centers in the cloud," Rovai says.

In 2019, former Google engineer Pete Warden explored a far less resource-intensive approach. He and his colleagues pared down traditional machine learning models by tweaking the numerical values that define how a model learns. They could cut unnecessary values and also reduce the precision of others without losing much overall accuracy. The result: models that could run on computer chips with limited memory and processing power.

TinyML models now rely on microcontroller chips similar to those found in everything from washing machines to car airbags. Warden and others first envisioned tinyML consuming so little power—less than 1 milliwatt—that devices could run for a year or more on coin-cell batteries and virtually forever using solar power. "There's some magic stuff that happens once you get below 1 milliwatt," Warden says. They haven't reached that threshold widely yet, but tinyML devices can run for weeks or sometimes months on AA batteries.

The chips themselves are cheap and commercially available from several different manufacturers. As a result, most tinyML devices now cost from a few dollars to tens of dollars, depending on how powerful they are. They often include not just the chip, but also cameras and sensors to detect images and sounds for the AI models to interpret. Much of the software, hardware, and data sets that researchers need to get started with tinyML are open source, which means they can freely access and modify them.

TinyML models typically use less data than larger varieties, ingesting thousands of images or sounds instead of the millions that LLMs often require. Murugan, for instance, says he used 20,000 images of cashew Anthracnose disease collected from various public domain sources to train his model, although he is supplementing that data set with his own photos to improve the model's accuracy. Once trained, the model does all its computation on the device, so tinyML systems can provide results within milliseconds with no need to go to the cloud and back. With no need to connect to the internet, tinyML uses less power and also tends to be a boon for privacy and security.

An artificial intelligence model running on tiny machine learning devices is helping cleanup efforts in Malaysia quickly classify types of trash hampering the growth of mangroves, including plastic food containers (●), plastic bags (●), food wrappers (●), clear plastic water bottles (●), or other kinds of bottles (●).

João Yamashita first came across tinyML in 2020 while completing his undergraduate degree remotely because of the COVID-19 pandemic. An electronics engineer at UNIFEI, Yamashita spent this time at his home in Mogiana Paulista, a coffee-growing region in the southeastern part of Brazil.

Yamashita realized that many small farmers were struggling to diagnose diseases in their coffee plants. Specialists to do this diagnosis didn't come cheap and wouldn't travel during the pandemic. Yamashita turned to tinyML for a possible solution.

After using public data sets to train a generic AI model, Yamashita went to the farmers to find out about coffee diseases specific to Brazil. "At first they were very skeptical," he says. "A lot of them don't even have cellphones," so an AI model to autodetect disease was very new, he says.

To fine-tune his model, which runs on a tinyML device about the size of a pack of cards, Yamashita took photos of healthy and diseased coffee leaves from rows of coffee plants growing on hillside plantations in his home region. He also collected samples and photographed them later under more controlled light and background conditions. The resulting model could identify a variety of coffee diseases with 96% to 98% accuracy, including fungal diseases such as Phoma, sooty mold, and rust, as well as the leaf miner moth. A farmer could point the device's camera at a leaf and its screen would show the disease name and a score indicating the model's confidence. "When I showed [farmers] the device working … they looked like they were amazed," says Yamashita, who is now working on tinyML devices for other uses.

The device is practical for poor farmers, Yamashita says. It doesn't need internet access, lasts a week or more without needing to charge its battery, and costs less than $20. It's just the sort of bespoke, homegrown solution tinyML is good for, and explains why agriculture has been a popular use for the technology.

Beyond agriculture, researchers are also developing tinyML devices for health care applications, from detecting atrial fibrillation—a type of abnormal heart rhythm—in Brazil to anemia in Peru. And multiple groups have used the technology to distinguish mosquito species by the buzzing of their wings, making it unnecessary to collect mosquitoes in traps and manually identify each one. The result is speedier alerts that could help with control of species that spread disease.

Rovai, for example, trained an AI model to identify two species of *Aedes* mosquitoes—which can transmit dengue, Zika, and chikungunya viruses—with an accuracy of 98% on a proof-of-concept device whose battery can last for up to 4 days in the field. He envisions the device being of great use in Brazil, where dengue affects more than 1 million people each year. In Kenya, similar projects are using tinyML to automatically classify mosquitoes that carry malaria. "Being able to categorize that in an automatic way is a huge advantage for people working in the field," says Marco Zennaro, a computer scientist at the Abdus Salam International Centre for Theoretical Physics (ICTP) who worked with Rovai on the *Aedes* mosquito detection project.

Similar devices are also finding their way into environmental applications. Researchers have tacked tinyML devices to the back of tortoise shells in Argentina to track how and where the animals move. And in Malaysia, Rosdiadee Nordin, an engineer at Sunway University, is using tinyML devices to monitor rivers for plastic trash that might hinder the growth of delicate young mangroves. He and a team of volunteers gathered 9000 images of plastic trash to train AI models until they could classify it, distinguishing clusters of plastic bottles from clumps of plastic bags. They plan to make the data publicly accessible to help track the location, quantity, and types of plastic waste. "This will help not only [those who pick up the trash], but also local council or environmental agencies to further plan their waste collection activities," Nordin says.

Nordin's work extends to Tasik Chini Lake in eastern Malaysia, which supplies the local Indigenous community with both drinking water and fish. Nordin deployed solar-powered water quality sensors to de-

IMAGE: SITI ZULAIKHA; ROBOFLOW

tect pollution in the lake and send that information to tinyML devices, where AI models use it to make predictions about water quality. "If we make the water quality data accessible [to] the community, they will be able to understand whether the water is safe for them to consume," Nordin says.

The lake lacks the internet and cellular connectivity needed to broadcast data to the researchers. To avoid having to travel to each device and manually download the data, Nordin had to improvise. He turned to LoRa, a long range wireless transmission protocol that uses relatively little power and bandwidth compared with Wi-Fi and can work over tens of kilometers.

Other tinyML researchers are eyeing the same system to send data from tinyML devices used in agriculture to farmers' computers or phones. Without a way to download the data remotely, "the farmer will have to go through all these devices; it is time consuming, and it is tedious," says James Adeola, a Ph.D. student in computer science at the University of Abomey-Calavi who is developing tinyML devices that can detect diseases in cotton and reduce the need for pesticide use. Farmers "will be very happy if this solution can be implemented," he says.

SIMPLE AS TINYML DEVICES appear, developing them can be a challenge. For one, it requires expertise in multiple skill sets. "It's combining hardware, software, and machine learning," Basikolo says. "Very few people can do all these, so combining all these skills also takes time."

Researchers are trying to disseminate that expertise by running tinyML courses and workshops in Morocco, Brazil, Nigeria, South Africa, Rwanda, Malaysia, and other countries in the Global South. In 2021, for example, Harvard University and ICTP launched the tinyML for development academic network, which now encompasses 50 academic institutions across the Global South. The organizers, including Zennaro and Vijay Janapa Reddi, a computer scientist at Harvard University, began by donating tinyML kits to partner institutions. "When we started this initiative, we saw that the main issue was getting the hardware in the hands of people," because low-cost in the United States can still be expensive elsewhere, Zennaro says.

The resulting kits have been used to train students at universities in Malaysia, Saudi Arabia, and elsewhere. In just an hour or two, students were able to train tinyML devices to recognize words and phrases in their own local languages, "which is … empowering for communities," Zennaro says.

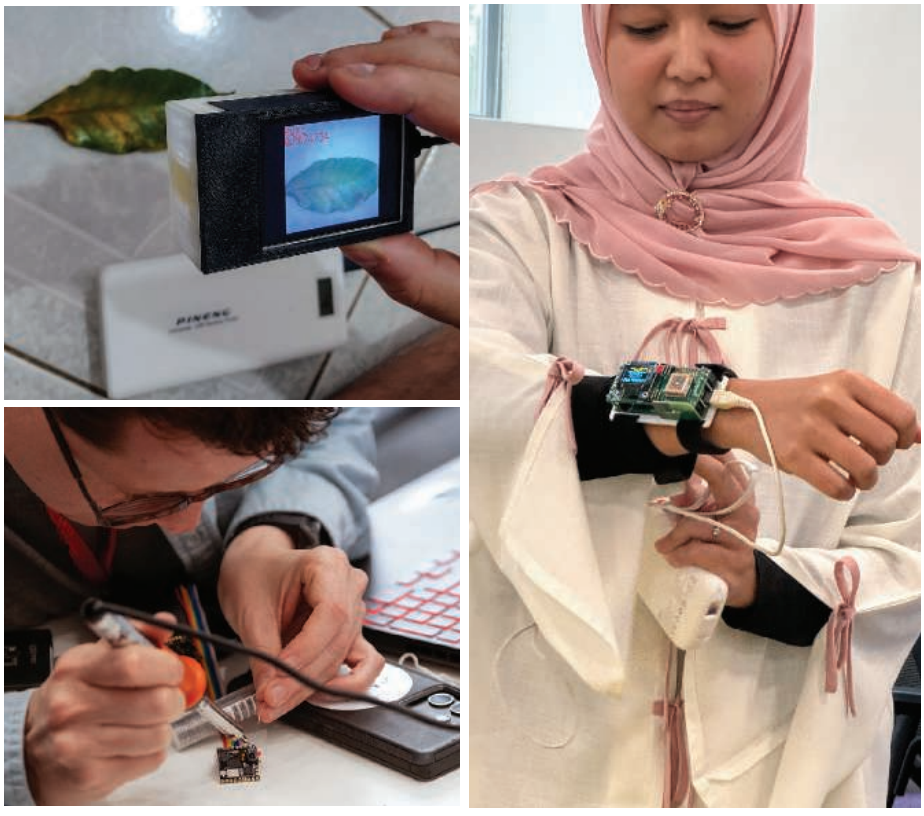
By nature, tinyML's capabilities are limited. "I think tinyML is very good at tiny problems," Yamashita says. The microcontrollers that run the devices have extremely limited memory and computing power, which makes them more suited for highly specific tasks than as generalizable commercial products. A device could be great at identifying words in one language or identifying diseases of one plant species, but it's less likely to serve as a universal translator or identify diseases across all plants.

Yet the appeal of smaller models that run directly on devices is dawning on many tech companies in the Global North, including Apple and Microsoft. For some applications, such as taking an order at a McDonald's, specialized tinyML models that can run on local devices rather than in the cloud, at lower cost and power use, may be preferable to running an expensive general LLM. "Sometimes these large language models are doing way more than you might need," says Jean Louis, a computer science Ph.D. student at the University of Florida; they might be the equivalent of using a sledgehammer to crack a nut.

Meanwhile, tinyML devices themselves are rapidly becoming more powerful. Just a few years ago it was considered "ridiculous" to be able to run advanced machine learning models on microcontrollers, says Peter Ing, an electronics engineer and member of the TinyML Open Education Initiative. Warden has already gotten a simple LLM to run on a device that is only slightly more expensive and power hungry than tinyML devices. He envisions more large AI models migrating to these smaller, power-efficient devices rather than relying solely on data centers.

At the same time, the simplest tinyML devices are likely to become more prevalent as microcontrollers continue to become cheaper and more powerful, with some already being developed specifically to run AI. "It's just reached a maturity point where we are now seeing solutions and technology that can be commercialized," says Reddi, who runs a free, massively online course on tinyML and has written an open-source book about it. And even though each tinyML device may be relatively small and specialized, many such devices talking to each other could help tackle bigger and more complex tasks.

As Yamashita puts it, "[TinyML] will enable AI to, in fact, go everywhere." ■



Tiny machine learning (tinyML) projects around the world aim to tackle challenges in the Global South, including identifying diseases such as rust in coffee leaves in Brazil (top left). TinyML devices to detect mosquitoes in the field (bottom left) and heart rate and other physiological readings (right) are also being developed in workshops.

PHOTOS: (CLOCKWISE FROM TOP LEFT) JOÃO YAMASHITA; SERGE STINCKWICH; RODRIGO NEUMANN

Sandeep Ravindran is a science journalist near Washington, D.C.