# FORSCHUNG KOMPAKT

**Next Generation Computing**

## Energieeffiziente KI-Chips für die Erkennung von Vorhofflimmern

**KI-Systeme können die Gesundheitsversorgung verbessern, Heilungschancen für Patienten erhöhen und Ärzte bei ihren Diagnosen unterstützen. Die Crux: Künstliche Intelligenz verbraucht enorm viel Strom. Die Fraunhofer-Institute für Integrierte Schaltungen IIS und für Techno- und Wirtschaftsmathematik ITWM haben Lösungen für energiesparsame KI-Chips entwickelt, die künftig dabei helfen können, Vorhofflimmern – eine spezielle Herzrhythmusstörung – frühzeitig zu erkennen. Für ihre Ideen wurden die beiden Institute im Pilotinnovationswettbewerb »Energieeffiziente KI-Systeme« des Bundesministeriums für Bildung und Forschung BMBF mit einem 1. Platz ausgezeichnet.**

Vorhofflimmern ist eine der häufigsten Herzrhythmusstörungen. Wird die Erkrankung nicht rechtzeitig erkannt, kann sie einen Schlaganfall auslösen. Eine Möglichkeit, EKGs über einen langen Zeitraum aufzuzeichnen und so die Chance zu erhöhen, das Herzstolpern zu erkennen, bieten Wearables wie etwa Smartwatches, die der Patient am Handgelenk trägt. Doch damit die mobile Diagnose praktisch umsetzbar ist, müssen die aufgezeichneten EKG-Daten energieeffizient ausgewertet werden können. Das Problem: Die Algorithmen zur Auswertung der Patientendaten können sehr rechenintensiv sein, was einen hohen Energieverbrauch zur Folge hat. Die Laufzeit und damit die Zuverlässigkeit eines mobilen Systems ist aber von dessen Energieverbrauch abhängig. Für mobile Anwendungen hat deswegen die energieeffiziente Ausführung der Auswertungsalgorithmen auf der Hardware höchste Priorität.

### Erster Platz für zwei Fraunhofer-Institute

Dies war Anlass für das Bundesministerium für Bildung und Forschung BMBF, den Pilotinnovationswettbewerb »Energieeffiziente KI-Systeme« zu starten. Nur wenn der Energieverbrauch heutiger Mikroelektronik gesenkt wird, schafft Künstliche Intelligenz (KI) Nutzen und hält Einzug in medizinische, industrielle und andere Anwendungen. Die Wettbewerbsaufgabe lautete, dass der KI-Chip mit einer Genauigkeit von mindestens 90 Prozent Vorhofflimmern erkennen

**Kontakt**
**Janis Eitner** | Fraunhofer-Gesellschaft, München | Kommunikation | Telefon +49 89 1205-1333 | presse@zv.fraunhofer.de
**Thoralf Dietz** | Fraunhofer-Institut für Integrierte Schaltungen IIS | Telefon +49 9131 776-1630 | Am Wolfsmantel 33 | 91058 Erlangen | www.iis.fraunhofer.de | thoralf.dietz@iis.fraunhofer.de
**Ilka Blauth** | Fraunhofer-Institut für Techno- und Wirtschaftsmathematik ITWM | Telefon +49 631 31600-4674 | Fraunhofer Platz 1 | 67663 Kaiserslautern | www.itwm.fraunhofer.de | ilka.blauth@itwm.fraunhofer.de

soll, dies in Echtzeit klassifiziert und dabei so wenig wie möglich Energie verbraucht. Die Anzahl der Fehlalarme darf 20 Prozent nicht überschreiten. Für die Umsetzung der Aufgabe bekamen die teilnehmenden Teams 16 000 einzelne EKG-Aufnahmen von je zwei Minuten Länge von der Berliner Charité gestellt. 8000 der Aufnahmen waren von Patienten mit Vorhofflimmern, die restlichen 8000 von Gesunden. Sowohl das Fraunhofer IIS als auch das Fraunhofer ITWM belegten den ersten Platz, wenn auch in unterschiedlichen Kategorien und mit unterschiedlichen Ansätzen. Mit ihrem Preis konnten die Institute unter Beweis stellen, dass Fraunhofer beim Einsatz von KI und in der Mikroelektronik ganz vorne mitspielt in Deutschland.

Das Team des Fraunhofer IIS, geleitet von Dr. Marco Breiling, gewann gemeinsam mit den Forschenden der Friedrich-Alexander-Universität Erlangen-Nürnberg um Dr. Marc Reichenbach und Prof. Dietmar Fey in der Kategorie ASIC 130 Nanometer (englisch: Application-Specific Integrated Circuit, anwendungsspezifischer integrierter Schaltkreis) mit dem Projekt »Low-Power Low Memory Low-Cost EKG-Signalanalyse mit ML-Algorithmen – Lo3-ML«. Das Fraunhofer ITWM gewann in Zusammenarbeit mit der Technischen Universität Kaiserslautern in der Kategorie FPGA (englisch: Field Programmable Gate Array, ein programmierbarer Logik-Schaltkreis) mit dem Projekt: »Holistischer Ansatz zur Optimierung von FPGA Architekturen für tiefe neuronale Netze via AutoML – Automatisches Maschinenlernen (HALF)«. Mit ihren Siegen haben die beiden Fraunhofer-Institute die Chance erhalten, ihre Schaltungen und Tools mit jeweils einer Million Euro weiterzuentwickeln.

## Projekt Lo3-ML – Signalverarbeitung fällt in Schlafmodus

Um zu erkennen, ob der Patient gesund oder krank ist, setzen die Forscher am Fraunhofer IIS in Lo3-ML auf Deep Learning, einer speziellen Methode des Maschinellen Lernens, die Neuronale Netze mit verschiedenen Eingabe- und Ausgabeschichten – auch als Layer bezeichnet – verwendet. Das digital vorliegende EKG-Signal wird in das Neuronale Netz eingegeben, die Abschnitte des Signals werden gefiltert, die einzelnen Signalanteile gewichtet und in mehreren Schichten aufsummiert. Die IIS-Forscher sprechen auch von einem ternären Neuronalen Netz, da die einzelnen Werte der Zeitreihe mit den ternären Gewichtswerten +1, 0 und -1 gewichtet werden. »In der ersten Schicht des Neuronalen Netzes wird ein gewisses Signalverhalten erkannt, in der zweiten Schicht werden die Merkmale in Beziehung zueinander gesetzt. Insgesamt kommen sechs Schichten zum Einsatz. Erst in der letzten sechsten Schicht entsteht ein komplexes Bild des EKG-Signals, das eine vorliegende Erkrankung anzeigt«, erläutert Marco Breiling, Wissenschaftler am Fraunhofer IIS. Mit einem Trick gelang es

dem Forscherteam um Breiling diese Zeitreihensignale, also die digitale Darstellung des EKG-Signals, besonders energieeffizient [BM1]zu verarbeiten: Die Signalverarbeitung als ein Teil des KI-Chips wird schlafen gelegt, solange sie nicht benötigt wird. Dadurch lässt sich 95 Prozent der Energie einsparen. »Der Chip sammelt 12,7 Sekunden lang das EKG-Signal ein und verarbeitet es dann innerhalb von nur 24 Millisekunden, sprich in 0,2 Prozent der Zeit. Die Verarbeitung schläft also über 99,8 Prozent der Zeit und beansprucht dabei kaum Energie. Dank nicht-flüchtiger RRAM-Speicher, die Bestandteil des Systems sind, kann die Signalverarbeitung gleich nach dem Aufwecken nach fast 12,7 Sekunden ohne Energieverbrauch wieder aufgenommen werden«, erläutert Breiling die Funktionsweise. Die RRAMs speichern die ternären Gewichte besonders effizient.

Zur erheblichen Energieeinsparung tragen darüber hinaus systolische Arrays bei – eine spezielle Architektur des Chips. »Für den permanenten Betrieb benötigt unser Chip eine derart geringe Leistung, dass eine Solarzelle mit einer Fläche von 6 mm x 6 $mm^2$ ausreichen würde, die bei Mondschein betrieben wird. Alternativ könnte der Chip mit der allerkleinsten am Markt verfügbaren Knopfzelle 330 Tage in Folge EKGs auswerten«, sagt der Forscher. Die entwickelte Schaltung eignet sich nicht nur für den medizinischen Einsatz, sondern auch für andere Anwendungen, bei denen Zeitreihensignale verarbeitet werden, wie Condition Monitoring und Predictive Maintenance.

## Projekt HALF – kurz für Holistisches AutoML für FPGAs

Das Forscherteam des Fraunhofer ITWM in Kaiserslautern berücksichtigt im Wettbewerb sowohl den Energieverbrauch der Hardware als auch die neuronale Netzwerktopologie. Im Vorfeld steht die Überlegung, welches Netzdesign die besten Voraussetzungen für die Aufgabe bietet. Für den Wettbewerb muss das Neuronale Netz nicht nur den Faktor Performance berücksichtigen, sondern um den Faktor Energieeffizienz ergänzt werden. Energieeffizienz könnte derart beschrieben werden, dass nur die minimal nötige Anzahl an Rechenoperationen aufgebracht werden soll, um das Vorhofflimmern zu erkennen.

## KI-Modell entscheidet über Energieverbrauch der Hardware

Doch wie findet man genau die Netze, die den definierten Ansprüchen und Vorgaben entsprechen? »Hier gibt es verschiedene Suchstrategien, wobei wir einen evolutionären Ansatz verwenden. Wir starten mit zehn verschiedenen zufällig gewählten Netzen, trainieren sie und prüfen, wie gut sie funktionieren. Anschließend wählen wir die beiden besten Netze aus und mutieren sie, sodass

zehn neue Netzvarianten entstehen. Diesen Vorgang wiederholen wir so oft, bis wir das beste Netz gefunden haben. Dieses Verfahren bezeichnet man als automatisiertes Maschinelles Lernen«, erläutert Dr. Jens Krüger, der am Fraunhofer ITWM im Competence Center – High Performance Computing forscht und das Projekt gemeinsam mit Prof. Dr.-Ing. Norbert Wehn von der TU Kaiserslautern geleitet hat. Dieses als automatisiertes Maschinelles Lernen bezeichnete Verfahren erweitern die Forschenden um einen holistischen Ansatz, der nicht nur das Neuronale Netz, sondern auch die Hardware betrachtet, da das KI-Modell den Energieverbrauch der Hardware beeinflusst.

Krüger und sein Team verwenden programmierbare Chips, FPGAs (Field Programmable Gate Arrays), in denen die Neuronalen Netze abgebildet werden und mit denen eine Vielzahl von Schaltungen realisiert und die bestmögliche Ausführung eines optimalen Algorithmus erzielt werden kann. Das FPGA lässt sich beliebig oft neu programmieren und zeichnet sich durch verschiedene Eigenschaften aus, die bei der Suche nach dem optimalen Neuronalen Netz betrachtet werden. »Insofern spiegelt der Projektname HALF – Holistisches AutoML für FPGAs – den Kernaspekt unseres Ansatzes wider«, sagt der Forscher. Mit einem an der TU Kaiserslautern entwickelten Software-Tool wird das Neuronale Netz auf das FPGA übertragen und ist dann in der Lage, die EKG-Daten automatisch auszuwerten. Durch diese Vorgehensweise ist eine neue vereinheitlichende Methodik entstanden, die nicht nur energieeffizienter als bislang ist, sondern auch eine Reduzierung der Entwicklungszeit für optimale neuronale Netzwerktopologien und entsprechende FPGA-Implementierungen ermöglicht. Die entwickelten Softwarewerkzeuge eignen sich nicht nur für FPGAs, sondern für verschiedenste Chips und Umgebungen.


**Abb. 1 Für ihre Forschungsarbeiten im Pilotinnovationswettbewerb »Energieeffizientes KI-System« wurden das Fraunhofer ITWM und das Fraunhofer IIS mit einem ersten Platz ausgezeichnet.**

**© Fraunhofer**

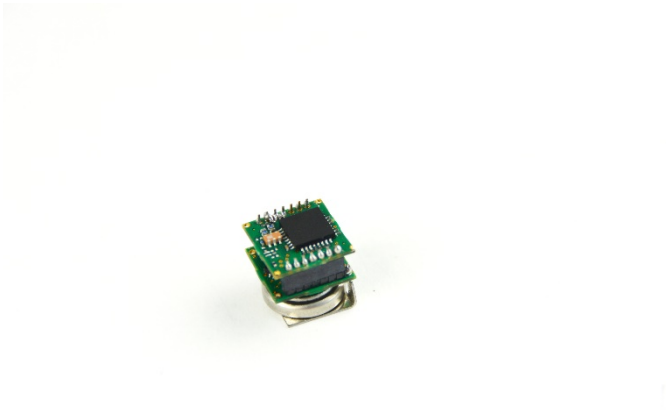

**Abb. 2 Ultra-low-power KI in der Edge.**

**© Fraunhofer IIS**