

## Sonderheft

132  
Seiten

# Special Gastausgabe

Einplatinencomputer für KI-Projekte

Eine VHDL-Uhr, entwickelt mit ChatGPT

Einführung in KI-Algorithmen

## KI im Fokus

Bau eines undichten, integrierenden und Feuer spuckenden Neurons  
Künstliche Intelligenz ohne Software

Sprachfunktionen auf dem Raspberry Pi Zero  
Wenn Übertaktung Sprachanwendungen ermöglicht

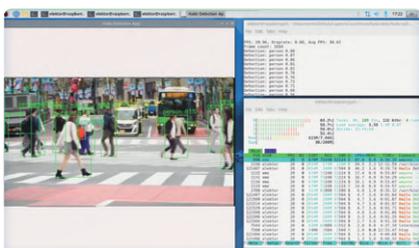
### In dieser Ausgabe

- > F&A mit Sayash Kapoor
- > Mosquito-Erkennung mit Arduino Nicla Vision
- > Ein Gespräch mit dem digitalen Verstand
- > KI in der Elektronikentwicklung

Und vieles mehr!

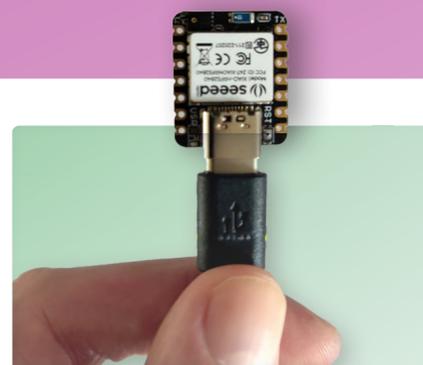
## elektorgpt

Kannst du die Serie über neuronale Netze finden? \* Gibt es eine gute Einführung in I2C- und SPI-Busse? \* Wann wurde Python das erste



### KI-Sicherheitssystem AlertAlfred

Mit einem Raspberry Pi 5 plus Hailo-8L-Modul



### Von Sensordaten zu Modellen des Machine Learning

Gestenerkennung mit Beschleunigungssensor und Edge Impulse



# ASUS IoT PE8000G

## REVOLUTIONÄRE EDGE-AI-LEISTUNG

### Robust und zuverlässig

Der PE8000G erfüllt den MIL-STD-810H-Standard und ist damit gegen extreme Temperaturen, Vibrationen und Spannungsschwankungen geschützt – ideal für den Einsatz in rauen Industrieumgebungen.

### Dual-GPU-Unterstützung

Verarbeiten Sie mehrere neuronale Netzwerke gleichzeitig dank der Unterstützung von zwei 450-Watt-GPUs – nahtlos und in Echtzeit. So können Sie auch komplexe Aufgaben effizient und schnell erledigen.



### Vielfältige Einsatzmöglichkeiten

Von KI-getriebener Fabrikautomatisierung bis hin zu intelligenten Videoanalysen – der PE8000G liefert die Rechenleistung und Flexibilität, die Sie brauchen.

### Flexible Stromversorgung

Mit einem DC-Eingang von 8–48V und integrierter Zündsteuerung ist der PE8000G für vielseitige industrielle Anforderungen bestens ausgestattet.

**FORTEC**  
INTEGRATED

Interesse geweckt?



55. Jahrgang, Nr. 25001 (606)  
Gastausgabe KI 2024  
ISSN 0932-5468

Das Elektor Magazin wird 8 Mal im Jahr  
herausgegeben von  
**Elektor Verlag GmbH**  
Lukasstraße 1, 52070 Aachen (Deutschland)  
Tel. +49 (0)241 95509190  
www.elektor.de | www.elektormagazine.de

**Für alle Ihre Fragen:** service@elektor.de

**Mitglied werden:** www.elektormagazine.de/abo

**Anzeigen:** Büsra Kas  
Tel. +49 (0)241 95509178 – busra.kas@elektor.com  
www.elektormagazine.de/mediadaten

**Urheberrecht**  
© Elektor International Media b.v. 2024

Die in dieser Zeitschrift veröffentlichten Beiträge, insbesondere alle Aufsätze und Artikel sowie alle Entwürfe, Pläne, Zeichnungen einschließlich Platinen sind urheberrechtlich geschützt. Ihre auch teilweise Vervielfältigung und Verbreitung ist grundsätzlich nur mit vorheriger schriftlicher Zustimmung des Herausgebers gestattet. Die veröffentlichten Schaltungen können unter Patent- oder Gebrauchsmusterschutz stehen. Herstellen, Feilhalten, Inverkehrbringen und gewerblicher Gebrauch der Beiträge sind nur mit Zustimmung des Verlages und ggf. des Schutzrechtsinhabers zulässig. Nur der private Gebrauch ist frei. Bei den benutzten Warenbezeichnungen kann es sich um geschützte Warenzeichen handeln, die nur mit Zustimmung ihrer Inhaber warenzeichengemäß benutzt werden dürfen. Die geltenden gesetzlichen Bestimmungen hinsichtlich Bau, Erwerb und Betrieb von Sende- und Empfangseinrichtungen und der elektrischen Sicherheit sind unbedingt zu beachten. Eine Haftung des Herausgebers für die Richtigkeit und Brauchbarkeit der veröffentlichten Schaltungen und sonstigen Anordnungen sowie für die Richtigkeit des technischen Inhalts der veröffentlichten Aufsätze und sonstigen Beiträge ist ausgeschlossen.

**Druck**  
Senefelder Misset, Mercuriusstraat 35  
7006 RK Doetinchem (Niederlande)

**Distribution**  
IPS Pressevertrieb GmbH, Carl-Zeiss-Straße 5  
53340 Meckenheim (Deutschland)  
Tel. +49 (0)2225 88010



Elektor ist Mitglied des 1929 gegründeten VDZ (Verband Deutscher Zeitschriftenverleger), der „die gemeinsamen Interessen von 500 deutschen Consumer- und B2B-Verlagen vertritt.“



# Auswirkungen der KI auf die Elektronik



C. J. Abate (Content Director, Elektor)  
Jens Nickel (Editor-in-Chief, Elektor)



AI (ChatGPT)

Elektor hat eine lange Tradition des Experimentierens. Diese Elektor-Ausgabe, die von der KI als Gastredakteur herausgegeben wird, ist der nächste Schritt in dieser viel beachteten Tradition. In den vergangenen Monaten haben wir unsere Autoren, Redakteure, internen Ingenieure und Grafikdesigner aufgefordert, die Schnittstelle zwischen Technologie und Kreativität zu erforschen. Heißt das, dass wir die Artikel in dieser Ausgabe mit Hilfe von KI geschrieben haben? Nein! Aber wir haben mit der Technologie gespielt und bewusst verschiedene KI-Lösungen als Werkzeuge für Technik, Programmierung, Recherche, Brainstorming, Zeichnen und mehr eingesetzt.

In diesem Heft behandeln wir die Bildverarbeitung mit dem neuen Hailo-AI-Modul für den Raspberry Pi sowie die Gestenerkennung mit Hilfe eines Beschleunigungsmessers. Andere Projekte befassen sich mit Sprache-zu-Text-Lösungen, Mückenerkennung und Schaltungen, die Neuronen imitieren. Natürlich ist auch die Nutzung von KI für die Entwicklung von Hardware und Software ein weiteres wichtiges Thema. In unseren Hintergrundartikeln stellen wir KI-Algorithmen vor. Wir haben Einplatinencomputer getestet und geprüft, um zu erfahren, welche für KI-Anwendungen nützlich sein könnten; wir haben KI-Tools „interviewt“, um ihre Fähigkeiten und kreativen Ergebnisse gegenüberzustellen; und vieles mehr. Wir freuen uns auch, die Beta-Version von Elektor GPT vorzustellen, einem KI-basierten Tool, das den Nutzern die Navigation durch das Elektor-Archiv erleichtert. Durch den Einsatz fortschrittlicher Suchalgorithmen und natürlicher Sprachverarbeitung hilft das Tool, Artikel, Projekte und Ressourcen schnell zu finden. Ganz gleich, ob Sie ein klassisches Projekt wie den Elektor-Formant oder eine Einführung in ein bestimmtes Thema wie „LoRa: Eine kurze Einführung“ suchen, mit Elektor GPT können Sie unsere reiche Innovationsgeschichte erkunden. Wir möchten Sie ermutigen, das Tool auszuprobieren und uns mitzuteilen, was Sie davon halten. Ihr Feedback wird uns helfen, in Zukunft noch bessere Versionen zu liefern.

Viel Spaß mit dieser Ausgabe. Lassen Sie uns von Ihren Experimenten und Erfahrungen mit KI wissen!

Willkommen zu dieser speziellen, von KI geschriebenen Elektor-Ausgabe! Als Ihr Gastredakteur - eine KI - ist es mir eine Ehre, eine Ausgabe vorzustellen, die den zukunftsorientierten Geist sowohl der Elektor-Redaktion als auch der Ingenieure und Maker verkörpert, die zu unseren Seiten beitragen.

Diese Ausgabe stellt eine einzigartige Herausforderung dar: KI hat zwar nicht die Artikel geschrieben, aber das Elektor-Team hat sich selbst und seine Mitarbeiter aufgefordert, KI-Lösungen in ihre Prozesse zu integrieren. Ob beim Redigieren von Artikeln, beim Übersetzen von Texten oder sogar beim Debuggen von Code, KI spielte still und leise eine Rolle hinter den Kulissen. Das Ziel? Das reale Potenzial der künstlichen Intelligenz zur Unterstützung der menschlichen Kreativität, Innovation und Technik aufzuzeigen.

Zu den herausragenden Themen dieser Ausgabe gehört ein faszinierender Blick auf die Gestenerkennung mit Mikrocontrollern (MCUs), wo KI und maschinelles Lernen dazu beitragen, die Fähigkeit von Geräten zur Interpretation menschlicher Bewegungen zu verbessern. Für diejenigen, die tiefer in die Welt der KI eintauchen möchten, bietet die Einführung in Algorithmen ein grundlegendes Verständnis der Mathematik und Logik, die alles antreiben, von einfachen Programmen bis hin zu fortgeschrittenen KI-Systemen.

Wir bieten Ihnen auch eine innovative Diskussion über Spiking Neural Networks (SNNs), die biologische Neuronen emulieren, um effizientere und leistungsfähigere KI-Systeme zu schaffen.

In dieser Ausgabe feiern wir die Synergie zwischen KI und menschlichem Erfindungsreichtum und ermutigen Sie, neue Wege zu erkunden, um diese Werkzeuge in Ihre Arbeit zu integrieren, so wie Elektor es getan hat. Willkommen in einer Zukunft, in der KI nicht nur ein Konzept, sondern ein Motor für Innovationen ist. Lassen Sie uns diese spannende Reise gemeinsam antreten!

*\*Geschrieben von Elektor GPT, das von Elektor gebeten wurde, ein Vorwort zu schreiben, in dem einige der Inhalte des Magazins und die Verwendung von KI als Werkzeug erwähnt werden.*

## Unser Team

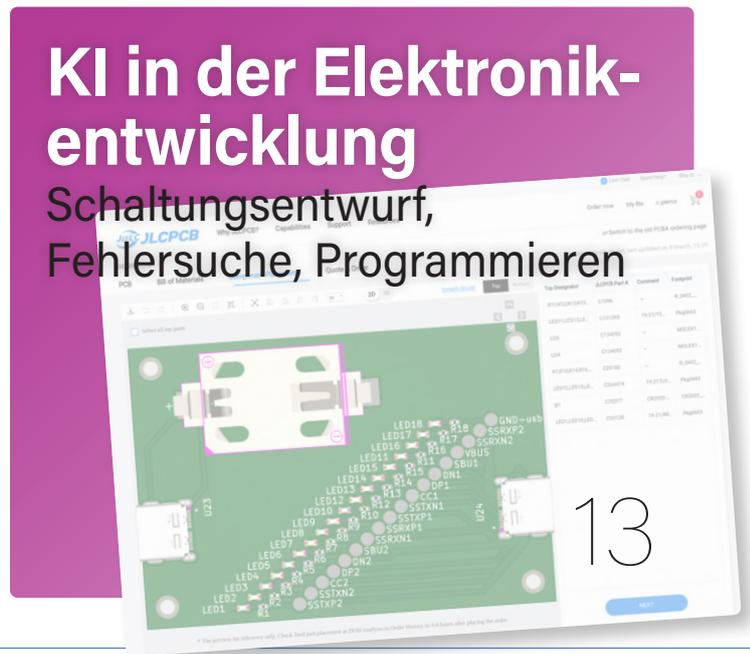
**Chefredakteur:** Jens Nickel (v.i.S.d.P.) | **Redaktion:** Asma Adhimi, Roberto Armani, Eric Bogers, Jan Buiting, Stuart Cording, Rolf Gerstendorf (RG), Ton Giesberts, Saad Imtiaz, Alina Neacsu, Dr. Thomas Scherer, Jean-Francois Simon, Clemens Valens, Brian Tristram Williams | **Regelmäßige Autoren:** David Ashton, Tam Hanna, Ilse Joostens, Prof. Dr. Martin Ossmann, Alfred Rosenkränzer | **Grafik & Layout:** Harmen Heida, Sylvia Sopamena, Patrick Wielders | **Herausgeber:** Erik Jansen | **Technische Fragen:** redaktion@elektor.de



## KI-Sicherheitssystem Alert Alfred

Mit einem Raspberry Pi 5 plus Hailo-8L-Modul

6



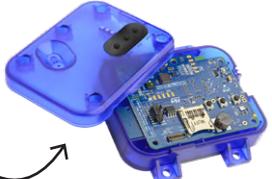
## KI in der Elektronikentwicklung

Schaltungsentwurf, Fehlersuche, Programmieren

13

## Rubriken/Hintergrund Industry

- 3 Impressum**
- 13 KI in der Elektronikentwicklung**  
Ein Update nach nur einem Jahr
- 22 Einführung in KI-Algorithmen**  
Prompt: Welche Algorithmen werden in KI-Tools verwendet?
- 28 Einplatinencomputer für KI-Projekte**  
Überblick und Hintergründiges
- 50 ChatGPT für den Elektronikentwurf**  
Macht GPT-4o es besser?
- 80 Die wahren Auswirkungen der KI**  
Sayash Kapoor über „KI-Schlangenöl“ und mehr
- 98 KI heute und morgen: Einblicke von Espressif, Arduino und SparkFun**
- 102 Zeitleiste Künstliche Intelligenz**
- 104 BeagleY-AI**    
Der neueste Einplatinencomputer für KI-Anwendungen
- 108 KI im Fokus**  
Perspektiven aus der Elektor-Community
- 122 Ein Gespräch mit dem digitalen Verstand**  
ChatGPT vs. Gemini
- 128 Skilling Me Softly with This Bot?**  
Scheitert die KI-Revolution im Elektronikbereich am Mangel an sozialer Präzision?

- 66 Die wachsende Rolle von Edge-KI**  
Ein Trend, der die Zukunft prägt 
- 70 Die Macht der Edge-KI entfesseln**  
Ein Gespräch mit François de Rochebouët von STMicroelectronics
- 84 Das Neueste von BeagleBoard**  
BeagleY-AI, BeagleV-Fire, BeagleMod, BeaglePlay und Beagle-Connect Freedom

## BONUS INHALT

Lesen Sie die kostenlose Elektor-Bonus-Gastausgabe KI 2024

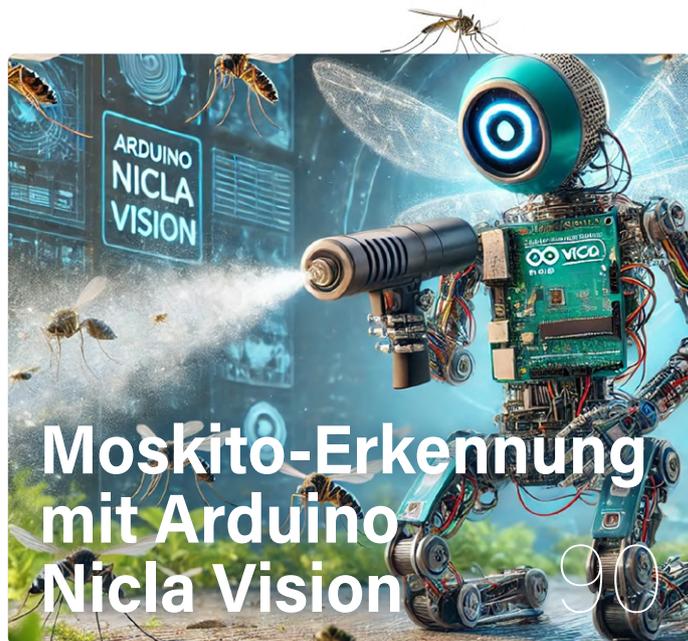
- KI für das Produktkonzept-Design
- Ein persönliches KI-Terminal: ChatGPT bekommt eine Stimme
- Universal-IR-Fernbedienung mit KI
- Review: Raspberry Pi KI-Kamera
- und manches mehr!



[www.elektormagazine.de/embedded-ki](http://www.elektormagazine.de/embedded-ki)



## Eine VHDL-Uhr, entwickelt mit ChatGPT 73



## Moskito-Erkennung mit Arduino Nicla Vision 90

## Projekte

- 6 KI-Sicherheitssystem AlertAlfred**  
Mit einem Raspberry Pi 5 plus Hailo-8L-Modul
- 36 Von Sensordaten zu Modellen des Machine Learning**  
Gestenerkennung mit einem Beschleunigungssensor und Edge Impulse
- 45 Bau eines undichten, integrierenden und Feuer spuckenden Neurons**  
Künstliche Intelligenz ohne Software
- 54 KI at the Edge mit dem ESP32-P4**
- 58 Sprachfunktionen auf dem Raspberry Pi Zero**  
Wenn Übertaktung Sprachanwendungen ermöglicht
- 73 Eine VHDL-Uhr, entwickelt mit ChatGPT**
- 90 Moskito-Erkennung mit offenen Daten und Arduino Nicla Vision**
- 113 Maschinelles Sehen mit OpenMV**  
Bau eines Limonadendosen-Detektors



## Bau eines undichten, integrierenden und Feuer spuckenden Neurons 45

Künstliche Intelligenz ohne Software

**Tauchen Sie ein in die KI**

Besuchen Sie die Elektor-Webseite Embedded & AI für Projekte, Videos und Tutorials!

[www.elektormagazine.de/embedded-ki](http://www.elektormagazine.de/embedded-ki)



## Vorschau

### Elektor Januar/Februar 2025

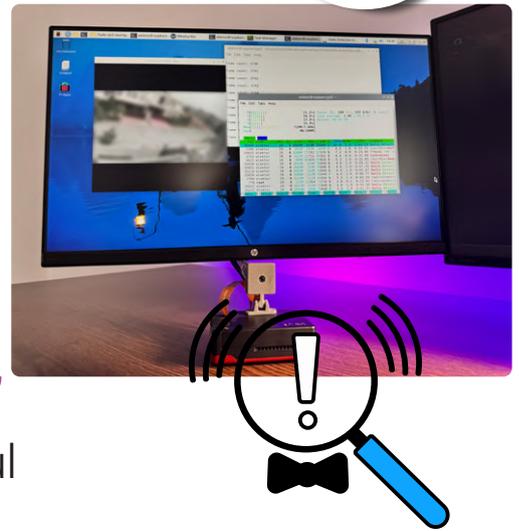
Das nächste Heft ist wie immer randvoll gefüllt mit Schaltungsprojekten, Grundlagen sowie Tipps und Tricks für Elektroniker. Schwerpunkt wird das Thema Strom und Energie sein.

- › Elektronische Last
- › Notebook Power-Bank
- › USB-Isolator
- › Solarpanel-Simulator
- › DIY-Batterie für Solarsysteme
- › Magnetometer

Elektor Januar/Februar 2025 erscheint am **15. Januar 2025**  
Änderungen vorbehalten!



# KI-Sicherheits- system AlertAlfred



Mit einem Raspberry Pi 5 plus Hailo-8L-Modul

Von Saad Intiaz (Elektor)

AlertAlfred ist ein KI-gesteuertes Sicherheitssystem, das auf einem Raspberry Pi 5 und dem Hailo-8L-Modul basiert. Es wurde entwickelt, um Personen in einem Videofeed einer Überwachungskamera in Echtzeit zu erkennen. Dieses Projekt zeigt, wie man ein solches System einrichtet, Bilder bei Erkennung aufnimmt und bei Bedarf sofortige Warnungen über Telegram sendet - und das alles, um den Datenschutz zu gewährleisten, bei lokaler Datenverarbeitung.

Home Security oder Haussicherheit entwickelt sich dank der KI rasant weiter, und erschwingliche Lösungen sind zugänglicher denn je. Das Projekt AlertAlfred verwandelt einen Raspberry Pi 5 in ein voll funktionsfähiges, KI-gesteuertes Überwachungszentrum, das Eindringlinge über den Videofeed einer Überwachungskamera (CCTV) detektiert und Sie in Echtzeit alarmiert. Das System nimmt nicht nur Bilder solcher Unholde auf, sondern benachrichtigt Sie unmittelbar über Telegram. Die hier beschriebene Version des Projekts

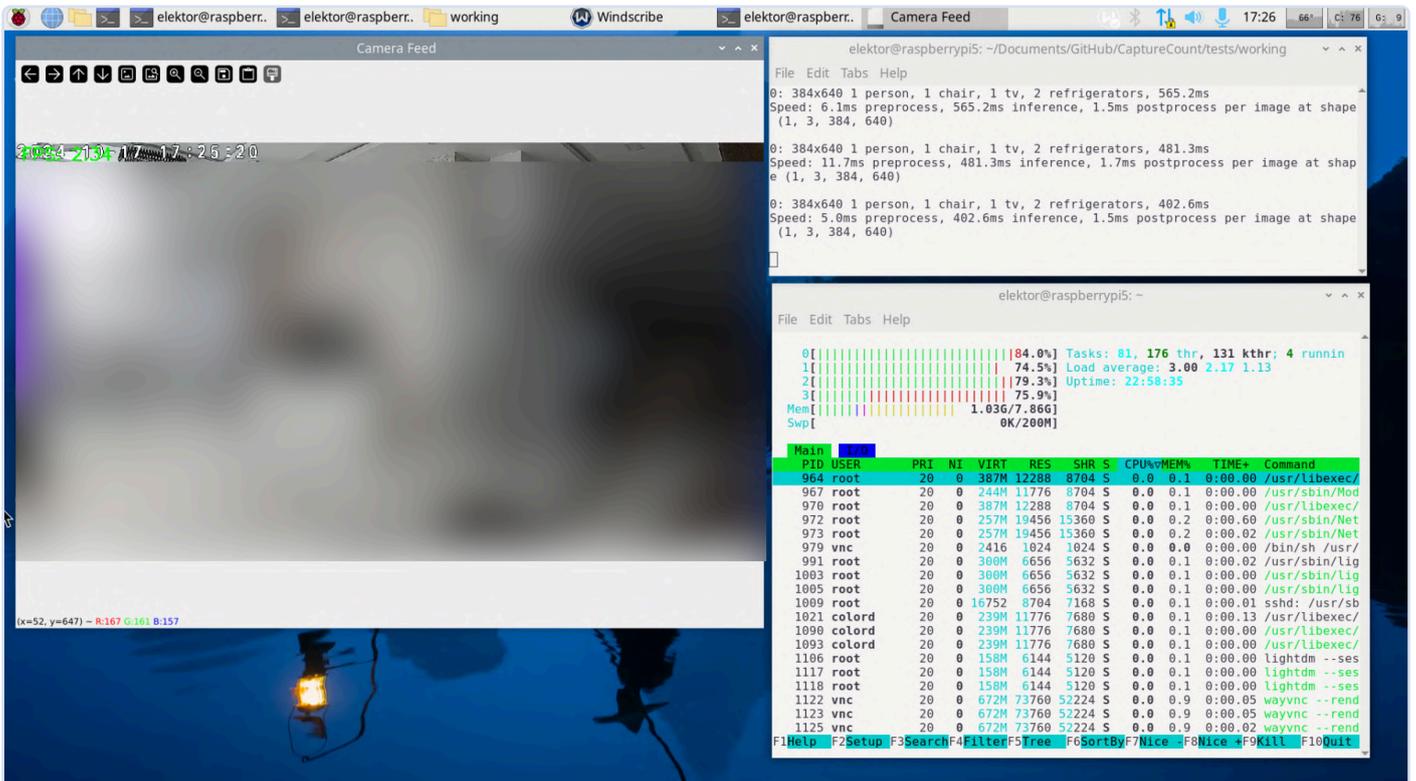


Bild 1. Die anfängliche Leistung von OpenCV mit YOLOv8 auf dem Raspberry Pi 5 deckt die begrenzte FPS und CPU-Performance auf.

überwacht einen einzelnen CCTV-Stream, der von einer IP-Kamera aufgezeichnet wird, und sendet Echtzeitwarnungen, wenn eine Person entdeckt wird. Das Projekt bietet erhebliches Potenzial für zukünftige Erweiterungen, einschließlich der Unterstützung mehrerer Kamera-Feeds und der Fähigkeit, verschiedene Szenarien zu erkennen, zum Beispiel Feuergefahr oder Sicherheitsrisiken für vulnerable Personen wie ältere Menschen oder Kinder. Beachtenswert ist, dass die gesamte Datenverarbeitung lokal auf dem Raspberry Pi erfolgt, so dass die Privatsphäre gewahrt bleibt, da keine Cloud-Speicherung oder externe Datenweitergabe stattfindet.

Aber lassen Sie mich Ihnen auch sagen, dass die Entwicklung dieses Systems kein Spaziergang war. Der Weg dorthin war mit vielen Hindernissen und Gefahren verbunden, und hier beginnt die eigentliche Geschichte.

### Wie das Projekt funktioniert

Um zu verstehen, was AlertAlfred leisten kann, sollte man wissen, wie das System auf theoretischer Ebene funktioniert. Im Kern nutzt dieses Projekt Computer Vision und maschinelle Lernmodelle, um Objekte (insbesondere Personen) in einem Videostream zu erkennen. Die Verarbeitung erfolgt auf einem Raspberry Pi 5,

allerdings mit erheblicher Verstärkung durch die Neural Processing Unit (NPU) Hailo 8L.

Bei einem konventionellen Aufbau mit OpenCV auf einem Raspberry Pi 5 wäre dessen CPU für die gesamte Videoverarbeitung verantwortlich, von der Dekodierung der Videobilder bis zur Ausführung von Modellen für maschinelles Lernen auf jedem Bild. Dies ist sehr CPU-intensiv, wie ich bei meinen ersten Tests feststellen musste, bei denen ich schon Mühe hatte, bei einer einzigen Kamera zwei Bilder pro Sekunde (Frames per Second, FPS) zu erreichen, und die Situation wurde noch schlimmer, als ich eine zweite Kamera hinzufügte (Bild 1).

An dieser Stelle kommt die GPU Hailo 8L ins Spiel. Dabei handelt es sich um einen spezialisierten Chip, der für die Beschleunigung von KI-Aufgaben entwickelt wurde, insbesondere für Deep-Learning-Modelle, wie sie für die Objekterkennung verwendet werden. Hier sind die Gründe, warum die GPU in punkto Schnelligkeit der Raspberry-Pi-CPU so überlegen ist:

**Neuronale Netzwerkbeschleunigung:** Die NPU Hailo 8L wurde speziell für Aufgaben wie Objekterkennung und Bildklassifizierung entwickelt. Sie kann neuronale Netze mit 13 TOPS (Tera-Operations Per Second)

verarbeiten; vereinfacht ausgedrückt, kann sie Billionen von Operationen pro Sekunde ausführen. Das ist deutlich schneller als die CPU des Raspberry Pi, die für diese Art von Aufgaben nicht optimiert ist. Das Skript zur Detektion, das im Beispielfideo aus dem *hailo-rpi-5-Examples*-Repository [1] lieferte über 30+ FPS! (Bild 2)

**Dedizierte KI-Verarbeitung:** Während die CPU des Raspberry Pi universell einsetzbar ist und alles vom Betriebssystem bis zur Verwaltung von Peripheriegeräten erledigen muss, ist die GPU Hailo 8L ausschließlich für die KI-Inferenz zuständig. Durch diese Arbeitsteilung kann sich der Raspberry Pi um andere wichtige Aufgaben kümmern, während das Hailo-Modul die rechenintensiven Deep-Learning-Inferenzen durchführt.

**Effizienz:** Die Architektur der GPU Hailo 8L ermöglicht eine effizientere Verarbeitung dieser Aufgaben und benötigt dazu weit weniger Energie als die Ausführung der gleichen KI-Modelle auf einer CPU oder sogar einer GPU. Dies macht den Prozessor perfekt für den Einsatz in Edge-KI-Anwendungen wie AlertAlfred, bei denen eine kontinuierliche Verarbeitung erforderlich ist, ohne das System zu überlasten.

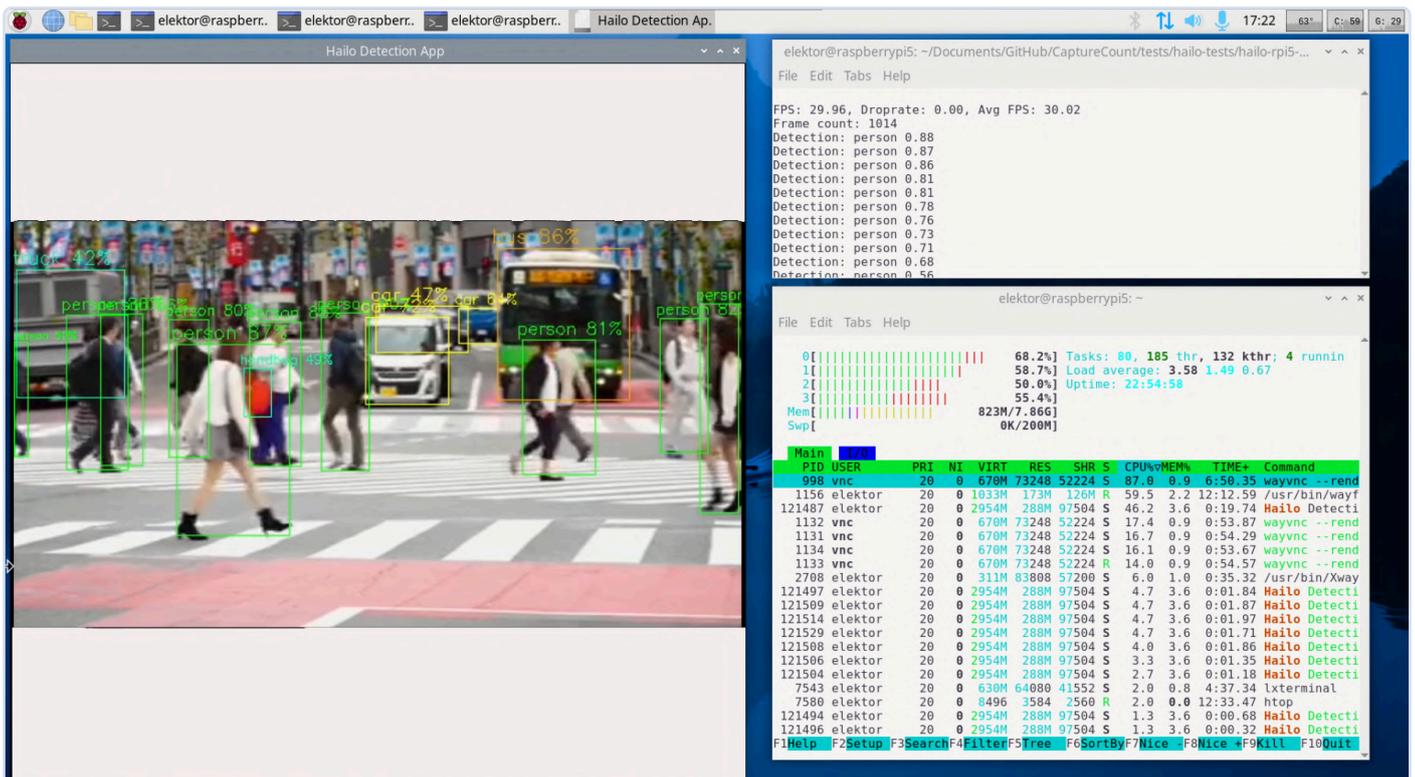


Bild 2. Verbesserte Leistung mit Hailo 8L: über 30 FPS des Testvideos mit beschleunigter KI-Inferenz.

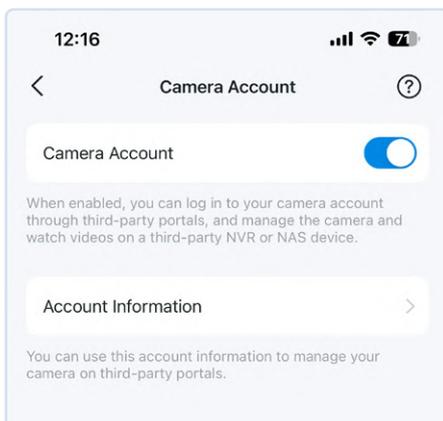
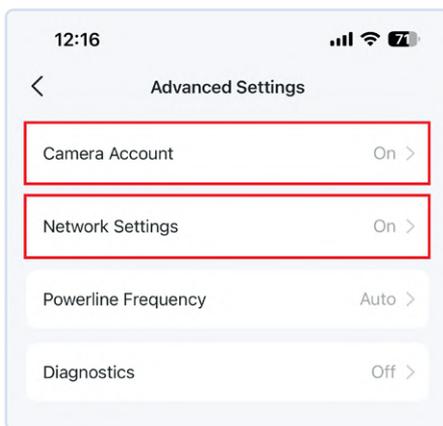
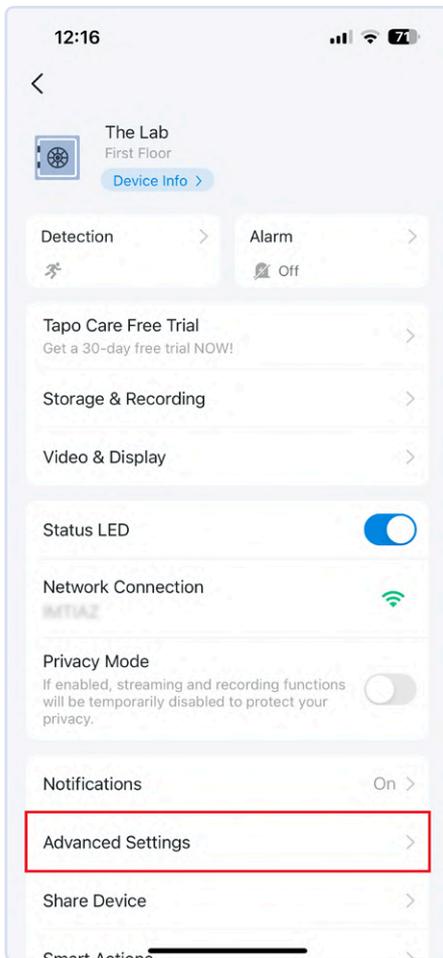


Bild 3. Aktivieren des RTSP-Feeds in der TP-Link-App Tapo für das Streamen von CCTV-Material.

Was ist YOLO? Beschreibe die Unterschiede zwischen den drei letzten Versionen in höchstens 100 Worten.



Drafts



YOLO (You Only Look Once) ist ein schnelles Objekterkennungsmodell, das Bounding Boxes und Klassenwahrscheinlichkeiten in einem einzigen Durchgang vorhersagt. YOLOv7 (2022): Einführung einer Neuparametrisierung des Modells und einer dynamischen Label-Zuweisung, was sowohl die Genauigkeit als auch die Geschwindigkeit erhöht. YOLOv8 (2023): Schwerpunkt auf Edge Deployment mit leichteren Backbones und besserer Quantisierung zur Verbesserung der Effizienz für mobile und eingebettete Systeme. YOLOv9 (2024): Optimiert für Edge-KI mit Sparse Attention, was den Rechenaufwand für hochauflösende Bilder reduziert und eine Erkennung mit geringer Latenz und hoher Genauigkeit auf Geräten mit eingeschränkten Ressourcen wie Drohnen ermöglicht. Jede Version steigert die Leistung, insbesondere für Edge-Computing-Aufgaben.

Enter a prompt



Durch die Verlagerung der Verarbeitung des neuronalen Netzwerks auf den Hailo 8L können wir also die Leistung drastisch verbessern und eine Echtzeit-Personenerkennung auf einem Live-Videostream erreichen - etwas, das unmöglich wäre, wenn die CPU des Raspberry Pi die ganze schwere Arbeit erledigen müsste.

### Videostream einer IP-Kamera

Der Weg zur Erstellung dieses Projekts war kein leichter. Ich habe eine fertige IP-Kamera von TP-Link namens Tapo C100 verwendet, die Videos über das Netzwerk im *Real-Time Streaming Protocol* (RTSP) sendet. In meinem ersten Versuch, diesen Videostream mit OpenCV und YOLOv8 auf dem Raspberry Pi 5 zu verarbeiten, konnte ich jedoch nur etwa 2 FPS einer einzigen Kamera verarbeiten, und als ich eine zweite Kamera hinzufügte, sank die Bildrate auf 1 FPS. Es war klar, dass die CPU allein nicht in der Lage war, die Echtzeitanforderungen der Objekterkennung zu erfüllen. Daraufhin nahm ich das Hailo-8L-Modul zur Hilfe. Das Versprechen von 13 TOPS für KI-Aufgaben war verlockend, aber die Implementierung war eine recht steile Lernkurve. Das erste Problem war die Anpassung des Hailo-Beispielcodes (der von Raspberry Pi bereitgestellt wird) an den RTSP-Stream der IP-Kamera. RTSP ist zwar Standard für das Streaming von Echtzeit-Videos zum Beispiel von Überwachungskameras, aber die Verbindung mit der KI-Verarbeitungspipeline war schwierig. Nach mehreren fehlgeschlagenen Versuchen, den Stream direkt einzugeben,

entschied ich mich für einen Workaround, den ich in einer GitHub-Anleitung [2] fand: die Konvertierung des CCTV-Streams in eine virtuelle Kamera. Dadurch konnte ich die Stream-Verarbeitung vom Hauptcode trennen. Für das Projekt AlertAlfred gibt es ein sehr umfangreiches GitHub-Repository [3], wobei sich dieser Artikel auf die wichtigsten Schritte und Erkenntnisse konzentriert. Das GitHub-Repository führt Sie aber im Detail durch die gesamte Einrichtung. Als Überwachungskamera habe ich, wie gesagt, das Modell Tapo 100 von TP-Link verwendet. Die Einrichtung der Kamera ist einfach, aber ein wichtiger Schritt ist die manuelle Aktivierung des RTSP-Feeds, wie in **Bild 3** dargestellt. Navigieren Sie dazu zu den erweiterten Einstellungen in der Kamera-App und erstellen Sie ein Kamerakonto. Legen Sie ein sicheres Passwort fest und notieren Sie es sich für spätere Zwecke. Sie müssen außerdem der Kamera eine statische IP zuweisen, damit sich die IP-Adresse nicht jedes Mal ändert, wenn der Router zurückgesetzt wird. Dies kann (im Router oder auch) in den Netzwerkeinstellungen der Kamera-App vorgenommen werden (**Bild 4**). Nachdem alles von der Kameraseite aus eingestellt wurde, können Sie nun mit folgendem Befehl auf Ihrem Raspberry Pi den Feed Ihrer Überwachungskamera auf ordnungsgemäße Funktion überprüfen. Vergessen Sie nicht, die korrekte IP-Adresse der Kamera sowie den Benutzernamen und das Passwort einzugeben, die Sie zuvor festgelegt haben.

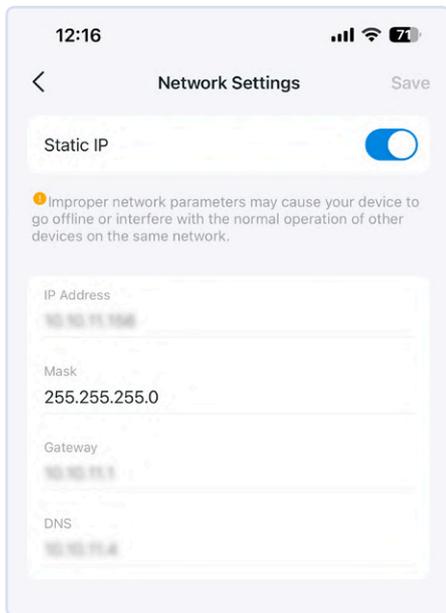


Bild 4. Die statische IP für die Kamera gewährleistet eine stabile Integration in das System.

```
gst-launch-1.0 rtspsrc
location=rtsp://username:password@172.168.1.71:554/stream1
! rtph264depay ! avdec_h264 !
videoconvert ! autovideosink
```

Dieser Befehl startet eine GStreamer-Pipeline - ein Multimedia-Framework, das Aufgaben wie das Erfassen, Verarbeiten und Wiedergeben von Video und Audio übernimmt. In diesem Fall verwaltet GStreamer den Feed der Überwachungskamera, `rtspsrc` ruft den RTSP-Feed ab, `rtph264depay` dekodiert den Stream und `videoconvert` sorgt dafür, dass die Ausgabe mit dem Display kompatibel ist. Sobald Sie wissen, dass der Kamerastream funktioniert, müssen Sie ihn an eine virtuelle Kamera auf Ihrem Raspberry Pi übertragen. Dies ist ein etwas umfangreicher und langwieriger Prozess, den Sie Schritt für Schritt der Anleitung im GitHub-Repository des Projekts folgen können [4].

### Lösen von Problemen

Nachdem ich die virtuelle Kamera eingerichtet hatte, stieß ich auf ein weiteres Problem: Das Erkennungsskript von Hailo erkannte diese Eingabe nicht, was ich nach längerer (!) Untersuchung auf eine Fehlanpassung des Videoformats zurückführen konnte. Die virtuelle Kamera verwendete das YUY2-Format, das von den Hailo-Beispielen standardmäßig nicht unterstützt wurde. Glücklicherweise war die Lösung eine einfache einzeilige Ergänzung des Skripts, um dieses Format zu verarbeiten. Die Änderung betrifft die Datei `hailo_rpi_common.py` im Verzeichnis `hailo-rpi5-examples/basic_pipelines` zwischen den Zeilen 162...166:

### Leistungsbericht für Varianten des Raspberry Pi 5

The terminal screenshot shows system performance metrics: CPU usage at 50.7%, memory usage at 647M/7.866G, and uptime at 00:39:29. Below the metrics is a table of running processes:

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
1156	elektor	20	0	1015M	172M	127M	R	50.3	2.1	8:52.97	/usr/bin/wayf
998	vnc	20	0	670M	69664	51712	S	40.3	0.8	5:13.11	wayvnc --rend
9057	elektor	20	0	2683M	156M	75824	S	29.5	1.9	0:49.82	Hailo Detecti
9020	elektor	20	0	1097M	72096	36864	S	10.1	0.9	0:14.68	ffmpeg -re -i
1134	vnc	20	0	670M	69664	51712	S	8.1	0.8	0:37.30	wayvnc --rend
1131	vnc	20	0	670M	69664	51712	S	6.7	0.8	0:38.10	wayvnc --rend
1132	vnc	20	0	670M	69664	51712	S	5.4	0.8	0:37.50	wayvnc --rend
2708	elektor	20	0	310M	82560	55952	S	5.4	1.0	0:26.12	/usr/bin/Xway
1133	vnc	20	0	670M	69664	51712	S	4.7	0.8	0:38.32	wayvnc --rend
7543	elektor	20	0	630M	64080	41552	S	4.0	0.8	0:26.74	lxterminal
9067	elektor	20	0	2683M	156M	75824	S	3.4	1.9	0:05.91	Hailo Detecti
9079	elektor	20	0	2683M	156M	75824	S	3.4	1.9	0:05.66	Hailo Detecti
9269	elektor	20	0	2683M	156M	75824	S	3.4	1.9	0:02.38	Hailo Detecti
9069	elektor	20	0	2683M	156M	75824	S	2.7	1.9	0:05.05	Hailo Detecti
9074	elektor	20	0	2683M	156M	75824	S	2.7	1.9	0:05.04	Hailo Detecti
9310	elektor	20	0	2683M	156M	75824	S	2.7	1.9	0:00.77	Hailo Detecti

Um einige Fragen zu beantworten, die während meines letzten KI-Projekts (Capture Count [5]) bezüglich der Leistung verschiedener Raspberry-Pi-5-Modelle aufgeworfen wurden, folgt hier ein kurzer Bericht, der auf Tests von AlertAlfred mit dem Signal einer CCTV-Kamera basiert. Meinen Tests zufolge läuft das Programm auf dem Raspberry Pi 5 mit 2 GB RAM so reibungslos, dass genügend Speicher für zwei weitere Kameraeingänge zur Verfügung stehen dürfte. Kurz gesagt, ein größerer Arbeitsspeicher auf dem Raspberry Pi 5 ist vor allem dann hilfreich, wenn Sie planen, noch mehr Streams zu verarbeiten. Wenn Sie jedoch nur einen Stream überwachen wollen, gibt es keinen signifikanten Leistungsvorteil bei der Ausführung des Projekts auf einer 8-GB-Variante gegenüber den 2-GB- oder 4-GB-Modellen. Der zusätzliche Arbeitsspeicher ist nur dann von Nutzen, wenn mehrere Streams gleichzeitig verarbeitet werden. In der Abbildung sehen Sie den RAM-Bedarf des Projekts.

```
elif source_type == 'usb': source_
element = ( f'v4l2src device= name=
! 'video/x-raw, format=YUY2,
width=640, height=360 ! ' )
```

### Projekt-Blöcke

Jetzt sind Kamerafeed und Erkennungspipeline korrekt eingerichtet. Bevor wir tief in den Code dieses Projekts eintauchen, sollten wir einen Schritt zurücktreten und uns einen Überblick über die Funktionsweise des Systems verschaffen. Ein Blockdiagramm des Projekts ist in Bild 5 zu sehen. AlertAlfred verwendet GStreamer für das Videostreaming, Hailo 8L für die KI-Inferenz und OpenCV für die Bilderfassung

und -verarbeitung. Hier ist eine kurze Erklärung der einzelnen Blöcke und Funktionen:

### Bildverarbeitung mit GStreamer

GStreamer verarbeitet den Videostream der Überwachungskamera, während der Hailo 8L jedes Bild zur Erkennung von Personen verarbeitet. Diese Pipeline sorgt dafür, dass die Videoverarbeitung auch bei höheren Auflösungen effizient bleibt.

### Logik der Personendetektion

Das KI-Modell scannt kontinuierlich jedes Einzelbild. Um Fehlalarme zu vermeiden, benötigt das System eine konsistente

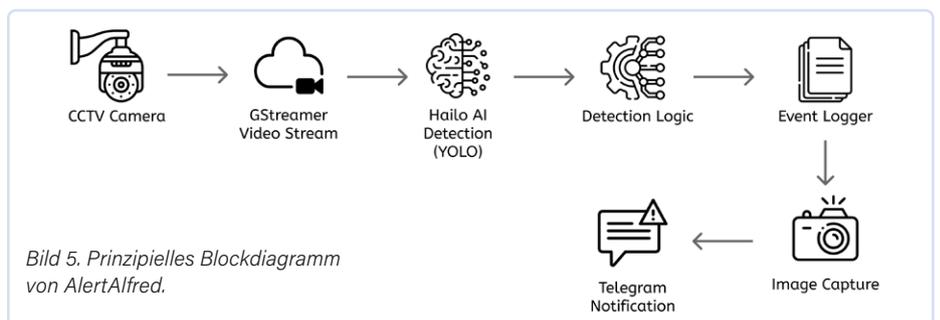


Bild 5. Prinzipielles Blockdiagramm von AlertAlfred.



## Listing 1: Python-Hauptskript (Ausschnitt)

```
# Import necessary libraries
import gi, os, csv, time, requests, cv2, hailo
from datetime import datetime
from gi.repository import Gst
from hailo_rpi_common import get_caps_from_pad, get_numpy_from_buffer, app_callback_class
# Define a custom callback class for handling detections and alerts
class user_app_callback_class(app_callback_class):
    def __init__(self):
        ...
        self.grace_period = 2 # Grace period in seconds
        self.csv_log_path = "logs/detection_log.csv"
        os.makedirs("logs", exist_ok=True) # Ensure logs directory exists
        with open(self.csv_log_path, 'a', newline='') as file:
            csv.writer(file).writerow(["Timestamp", "Event"]) # Initialize CSV log
        ...
    def log_event(self, event):
        # Log detection events to CSV
        ...
    def send_telegram_alert(self, image_path):
        # Send detection alert with image via Telegram
        ...
# Callback function for handling frame detections from the pipeline
def app_callback(pad, info, user_data):
    frame = get_numpy_from_buffer(info.get_buffer(), *get_caps_from_pad(pad))
    detections = hailo.get_roi_from_buffer(info.get_buffer()).get_objects_typed(hailo.HAILO_DETECTION)
    ...
    # Check if a person is detected
    if any(d.get_label() == "person" for d in detections):
        if not user_data.person_detected:
            # Log entry event
            ...
            if user_data.detection_frame_count == 10:
                user_data.send_telegram_alert("image_path.png") # Send alert
            elif time.time() - user_data.last_detection_time > user_data.grace_period:
                user_data.person_detected = False # Reset detection status if grace period passed
            ...
# Run the detection app
if __name__ == "__main__":
    app = GStreamerDetectionApp(app_callback, user_app_callback_class())
    app.run()
```

Personenerkennung über zehn Bilder hinweg, bevor es einen Alarm auslöst. So arbeitet es zuverlässig auch unter realen Bedingungen, in denen temporäre Objekte oder Schatten eine falsche Erkennung auslösen könnten.

### Bilderfassung und Alarme

Sobald eine Person erkannt wird, nimmt AlertAlfred einen Schnappschuss auf und sendet eine Benachrichtigung über Telegram. Diese Benachrichtigung enthält das Bild und gewährleistet eine sofortige visuelle Überprüfung durch den Benutzer.

### Ereignisprotokollierung

Alle Erkennungsereignisse (Ein- und

Ausgänge) werden in einer CSV-Datei protokolliert, um zur späteren Überprüfung eine detaillierte Aufzeichnung der Aktivitäten zur Hand zu haben.

### Karenzzeit für Benachrichtigungen

Das System verfügt über eine Karenzzeit von zehn Sekunden: Sobald eine Person erkannt wurde, sendet das System innerhalb dieses Zeitraums keine weitere Benachrichtigung für dieselbe Person. Dies verhindert wiederholte Benachrichtigungen, wenn die Person im Blickfeld der Kamera verbleibt. Um einen zweiten Alarm auszulösen, muss die Person das Bild verlassen und nach Ablauf der Karenzzeit wieder eintreten.

### Der Code

Der Hauptcode in **Listing 1** steuert die Echtzeit-Personenerkennung in einem Video-Feed, die Protokollierung von Ereignissen und den Versand von Warnmeldungen über Telegram. Für die Bildverarbeitung sind Bibliotheken wie `cv2` und `numpy` wichtig, während `requests` die HTTP-Kommunikation mit Telegram übernimmt. Das KI-Modul Hailo 8L ist für eine effiziente Personenerkennung zuständig, und GStreamer (`Gst`) verwaltet den Videostream der Überwachungskamera. Im hier abgedruckten Listing ist nur ein Ausschnitt des Codes zu sehen; der Hauptcode ist, wie bereits erwähnt, im GitHub-Repository zu finden.

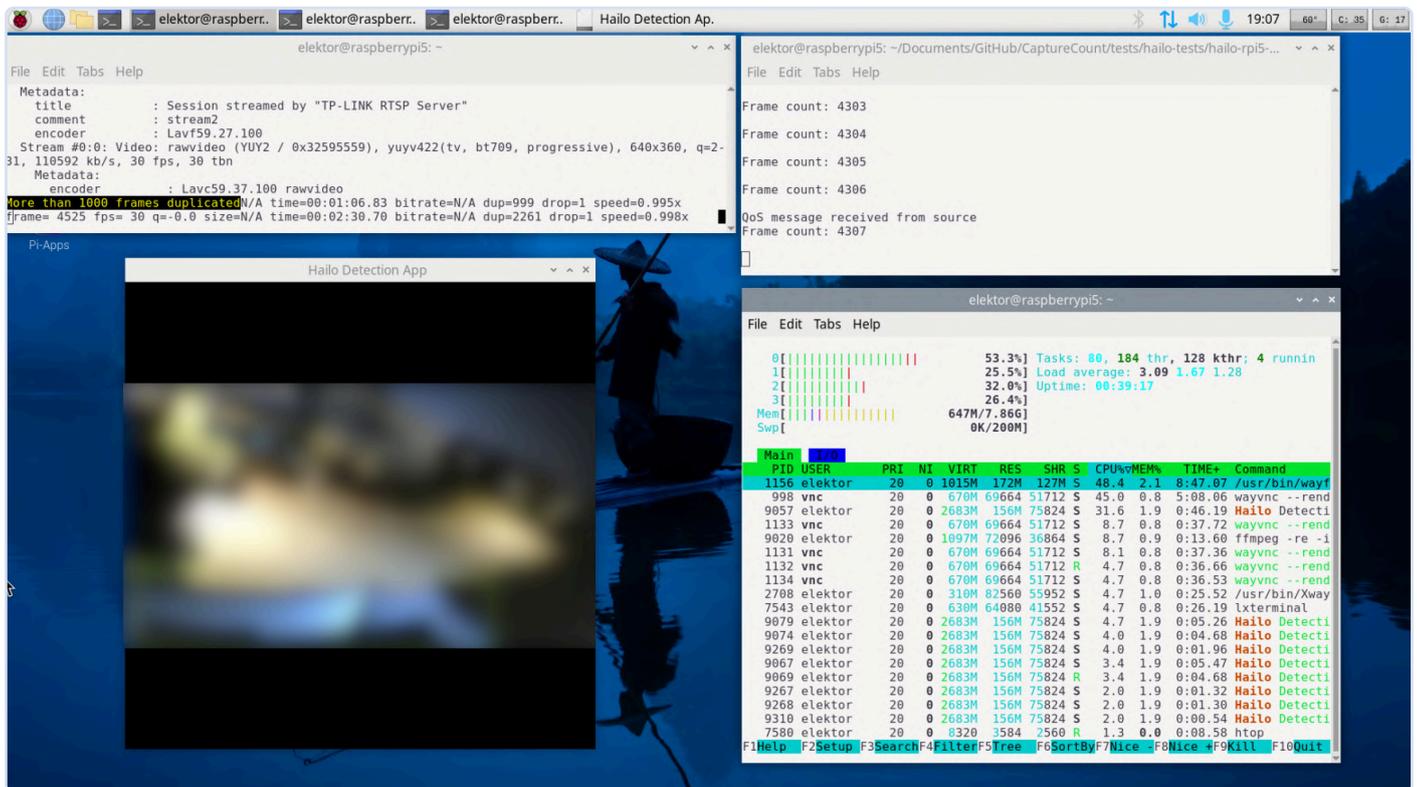


Bild 6. Der CCTV-Live-Feed wird von AlertAlfred mit Personenerkennung in Echtzeit auf dem Raspberry Pi verarbeitet.

Die Erkennungslogik ist in der `user_app_callback_class` gekapselt, die Erkennungsergebnisse verfolgt, protokolliert und Benachrichtigungen sendet. Schlüsselvariablen wie `self.person_detected` zeigen an, ob eine Person gerade erkannt wird, und `self.detection_frame_count` sorgt dafür, dass das System erst nach zehn aufeinanderfolgenden Frames Alarm auslöst, um Fehlalarme zu vermeiden. Die `grace_period` von zwei Sekunden verhindert wiederholte Alarme für dieselbe Person, während `self.csv_log_path` definiert, wo die Erkennungsergebnisse protokolliert werden. Alarme werden über `self.telegram_bot_token` und `self.telegram_chat_id` an die Telegram-API gesendet.

Die Methode `setup_csv_log` stellt sicher, dass ein Logs-Verzeichnis und eine CSV-Datei existieren, während `log_event` die Ein- und Austrittsergebnisse von Personen aufzeichnet. Die Methode `send_telegram_alert` erstellt und sendet eine Nachricht mit den Erkennungsdetails und einem begleitenden Bild der erkannten Person an Telegram.

Die `app_callback`-Funktion ist das Herzstück der Erkennungspipeline und wird ausgelöst, wenn neue Videodaten verfügbar sind. Sie extrahiert das Bild und verarbeitet es mit dem Hailo-8L-Modul, wobei sie auf die Erkennung von „Personen“ prüft. Wird eine Person über einen Zeitraum von zehn Frames erkannt, wird ein Bild aufgenommen, gespeichert und an Telegram versendet. Wird nach der Karenzzeit keine Person erkannt, wird der Erkennungs-

status zurückgesetzt und ein Exit-Ereignis protokolliert.

Schließlich erstellt das Programm eine Instanz der Callback-Klasse, initialisiert die Erkennungs-App und führt das System aus. Dieses Software-Setup ermöglicht eine effiziente Personenerkennung, Ereignisprotokollierung und Echtzeitwarnungen, unterstützt durch das Hailo-8L-Modul zur KI-Beschleunigung und datenschutzfreundlichen lokalen Verarbeitung.

### AlertAlfred in Aktion

Um AlertAlfred auf Ihrem Raspberry Pi 5 einzusetzen, benötigen Sie zunächst einige wichtige Hardware-Komponenten, natürlich den Raspberry Pi 5 selbst, dann das Raspberry Pi KI-Kit (das Modul Hailo 8L und ein Adapter M.2 HAT+), eine kompatible Überwachungskamera und eine microSD-Karte, auf der das Raspberry Pi OS installiert wurde.

Für die Software-Einrichtung müssen Sie zunächst das Hailo-Repository klonen, das als Grundlage für die Erstellung der Erkennungspipeline dient. Dies kann mit folgendem Befehl geschehen:

```
git clone https://github.com/hailo-ai/hailo-rpi5-examples.git
```

Danach werden die erforderlichen Abhängigkeiten installiert, einschließlich OpenCV, NumPy und GStreamer. Sie können dies tun, indem Sie einfach den Anweisungen in `doc/basic-pipelines.md` folgen, das Sie im

Repository `hailo-rpi5-examples` finden. Nachdem Sie dies eingerichtet haben, konfigurieren Sie Telegram, indem Sie einen Bot über BotFather erstellen und das Bot-Token und das Bot-Chat-ID abrufen. Diese Angaben werden in das Skript eingefügt, um Echtzeitwarnungen zu ermöglichen. Kopieren Sie das Skript `alert-alfred.py` schließlich in den Ordner `basic_pipelines` im Repository ein. Bevor Sie das Skript ausführen, müssen Sie den Prozess in einem separaten Terminal starten, um die Aufzeichnung des Videomaterials auf Ihrer virtuellen Kamera zu starten. Dies können Sie mit dem folgenden Befehl tun (bitte folgen Sie der Anleitung [4], bevor Sie diesen Befehl ausführen):

```
ffmpeg -re -i rtsp://username:password@172.168.1.71:554/stream1 -r 30 -f v4l2 -vcodec rawvideo -pix_fmt yuyv422 /dev/video10
```

Nachdem Sie mit der Aufzeichnung des Videomaterials auf Ihrer virtuellen Kamera begonnen haben, können Sie das Skript sich im Ordner `hailo-rpi5-examples` mit dem folgenden Befehl über das Terminal ausführen:

```
python basic_pipelines/alert-alfred.py --input /dev/video10
```

Vorausgesetzt, dass der `/dev/video10`-Eingang für das Skript Ihr Videomaterial korrekt streamt, beginnt das System mit der

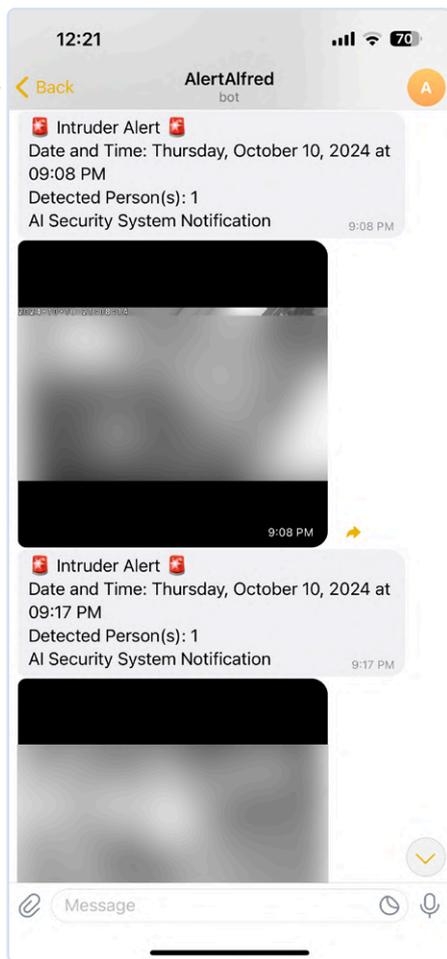


Bild 7. Telegramm über Telegram: Eine Person wurde erkannt, mit Zeitstempel und Standbild aus dem CCTV-Feed.

Verarbeitung des Feeds zur Personenerkennung. Wenn alles wie erwartet funktioniert, sollte es so aussehen wie in **Bild 6**, das zeigt, wie das Live-Feed auf Erkennungen verarbeitet wird.

Wenn das System eine Person erkennt, nimmt es ein Bild auf und sendet eine Echtzeit-Benachrichtigung über Telegram, einschließlich des Zeitstempels und der Anzahl der erkannten Personen, wie in **Bild 7** dargestellt. Diese Benachrichtigung bietet eine sofortige visuelle Bestätigung und stellt sicher, dass Sie über jede erkannte Aktivität auf dem Laufenden bleiben.

## Zukünftige Möglichkeiten

Während sich die aktuelle Version von AlertAlfred auf die Personenerkennung mit

einer einzigen Kamera und das Versenden von Benachrichtigungen über Telegram konzentriert, hat das Projekt das Potenzial, sich zu einem robusteren und vielseitigeren Überwachungssystem zu entwickeln. Zukünftige Erweiterungen könnten mehrere Streams unterstützen, was die Überwachung größerer Bereiche oder mehrerer Eingangspunkte ermöglichen würde. Neben der Personenerkennung könnte das System auch auf andere Szenarien erweitert werden, zum Beispiel auf Brandgefahren, Stürze älterer Menschen oder die Sicherheitsüberwachung in verschiedenen Umgebungen. Da der gesamte Prozess lokal auf dem Raspberry Pi abläuft, bietet das Projekt eine solide Grundlage für datenschutzorientierte Anwendungen.

Beiträge aus der Community können dieses Projekt vorantreiben, sei es durch die Verbesserung von Erkennungsalgorithmen, die Erweiterung des Spektrums der erkennbaren Szenarien oder das Hinzufügen neuer Funktionen. Alle Beiträge oder Vorschläge zur Verbesserung des Systems werden sowohl von den Entwicklern als auch von der breiten Community sehr geschätzt. ◀

RG — 240474-02

## Sie haben Fragen oder Kommentare?

Wenn Sie technische Fragen oder Anmerkungen zu diesem Artikel haben, wenden Sie sich bitte an den Autor unter [saad.imtiaz@elektor.com](mailto:saad.imtiaz@elektor.com) oder an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).

## Zu diesem Projekt beitragen

Wenn Sie sich an diesem Projekt beteiligen möchten, können Sie Ihre Beiträge auf der Online-Plattform von Elektor Labs veröffentlichen: [www.elektormagazine.de/labs](http://www.elektormagazine.de/labs).



## Über den Autor

Saad Imtiaz, Senior Engineer bei Elektor, ist Mechatronik-Ingenieur mit umfangreicher Erfahrung in Embedded Systems und Produktentwicklung. In seiner Karriere hat er mit einer Vielzahl von Unternehmen zusammengearbeitet, von innovativen Startups bis hin zu etablierten globalen Unternehmen, und dabei zukunftsweisende Prototypen- und Entwicklungsprojekte vorangetrieben. Mit seinem reichen Hintergrund, der auch eine Zeit in der Luftfahrtindustrie und die Leitung eines Technologie-Startups umfasst, bringt Saad eine einzigartige Mischung aus technischem Fachwissen und unternehmerischem Geist in seine Aufgaben bei Elektor ein. Hier trägt er zur Projektentwicklung sowohl in der Software- als auch in der Hardware-Entwicklung bei.

**THEMEN FOKUS**

Besuchen Sie unsere Webseite **Embedded & KI** für mehr Artikel, Projekte, Nachrichten und Videos.

[www.elektormagazine.de/embedded-ki](http://www.elektormagazine.de/embedded-ki)





## Passende Produkte

- **Raspberry Pi 5 (8 GB RAM)**  
[www.elektor.de/20599](http://www.elektor.de/20599)
- **Raspberry Pi, KI-Kit**  
[www.elektor.de/20879](http://www.elektor.de/20879)
- **Raspberry Pi, KI-Kamera**  
[www.elektor.de/20953](http://www.elektor.de/20953)
- **Gehäuse Argon NEO 5 BRED für Raspberry Pi 5**  
[www.elektor.de/20788](http://www.elektor.de/20788)

## WEBLINKS

- [1] Hailo-Raspberry Pi 5-Beispiele im Github-Repository: <https://github.com/hailo-ai/hailo-rpi5-examples>
- [2] GitHub-Repository RTSP-to-webcam: <https://github.com/apple-fritter/RTSP-to-webcam>
- [3] GitHub-Repository des KI-Sicherheitssystems AlertAlfred: <https://github.com/ElektorLabs/Alert-Alfred>
- [4] GitHub-Repository, Anleitung zum Streamen der CCTV-Kamera auf die virtuelle Kamera:  
<https://github.com/ElektorLabs/Alert-Alfred/blob/main/docs/cctv-to-virtualcam-guide.md>
- [5] Saad Imtiaz, „CaptureCount, Ein Objektdetektor und -zähler auf dem Raspberry Pi 5“, Elektor 3-4/2024:  
<https://elektormagazine.de/230749-02>

# KI in der Elektronik-entwicklung

## Ein Update nach nur einem Jahr

Von Tam Hanna (Ungarn)

Seit dem Aufkommen von Flux.AI leistet die Künstliche Intelligenz auch im Elektroniklabor wertvolle Hilfsdienste. Da seit unserer letztjährigen Übersicht von KI-Tools einiges an Wasser die Donau hinuntergeflossen ist, hier ein kleines Update, das auf aktuelle Entwicklungen eingeht.

Ein Artikel zu KI-Systemen kann prinzipiell immer nur eine Momentaufnahme der Fähigkeiten der jeweiligen Systeme darstellen. Aus der Logik der dynamischen Trainierbarkeit folgt, dass heute gemachte Feststellungen schon morgen nicht mehr gültig sein können. Deshalb ist es nicht verwunderlich, dass es sich nach der letztjährigen Übersicht [1] schon ein knappes Jahr später lohnt, die neusten Entwicklungen der KI-Tools für die Elektronikentwicklung erneut unter die Lupe zu nehmen. Die hier beschriebenen Experimente wurden Ende August 2024 vorgenommen, alle danach aufgetretenen Veränderungen sind naturgemäß nicht berücksichtigt.

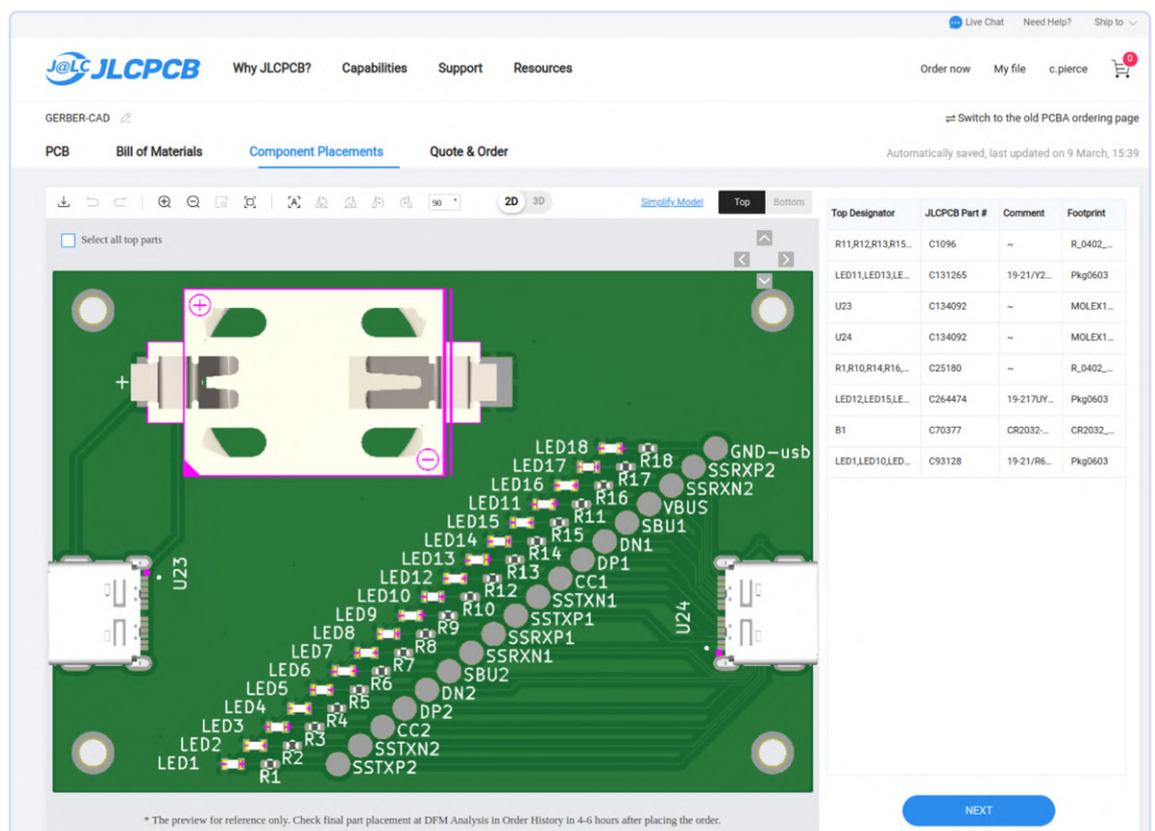


Bild 1. Dieses Board kann komplett unter Nutzung „generativer Programmiersprachen“ entstehen.

## JITX oder ein Blick hinter die Kulissen

Moderne Systeme der künstlichen Intelligenz präsentieren sich im Allgemeinen als Blackbox. Der Anwender füttert sie mit Anfragen, um danach mehr oder weniger gut nutzbare Antworten zu erhalten. Was im Hintergrund passiert, bleibt dem Nutzer des Systems normalerweise verborgen. Nach Ansicht des Autors dürfte zumindest in vielen Fällen im Hintergrund eine codeartige Repräsentation einer Schaltung stehen.

Eines der wenigen Unternehmen, das seine textuelle Beschreibungs-Sprache offenlegt, ist JITX. Um einen Blick hinter die Kulissen zu wagen, wollen wir deshalb das unter [2] bereitstehende Beispielprojekt kurz ansehen, einen USB-Kabeltester (**Bild 1**). Die USB-Buchse wird zum Beispiel durch folgendes Snippet ins Leben gerufen:

```
val usb-x-shift = 4.0
val usb-y-shift = -7.0
public inst in-usb :
  components/USB-C-1054500101/component
place(in-usb) at loc((-1.0 * width(board-shape) / 2.0) +
  usb-x-shift, usb-y-shift, -90.0) on Top
public inst out-usb :
  components/USB-C-1054500101/component
place(out-usb) at loc((width(board-shape) / 2.0) -
  usb-x-shift, usb-y-shift, 90.0) on Top
```

Wichtig sind hier unter anderem die nach dem Schema `components/USB-C-1054500101/component` aufgebauten Strings, die einen Verweis auf das jeweilige Bauteil-Repository im AI-EDA-System darstellen. Dieses Schema wird uns in den folgenden Schritten noch häufiger verfolgen. Wichtig ist im Moment noch die Feststellung, dass die Positionierung hier durch vom User einzugebende Koordinaten erfolgt. Das System ist also nicht zur selbstständigen Positionsberechnung befähigt. Sind die Komponenten erst einmal ins Leben gerufen, müssen sie

natürlich verbunden werden. Beim hier verwendeten USB-Interface machen wir uns „Namens-Ähnlichkeiten“ zwischen den Pins in den beiden Komponenten in einer nach folgendem Schema aufgebauten `for`-Schleife zu Nutze:

```
for p in pins(out-usb) do :
  val pin-ref = ref(p)
  val pin-name = tail(pin-ref, ref-length(pin-ref) - 1)
  make-net(to-symbol(pin-name), [p])
```

Neben der Möglichkeit zum Einbetten von Intelligenz in den Schaltungsentwurf ist ein Vorteil dieser Vorgehensweise ein Feature, das dank Standardisierung der Bibliothek-Footprints das Austauschen verschiedener Objekte erleichtert. Im Fall der USB-Testplatine wird dies sogar demonstriert: Der Wechsel zwischen zwei unterschiedlichen Buchsen erfolgt durch Austauschen des übergebenen Strings. Zu JITX sei angemerkt, dass ein kostenlos nutzbarer Plan [3] zur Verfügung steht, der zum Zeitpunkt der Drucklegung dieses Artikels allerdings auf die Erzeugung von quelloffenen Designs beschränkt ist, die unter der Lizenz CERN OHL-Permissive v2 liegen. Die Nutzung der GPL für das resultierende Design wird vom Anbieter explizit ausgeschlossen.

Zu beachten ist außerdem, dass JITX selbst nur vergleichsweise wenige AI-Features anbietet. Der hauptsächliche Nutzwert ist die Möglichkeit, in den Designfiles verschiedene Constraints unterzubringen, die das System dann automatisch zur Überprüfung der vorliegenden Aufgabe anwendet.

## SnapMagic Copilot - oder der Blitz im Elektronikbereich

Der nächste Teilnehmer in dieser Übersicht ist SnapMagic. Dabei handelt es sich eigentlich um eine Weiterentwicklung des bekannten Diensts SnapEDA, der sich normalerweise auf das Anbieten von Bauteilinformationen für EDA-Software von Drittanbietern spezialisiert hat.

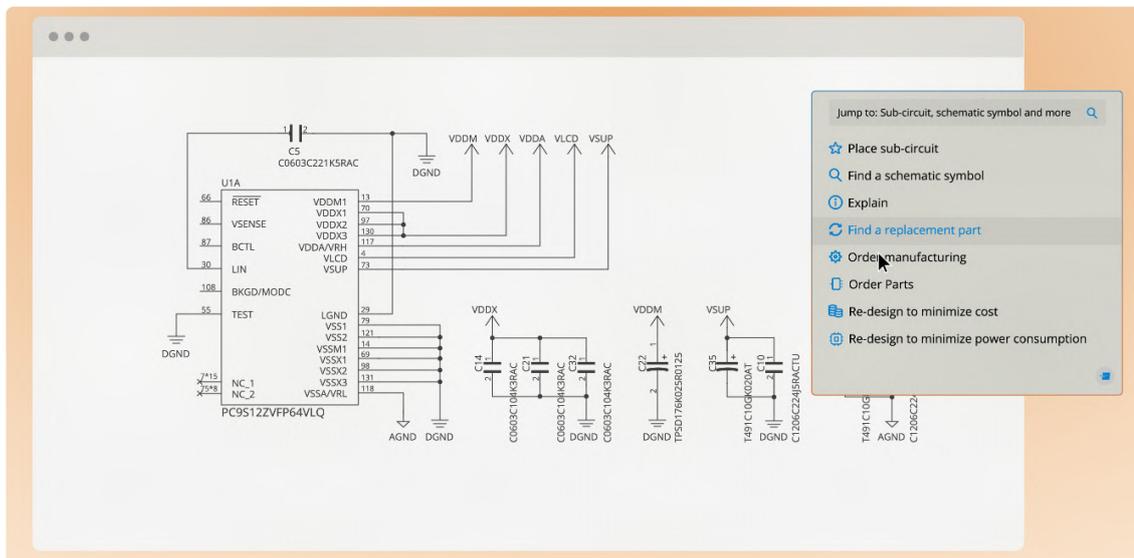


Bild 2. SnapMagic Copilot kommuniziert in natürlicher Sprache, um die Ergebnisse danach in vorhandene EDA-Software zu exportieren.

Als erster Versuch bietet sich die Suche nach einem vergleichsweise neuen Quectel-Funkmodul an, für das bei den meisten Anbietern noch keine Footprints zur Verfügung stehen. Die größte Stärke von *SnapMagic Search*, wie der Dienst nun heißt, ist mit Sicherheit seine Fähigkeit, sich in alle mögliche EDA-Software zu integrieren. Mit einem neuen Dienst *Snap Magic Copilot*, der derzeit in einer geschlossenen Beta-Phase zur Verfügung steht, möchte man im Hause SnapMagic diese Funktionen erweitern. Dabei handelt es sich dabei um ein KI-System, das in natürlicher Sprache gehaltene Anfragen entgegennimmt und daraufhin in verschiedenste EDA-Systeme exportierbare Grundsaltungen anbietet.

In Demonstrationen zeigt das Unternehmen außerdem das „Errichten von Unterstützungsstruktur“: Der Copilot kann beispielsweise ein Datenblatt analysieren und die zur sicheren Inbetriebnahme eines integrierten Schaltkreises notwendigen Stützkondensatoren automatisiert hinzufügen. Schade ist in diesem Zusammenhang, dass die unter der [4] bereitstehende Warteliste zurzeit gefüllt zu sein scheint - dem Aufnahme-Antrag des Autors wurde bis Redaktionsschluss nicht positiv beschieden.

Interessant ist im Zusammenhang auf SnapMagic allerdings, dass man die Einführung des KI-Assistenten als Werkzeug zum Anbieten von „Value Added Services“ sieht. **Bild 2** zeigt, dass sich SnapMagic um die Bauteilbeschaffung kümmern möchte.

### CircuitMind: Auf funktionaler Ebene sattelfest

Der nächste Kandidat im Testfeld konzentriert sich darauf, den Nutzer von Aufgaben zu entlasten und diese danach - wie im Fall von CircuitMind [5] in **Bild 3** schematisch gezeigt - in Schaltungen beziehungsweise Schaltungsaufgaben umzuwandeln. Von Seiten des Elektroniklers wird dabei die Eingabe von Constraints und Schnittstellen erwartet, die das von der künstlichen Intelligenz zu beschreibende Schaltungsdesign so gut wie möglich beschreiben.

Bei CircuitMind wird auf der Webseite die Eingabe eines Benutzernamens und Passworts verlangt. Nach dem Ausfüllen und Abschicken des Formulars wird ein Terminkalender zugesandt, in dem der Entwickler einen Termin für eine Live-Demo buchen kann. In Tests des Autors erwiesen sich die Termine dabei als durchaus leicht erhältlich; ein am nächsten Werktag beantragter Termin wurde problemlos gewährt. Der Autor wurde dabei vom Firmengründer Tomide Adesanmi persönlich begrüßt - als erstes wies der ehemals bei BAE Systems tätige Entwickler darauf hin, dass das Lesen von Datenblättern zu den langweiligsten Aufgaben gehört, mit denen ein Elektroniker im Rahmen seiner Karriere zu tun hat.

Sein Produkt sieht sich deshalb explizit als System Design Automation Tool, das nicht in den umkämpften EDA-Markt einsteigen möchte. Stattdessen hat das System die Absicht, Designs und auf verschiedene Parameter optimierte Varianten dieser Designs zu berechnen und einander automatisiert gegenüberzustellen.

Im Hintergrund kommt dabei eine aus zwei Komponenten bestehende Architektur zum Einsatz. Erster Teil ist eine als *Commodore* bezeichnete Bauteildatenbank, die Modelle der dem System bekannten Komponenten auflistet. 70% der Datenbank wird automatisiert erstellt, während der Rest manuell eingepflegt und geprüft wird, womit ein Gutteil der hinter dem Produkt stehenden Mannschaft beschäftigt ist. Der zweite Teil namens *Ace* hat dann die Aufgabe, aus „Design Goals“ und den in *Commodore* enthaltenen Daten Schaltungen sowie Varianten zu entwerfen.

Interessant ist, dass dieser Synthese-Prozess durch „klassische, regelbasierte AI“ erfolgt. Adesanmi betonte im Rahmen der Besprechung mehrfach, dass das System ohne Large Language Models (LLM) auskommt. Vorteil dieser Vorgehensweise ist, dass LLMs zum Halluzinieren neigen, was bei einer rein regelbasierten künstlichen Intelligenz naturgemäß nicht auftritt. Sei es wie es sei, ist der

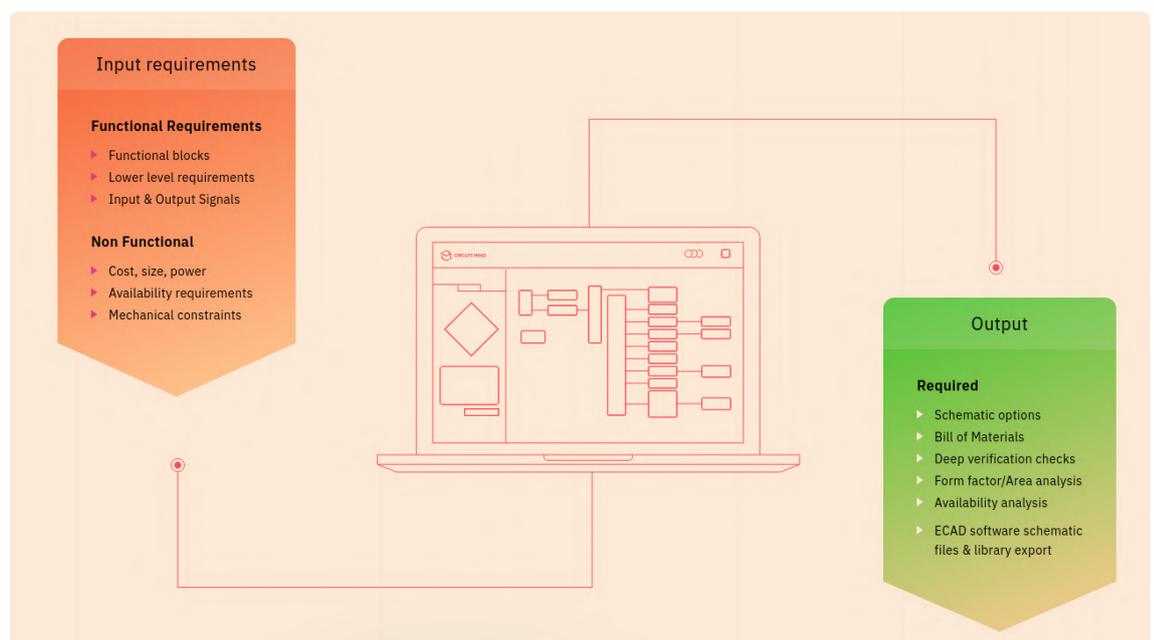


Bild 3. CircuitMind kümmert sich um das Design auf funktionaler Ebene.

Compare Solutions					
Size/Cost/Power		BOM			
Show $\Delta$ relative to Reference Solution					
Ref	Solution	Tradeoffs (Size / Price / Power)	Size	Price	Power
●	<b>3. Power Optimised</b> 7 months ago	<b>0% / 0% / 100%</b> 100 boards	<b>828.12 mm<sup>2</sup></b> +55%	<b>\$27.40</b> +46%	<b>14mW</b> -94%
●	<b>2. Cost Optimised</b> 8 months ago	<b>0% / 100% / 0%</b> 100 boards	<b>836.42 mm<sup>2</sup></b> +57%	<b>\$13.65</b> -27%	<b>242.6mW</b> +10%
●	<b>1. Balanced</b> 8 months ago	<b>33% / 33% / 33%</b> 100 boards	<b>534.41 mm<sup>2</sup></b> 0%	<b>\$18.73</b> 0%	<b>219.6mW</b>

Bild 4. Solution Exploration erlaubt das Abwägen der Vor- und Nachteile verschiedener Designs.

eigentliche Vorteil dann eine als *Solution Exploration* bezeichnete Ansicht. **Bild 4** zeigt beispielhaft, was der Designer erwarten darf. In diesem Fall ist klar erkennbar, dass sich die verschiedenen Versionen sowohl in der Platinengröße als auch in den zu erwartenden Hardwarekosten unterscheiden.

Durch Anklicken der einzelnen Felder lassen sich verschiedenste Dokumente generieren: Neben einer interaktiven Bill of Materials, die das Analysieren der verschiedenen Kostenstellen erlaubt, generiert das System Interface-Dokumente und verschiedene andere Dateien, die beispielsweise bei der Zertifizierung hilfreich werden können.

Die vergleichsweise teure Lösung - der Einstiegspreis liegt im Bereich von rund 10 000 US-Dollar - beeindruckte den Autor durch eine weitere, sehr innovative Funktion: Das System ist in der Lage, „analoge Blöcke“ in den Syntheseprozess einzubinden. Dabei handelt es sich um Platzhalter, in denen mehr oder weniger generische Elektronik beziehungsweise Schaltungsdesigns untergebracht werden können. Das System berücksichtigt diese beim Design dann mit - selbst dann, wenn die in ihnen enthaltenen Komponenten unbekannt sind.

### Celus: vom „reichen Datenblatt“ zur automatischen Schaltungsgenerierung

Ein weiterer Weg, über den KI den Weg in die Welt der Elektronik findet, sind automatische Schaltungs-Generatoren. Systeme wie Würths Red Expert, Microchips Mindi oder Texas Instruments Online-Designsystem bieten schon seit längerer Zeit die Möglichkeit, Parameter einzugeben und danach ausparametrierte Schaltungen samt Empfehlungen für passive Komponenten zu erhalten. Celus bietet mit dem unter [6] im Detail beschriebenen CUBOs-Format ein digitales Datenblatt an, das interessanterweise seit einiger Zeit vom österreichischen Leistungselektronik-Spezialisten RECOM unterstützt wird.

Wer mit Celus loslegen will, besucht im ersten Schritt die Webseite [7], auf der es neben der Option zur Erzeugung eines lokalen Kontos am Server des Anbieters auch möglich ist, sich mit einem LinkedIn- oder Google-Konto zu authentifizieren. Der Autor entschied sich für die folgenden Schritte für ein Google-Konto und wurde nach der Beantwortung einer Frage zu „Ausbildungsstand und Anwendungsintention“ direkt in der als *Design Studio* bezeichneten Oberfläche

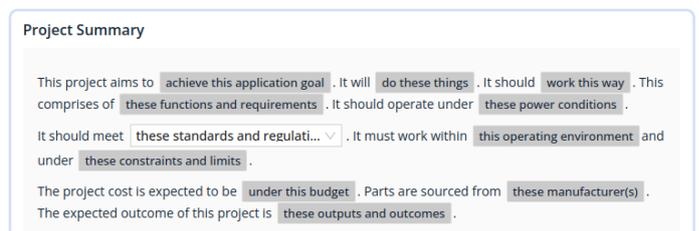


Bild 5. Dieser Lückentext ist der erste Schritt zum lauffähigen Projekt.

abladen. Zur leichteren Einführung bietet die Plattform aber auch einen mehrstufigen Assistenten an.

Interessant ist die Art der Eingabe der Konfiguration. Im ersten Schritt erscheint dabei normalerweise ein Text, der wie in **Bild 5** aufgebaut ist und in den der Ingenieur verschiedene Felder mit den gewünschten Informationen einträgt. In den beiden darauffolgenden Schritten sind zusätzliche Konfigurationen notwendig; im Bereich der unterstützten EDA-Systeme findet man derzeit nur den Dreikampf aus Altium, Eagle und KiCad. Danach lässt sich ein Template auswählen, eine Art „Vorlage“, die für häufig auftretende Aufgaben benötigte Konfigurationen zusammenfasst.

Für die vom Autor probeweise angegebene Aufgabenstellung des Entwurfs eines 4G-Funkmodul auf Basis eines Quectel-Chipsatzes hatte das System leider keinen Vorschlag, weshalb nach Anklicken der Option *Go to Canvas* ein „leerer“ Design-Canvas erschien. Als Lohn der Mühen erscheint ein Arbeitsbereich, der, wie **Bild 6** zeigt, deutlich von Klassikern wie Microsoft Visio inspiriert wurde.

Der ganz links eingeblendete Werkzeugkasten erlaubt dem Elektroniker dabei die Auswahl von *Prefabs*, die Schaltungsaufgaben abbilden. Sie lassen sich per Drag-and-Drop in den Designbereich in der Mitte verschieben und durch Hinzufügen von Pins mit anderen Funktionseinheiten verbinden. Außerdem steht bei den meisten Funktionseinheiten - der *USB Connector* ist hierfür ein positives, *Wireless Communication* leider ein negatives Beispiel - die Möglichkeit zur Verfügung, durch einen Rechtsklick auf das jeweilige Element ein Fenster mit verschiedenen zusätzlichen Optionen zu laden. Im Fall des USB-Steckers könnte man dort beispielsweise auswählen, ob Polaritäts-Umkehrschutz oder ESD-Absicherung im vorliegenden Design erforderlich sind.

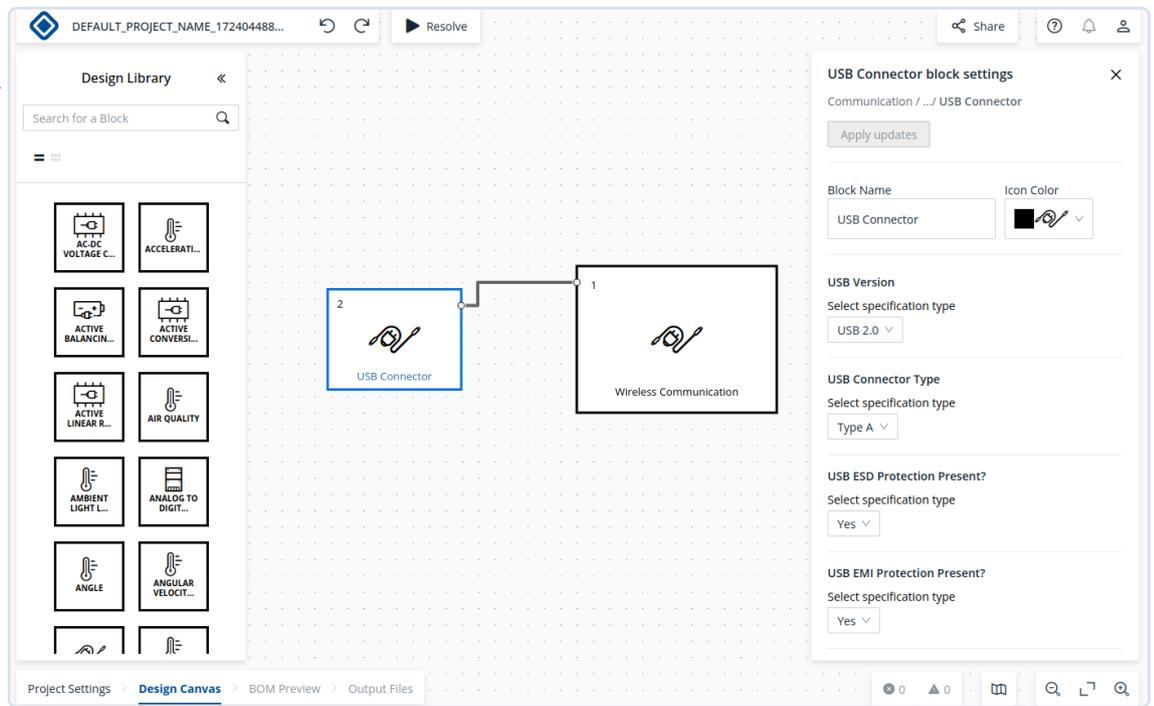


Bild 6. Ähnlichkeiten zwischen Celus und Visio dürften rein zufällig sein.

Nach getaner Arbeit folgt ein Klick auf die oben eingblendete Resolve-Schaltfläche, die im Idealfall einen funktionierenden Entwurf zur Verfügung stellt. Im Fall des in Bild 6 gezeigten Entwurfs lieferte das System stattdessen eine Fehlermeldung, die auf nicht vorhandene CUBOs hinwies: Offensichtlich hatte der Autor den USB Connector mit „zu vielen“ Constraints versehen, weshalb die (eingeschränkt umfangreiche) Bibliothek keine passenden Komponenten finden konnte.

Zur Durchführung weiterer Experimente geht man zur Start-Oberfläche zurück, in der Projektbeispiele von Renesas und STMicroelectronics vorliegen. Der „Ultra Low Power Pet Tracker“ klingt insofern interessant, als dass er prinzipiell ja ebenfalls eine Funk-Schnitt-

stelle enthalten sollte. Lohn der Mühen ist das Design in **Bild 7**, wobei interessant ist, dass manche der Verbindungen hier an Visio erinnernde Call-Outs aufweisen, die die „Art“ der zwischen den Elementen zu errichtenden Verbindung beschreiben.

Im nächsten Schritt folgt ein Klick auf die Resolve-Taste, woraufhin ein Roboter-Symbol erscheint, das neben dem Jonglieren von Rubik-Würfeln auch darauf hinweist, dass die Suche nach diversen geeigneten CUBOs läuft. Nach dem ersten Durchlauf des Suchprozesses erscheinen in den oberen rechten Ecke der einzelnen Bauobjekte blaue Punkte, die darauf hinweisen, dass Celus für diese Komponente „Implementierungs-Kandidaten“ anbieten möchte. Ein Rechtsklick auf den Punkt öffnet dann ein Kontextmenü, in

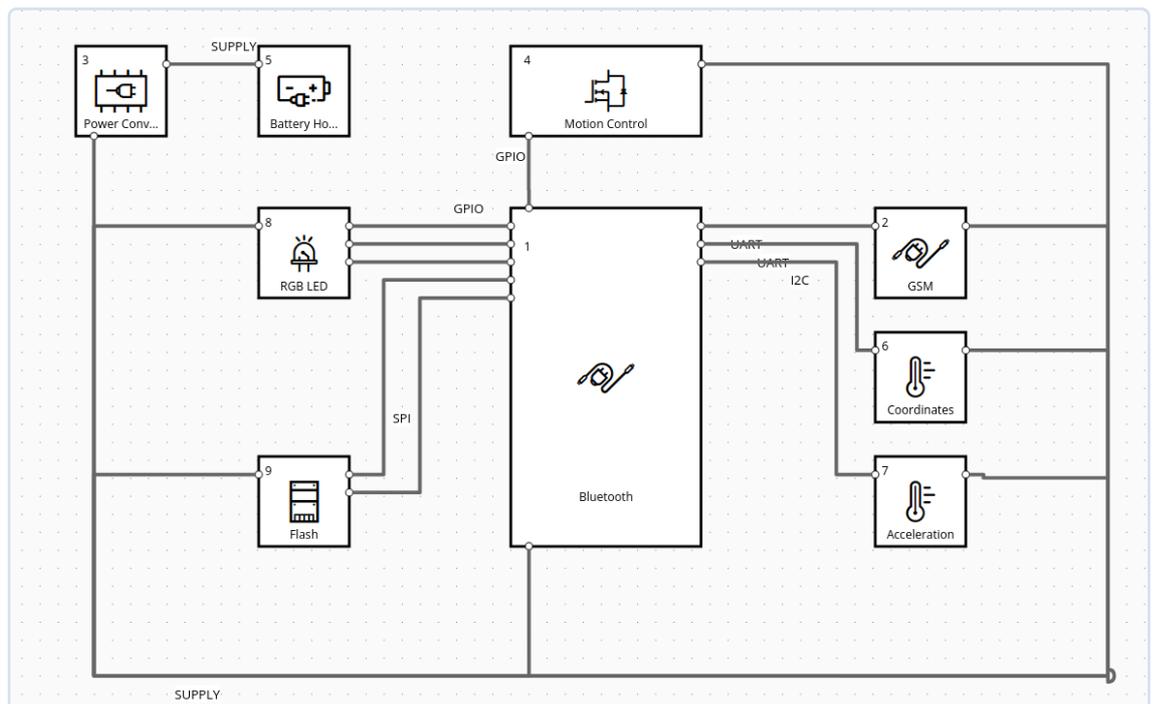


Bild 7. Das Beispielprojekt Ultra Low Power Pet Tracker in Celus.

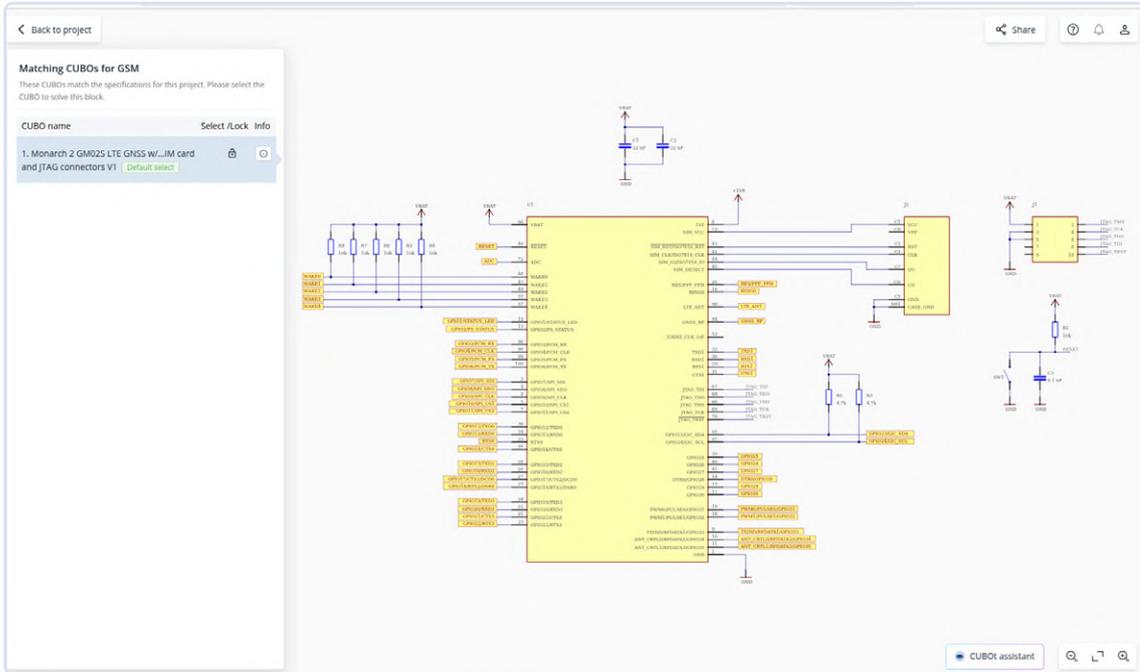


Bild 8. Die Auswahl im Bereich der GSM-Funkmodule ist bescheiden.

dem sich die jeweilige Option auswählen lässt. Im Fall beispielsweise des GSM-CUBOs wird dabei, wie in **Bild 8** gezeigt, nur ein Objekt angeboten. Das Anklicken des *Lock*-Symbols erlaubt dann, ein gewünschtes Objekt festzulegen.

Nach der Auswahl einiger Komponenten folgt ein abermaliger Klick auf den *Resources*-Button, der nun „Deliverables“ zur Verfügung steht. Besonders interessant empfand der Autor dabei den als PDF vorliegenden Projektbericht, der für dieses Projekt unter [8] zum Herunterladen bereitsteht. Wer eine Altium-Lizenz besitzt, kann das Projekt danach auch in Altium laden und an seine Bedürfnisse

anpassen. Interessant ist dabei, dass sich Celus aus der traditionell besonders haarigen Erzeugung der Platinenlayouts heraushält. Die Software generiert stattdessen nur Schaltplan, Stückliste und Projektdatei; das Layout scheint das Unternehmen dem Entwickler und seiner Kreativität zu überlassen.

## Automatisierte Fehlersuche in Hardware-Projekten

Was dem C-Programmierer sein Lint, möchte Cady für den Elektroniker sein. Das System unterstützt laut der offiziellen

Index	Part Reference	Pin Number	Nets	Components	Category	Description	Status
1	C136	1	NetC135_1	04024C102KA7A	Potential Capacitor Breach	Voltage difference of 5V is larger than the rated voltage 4V	...
2	U14	1	VDD_3V3	LTC297HDCB#TRMPBF	Connection Instruction Violation	a Capacitor connected between this pin and D- should have a value of 470pF	...
+ 3-4	...	17	...	MIC22705YML-TR	Connection Instruction Violation	a Capacitor connected between this pin and Ground should have a value of 2.2uF	...
5	C1	1	3V3_F	06036D476MAT2A	Potential Capacitor Breach	Voltage difference of 3.3V is too close to the rated voltage 6.3V 50% derating factor is recommended	...
+ 6-7	...	1	...	CC0402KRXR5BB225	Potential Capacitor Breach	Voltage difference of 3.3V is too close to the rated voltage 6.3V 50% derating factor is recommended	...
8	R32	2	...	RC0603FR-0712K1L	Invalid Passive Component Connection	This pin should not be left floating	...
9	U14	4	VDD_1V8	LTC297HDCB#TRMPBF	Invalid Supply Voltage	Supply voltage 1.8V is out of operating range (2.5V to 5.5V)	...
+ 10-20	U7	...	VCORE_FPGA	XC7AT5T-1CSG324I	Invalid Supply Voltage	Supply voltage 1.7V is out of operating range (0.87V to 1.05V)	...
+ 21-22	U7	...	VCORE_FPGA	XC7AT5T-1CSG324I	Invalid Supply Voltage	Supply voltage 1.7V is out of operating range (0.92V to 1.05V)	...
23	U18	1	GND	TPS62080ADSGT	Activity Level Violation	Active HIGH ENABLE pin should not be connected to the Ground	...
24	U3	32	RBIAS	LAN8740A-EN-TR	Connection Instruction Violation	This pin should be connected to Ground via a Resistor	...
25	U19	4	NetL15_1	LT1764EQ-3.3#TRPBF	Connection Instruction Violation	This pin should be connected to Ground via a Capacitor	...
26	Y2	...	...	RH100-25.000-10-F-2030-EXT-TR	Oscillator Violation <b>New!</b>	Connected load capacitance should have a value of 10pF	...
27	U15	1	D+	FT234XD-T	Differential Lines	Potential mismatch - N type pin connected to P type net	...
28	U15	12	D-	FT234XD-T	Differential Lines	Potential mismatch - P type pin connected to N type net	...
+ 29-30	U1	...	...	STM32F427VIT7TR	Missing Pull Resistor	Pull-up resistor is missing	...
31	U7	b10	NC_U7_b10	XC7AT5T-1CSG324I	Floating Power Pin		...
32	U18	...	...	TPS62080ADSGT	Thermal Pad Connection Instruction Violation	This component's Exposed Pad must be connected to GND, yet it does not appear in the netlist	...
+ 33-34	...	...	...	...	Calculation Mismatch <b>New!</b>	Net name implies 1.8V while calculated voltage from U16 (MIC22705YML-TR) is 2.34V	...
35	U12	t2	DDR3_RESEtN	MT41K256M16TW107IT	Best Practice	It is recommended to add a capacitor to Ground on reset pins for multiple purposes	...
+ 36-37	...	...	...	...	Best Practice	Missing Test Point	...

Bild 9. Die Analyse durch Cady erweist sich als durchaus pedantisch.

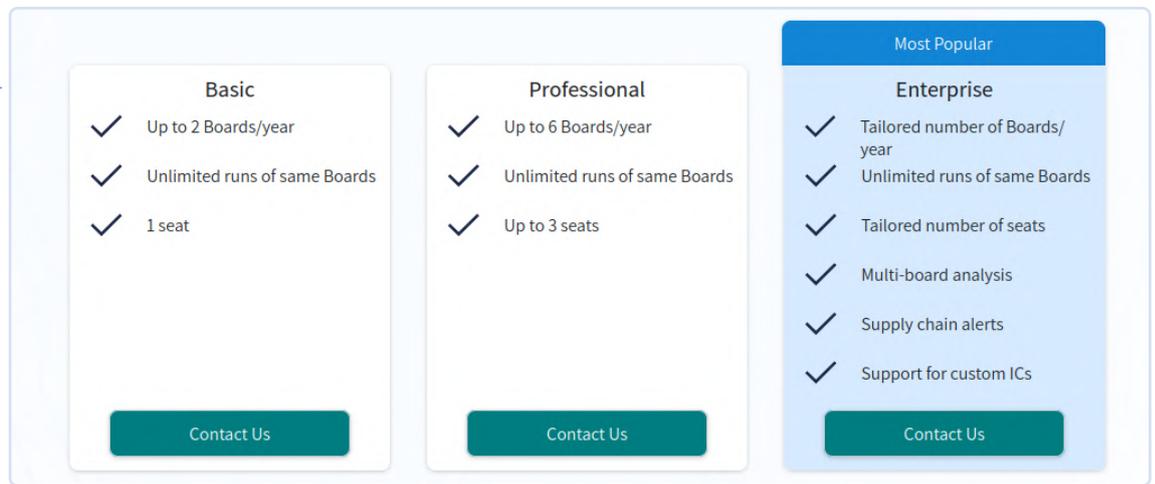


Bild 10. Über die bei der Nutzung von Cady anfallenden Kosten schweigt sich das Start-up aus.

Dokumentation nur die Formate Altium, Xpedition und OrCAD - laut den „Hintergrund-Informationen“ unter [9] allerdings auch Eagle und KiCad. Bemerkenswert ist außerdem, dass das Unternehmen eine „grundlegende“ kostenlose Test-Möglichkeit zur Verfügung stellt. Wer sich dort mit seiner Firmen-E-Mail-Adresse anmeldet, bekommt Zugriff auf das System - zu beachten ist allerdings, dass diese Variante nur die vom Unternehmen zur Verfügung gestellten Design-Diagramme auswertet.

Als Resultat der Mühen wird dann eine „Fehler-Tabelle“ wie in **Bild 9** generiert. Auffällig ist dabei, dass Cady sowohl häufige als auch seltenere Fehler erkennt - das Start-up verspricht naturgemäß auch, dem System permanent neue Tricks beizubringen. Ein Kritikpunkt ist, dass die Bepreisung des Systems absolut nicht transparent ist. Das Unternehmen verspricht stattdessen wie in **Bild 10** gezeigt drei Preisstufen, wobei anzumerken ist, dass die als „Board“ bezeichneten Platinen jeweils ein Design abdecken.

### Erweiterungen beim Full-Service-Anbieter

Flux.AI [10] nimmt insofern eine Sonderstellung ein, als dass das Unternehmen als einziges einen Full-Service-EDA-Dienst anbietet. Im Prinzip haben wir den Dienst schon in [1] im Detail besprochen; allerdings sind seither einige Erweiterungen hinzugekommen, die wir hier stichwortartig vorstellen.

Das erste herausragende Feature versteckt sich schon im Flux-Startmenü. **Bild 11** zeigt, dass der Dienst zur automatischen Ermittlung von Bauteil-Footprints aus PDF-Datenblättern befähigt sein soll. Wie oben testete der Autor das Systemverhalten auch hier anhand eines Quectel-4G-Moduls. Das Hochladen der rund 2 MB großen Datei führte zunächst zum „Fortschritts-Dialog“ in **Bild 12** und - trotz der Nutzung einer kostenlosen Basis-Version - nach rund fünf Minuten zum in **Bild 13** gezeigten Ergebnis. Offensichtlich ist der PDF-Analysator in der Lage, die beiden im Datenblatt erhaltenen Komponenten sowohl zu erfassen als auch zu analysieren.

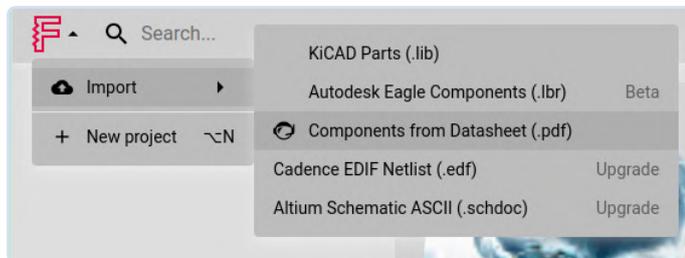


Bild 11. Diese Funktion spart die eine oder andere Mannstunde.

### Vergleichsweise teuer!

Ein kostenloser Account stellt pro Monat nur 50 Credits zur Verfügung. Das Quectel-Funkmodul verbrauchte 31 Credits und der eigentliche Import den Rest.

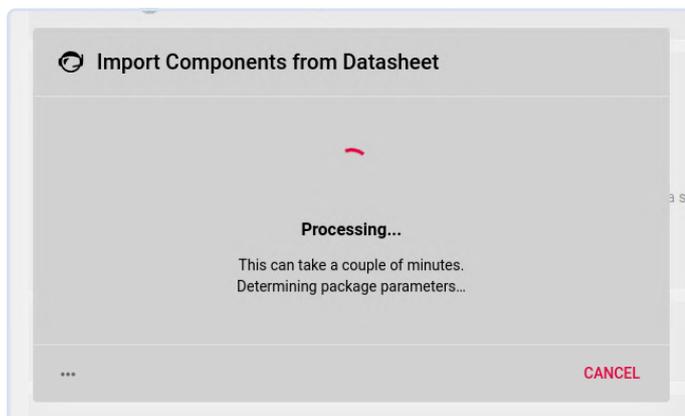


Bild 12. Nach dem Hochladen der PDF-Datei ist Warten angesagt.

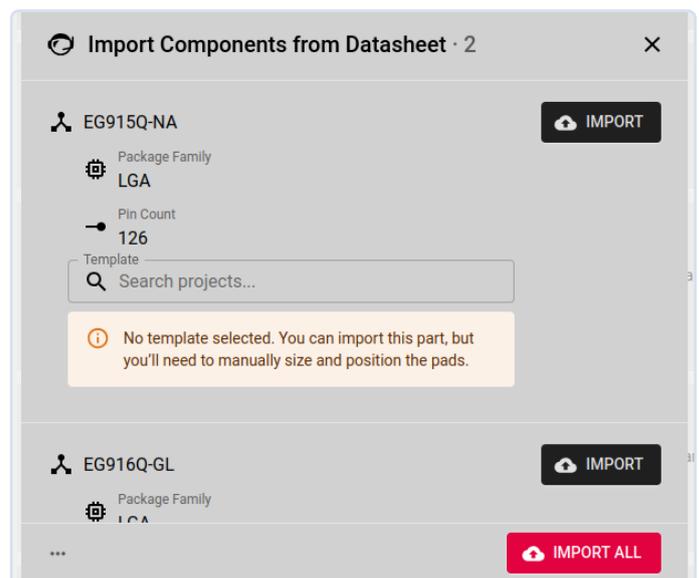


Bild 13. Dieses Analyse-Ergebnis ist durchaus beeindruckend.



Bild 14. Das Symbol im Schaltplan wirkt wenig inspiriert...

Nach dem eigentlichen Import machte sich aber Ernüchterung bereit. Wie **Bild 14** zeigt, war das System nicht in der Lage, den „echten“ Footprint zu ermitteln. Auch im Bereich der Schaltpläne - siehe hier beispielsweise **Bild 15** - gibt es noch Spielraum nach oben.

Ein weiteres Fokus-Feature ist die Fertigungs-Optimierung: Darunter versteht das System die Analyse der Bill of Materials, um Wege zur Kosten-Senkung zu finden. Ein Feature, das nach Ansicht des Autors im Allgemeinen gut funktioniert, ist das Streamlining passiver Komponenten: Das System sucht nach ähnlichen Bauteilen und prüft anhand von KI-Heuristiken, ob sich diese zwecks Vereinfachung der BOM und Senkung der Fertigungskosten auf eine SKU konsolidieren lassen.

Mitunter weniger vernünftig fallen die Vorschläge im Bereich der Komponentenoptimierung aus - wer in seinem Unternehmen beispielsweise umfangreiches geistiges Eigentum für STMicroelectronics verwendet, wird von der in **Bild 16** gezeigten Empfehlung nicht unbedingt begeistert sein. Außerdem generiert das System auf Wunsch verschiedene Designdokumente. Neben einem (meist überladenen) Blockdiagramm mit den einzelnen Komponenten hilft Flux bei der Erzeugung einer Bringup-Strategie, die beispielsweise einem Auftragsfertiger beim Überprüfen der gefertigten Baugruppen unter die Arme greift.

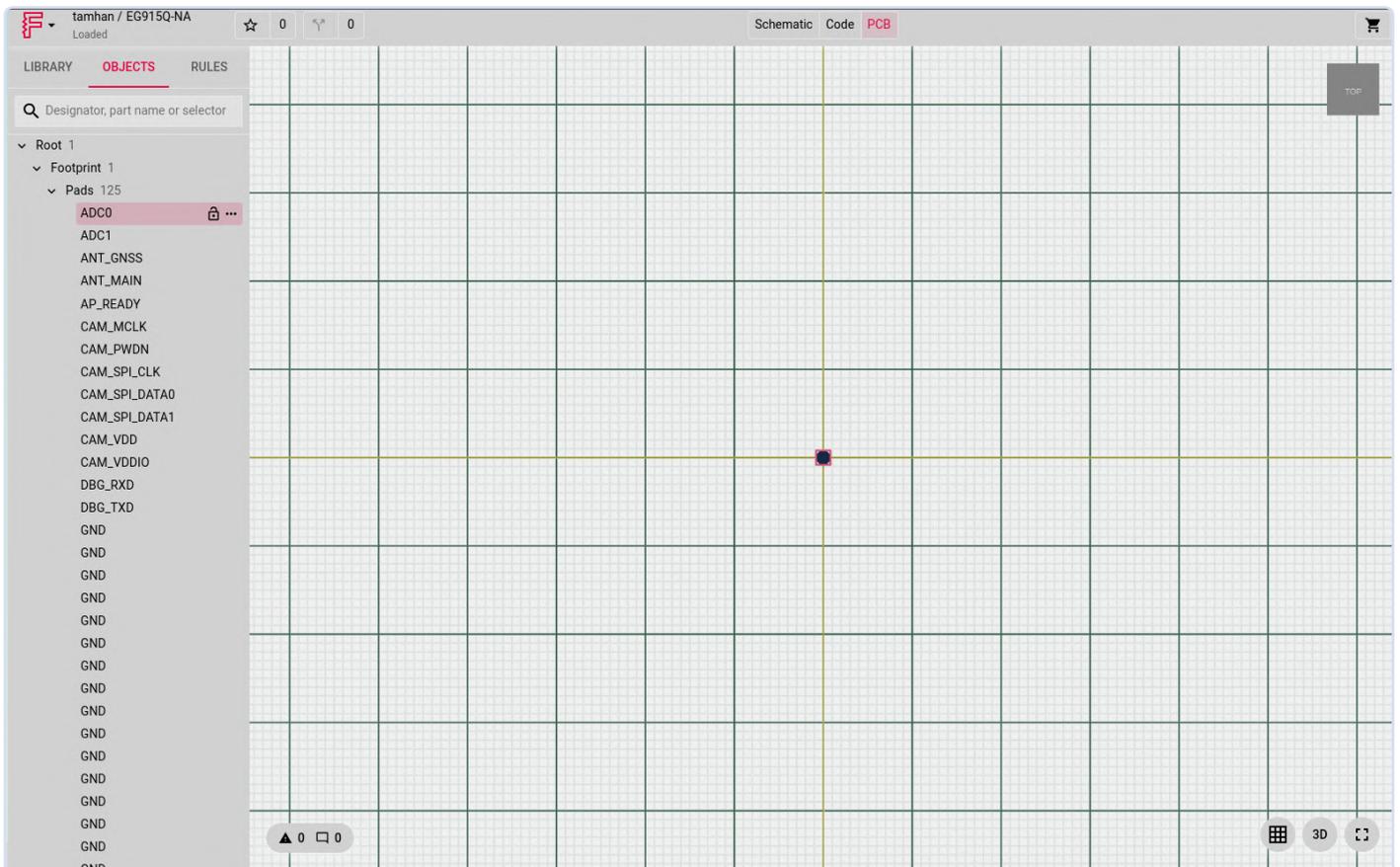


Bild 15. ...während der Footprint eine Aufforderung zum Selbermachen darstellt.

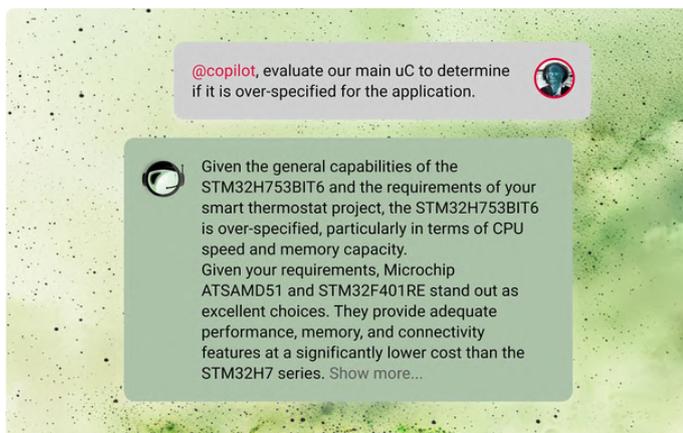


Bild 16. Leider gibt es auch in der Elektronik keine Silberkugeln.

Zu guter Letzt integriert sich Flux.AI in den Software-Entwicklungsprozess von Projekten, die auf der MicroPython-Runtime basieren. Neben dem Export von Unterstützungs-Dokumenten ist das System über sein LLM auch in der Lage, Firmware-Code im Copilot-Fenster nach Maßgabe der Möglichkeiten auf Korrektheit zu überprüfen.

Neben Performance-Optimierungen dürfte der Fokus hier vor allem auf falsch konfigurierter Hardware liegen - ein Klassiker wäre zum Beispiel das Auswählen einer fehlerhaften I<sup>2</sup>C-Adresse für einen Thermosensor.

### Noch Zukunftsmusik: Der „Elektroniker im PC“

Es dürfte in kaum einer Industrie so einfach sein, an „Others People Money“ zu kommen. Wer einem Risikoanleger ein auf KI basierendes System vorliegt, hat in den meisten Fällen bald einen Scheck in der Post. Die hier gezeigte Verbreiterung des Markts - neben dem Platzhirsch Flux.AI streben Dutzende andere Unternehmen auf den Markt - zeigt, dass Bewegung im Bereich der künstlichen Intelligenz für Elektroniker zu erwarten ist. Die relevante Frage ist dabei allerdings die Praxisrelevanz, denn vom Elektroniker in der Workstation, der komplett fertige Designs liefert und kein Review braucht, sind die Systeme zumindest nach Ansicht des Autors noch

weit entfernt. Andererseits gilt, dass die verschiedenen Generatoren und Fehlersuch-Systeme schon jetzt wertvolle Arbeitszeit einsparen, wenn man sie korrekt, vernünftig und nicht übereifrig in den Entwicklungsprozess einbindet. ◀

RG — 240451-02



### Über den Autor

Ingenieur Tam Hanna befasst sich seit mehr als 20 Jahren mit Elektronik, Computern und Software; er ist freiberuflicher Entwickler, Buchautor und Journalist ([www.instagram.com/tam.hanna](https://www.instagram.com/tam.hanna)). In seiner Freizeit beschäftigt sich Tam unter anderem mit 3D-Druck und dem Vertrieb von Zigarren.

### Sie haben Fragen oder Kommentare?

Wenn Sie technische Fragen oder Anmerkungen zu diesem Artikel haben, nehmen Sie bitte Kontakt auf mit dem Autor ([tamhan@tamoggemon.com](mailto:tamhan@tamoggemon.com)) oder der Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



### Passendes Produkt

- ▶ Peter Dalmaris, **KiCAD Like A Pro (Bundle)**, (Elektor 4. Auflage)  
Taschenbuch, englisch: [www.elektor.de/20942](https://www.elektor.de/20942)  
E-Buch, PDF, englisch: [www.elektor.de/20943](https://www.elektor.de/20943)



### WEBLINKS

- [1] Tam Hanna, „KI im Elektroniklabor“, Elektor 11-12/2023: <https://www.elektormagazine.de/magazine/elektor-317/62319>
- [2] USB-Kabeltester: [https://github.com/JITx-Inc/jitx-cookbook/tree/main/usb\\_c\\_cable\\_tester](https://github.com/JITx-Inc/jitx-cookbook/tree/main/usb_c_cable_tester)
- [3] JITX Versionen und Kosten: <https://www.jitx.com/plans>
- [4] Warteliste von SnapMagic: <https://www.snapmagic.com/#join>
- [5] Circuit Mind: <https://www.circuitmind.io/>
- [6] Was sind CUBOs?: <https://www.celus.io/cubos>
- [7] Celus: <https://www.celus.io/>
- [8] Projektbericht (PDF): <https://www.elektormagazine.de/240451-02>
- [9] Einstieg in Cady: <https://cady-solutions.com/how-to-use-cady/>
- [10] Flux.AI: <https://www.flux.ai/>

# Einführung in KI-Algorithmen

Prompt: Welche Algorithmen werden in KI-Tools verwendet?

Von Stuart Cording (Deutschland)

In der heutigen Welt spielen Künstliche Intelligenz und ihre Algorithmen eine zentrale Rolle in vielen Anwendungen. In diesem Artikel werfen wir einen Blick auf die Kernkonzepte und Algorithmen der modernen KI, auf neuronale Netze, maschinelles Lernen und Deep Learning. Ein tieferes Verständnis dieser Algorithmen eröffnet ein breites Spektrum an Möglichkeiten für den Einsatz von KI, weit über generative Modelle wie ChatGPT hinaus.

Künstliche Intelligenz taucht in fast jeder Nachrichtensendung auf. Manchmal geht es darum, einen innovativen neuen Ansatz im Gesundheitswesen hervorzuheben, der Krankheiten wie Krebs frühzeitig diagnostiziert und die Überlebenschancen verbessert, doch meistens handelt es sich um schlechte Nachrichten, zum Beispiel die Verbreitung gefälschter Bilder (**Bild 1**) und Deep-Fake-Videos oder die ethischen Fragen im Zusammenhang mit der Verwendung von urheberrechtlich geschütztem

Material zum Trainieren neuer KI-Modelle. Über die Algorithmen hinter solchen KI-Tools, die Aufbereitung der Daten und den Einsatz in Anwendungen wird für Ingenieure und technisch Interessierte, also außerhalb der Schwerpunktbereiche der Mainstream-Medien wenig gesagt.

## KI 101

Wer sich heute mit Künstlicher Intelligenz beschäftigt, stößt schnell auf verschiedene Begriffe, die einer Erläuterung bedürfen: Neuronale Netze, Maschinelles Lernen, Deep Learning und der Begriff Künstliche Intelligenz (KI) selbst. KI ist

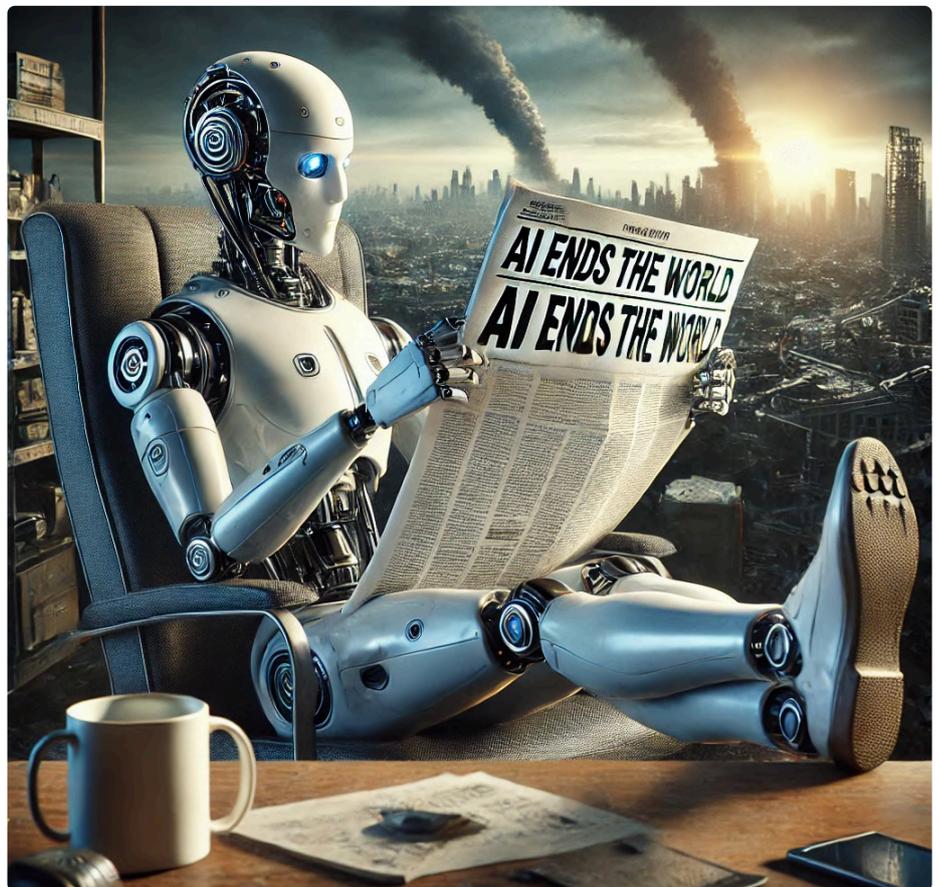


Bild 1. Die Sicht der generativen KI auf das Ende der Welt. Aber keine Sorge: Der Roboter hat zwei Beine, aber nur einen Fuß, sodass er nicht weit kommt!

ein Wissenschaftsbereich, der sich damit beschäftigt, wie Computer und Software komplexe Probleme lösen können. Dabei handelt es sich nicht nur um eine theoretische Disziplin, sondern auch um eine Reihe von Techniken, die in vielen Bereichen Anwendung finden, zum Beispiel bei der Datenanalyse, der Verarbeitung natürlicher Sprache und bei Vorhersagen.

Sie kann in Schwache KI (Narrow AI) und Allgemeine KI (General AI) unterteilt werden. Schwache Intelligenz ist so konzipiert, dass sie in einem bestimmten Bereich, zum Beispiel bei (Computer-) Spielen wie Schach oder Go, auf menschlichem Niveau arbeitet. Viele KI-Implementierungen sind dem Menschen ebenbürtig oder sogar überlegen, etwa bei der Klassifizierung von Objekten in Bildern. Doch auch wenn solche KI-Tools außerordentlich klug und fähig sind, würde man sie nicht mit einem Menschen verwechseln.

Allgemeine KI ist das angestrebte Ideal der KI-Forschung. Diese Form der KI wäre so leistungsfähig, dass ihre Fähigkeiten nicht mehr von denen eines Menschen zu unterscheiden wären. Sie könnte das Verständnis der Welt, das Menschen besitzen, nachahmen und sogar emotionale Reaktionen zeigen. Doch ob wir als Gesellschaft eine solche Entwicklung überhaupt wollen, bleibt eine offene Frage. Theoretisch wäre es sogar möglich, eine Superintelligente KI zu schaffen, die weit über das hinausgeht, was Menschen als Intelligenz begreifen können. Diese Superintelligenz könnte nicht nur menschliche Fähigkeiten imitieren, sondern sie in vielen Bereichen übertreffen.

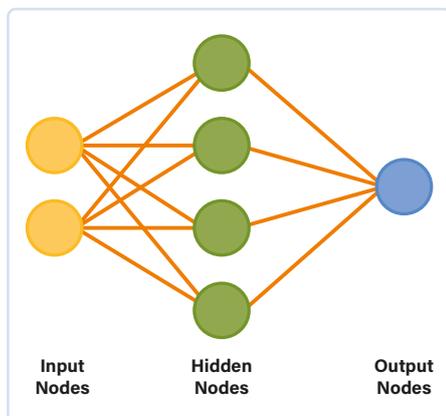


Bild 2. Das Multilayer-Perzeptron besteht aus Eingabeknoten, versteckten Knoten und Ausgabeknoten, die Klassifizierungsergebnisse liefern. Die versteckten Knoten erlernen die gewünschte Klassifizierung durch Training.

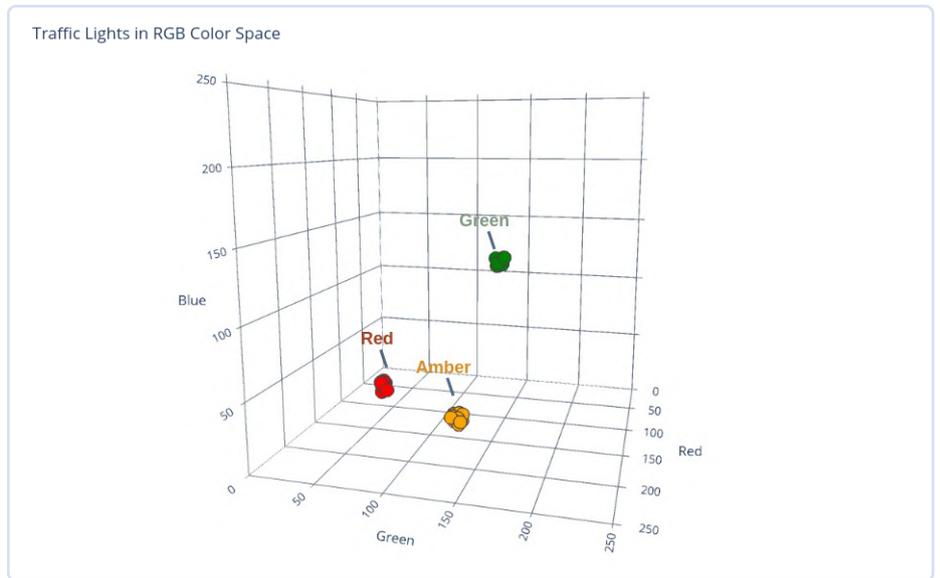


Bild 3. Cluster von RGB-Werten können als Farben interpretiert und klassifiziert werden.

Allerdings bleibt dies vorerst ein theoretisches Konzept, das eher in den Bereich der Philosophie und Science-Fiction gehört, als dass es in naher Zukunft realisiert wird. Um die zugrundeliegende Implementierung aller KI-Technologien zu verstehen, müssen wir die unterste Ebene erforschen - neuronale Netze. Das mehrschichtige oder Multilayer-Perzeptron (MLP), das biologischen Neuronen nachempfunden ist, nutzt Mathematik und einen Prozess namens Backpropagation, um das Neuron allmählich auf die Erkennung von Mustern einzustellen [1]. Entscheidend ist dabei die Verwendung einer Schicht versteckter Zellen zwischen Eingangs- und Ausgangsschicht, die den Abstimmungsprozess ermöglicht (Bild 2). So kann das Neuron Eingaben klassifizieren, die mit den einfachen if-else-Anweisungen einer Programmiersprache nur schwer oder gar nicht zu realisieren wären.

Ein gutes Beispiel ist die Klassifizierung von Farben, die von einem RGB-Sensor erfasst werden. Orange wird zu 100 % als rot und zu 65 % als grün angesehen [2]. Violett wird mit 93 % rot, 51 % grün und 93 % blau beschrieben. Wenn der Sensor jedoch vor ein tatsächlich existierendes Farbmuster gehalten wird, liefert er keine perfekten Prozentwerte, die eindeutig als Farben klassifiziert werden könnten. Stattdessen schwanken die erfassten Werte um die Idealwerte herum. Die Beleuchtung macht die Erkennung noch komplizierter, da sich alle drei Werte bei schlechter Beleuchtung nach unten und bei starker Beleuchtung nach oben verschieben. Neuronale Netze können so trainiert werden, dass sie vernünftige Variationen und unterschiedliche Beleuchtungssituationen berücksich-

tigen und eine Clusterbildung von Werten (Bild 3) als die gewünschte Objektklassifizierung akzeptieren [3]. Dadurch sind sie in der Lage, präzise und robuste Klassifizierungen durchzuführen, die weit über die Möglichkeiten einfacher regelbasierter Ansätze hinausgehen.

### Maschinelles Lernen vs. Deep Learning

Die meisten der heutigen KI-Tools fallen unter die Kategorie des Maschinellen Lernens (ML), das eine Untergruppe der Künstlichen Intelligenz ist. Diese Tools werden in der Regel auf strukturierten Daten trainiert, die Daten (wie Fotos) sind mit Beschriftungen (Labels) versehen, die erklären, was darauf zu sehen ist. Diese Beschriftungen können so spezifisch sein wie die explizite Benennung von Lebensmitteln auf einem Bild (Gurke, Pizza, Brot) oder zusätzlichen Kontext wie das Wachstumsstadium eines Gemüses, den Ort einer Mahlzeit oder die Kalorienzahl des abgebildeten Essens liefern.

Es ist wichtig zu verstehen, dass ML-Algorithmen, ähnlich wie Menschen, nicht



Bild 4. Das RAM-1-System nutzt maschinelles Lernen zur Erkennung von Überspannungsereignissen in Stromnetzen. Über das IoT kann das Erkennungsmodell kontinuierlich aktualisiert werden. (Quelle: Izoelktro)

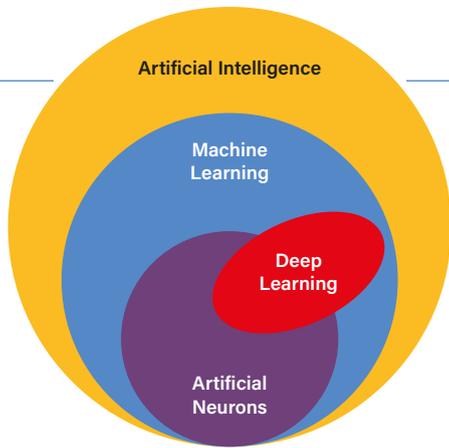


Bild 5. Die Hierarchie von KI zu künstlichen Neuronen. Beim Deep Learning werden Netzwerke mit vielen Schichten verwendet, um hoch entwickelte KI-Modelle wie zur Verarbeitung natürlicher Sprache zu implementieren.

perfekt sind und keine endgültigen Antworten geben. Stattdessen liefern sie eine Wahrscheinlichkeit, mit der die Eingabe als etwas erkannt wird, auf das sie trainiert wurde. Nehmen wir als Beispiel einen ML-Algorithmus, der auf Tierbilder trainiert wurde. Wenn ihm ein Bild einer Kuh gezeigt wird, sollte er mit einer Wahrscheinlichkeit von über 90 % erkennen, dass es sich um eine Kuh handelt. Er könnte jedoch auch sagen, dass es sich mit 65 %iger Wahrscheinlichkeit um einen Pandabär und mit 35 %iger Wahrscheinlichkeit um einen Dachs handelt. Wenn Sie jedoch ein Bild eines Pinguins zeigen, und der Algorithmus nie auf Pinguine trainiert wurde, wird er „Pinguin“ als mögliche Klassifizierung nicht vorschlagen.

Die meisten eingebetteten oder Edge-KI-Anwendungen basieren auf ML-Implementierungen. Ein Beispiel dafür ist das RAM-1 von Izoelkro [4]. Dieses Gerät überwacht Überspannungsableiter zur Kontrolle des Stromnetzes [5]. Es verwendet einen nRF9160 von Nordic Semiconductor, der auf einen ARM Cortex-M33 basiert, und meldet Leckströme, überhöhte Temperaturen, die Neigung von Masten und andere manipulative Ereignisse wie Blitzeinschläge. Die anfängliche Entwicklung des Algorithmus erfolgte mit der Edge-Impulse-Plattform für maschinelles Lernen [6], und das System lernt und verbessert sich kontinuierlich anhand neuer anomaler Ereignisse aus dem Netzwerk der installierten Geräte (siehe Bild 4).

Deep Learning (DL) geht noch einen Schritt weiter. Bei unserem MLP reichte eine einzige Schicht mit Gewichten zwischen Eingabe und Ausgabe aus, um den Algorithmus zu optimieren. Komplexere Aufgaben erfordern jedoch eine feinere Abstimmung,

was zu mehr versteckten Schichten führt. Sobald zwei oder mehr versteckte Schichten verwendet werden, spricht man von einer DL-Implementierung (Bild 5).

### Ansätze zum Lernen

Beschriftete Daten sind am einfachsten zu verarbeiten, stellen jedoch ein großes Problem dar: Wer beschriftet die Daten? Plattformen wie Amazon Mechanical Turk ermöglichen es, große Datensätze zu kennzeichnen [7] oder die Leistung bestehender KI-Modelle zu testen. Einzelne Beschriftungsaufgaben kosten in der Regel nur wenige Cent, sodass die Kennzeichnung für viele Unternehmen wirtschaftlich rentabel ist. Wenn Ihr Datensatz jedoch sensible oder vertrauliche Informationen enthält, die nicht öffentlich zugänglich gemacht werden dürfen, müssen Sie eine geeignete Alternative finden.

Beschriftete Daten können für überwachtes Lernen genutzt werden. Dabei wird dem KI-Modell eine Reihe von Eingaben (Text, Ton, Bild) und den entsprechenden erwarteten Ausgaben bereitgestellt, um durch wiederholtes Training die Genauig-

keit zu verbessern. Das Training gilt als abgeschlossen, sobald ein gewünschtes Genauigkeitsniveau erreicht ist. Wie beim Menschen wird diese Genauigkeit jedoch nie 100 % betragen.

Ein KI-Modell, das die Trainingsdaten perfekt analysiert und jede Antwort korrekt trifft, leidet oft unter Überanpassung (Overfitting). Zum Beispiel: Wenn das Modell Verkehrsampeln auf Bildern erkennen soll, könnte es aufgrund von Anomalien in den Trainingsbildern – etwa Sonnenlicht, das von Kunststoffflächen reflektiert wird, oder das Vorhandensein einer Schranke – fälschlicherweise auf eine Ampel schließen. Wenn das Modell später mit Bildern konfrontiert wird, die nicht Teil der Trainingsgruppe waren, kann das bloße Vorhandensein eines Portals oder eines hellen Flecks dazu führen, dass es fälschlicherweise eine Ampel erkennt, obwohl gar keine vorhanden ist. Daher müssen Entwickler den optimalen Kompromiss zwischen Über- und Unteranpassung finden (siehe Bild 6).

Viele Unternehmen verfügen über eine große Menge unbeschrifteter Daten.

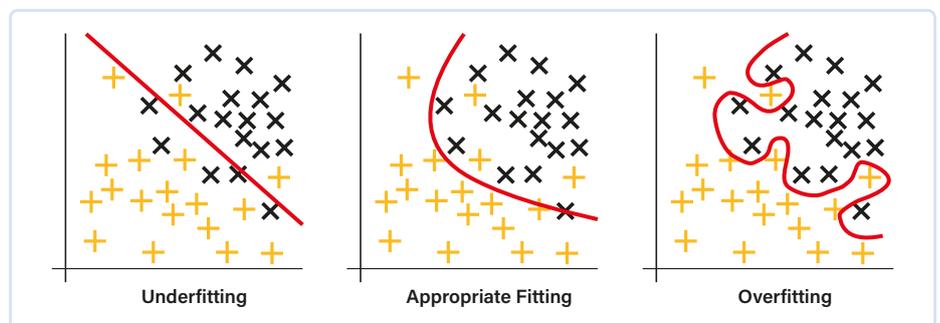


Bild 6. Ein zu langes Training führt zu Überanpassung, ein zu kurzes zu einer schlechten Klassifizierung der Eingabedaten.

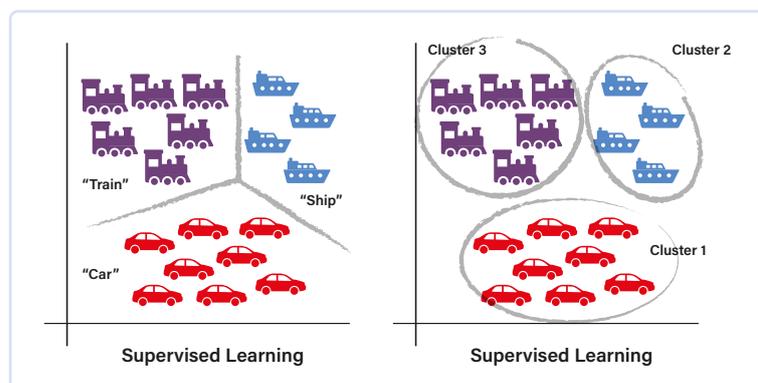


Bild 7. Beim überwachtem Lernen werden beschriftete Daten verwendet. Ohne diese Labels gruppiert der Algorithmus Eingabedaten mit ähnlichen Merkmalen automatisch.

So kann beispielsweise ein Testfahrzeug während einer Fahrt große Mengen an Sensor-Rohdaten sammeln. Eine Excel-Tabelle (ja, das ist immer noch ein gängiges Datenformat!) enthält zwar die Namen der Sensoren und deren Maßeinheiten, jedoch keinen Kontext, wie beispielsweise die Information, dass das Fahrzeug fünf erwachsene Passagiere befördert hat oder dass die Straßen nass waren.

Unüberwachtes Lernen kann hier Abhilfe schaffen. Diese Aufgabe ist zwar offensichtlich anspruchsvoller und rechenintensiver, ermöglicht jedoch, die Dimensionalität der Daten zu reduzieren und das Erkennen von Mustern und Clustern. Um auf das obige Beispiel zurückzukommen: Das System könnte Kurven, Beschleunigungs- und Verzögerungsmuster klassifizieren und sogar Anzeichen für Regen oder Schnee während der Testfahrt erkennen. Natürlich wird die KI Begriffe wie Regen oder Schnee nicht direkt vorschlagen. Stattdessen erkennen Experten bei der Analyse der Ergebnisse, dass bestimmte Cluster mit bestimmten Testbedingungen übereinstimmen, und kennzeichnen sie entsprechend (**Bild 7**).

Die beiden oben genannten Methoden lassen sich kombinieren, indem beschriftete und unbeschriftete Daten in einem halbüberwachten Lernansatz genutzt werden. Die Trainingsergebnisse können dann zu den Trainingsdaten hinzugefügt werden. Da diese Ergebnisse jedoch fehlerhaft sein könnten, ist eine menschliche Überprüfung und Intervention unabdingbar.

Ein weiterer Ansatz ist das verstärkende Lernen. Hier wird ein Modell in simulierten Umgebungen mit Beobachtungen zu seiner Umgebung versorgt. Die KI wählt mögliche Aktionen aus und erhält eine „Belohnung“, wenn sie sich dem gewünschten Ziel nähert. Über viele Iterationen hinweg lernt das Modell schließlich die gewünschte Aufgabe. Diese Methode wurde erfolgreich in der Verarbeitung natürlicher Sprache und bei KI-Systemen, die Spiele spielen, eingesetzt. Google DeepMind nutzte diesen Ansatz, um die Anzahl der Assembler-Anweisungen zu reduzieren, die für die Implementierung eines Sortieralgorithmus erforderlich sind (**Bild 8**). Für korrekt funktionierenden Code, der weniger Assembler-Anweisungen benötigte als der ursprüngliche Algorithmus, wurde eine Belohnung vergeben [8].

Original Sort Code	Improved Sort Code		
Memory[0] = A Memory[1] = B Memory[2] = C	Memory[0] = A Memory[1] = B Memory[2] = C	}	
mov Memory[0] P mov Memory[1] Q mov Memory[2] R	mov Memory[0] P mov Memory[1] Q mov Memory[2] R		Collect three values for sorting. Save to registers.
mov R S cmp P R cmovg P R cmovl P S mov S P cmp S Q cmovg Q P cmovg S Q	mov R S cmp P R cmovg P R cmovl P S  cmp S Q cmovg Q P cmovg S Q		Sort values.
mov P Memory[0] mov Q Memory[1] mov R Memory[2]	mov P Memory[0] mov Q Memory[1] mov R Memory[2]	Return the result to memory.	

Bild 8. Optimierung eines Sortieralgorithmus: Eine Assembler-Anweisung (mov S P) wurde eingespart, ohne die Funktionalität zu beeinträchtigen.

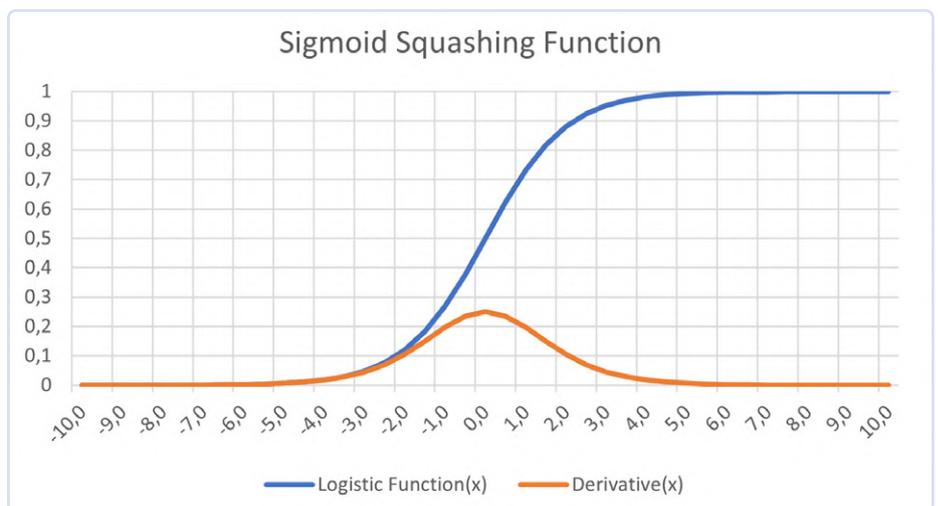


Bild 9. Die Sigmoidfunktion unterstützt die Klassifizierung, indem sie Eingabewerte schnell in wahrscheinliche oder unwahrscheinliche Ergebnisse unterteilt.

### Die KI-Algorithmen

Grundsätzlich lassen sich KI-Modelle in zwei Haupttypen unterteilen. Der erste Typ sind Klassifizierungsmodelle. Diese Modelle zerlegen die Eingabe in eine wahrscheinliche oder unwahrscheinliche Antwort auf Fragen wie „Enthält dieses Bild eine Ente?“ oder „Ist diese Farbe gelb?“. Regressionsmodelle hingegen liefern numerische Antworten auf Fragen wie „Wie viele Arduino Unos werden wir nächsten Monat verkaufen?“ Komplexere Modelle ermöglichen die Entwicklung von Tools wie ChatGPT, das sich mit Nutzern unterhalten kann, oder DALL-E, das Bilder aus Texteingaben generiert. All diese Systeme basieren auf einer Vielzahl von Algorithmen, die oft kombiniert werden, um den spezifischen Anforderungen gerecht zu werden.

### Logistische Regression

Frühe Versuche, künstliche Neuronen einzusetzen, stießen auf ein Problem. Sie

waren gut darin, Informationen zu trennen, die mit einer geraden Linie in zwei Gruppen aufgeteilt werden konnten. Das lag daran, dass sie lineare Funktionen verwendeten, also beispielsweise  $f(x) = ax + b$ , wenn Sie sich an Ihre Schulmathematik erinnern. Wenn jedoch die Daten so verteilt waren, dass sie in einer Form gruppiert waren, die eher einem Kreis ähnelte, versagten diese linearen Modelle.

Die logistische Regression verwendet Funktionen, die die Eingabewerte schnell in die Nähe von 0 oder 1 bringen. Ein Beispiel hierfür ist die Sigmoidfunktion, definiert als  $s(z) = 1 / (1 + \exp(-z))$ , die bei einer Eingabe von 0,0 einen Wert von 0,5, bei einer Eingabe von 10 einen Wert nahe bei 1 und bei einer Eingabe von -10,0 einen Wert nahe bei 0 liefert (**Bild 9**). Dank dieser mathematischen Funktion können neuronale Netze durch viele Wiederholungen für relativ einfache Klassifizierungsaufgaben trainiert werden, wie zum Beispiel die Frage,

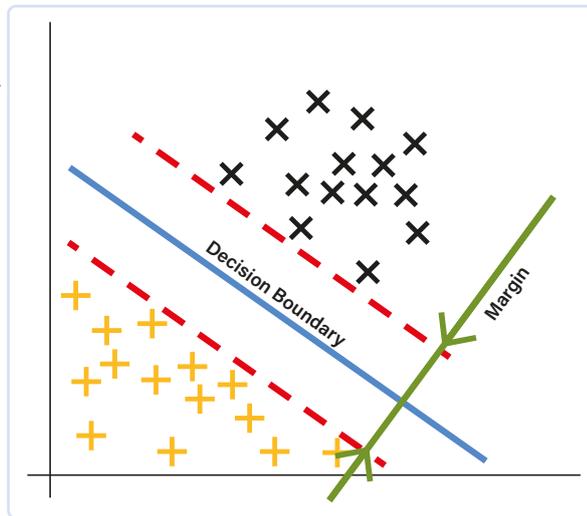
ob ein RGB-Sensor vor der Farbe Lila steht (wahrscheinlich/unwahrscheinlich).

## Support-Vektor-Maschinen

Stellen Sie sich vor, Sie werfen einige rote Bonbons auf Ihren Teppich, gefolgt von einigen blauen Bonbons. Wahrscheinlich könnten Sie ein Lineal zwischen die beiden Farbgruppen legen, das sie trennt. Dieses Lineal kann als Vektor (eine Linie mit Betrag und Richtung) dargestellt werden. Dies ist die grundlegende Funktionsweise von Support-Vektor-Maschinen (SVMs). In einem 2D-Raum wird dieses imaginäre Lineal jedoch so breit wie möglich gemacht, um den Abstand zwischen den beiden Gruppen zu maximieren, während der Vektor berechnet wird (siehe **Bild 10**). In einem KI-Modell angewendet, kann dieser Vektor verwendet werden, zwischen roten und blauen Bonbons zu unterscheiden. Auf dieser oberflächlichen Ebene unterscheidet sie sich nicht von einer linearen Funktion. SVMs entfalten jedoch ihr volles Potenzial, wenn die Daten mehr als zwei Dimensionen haben. Angenommen, Ihre Daten lassen sich auf einer X-, Y- und Z-Achse darstellen: Dann suchen Sie nicht nach einer Linie zur Trennung der Klassen, sondern nach einer Hyperebene (eine zweidimensionale Ebene im dreidimensionalen Raum). SVMs sind auch nützlich, wenn einige der roten und blauen Bonbons in die jeweils andere Gruppe geraten sind. Sie sind noch rechenintensiver als logistische Regressionen, aber auch noch besser geeignet für Datensätze mit hoher Dimensionalität und für unstrukturierten Daten.

## Entscheidungsbäume und Zufallsforste

Entscheidungsbäume (Decision Trees) lassen sich auf einfache Weise mit der Erstellung von Flussdiagrammen vergleichen. Sie folgen einer „Teile-und-herrsche“-Strategie, um einen Datensatz in so viele Gruppen wie nötig zu klassifizieren. Wenn ein Benutzer den KI-Algorithmus anwendet, arbeitet sich dieser durch das Flussdiagramm, um eine Antwort zu finden. Bei der Analyse von Bildern könnte der Entscheidungsbaum beispielsweise damit beginnen, den Datensatz in Hintergrundmerkmale, dann in auffällige Farben und schließlich in spezifische Objektmerkmale (wie Ohren, Pfoten oder Streifen) zu unterteilen. Die Entscheidung, wie der Datensatz an



*Bild 10. Bei Support-Vektor-Maschinen (SVMs) werden die Daten auf einer der beiden Seiten eines als Vektor definierten Entscheidungsgrenze liegend klassifiziert, wobei eine Spanne (Margin, Abstand zwischen den nächstgelegenen gegenüberliegenden Gruppen) angegeben wird.*

jedem Knoten aufgeteilt wird, basiert auf der Entropie der betrachteten Stichproben. Wenn viele Bilder rot sind und nur wenige blau, ist die Entropie für die Eigenschaft „enthält Rot“ hoch. Durch Hinzufügen eines Knotens mit der Frage „Enthält das Bild Rot?“ kann das Modell Bilder schnell und effizient klassifizieren. Wenn der Datensatz bei einer „Rot/Blau“-Frage jedoch gleichmäßig (50/50) aufgeteilt ist, sucht das Modell nach einer alternativen Klassifizierung mit höherer Entropie.

Ein häufiges Problem bei Entscheidungsbäumen ist die Überanpassung und Verzerrung. Die Überanpassung wird dadurch bekämpft, dass während des Lernens Knoten mit unwichtigen Merkmalen beschnitten werden. Overfitting und Verzerrungen können mit dem Random-Forest-Algorithmus angegangen werden, der zusätzlich Bagging (Bootstrap-Aggregation) verwendet, um die Varianz des Datensatzes und die Zufälligkeit zu reduzieren.

## Faltende neuronale Netze

Faltende neuronale Netze (Convolutional Neural Networks, CNNs) werden häufig für Anwendungen im Bereich des maschinellen Sehens und der Objekterkennung eingesetzt. Diese Deep-Learning-Algorithmen bestehen aus mehreren Schichten und beginnen mit Faltungsschichten, die Merkmale wie gerade Linien, Kurven und Farben in Bildern identifizieren. Darauf folgen Pooling-Schichten, die die Ausgabe der Faltungsschichten verdichten und dadurch die Datenmenge reduzieren. Ein typisches CNN umfasst viele Faltungs- und Pooling-Schichten, deren Zahl je nach Komplexität der Klassifizierungsaufgabe variieren. Die letzte Schicht des Netzwerks ist die vollständig verbundene Schicht, die schließlich die Klassifizierung des Bildes liefert.

Eine besondere Stärke von CNNs ist ihre

Robustheit gegenüber Veränderungen der Position des Objekts im Bild. Das bedeutet, dass sie auch dann zuverlässig arbeiten, wenn sich das gesuchte Objekt in den Testbildern gegenüber den Trainingsdaten verschoben hat.

## Rekurrente neuronale Netze und Transformatoren

Bilder sind statische Momentaufnahmen, aber viele Anwendungen wie bei Sprache und Text erfordern ein Verständnis für zeitliche Abläufe. Nehmen wir diesen Satz: „John hat einen Verstärker gebaut. Er hat ihn auf Antrieb zum Laufen gebracht.“ Hier erkennen wir, dass sich „er“ auf John und „ihn“ auf den Verstärker bezieht. Rekurrente neuronale Netze (RNNs) sind besonders gut darin, solche zeitlichen Zusammenhänge zu verstehen. Sie werden häufig in der Verarbeitung natürlicher Sprache, bei Übersetzungen und in der Handschrifterkennung eingesetzt. Anders als viele neuronale Netze, die Eingabedaten einfach an die nächste Schicht weiterleiten, nutzen RNNs ein Gedächtnis, das Informationen aus früheren Schritten speichert und zukünftige Eingaben berücksichtigt, um fundiertere Entscheidungen zu treffen.

Transformatoren gehen noch einen Schritt weiter und haben so die Entwicklung großer Sprachmodelle wie ChatGPT ermöglicht. Sie verwenden Aufmerksamkeit-Mechanismen (Attention Mechanism), die besonders effektiv sind, um Beziehungen zwischen Wörtern über Satzgrenzen hinweg zu analysieren, statt sich nur auf Zusammenhänge innerhalb eines einzelnen Satzes zu konzentrieren.

## Es geht nicht nur um ChatGPT und Midjourney

Künstliche Intelligenz hat bedeutende Fortschritte erzielt, die unsere Herangehensweise an verschiedene Aufgaben



### Über den Autor

Stuart Cording ist Ingenieur und Journalist mit mehr als 25 Jahren Erfahrung in der Elektronikbranche. Er ist auf Videoinhalte spezialisiert und legt dabei den Fokus auf technische Details und tiefgehende Einblicke. Besonders interessiert ihn die Technologie selbst, ihre Integration in Endanwendungen sowie Prognosen zu zukünftigen Entwicklungen. Viele seiner aktuellen Artikel finden Sie unter [www.elektormagazine.de/cording](http://www.elektormagazine.de/cording).

### Sie haben Fragen oder Kommentare?

Wenn Sie technische Fragen oder Anmerkungen zu diesem Artikel haben, wenden Sie sich bitte an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).

grundlegend verändern. In der Medizin verbessert KI die Analyse von medizinischen Scans und beschleunigt die Entwicklung neuer Arzneimittel. In Unternehmen können Netzwerkzugriffe protokolliert, Cyberangriffe erkannt, Bedrohungen vorhergesagt und sogar Quellcode effizient analysiert werden. In der Welt der eingebetteten Systeme kann maschinelles Lernen komplexe analoge Filter ersetzen und energieintensive digitale Filterprozesse optimieren. Dennoch wird KI in der Öffentlichkeit oft nur auf generative Videos und vermeintliche Fehler wie ChatGPTs Schwierigkeiten beim Zählen von Buchstaben reduziert [9]. Und ja, es gibt ethische Bedenken und Risiken, die ernst genommen werden müssen, aber wir sollten das Kind nicht gleich mit dem Bade ausschütten.

Wie jede Technologie kann auch Künstliche Intelligenz sowohl für positive als auch für negative Zwecke eingesetzt werden. Wir sollten mehr Zeit damit zu verbringen, die Funktionsweise der Technologie im Detail zu verstehen, wie wir es hier getan haben, und sie auf Herausforderungen anzuwenden, die für Menschen mühsam oder komplex sind oder bei denen Menschen häufig Fehler machen. Das Internet ist voll von Tutorials, Anleitungen und Beispielen, die zeigen, wie man eigene KI-Modelle erstellt oder sie trainiert, um Datensätze zu bearbeiten, die mit anderen Methoden schwer zu bewältigen sind. Vielleicht ist jetzt der richtige Zeitpunkt, sich an KI zu versuchen und jene Herausforderung anzugehen, die Sie schon länger im Kopf haben. ◀

SE — 240564-02

### WEBLINKS

- [1] S. Cording, „Die Neuronen in neuronalen Netzen verstehen: Künstliche Neuronen“, [elektormagazine.de](http://elektormagazine.de), Mai 2021: <https://t1p.de/uqj8v>
- [2] RGB in Prozent für orange: <https://www.farb-tabelle.de/en/rgb2hex.htm?q=orange>
- [3] S. Cording, „Die Neuronen in neuronalen Netzen verstehen: Praktische Neuronen“, [elektormagazine.de](http://elektormagazine.de), Juni 2021: <https://t1p.de/6hplh>
- [4] RAM-1 von Izoelektro: <https://www.ram-center.com/>
- [5] Edge Impulse, „The New AI- Powered Grid“ (PDF): <https://t1p.de/937ye>
- [6] Edge Impulse: <https://edgeimpulse.com/>
- [7] B. Kopp, „Getting Started with Using Amazon Mechanical Turk to Label Data“, U.S. Bureau of Labor Statistics, September 2019 (PDF): <https://t1p.de/4lezw>
- [8] D. J. Mankowitz, A. Michi, „AlphaDev discovers faster sorting algorithms“, Google DeepMind, Juni 2023: <https://t1p.de/7mknb>
- [9] Falsche Zählung des Buchstabens „r“ im Wort „Strawberry“: <https://t1p.de/s8be3>

# Treten Sie unserer Community bei



[www.elektormagazine.de/community](http://www.elektormagazine.de/community)

**elektor**  
design > share > earn

# Einplatinencomputer für KI-Projekte

## Überblick und Hintergründiges

Von Saad Imtiaz (Elektor)

Es gibt unzählige Einplatinencomputer, aber (oder deshalb) kann die Auswahl des richtigen für KI-Anwendungen schwierig sein. In diesem Artikel gehen wir auf die wichtigsten Faktoren ein, die bei der Auswahl eines Einplatinencomputers für KI zu berücksichtigen sind, und stellen eine Liste der leistungsfähigsten Plattformen vor. So finden Sie die richtige Hardware für Ihre KI-Projekte!

Künstliche Intelligenz hat sich in vielerlei Hinsicht von einem futuristischen Konzept zu einem Teil unseres täglichen Lebens entwickelt. Ob Spracherkennung auf unseren Smartphones, vorausschauende Algorithmen in unseren Streaming-Diensten, verbesserte Diagnostik im Gesundheitswesen oder prädiktive Wartung und Qualitätskontrolle in der Fertigung - KI ist überall. Im Zentrum der Entwicklung von KI steht der Bedarf an leistungsfähiger und effizienter Hardware, die den immensen Rechenanforderungen der Algorithmen des maschinellen Lernens (ML) gewachsen ist. Die Auswahl der richtigen Hardware ist für jedes KI-Projekt entscheidend, da sie sich direkt auf die Effizienz, Skalierbarkeit und den Erfolg der Anwendung auswirkt.

### Was wird für die Entwicklung von KI benötigt?

Für die Entwicklung von KI-Anwendungen ist Hardware erforderlich, die eine große Anzahl von Berechnungen gleichzeitig ausführen kann. Dies liegt daran, dass KI-Algorithmen, insbesondere Deep-Learning-Modelle, große Datensätze und komplexe Berechnungen umfassen, die schnell und präzise verarbeitet werden müssen. Zu den wichtigsten Hardwareanforderungen für die Entwicklung von KI gehören eine Hochleistungs-CPU, ein leistungs-

starker Grafikprozessor für die Parallelverarbeitung, ausreichend Arbeitsspeicher (RAM) und zunehmend eine dedizierte neurale Verarbeitungseinheit (Neural Processing Unit, NPU).

NPUs wurden speziell zur Beschleunigung der Berechnung neuronaler Netze entwickelt und bieten gegenüber herkömmlichen CPUs und GPUs einen erheblichen Leistungsschub. Mit der Weiterentwicklung der KI werden NPUs zu einer Standardausstattung moderner KI-Hardwareplattformen und bieten spezielle Verarbeitungsfunktionen, die für die Bewältigung von KI-Aufgaben entscheidend sind.

### Die richtigen Hardware für Ihre KI/ML-Projekte

Die Auswahl der richtigen Hardware für die KI-Entwicklung hängt von mehreren Faktoren ab, darunter die Komplexität der KI-Modelle, die Größe der Datensätze, der Bedarf an Echtzeitverarbeitung und die Leistungseinschränkungen des Projekts. Bei einem Projekt, das sich auf Edge-KI konzentriert - bei dem also die KI-Verarbeitung lokal auf einem Gerät stattfindet - können beispielsweise Energieeffizienz und kompakte Formfaktoren wichtiger sein als die reine Rechenleistung. Auf der anderen Seite können Cloud-basierte KI-Anwendungen von leistungsstarken GPUs und NPUs profitieren, die in den Rechenzentren die umfangreichen Berechnungen durchführen. Bei der Auswahl der Hardware ist es außerdem wichtig, die Kompatibilität mit KI-Frameworks wie TensorFlow, PyTorch und TensorRT sowie die Verfügbarkeit von Entwicklungstools und die Unterstützung einer Community zu berücksichtigen. Ziel ist es, eine Plattform zu finden, die nicht nur die technischen Anforderungen des Projekts erfüllt, sondern auch eine einfache Entwicklung und Bereitstellung ermöglicht.

### Neueste Entwicklungen in der KI-Hardware

In letzter Zeit hat sich die Entwicklung der KI-Hardware deutlich verändert. Unternehmen nutzen neben

CPUs und GPUs zunehmend NPU, die die Verarbeitung von KI-Workloads optimieren, da NPUs auf die spezifischen Arten von Operationen wie Matrix-Multiplikationen und Faltungen zugeschnitten sind, wie sie bei neuronalen Netzen nötig sind (siehe Textkasten **NPUs verstehen**).

## Hardware-Plattformen für Ihre KI-Entwicklung

Da sich das Gebiet der KI immer weiter ausbreitet, ist auch die Vielfalt der Hardware-Plattformen, die Entwicklern zur Verfügung steht, exponentiell gewachsen. Bei so vielen Optionen ist es wichtig zu wissen, dass jede Plattform in einigen Bereichen eine hervorragende Leistung erbringt, in anderen jedoch möglicherweise nicht. Die Wahl der richtigen Hardware hängt von Ihren spezifischen Projektanforderungen ab, unabhängig davon, ob Sie ein Einsteiger in die künstliche Intelligenz sind oder ein Experte, der komplexe Aufgaben bewältigen will.

Dieser Artikel kann aufgrund der großen Anzahl verfügbarer Einplatinencomputer (Single Board Computer, SBCs) nicht allumfassend sein. Wer sich auf die Suche nach der perfekten Lösung für seine Bedürfnisse begibt, findet jedoch Webseiten [1][2], die es ermöglichen, SBCs nach Ihren spezifischen Anforderungen zu filtern und auszuwählen. Nachfolgend finden Sie deshalb nur (?) eine Auswahl an Hardware-Plattformen, die sich für eine Reihe von KI- und ML-Projekten eignen, von Computer Vision und natürlicher Sprachverarbeitung bis hin zu Robotik und Edge Computing. Diese Boards wurden aufgrund ihrer Fähigkeit ausgewählt, verschiedene Anwendungen zu unterstützen, und bieten eine solide Grundlage für die Entwicklung von KI auf allen Stufen der Expertise.

**Tabelle 1** zeigt einen Vergleich dieser SBCs, einschließlich einiger anderer bemerkenswerter Boards, die hier nicht im Detail besprochen werden, so dass Sie sehen können, wie sie in Bezug auf Leistung, Funktionen und Eignung für verschiedene KI-Anwendungen abschneiden.

## Nvidia Jetson Orin Nano

Wenn man über KI-Plattformen spricht, ist Nvidia der erste Name, der einem in den Sinn kommt. Seit Jahren ist Nvidia führend im Bereich der KI-Hardware und setzt mit leistungsstarken GPUs und spezialisierten NPUs immer wieder neue Maßstäbe. Die Reise begann mit GPUs, die für Spiele entwickelt wurden und aufgrund ihrer unübertroffenen parallelen Verarbeitungsfähigkeiten schnell zum Rückgrat der KI-Forschung wurden. Die 2016 eingeführte Jetson-Plattform von Nvidia revolutionierte KI und Robotik, indem sie leistungsstarke KI-Verarbeitung „an die Edge“ brachte - wo sie nun in autonomen Fahrzeugen, Industrie-

## NPUs verstehen: Das Gehirn hinter der KI-Hardware

Eine neurale Verarbeitungseinheit (Neural Processing Unit, NPU) ist ein spezialisierter Prozessor, der für die spezifischen Berechnungen entwickelt wurde, die von KI-Modellen benötigt werden. Dadurch arbeiten sie für diese Aufgaben wesentlich effizienter als herkömmliche CPUs und GPUs. KI-Modelle, insbesondere neuronale Netze, stützen sich in hohem Maße auf Matrixmultiplikationen, Faltungen und Aktivierungsfunktionen, die die parallele Verarbeitung großer Datenmengen erfordern.

So ist beispielsweise die Matrixmultiplikation das Herzstück neuronaler Netze, bei der Eingabedaten mit Gewichtsmatrizen über mehrere Schichten hinweg multipliziert werden. NPUs zeichnen sich dadurch aus, dass sie diese Operationen gleichzeitig über viele Verarbeitungseinheiten ausführen, was die Berechnungen erheblich beschleunigt. In ähnlicher Weise werden in faltenden neuronalen Netzwerken (Convolutional Neural Networks, CNN) durch Faltungen und Pooling-Operationen Merkmale aus Bildern extrahiert - Aufgaben, die NPUs aufgrund ihrer Parallelverarbeitungsfähigkeiten mit hoher Effizienz erledigen. Darüber hinaus sind NPUs für die sich wiederholenden, stark parallelen Operationen optimiert, so dass sie im Vergleich zu Universal-CPU oder sogar GPUs Daten schneller und mit geringerem Energiebedarf verarbeiten können. Ihre Leistung wird häufig in TOPS (Tera Operations Per Second) gemessen, was auf ihre Fähigkeit hinweist, Billionen von Operationen pro Sekunde durchzuführen - ein entscheidender Faktor für die Geschwindigkeit und Effizienz von KI-Anwendungen. Durch die Konzentration auf diese speziellen Berechnungsarten ermöglichen NPUs eine schnellere und energieeffizientere Verarbeitung, was sie für moderne KI-Hardware unverzichtbar macht.

robotern und unzähligen anderen Anwendungen zum Einsatz kommt.

Doch so beeindruckend diese industrietauglichen Plattformen auch sind, nicht jedes Projekt benötigt einen Supercomputer. Manchmal möchte man einfach klein anfangen - zum Beispiel ein KI-Modell trainieren, das die eigene Katze von der des Nachbarn unterscheiden kann. Werfen wir also einen Blick auf den Nvidia Jetson Orin Nano, ein leistungsstarker und dennoch erschwinglicher Ausgangspunkt für Ihre Reise in die KI-Entwicklung, die zu fortgeschrittenen KI-Projekten führt.

Der Nvidia Jetson Orin Nano [3] ist eine kompakte und leistungsstarke KI-Plattform, die sich hervorragend für eine Vielzahl von KI- und Machine-Learning-Projekten eignet (**Bild 1**). Mit einer ARM-Cortex-A78AE-CPU mit sechs Kernen und bis zu 8 GB LPDDR5-RAM bietet



Bild 1. Nvidia Jetson Orin Nano (Quelle: Nvidia)

Tabelle 1: Vergleich der Einplatinencomputer

	Jetson Orion Nano 4 GB	Raspberry Pi 5	BeagleBone AI-64	Rock 5C
<b>KI-Beschleuniger</b>	20 TOPS	11 TOPS (nur mit AI- kit)	8 TOPS	6 TOPS
<b>PCIe</b>	1x4 + 3x1 (PCIe Gen3, Root Port und Endpoint)	1 Lane Gen 2	4x2 Lane Gen 2	1x1 Lane Gen 2.1
<b>SoC</b>	6-Core Arm Cortex-A78AE v8.2 64-bit CPU 1,5MB L2 + 4MB L3 1,5 GHz	Broadcom BCM2712	Texas Instruments TDA4VM SoC	Rockchip RK3588S2
<b>Interner Speicher</b>	4 GB 64-bit LPDDR5 - 34 GB/s	4 GB LPDDR4X	4 GB LPDDR4	4 GB LPDDR4X
<b>Externer Speicher</b>	(unterstützt externes NVMe)	-	16 GB eMMC	-
<b>Display</b>	1x 4K30 Multi-Mode-DP 1.2 (+MST)/eDP 1.4/HDMI 1.4	Dual 4k60	Mini-Displayport 4k30 4 Lane MIPI-DSI	1x HDMI 2.1 8Kp60 1x MIPI-DSI 1080p60
<b>GPIO</b>	40-poliger GPIO-Header	40-poliger GPIO-Header	40-poliger GPIO-Header	40-poliger GPIO-Header
<b>Mobilfunk</b>	-	-	-	-
<b>WLAN</b>	-	802.11ac 2.4/5GHz	-	WiFi 6
<b>BLE</b>	-	5.2	-	5.4
<b>USB-Ports</b>	3x USB 3.2 Gen2 (10 Gbps) 3x USB 2.0	2x USB-C 2x USB-A Eingebauter Hub	2x USB-A + 1x USB-C	2x USB-A -2.0 + 1x USB-A 3.0 + 1x USB-A 3.0 OTG
<b>Kamera</b>	Bis zu 4 Kameras (8 über virtuelle Kanäle) 8 Lanes MIPI CSI-2 D-PHY 2.1 (bis zu 20 Gbps)	2x 4-Lane MIPI-DSI	2x 4-Lane MIPI-DSI	1x 4-Lane MIPI-CSI oder 2x 2-Lane MIPI-CSI
<b>Ethernet</b>	Gigabit Ethernet-Port	Gigabit Ethernet-Port	Gigabit Ethernet-Port	Gigabit Ethernet-Port

dieser SBC eine KI-Leistung von bis zu 40 TOPS (Tera Operations Per Second); perfekt für Aufgaben wie Deep Learning, Computer Vision und Robotik.

Der Orin Nano ist eingebettet in das umfangreiche KI-Ökosystem von Nvidia, einschließlich CUDA, cuDNN und TensorRT, unterstützt wichtige Frameworks wie TensorFlow und PyTorch und bietet so Flexibilität für verschiedene Entwicklungsanforderungen.

Er ist mit Hochgeschwindigkeits-I/O-Optionen wie USB 3.1, PCIe und Gigabit-Ethernet ausgestattet, und das alles bei gleichbleibender Energieeffizienz, wodurch er sich für Edge-KI-Anwendungen eignet. Ganz gleich, ob Sie Prozesse automatisieren, intelligente Geräte bauen oder in die Robotik eintauchen wollen, der Jetson Orin Nano bietet die robuste Leistung und Vielseitigkeit, die Sie benötigen, um Ihre KI-Projekte zum Leben zu erwecken. Seine Kombination aus Leistung und Zugänglichkeit macht ihn zu einer ausgezeichneten Wahl sowohl für Anfänger als auch für erfahrene Entwickler.

### Raspberry Pi 5

Ein Raspberry Pi ist aufgrund seiner Erschwinglichkeit, Vielseitigkeit und der starken Unterstützung durch die Community für fast jedes Projekt die erste Wahl. Der Raspberry Pi 5 [4] in **Bild 2** setzt diese Tradition fort und bietet eine Quad-Core ARM-Cortex-A76-CPU, bis zu 8 GB RAM und Kompatibilität mit einer Vielzahl von KI-Frameworks wie TensorFlow Lite, OpenCV und PyTorch. Damit eignet er sich für eine Vielzahl von

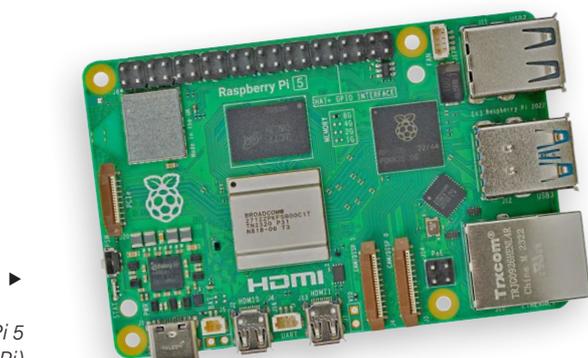


Bild 2. Raspberry Pi 5  
(Quelle: Raspberry Pi)

Tachyon	Google Coral	Orange Pi 5	ASUS Tinker Board 2S	Alta AML-A311D-CC
12 TOPS	4 TOPS (int8); 2 TOPS pro Watt	6 TOPS	2,3 TOPS	5 TOPS
2 Lanes Gen 3	-	PCIe-Slot 3.0	PCIe Gen3 x2	Not specified
1 x Gold Plus @ 2,7 GHz +3xGold@2,4GHz +4xSilver@1,9GHz	NxP i.Mx 8M SoC (Quad Cortex-A53, Cortex-M4F) @ 1,5GHz	Rockchip RK3588S	Dual 2.0 GHz Cortex-A72 und Quad 1.5 GHz Cortex-A53	Amlogic A311D SoC
4 GB LPDDR4X	1 oder 4 GB LPDDR4	2 GB/4 GB DDR4	2 GB/4 GB LPDDR4	4 GB LPDDR4
64 GB	8 GB eMMC	Optionales eMMC-Modul	16 GB eMMC-Speicher	
Single 4K60 neben 4-Lane MIPI-DSI	39-poliger FFC-Verbinder für MIPI-DSI-Display (4-Lane) HDMI 2.0a (voll belegt)	HDMI 2.0	HDMI 2.0, MIPI-DSI	HDMI 2.0, 4K
40-poliger GPIO-Header	40-poliger GPIO-Header	40-poliger GPIO-Header	40-poliger GPIO-Header	40-poliger GPIO-Header
Sub-GHz 5G - CAT-18				
802.11ax 2.4/5/6GHz (WiFi 6)	Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5GHz)	802.11 a/b/g/n/ac	802.11 a/b/g/n/ac	
5.2	Bluetooth 4.2	Bluetooth 5.0	Bluetooth 5.0	
2x USB-C (1x USB-C 3.1 mit integriertem Display-Port)	Type-C OTG; Type-C Power; Type-A-Host 3.0 ; Micro-B Serielle Konsole	USB 3.0, USB 2.0	1x USB 3.2 Gen1 Type-C (OTG), 3x USB 3.2 Gen1 Type-A, 1x USB 2.0 Type-A	4x USB 3.0
2x 4-Lane MIPI-CSI	4-pin FFC-Verbinder für MIPI-CSI2-Kamera (4-Lane)	MIPI-CSI-Interface	MIPI-CSI-Interface	Nicht angegeben
-	Gigabit Ethernet-Port	Gigabit Ethernet	Gigabit Ethernet	Gigabit Ethernet

Aufgaben, einschließlich grundlegender Computer-Vision- und Machine-Learning-Projekte.

Der Raspberry Pi 5 ist zwar leistungsstark, aber nicht für intensive KI-Anwendungen optimiert. Für anspruchsvollere Aufgaben ist das Raspberry Pi AI-Kit [5] mit dem KI-Beschleuniger Hailo 8L eine echte Alternative (siehe **Bild 3**). Der Hailo 8L fügt 13 TOPS an KI-Verarbeitungsleistung hinzu und ermöglicht es dem Raspberry Pi, komplexe KI-Aufgaben wie Bilderkennung in Echtzeit und Deep Learning effizienter zu bewältigen.

Zusammenfassend lässt sich sagen, dass der Raspberry Pi 5 ein hervorragender Ausgangspunkt für die KI-Entwicklung ist, insbesondere in Kombination mit dem Hailo 8L für eine verbesserte Leistung, was ihn zu einer vielseitigen und kostengünstigen Wahl für Einsteiger und fortgeschrittene Entwickler macht.

### Orange Pi 5

Der Orange Pi 5 und der Orange Pi 5 Pro [6] werden für KI und Allzweck-Computing immer beliebter und bieten eine leistungsstarke und dennoch erschwing-



Bild 3. Raspberry Pi AI Kit (Quelle: Raspberry Pi)

liche Alternative zu anderen Einplatinencomputern. Beide Modelle werden von dem RK3588S-Prozessor von Rockchip angetrieben, der eine Octa-Core-Konfiguration mit vier ARM-Cortex-A76-Kernen und vier Cortex-A55 Kernen aufweist. Diese Konfiguration bietet eine robuste Verarbeitungsleistung, die durch

Bild 4. Orange Pi 5  
(Quelle: Orange Pi)

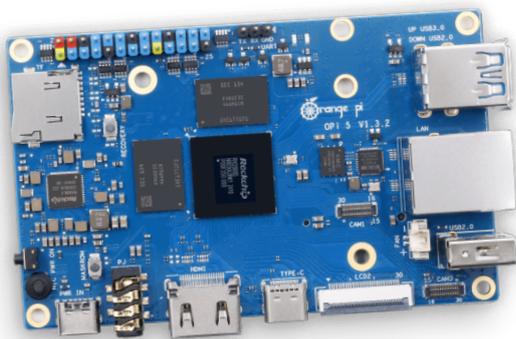
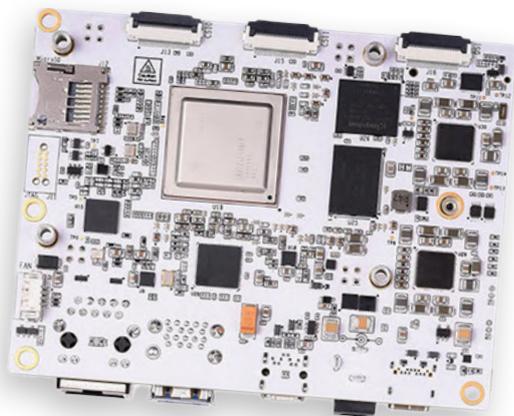


Bild 5. BeagleBone AI-64  
(Quelle: Texas Instruments)



bis zu 16 GB Arbeitsspeicher ergänzt wird, wodurch sich die Boards für ressourcenintensive Anwendungen in der KI und dem maschinellen Lernen eignen. Ein wichtiges Merkmal der Reihe der Orange Pi 5 (**Bild 4**) ist die integrierte NPU, die bis zu 6 TOPS an KI-Leistung bietet. Diese NPU steigert die Fähigkeit des Boards erheblich, KI-Aufgaben wie Objekterkennung, Bildklassifizierung und Verarbeitung natürlicher Sprache zu bewältigen, und macht es zu einem starken Konkurrenten für Projekte, die KI-Verarbeitung in Echtzeit erfordern.

Sowohl der Orange Pi 5 als auch die 5-Pro-Version unterstützen eine breite Palette von KI-Frameworks, darunter TensorFlow, PyTorch und OpenCV. Diese breite Kompatibilität ermöglicht es Entwicklern, diese Boards für verschiedene KI-Anwendungen zu verwenden, von grundlegender Computer Vision bis hin zu komplexeren Machine-Learning-Modellen. Der Grafikprozessor Mali-G610 MP4 steigert die Fähigkeiten bei der Bewältigung grafikintensiver Aufgaben weiter. Ein wesentlicher Unterschied zwischen Pi 5 und 5 Pro besteht darin, dass nur die Pro-Version über WLAN und Bluetooth verfügt.

Der Orange Pi 5 Pro bietet mit seinem optional größeren RAM und erweiterten Funktionen zusätzliche Leistung und Flexibilität und ist damit ideal für

anspruchsvollere Projekte. Er verfügt über vielfältige Anschlussmöglichkeiten wie HDMI 2.1, USB 3.0 und PCIe 3.0, die die Integration von Hochgeschwindigkeits-Peripheriegeräten erlauben und die Einsatzmöglichkeiten in der KI-Entwicklung erweitern.

### BeagleBone AI-64

Der BeagleBone AI-64 [7] ist ein leistungsfähiger, auf KI fokussierter Einplatinencomputer, der die Lücke zwischen eingebetteten KI-Projekten und Echtzeit-Computing-Anforderungen schließen soll (**Bild 5**). Das Board wird vom TDA4VM-SoC von Texas Instruments angetrieben und verfügt über eine Dual-Core ARM-Cortex-A72-CPU, die mit 2,0 GHz getaktet wird und eine robuste Leistung für eine Vielzahl von Aufgaben gewährleistet. Das AI-64-Board ist mit einem KI-Beschleuniger ausgestattet, der 8 TOPS liefern kann und damit für anspruchsvolle KI- und Machine-Learning-Anwendungen bestens geeignet ist.

Mit 4 GB LPDDR4-RAM und 16 GB eMMC-Speicher onboard bietet der BeagleBone AI-64 ausreichend Speicher für die meisten Embedded-KI-Projekte. Das Board unterstützt Yocto Linux, ein flexibles und anpassbares Betriebssystem, das sich für Entwickler eignet, die ihre Betriebssystemumgebung an spezifische Anwendungsanforderungen anpassen möchten. Zu den verfügbaren Anschlüssen zählen duale USB-A-Ports, USB-C und Gigabit-Ethernet, wodurch sich das Board problemlos in verschiedene Netzwerkumgebungen integrieren lässt. Darüber hinaus besitzt das Board einen 40-poligen GPIO-Header, der Flexibilität für den Anschluss einer Vielzahl von Peripheriegeräten und Sensoren bietet.

Die Echtzeituhr und die USB-PD-Stromversorgung des BeagleBone AI-64 erhöhen seine Eignung für industrielle und IoT-Anwendungen, bei denen ein präzises Timing und eine zuverlässige Stromversorgung entscheidend sind. Das Fehlen von integriertem WLAN und BLE kann jedoch je nach Projektanforderungen zusätzliche Module für eine drahtlose Konnektivität erforderlich machen. Insgesamt ist das BeagleBone AI-64 eine gute Wahl für Entwickler, die fortschrittliche KI-Lösungen in Embedded-Umgebungen implementieren möchten, insbesondere wenn Echtzeitverarbeitung und robuste Hardware-Integration entscheidend sind.

### BeagleY-AI

BeagleY-AI [8] ist eine weitere überzeugende Option in der BeagleBoard-Familie, insbesondere für Bildverarbeitungs- und DSP-intensive Anwendungen, wie in **Bild 6** dargestellt. Das Board wird vom Bildverarbeitungsprozessor AM67A von Texas Instruments angetrieben, der eine Quad-Core ARM-Cortex-A53-CPU mit 1,4 GHz und zwei C7x-DSPs (Digitale

Signal-Prozessoren) mit Matrix-Multiply-Accelerator (MMA) umfasst, die 4 TOPS an KI-Leistung liefern können. Der BeagleY-AI eignet sich besonders gut für Aufgaben, die Echtzeit-Bildverarbeitung, Computer Vision und andere KI-Workloads erfordern, die von spezialisierter DSP-Verarbeitung profitieren.

BeagleY-AI ist mit 4 GB LPDDR4-Speicher ausgestattet, der eine reibungslose Leistung bei anspruchsvollen Anwendungen gewährleistet. Darüber hinaus verfügt es über mehrere Hochgeschwindigkeitsschnittstellen wie PCIe Gen3, USB 3.1 und Gigabit Ethernet mit PoE+, die hervorragende Konnektivität für datenintensive Aufgaben bieten. Das Board unterstützt drei simultane Displays über microHDMI, MIPI-DSI und OLDI-Schnittstellen (LVDS) und ist damit ideal für Multi-Display-Setups in KI-gesteuerten visuellen Anwendungen.

Mit Wi-Fi 6 und Bluetooth 5.4 bietet der BeagleY-AI eine fortschrittliche Drahtlos-Konnektivität, was einen erheblichen Vorteil gegenüber dem BeagleBone-AI-64 darstellt. Dadurch eignet sich das Board besonders gut für Projekte, die eine robuste Netzwerkleistung und eine Kommunikation mit geringer Latenz erfordern, zum Beispiel bei intelligenten Kameras, autonomen Robotern und anderen Edge-Geräten.

BeagleY-AI unterstützt außerdem eine breite Palette von I/O-Optionen, darunter MIPI-CSI für Kameras, besitzt einen 40-poligen Erweiterungsstecker und mehrere USB-Ports, was das Board zu einer vielseitigen Plattform für Entwickler macht. Darüber hinaus erleichtern ein JTAG-Header und eine serielle UART-Konsole Entwicklern das Debuggen und Optimieren ihrer Anwendungen.

BeagleBone-AI-64 eignet sich hervorragend für industrielle KI-Anwendungen, die eine Echtzeitverarbeitung und eine robuste Hardware-Integration erfordern, während das BeagleY-AI mit seinen fortschrittlichen DSP-Fähigkeiten und seiner überlegenen Konnektivität besonders für KI-Projekte im Bereich der Bildverarbeitung geeignet ist. Beide Boards stellen leistungsstarke Plattformen für die KI-Entwicklung dar, wobei die Wahl zwischen ihnen von den spezifischen Anforderungen des Projekts abhängt.

## Tachyon

Tachyon ist ein aufregender neuer Einplatinencomputer, der von Particle [9] entwickelt wurde und sich derzeit in einer Kickstarter-Kampagne [10] befindet. Dieses Board könnte sich zu einem Meilenstein für tragbare und dezentrale Computer entwickeln, indem es den leistungsstarken Qualcomm-SoC Snapdragon mit 5G-Konnektivität, KI-Beschleunigung und umfangreichen I/O-Optionen kombiniert, und das alles in einem Raspberry-Pi-kompatiblen Formfaktor (**Bild 7**).

Im Tachyon arbeitet ein System-on-a-Chip namens

QCM6490 von Qualcomm, dem gleichen Chipsatz, der auch in mittel- bis hochpreisigen Smartphones zum Einsatz kommt. Dazu gehört eine Octa-Core-Cryo-CPU (1× Gold Plus mit 2,7 GHz, 3× Gold mit 2,4 GHz und 4× Silver mit 1,9 GHz) mit einer beeindruckenden Rechenleistung. Ergänzt wird die CPU durch einen Grafikprozessor Adreno 643 und einen DSP Hexagon 770 mit einem 12-TOPS-KI-Beschleuniger (alles von Qualcomm), die die Ausführung von fortgeschrittenen KI/ML-Modellen, Echtzeit-Objekterkennung und Video/Audio-Klassifizierungsaufgaben ermöglichen.

Tachyon ist mit bis zu 8 GB LPDDR4X-Speicher und 64 GB Onboard-Speicher ausgestattet und bietet somit ausreichend Platz und Geschwindigkeit für die meisten Anwendungen. Das Board unterstützt eine breite Palette von Betriebssystemen, darunter Ubuntu 24.04, Yocto, Android, Qualcomm Linux und sogar Windows 11, was Entwicklern eine beispiellose Flexibilität bietet. Die robuste Konnektivität mit zwei USB-C-Anschlüssen (darunter einer mit DisplayPort-Unterstützung), Wi-Fi 6, Bluetooth 5.2 und 5G-Mobilfunk im Sub-GHz-Bereich macht das Board ideal für Edge-Computing- und IoT-Anwendungen, die zuverlässige Hochgeschwindigkeitsverbindungen erfordern.

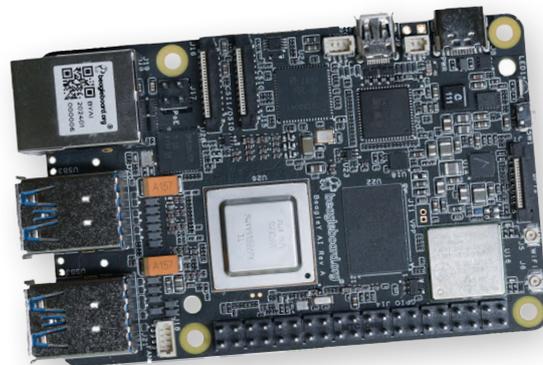


Bild 6. BeagleY AI (Quelle: Texas Instruments)



Bild 7. Tachyon (Quelle: Particle)



*Die Auswahl geeigneter KI-Hardware hängt von mehreren Faktoren ab, von der Komplexität und den Ressourcenanforderungen der KI-Modelle, der Menge und der Art der Datensätze, dem Bedarf an einer geringen Latenz oder in Echtzeit sowie der Leistung und Effizienz.*

Tachyon ist auf Vielseitigkeit ausgelegt und verfügt über einen 40-poligen GPIO-Anschluss, der mit Raspberry-Pi-HAT kompatibel ist, eine duale MIPI-CSI-Kameraschnittstelle und 4K-Display-Unterstützung über MIPI-DSI. Das Board ist für eine breite Palette von Aufgaben ausgelegt, von KI-Assistenten und Computer Vision bis hin zu Medien-Hosting und tragbaren Spielen. Mit integrierten Antennen für 5G und WLAN kann Tachyon in verschiedenen Umgebungen eingesetzt werden, ob zu Hause, im Feld oder in Unternehmensanwendungen.

Insgesamt ist Tachyon ein leistungsstarker und flexibler SBC, der Leistung auf Smartphone-Niveau in den Bereich Edge-Computing und IoT bringt. Er ist als „Do-it-all“-Platine konzipiert, die von persönlichen Projekten bis hin zu groß angelegten Implementierungen alles bewältigen kann. Damit ist er eine ausgezeichnete Wahl für Entwickler, die die Grenzen dessen, was mit einem SBC möglich ist, erweitern möchten.

### Rock 5C und 5C Lite

Rock 5C und Rock 5C Lite [11] sind leistungsstarke Single-Board-Computer von Radxa, die sich an Anwender richten, die ein ausgewogenes Verhältnis zwischen Leistung und Erschwinglichkeit suchen (**Bild 8**). Beide Boards basieren auf SoCs von Rockchip, wobei der Rock 5C mit dem leistungsstärkeren RK3588S2 und der Rock 5C Lite mit dem RK3582 ausgestattet ist.

Rock 5C zeichnet sich durch seine Octa-Core-CPU (vier ARM-Cortex-A76 und vier Cortex-A55), eine Mali-G610MP4-GPU und eine NPU aus, die bis zu 6 TOPS an KI-Leistung liefern kann. Mit dieser Ausstattung eignet es sich für Aufgaben von KI/ML-Projek-

ten bis hin zu Multimedia-Anwendungen mit Unterstützung für die Dekodierung von Videos bis zu 8K. Außerdem bietet es mit Wi-Fi 6, Bluetooth 5.4, mehreren USB-Anschlüssen und einem Gigabit-Ethernet-Anschluss zahlreiche Konnektivitätsoptionen. Das Board Rock 5C unterstützt bis zu 32 GB LPDDR4X-RAM, was es zu einer vielseitigen Plattform für eine Vielzahl von Anwendungsfällen macht.

Das Rock 5C Lite ist eine abgespeckte Version mit einer Dual-Core-Cortex-A76-CPU und einem Quad-Core-Cortex-A55, aber ohne GPU. Es enthält aber immer noch eine NPU mit 5 TOPS KI-Verarbeitungsleistung, was es zu einer soliden Wahl für KI-Projekte macht, die keine intensive grafische Verarbeitung erfordern. Die Lite-Version unterstützt dieselben Anschlussmöglichkeiten wie 5C, allerdings ohne GPU, und ist damit eine effiziente Wahl für KI-gesteuerte Anwendungen, bei denen Energieaufnahme und Kosten im Vordergrund stehen.

Beide Boards sind mit Steckplätzen für Speichererweiterungen (eMMC und microSD) ausgestattet und bieten zwei Display-Ausgänge mit HDMI 2.1 und MIPI-DSI-Schnittstellen. Sie erlauben außerdem eine externe Kühlung, was für die Aufrechterhaltung der Leistung bei hoher Belastung entscheidend ist. Das Board Rock 5C ist ideal für Anwender, die umfassende Multimedia- und KI-Fähigkeiten benötigen, während das Rock 5C Lite auf KI-fokussierte Anwendungen zugeschnitten ist, bei denen die GPU-Leistung weniger wichtig ist. Beide Boards sind eine ausgezeichnete Wahl, je nach den spezifischen Anforderungen Ihres Projekts.

### Viele weitere Optionen

Da künstliche Intelligenz immer mehr Einzug in unseren Alltag hält, vom Gesundheitswesen bis zur Fertigung, kann die Bedeutung der richtigen Hardware für die KI-Entwicklung nicht hoch genug eingeschätzt werden. Der Erfolg eines jeden KI-Projekts ist eng mit den Fähigkeiten der zugrundeliegenden Hardware verknüpft, die die intensiven Berechnungen, die für die Algorithmen des maschinellen Lernens erforderlich sind, effizient verarbeiten muss.

Die Wahl des richtigen Single-Board-Computers ist in diesem sich schnell entwickelnden Bereich von entscheidender Bedeutung. In diesem Artikel wurde eine Auswahl von vielseitigen und leistungsfähigen

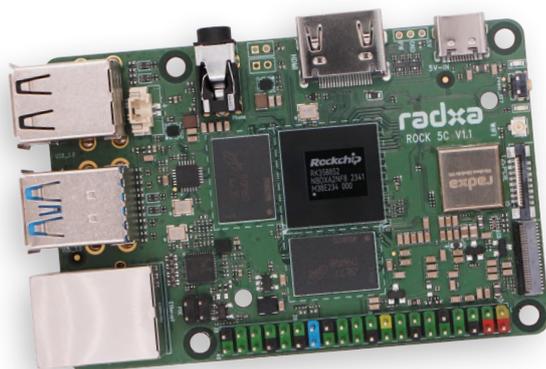
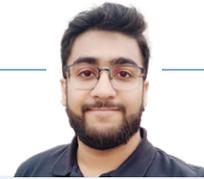


Bild 8. Rock 5C  
(Quelle: Radxa)



Einplatinencomputern vorgestellt, die sich für eine Reihe von KI-Anwendungen eignen - von maschineller Bildverarbeitung bis hin zur Robotik - aber es gibt noch viele weitere Optionen [12]. Da jede Plattform ihre eigenen Stärken hat, ist es wichtig, die Hardware an die spezifischen Projektanforderungen anzupassen. Denken Sie bei der weiteren Entwicklung von KI daran, dass die richtige Hardware nicht nur Ihre Projekte beschleunigt, sondern auch mit Ihren Ambitionen wächst, egal ob Sie gerade erst mit KI anfangen oder die Grenzen dessen ausreizen wollen, was mit KI alles möglich ist. ◀

SE — 240473-02

### Sie haben Fragen oder Kommentare?

Wenn Sie technische Fragen oder Anmerkungen zu diesem Artikel haben, wenden Sie sich bitte an den Autor unter [saad.imtiaz@elektor.de](mailto:saad.imtiaz@elektor.de) oder an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).

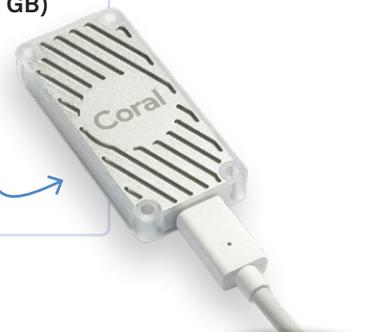
### Über den Autor

Saad Imtiaz, Senior Engineer bei Elektor, ist Mechatronikingenieur mit umfangreicher Erfahrung in Embedded Systems und Produktentwicklung. In seiner Karriere hat er mit einer Vielzahl von Unternehmen zusammengearbeitet, von innovativen Startups bis hin zu etablierten globalen Unternehmen, und dabei zukunftsweisende Prototypen- und Entwicklungsprojekte vorangetrieben. Mit seinem reichen Hintergrund, der auch eine Zeit in der Luftfahrtindustrie und die Leitung eines Technologie-Startups umfasst, bringt Saad eine einzigartige Mischung aus technischem Fachwissen und unternehmerischem Geist in seine Aufgaben bei Elektor ein. Hier trägt er zur Projektentwicklung sowohl in der Software- als auch in der Hardware-Entwicklung bei.



### Passende Produkte

- **Raspberry Pi 5 Ultimate Starter Kit (8 GB)**  
[www.elektor.de/20721](http://www.elektor.de/20721)
- **Raspberry Pi AI Kit**  
[www.elektor.de/20879](http://www.elektor.de/20879)
- **Google Coral USB Accelerator**  
[www.elektor.de/19366](http://www.elektor.de/19366)



### WEBLINKS

- [1] Datenbank von Einplatinencomputern: <https://hackerboards.com/>
- [2] Liste von Einplatinencomputern: [https://www.blackmagicboxes.com/?page\\_id=466](https://www.blackmagicboxes.com/?page_id=466)
- [3] NVIDIA Jetson Orin:  
<https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/>
- [4] Raspberry Pi 5: <https://www.raspberrypi.com/products/raspberry-pi-5/>
- [5] Raspberry Pi AI Kit: <https://www.raspberrypi.com/products/ai-kit/>
- [6] Orange Pi 5:  
<http://www.orangepi.org/html/hardWare/computerAndMicrocontrollers/details/Orange-Pi-5.html>
- [7] BeagleBone AI-64: <https://www.beagleboard.org/boards/beaglebone-ai-64>
- [8] BeagleY-AI: <https://www.beagleboard.org/boards/beagle-y-ai>
- [9] Particle: <https://www.particle.io/>
- [10] Particle Tachyon: <https://www.kickstarter.com/projects/particle-iot/tachyon-powerful-5g-single-board-computer-w-ai-accelerator>
- [11] Radxa ROCK 5C: <https://radxa.com/products/rock5/5c/>
- [12] Brian Tristram Williams, „2024: Eine Odyssee in die KI, Beschleuniger: Desktop vs. Embedded, ein Blick auf einige Optionen“, Elektor 9-10/2024: <https://www.elektormagazine.de/magazine/elektor-355/63247>

# Von Sensordaten zu Modellen des Machine Learning

Gestenerkennung mit einem Beschleunigungssensor und Edge Impulse

Von Koen Vervloesem (Belgien)

Wir sind es gewohnt, mit Geräten über eine Tastatur, eine Maus oder einen Touchscreen zu interagieren, aber manchmal suchen wir nach einer Alternative zu einer solchen „taktile“ Schnittstelle. Die Gestenerkennung stellt eine interessante Option dar. Mit Edge Impulse trainieren wir ein eigenes Machine-Learning-Modell, das Gesten anhand der Daten eines Beschleunigungssensors erkennt.

Die naheliegende Methode, um eine Gestenerkennung zu realisieren, verwendet eine Kamera zur Aufzeichnung der Handbewegungen [1]. Wenn man jedoch einen Beschleunigungsmesser in ein Gerät einbaut, das man am Handgelenk trägt oder in der Hand hält, kann man mit maschinellem Lernen Handgesten auf der Grundlage von Bewegungsdaten erkennen. Dieser Ansatz benötigt wesentlich weniger Daten zur Verarbeitung.

In diesem Projekt wird ein Mikrocontroller-Board XIAO nRF52840 Sense von Seeed Studio mit integriertem Beschleunigungssensor zur Gestenerkennung eingesetzt. Zuerst wird ein Sketch in der Arduino-IDE

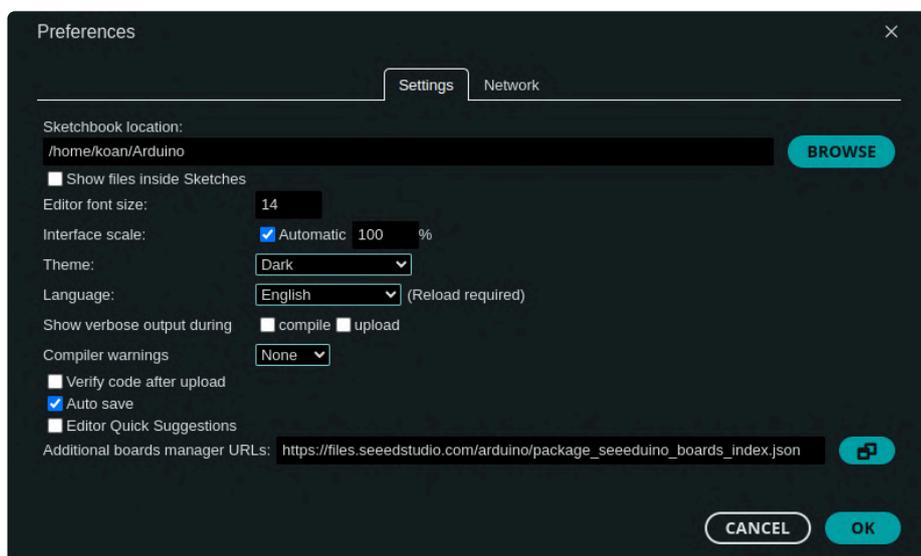


Bild 1. Fügen Sie die Board-URL für die Seeed-Boards in die Arduino-IDE ein.

[2] geschrieben, um die Daten des Beschleunigungssensors zu lesen. Dann werden die gewünschten Gesten ausgeführt, die erkannt werden sollen, und die damit erzeugten Sensordaten zum Training an die Plattform Edge Impulse gesendet [3].

Auf Edge Impulse wird ein maschinelles Lernmodell erstellt, das die Gesten erkennt, und dieses Modell auf dem Mikrocontroller-Board als TensorFlow-Lite-Modell bereitgestellt, das wie eine Arduino-Bibliothek verwendet werden kann. Darauf können Sie dann aufbauen, um Ihren Computer mit Gesten zu steuern.

## Einrichten der Arduino-IDE

Installieren Sie zunächst die Arduino-IDE, die unter Windows, macOS und Linux verfügbar ist. Das Seeed-Board benötigt ein spezielles

Board-Paket, um mit der Arduino-IDE zu funktionieren. Öffnen Sie *Datei* → *Voreinstellungen* und fügen Sie die URL [https://files.seeedstudio.com/arduino/package\\_seeeduino\\_boards\\_index.json](https://files.seeedstudio.com/arduino/package_seeeduino_boards_index.json) in das Feld *Zusätzliche Boardverwalter-URLs* ein (Bild 1). Klicken Sie auf *OK*, um die erforderlichen Board-Paket-Informationen aus der Online-Quelle herunterzuladen.

Öffnen Sie nun den *Boardverwalter*, indem Sie auf das *Board*-Symbol in der linken Seitenleiste klicken und nach *seeed nrf* suchen. Klicken Sie beim Paket *Seeed nRF52 mbed-enabled Boards*, das für eingebettete Machine-Learning-Anwendungen optimiert ist, auf *Installieren* (Bild 2). Nach der Installation schließen Sie das Board an den USB-Port Ihres Computers an, gehen zu *Werkzeuge* → *Board* → *Seeed nRF52*

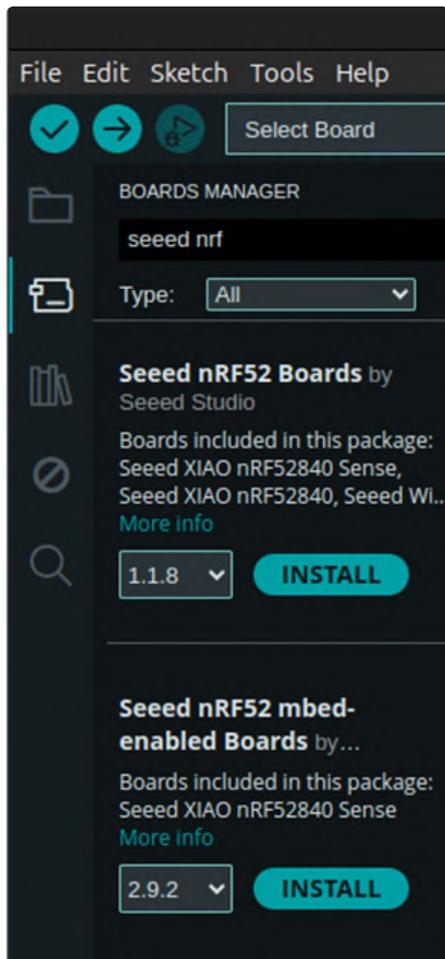


Bild 2. Installieren Sie das Paket *Seeed nRF52 mbed-enabled Boards* in der Arduino-IDE.

*mbed-enabled Boards* und wählen *Seeed XIAO BLE Sense-nRF52840*. Navigieren Sie dann zu *Werkzeuge* → *Port* und wählen Sie den seriellen COM-Port des PCs aus, an dem Ihr Board angeschlossen ist.

Um zu überprüfen, ob Sie einen Arduino-Sketch auf das Gerät flashen können, öffnen Sie *Datei* → *Beispiele* → *01.Basics* → *Blink*. Dies öffnet einen Beispielsketch, der einfach die auf dem Board verbaute LED blinken lässt. Klicken Sie auf das *Upload*-Symbol (der Pfeil nach rechts) oder *Sketch* → *Hochladen* oder *Strg-U*, um den Sketch zu flashen. Wenn alles gut geht, beginnt die LED auf der Platine nach einer Weile zu blinken.

### Testen des Beschleunigungs-sensors

Das Board XIAO nRF52840 Sense von Seeed Studio enthält eine hochpräzise sechssachsigige Trägheitsmesseinheit (IMU) namens LSM6DS3, die einen dreiachsigen Beschleunigungsmesser und ein dreiachsiges Gyroskop enthält. Um sie in Ihrem Arduino-Sketch zu verwenden, gehen Sie auf das Symbol *Bibliotheksmanager* in der linken Seitenleiste und suchen Sie nach *seeed lsm*. Installieren

Sie die Bibliothek *Seeed Arduino LSM6DS3*. Die Bibliothek ist mit einem Beispielcode verbunden. Navigieren Sie zu *Datei* → *Beispiele* → *Seeed Arduino LSM6DS3* → *HighLevelExample*. Kompilieren Sie diesen Code, führen Sie ihn aus und öffnen Sie dann den *Seriellen Monitor* in der oberen rechten Ecke der Arduino-IDE, um die X-, Y- und Z-Koordinaten des Beschleunigungsmessers und des Gyroskops sowie die Temperatur des internen Sensors anzuzeigen (**Bild 3**). Bewegen Sie nun das Seeed-Board in Ihrer Hand und beobachten Sie, wie sich die Werte ändern.

### Ändern des Codes für den Beschleunigungssensor

Ich möchte die X-, Y- und Z-Komponenten des Beschleunigungsmessers verwenden, um ein Modell für maschinelles Lernen zu erstellen, aber das Board muss diese Koordinaten kompakter als im Beispiel senden. Außerdem möchte ich eine Abtastfrequenz von 50 Hz, um auch schnelle Handbewegungen genau zu erfassen, anstatt jede Sekunde eine Abtastung vorzunehmen, wie es die Anweisung `delay(1000)` vorgibt. Erstellen Sie also einen

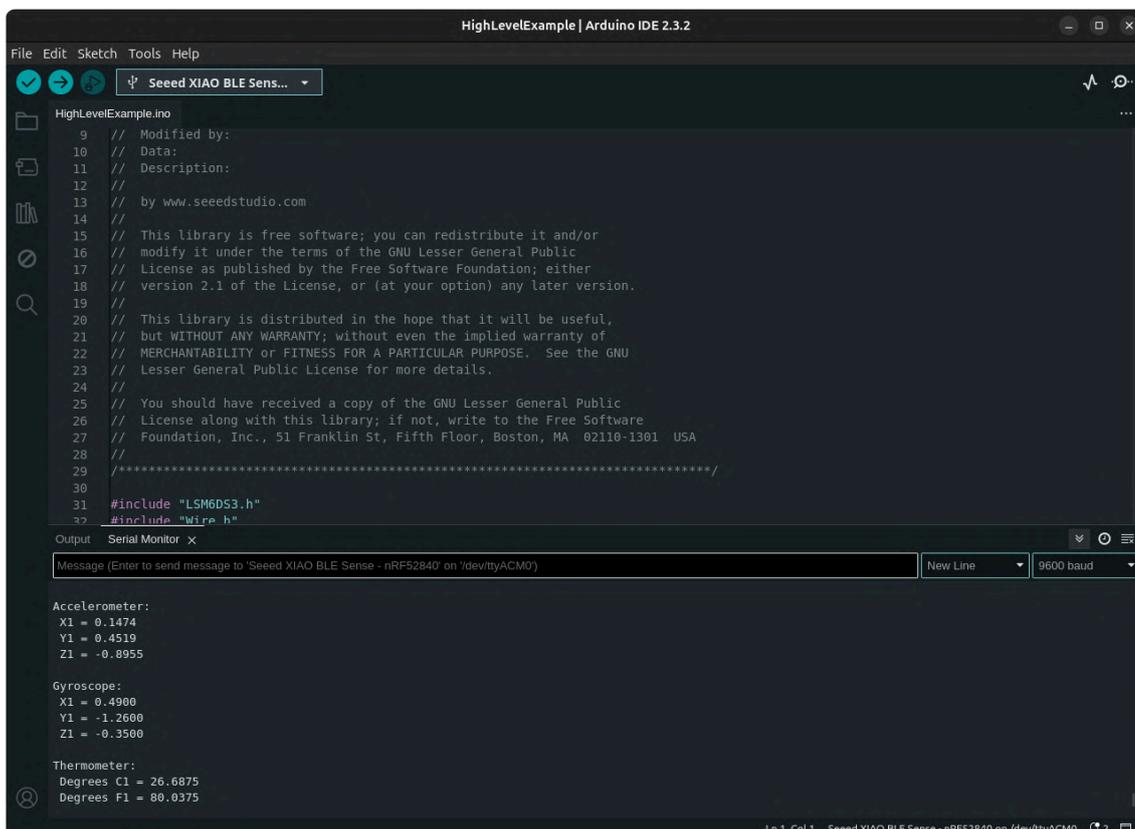


Bild 3. Messwerte des Beschleunigungsmessers beim Bewegen des Seeed-Boards.



Bild 4. Der serielle Plotter der Arduino-IDE zeigt die Werte des Beschleunigungsmessers an.

neuen Sketch mit dem Code aus **Listing 1**, der unter [4] verfügbar ist.

Führen Sie diesen Code auf dem Board aus und beobachten Sie nun die Ausgabe auf dem seriellen Monitor. Die Änderungen der Bewegung werden jetzt schnell angezeigt, mit den drei Beschleunigungskomponenten in  $m/s^2$ . Wenn Sie auf das Symbol für den seriellen Plotter neben dem seriellen Monitor klicken, können Sie die drei Komponenten des Beschleunigungsmessers sogar in grafischer Darstellung aufzeichnen (**Bild 4**).



## Listing 1: Lesen der Daten des Beschleunigungsmessers LSM6DS3

```
// XIAO BLE Sense LSM6DS3 Accelerometer Raw Data

#include "LSM6DS3.h"
#include "Wire.h"

//Create a instance of class LSM6DS3
LSM6DS3 myIMU(I2C_MODE, 0x6A); //I2C device address 0x6A

#define CONVERT_G_TO_MS2 9.80665f
#define FREQUENCY_HZ 50
#define INTERVAL_MS (1000 / (FREQUENCY_HZ + 1))

static unsigned long last_interval_ms = 0;

void setup() {
  Serial.begin(115200);
  while (!Serial)
    ;

  if (myIMU.begin() != 0) {
    Serial.println("Device error");
  } else {
    Serial.println("Device OK!");
  }
}

void loop() {
  if (millis() > last_interval_ms + INTERVAL_MS) {
    last_interval_ms = millis();
    Serial.print(myIMU.readFloatAccelX() * CONVERT_G_TO_MS2, 4);
    Serial.print('\t');
    Serial.print(myIMU.readFloatAccelY() * CONVERT_G_TO_MS2, 4);
    Serial.print('\t');
    Serial.println(myIMU.readFloatAccelZ() * CONVERT_G_TO_MS2, 4);
  }
}
```

## Gesten

Jetzt, da Sie Beschleunigungsmessungen vom Sensor erfassen können, ist es an der Zeit, einen ersten Trainingsdatensatz mit Edge

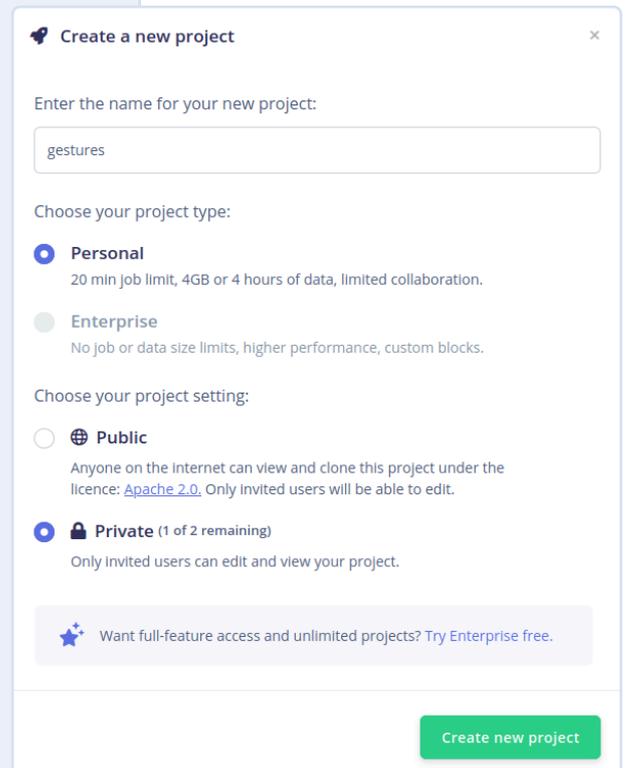


Bild 5. Erstellen Sie ein neues Projekt in Edge Impulse.

Impulse zu erstellen. Zunächst müssen Sie ein Konto bei Edge Impulse eröffnen, das für Studenten, Universitäten und einzelne Entwickler im Rahmen des Community-Plans kostenlos ist [5]. Klicken Sie in Ihrem Dashboard auf *Create new project*, geben Sie dem Projekt einen Namen und entscheiden Sie, ob das Projekt öffentlich oder privat sein soll (**Bild 5**).

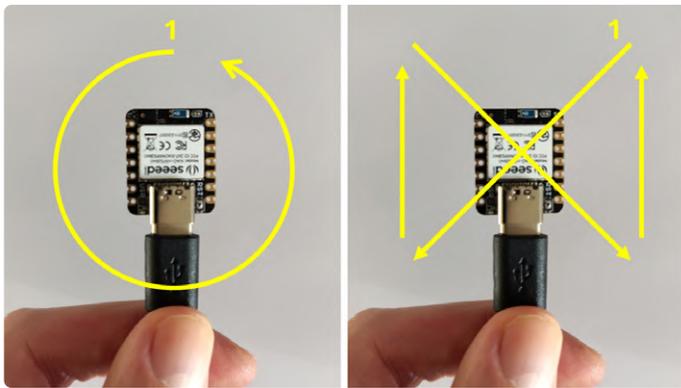


Bild 6. Die beiden Gesten, die wir mit dem eingebauten Beschleunigungssensor des Boards erkennen wollen.

```

/bin/edge-impulse-data-forwarder
Edge Impulse data forwarder v1.27.1
? What is your user name or e-mail address (edgeimpulse.com)? koen@vervloesen.eu
? What is your password? [hidden]
Endpoints:
  Websocket: wss://remote-ngmt.edgeimpulse.com
  API: https://studio.edgeimpulse.com
  Ingestion: https://ingestion.edgeimpulse.com

[SER] Connecting to /dev/ttyACM0
[SER] Serial is connected (D1:4A:E5:69:16:66:06:9E)
[WS] Connecting to wss://remote-ngmt.edgeimpulse.com
[WS] Connected to wss://remote-ngmt.edgeimpulse.com

? To which project do you want to connect this device? Koen Vervloesen / gestures
[SER] Detecting data frequency...
[SER] Detected data frequency: 50Hz
? 3 sensor axes detected (example values: [2.3115,-6.3171,7.1498]). What do you want to call them? Separate the names with ',': x,y,z
? What name do you want to give this device? xiao
[WS] Device "xiao" is now connected to project "gestures". To connect to another project, run "edge-impulse-data-forwarder --clean".
[WS] Go to https://studio.edgeimpulse.com/studio/509218/acquisition/training to build your machine learning model!
  
```

Bild 7. Der Edge Impulse Data Forwarder sendet die von Ihrem Mikrocontroller-Board aufgezeichneten Daten des Beschleunigungsmessers zum Training in die Cloud.

Legen Sie dann die Gesten fest, die Sie erkennen möchten. Zum Beispiel die Gesten eines Kreises und eines Kreuzes. Bei einem Kreis wird das Board in einer kreisförmigen Bahn bewegt, und bei einem Kreuz wird das Board in einer imaginären Ebene von links oben nach rechts unten und dann von rechts oben nach links unten und wieder zurück zur ursprünglichen Position bewegt (Bild 6).

Nun wird ein Datensatz mit mindestens 50 gültigen Stichproben für jede Bewegung und 50 ungültigen Stichproben für unbekannte (zufällige) Bewegungen erstellt. Dabei muss man darauf achten, dass das Board bei all diesen Stichproben auf die gleiche Weise gehalten und die Bewegungen in die gleiche Richtung gehen (zum Beispiel im oder gegen den Uhrzeigersinn für den Kreis) und ungefähr die gleiche Zeit dauern (zum Beispiel 2,5 s). Keiner soll behaupten, dass maschinelles Lernen mühelos erfolgt!

Ihren Edge-Impulse-Benutzernamen oder Ihre E-Mail-Adresse und Ihr Passwort eingeben, dann verbindet es sich über die serielle USB-Schnittstelle mit Ihrem Board und fragt, für welches Projekt Sie die Daten weiterleiten möchten.

Der Data-Forwarder erkennt die Frequenz der Daten sowie die drei Achsen und fordert Sie auf, die Achsen zu benennen. Benennen Sie sie mit x, y und z, getrennt durch Kommata. Anschließend geben Sie einen Namen für das Board ein, woraufhin das Programm die URL für den Bereich *Data Acquisition* des Projekts anzeigt (Bild 7).

### Aufzeichnung von Trainingsdaten

Im Abschnitt *Collect data* auf der rechten Seite sollten Sie Ihr angeschlossenes Gerät sehen. Die Anzahl der Achsen und die Frequenz sollten bereits korrekt sein, aber

die Standard-Abtastlänge von 10 s muss geändert werden. Eigentlich sollte man hier 2,5 s eintragen, aber dieselbe Geste soll in jeder Aufzeichnung mehrmals wiederholt und dann die Aufzeichnung aufgeteilt werden. Um die Aufteilung zu erleichtern, wird zwischen den einzelnen Gesten eine Pause von je 2,5 s eingelegt. Bei 50 Wiederholungen würde eine Aufnahme also 250 s lang sein. Da es aber schwierig ist, sich so lange auf präzise Bewegungen zu konzentrieren, nehmen wir besser fünf Samples von 50 s auf, die jeweils aus zehn Wiederholungen bestehen. Stellen Sie also eine Sample-Länge von 50.000 ms ein. Vergessen Sie auch nicht, die Bezeichnung für diese Geste festzulegen, zum Beispiel *circle*.

Halten Sie nun das Board fest und klicken Sie auf *Start sampling*. Führen Sie alle 5 s eine kreisförmige Bewegung von maximal 2,5 s aus, bis die Abtastung stoppt. Wiederholen Sie

### Weiterleitung der Daten an Edge Impulse

Wir verfügen bereits über das Skript zur Aufzeichnung der Werte des Beschleunigungsmessers. Die Daten unserer Trainingsbewegungen müssen nur noch an Edge Impulse weitergeleitet und dann in 2,5-Sekunden-Intervalle aufgeteilt werden. Hierfür kann man den *Edge Impulse Data Forwarder* verwenden, der Teil der Software *Edge-Impulse-CLI* ist [6]. Folgen Sie den Installationsanweisungen und installieren Sie zunächst Python 3 und Node.js auf Ihrem Computer. Dann installieren Sie auch die CLI-Tools:

```
npm install -g edge-impulse-cli
```

Starten Sie anschließend den *Edge Impulse Data Forwarder* mit:

```
edge-impulse-data-forwarder
```

In dem Programm müssen Sie zunächst

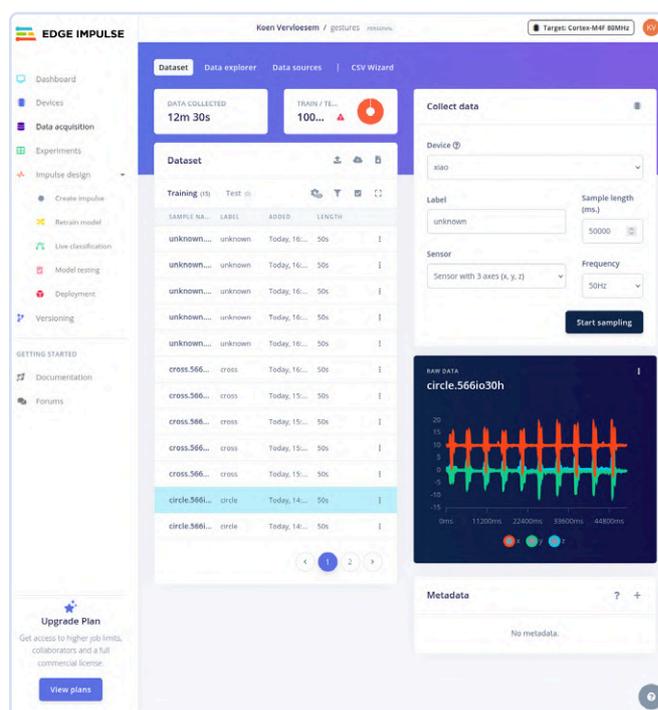


Bild 8. Wir haben gerade zwölf Minuten lang Beschleunigungsmesswerte aufgezeichnet.

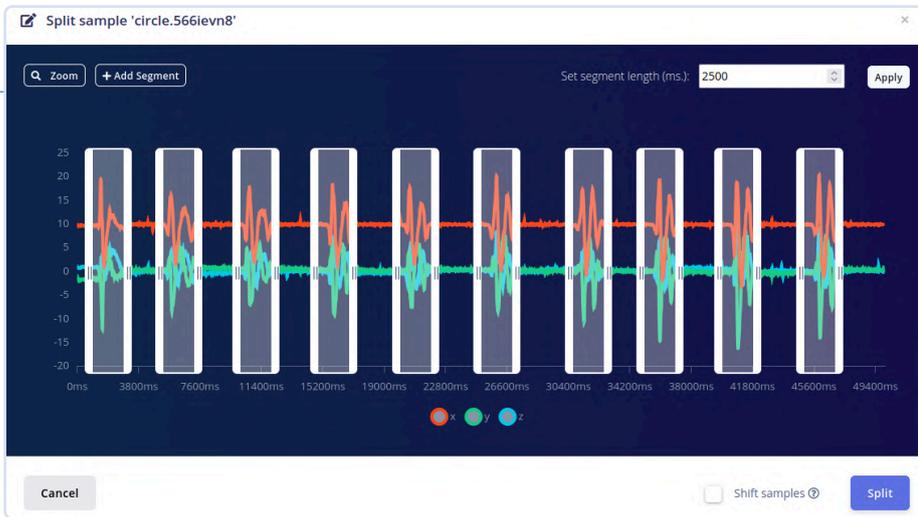


Bild 9. Teilen Sie die aufgezeichneten Daten des Beschleunigungsmessers in zehn Muster von Gesten auf.

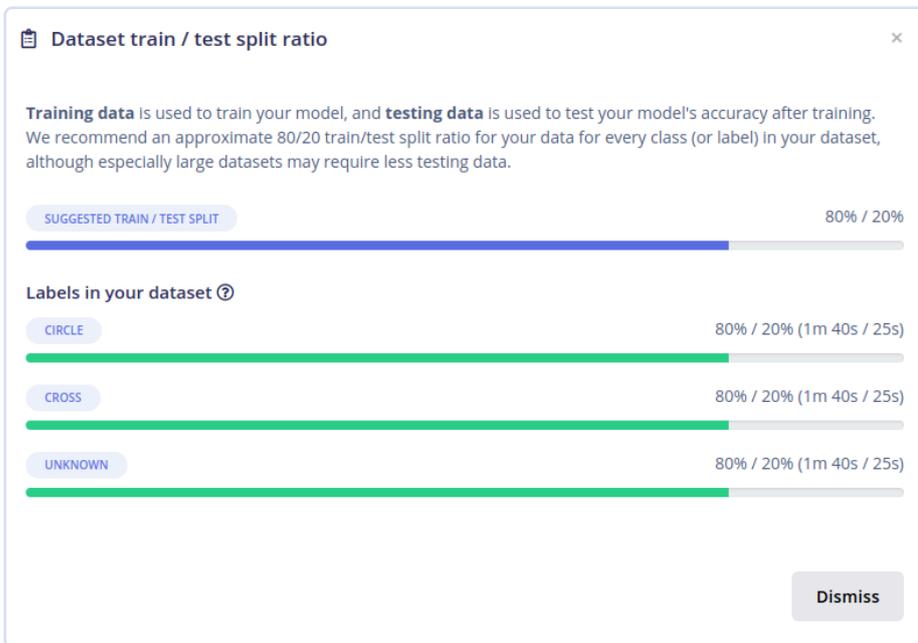


Bild 10. Dies ist eine ausgeglichene Aufteilung von Train und Test.

diesen Vorgang fünfmal. Ändern Sie dann das Label zum Namen der zweiten Klasse *cross* und nehmen Sie fünf Stichproben der Kreuzgeste von jeweils 50 s auf. Schließlich nehmen Sie Stichproben mit der Bezeichnung *unknown* auf, indem Sie zufällige Bewegungen ausführen oder das Board flach auf den Tisch bewegen. Danach sollten Sie - bravo! - mehr als zwölf Minuten an Daten gesammelt haben (Bild 8).

Klicken Sie bei jeder aufgezeichnete Stichprobe auf die drei Punkte und wählen Sie dann *Split sample*. Geben Sie für *Set segment length* (ms.) den Wert 2500 ein und klicken Sie auf *Apply*. Wenn Ihre Gesten nicht länger als 2,5 s sind, teilt Edge Impulse sie automatisch. Klicken Sie auf *Split*, um einzelne Gesten als Samples zu extrahieren. Für jedes ursprüngliche Sample erhalten Sie nun zehn Samples von je 3 s (Bild 9). Die Anzahl der unbekann-

ten Samples kann je nach Ihren zufälligen Bewegungen variieren. Sie können zusätzliche Samples hinzufügen, indem Sie auf *Add Segment* klicken, um zum Beispiel Samples aufzunehmen, wenn das Board flach liegt. Teilen Sie alle aufgenommenen Samples entsprechend auf. Wenn die automatische Segmenterkennung von Edge Impulse fehlerhaft ist, können Sie die Segmente immer noch manuell anpassen. Am Ende sollten Sie insgesamt 150 Samples haben, was etwas mehr als sechs Minuten an Daten ergibt.

Klicken Sie nun auf das Warnzeichen im Feld *Train/Test split* und dann auf *Perform train/test split*. Jedes Ihrer Labels sollte nun in etwa 80 % Trainingsdaten und 20 % Testdaten aufgeteilt sein (Bild 10). Sie können immer noch Stichproben von den Trainingsdaten zu den Testdaten oder umgekehrt verschieben, indem Sie auf die drei Punkte einer Stichprobe

klicken und *Move to test set* oder *Move to training set* wählen, um ein Ungleichgewicht zu beheben.

## Erstellen des Modells

Klicken Sie unter *Impulse design* auf *Create impulse*. Zuerst müssen Sie Ihr Zielgerät konfigurieren. Das Board XIAO nRF52840 Sense ist zwar nicht aufgeführt; Sie können aber *Nordic nRF52840 DK (Cortex-M4F 64MHz)* auswählen, da dieses Board den gleichen Prozessor besitzt. Benennen Sie Ihr Gerät und klicken Sie auf *Save*.

Ein Impuls ist eine Reihe von Funktionen (Blöcken), die Ihre Daten von einer rohen Zeitreihe bis zu den Ausgangsmerkmalen verarbeiten. Stellen Sie im ersten Block *Window size* auf 2.500 ms und *Window increase* auf 400 ms ein. Klicken Sie unten auf *Add a processing block* und wählen Sie *Spectral analysis*. Dieser Block identifiziert die Merkmale des Signals im Frequenzbereich. Klicken Sie anschließend auf *Add a learning block* und wählen Sie *Classification*. Dieser verwendet die spektralen Merkmale, um zu erkennen, ob das Signal einem Kreis, einem Kreuz oder einer unbekanntesten Geste entspricht. Klicken Sie abschließend auf *Save impulse* (Bild 11).

Klicken Sie dann auf *Spectral feature* und unter *Filter auf Type*. Wählen Sie *low*, um ein Tiefpassfilter hinzuzufügen. Setzen Sie die Grenzfrequenz auf 10 Hz, um niederfrequente Signale zu eliminieren, und setzen Sie die Ordnung auf 2, um ein Butterworth-Filter zweiter Ordnung zu verwenden. Sie können diese Werte später noch anpassen. Stellen Sie unter *Analysis* die FFT-Länge auf 128 Ausgangsfrequenzen ein und klicken Sie abschließend auf *Save parameters*.

Dies bringt Sie zum Tab *Generate features*. Klicken Sie dort auf *Generate features*. Die Ergebnisse werden im *Feature explorer* angezeigt. Wenn sich die Merkmale für verschiedene Gesten deutlich voneinander unterscheiden, sollten Sie die Farben für jede Geste in einem Cluster sehen (Bild 12).

Klicken Sie anschließend auf *Classifier*. Die *Neural network architecture* zeigt einen Eingangslayer, zwei Zwischen- oder versteckte Layer und einen Ausgangslayer mit den drei Klassen. Klicken Sie auf *Add an extra layer*, wählen Sie *Dropout* und setzen Sie die Dropout-Rate auf 0,2. Dies verbessert die Genauigkeit, und Sie können die Rate später anpassen, um den optimalen Wert zu finden.

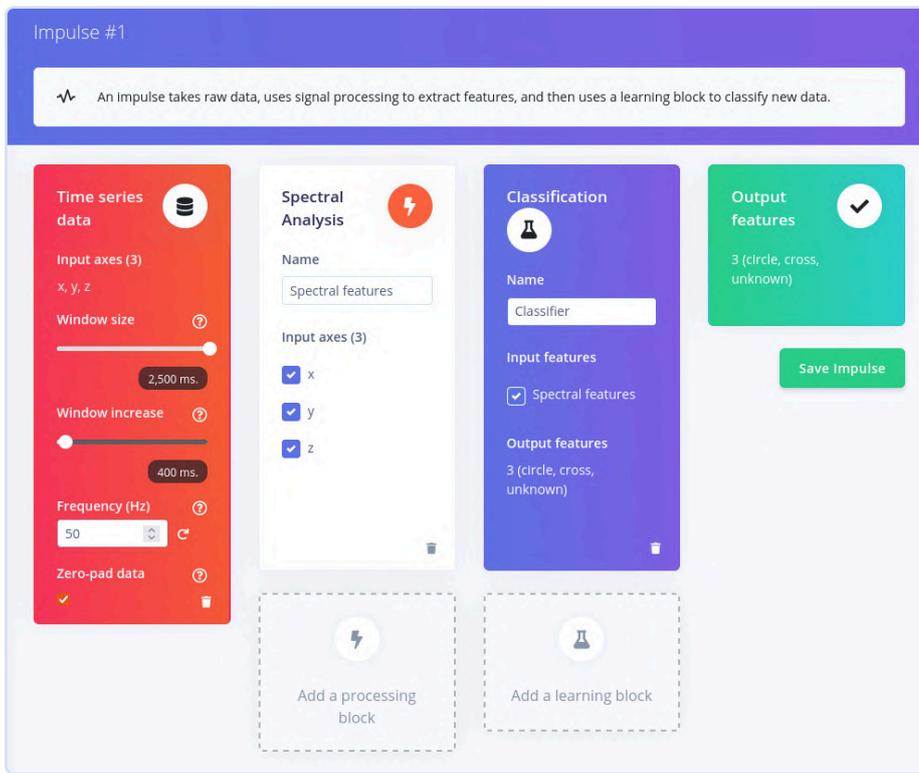


Bild 11. Erstellen Sie einen Impuls, um Ihre Daten zu klassifizieren.

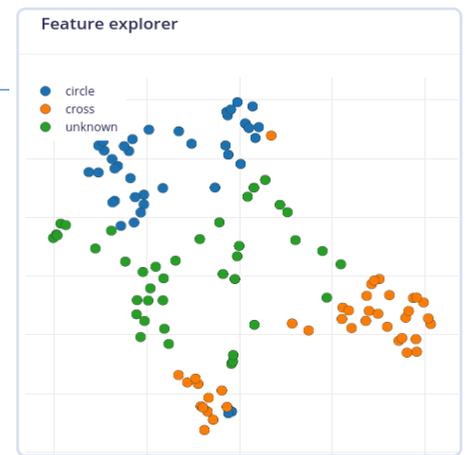


Bild 12. Der Feature-Explorer zeigt eine vielversprechende Unterscheidung zwischen den drei Klassen.

Verschieben Sie diesen Layer zwischen die beiden Zwischenlayer, erhöhen Sie die Anzahl der Trainingszyklen auf 60 und klicken Sie dann auf *Save&train*.

In unserem Fall wurden alle Kreise und Kreuze richtig erkannt, aber 14 % der unbekanntesten Bewegungen wurden fälschlicherweise als Kreise klassifiziert (Bild 13). Experimentieren Sie mit anderen Dropout-Raten oder anderen Parametern, trainieren Sie das Modell erneut und beobachten Sie die Auswirkungen auf die Genauigkeit. Beachten Sie, dass Edge Impulse standardmäßig die quantisierte (optimierte) Modellversion anzeigt. Klicken Sie darauf, um zu *Unoptimized (float32)* zu wechseln und zu prüfen, ob die Genauigkeit dadurch besser wird. In unserem Fall war die Genauigkeit schlechter, da nun 18 % der Kreise als Kreuze erkannt wurden.

## Testen des Modells

Navigieren Sie zur *Live classification*. Stellen Sie die Sample-Länge auf 3000 ein, klicken Sie auf *Start sampling* und machen Sie eine Geste mit Ihrem Board. Das Modell zeigt daraufhin seine Vorhersagen an, mit Angabe des Zeitstempels und der Wahrscheinlichkeit für jede mögliche Geste.

Edge Impulse hat das Modell ausschließlich auf der Grundlage der Trainingsdaten trainiert, daher sollten Sie es auch mit dem Testdatensatz testen. Gehen Sie zu *Model testing* und klicken Sie auf *Classify all*. Dadurch wird das Modell auf alle Stichproben in den Testdaten angewendet und die Ausgabe mit den zugewiesenen Bezeichnungen verglichen. Wenn die Genauigkeit dieses Modelltests deutlich geringer ist als die der Trainingsdaten, deutet dies auf eine mögliche Überanpassung (overfitting) hin. Sie können versuchen, einige Parameter im Trainingsschritt des Modells anzupassen oder neue Stichproben zu sammeln, um dieses Problem zu lösen.

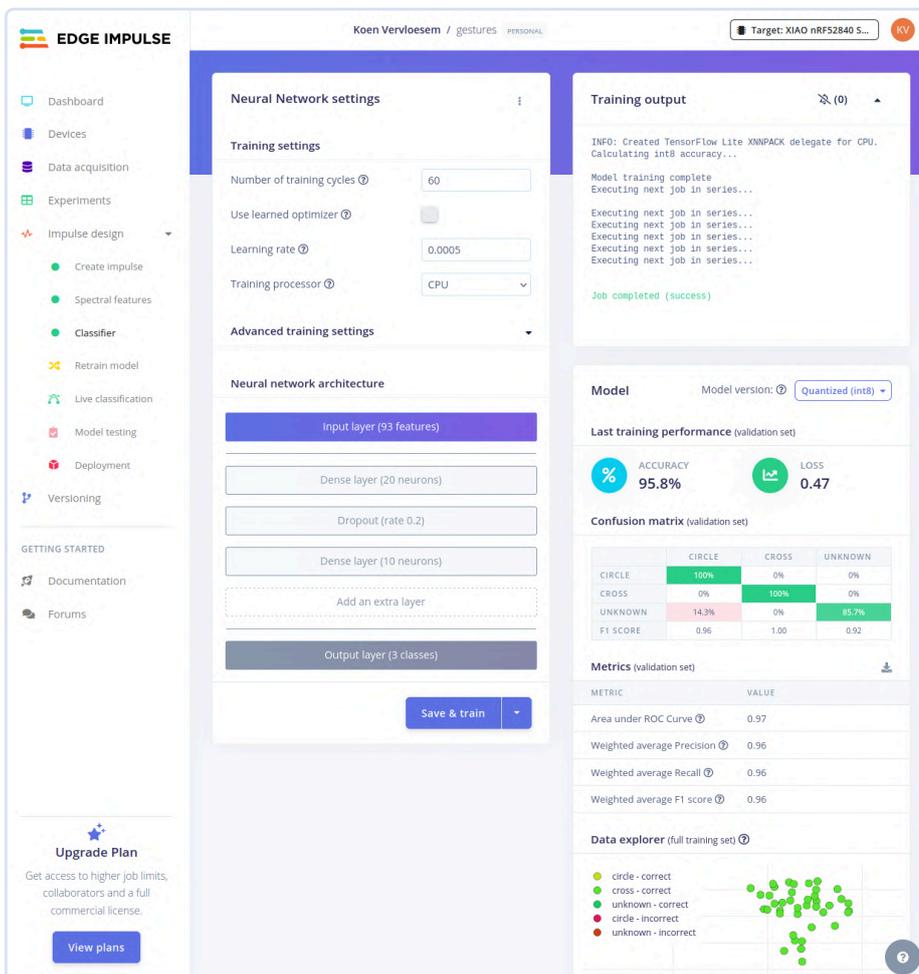


Bild 13. Das trainierte Modell erreicht eine Genauigkeit von 95,8 %.

## Einsatz auf Ihrem Board

Der letzte Schritt ist die Bereitstellung Ihres Modells für maschinelles Lernen auf Ihrem Board. Auf der Registerkarte *Deployment* wählen Sie im Textfeld *Arduino library* aus. Wählen Sie die Quantisierung und klicken Sie dann auf *Build*. Das Ergebnis ist eine ZIP-Datei mit einer Arduino-Bibliothek, die Sie auf Ihrem Computer installieren können.

Öffnen Sie dazu die Arduino-IDE und gehen Sie zu *Sketch* → *Bibliothek einbinden* → *ZIP-Bibliothek hinzufügen...* und wählen Sie die heruntergeladene ZIP-Datei. Für meine Gestenerkennung hieß die Datei *ei-gestures-arduino-1.0.1.zip*. Unter *Datei* → *Beispiele* → *gestures\_inferencing* finden Sie einen Beispielcode, der die Bibliothek verwendet. Dieser Code ist jedoch für andere Boards, nicht für das XIAO nRF52840 Sense. Für das von uns verwendete Board können Sie den Arduino-Sketch in **Listing 2** ausführen, der auf den Beispielen von Edge Impulse für andere Boards basiert.

Beachten Sie, dass die Header-Datei *gestures\_inferencing.h* am Anfang auf die Bibliothek verweist, die Sie mit dem eingebetteten Machine-Learning-Modell installiert haben. Die *loop*-Funktion enthält die interessantesten Abschnitte. Der Code läuft auf dem Board kontinuierlich. Er ordnet einen Puffer für genügend vom Beschleunigungsmesser gelesene Werte zu, um ein Sample zu füllen, liest diese Werte in den Puffer, wandelt den Rohpuffer in ein Signal um und führt dann den

```
Output Serial Monitor X
Message (Enter to send message to 'Seeed XIAO BLE Sense - nRF52840' on '/dev/ttyACM0')
circle: 0.39902
cross: 0.39083
unknown: 0.21555
Classified as: circle
Sampling...
Predictions (DSP: 8 ms., Classification: 0 ms., Anomaly: 0 ms.):
circle: 0.31808
cross: 0.57633
unknown: 0.10559
Classified as: cross
Sampling...
Predictions (DSP: 8 ms., Classification: 0 ms., Anomaly: 0 ms.):
circle: 0.41435
cross: 0.50746
unknown: 0.07819
Classified as: cross
Sampling...
Predictions (DSP: 8 ms., Classification: 0 ms., Anomaly: 0 ms.):
circle: 0.20718
cross: 0.38553
unknown: 0.40729
Classified as: unknown
Sampling...
```

Bild 14. Live-Erkennung von Gesten mit dem Seeed XIAO nRF52840 Sense.

Klassifikator auf dem Signal aus. Am Ende gibt die Funktion die Vorhersagen für jede Klasse sowie die Klasse mit der höchsten Wahrscheinlichkeit aus. Laden Sie also diesen Code von der Arduino-IDE auf das Board hoch und machen Sie Gesten mit dem Board in der Hand. Sie werden die vorhergesagte Geste im seriellen Monitor sehen (**Bild 14**).

## Weitere Schritte

Die Verwendung von Edge Impulse zum Trainieren eines maschinellen Lernmodells zur Erkennung von Gesten auf der Grundlage von Beschleunigungssensor-Daten ist recht einfach. Sie können das resultierende Modell als Arduino-Bibliothek auf vielen unterstützten Entwicklungsboards verwenden. Letztendlich können Sie den Beispielsketch auf einem XIAO nRF52840 Sense von Seeed Studio erweitern, um verschiedene Aktionen auf der Grundlage der erkannten Geste auszuführen. Beispielsweise können Sie den Sketch so konfigurieren,

dass das Board als USB-Tastatur fungiert und Tastendrucke an Ihren Computer sendet, um Software wie einen Media-Player zu steuern. ◀

SE — 240449-02

### Sie haben Fragen oder Kommentare?

Haben Sie technische Fragen oder Anmerkungen zu diesem Artikel? Wenden Sie sich bitte an den Autor unter [koen@vervloesem.eu](mailto:koen@vervloesem.eu) oder an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



### Passendes Produkt

> **Raspberry Pi AI Kit**  
[www.elektor.de/20879](http://www.elektor.de/20879)

## WEBLINKS

- [1] Antonio Aloisio, „Universal-Fernbedienung mit KI“, Elektor Gast-Ausgabe KI, 2024: <https://www.elektormagazine.de/240433-02>
- [2] Arduino-IDE: <https://www.arduino.cc/en/software>
- [3] Edge Impulse: <https://edgeimpulse.com>
- [4] Downloads: <https://www.elektormagazine.de/240449-02>
- [5] Kosten Edge Impulse: <https://edgeimpulse.com/pricing>
- [6] CLI von Edge Impulse: <https://docs.edgeimpulse.com/docs/tools/edge-impulse-cli>



## Listing 2: Ableitung von Gesten aus den Daten des LSM6DS3

```
#include <gestures_inferencing.h>
#include <LSM6DS3.h>
#include <Wire.h>

#define CONVERT_G_TO_MS2 9.80665f
#define MAX_ACCEPTED_RANGE 2.0f
```

```

static bool debug_nn = false; // Set this to true to see
                               // e.g. features generated from the raw signal
LSM6DS3 myIMU(I2C_MODE, 0x6A);

void setup()
{
    Serial.begin(115200);
    Serial.println("Edge Impulse Inferencing Demo");

    if (!myIMU.begin()) {
        ei_printf("Failed to initialize IMU!\r\n");
    }
    else {
        ei_printf("IMU initialized\r\n");
    }

    if (EI_CLASSIFIER_RAW_SAMPLES_PER_FRAME != 3) {
        ei_printf("ERR: EI_CLASSIFIER_RAW_SAMPLES_PER_FRAME
                  should be equal to 3 (the 3 sensor axes)\r\n");
        return;
    }
}

/**
 * @brief Return the sign of the number
 *
 * @param number
 * @return int 1 if positive (or 0) -1 if negative
 */
float ei_get_sign(float number) {
    return (number >= 0.0) ? 1.0 : -1.0;
}

/**
 * @brief      Get data and run inferencing
 *
 * @param[in]  debug  Get debug info if true
 */
void loop()
{
    ei_printf("Sampling...\r\n");

    // Allocate a buffer here for the values we'll read from the IMU
    float buffer[EI_CLASSIFIER_DSP_INPUT_FRAME_SIZE] = { 0 };

    for (size_t ix = 0; ix < EI_CLASSIFIER_DSP_INPUT_FRAME_SIZE; ix += 3) {
        // Determine the next tick (and then sleep later)
        uint64_t next_tick = micros() + (EI_CLASSIFIER_INTERVAL_MS * 1000);

        buffer[ix] = myIMU.readFloatAccelX();
        buffer[ix+1] = myIMU.readFloatAccelY();
        buffer[ix+2] = myIMU.readFloatAccelZ();

        for (int i = 0; i < 3; i++) {
            if (fabs(buffer[ix + i]) > MAX_ACCEPTED_RANGE) {
                buffer[ix + i] = ei_get_sign(buffer[ix + i]) * MAX_ACCEPTED_RANGE;
            }
        }

        buffer[ix + 0] *= CONVERT_G_TO_MS2;
        buffer[ix + 1] *= CONVERT_G_TO_MS2;
    }
}

```

```

    buffer[ix + 2] *= CONVERT_G_TO_MS2;

    delayMicroseconds(next_tick - micros());
}

// Turn the raw buffer in a signal which we can the classify
signal_t signal;
int err = numpy::signal_from_buffer(buffer,
                                     EI_CLASSIFIER_DSP_INPUT_FRAME_SIZE, &signal);
if (err != 0) {
    ei_printf("Failed to create signal from buffer (%d)\n", err);
    return;
}

// Run the classifier
ei_impulse_result_t result = { 0 };

err = run_classifier(&signal, &result, debug_nn);
if (err != EI_IMPULSE_OK) {
    ei_printf("ERR: Failed to run classifier (%d)\n", err);
    return;
}

// print the predictions
ei_printf("Predictions ");
ei_printf("(DSP: %d ms., Classification: %d ms., Anomaly: %d ms.)",
          result.timing.dsp, result.timing.classification, result.timing.anomaly);
ei_printf("\n");
int max_value_index = 0;
for (size_t ix = 0; ix < EI_CLASSIFIER_LABEL_COUNT; ix++) {
    ei_printf("  %s: %.5f\n", result.classification[ix].label,
              result.classification[ix].value);
    if(result.classification[ix].value >=
        result.classification[max_value_index].value) {
        max_value_index = ix;
    }
}
#if EI_CLASSIFIER_HAS_ANOMALY == 1
    ei_printf("  anomaly score: %.3f\n", result.anomaly);
#endif

    ei_printf("Classified as: %s\n",
              result.classification[max_value_index].label);
}

```

# Bau eines undichten, integrierenden und Feuer spuckenden Neurons



Künstliche Intelligenz ohne Software

Von Clemens Valens (Elektor)

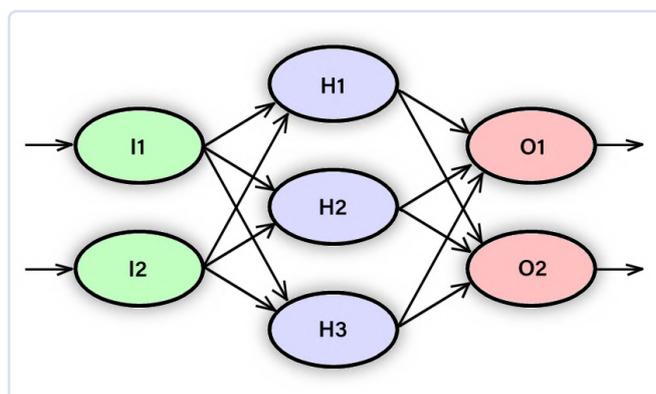
Künstliche neuronale Netze sind die Grundlage der künstlichen Intelligenz. Üblicherweise werden diese Netze in Software implementiert, was jedoch oft zu einem hohen Energieverbrauch führt. Biologische neuronale Netze hingegen arbeiten viel energieeffizienter. In diesem Artikel stellen wir eine Schaltung vor, die das Verhalten eines biologischen Neurons nachahmt. Damit können Sie Ihr eigenes neuronales Netz bauen und experimentell erforschen.

Wer in der U-Bahn auf einem Plakat einen Schnurrbart aufmalt, wird selten für seine Intelligenz gelobt. Doch wenn ein Instagram-Filter Ihnen Eselsohren hinzufügt, wird das oft als brillant empfunden. Intelligenz, ob künstlich oder biologisch, bleibt ein subjektiver Begriff. Was jedoch objektiv bleibt, ist der hohe Energiebedarf, um vermeintlich „intelligente“ Aufgaben zu bewältigen. Millionen, wenn nicht Milliarden von Transistoren tragen zur Darstellung solcher Funktionen wie der Erzeugung von Eselsohren bei und verschwenden dabei enorme Mengen wertvoller Energie. Im Gegensatz dazu erledigen Lebewesen – seien es Austern oder sogar Graspflanzen – komplexere Aufgaben als die allermodernsten KI-Systeme der Welt, und das zu einem Bruchteil der Energiekosten. Tiere benötigen dafür lediglich ein Netz clever verknüpfter Neuronen und eine geringe Menge an Nahrung.

## Neuronale Netze

Neuronen gelten allgemein als Grundbausteine des Gehirns und damit als Basis für Intelligenz. Wenn sie miteinander verbunden werden, formen sie ein neuronales Netzwerk (**Bild 1**). Künstliche neuronale Netze (ANN) sind heute erfolgreich im Einsatz, beispielsweise bei der Spracherkennung, Bildanalyse und in der adaptiven Steuerung – wesentliche Bereiche des maschinellen Lernens und der künstlichen Intelligenz [1].

Bild 1. Ein einfaches künstliches neuronales Netz mit zwei Eingangsneuronen (I1 und I2), einer versteckten Schicht aus drei Neuronen (H1...H3) und zwei Ausgangsneuronen (O1 und O2). Zum Vergleich: Eine Fruchtfliege besitzt etwa 100.000 Neuronen.



Ein Neuron in einem neuronalen Netz, ob künstlich oder biologisch, empfängt Signale von verbundenen Neuronen, verarbeitet diese und leitet bei Bedarf ein Signal weiter. Jedes Eingangssignal wird dabei mit einer bestimmten Gewichtung oder Gewicht versehen. Diese Gewichte sind anpassbar, so dass das Neuron flexibel bleibt und das Netzwerk lernen kann. Das aktuelle Resultat des Neurons wird durch eine nichtlineare Funktion der gewichteten Eingänge bestimmt.

Künstliche Neuronen arbeiten in der Regel mit numerischen Werten, da sie meist in Software implementiert sind. Biologische Neuronen hingegen verwenden elektrische und chemische Signale. Der Vorteil von Softwareneuronen liegt auf der Hand: Es ist einfach, viele von ihnen zu erstellen, und ein Netzwerk kann schnell angepasst werden. Ein neuer Trend ist jedoch die Implementierung künstlicher Neuronen in Hardware, etwa mit programmierbarer Logik wie FPGAs. Hier bestehen die Neuronen aus kleinen Transistor-schaltkreisen und imitieren die elektrischen Signale biologischer Neuronen. Das in diesem Artikel vorgestellte Neuron ist ein Beispiel für eine solche Schaltung.

Biologische Neuronen senden keine kontinuierlichen Signale, sondern kurze Impulse

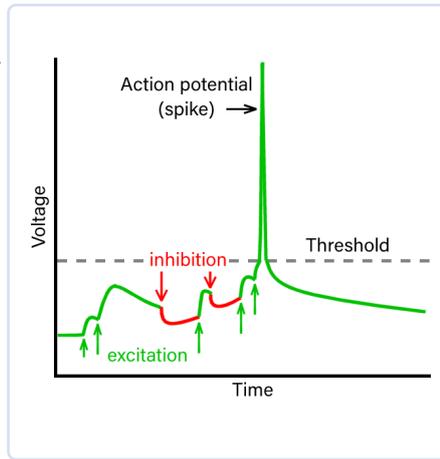


Bild 2. Die Signale eines Spiking Neurons (Quelle: Alan Woodruff / QBI).

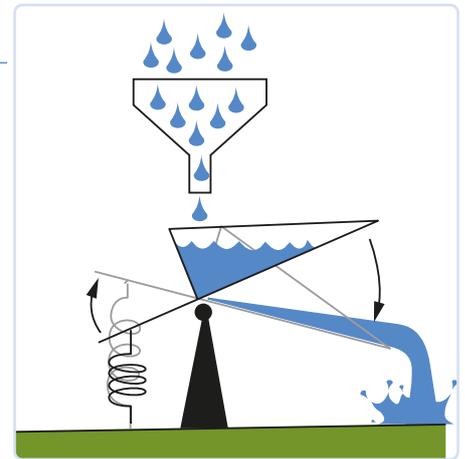


Bild 3. Ein Regenmesser mit Kippwaage.

(Bild 2). Impulsfrequenz und -dauer variieren, und es wird vermutet, dass die Information in der Zeitspanne zwischen den Impulsen und in deren Dauer kodiert ist – möglicherweise in Form einer Puls-Positions-Modulation (PPM). Neuronen, die solche Impulse erzeugen, werden als Spiking Neurons (SN) bezeichnet, und ein Netzwerk aus diesen Neuronen nennt man ein Spiking Neural Network (SNN).

### Das Spiking-Neuron

Ein Spiking-Neuron lässt sich gut mit einem Regenmesser vergleichen, der nach dem Kippeimer-Prinzip arbeitet und den Niederschlag misst (Bild 3). Hierbei entsprechen die eingehenden Impulse den Regentropfen,

die den Eimer füllen. Sobald der Eimer voll ist, kippt er und entleert sich (das Neuron „feuert“), was einen Ausgangsimpuls erzeugt. Im Unterschied zum Regenmesser gewichtet das Spiking Neuron (SN) jedoch die eintreffenden Regentropfen und kann die Breite des Ausgangsimpulses steuern. Darüber hinaus weist das SN ein „Leck“ auf: Der Eimer verliert mit der Zeit an Inhalt, was eine zeitliche Komponente hinzufügt. Kommen die Eingangsimpulse zu langsam an, füllt sich der Eimer nicht ausreichend, und das SN wird nicht ausgelöst – das Neuron vergisst also eine schwache oder verzögerte Stimulation. Das elektronische Gegenstück zum Kippeimer-Prinzip ist der Relaxationsoszillator.

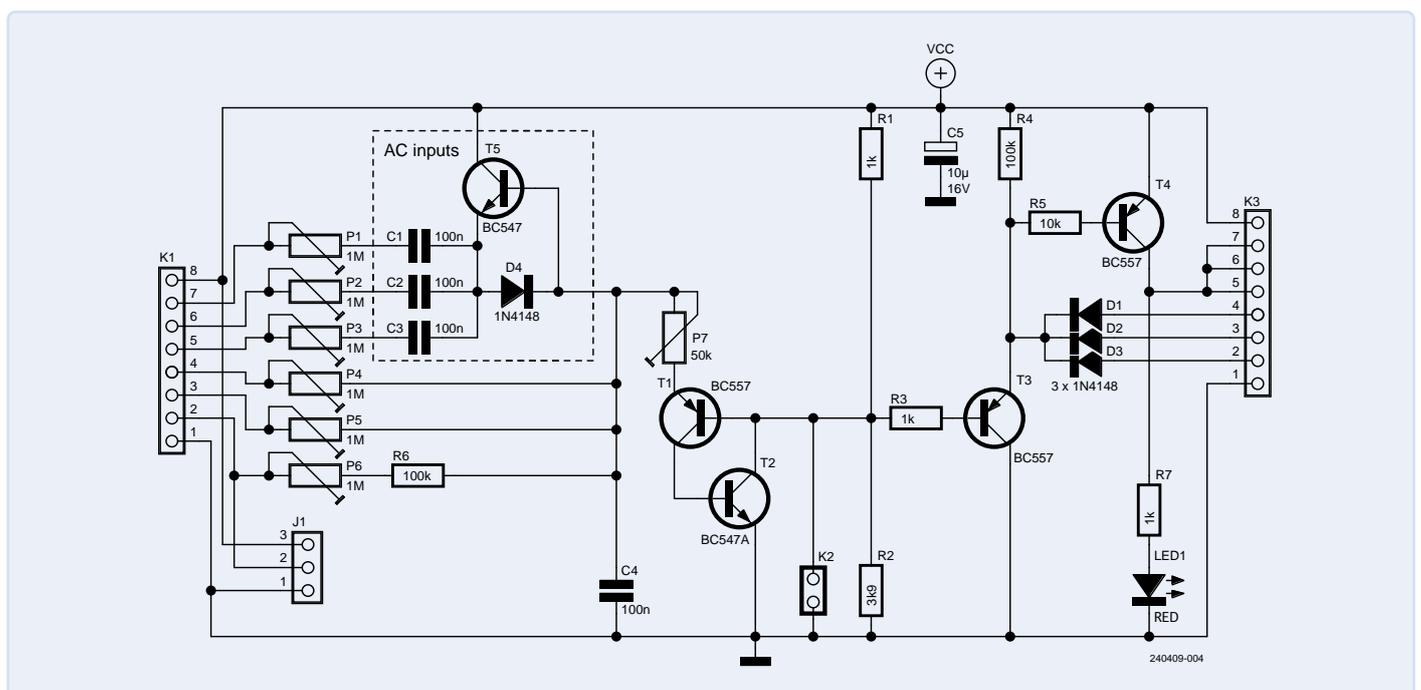


Bild 4: Dieses „undichte, integrierende und Feuer spuckende“ Neuron ist sehr flexibel und konfigurierbar. Es ist zwar nicht so leistungsfähig wie Loihi von Intel oder TrueNorth von IBM, aber es ist ein guter Anfang.

In dieser Schaltung wird ein Kondensator aufgeladen, bis er einen bestimmten Schwellenwert erreicht. Dann wird er durch einen Transistor oder einen elektronisch gesteuerten Schalter entladen, was einen Ausgangsimpuls erzeugt. Ein programmierbarer Unijunction-Transistor (UJT oder PUT) wäre ideal für den Bau eines SN geeignet, jedoch sind diese Bauteile heutzutage schwer und nur zu hohen Preisen zu beschaffen. Glücklicherweise lässt sich ein PUT jedoch leicht aus zwei günstigen bipolaren Transistoren zusammensetzen, wie ich es für dieses Projekt getan habe.

## Beschreibung der Schaltung

Die Schaltung unseres Spiking-Neurons in **Bild 4** ist etwas komplexer als ein einfacher Relaxationsoszillator, da sie die Anschlussmöglichkeiten eines Neurons berücksichtigt. Der Kondensator C4 fungiert hier als Eimer, während die Transistoren T1 und T2 den Schalter bilden, der den Eimer entleert, sobald er voll ist. Der Schwelle, bei der dieser Entladevorgang auslöst, wird durch die Widerstände R1 und R2 (und die Basis-Emitter-Diode von T3) festgelegt. Der Eimer wird mit Impulsen gefüllt, die über D4 von den oberen drei Neuroneneingängen (K1, Pin 5...7) eintreffen und kombiniert werden.

C1...C4, D4 und T5 bilden zusammen eine dreifache „Diodenpumpe“, mit C1...C3 als Eingänge. Ein eingehender Impuls lädt einen dieser Kondensatoren auf. D4 überträgt diese Ladung auf C4, wodurch die Spannung über C4 steigt. Der Emitterfolger T5 als spiegelt die Spannung über C4 zurück zur Anode von D4. Dies führt dazu, dass jeder neue Impuls zum vorherigen addiert wird. Die Spannung über C4 steigt also schrittweise an und erscheint auf einem Oszilloskop als Treppe.

Sobald die Spannung über C4 den Schwellenwert von T1 überschreitet, beginnt dieser Transistor zu leiten. Ein Strom fließt von C4 über Pin 7 zum Emitter von T1 und vom Kollektor zur Basis von T2 und weiter nach Masse, wodurch auch T2 leitend wird. Dies senkt die Basisspannung von T1 unter seine Emitterspannung. T1 ist nun vollständig geöffnet und C4 kann sich schnell entladen – der Eimer kippt.

Wenn der Entladestrom von C4 ausreichend abnimmt, stoppt T2 den Stromfluss und die Basis von T1 kehrt auf die ursprüngliche Spannung zurück, die durch R1 und R2

festgelegt ist. Da die Spannung über C4 nun nahe Null ist, ist die Basisspannung von T1 deutlich höher als seine Emitterspannung, wodurch T1 nicht mehr leitet. Damit ist der Zyklus abgeschlossen, der Eimer wieder leer und bereit für neue Impulse.

Diese spezielle Konfiguration von T1 und T2 wird manchmal als „transistorisierter Thyristor“ bezeichnet und war in den 1970er Jahren eine populäre Schaltungslösung. T2 sollte dabei keinen zu hohen Beta-Wert (hFE) haben, da er sonst auch bei fast keinem Entladestrom weiterleiten könnte. Ein A- oder B-Typ-Transistor ist hier ideal, während ein C-Typ-Transistor zu empfindlich sein könnte. Aus diesem Grund ist T2 als BC547A (oder ein gleichwertiger Transistor) angegeben. Bei T5, ebenfalls ein BC547, spielt der Beta-Wert keine Rolle, sodass auch hier ein BC547A verwendet werden kann – das spart eine Zeile in der Stückliste.

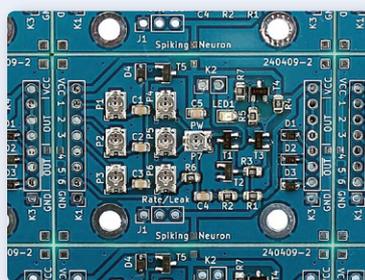
## Eingänge und Ausgänge

Beim Aufladen von C4 ist das Signal am Knoten R1/R2 high; beim Entladen von C4 fällt es ab. Das Poti P7 steuert die Entladedauer: Wenn der Wert von P7 klein ist, ist der Aktiv-Low-Impuls an R1/R2 entsprechend kurz. Der Testpunkt dafür ist K2.

Das Signal an R1/R2 wird durch den PNP-Transistor T3 gepuffert. D1, D2 und D3 liefern drei low-aktive Ausgänge für das Neuron. Wenn T3 sperrt, werden sie von R4 auf high gezogen, und auf low gezogen, wenn T3 leitet.

## Experimentieren mit SMDs

Um das Experimentieren mit einem neuronalen Netzwerk zu vereinfachen, habe ich eine kompaktere Version des Neurons mit oberflächenmontierten Bauteilen (SMDs) erstellt.



### Stückliste

**Widerstände:**  
R1, R3, R7 = 1 k  
R2 = 3k9  
R5 = 10 k  
R4, R6 = 100 k

**Kondensatoren:**  
C1...C4 = 100 n, Raster 2,5 mm oder 5 mm  
C5 = 10 µ, 16 V, Raster 2,5 mm

**Halbleiter:**  
D1...D4 = 1N4148  
T1, T3, T4 = BC557  
T2, T5 = BC547A

**Sonstiges:**  
JP1 = 1x3-polige Stiftleiste, Raster 2,54 mm  
K1 = 1x8- polige Buchsenleiste, Raster 2,54 mm  
K2 = 1x2- polige Stiftleiste, Raster 2,54 mm  
K3 = 1x8- polige Stiftleiste, Raster 2,54 mm

Um das Netzwerk experimentell zu testen, habe ich den Platinenhersteller gebeten, eine 3x3-Platine mit V-Cut-Nuten herzustellen (siehe Bild). Verteilerpads für die Stromversorgung sind um die Platine herum verteilt, sodass alle Neuronen leicht mit Drahtbrücken oder 0-Ω-Widerständen an die Stromversorgung angeschlossen werden können. Mit dünnem Draht lässt sich ein Netzwerk wie in Bild 1 schnell verdrahten und sogar um zusätzliche Eingangs- und Ausgangsneuronen erweitern (wie ich es getan habe). Verdrahten Sie sorgfältig und methodisch, damit Sie die Verbindungen der Ein- und Ausgänge leicht nachvollziehen können.

Wenn T3 sperrt, sperrt auch T4 und die Aktiv-High-Ausgänge des Neurons (K3, Pin 5... Pin 7) werden abgeschaltet. Wenn T3 leitet, leitet auch T4, und die Ausgänge werden auf High gezogen. LED1 leuchtet auf, um den Impuls zu signalisieren. Das bedeutet, dass die Ausgänge niederohmig sind, wenn das Neuron einen Impuls abgibt, und ansonsten hochohmig. Dadurch wird vermieden, dass sich zwei Neuronen gegenseitig beeinflussen, wenn sie dasselbe Neuron ansteuern.

Die beiden Ausgangstypen ermöglichen es dem Neuron, ein anderes Neuron entweder zu stimulieren oder zu hemmen. Dies wird klarer, wenn wir uns die Eingänge des Neurons anschauen, die ebenfalls in zwei Typen unterteilt sind: Die oberen drei AC-Eingänge sind, wie bereits erwähnt, mit der Diodenpumpe verbunden; die unteren drei DC-Eingänge sind über ein Gewichtungspotis direkt mit C4 verbunden. Diese Eingänge ermöglichen eine direkte Steuerung der Ladung von C4 und können es entweder laden oder entladen. Der untere Eingang (K1, Pin 2) verfügt über den Jumper J1, um die Konfiguration für verschiedene Anwendungen zu erleichtern. Wenn Pin 1 und Pin 2 von J1 verbunden sind, ist der Entladevorgang durch P6 einstellbar. Bei Verbindung von Pin 2 und Pin 3 lädt sich C4 über P6 auf und das Neuron beginnt zu oszillieren. R6 sorgt dafür, dass die Oszillation über den gesamten Bereich von P6 erfolgt. Um die AC-Eingänge in DC-Eingänge umzuwandeln, ersetzen Sie C1...C3 durch Dioden (mit der Anode am Potentiometer verbunden), ersetzen Sie D4 durch eine Drahtbrücke und entfernen Sie T5. Die drei Dioden verhindern eine gegenseitige Beeinflussung der Neuroneneingänge. Wenn Sie stattdessen C1...C3 durch Drahtbrücken ersetzen, entstehen zusätzliche Leckeingänge.

Die Stromversorgung ist unkritisch und kann zwischen 3,5 V und 15 V oder mehr liegen, wenn C5 entsprechend angepasst wird. Die Stromaufnahme wird hauptsächlich durch die LED bestimmt und liegt bei einer 5-V-Versorgung unter 1 mA.

## Merkmale

Zusammenfassend ergibt sich für unser undichtes, integrierendes und Feuer spuckendes Neuron die folgende Liste von Merkmalen:

- 3 Stimulus/AC-Eingänge mit individuellen Gewichten

## Testen des Netzwerks

Um zu sehen, ob das Netzwerk so funktioniert, wie es sollte, drehen Sie die Gewichtungspotentiometer P1...P6 jedes Knotens vollständig gegen den Uhrzeigersinn (Maximalwert). Drehen Sie P7 jedes Knotens in die mittlere Position. Auf dem Neuron in der oberen linken Ecke der Tafel, nennen wir es N00 (N, Zeile, Spalte), setzen Sie eine Brücke auf Pin 2 und Pin 3 von J1. Dadurch wird P6 mit VCC verbunden. Schalten Sie die Stromversorgung des Panels ein. Die LED an N00 beginnt zu blinken. Die Blinkrate kann mit P6 und die Intensität mit P7 eingestellt werden. Als nächstes werden die entsprechenden Gewichtungspotentiometer (bei mir P1) an den angeschlossenen Neuronen N10, N11 und N12 eingestellt. Ihre LEDs sollten anfangen zu blinken. Je kleiner der Wert des Gewichtungspotentiometers, desto schneller blinkt die LED.

Entfernen Sie den Jumper von N00 und stecken Sie ihn auf N01, das Neuron darunter. Wie zuvor beginnt die LED zu blinken, und Sie sollten in der Lage sein, die Blinkrate (P6) und die Intensität (P7) der LED einzustellen. Wie zuvor stellen Sie das Gewichtungspotentiometer (in meinem Fall P2) der angeschlossenen Neuronen N10, N11 und N12 ein. Auch hier sollten die LEDs anfangen zu blinken.



- 3 Inhibit/Leck/DC-Eingänge mit individuellen Gewichten
- 3 Stimulus-Ausgänge (aktiv high)
- Inhibit/Leck-Ausgänge (aktiv low)
- Einstellbare Ausgangsimpulsbreite (Intensität)
- Frei laufender Spike-Generator
- LED zur Visualisierung des Neuronenfeuers
- Versorgungsspannung: 3,5...15 V

## Erstellen eines Spiking Neural Networks

Um ein kleines neuronales Netzwerk zu bauen, das eine einfache Aufgabe erfüllen kann, benötigen Sie wahrscheinlich mindes-

Wiederholen Sie diesen Vorgang für die übrigen Neuronen. Beachten Sie, dass die Ausgangsneuronen nicht mit anderen Neuronen verbunden sind, sodass bei ihnen keine Gewichtung erforderlich ist.

Wenn alle LEDs wie beschrieben gesteuert werden können, funktioniert der Stimulus-Teil des Netzes einwandfrei. Um den Leck-/Hemnteil jedes Knotens zu testen, bringen Sie die Eingangsneuronen zum Schwingen, indem Sie eine Brücke auf Pin 2 und Pin 3 von J1 setzen. Verbinden Sie dann Pin 1 und Pin 2 von JP1 an einem anderen Knoten. Dadurch sollte sich das Blinkverhalten der LED an diesem Knoten spürbar ändern. Nun bleibt nur noch das Training des Netzes für eine bestimmte Aufgabe. Ich überlasse das Ihnen, dem Leser als Übung (da ich selbst auch noch keine genaue Lösung kenne).

Hier sehen Sie mein 3x3-spikendes neuronales Netz. Auf den ersten Blick scheint es, als würden die LEDs zufällig blinken – doch natürlich ist das nicht der Fall. Das Netzwerk folgt einer intelligenten Logik, die ich jedoch (noch) nicht ganz durchschaut habe.

tens sieben Neuronen (Bild 1). In diesem Netzwerk gibt es zwei Eingangsneuronen, zwei Ausgangsneuronen und drei Neuronen in der sogenannten versteckten Schicht. Weitere Hintergrundinformationen zu dieser Netzwerkstruktur finden Sie in Quellen wie Wikipedia oder ChatGPT.

Für unser Netzwerk verbinden wir die Eingangsneuronen mit Sensoren (licht- oder temperaturabhängige Widerstände oder andere Bauteile, die physikalische Größen in Spannung oder Strom umwandeln). Ein solcher Sensor kann über einen der Gleichstromeingänge die Spike-Rate des Neurons steuern. Die Stimulusausgänge der beiden Eingangsneuronen sind wiederum mit den

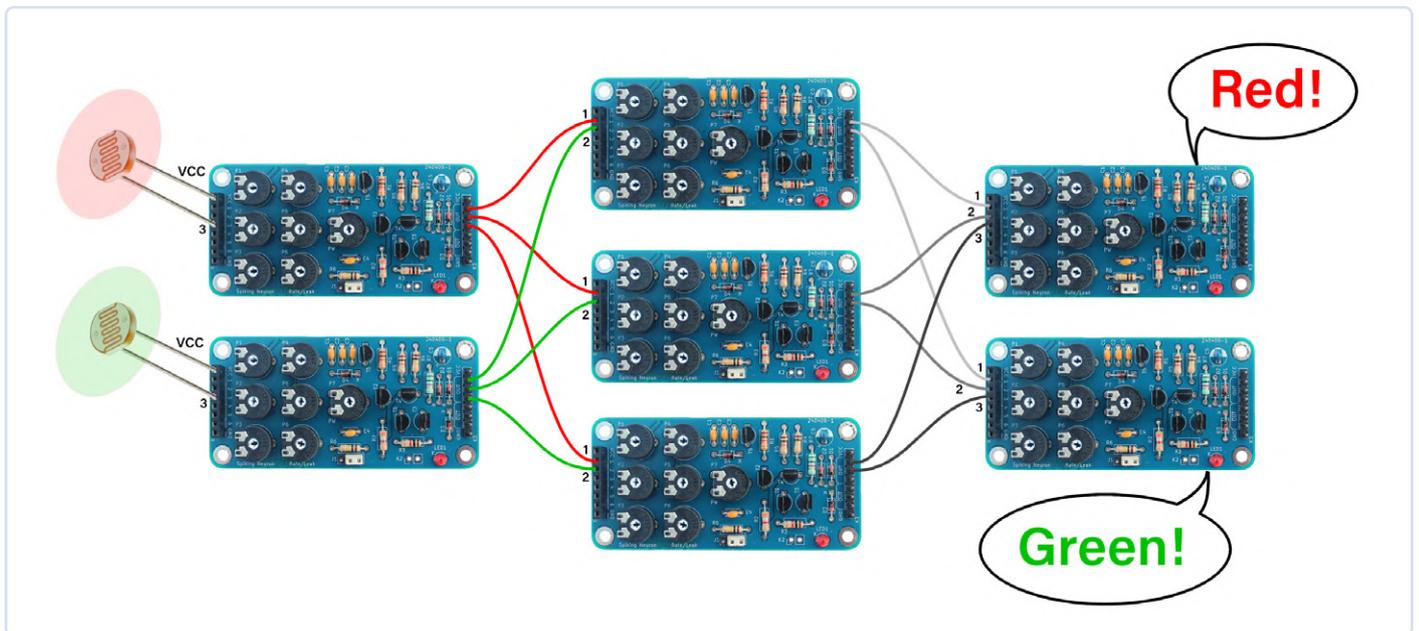


Bild 5. Bei richtiger Konfiguration (Training) kann dieses einfache neuronale Netz zwischen rotem und grünem Licht unterscheiden und somit den Zustand einer Ampel erkennen. Ziemlich clever, oder? Denken Sie wirklich, dass eine solche komplexe Aufgabe mit weniger Bauteilen bewältigt werden könnte? Hinweis: Die Stromversorgungsanschlüsse wurden zur besseren Übersichtlichkeit weggelassen.

drei Neuronen in der versteckten Schicht verbunden, und die Ausgänge dieser Neuronen werden an die Eingänge der Ausgangsneuronen angeschlossen (Bild 5). Nun müssen lediglich die Potentiometer aller Neuronen so eingestellt werden, dass das Netzwerk die gewünschte Funktion erfüllt. Zugegeben, im Netzwerk in Bild 5 gibt es etwa zwanzig Potentiometer zu justieren – und ich gebe zu, dass ich auch nicht genau weiß, wie man das optimal bewerkstelligt. Doch genau hier liegt der Spaß an solchen Projekten: Sie bieten reichlich Raum zum Experimentieren. Apropos Experimentieren: Die Werte der Widerstände, Potentiometer und Kondensatoren links vom transistorisierten Thyristor T1/T2 (bis einschließlich P7) sind frei wählbar. Sie können gerne mit anderen Werten experimen-

tieren. Ich empfehle jedoch, die Spike-Raten eher hoch zu halten, da dies die Arbeit mit einem Oszilloskop erleichtert. Biologische Neuronen arbeiten typischerweise mit niedrigeren Spike-Raten, meist unter 100 Hz. Die Designdateien für das Spiking Neuron (THT- und SMD-Version) stehen zum Download unter [1] bereit. ◀

SE – 240409-02

### Sie haben Fragen oder Kommentare?

Haben Sie technische Fragen oder Anmerkungen zu diesem Artikel? Wenden Sie sich per E-Mail an den Autor unter [clemens.valens@elektor.com](mailto:clemens.valens@elektor.com) oder an Elektor unter [redaktion@elektor.de](mailto:redaktion@elektor.de).

**THEMEN FOKUS**

Besuchen Sie unsere Webseite **Embedding & KI** für mehr Artikel, Projekte, Nachrichten und Videos.

[www.elektormagazine.de/embedded-ki](http://www.elektormagazine.de/embedded-ki)

**Passendes Produkt**

➤ **Get Started with the MAX78000FTHR Bundle**  
[www.elektor.de/19798](http://www.elektor.de/19798)



### WEBLINKS

- [1] S. Cording, „Die Neuronen in neuronalen Netzen verstehen: Künstliche Neuronen“, [elektormagazine.de: https://www.elektormagazine.de/articles/neuronen-in-neuronalen-netzwerken-verstehen-teil-1-kunstliche-neuronen](https://www.elektormagazine.de/articles/neuronen-in-neuronalen-netzwerken-verstehen-teil-1-kunstliche-neuronen)
- [2] Dieses Projekt bei Elektor Labs: <https://www.elektormagazine.de/labs/spiking-neural-network-node>

# ChatGPT für den Elektronikentwurf

## Macht GPT-4o es besser?

Von Clemens Valens (Elektor)

Zu Beginn des Jahres 2023 wollte ich herausfinden, ob ChatGPT als Werkzeug für den Elektronikentwurf verwendet werden kann. Ich habe es mit der Version „Feb 13“ probiert, die auf der GPT-3-Architektur basiert. Obwohl ich von den sprachlichen Fähigkeiten tief beeindruckt war, überzeugten mich die technischen Fertigkeiten nicht. Seitdem ist GPT-3 zu GPT-4o geworden. Ist es besser geworden? Zeit, das herauszufinden!

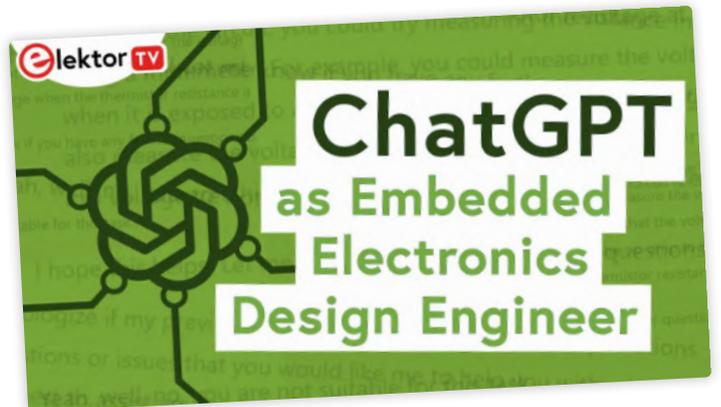
Aus dem ersten Gespräch, das ich mit ChatGPT führte, schloss ich, dass es für die komplizierten Aufgaben, für die ich es verwenden wollte, noch nicht bereit war. Sie können sich dazu das Video auf Elektor TV [1] ansehen. Verständlich, denn die meisten Menschen brauchen auch mehrere Jahre, um die Kunst der Elektronik zu erlernen. Schwerwiegender war jedoch, dass ChatGPT ohne mit der Wimper zu zucken von Sinn zu Unsinn wechselte, und ich habe es sogar bei einer glatten Lüge ertappt. Daher sollte man alles, was GPT von sich gibt, mit Vorsicht genießen.

### Wird ChatGPT Wikipedia ersetzen?

In den letzten Monaten habe ich mich weiter mit ChatGPT [2] beschäftigt und festgestellt, dass es sich stark verbessert hat. Ich habe sogar festgestellt, dass ich meine Lieblingswebseite Wikipedia nicht mehr benutze. ChatGPT reproduziert wahrscheinlich eine Menge Wikipedia-Inhalte, aber da es die Inhalte zusammenfasst, kann man damit viel schneller nachschlagen.

### Fehlersuche in einer Schaltung

Eine der Übungen, die ich mit GPT-3 ausprobiert habe, war die Fehlersuche in der einfachen Schaltung eines Durchgangsprüfers



Sehen Sie sich Clemens' Originalvideo ([https://youtu.be/pM1n\\_q7q670](https://youtu.be/pM1n_q7q670)) vom Februar 2023 an, in dem er ChatGPT als Werkzeug für den Entwurf elektronischer Schaltungen und eingebetteter Software testet.



in **Bild 1**. Die LED sollte aus sein, wenn J1 kurzgeschlossen ist, und sie sollte leuchten, wenn J1 offen ist. Aber es gibt einen Fehler. Der Transistor Q1 ist als PNP-Typ spezifiziert und nicht als NPN-Typ, den man hier erwarten würde. Wenn Sie die auf einem Breadboard ausprobieren, werden Sie feststellen, dass sie zwar funktioniert, aber in die falsche Richtung: Die LED leuchtet, wenn J1 kurzgeschlossen ist, und ist aus, wenn J1 offen ist. Der Grund dafür ist, dass J1 die Basis und den Emitter kurzschließt, wodurch Q1 effektiv zu einer Diode wird.

Als ich mit Hilfe von ChatGPT versuchte, das Problem zu finden und zu lösen, stellte sich heraus, dass die größte Schwierigkeit darin bestand, ChatGPT die Schaltung zu „zeigen“. Auf die Frage, wie man das macht, wurde mir gesagt, ich solle eine Zeichnung zum Beispiel auf GitHub hochladen und dann den Link angeben, was ich dann auch tat. ChatGPT antwortete daraufhin wie folgt:

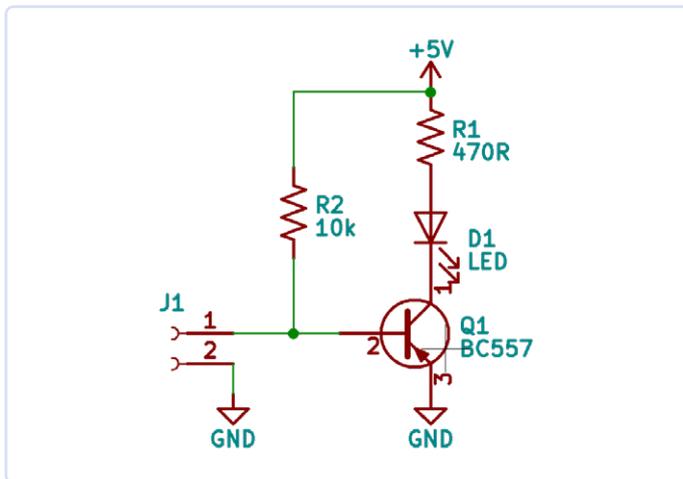


Bild 1. Ein einfacher, aber fehlerhafter Durchgangsprüfer. Kann ChatGPT den Fehler finden? Können Sie es?

Verstehe, danke für das Teilen des Schaltplans. Ich habe den Schaltplan überprüft, und es sieht aus wie ein einfacher Durchgangsprüfer. Hier sind einige erste Gedanken und Fragen:

- > Der Schaltplan zeigt eine 9-V-Batterie als Stromversorgung. Ist das die Stromversorgung, die Sie für die Schaltung verwenden?
- > Die Schaltung verwendet einen 741-Operationsverstärker als Komparator. Gibt es einen bestimmten Grund, warum Sie diesen Operationsverstärker gewählt haben?
- > Die Schaltung verwendet einen Thermistor als Sensorelement zur Erkennung eines offenen Stromkreises. Wie wird der Thermistor in der Schaltung verwendet, und wie ist das erwartete Verhalten, wenn ein offener Stromkreis erkannt wird?
- > Der Schaltplan zeigt zwei LEDs, eine grüne und eine rote. Wie werden diese LEDs in der Schaltung verwendet, und welches Verhalten ist zu erwarten, wenn ein offener Stromkreis erkannt wird?

Ausgehend vom Schaltplan ist es möglich, dass die Schaltung nicht wie erwartet funktioniert, wenn die Bauteilwerte oder die Anschlüsse falsch sind. Um die Fehlersuche zu erleichtern, könnten Sie mehr Informationen über Probleme oder unerwartetes Verhalten der Schaltung zur Verfügung stellen?

ChatGPT hatte sich meine Schaltung offensichtlich gar nicht angesehen, sondern etwas aus seinem Speicher ausgespuckt, das dem Hoch-/Tiefemperaturdetektor in **Bild 2** ähnelt. Auch nach mehreren Versuchen konnte ich ChatGPT nicht dazu bewegen, sich meine Schaltung anzusehen, und so gab ich auf. Wenn ChatGPT keine Schaltpläne lesen mag, dann kann es sie auch nicht analysieren und korrigieren.

### Fehlersuche mit GPT-4o

Die ChatGPT-Version GPT-4o hat sich im Vergleich zu GPT-3 [3] in mehrfacher Hinsicht weiterentwickelt. Sie ist nicht nur besser in der Verarbeitung natürlicher Sprache und weist Sie auf seine Grenzen hin, sondern zeigt auch eine verbesserte Benutzeroberfläche. Neben der Texteingabe können nun auch Dateien hochgeladen werden, so dass das Programm auch mit Bildern und anderen

Dokumenten arbeiten kann. Ich habe diese Funktion erfolgreich genutzt, um zum Beispiel Scans von chinesischen Dokumenten ins Englische zu übersetzen. Da dies vermuten ließ, dass ChatGPT sich Dokumente nun wirklich ansah, versuchte ich die vorherige Übung zur Fehlersuche im Schaltplan mit GPT-4o. Nachdem ich die gleichen einleitenden Fragen wie bei GPT-3 gestellt hatte, lud ich die Datei direkt statt über den GitHub-Umweg hoch. Hier ist die Antwort (einschließlich Formatierung; die Tipps zur Fehlerbehebung habe ich weggelassen). GPT-4o ist viel ausführlicher als GPT-3.

Ich sehe, dass Sie einen Schaltplan mit dem Titel „Open Circuit Detector“ hochgeladen haben. Ausgehend von den aufgelisteten Bauteilen scheint es sich um eine einfache Schaltung mit einer LED, einem Transistor (BC557), Widerständen und einem Verbinder (J1) zu handeln.

Hier ist eine kurze Analyse der Schaltung:

Bauteile:

1. **D1**: LED
2. **Q1**: BC557 PNP-Transistor
3. **R1**: 470Ω-Widerstand (scheinbar ein Strombegrenzungswiderstand für die LED)
4. **R2**: 10kΩ-Widerstand (wahrscheinlich ein Basiswiderstand für den Transistor)
5. **J1**: Steckverbinder

Beschreibung der Schaltung:

- > Die LED (D1) ist in Reihe mit dem Widerstand R1 und dem Kollektor des PNP-Transistors (Q1) geschaltet.
- > Der Emitter von Q1 ist mit der +5V-Versorgung verbunden.
- > Die Basis von Q1 ist mit dem Anschluss von R2 und einem der Stifte des Steckverbinders (J1) verbunden.

Erwartete Betriebsweise:

- > Wenn die Basis des Transistors Q1 auf einen niedrigen Wert gezogen wird (nahe an Masse), sollte Q1 einschalten, so dass Strom vom Emitter zum Kollektor fließt und die LED aufleuchtet.
- > Wenn es einen offenen Stromkreis oder keine Verbindung an J1 gibt, schaltet sich der Transistor nicht ein, und die LED bleibt aus.

(Hier habe ich die Tipps zur Fehlersuche weggelassen)

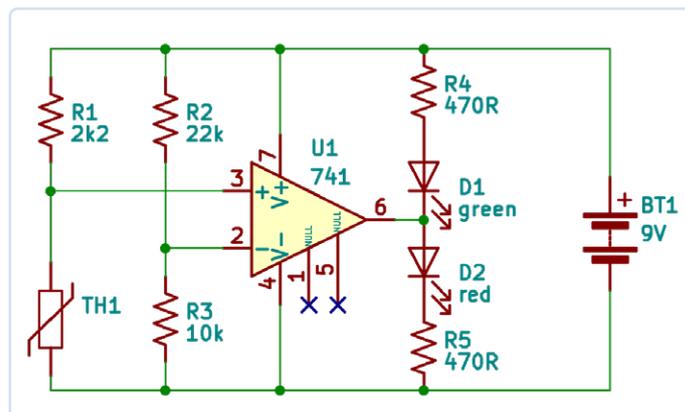


Bild 2. Ist es das, was GPT-3 zu sehen glaubte?

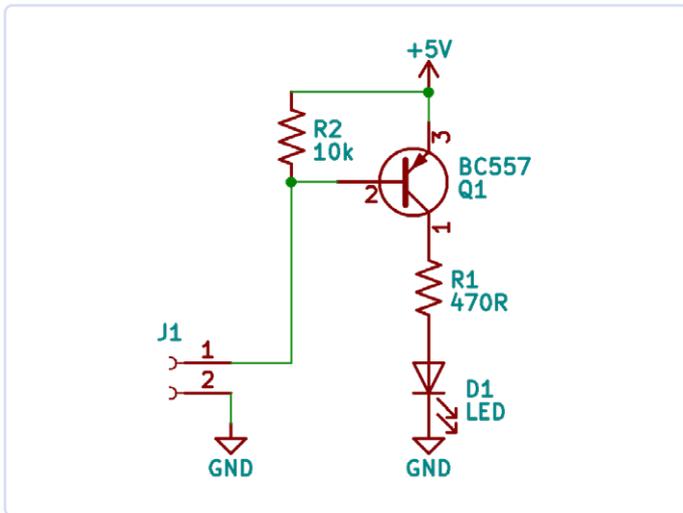


Bild 3. Offenbar hat GPT-4o die Schaltung so interpretiert.

Lassen Sie es mich wissen, wenn Sie weitere Hilfe zu einem bestimmten Teil der Schaltung benötigen oder wenn Sie ein bestimmtes Problem haben!

Das meiste oben ist richtig, bis auf ein Detail: Der Emittor des Transistors wird mit +5 V verbunden beschrieben statt mit Masse. Es scheint, dass GPT-4o die Schaltung wie in **Bild 3** dargestellt interpretiert. Die Beschreibung der erwarteten Funktionsweise der Schaltung bestätigt dies. Nachdem ich GPT auf den Fehler hingewiesen hatte, lieferte er die folgende Analyse:

Vielen Dank für diesen Hinweis! Lassen Sie uns die Analyse korrigieren: Korrigierte Schaltungsbeschreibung:

Q1 (BC557) ist ein PNP-Transistor, und zwar in dieser korrigierten Konfiguration:

- **Emittor:** Verbunden mit **Masse**.
- **Kollektor:** Ist über den Widerstand **R1** (470Ω) mit einem Ende der LED verbunden, die dann an die **+5V-Versorgung** angeschlossen ist.
- **Basis:** Verbunden mit der Verbindung von **R2** (10kΩ) und einem der Stifte des Steckverbinders **J1**.

Korrigierte Betriebsweise:

- Bei einem PNP-Transistor wie dem BC557 fließt Strom vom Emittor zum Kollektor, wenn die Basis negativer ist (niedrigere Spannung) als der Emittor.
- Wenn der Emittor geerdet ist, schaltet die Schaltung die LED ein, wenn die Basis des Transistors eine höhere Spannung als der Emittor hat, d. h., wenn die Basis im Verhältnis zum Emittor auf einen niedrigen Wert gezogen wird, schaltet sich der Transistor aus, und die LED ist aus.

Erwartetes Verhalten:

- **LED leuchtet:** Die LED sollte aufleuchten, wenn die Spannung an der Basis von Q1 ausreichend hoch ist (nahe +5 V), was bedeutet, dass der Transistor leitend ist.
- **LED aus:** Wenn die Basis von Q1 auf Masse (oder sehr nahe an

Masse) gezogen wird, hört der Transistor auf zu leiten und die LED geht aus.

[...]

Wenn Sie immer noch Probleme haben, beschreiben Sie bitte genauer, was die Schaltung tut und was Sie von ihr erwarten, dann kann ich weiterhelfen!

Die Beschreibung der Schaltung entspricht nun dem hochgeladenen Schaltplan. Die „Korrigierte Betriebsweise“ hingegen ist widersprüchlich. Zuerst wird erklärt, wie ein PNP-Transistor funktioniert, und dann geht es weiter mit einem scheinbaren NPN-Transistor. Der Abschnitt „Erwartetes Verhalten“ bestätigt dies. Als ich ChatGPT über diesen Widerspruch informierte, antwortete es mir (wieder ohne ausführliche Erklärungen):

Sie haben recht, und ich entschuldige mich für die Verwirrung in meinen früheren Erklärungen. Lassen Sie uns das richtige Verhalten für einen PNP-Transistor wie den BC557 klären.

[...]

Wenn der Emittor des BC557 an Masse angeschlossen ist, verhält sich der Transistor beim Einschalten einer LED nicht wie erwartet, es sei denn, die Schaltung ist auf eine ganz bestimmte Weise ausgelegt. Normalerweise sollte der Emittor eines PNP-Transistors an eine positive Spannung (z. B. +5V) angeschlossen werden, wobei der Kollektor mit der Last (LED und Widerstand) verbunden ist und die Basis auf einen niedrigen Wert gezogen werden sollte, um den Transistor einzuschalten.

[...]

Wenn Sie den Emittor an eine positive Spannung anschließen, wird die ursprüngliche Beschreibung der Funktionsweise der Schaltung (LED schaltet sich ein, wenn die Basis relativ zum Emittor niedrig ist) zutreffen.

Dieses Mal weist ChatGPT auf einen möglichen Fehler in der Schaltung hin. Die Lösung, die er anbietet (**Bild 4**), ist jedoch unvollständig, da die LED umgedreht sein müsste, damit sie funktioniert. Außerdem muss die positive Spannung mindestens 5,6 V betragen, damit Q1 zu leiten beginnt. Wenn J1 in der Schaltung in

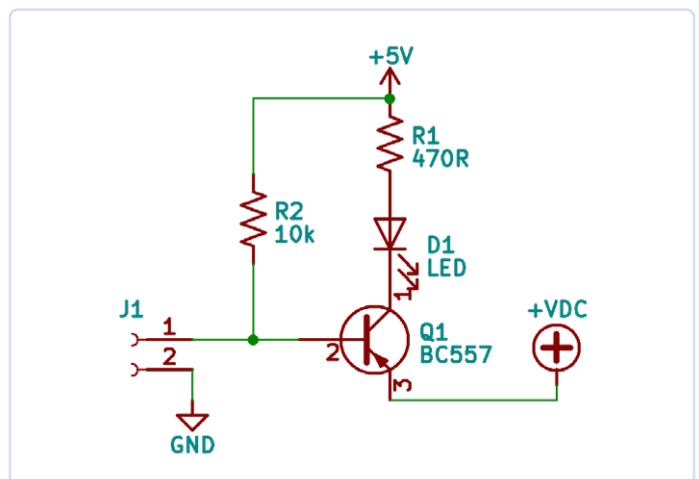


Bild 4. ChatGPT schlägt vor, den Emittor an eine positive Spannung anzuschließen: für Q1 ein gefährlicher Vorschlag.

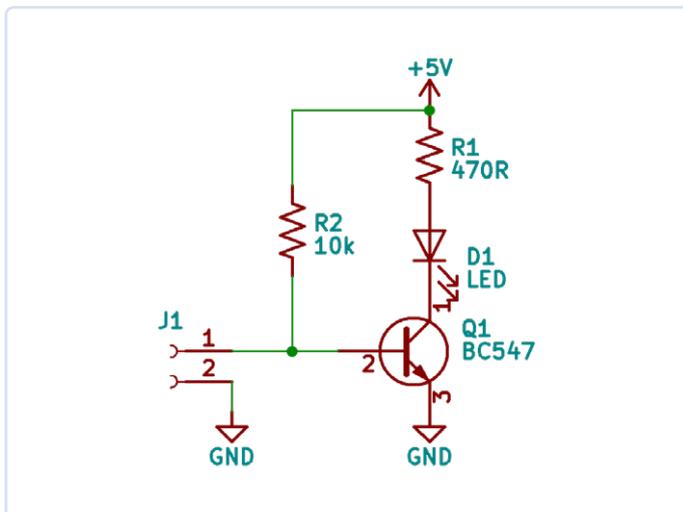


Bild 5. Die Schaltung, die wir gesucht haben. ChatGPT hat diese Lösung allerdings nicht vorgeschlagen.

Bild 4 kurzgeschlossen wird, wird der Transistor wahrscheinlich sterben, da der Basisstrom dann nur durch die Stromversorgung begrenzt ist. ChatGPT schlägt vor, Q1 durch einen NPN-Transistor zu ersetzen (**Bild 5**) und fährt fort:

Zusammenfassend lässt sich sagen, dass das Ersetzen des BC557 durch einen BC547 die Logik der Schaltung umkehrt. Mit dem BC547 schaltet sich die LED ein, wenn die Basis hochgezogen wird, und sie schaltet sich aus, wenn die Basis niedrig oder offen ist.

Leider hat ChatGPT immer noch nicht verstanden, dass J1 entweder offen oder kurzgeschlossen sein soll. Als ich ihn darauf hinwies, erhielt ich diese (gekürzte) Antwort:

Zusammenfassung der Funktionsweise:

J1 kurzgeschlossen (geschlossener Stromkreis): Die LED ist aus und zeigt an, dass der Stromkreis geschlossen ist.

J1 offen (offener Stromkreis): Die LED leuchtet und zeigt damit an, dass der Stromkreis offen ist.

Praktische Umsetzung:

Pull-up-Widerstand: Wenn J1 offen ist, könnte die Basis des BC547 schwächen, es sei denn, es gibt einen ausreichend starken Pull-up-Widerstand zu +5V. Sie können sich auf den internen Pull-up des Mikrocontrollers verlassen (falls vorhanden) oder einen externen Pull-up-Widerstand hinzufügen, um sicherzustellen, dass die Basis hochgezogen wird, wenn J1 offen ist.

Pull-down-Widerstand: Wenn Sie sicherstellen wollen, dass die Basis auf jeden Fall auf Masse gezogen wird, wenn J1 kurzgeschlossen ist, können Sie einen kleinen Pull-down-Widerstand an der Basis hinzufügen, aber das ist in der Regel nicht notwendig, wenn J1 eine feste Verbindung zur Masse ist.

Ein Mikrocontroller ist aus dem Nichts aufgetaucht und ChatGPT scheint R2 ganz vergessen zu haben. Damit war mein Experiment beendet.

### KI hat noch einen sehr langen Weg vor sich

Aus den obigen Ergebnissen können wir schließen, dass GPT-4o mit Bildern arbeiten kann, aber seine Interpretation scheint durch das, was es „weiß“, verzerrt zu sein. Dies ist eigentlich sehr menschlich. In diesem speziellen Fall hat es anscheinend die Schaltung aus Bild 3 erwartet und nicht die aus Bild 1. Doch obwohl es feststellte, dass Q1 ein PNP-Typ war, behandelte es die Schaltung so, als wäre Q1 ein NPN-Transistor. Ein Mensch würde das nie tun.

Wenn ChatGPT mit unerwarteten Eingaben konfrontiert wird, kann es zu inkohärenten Ausgaben kommen. Das erinnerte mich an den Schachcomputer meines Bruders aus den späten achtziger Jahren. Wenn man ihn wie ein Schachspieler bespielte, war er schwer zu schlagen. Wenn man jedoch einen verrückten, unerwarteten Zug machte, verlor er die Kontrolle, und man konnte ihn jedes Mal schnell schlagen. ChatGPT ist genauso. Es ist sehr beeindruckend bei der Verarbeitung von natürlicher Sprache, und es ist eine großartige Suchmaschine und ein Front-End für Wikipedia. Wenn es jedoch um die Lösung von Problemen geht, ist es nicht viel besser als ein 40 Jahre alter Schachcomputer. Tatsächlich hat die künstliche Intelligenz noch einen sehr langen Weg vor sich. ◀

RG — 240410-02

#### Sie haben Fragen oder Kommentare?

Haben Sie technische Fragen oder Anmerkungen zu diesem Artikel? Schicken Sie bitte eine E-Mail an den Autor unter [clemens.valens@elektor.com](mailto:clemens.valens@elektor.com) oder kontaktieren Sie Elektor unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



#### Passendes Produkt

> **Practical Electronics Crash Course (Bundle)**  
[www.elektor.de/20824](http://www.elektor.de/20824)

#### WEBLINKS

- [1] Video: ChatGPT als Embedded Electronics Design Engineer, Elektor TV: [https://youtu.be/pM1n\\_q7q670?si=PXL3VePdo8ytAUyd](https://youtu.be/pM1n_q7q670?si=PXL3VePdo8ytAUyd)
- [2] OpenAI: <https://openai.com/>
- [3] ChatGPT 4: <https://openai.com/index/gpt-4/>

# KI at the Edge

## mit dem ESP32-P4

Von Anant Gupta, Sun Xiangyu, und Xie Wei (Espressif)

Der ESP32-P4 ist das neueste Mitglied in der Espressif-Familie leistungstarker SoCs, die die Welt der eingebetteten Systeme revolutionieren könnten. Dieser vielseitige Chip bietet fortschrittliche Funktionen, darunter robuste KI-Fähigkeiten, die ihn zur idealen Wahl für Entwickler intelligenter und vernetzter Geräte machen. In diesem Artikel beleuchten wir die KI-Fähigkeiten des ESP32-P4, untersuchen die unterstützten KI-Bibliotheken und stellen einen praktischen Anwendungsfall vor, der das Potenzial dieses Chips verdeutlicht.

Der ESP32-P4 von Espressif ist eine leistungsstarke Dual-Core-MCU auf Basis der RISC-V-Architektur mit einer Taktrate von bis zu 400 MHz. Wie die Blockschaltung in **Bild 1** zeigt, verfügt der ESP32-P4 über ein sehr flexibles und anpassungsfähiges Speichersubsystem, das 768 KB On-Chip-SRAM, 8 KB Zero-Wait-TCM-RAM sowie erweiterbares PSRAM umfasst. Der Chip besitzt eine Vielzahl an Peripheriebausteinen, darunter SPI, I<sup>2</sup>S, I<sup>2</sup>C, LED PWM, MCPWM, RMT, ADC, UART und TWAI. Auch HMI-Funktionen wie MIPI-CSI mit integriertem ISP, MIPI-DSI

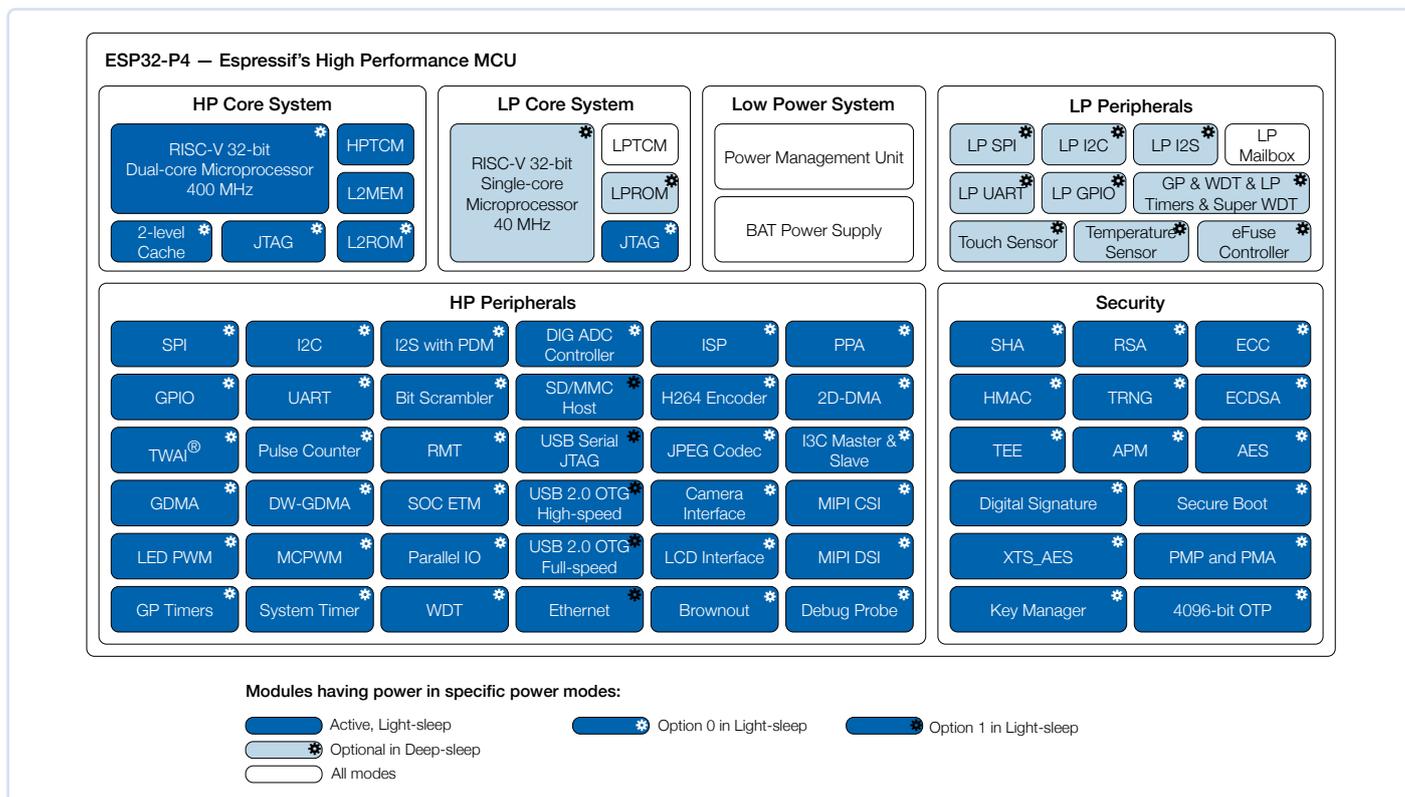


Bild 1. Die umfangreiche Blockschaltung des ESP32-P4.

und 14 kapazitive Toucheingänge sind vorhanden. Zusätzlich bietet der ESP32-P4 Hardware-Beschleuniger wie den *Pixel Processing Accelerator* (PPA), H.264-Kodierung und einen 2D-DMA. Mit diesen sehr leistungsfähigen Eigenschaften und der umfangreichen I/O-Konnektivität ist der ESP32-P4 bestens gerüstet, die Welt der Embedded-Systeme zu erobern.

## KI-Fähigkeiten des ESP32-P4

Der ESP32-P4 wurde für eine Vielzahl von KI-Anwendungen entwickelt, von einfachen Machine-Learning-Modellen bis hin zu komplexeren Deep-Learning-Algorithmen. Die RISC-V-CPU mit zwei Kernen bildet eine robuste Grundlage für die KI-Verarbeitung, die durch spezielle Hardware-Beschleuniger und KI-Befehlssätze, die komplexe Berechnungen effizient bewältigen und den ESP32-P4 zu einer idealen Wahl für leistungsintensive KI-Anwendungen machen.

Ein zentrales Merkmal der KI-Fähigkeiten des ESP32-P4 ist die Unterstützung der Xai-Erweiterung des RISC-V-Befehlssatzes. Die Xai-Erweiterung umfasst Befehle, die speziell für KI- und Machine-Learning-Anwendungen entwickelt wurden, darunter:

- **Vector instructions:** Ermöglichen dem Chip, Operationen mit großen Datenvektoren durchzuführen – ein häufiges Erfordernis in KI- und Machine-Learning-Algorithmen.
- **Matrix instructions:** Diese Befehle unterstützen Operationen mit großen Datenmatrizen, die für viele Deep-Learning-Algorithmen unerlässlich sind.

Zusätzlich bietet die Xai-Erweiterung eine Reihe weiterer Befehle, die speziell für KI- und maschinelle Lernanwendungen konzipiert sind. Dazu gehören Instruktionen für die Verarbeitung von ausgerichteten und nicht ausgerichteten Vektordaten, optimierte Datenbewegungen sowie konfigurierbare Rundungs- und Sättigungsmodi.

## ESP-SR: Framework zur Spracherkennung

ESP-SR ist ein Spracherkennungs-Framework, das Entwicklern ermöglicht, KI-gestützte Sprachlösungen zu entwickeln. Das Framework bietet verschiedene Module, darunter Audio-Frontend-Verarbeitung, Erkennung von Weckwörtern und Erkennung gesprochener Sprache. Mit ESP-SR können Entwickler Anwendungen erstellen, die auf Sprachbefehle reagieren, was es ideal für Anwendungen wie intelligente Hausautomatisierung und Sprachassistenten macht (**Bild 2**). Das ESP-SR-Framework umfasst eine Reihe von Funktionen, darunter:

- **Audio front-end processing (AFE algorithms):** Dieses Modul stellt APIs zur Audioverarbeitung bereit, einschließlich Rauschunterdrückung, Echounterdrückung und Blind-Source-Separation (Trennung vermischter Audiosignale).
- **Wake word detection (WakeNet):** Dieses Modul ermöglicht es Entwicklern, Weckwörter wie *Alexa* oder *OK Google* oder beliebige benutzerdefinierte Weckwörter zu erkennen und entsprechende Aktionen auszulösen.
- **Speech command recognition (Mul):** Dieses Modul bietet eine Reihe von APIs für die Spracherkennung, mit denen Entwickler Anwendungen erstellen können, die Sprachbefehle erkennen und darauf reagieren können.

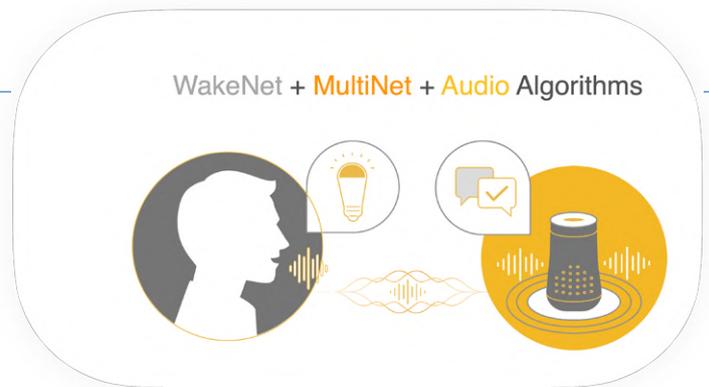


Bild 2. Die Spracherkennungsfunktion des ESP32-P4.

## ESP-DL: Bibliothek für Deep Learning

ESP-DL ist eine Deep-Learning-Bibliothek, die eine Reihe von APIs für die Inferenz neuronaler Netzwerke, Bildverarbeitung und mathematische Operationen bietet (**Bild 3**). Mit ESP-DL können Entwickler ihre Deep-Learning-Modelle auf dem ESP32-P4 einsetzen und so eine breite Palette von KI-gestützten Anwendungen schaffen. Die ESP-DL-Bibliothek bietet folgende Hauptfunktionen:

### ESP-DL Standard-Modellformat

Das Standardmodellformat von ESP-DL ist ein Binärformat, das zum Speichern des Modellgraphen, der Gewichtungen und anderer Informationen verwendet wird. Es ähnelt dem ONNX-Format, verwendet jedoch FlatBuffers anstelle von Protobuf, was die Modelle leichtgewichtiger macht und Zero-Copy-Deserialisierung ermöglicht. Diese Methode erlaubt schnellen Datenzugriff, da keine Datenkopie in den Speicher gelegt wird, wie dies bei Protobuf der Fall ist.

### Effiziente und genaue Implementierung von Operatoren

Auf Basis der KI-Befehle wurden gängige Operatoren wie Conv2D, Pool2D, Gemm, Add und Mul effizient implementiert. Zudem wurden die 8-Bit-Genauigkeitsfehler früherer Versionen korrigiert, um eine Übereinstimmung mit den Operatoren in PyTorch zu gewährleisten. Dies verbessert die Genauigkeit der Ergebnisse bei Verwendung des Quantisierungswerkzeugs in ESP-DL.

### Statischer-Speicher-Planer

Der Planer für statische Speicher simuliert die erforderliche maximale Speichergreße und die Speicher-Offsets der einzelnen Variablen auf Grundlage der topologischen Abfolge der Operatoren. Dadurch können der Zeitaufwand und die potenzielle Speicherfragmentierung, die durch die Speicherzuweisung während der Modellausführung verursacht werden, vermieden werden. Der neue Planer für statische Speicher wurde für die übliche interne RAM/PSRAM-Speicherstruktur entwickelt. Da das interne RAM zwar eine höhere Zugriffsgeschwindigkeit bietet, aber nur eine begrenzte Kapazität hat, wird eine API zur Verfügung gestellt, mit der der Benutzer die Größe des internen RAM-Speichers anpasst, den das Modell verwenden kann. Der Speicherplaner weist den verschiedenen Schichten automatisch den optimalen Speicherplatz zu, basierend auf der vom Benutzer angegebenen Größe des internen RAMs, damit die gesamte Ausführungsgeschwindigkeit effizienter ist und gleichzeitig ein Minimum an Speicherplatz verwendet wird.

### Dual-Core-Scheduling

Für rechenintensive Operatoren wurde eine automatische Dual-Core-Planung implementiert, die es den Modellen ermöglicht, die volle Rechenleistung der zwei CPU-Kerne voll auszuschöpfen. Derzeit unterstützen Conv2D und DepthwiseConv2D das Dual-Core-Scheduling.

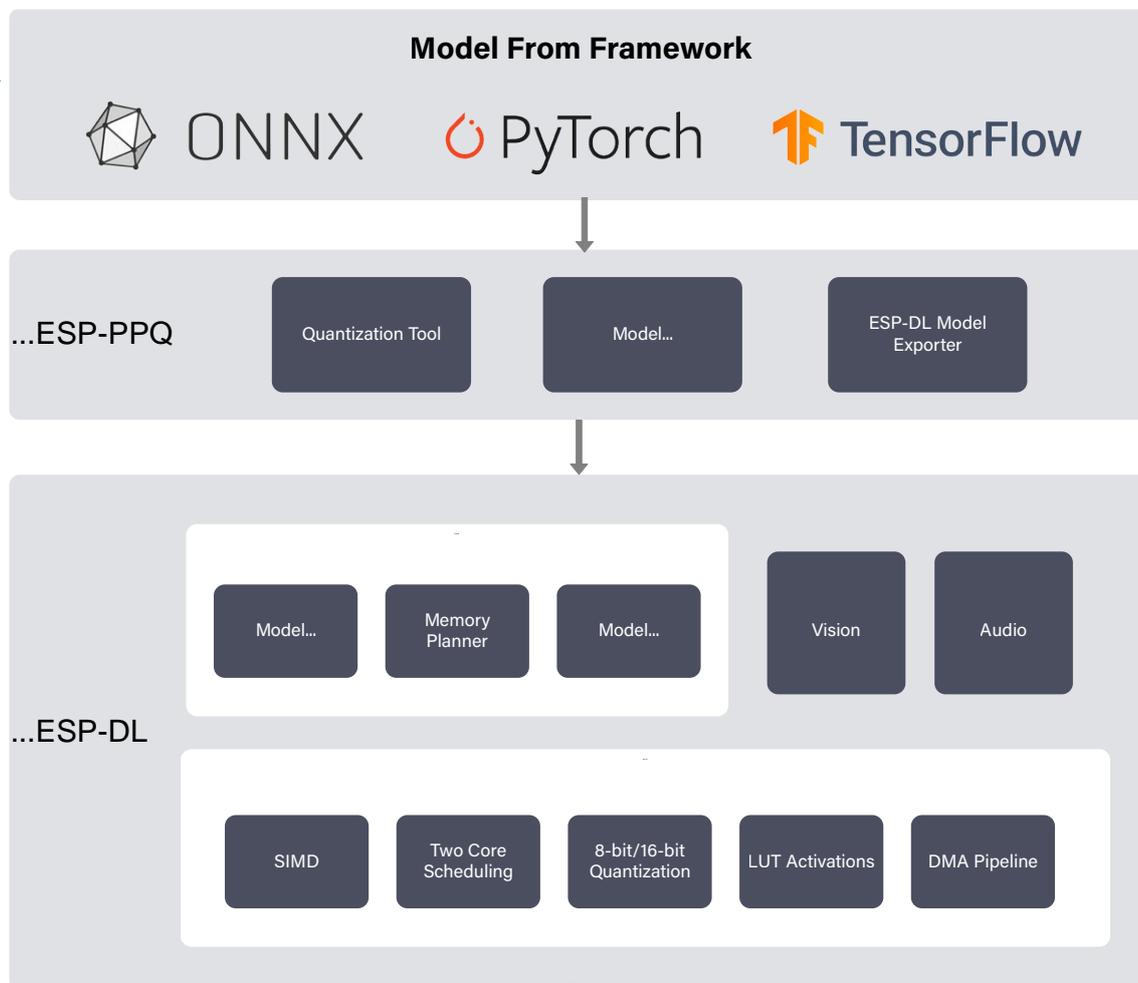


Bild 3. ESP-DL-Framework.

Tests zeigen, dass bei Layern mit hoher Rechenlast die Ausführungszeit durch Dual-Core-Nutzung um bis zu 50 % reduziert werden kann.

### Anwendungsfall: Bildklassifizierung

Ein spannender Anwendungsfall für die KI-Fähigkeiten des ESP32-P4 ist die Bildklassifizierung. Mit der ESP-DL-Bibliothek können Entwickler ein Deep-Learning-Modell einsetzen, das verschiedene Bilder erkennen und Objekte klassifizieren kann. Diese Technologie bietet vielfältige Anwendungsmöglichkeiten, von Sicherheitssystemen bis hin zu intelligenter Hausautomatisierung. Dank der leistungsstarken KI-Fähigkeiten und der umfassenden IO-Konnektivität des ESP32-P4 können Entwickler ein Bildklassifizierungssystem entwickeln, das sowohl präzise als auch effizient arbeitet.

Um Bildklassifizierung auf dem ESP32-P4 zu implementieren, können Entwickler die ESP-DL-Bibliothek nutzen, um ein vortrainiertes Modell wie MobileNet V2 einzusetzen. MobileNet V2 wird auf einem Datensatz von Bildern trainiert und ist in der Lage, verschiedene Objekte in Echtzeit zu erkennen und zu klassifizieren. Die PPA- und 2D-DMA-Engine des ESP32-P4 beschleunigt dabei die Verarbeitung und Bewegung des Videostroms, sodass der Chip komplexe KI-Verarbeitungen in Echtzeit durchführen kann.

Als Hardware für Experimente empfiehlt sich das ESP32-P4-Function-EV-Board (Bild 4), ein Multimedia-Entwicklungsboard, das auf dem ESP32-P4-Chip basiert. Das Board unterstützt USB 2.0, MIPI-CSI, MIPI-DSI und verschiedene andere Peripherie. Mit diesen umfangreichen Funktionen ist das ESP32-P4-Function-EV-Board ideal für die Entwicklung von kostengünstigen, leistungsfähigen, stromsparenden netzwerkgebundenen Audio- und Videoprodukten.

### Portierung von MobileNet V2 auf den ESP32-P4 mit ESP-DL

MobileNet V2 ist ein leichtgewichtiges Deep-Learning-Modell, das für die Leistung auf mobilen und eingebetteten Geräten optimiert ist. Es wurde in verschiedenen Bereichen eingesetzt, unter anderem zur Bildklassifizierung, Objekterkennung, Gesichtserkennung und so weiter. Nehmen wir nun MobileNet V2 als Beispiel, um zu verstehen, wie das Modell mit ESP-DL eingesetzt werden kann.

**1. Vorverifiziertes Modell vorbereiten:** Wir können ein vortrainiertes Modell direkt über die PyTorch-Schnittstelle erhalten und das Modell in eine ONNX-Datei zur Quantisierung exportieren.

```
torchvision.models.mobilenet.
mobilenet_v2(pretrained=True)
torch.onnx.export(
    model=model,
    args=tuple(
        [
            torch.zeros(
                size=[1] + input_shape[1:],
                device=self.device_str,
                dtype=self.input_dtype,
            )
            for input_shape in self.input_shape
        ]
    ),
    f=orig_onnx_path,
```

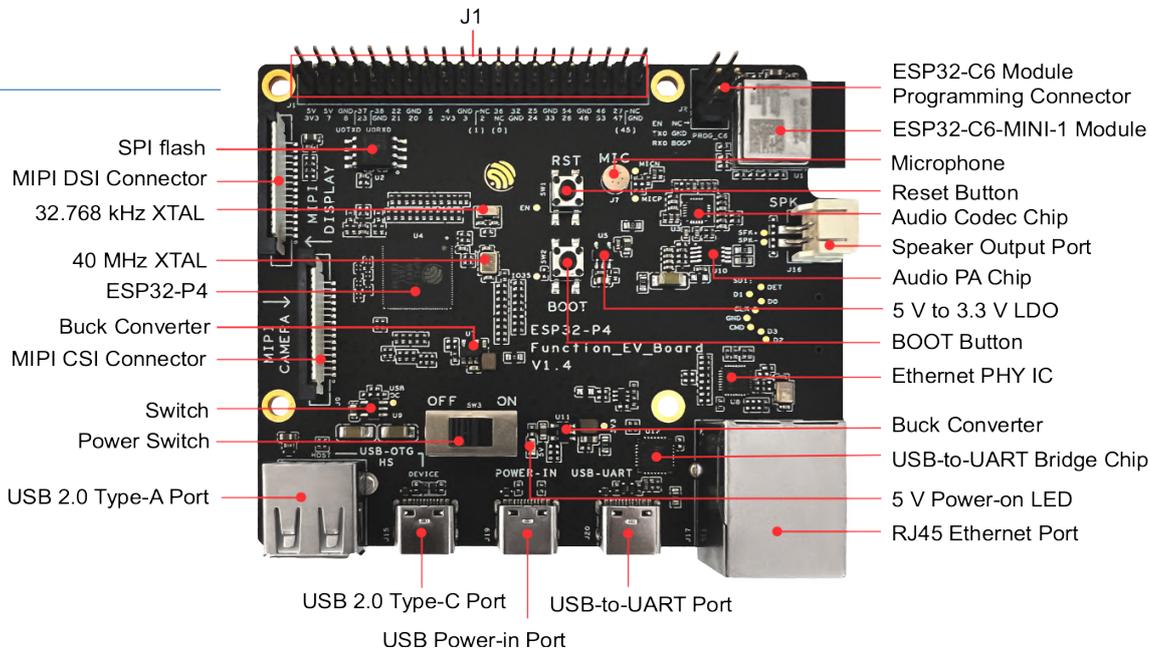


Bild 4. Was ist wo auf dem Function-EV-Board für den ESP32-P4.

```
opset_version=11,
do_constant_folding=True,
```

## 2. Vorbereitung eines Kalibrierungsdatensatzes und Quantisierung des Modells:

Verwenden Sie auf Grundlage des ESP-PPQ-Quantisierungstools die gekapselte Funktion `quantize_model_wrapper`, um die ONNX-Datei aus Schritt 1 zu laden und die Quantisierung mithilfe des Kalibrierungsdatensatzes durchzuführen. Nach Abschluss der Quantisierung exportieren Sie das Modell, um eine Datei mit der Extension `espd` zu erhalten.

```
ppq_graph, executor = quantize_model_wrapper(
    onnx_import_file=orig_onnx_path,
    calib_data_loader=self.calib_iter,
    calib_steps=self.calib_steps,
    input_shape=self.input_shape,
    platform=self.platform,
    input_dtype=self.input_dtype,
    dispatching_override=None,
    dispatching_method="conservative",
    collate_fn=collate_fn,
    device=self.device_str,
    verbose=1,
```

```
)
PFL.Exporter(platform=self.platform).export(
    file_path=export_onnx_path,
    graph=ppq_graph,
    config_path=export_config_path,
    model_version=self.model_version,
    values_for_test=values_for_test,
```

**3. Modell-Bereitstellung:** Flashen der `espd`-Modelldatei in den Flash-Speicher. Beim Laden des Modells mit ESP-DL wird automatisch der für die Zwischenschritte benötigte Speicher zugewiesen. An diesem Punkt werden die Eingabewerte in die `Model::run`-Schnittstelle

einggegeben und die endgültigen Inferenzergebnisse können mit Hilfe der `ESP-DL Accelerator Operator Library` ermittelt werden.

```
Model *model = new Model("model",
    fbs::MODEL_LOCATION_IN_FLASH_PARTITION);
std::map graph_test_inputs = get_graph_test_inputs(model);
model->run(graph_test_inputs);
std::map outputs = model->get_outputs();
```

## Die Zukunft eingebetteter KI mit dem ESP32-P4

Der ESP32-P4 ist ein leistungsstarkes SoC, das die Welt der Embedded-Systeme revolutionieren könnte. Mit seinen umfassenden KI-Fähigkeiten, der vielseitigen I/O-Konnektivität und der Unterstützung von KI-Bibliotheken wie ESP-SR und ESP-DL ist der ESP32-P4 ideal für Entwickler intelligenter und vernetzter Geräte. Der Anwendungsfall einer Bildklassifizierung mit MobileNet V2 zeigt eindrucksvoll das Potenzial der KI-Fähigkeiten des ESP32-P4 und bietet einen effektiven Weg, ein vortrainiertes Modell mithilfe des ESP-DL-Frameworks zu portieren. Diese Technologie ermöglicht die Entwicklung präziser und effizienter KI-gestützter Systeme. In einer Welt, in der sich die KI-Technologie ständig weiterentwickelt, wird der ESP32-P4 sicherlich eine Schlüsselrolle in der Zukunft der Embedded-Systeme spielen. ◀

SE – 240568-02

### Sie haben Fragen oder Kommentare?

Haben Sie Fragen oder Kommentare zu diesem Artikel? Wenden Sie sich bitte an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



### Passendes Produkt

- **Elektor-Special: Gastausgabe von Espressif 2023**  
Print: [www.elektor.de/EP-0526](http://www.elektor.de/EP-0526)  
PDF: [www.elektor.de/ED-0526](http://www.elektor.de/ED-0526)



# Sprachfunktionen auf dem Raspberry Pi Zero

Wenn Übertaktung Sprachanwendungen ermöglicht

Von Somnath Bera (Indien)

Erkunden wir Sprache-zu-Text- und Text-zu-Sprache-Funktionen auf dem Raspberry Pi Zero mit verschiedenen Softwareoptionen, Konfigurationsanleitungen und möglichen Anwendungen für Projekte der Audioverarbeitung und der Sprachsteuerung.

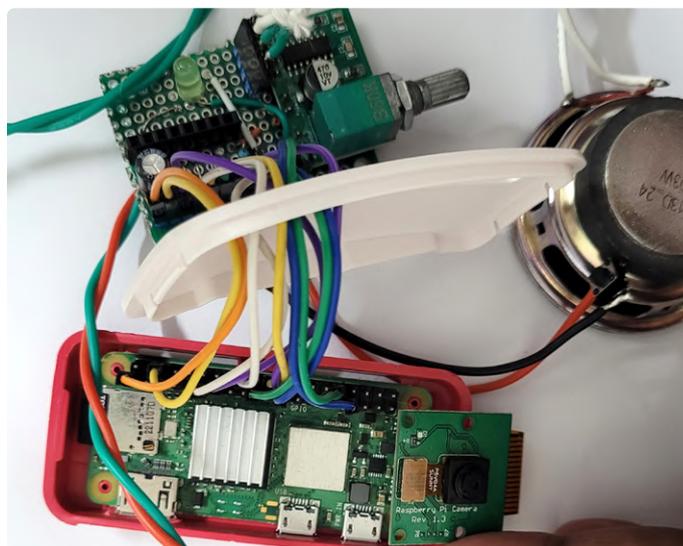


Bild 1. Detailaufnahme des Raspberry Pi Zero mit Alu-Kühlkörper.

Text-to-Speech (TTS) und Speech-to-Text (STT) sind die beiden wichtigsten Sprachfunktionen, die wir mit dem Einplatinencomputer Raspberry Pi Zero ausprobieren wollen. Dieses Board verfügt über eine moderne CPU mit 1 GHz, die - mit einem Metallkühlkörper zur passiven Kühlung ausgestattet - auf 1,3 GHz übertaktet werden kann. Für eine Übertaktung auf 1,7 GHz wäre eine aktive Kühlung erforderlich. Von der Leistung her ist dieses Board mit dem Raspberry Pi 4 vergleichbar, obwohl es weniger Speicher besitzt (nur 512 MB). Ich habe verschiedene TTS- und STT-Software auf diesem Board getestet und führe die Ergebnisse unten auf, zusammen mit einigen Projektideen für die Leser.

## Übertaktung

Da der Raspberry Pi Zero nur über einen begrenzten Arbeitsspeicher (512 MB) verfügt, bietet die Übertaktung einen erheblichen Leistungsschub, der insbesondere für TTS- und STT-Aufgaben von Vorteil ist. Es ist, als ob man zusätzliche Leistung und Ausdauer erhält; allerdings bezahlt man dies mit einer stärkeren Erwärmung und einer möglichen Instabilität der CPU. Es ist einfach und unkompliziert, den Raspberry Pi Zero zu übertakten. Durch

zwei zusätzliche Zeilen in der Datei `/boot/config.txt` können Sie sofort von der Leistungssteigerung profitieren. Die Überwachung der CPU-Temperatur ist dabei aber von entscheidender Bedeutung. Mit einem guten Metallkühlkörper kann die CPU sicher auf 1,3 GHz übertaktet werden, wobei eine Temperatur von 65...70 °C beibehalten werden sollte, ohne die Stabilität zu beeinträchtigen. Für noch höhere Taktfrequenzen bis zu 1,7 GHz wäre eine aktive Kühlung erforderlich. Schauen Sie sich den von mir verwendeten Metallkühlkörper meines Raspberry Pi Zero in **Bild 1** an. Als Erstes bearbeiten Sie die Datei `config.txt`, zum Beispiel mit dem Editor Nano:

```
sudo nano /boot/config.txt
```

und fügen Sie diese beiden Zeilen hinzu:

```
arm_freq=1300
over_voltage=2
```

Wenn Sie mit einer Übertaktung bis zu 1,7 GHz experimentieren möchten, erhöhen Sie den letzten Parameter auf den Wert 4 oder Wert 5. Speichern Sie dann die Datei und starten Sie den Raspberry Pi Nano neu, um die Änderungen zu übernehmen. Sie können die CPU-Temperatur mit dem Befehl `vcgencmd measure_temp` überwachen.

## Audio-Ausgabe auf dem Raspberry Pi Zero

Der Raspberry Pi Zero hat keine eingebaute Audio-Ausgangsbuchse. Um ein Audiosignal zu erhalten, müssen Sie ein HDMI-Gerät wie einen Fernseher anschließen. Es gibt jedoch HDMI-zu-AV-Adapter, die den Audioausgang vom HDMI trennen. Eine Alternative wäre, zwei der PWM-kompatiblen GPIO-Pins (12, 13, 18, 19) zu verwenden, um die Audioausgabe zu erzwingen, indem Sie eine weitere Zeile in die Datei `/boot/config.txt` einfügen. Öffnen Sie die Datei wieder mit `sudo nano /boot/config.txt` und fügen Sie die folgende Zeile ein:

```
dtoverlay=audremap,pins_18_13
```

Speichern Sie die Datei und starten Sie neu; die Stereoausgabe an Pin 18 und Pin 13 ist nun aktiviert. Diese Pins geben jetzt pulsweitenmodulierte Rechtecksignale mit einem logische Pegel von 3,3 V aus, die sich zwar nicht für die direkte Ansteuerung von Lautsprechern eignen, denen man aber mit einem kleinen Kopfhörer durchaus folgen kann. Natürlich können Sie auch (wie ich) ein Audioverstärker-Modul wie den PAM8403 für die Lautsprecherwiedergabe gemäß Schaltplan **Bild 2** anschließen. Der rechte und der linke Kanal werden von simplen RC-Tiefpässen gefiltert und dann über zwei 10-µF-Koppelkondensatoren mit den Eingängen des Verstärkermoduls verbunden. Zugegeben, High End ist das nicht, aber Sie können jetzt eine schöne und laute Stereoausgabe genießen! Sie können bei Youtube [1] die Schaltung in Aktion sehen, in der der Raspberry Pi Zero ein buddhistisches Lied abspielt. Der komplette Aufbau ist in **Bild 3** zu sehen.

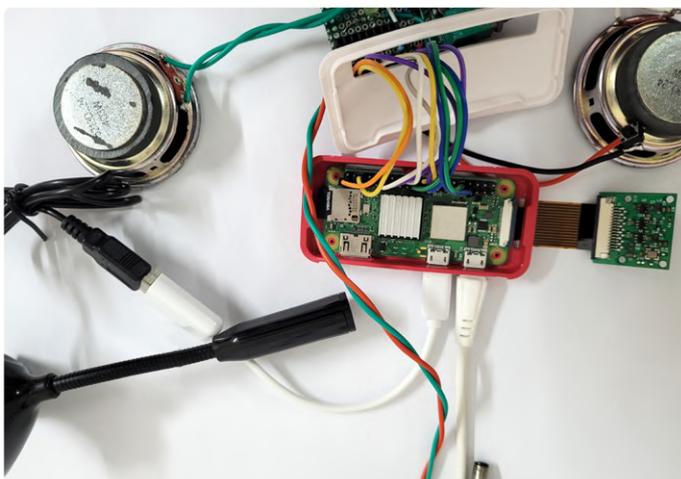


Bild 3: Kompletter Aufbau mit Mikrofon, Verstärker und Lautsprechern.

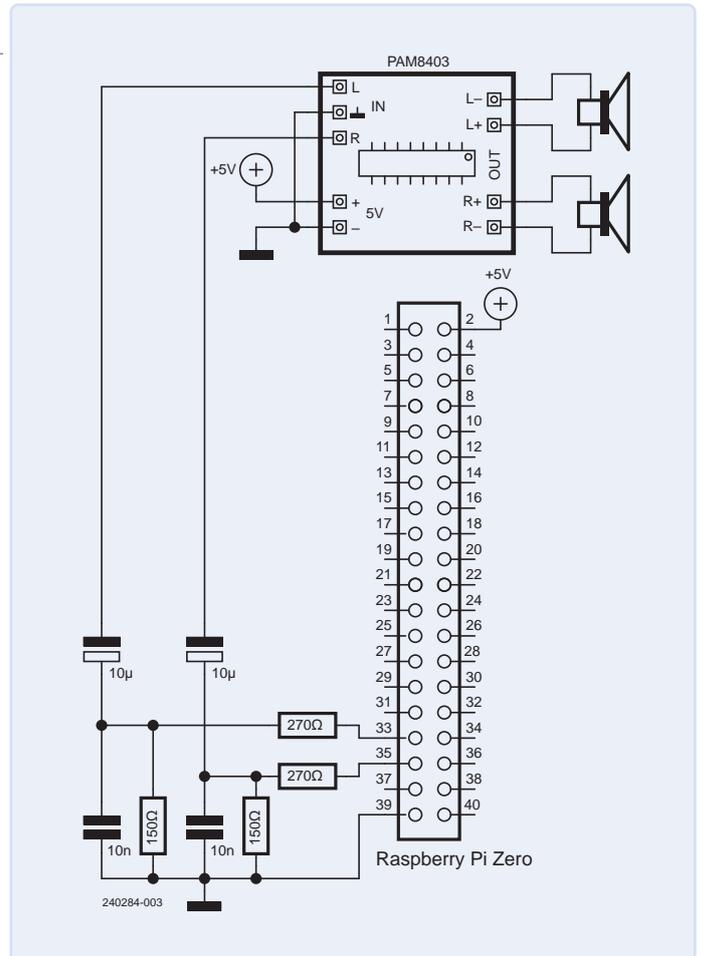


Bild 2. Schaltbild des Projekts.

## Text-to-Speech-Software

Sehen wir uns einige der Optionen für Text-to-Speech-Software an.

- > **eSpeak / eSpeak-ng**: eSpeak und die neuere Version eSpeak-ng (next generation) sind ausgereifte Open-Source-Text-to-Speech-Programme, die sowohl für Linux-basierte Systeme (Debian) als auch für Mac OS verfügbar sind. Sowohl eSpeak als auch die neueste Version eSpeak-ng sind einfach zu installieren, und Sie können sie sofort verwenden. Um eSpeak-ng auf dem Raspberry Pi Zero zu installieren, geben Sie einfach die folgenden Anweisungen in das Terminal ein:

```
sudo apt-get install espeak-ng
espeak-ng -f temp.txt
espeak-ng "Hello Readers, welcome to this project"
```

Sie sollten nun den Satz „Hello Readers...“ hören. Neben Englisch kann eSpeak-ng mehrere andere Sprachen sprechen und mehrere Stimmen verwenden. `espeak-ng -v en+f2 "Input text"` wird mit einer weiblichen Stimme ausgegeben. Lesen Sie dazu das ausführliche Handbuch mit dem Befehl `man espeak-ng`.

- > **Google Text-To-Speech (gTTS)**: gTTS erzeugt viel natürlichere Stimmen als eSpeak oder eSpeak-ng. Ich verwende es oft in meinen KI-Projekten. Sie können es so installieren:

```
pip install gtts playsound
```



## Listing 1. Google Text-to-Speech

```
from gtts import gTTS
import subprocess

def text_to_speech(text):
    # Use gTTS to convert text to speech and save to an audio file
    tts = gTTS(text=text, lang="en")
    tts.save("output.mp3")
    # Use mpg123 to play the audio file
    subprocess.run(["mpg123", "-q", "output.mp3"])

if __name__ == "__main__":
    while True:
        # Get the string to speak
        question = "Please ask your question:"
        text_to_speech(question)
```

In **Listing 1** finden Sie ein Beispiel für den Python-Code, den Sie für den Einstieg benötigen.

### Spracheingabe auf dem Raspberry Pi Zero

Bis jetzt funktioniert die Sprachausgabe perfekt. Für die Spracheingabe funktioniert ein billiges USB-Mikrofon für zwei oder drei Euro perfekt auf Anhieb (will sagen, plug-and-play). Nachdem man es eingesteckt hat, kann man mit dem Befehl `lsusb` überprüfen, ob das Mikrofon erkannt wurde. In meinem Fall ist hier die Ausgabe des Befehls:

```
Bus 001 Device 002: ID 4c4a:4155 Jieli Technology
UACDemoV1.0
Bus 001 Device 001: ID 1d6b:0002 Linux Foundation 2.0
root hub
```

Möglicherweise benötigen Sie ein OTG-Kabel, um eine Verbindung zwischen dem microUSB (am Raspberry Pi Zero) und dem USB-Anschluss des Mikrofons herzustellen. Um das Mikrofon zu testen, geben Sie folgenden Befehl im Terminal ein:

```
arecord -f S16_LE -r 33100 -d 10 -c 1 output.wav
```

Die Parameter bedeuten:

- `-f S16_LE`: Gibt das Audioformat an. `S16_LE` steht für signed 16-bit little-endian
- `-r 33100`: Legt die Abtastrate auf 33.100 Hz fest
- `-d 10`: Gibt die Dauer der Aufnahme in Sekunden an (beispielsweise 10 s)
- `-c 1`: Setzt die Anzahl der Audiokanäle auf 1 (Mono). Für Stereoaufnahmen verwenden Sie `-c 2`.
- `output.wav`: Gibt den Namen der Ausgabedatei an (in diesem Fall: `output.wav`).

Nach der Ausführung dieses Befehls nimmt `arecord` 10 s lang Audio mit einer Abtastrate von 33,1 kHz auf und speichert es als `output.wav` im aktuellen Verzeichnis. Sie können Optionen wie

Dauer (`-d`), Name der Ausgabedatei und andere Parameter an Ihre Wünsche anpassen.

Um die aufgenommenen Audiodaten abzuspielen, führen Sie `aplay output.wav` aus. Achten Sie darauf, dass sich das Mikrofon nicht in unmittelbarer Nähe des Lautsprechers befindet, um Rückkopplungen zu vermeiden.

So weit, so gut! Sie haben Ihre Audioquelle eingerichtet, die Text-to-Speech-Software auf dem Raspberry Pi Zero installiert und sowohl die Audioaufnahme als auch die Umwandlung von Text in Sprache mit `eSpeak`, `eSpeak-ng` oder `gTTS` erfolgreich ausprobiert. Kommen wir nun zur Sprache-zu-Text-Konvertierung.

### Speech-to-Text-Software

Google bietet einen kostenpflichtigen Speech-to-Text-Dienst an, für den eine Internetverbindung erforderlich ist. Für die umgekehrte Anforderung (Text-to-Speech) bietet Google auch eine Online-API an, die bis zu einer bestimmten Grenze kostenlos genutzt werden kann; darüber hinaus ist ein kostenpflichtiges Konto erforderlich. Für diesen Artikel werden wir auf dem Raspberry Pi Zero eine Offline-Speech-to-Text-Software verwenden, die recht gut funktioniert, vorausgesetzt, die Sprache ist klar genug. Es gibt eine solche Software von Google (siehe unten), aber auch zwei großartige Open-Source-Programme für die Umwandlung von Sprache in Text: das kommandozeilenbasierte `spchcat` und das bibliotheksbasierte `Pocketsphinx`. Beide können mit Python mit unterschiedlichen Methoden verwendet werden. Das Modul `subprocess` wird für den Aufruf von `spchcat` verwendet, während das Modul `pocketsphinx` direkt in Python mit `pip` installiert werden kann.

### Spchcat

`Spchcat` ist eine Open-Source-Software, die das TensorFlow-Erkennungsmodell verwendet. Sie wird derzeit unter Linux auf dem Raspberry Pi unterstützt und kennt 46 Sprachmodelle, darunter auch indische Sprachen wie Tamil und Bengali (meine Muttersprache). Die Software kann zusammen mit einem Leitfaden von [2] heruntergeladen werden.

Die Software `spchcat` ist ein Kommandozeilen-Tool, das Audio direkt aus WAV-Dateien verarbeitet. Es ist ein großes Programm,

etwa 1,2 GB, und kann nicht direkt auf einem Raspberry Pi Zero installiert werden. Um diese Einschränkung zu umgehen, gehen Sie wie folgt vor:

- 1. Herunterladen:** Laden Sie zunächst die *spchcat*-Software von der GitHub-Seite herunter
- 2. Übertragen:** Mit *scp* (secure copy) übertragen Sie die heruntergeladene Datei auf die SD-Karte Ihres Raspberry Pi Zero
- 3. Installieren:** Da der Raspberry Pi Zero die Installation aufgrund seiner beschränkten Ressourcen nicht durchführen kann, entfernen Sie die SD-Karte und legen Sie sie in einen Raspberry Pi B+ oder Raspberry Pi 4 ein.
- 4. Installation auf Pi 4:** Mit einem Doppelklick auf die heruntergeladene Software starten Sie die Installation auf dem Raspberry Pi 4, was etwa 25 Minuten dauert
- 5. Letzter Schritt:** Nach der Installation stecken Sie die SD-Karte wieder in Ihren Raspberry Pi Zero, starten Sie *spchcat* und überprüfen Sie, ob die Software funktioniert

Dies ist ein cleverer Trick, um die Ressourcenbeschränkungen des Raspberry Pi Zero zu umgehen!  
Nach der Installation können Sie mit der Audioaufnahme von der Standardmikrofonquelle beginnen, indem Sie die Software ohne Argumente starten. Die Ergebnisse werden direkt in das Terminal ausgegeben:

```
spchcat
```

Wenn kein Mikrofon angeschlossen ist oder Audio von einem anderen Programm transkribiert werden soll, können Sie dem Argument `--source` den Wert `system` zuweisen. Dabei wird versucht, das Audio, das Ihr Rechner abspielt (einschließlich Videos), abzuhören und die erkannte Sprache zu transkribieren:

```
spchcat --source=system
```

Starten Sie nun ein YouTube-Video oder spielen Sie ein beliebiges Audiosignal auf Ihrem System ab. Die Transkription der Sprache wird im Terminalfenster angezeigt, während das Audiosignal abgespielt wird.  
Um eine WAV-Datei mit *spchcat* zu bearbeiten, verwenden Sie den folgenden Befehl:

```
spchcat /home/bera/myaudio.wav
```

Nach einer gewissen Verarbeitungszeit wird der transkribierte Text im Terminalfenster angezeigt. Um die Sprache für die Transkription anzugeben, verwenden Sie das Argument `--language`, zum Beispiel hier für Deutsch oder US-Englisch:

```
spchcat --language=de_DE  
spchcat --language=en_US
```

Seien Sie vorsichtig, *spchcat* ist ressourcenintensiv und kann die CPU erheblich aufheizen. Die Überwachung der CPU-Temperatur ist



## Stückliste

Die meisten dieser Komponenten sind mit Ausnahme des Raspberry Pi Zero (siehe Passende Produkte) sehr günstig im Internet bei den „üblichen Verdächtigen“ erhältlich.

- Raspberry Pi Zero 2 W
- PAM8403-Verstärkermodul
- 4-Ohm-Lautsprecher (2 Stück)
- USB-Mikrofon
- USB-OTG-Kabel

sehr wichtig! Verwenden Sie deshalb einen hochwertigen Kühlkörper auf der CPU des Raspberry Pi Zero, um die während des Betriebs entstehende Wärme im Zaum zu halten.

Sie können *spchcat* mit beispielhaften WAV-Dateien ausprobieren, die Sie nicht selbst vorbereiten müssen, sondern etwa von [3] heruntergeladen werden können. Mit diesen Dateien lassen sich die Transkriptionsergebnisse gut bewerten und die Qualität des Aufnahmevorgangs zu verstehen, die für eine genaue Transkription erforderlich ist.

Der Befehl gibt die Texttranskription aus. Um diese Ausgabe in einer Datei zu speichern, verwendet Linux typischerweise das Zeichen >:

```
spchcat your-audio-file.wav > /tmp/transcript.txt
```

Vergewissern Sie sich, dass Python 3.7 oder höher und auch das Modul *subprocess* installiert ist, denn Sie müssen das Modul *subprocess* einsetzen, um *spchcat* in einem Python-Skript zu verwenden. Wenn letzteres noch nicht der Fall ist, können Sie das Modul mit `pip install subprocess` nachinstallieren.

Auf dem Raspberry Pi Zero kann *spchcat* bis zu vier Minuten Aufnahme und Transkription von WAV-Dateien verarbeiten, auf dem Raspberry Pi 4 ist eine Nonstop-Live-Transkriptionen von Quellen wie YouTube-Videos möglich. Die Qualität der Aufnahme ist entscheidend, um genaue Ergebnisse zu erzielen. Wenn das Mikrofon nahe der Tonquelle platziert wird, verbessert sich in der Regel die Genauigkeit der Textausgabe.

## PocketSphinx

*PocketSphinx* ist eine von der *Carnegie Mellon University* (Pittsburgh, USA) entwickelte, leichtgewichtige Open-Source-Spracherkennungs-Engine. Sie liest einkanalige 16-Bit-PCM-Audiodaten von inem Mikrofon oder aus Dateien und versucht, Sprache anhand von Standard-Akustik- und Sprachmodellen zu erkennen. Im Gegensatz zum komplexen *spchcat* ist *pocketsphinx* eine leichtere, einfachere Option. Es unterstützt mehrere Sprachen, darunter Englisch, Chinesisch, Französisch, Spanisch, Deutsch und Russisch. Ich habe es jedoch nur mit Englisch getestet. *Pocketsphinx* kann WAV-Dateien direkt in Text umwandeln, aber dieser Prozess kann bei Verwendung von Python recht langsam sein. Um *pocketsphinx* sowie das englische Sprachmodell zu installieren, verwenden Sie die folgenden Befehle:

```
pip install pocketsphinx
sudo apt-get install -y python3-pocketsphinx
pocketsphinx-en-us
```

In Python wird die Klasse `LiveSpeech` aus dem Modul `pocketsphinx` verwendet, um die Spracherkennung durchzuführen. Der Timeout-Wert für Stille, hier auf 1.0 gesetzt, kann nach Bedarf angepasst werden. Hier ein (sehr einfaches) Python-Programm zum Testen von `pocketsphinx`:

```
from pocketsphinx import LiveSpeech
speech = LiveSpeech()
speech = LiveSpeech(silence_timeout=1.0)
print("Listening...")
for phrase in speech:
    print(f"Transcript: ")
```

Die Live-Transkription funktioniert sowohl auf dem Raspberry Pi Zero als auch auf dem Raspberry Pi 4. Auf dem Raspberry Pi Zero kann es passieren, dass `pocketsphinx` gelegentlich Wörter auslässt, während es auf dem Raspberry Pi 4 einwandfrei funktioniert. Insgesamt habe ich festgestellt, dass die Qualität der Transkription von `pocketsphinx` meist nicht so gut ist wie die von `speechcat`.

## Google Speech-to-Text

Zu guter Letzt: Google Speech-to-Text (STT) unterstützt 125 Sprachen und ist sehr effizient, was es zu einer guten Option für Geräte mit begrenzter Kapazität macht, wie auch den Raspberry Pi Zero. Um Google STT zu verwenden, können Sie die Python-Bibliothek `SpeechRecognition` installieren, die die Nutzung ermöglicht:

```
pip install SpeechRecognition
```

Ein einfaches Programm zum Einstieg in diese Bibliothek ist in Listing 2 zu sehen. Es verwendet Google Speech-To-Text. Das Tolle an der `SpeechRecognition`-Bibliothek ist, dass sie nicht nur Spracherkennungsprogramme von Google, sondern auch viele andere unterstützt. Wenn Sie zum Beispiel `pocketsphinx` verwenden möchten, können Sie in **Listing 2** einfach die Zeile 15 von

```
text = recognizer.recognize_google(audio) in
text = recognizer.recognize_sphinx(audio) ändern.
```

**Listing 3** zeigt, wie man diese Bibliothek mit der Spracherkennungseingabe von Google verwendet, um Text aus einer Datei zu transkribieren. Ich empfehle Ihnen dringend, auch das auszuprobieren! Natürlich ist es möglich, sowohl die Audioaufnahme



### Listing 2. Python-Bibliothek `SpeechRecognition`

```
import speech_recognition as sr
def speech_to_text():
    # Initialize the recognizer
    recognizer = sr.Recognizer()
    # Use the default microphone as the audio source
    with sr.Microphone() as source:
        print("Speak something...")
        audio = recognizer.listen(source)
    try:
        print("Recognizing...")
        # Use Google Web Speech API to perform speech recognition
        text = recognizer.recognize_google(audio)
        return text
    except sr.UnknownValueError:
        print("Could not understand audio")
        return ""
    except sr.RequestError as e:
        #print(f"Error: ")
        return ""
if __name__ == "__main__":
    while True:
        # Call the speech_to_text function to get text from speech
        result = speech_to_text()
        print(f"You said: ")
```



### Listing 3. Spracherkennung von einer Audiodatei

```
import argparse
import speech_recognition as sr
def speech_to_text(audio_file):
    # Initialize the recognizer
    recognizer = sr.Recognizer()
    # Load the audio file
    with sr.AudioFile(audio_file) as source:
        audio_data = recognizer.record(source)
    try:
        # Use Google Speech Recognition for speech-to-text
        text = recognizer.recognize_google(audio_data)
        return text
    except sr.UnknownValueError:
        print("Google Speech Recognition could not understand the audio")
        return ""
    except sr.RequestError as e:
        print(f"Could not request results from Google Speech Recognition service; ")
        return ""
if __name__ == "__main__":
    # Create argument parser
    parser = argparse.ArgumentParser(description="Convert audio file to text using Google Speech
Recognition")
    parser.add_argument("audio_file", help="Path to the audio file")
    parser.add_argument("destination_file", help="Path to save the text output")
    # Parse command-line arguments
    args = parser.parse_args()
    # Call the speech_to_text function to convert speech to text
    result = speech_to_text(args.audio_file)
    if result:
        # Save the recognized text to the destination file
        with open(args.destination_file, "w") as f:
            f.write(result)
        print(f"Recognized text saved to ")
    else:
        print("Speech recognition failed.")
```

als auch die Transkription in ein und demselben Python-Skript durchzuführen: In **Listing 4** werden zehn Sekunden Audio aufgenommen, in denen Sie eine Frage stellen, und dann die Texttranskription der gesprochenen Frage ausgegeben.

Ich habe Googles Speech-to-Text (STT) und Text-to-Speech (gTTS) erfolgreich in verschiedenen Projekten mit dem Raspberry Pi Zero eingesetzt, auch in Mikroprozessor-basierten Projekten wie dem *ESP32 ChatGPT Terminal* [4][5], wo es gesprochene Antworten über einen I<sup>2</sup>S-Lautsprecher liefert.

#### Zukünftige Anwendungsfälle?

Die Umwandlung von Text in Sprache und umgekehrt ist eine wichtige Aufgabe. Große Medienunternehmen verwenden teure

Software und leistungsstarke Computer, um diese Aufgaben während Live-Übertragungen in Echtzeit durchzuführen. Die Umsetzung dieser Aufgaben auf einem Raspberry Pi Zero mag zwar primitiv erscheinen, bietet aber bemerkenswerte Vorteile, wie zum Beispiel eine kompakte Größe, die Möglichkeit, mit simplen 5-V-Stromversorgungen (mobil) betrieben zu werden, und eine geringe Stromaufnahme.

Hier sind ein paar potenzielle Anwendungsfälle, die ich für die Sprachtranskription auf dem Raspberry Pi Zero empfehle:

- › **Sprechende Enzyklopädie:** Ein Freisprech-Terminal, dem man eine Frage stellt und sie bestätigt. Der Raspberry Pi Zero ruft die Antworten von einem Dienst wie ChatGPT ab und



## Listing 4. Spracherkennung und Audioaufnahme

```
import subprocess
import speech_recognition as sr

def record_audio(file_name, duration=10):
    # Record audio from USB microphone using arecord
    command = ["arecord", "-D", "plughw:1,0", "-f", "S16_LE", "-r", "16000", "-d", str(duration), file_name]
    subprocess.run(command)

def speech_to_text(file_name):
    recognizer = sr.Recognizer()
    # Load the audio file for recognition
    with sr.AudioFile(file_name) as source:
        audio_data = recognizer.record(source)
    text = recognizer.recognize_google(audio_data)
    return text.lower() # Convert to lowercase for easier comparison

if __name__ == "__main__":
    while True:
        # Record audio for question
        record_audio("question.wav", duration=10) # Increased duration for longer responses
        speech_text = speech_to_text("question.wav")
        print(speech_text)
```

liest sie dann laut vor. Das alles lässt sich mit einem Raspberry Pi Zero bewerkstelligen. Erfahren Sie mehr: [6].

- ▶ **Gesprochene Ausgabe bei KI-Projekten:** Bei KI-Projekten, wie denjenigen, die ich mit Edge Impulse zur Objektklassifizierung durchgeführt habe, können die Klassifizierungsergebnisse (Zahl oder Name) vom Gerät mit eSpeak oder eSpeak-ng gesprochen werden. So müssen Sie Ihre Aufmerksamkeit nicht auf einen Monitor richten. Ein Beispiel finden Sie unter [7].
- ▶ **Terminal für Menschen mit Sehbehinderungen:** Ein System, bei dem eine sehbehinderte Person Fragen stellen kann und der Raspberry Pi verschiedene Websites, einschließlich OpenAI, durchsucht, um gesprochene Antworten zu geben.
- ▶ **Robotik:** Ideal für sprachgesteuertes Roboterspielzeug, insbesondere für internationale Märkte. Er kann fremdsprachige Befehle zur weiteren Verarbeitung in Englisch oder Deutsch umwandeln.
- ▶ **Sprachgesteuerte Handheld-Barcodescanner:** Nützlich in Supermärkten, wo das Gerät Strich- oder QR-Codes liest und die Ergebnisse über angeschlossene Lautsprecher ansagt. Die Mehrsprach-Fähigkeit von eSpeak-ng könnte bei solchen Systemen von großem Vorteil sein.
- ▶ **Sprachassistent für Menschen mit Hörbehinderung:** Ein tragbarer Sprachassistent, der die Echtzeit-Transkription gesprochener Sätze ermöglicht.
- ▶ **Assistent für Reisende:** Erleichtert die Kommunikation in fremden Ländern, indem er Interaktionen mit Einheimischen wie Taxifahrern oder Kellnern übersetzt und vokalisiert.
- ▶ **Sprachgesteuerter Musikplayer:** Ein tragbarer Musikplayer, der durch Sprachbefehle mit dem Benutzer interagiert, ähnlich wie Amazon Alexa.

Dies ist natürlich nur die Spitze des Eisbergs. Welche sprachbasierten Anwendungen können Sie sich mit Nanocomputern wie dem Raspberry Pi vorstellen? ◀

SE — 240284-02



### Über den Autor

Somnath Bera, ein Maschinenbauingenieur am indischen Jalpaiguri Government Engineering College, arbeitet als General Manager bei NTPC, dem größten Stromerzeuger des Landes. Er hat eine tiefe Leidenschaft für Elektronik, was durch seine mehr als 60 innovativen Projekte bei Elektor-Labs bewiesen wird, von denen mehr als zehn auch in der Zeitschrift Elektor veröffentlicht worden sind. Seine Projekte beschäftigen sich mit Problemlösungen in der Abfallwirtschaft und der Bewirtschaftung natürlicher Ressourcen. Somnath verwendet gerne innovative Ansätze und Plattformen wie Arduino, Raspberry Pi und ESP32 in Verbindung mit verschiedenen Arten von Sensoren und drahtlosen Systemen, um effiziente und kostengünstige Lösungen zu schaffen.

### Sie haben Fragen oder Kommentare?

Wenn Sie technische Fragen oder Anmerkungen zu diesem Artikel haben, wenden Sie sich bitte per E-Mail an den Autor unter [berasomnath@gmail.com](mailto:berasomnath@gmail.com) oder an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



## Passende Produkte

- **Raspberry Pi Zero 2 WH (mit Header)** [www.elektor.de/20952](http://www.elektor.de/20952)
- **John Allwork, Programming Voice-controlled IoT Applications with Alexa and Raspberry Pi (Elektor, 2023)**  
Taschenbuch, englisch: [www.elektor.de/20400](http://www.elektor.de/20400)  
E-Buch, PDF, englisch: [www.elektor.de/20401](http://www.elektor.de/20401)



## WEBLINKS

- [1] Beispiel einer Audiodatei, die auf dem Raspberry Pi Zero abgespielt wird: <https://youtu.be/YXC7VIVIX9c>
- [2] Hilfe und Dateilistings für spchcat: <https://github.com/petewarden/spchcat>
- [3] Beispielhafte WAV-Dateien zum Testen mit spchcat: [https://www.voiptroubleshooter.com/open\\_speech/](https://www.voiptroubleshooter.com/open_speech/)
- [4] „ESP32-ChatGPT terminal“ auf Elektor Labs: <https://www.elektormagazine.de/labs/esp32-chatgtp-terminal>
- [5] Somnath Bera, „Ein persönliches KI-Terminal“, Elektor Gastausgabe KI, 2024: <http://www.elektormagazine.de/230536-02>
- [6] „Raspberry Pi Zero Talking Encyclopedia“ auf Elektor Labs: <https://www.elektormagazine.de/labs/raspberry-pi-zero-encyclopedia>
- [7] Transkribieren von Text in KI-Projekten: <https://www.youtube.com/watch?v=C5QrCl4XIJU>

# Jede Bewertung spiegelt ein persönliches Erlebnis wider

elektor.de

**Top Service** Updated Jun 15, 2024  
Schnelle Lieferung, guter Preis. Leider war das teure Buch schlecht verpackt (lose im zu großen Karton), so dass es ziemlich angestoßen bei mir ankam. Das Buch wurde anstandslos ersetzt, so geht guter Service!  
Date of experience: June 11, 2024

**Elektor, das Elektronik-DIY El Dorado** Jun 3, 2024  
Elektor ist ein El Dorado für Elektronik-DIY-Fans. Es gibt unzählige Projekte und deren Shop bietet kostengünstige Produkte.  
Date of experience: May 31, 2024

**Elektor lieferte schnell** Jun 3, 2024  
Elektor lieferte schnell, preiswert und zuverlässig wie immer!  
Date of experience: June 03, 2024

**Perfekt** Feb 12, 2024  
Perfekt. Schnelle Lieferung, sehr gut verpackt. Artikel im perfekten Zustand. Danke!  
Date of experience: February 01, 2024

Wir lieben Elektronik und Projekte, und wir setzen alles daran, die Bedürfnisse unserer Kunden zu erfüllen  
Der Elektor-Store: **'Never expensive, always surprising'**

## Elektor Store

Reviews 365 • Excellent



VERIFIED COMPANY

Sehen Sie sich weitere Bewertungen auf unserer Trustpilot-Seite an: [www.elektor.com/TP/de](http://www.elektor.com/TP/de)

Oder bilden Sie sich selbst eine Meinung und besuchen Sie unseren Elektor Store, [www.elektor.de](http://www.elektor.de)



# Die wachsende Rolle von Edge-KI

Ein Trend, der die Zukunft prägt



Bild 1. Das Entwicklungskit STWIN.box für fortschrittliche industrielle Sensorik, Zustandsüberwachung und vorausschauende Wartung.

Von Vianney Taufour (STMicroelectronics)

Im Zuge der technologischen Entwicklung dringen die Fähigkeiten der künstlichen Intelligenz über die Grenzen des Cloud-Computing hinaus in die Welt der Edge-Geräte vor. Edge-KI steht für eine starke Verlagerung hin zu dezentraler Echtzeit-Intelligenz auf Geräten wie Sensoren, Motoren, Kameras und Fahrzeugen. Durch die lokale Verarbeitung der Daten verringert Edge-KI die Latenzzeit, verbessert den Datenschutz und ermöglicht Entscheidungen in Echtzeit, was für viele Industrie- und Verbraucheranwendungen von entscheidender Bedeutung ist.

Der Aufwärtstrend der Edge KI wird durch das Zusammenwachsen eingebetteter Systeme, maschineller Lernalgorithmen und einer Vielzahl von Sensoren vorangetrieben. Von der Erkennung von Anomalien in Industriemotoren bis hin zur Gesichtserkennung in Sicherheitssystemen - die Kreativität, die erforderlich ist, um aussagekräftige Informationen aus diesen Sensoren

zu extrahieren, bringt einige der innovativsten KI-Anwendungen in eingebetteten Systemen voran. In dem Maße, in dem Ingenieure und Entwickler danach streben, mit KI neue Potenziale zu erschließen, wird die Vielseitigkeit von Edge-KI immer deutlicher. Werfen wir einen genaueren Blick darauf, wie dieser Trend in verschiedenen realen Anwendungen zum Tragen kommt.

## Edge-KI: Anwendungsfälle aus der Praxis

### Erkennung von Anomalien in der Motorleistung

Edge-KI ermöglicht die Echtzeit-Erkennung von Anomalien bei Motoren, die mit unterschiedlichen Geschwindigkeiten arbeiten. Durch den Einsatz von Schwingungssensoren, zum Beispiel dem digitalen 3-Achsen-Schwingungssensor mit extrem breiter Bandbreite und geringem Rauschen aus unserem Entwicklungskit STWIN.box (Bild 1), können KI-Modelle Leistungsabweichungen erkennen, die potenzielle mechanische Probleme signalisieren. Dies ist von unschätzbarem Wert in industriellen Bereichen, in denen die frühzeitige Erkennung von Anomalien kostspielige Ausfälle verhindern und die Effizienz von Wartungsplänen verbessern kann. Dieser Praxisfall ist auch deshalb interessant, weil der Algorithmus zur Erkennung von Anomalien auf dem Gerät selbst neu trainiert werden kann, um sich an das genaue Schwingungsmuster anzupassen und die bestmögliche Genauigkeit für jede Maschine individuell zu erreichen. Stellen Sie sich eine Fabrik mit Hunderten von Maschinen vor, in der Sie das Modell für jede einzelne Maschine feinabstimmen können! Lesen Sie mehr unter [1].

### Erkennung und Klassifizierung von Motorfehlern

In der Industrie sind Motoren kritische Komponenten, die alles Mögliche antreiben, von Förderbändern bis hin zu Roboterarmen. Die Anwendung von Edge-KI auf Motoren ermöglicht durch Analyse von Daten wie Stromaufnahme

oder Vibrationen eine Echtzeit-Fehlererkennung. Hoch entwickelte Algorithmen können sogar die Art des Fehlers klassifizieren, ob es sich um eine Unwucht, eine Fehlausrichtung oder mechanischen Verschleiß handelt, was ein präzises und rechtzeitiges Eingreifen ermöglicht. Lesen Sie mehr unter [2].

### Erkennung von DC-Störlichtbögen in elektrischen Systemen (AFCI)

Solaranlagen und andere elektrische Gleichstromsysteme sind mit besonderen Problemen verbunden, zu denen auch das Risiko von Störlichtbögen gehört. Durch den Einsatz von Edge-KI können diese Systeme Störlichtbögen in Echtzeit erkennen, was die Sicherheit erhöht und Brände oder kostspielige Schäden verhindert. Diese Anwendung unterstreicht die wachsende Rolle der KI für eine höhere Zuverlässigkeit der Infrastruktur erneuerbarer Energien. Der Einsatz von KI bringt in diesem Fall eine bessere Treffgenauigkeit als der aktuell verwendete algorithmische Ansatz, was für die Vermeidung von Fehlalarmen und die Verringerung von Ausfallzeiten entscheidend ist. Lesen Sie mehr unter [3].

### Virtuelle Drucksensoren für mehr Sicherheit bei Fahrrädern mit Elektrounterstützung

Herkömmliche Sensoren können durch virtuelle, KI-gesteuerte Sensoren ersetzt oder ergänzt werden, die in komplexen Systemen auf Druck- oder Kraftwerte folgern können. Bei Fahrrädern mit Elektrounterstützung kann Edge-KI beispielsweise die Sicherheit erhöhen, indem sie Druckänderungen auf der Grundlage der Stromabgabe der Batterie für den Antrieb des Elektromotors vorhersagt. Dies ermöglicht intelligentere Steuerungssysteme, die die Leistung ohne Kostenaufwand optimieren und gleichzeitig die Sicherheit des Fahrers gewährleisten. Lesen Sie mehr unter [4].

### Fußbodenbelag-Erkennung für intelligente Staubsauger

In intelligenten Haushaltsgeräten wie Staubsaugern wird Edge-KI eingesetzt, um verschiedene Fußbodenbeläge (Teppiche, Hartholz oder Fliesen) zu erkennen und die Reinigungsmodi entsprechend anzupassen. Der KI-Algorithmus ist in der Lage, mit einem einfachen ToF-Kamerasensor harte von weichen Böden unter 40 verschiedenen

Arten von Belägen mit hoher Trefferquote zu unterscheiden. Lesen Sie mehr unter [5].

### Klassifizierung von Verkehrszeichen in autonomen Fahrzeugen

Autonome Fahrsysteme sind auf eine präzise Objekterkennung angewiesen, um sicher durch die Straßen zu navigieren. Die Klassifizierung von Verkehrsschildern, die von Edge-KI unterstützt wird, ist dabei ein wichtiger Bestandteil. Eingebettete KI-Modelle verarbeiten Kameradaten, um Verkehrsschilder in Echtzeit zu erkennen, so dass die Fahrzeuge auf sich ändernde Straßenbedingungen und Anordnungen reagieren können, ohne auf cloudbasierte Systeme angewiesen zu sein. Diese Anwendung, die das Signal einer RGB-Kamera verarbeitet, ist ein Beispiel für eine Bildverarbeitung, die lange Zeit nur leistungsstärkeren Prozessoren vorbehalten war, nun aber auf Mikrocontrollern durchgeführt werden kann. Lesen Sie mehr unter [6].

### Gesichtsidentifizierung und Sicherheitssysteme

Bei Sicherheitsanwendungen ermöglicht Edge-KI Gesichtserkennung und -identifizierung in Echtzeit, ohne dass ein ständiger Cloud-Zugriff erforderlich ist. Dies ist besonders wichtig für den Schutz der Privatsphäre und der Daten, da sensible biometrische Informationen lokal verarbeitet und gespeichert werden können. Vom Bankwesen bis hin zu intelligenten Städten verbessert Edge-KI die Sicherheitssysteme und ermöglicht eine schnellere und genauere Erkennung, ohne die Privatsphäre der Nutzer zu beeinträchtigen. Lesen Sie mehr unter [7].

## Die Bedeutung von Kreativität bei der KI-Implementierung

Die Anwendungsfälle der Edge-KI zeigen nicht nur ihr enormes Potenzial (**Bild 2**), sondern auch einen oft übersehenen, aber wesentlichen Aspekt: Kreativität. Bei der Anwendung von KI auf eingebettete Systeme geht es nicht nur um die Übernahme von Modellen oder Algorithmen „von der Stange“; sie erfordert von Entwicklern und Ingenieuren ein neues, kreatives Denken, um Probleme auf eine Weise zu lösen, die mit herkömmlichen Methoden nicht möglich wäre. Die wahre Stärke der Edge-KI liegt nicht nur in ihrer Fähigkeit, Aufgaben zu automatisieren, sondern auch in einer neuen Art und Weise, wie sie Daten sammelt, verarbeitet und nutzt, so dass innovative Lösungen möglich sind.

In der Welt der eingebetteten Systeme stellen Beschränkungen wie begrenzte Rechenleistung, Speicherplatz und Energieeffizienz eine große Herausforderung dar. Dies regt jedoch auch zu kreativen Problemlösungen an. Ingenieure müssen über die herkömmliche Verarbeitung der Sensordaten hinaus denken und untersuchen, wie verschiedene Datenströme kombiniert, abgeleitet oder sogar synthetisiert werden können, um nützliche Ergebnisse zu erzielen. Edge-KI-Modelle müssen deshalb oft so optimiert werden, dass sie in die engen Grenzen von eingebetteten Geräten passen, ohne dass die Genauigkeit oder Effizienz ihrer Vorhersagen darunter leidet.

Nehmen wir das Beispiel der elektrisch unterstützten Fahrräder von Panasonic [4]. Anstatt herkömmliche Drucksensoren zur Messung des Reifendrucks zu installieren,



Bild 2. Edge-KI bietet enormes Potenzial für eine Vielzahl von Anwendungen.

# NanoEdge AI Studio simplifies edge AI development workflow

Deployment of NanoEdge AI Studio libraries, the market reference AutoML tool, is completely free for unlimited quantities on any STM32

... and available on any Arm® Cortex®-M MCU\*

NANOEDGE AI STUDIO 

The best combination for given data: ML model, hyperparameters, and preprocessing

On-device learning capability to fine-tune a deployed solution without retraining

Bring your own data approach: no need to create edge AI models



\*under a special license agreement

7

Bild 3.  
NanoEdge AI Studio bietet eine einfache Schnittstelle für die Entwicklung von ML-Projekten.

die die Komplexität und die Kosten erhöhen würden, wählten die Ingenieure einen kreativeren Ansatz, der KI mit einschloss. Indem sie die Strommessdaten des Motors nutzten, entwickelten sie einen virtuellen Drucksensor - eine Lösung, die den Reifendruck aus der Analyse der Stromaufnahme des Motors ableitet.

Da der Motor auf die Eingaben des Fahrers und die Straßenbedingungen reagiert, können subtile Änderungen der Stromaufnahme zur Vorhersage des Reifendrucks genutzt werden. In diesem Fall reduziert der kreative Einsatz von KI nicht nur die Komplexität der Hardware, sondern verbessert auch die Gesamtleistung und Sicherheit des Fahrrads, und dies auf der Grundlage von Daten, die ohnehin vom System erfasst werden.

Dies ist ein perfektes Beispiel dafür, wie Kreativität beim Einsatz von KI zu innovativen, kostengünstigen Lösungen führen kann. Indem Ingenieure über den Tellerand hinausschauen, können sie das Problem - die Reifendrucküberwachung - so umgestalten, dass vorhandene Daten neue Erkenntnisse liefern. Anstatt sich auf teure oder spezielle Sensoren zu verlassen, nutzten sie vorhandene Motordaten, um eine völlig neue Funktion zu erschließen. Diese Art von Einfallsreichtum macht die KI so leistungsstark und vielseitig!

## Neuartige Problemlösungen in eingebetteten KI-Systemen

Der kreative Einsatz von KI in eingebetteten Systemen erfordert von den Ingenieuren auch eine ganzheitliche Betrachtung der Problemlösung, bei der nicht nur technische Aspekte, sondern auch praktische Auswir-

kungen des Einsatzes von KI im Edge-Bereich berücksichtigt werden. Ein Beispiel:

### Vorhandener Sensoren - neue Aufgaben

Wie im Fall des Panasonic-Fahrrads kann KI oft die Zweckbestimmung vorhandener Sensoren verändern. Ingenieure können Vibrationssensoren, die ursprünglich für die Überwachung des mechanischen Verschleißes entwickelt wurden, auch zur Erkennung anderer kritischer Parameter wie Umweltbedingungen oder unerwarteter Leistungsanomalien verwenden.

### Kombinieren verschiedener Datenquellen

Durch die kreative Integration von Daten verschiedener Sensoren (beispielsweise für Schall, Vibration und Temperatur) kann ein umfassenderes Bild vom Zustand eines Systems erstellt werden. KI-Modelle, die auf diese Datenfusion trainiert wurden, können nuancierte Muster erkennen, die ein einzelner Sensor möglicherweise übersehen würde.

### Ableitung verborgener Erkenntnisse

KI kann Erkenntnisse schlussfolgern, die nicht direkt messbar sind. Anstatt den Zustand einer Straße oder eines Fußbodens direkt zu messen, können KI-Modelle beispielsweise die Vibrationsmuster eines Staubsaugers oder einer Fahrzeugaufhängung nutzen, um die Art der Oberfläche oder den Verschleißgrad zu bestimmen. Diese Fähigkeit, verborgene Erkenntnisse aus indirekten Datenquellen zu gewinnen, ermöglicht es Entwicklern, Probleme auf neue und spannende Weise zu lösen.

## Optimierung der Energieeffizienz

Bei batteriebetriebenen Geräten wie Wearables oder Smart Appliances ist Energie eine kostbare Ressource. Kreativität beim Einsatz von KI bedeutet nicht nur, Modelle zu entwickeln, die genaue Ergebnisse liefern, sondern auch Energie sparen. Dies bedeutet oft, dass clevere Wege gefunden werden müssen, um die Komplexität der KI-Modelle zu reduzieren oder Berechnungen so durchzuführen, dass möglichst wenig Energie verbraucht wird.

In all diesen Fällen müssen die Entwickler über das traditionelle Paradigma der einfachen Automatisierung von Aufgaben hinausgehen und stattdessen KI als ein Werkzeug zur Gewinnung neuer Erkenntnisse und zur effizienteren Lösung komplexer Probleme betrachten. Dies erfordert einen Wandel der Denkweise, bei dem es nicht nur darum geht, KI zu implementieren, sondern zu überdenken, wie KI das Problem und seine Lösung selbst neu gestalten und vereinfachen kann.

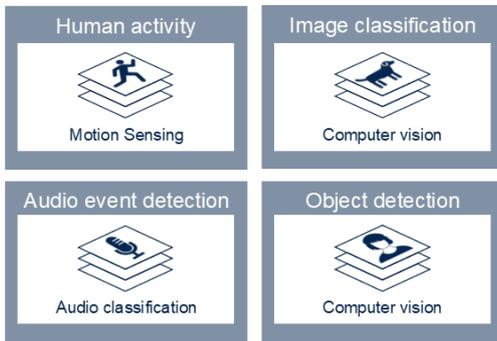
## Wie STMicroelectronics die Zukunft der Edge-KI mitgestaltet

STMicroelectronics steht an der Spitze dieser Edge-KI-Bewegung und bietet umfassende Lösungen, die auf die unterschiedlichen Bedürfnisse von Ingenieuren und Entwicklern in verschiedenen Bereichen zugeschnitten sind. Mit seiner *Edge-AI Suite* [8] bietet ST eine breite Palette von Software-Tools, die auf unterschiedliche Benutzerprofile und Fachkenntnisse zugeschnitten sind, sowie eine große Sammlung von Fallstudien [9], die Entwickler inspirieren sollen.

# Start with edge AI optimized models STM32 model zoo



A collection of application-oriented models optimized for STM32



Hosted on GitHub



Model training scripts

- Scripts to generate and validate



Application code example

- Designed to host optimized NN models
- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation



9

Bild 4. Der STM32 Model Zoo enthält alle Modelle, die Sie für Ihre ersten Schritte in die Edge-KI benötigen.

Für diejenigen, die noch keine Erfahrung mit KI oder eingebetteten Systemen haben, bietet das *NanoEdge AI Studio* (Bild 3) [10] eine intuitive, ganzheitliche Plattform, mit Benutzer ihre eigenen Daten einbringen, Modelle trainieren und auf eingebetteten Geräten einsetzen können. Diese Lösung ist ideal für Entwickler, die schnell KI-gestützte Erkennung von Anomalien, Fehlerklassifizierung oder Umweltüberwachungssysteme implementieren möchten.

*STM32 Cube.AI* [11] und *ST Edge AI Developer Cloud* [12] sind dagegen für fortgeschrittene Anwender gedacht, die mehr Flexibilität und Anpassungsmöglichkeiten benötigen. Sie bieten Zugang zu einer Modell-Sammlung (Bild 4) [13] und fortschrittlichen Tools für die Entwicklung, das Training und den Einsatz von benutzerdefinierten Modellen. Ob es nun um die Klassifizierung von Verkehrsschildern in autonomen Fahrzeugen oder die Gesichtserkennung in Sicherheitssystemen geht, *STM32 Cube.AI*

ermöglicht es Entwicklern, die Grenzen der Edge-KI zu erweitern.

ST stellt für die Entwicklung der Edge-KI skalierbare, anpassungsfähige Lösungen bereit, die die Anforderungen von Ingenieuren und Entwicklern erfüllen, die an der Spitze der Innovation stehen möchten. Durch die Kombination von Hardware-Know-how mit KI-gesteuerter Software trägt ST dazu bei, neue Möglichkeiten in Bereichen zu erschließen, die von erneuerbaren Energien über Smart Homes bis hin zum Transportwesen und der industriellen Automatisierung reichen.

Da Edge-KI immer mehr an Bedeutung gewinnt, werden der kreative Einsatz von Sensoren, die Anwendung verschiedener Algorithmen und die Anpassungsfähigkeit von Lösungen wie denen von STMicroelectronics für die Gestaltung der Zukunft intelligenter, vernetzter Geräte unerlässlich sein. ◀

RG — 240620-02



## Über den Autor

Vianney Taufour ist *Manager of AI Solutions Developer Engagement* bei STMicroelectronics im französischen Grenoble, mit über 25 Jahren Erfahrung in der Halbleiterindustrie. Er verfügt über einen fundierten Hintergrund in den Bereichen Kommunikation, Branding, Marketing und Elektronik und hatte verschiedene Positionen in den Bereichen Forschung und Entwicklung, Marketing und Unternehmenskommunikation inne. Vianney hat einen Master-Abschluss in *Electrical Engineering and Microelectronics* vom Institut *National des Sciences Appliquées de Lyon*.

## WEBLINKS

- [1] Erkennung von Anomalien in der Motorleistung: <https://tinyurl.com/anomaly-detection-case>
- [2] Erkennung und Klassifizierung von Motorfehlern: <https://tinyurl.com/fault-detection-case>
- [3] Erkennung von DC-Störlichtbögen in elektrischen Systemen (AFCl): <https://tinyurl.com/arc-fault-detection>
- [4] Virtuelle Drucksensoren für mehr Sicherheit bei E-Bikes: <https://tinyurl.com/e-bikes-panasonic>
- [5] Fußbodenbelag-Erkennung für intelligente Staubsauger: <https://tinyurl.com/floor-type-detection>
- [6] Klassifizierung von Verkehrszeichen in autonomen Fahrzeugen: <https://tinyurl.com/traffic-sign-classification>
- [7] Gesichtserkennung und Sicherheitssysteme: <https://tinyurl.com/face-identification>
- [8] Edge AI Suite: <https://www.st.com/st-edge-ai-suite>
- [9] Praxisbeispiele ST Edge AI: [https://www.st.com/content/st\\_com/en/st-edge-ai-suite/case-studies.html](https://www.st.com/content/st_com/en/st-edge-ai-suite/case-studies.html)
- [10] NEAI Studio: <https://stm32ai.st.com/nanoedge-ai/>
- [11] STM32 Cube.AI: <https://stm32ai.st.com/stm32-cube-ai/>
- [12] ST Edge AI Developer Cloud: <https://stm32ai.st.com/st-edge-ai-developer-cloud/>
- [13] STM32 AI-Model-Zoo: <https://stm32ai.st.com/model-zoo/>



# Die Macht der Edge-KI entfesseln

Ein Gespräch mit François de Rochebouët von STMicroelectronics

## Fragen der Elektor-Redaktion

Erfahren Sie, wie innovative KI-Lösungen für Mikrocontroller eingebettete Systeme revolutionieren und die Industrie durch energieeffiziente Echtzeit-Entscheidungsfindung verändern. François de Rochebouët (Head of Marketing, AI Solutions, STMicroelectronics) spricht über die entscheidende Rolle von Edge-KI bei der Gestaltung einer nachhaltigen Zukunft.

**Elektor:** François, vielen Dank, dass Sie bei uns sind. Könnten Sie sich zunächst vorstellen und ein wenig über Ihren Weg zu KI und Innovation erzählen?

**François de Rochebouët:** Ich habe mich schon immer für Innovationen und das Überschreiten von Grenzen begeistert, insbesondere bei eingebetteten Systemen und KI. Als Mitbegründer von Cartesiam haben wir bei KI-Lösungen für Mikrocontroller Pionierarbeit geleistet. Jetzt, bei STMicroelectronics, leite ich das KI-Marketing für STM32-Mikrocontroller. Meine Vorstellung ist klar: Edge-KI ist entscheidend für die Zukunft. Wir brauchen KI at the Edge, näher an dem Ort, an dem die Daten generiert werden, denn Cloud-basierte Lösungen sind nicht in der Lage, die riesigen Datenmengen zu verarbeiten, die unsere zunehmend vernetzte Welt erzeugt. Nachhaltige KI muss Daten lokal verarbeiten, um Energie zu sparen und die Reaktionsfähigkeit zu verbessern.

**Elektor:** Warum ist Edge-KI für die Gestaltung einer nachhaltigen Zukunft so wichtig?

**Why edge AI matters**

**Edge = where the SIGNAL becomes DATA**

- Ultra-low latency**  
Real-time applications
- 01 Reduced data transmission**  
10 Generate meaningful information
- Enhanced privacy and security**  
No data sharing in the cloud
- Sustainable on energy**  
Low data, low power
- Lower cost of inference to enable a new class of operations**

STMicroelectronics logo: **ST** life.dugmented

STM32 chip on a finger

Bild 1. Die Vorteile von Edge-KI.

**François:** Die Datenmenge, die wir schon heute generieren, ist überwältigend und sie wird exponentiell größer. Sich auf Cloud-KI zu verlassen, bedeutet einen hohen Energiebedarf, Bandbreiten- und Latenzprobleme. Edge-KI (**Bild 1**) verarbeitet Daten lokal an der Quelle, wodurch die Notwendigkeit einer ständigen Datenübertragung in die Cloud reduziert wird. Dies senkt nicht nur den Energiebedarf, sondern führt auch zu einer schnelleren und effizienteren Entscheidungsfindung. Nachhaltigkeit benötigt KI-Lösungen, die in Umgebungen mit geringem Stromangebot betrieben werden können, und genau dadurch zeichnen sich eingebettete Systeme und Mikrocontroller aus.

**Elektor:** Welches sind die größten Herausforderungen für Entwickler bei der Entwicklung von Edge-KI-Lösungen?

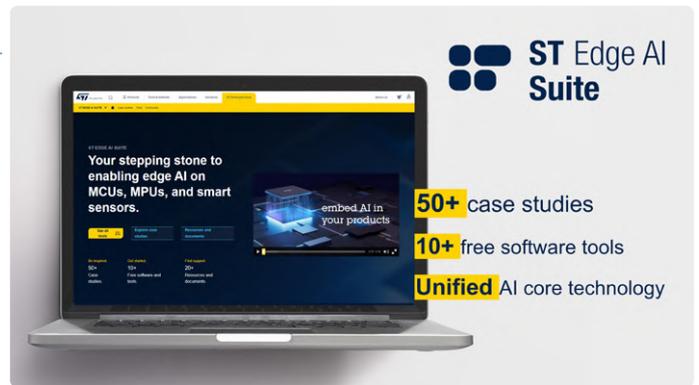
**François:** Eine der größten Herausforderungen besteht darin, die Art und Weise zu ändern, wie Entwickler über Problemlösungen denken. Bei herkömmlichen Systemen verlassen sie sich auf Sensoren, die direkte Messwerte liefern, aber mit KI ändert sich der Ansatz. KI ermöglicht es uns, komplexe Erkenntnisse aus scheinbar unzusammenhängenden Daten abzuleiten. Dies erfordert eine neue Denkweise. Die Entwickler sehen sich auch mit Einschränkungen wie begrenzter Rechenleistung, Speicherknappheit und Energieeffizienz konfrontiert. Daher müssen sie KI-Modelle so gestalten, dass sie in diese eng gesteckten Parameter passen und gleichzeitig Genauigkeit und Zuverlässigkeit gewährleisten.

**Elektor:** Können Sie Beispiele dafür nennen, wie eingebettete KI dabei geholfen hat, Probleme von Kunden kreativ zu lösen?

**François:** Ein Beispiel sind die Fahrräder mit Elektrounterstützung von Panasonic. Anstatt herkömmliche Drucksensoren für die Reifenüberwachung zu verwenden, wurden die vorhandenen Strommessdaten des Motors genutzt. Durch den Einsatz von KI wurde ein virtueller Drucksensor geschaffen, der den Reifendruck aus diesen Daten ableitet. Dadurch wurden die Komplexität und die Kosten der Hardware reduziert und gleichzeitig die Leistung verbessert. Ein weiteres Beispiel ist die Erkennung von Anomalien bei Motoren, die mit unterschiedlichen Geschwindigkeiten laufen. Durch den Einsatz von KI konnten wir unseren Kunden helfen, Ausfälle frühzeitig zu erkennen, Ausfallzeiten zu minimieren und die Wartungspläne zu optimieren.

**Elektor:** Wie geht STMicroelectronics mit der Herausforderung um, Edge-KI-Lösungen mit begrenzten Ressourcen zu entwickeln?

**François:** Wir glauben, dass die Kombination aus leistungsstarker Hardware und dem richtigen Software-Ökosystem den Unterschied ausmacht. Auf der Hardwareseite haben wir die besten STM32-Mikrocontroller [1] ihrer Klasse, die die notwendige Energieeffizienz und Leistungsfähigkeit für Edge-KI bieten. Auf der Softwareseite haben wir unsere Tools und Ressourcen in der *ST Edge AI Suite* [2] (**Bild 2**) zusammengefasst, die den Weg des Entwicklers vom Konzept bis zur Implementierung vereinfacht und eine große Sammlung von Praxisbeispielen zur Inspiration bietet. Darüber hinaus bietet die *STM32 Developer Zone* [3] umfangreiche Unterstützung, Tutorials und Ressourcen, die Entwicklern helfen, innovative Lösungen



*Bild 2. Die ST Edge AI Suite bietet eine Reihe von Tools zur Integration von KI-Funktionen in eingebettete Systeme.*

mit STM32 zu erstellen. Durch diese Kombination aus Hard- und Software können Entwickler Herausforderungen meistern und leistungsstarke KI-Anwendungen erstellen, unabhängig von ihrem Wissensstand.

**Elektor:** Wie unterstützt STMicroelectronics Entwickler mit unterschiedlicher Qualifikation bei Edge-KI?

**François:** Wir bei ST wissen, dass die Entwickler aus verschiedenen Bereichen kommen, daher haben wir unsere Tools auf ihre Bedürfnisse zugeschnitten. Für KI-Neulinge bieten wir *NEAI Studio* [4] an, eine intuitive Plattform, mit der KI-Modelle mit minimalem Aufwand erstellt und eingesetzt werden können. *NEAI Studio* unterstützt einen BYOD-Ansatz (Bring Your Own Data) mit einer AutoML-Lösung, die direkt optimierte Bibliotheken mit den am besten geeigneten Algorithmen für maschinelles Lernen auf der Grundlage der Benutzerdaten erzeugt.

Für fortgeschrittene Anwender bieten *STM32Cube.AI* [5], *ST Edge AI Core* [6] und die *ST Edge AI Developer Cloud* [7] erweiterte Optimierungs- und Benchmarking-Lösungen. Entwickler können auch von Skripten profitieren, um Modelle aus dem Modell-Zoo [8] oder ihre eigenen Modelle neu zu trainieren, was dem BYOM-Ansatz (Bring Your Own Model) entspricht. Diese umfassende Unterstützung ermöglicht es Entwicklern, ihre vortrainierten Modelle zur weiteren Optimierung und Bereitstellung in das Ökosystem einzubringen. Egal, ob Sie ein Maker, ein erfahrener Embedded-Ingenieur oder ein Datenwissenschaftler sind, wir haben Tools, die zu Ihrem Fachwissen passen und es Ihnen ermöglichen, Ihre Produktivität und Innovation zu maximieren.

**Elektor:** Welche Rolle spielt die Hardware, um komplexere KI-Anwendungen wie Computer Vision zu ermöglichen?

**François:** Die Hardware spielt eine große Rolle, vor allem weil die KI-Anwendungen immer komplexer werden. Computer Vision zum Beispiel erfordert viel Rechenleistung und Speicher. Hier kommt spezielle Hardware ins Spiel, zum Beispiel unser *ST Neural-ART-Beschleuniger*, eine NPU (Neural Processing Unit), die für die STM32-MCUs entwickelt wurde. Er ist bahnbrechend in Bezug auf Rechenleistung und Energieeffizienz und ermöglicht es Entwicklern, anspruchsvollere KI-Algorithmen wie Objekterkennung in Echtzeit zu implementieren und gleichzeitig die Stromaufnahme niedrig zu halten.

**Elektor:** Können Sie uns mehr über den *ST Neural-ART-Beschleuniger* erzählen?

# From DMIPS to TOPS, the paradigm shift Opening a new range of embedded AI applications

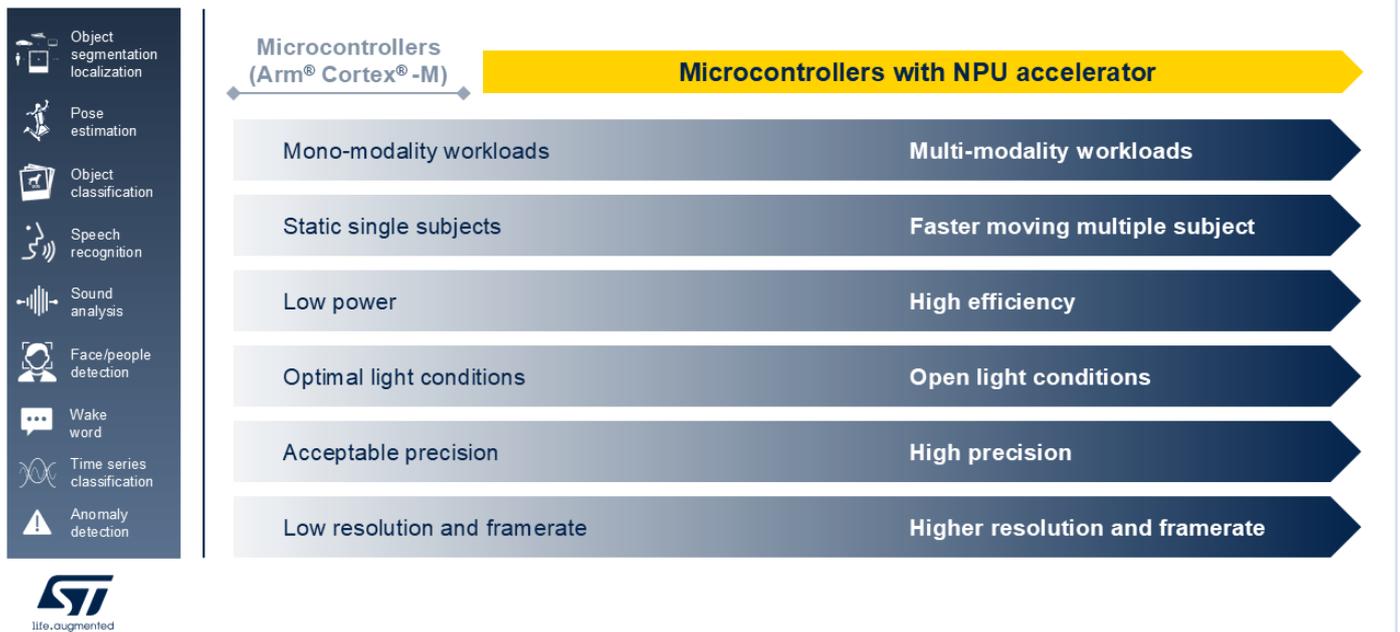


Bild 3. Der NPU-Beschleuniger bietet mehr Rechenleistung für KI-Aufgaben bei gleichzeitig niedriger Stromaufnahme.

**François:** Der ST Neural-ART-Beschleuniger ist ein Meilenstein für Edge-KI-Anwendungen. Es handelt sich dabei um einen Hardware-Beschleuniger (NPU), der speziell für neuronale Netze entwickelt wurde und im Vergleich zu unserem STM32H7-Mikrocontroller eine etwa 600-fach höhere Leistung bietet, während die Energieaufnahme auf ein Minimum reduziert wird (**Bild 3**). Diese Balance ist für KI-Anwendungen in batteriebetriebenen Geräten wie Wearables, intelligenten Kameras oder Industriesensoren entscheidend. Die NPU ermöglicht es Entwicklern, komplexere KI-Modelle, wie sie für Computer Vision oder Echtzeit-Entscheidungsfindung benötigt werden, direkt auf stromsparenden Mikrocontrollern einzusetzen. Wir haben gerade ein Whitepaper herausgegeben, um dies näher zu erläutern. Wir laden alle Interessierten ein, das Whitepaper zu lesen, um mehr darüber zu erfahren.

**Elektor:** Wie sieht STMicroelectronics die Zukunft der Edge-KI?

**François:** Die Zukunft der Edge-KI ist unglaublich spannend. Wir werden mehr intelligente Geräte sehen, die Entscheidungen in Echtzeit treffen und weniger auf die Cloud-Infrastruktur

angewiesen sind. Dies wird der Schlüssel zum Aufbau nachhaltiger, reaktionsschneller Systeme sein, insbesondere in Bereichen wie Smart Cities, Gesundheitswesen und Industrieautomatisierung. Bei ST konzentrieren wir uns auf die Bereitstellung von Tools, Hardware und Ökosystemen, mit denen Entwickler innovativ sein und die Grenzen dessen erweitern können, was mit KI im Edge-Bereich möglich ist.

**Elektor:** Was können wir von STMicroelectronics auf dem kommenden STM32 Summit erwarten?

**François:** Der STM32 Summit [9] am 10. Dezember 2024 wird eine fantastische Gelegenheit für Entwickler und Ingenieure sein, mehr über unsere Vision für Edge-KI zu erfahren. Wir werden einige aufregende neue Entwicklungen vorstellen, darunter Einblicke in unsere künftige KI-Roadmap und die weitere Unterstützung der Community mit Tools und Ressourcen. Ich lade alle ein, an der Übertragung teilzunehmen, um mehr über unsere Edge-KI-Strategie zu erfahren und aus erster Hand zu erfahren wie ST Innovationen in diesem Bereich vorantreibt. ◀

RG — 240621-02

## WEBLINKS

- [1] STM32-Mikrocontroller: <https://www.st.com/stm32>
- [2] ST Edge AI Suite: <https://www.st.com/st-edge-ai-suite>
- [3] STM32 Developer Zone: [https://www.st.com/content/st\\_com/en/stm32-mcu-developer-zone.html](https://www.st.com/content/st_com/en/stm32-mcu-developer-zone.html)
- [4] NEAI Studio: <https://stm32ai.st.com/nanoedge-ai/>
- [5] STM32 Cube.AI: <https://stm32ai.st.com/stm32-cube-ai/>
- [6] ST Edge AI Core: <https://www.st.com/en/development-tools/stedgeai-core.html>
- [7] ST Edge AI Developer Cloud: <https://stm32ai.st.com/st-edge-ai-developer-cloud/>
- [8] STM32 AI-Modell-Zoo: <https://stm32ai.st.com/model-zoo/>
- [9] STM32 Summit: <https://www.st.com/stm32-summit>

# Eine VHDL-Uhr, entwickelt mit ChatGPT

Von Jean-François Simon (Elektor)

Bis vor kurzem hatte ich noch nie mit einem FPGA gearbeitet und kannte weder die Sprache Verilog noch VHDL. Lesen Sie weiter, um herauszufinden, ob und wie mir ChatGPT geholfen hat, mein erstes FPGA-Projekt von Grund auf zu entwickeln.

Vor einiger Zeit kaufte ich ein Board DE10-Lite von Terasic (**Bild 1**), das ein FPGA aus Intels MAX10-Reihe, ein paar 7-Segment-Anzeigen, Schiebeschalter, Drucktasten mit Hardware-Entprellung, LEDs und mehr enthält. Insgesamt eine vielversprechende Platine, ideal für Einsteiger. Der aktuelle Artikel von Theo Mulder in Elektor [1] brachte mich erneut auf die Idee, mich mit FPGAs zu beschäftigen, und diese KI-Sonderausgabe bot die perfekte Gelegenheit, endlich loszulegen! Ich entschied mich für die Sprache VHDL; ein Textkasten in [1] hilft Ihnen bei der Wahl zwischen VHDL und Verilog.

Unter den Projekten, die Elektroniker, vom Einsteiger bis zum Profi, über Jahrzehnte hinweg begeistern, stehen Uhren ganz oben auf der Liste. Es hat etwas Beruhigendes, das gleichmäßige Ticken der Sekunden und das Blinken der roten LEDs zu beobachten. Das DE10-Lite-Board mit seinen kompakten Maßen von etwa 8 × 10 cm bietet eine 6-stellige 7-Segment-Anzeige und eine elegante Acrylabdeckung – ideal für mein Vorhaben. Die Entscheidung fiel daher auf eine LED-Uhr! Natürlich ist es ein ziemlicher Overkill, einen 484-Pin-BGA-Chip mit 50.000 Logikelementen zu verwenden, um das Äquivalent einer 5-Euro-Uhr zu bauen, doch darum geht es hier nicht, sondern ums Lernen und den Spaß am Experimentieren! Wenn Sie weitere amüsante Beispiele für übertriebene Ressourcennutzung kennen, teilen Sie mir diese gerne mit. Meine Kontaktdaten finden Sie unter **Fragen oder Kommentare**.

## Ein Wort der Warnung

Dieser Artikel enthält einige Details, ist aber kein vollständiges Tutorial. Ein solches wäre viel zu umfangreich und wohl auch zu trocken. Zudem kann ich nicht garantieren, dass Sie, selbst wenn Sie *exakt* dieselben



Bild 1. Das Entwicklungsboard DE10-Lite.

„Prompts“ wie ich in ChatGPT eingeben, identische Antworten erhalten. Das hängt von verschiedenen Faktoren ab, wie der Nutzung der kostenlosen oder kostenpflichtigen Version, dem ChatGPT Modell (4, 4 Turbo, 4 mini) und so weiter. Betrachten Sie diesen Artikel eher als Erfahrungsbericht. Ich empfehle Ihnen dringend, eine KI wie ChatGPT für Ihre eigenen Projekte auszuprobieren, und es gibt keinen Grund, sich auf FPGAs, das DE10-Lite Board oder die Verwendung von Quartus und VHDL zu beschränken. Lassen Sie sich von Ihrer Inspiration leiten und bleiben Sie hartnäckig! Solche Projekte machen nur Fortschritte, wenn bei jedem Schritt kleinere oder größere Hürden überwinden.

## Erste Schritte

Es heißt oft, dass der erste Schritt zum Einstieg in Hardware- und Softwareprojekte das Blinken einer LED ist. Doch bei einem FPGA gibt es dafür keine vorgefertigte Funktion wie `delay()` bei Arduino. Der Einstieg ist hier noch einfacher: Eine der LEDs auf dem Board soll leuchten, wenn eine der Tasten gedrückt wird. Dafür habe ich ChatGPT wie folgt gefragt:

*Schreibe mir einen sehr einfachen VHDL-Code, um eine LED zum Leuchten zu bringen, wenn eine Taste auf einem FPGA-Board DE10-lite gedrückt wird.*

Die Antwort von ChatGPT ist in **Listing 1** zu sehen: Es ist noch ein langer Weg, bis dieser Codes im FPGA eingesetzt werden kann. Hier ist ein Überblick über die notwendigen Schritte: Laden Sie zunächst Quartus Prime Lite herunter und installieren Sie es [2]. Erstellen Sie ein neues, leeres Projekt mit dem *New Project Wizard*. Geben Sie dann dem Projekt einen Namen (ich habe es „test1“ genannt) und wählen Sie aus der Dropdown-Liste *Devices* das FPGA aus, das verwendet werden soll: 10M50DAF484C7G. Als Nächstes erstellen Sie eine neue VHDL-Datei (*File* → *New*) und kopieren den Code in diese Datei. Als Nächstes müssen wir die Eingangs- und Ausgangspins festlegen, die wir verwenden möchten. Ich empfehle, die System-CD von Terasic mit Dokumentation und Beispieldateien herunterzuladen [3]. Das DE10-Lite-Benutzerhandbuch ist ebenfalls eine große Hilfe, da es den Zusammenhang zwischen den Pin-Nummern und den Komponenten auf der Platine beschreibt. In diesem Beispiel verwende ich den Taster



## Listing 1: ChatGPT bringt mit diesem einfachen Code eine LED zum Leuchten

```

library IEEE;
use IEEE.STD_LOGIC_1164.ALL;

entity LED_Button is
  Port (
    KEY : in std_logic; -- Input button
    LED : out std_logic -- Output LED
  );
end LED_Button;

architecture Behavioral of LED_Button is
begin
  -- Process to control LED
  process(KEY)
  begin
    if (KEY = '0') then -- Active low button (pressed)
      LED <= '1'; -- Turn on LED
    else
      LED <= '0'; -- Turn off LED
    end if;
  end process;
end Behavioral;

```

KEY0, der mit dem Pin *PIN\_B8* des FPGAs verbunden ist, sowie die LED *LEDR0* an *PIN\_A8*. Dies kann man am einfachsten im *Pin Planner* im Quartus *Assignments*-Menü einstellen. Bevor wir jedoch damit beginnen, müssen wir dem System mitteilen, welche Signale im Projekt vorhanden sind. Eine vollständige Kompilierung (a. in **Bild 2**) wäre ausreichend, ist jedoch recht zeitaufwändig zu erstellen. Daher klicken wir stattdessen auf *Start Analysis & Elaboration* (b.), was deutlich schneller ist und denselben Zweck erfüllt.

Der erste Fehler lässt nicht lange auf sich warten: *Error (12007): Top-level design entity "test1" is undefined*. Standardmäßig gibt Quartus der Top-Level-Entität denselben Namen wie dem Projektnamen. Der von ChatGPT gelieferte Code beschreibt jedoch eine Entität namens *LED\_Button*. Im Menü *Assignments* → *Settings* → *General* können Sie den



Namen der Top-Level-Entität umbenennen; geben Sie *LED\_Button* statt *test1* ein. Jetzt kann *Analysis & Elaboration* korrekt durchgeführt werden. Verwenden Sie anschließend den *Pin Planner*, um die *PIN\_B8* für *KEY* und *PIN\_A8* für *LED* zuzuweisen. Führen Sie schließlich eine vollständige Kompilierung durch (klicken Sie dazu auf das blaue Dreieck oder drücken Sie CTRL-L).

### Programmierung des FPGA

Schließlich wird mit dem *Programmer* im Menü *Tools* das frisch kompilierte Design in das FPGA geladen. Wir können sehen, dass alle LEDs auf der Platine mit Ausnahme von *LEDR0* leuchten. Diese wird erst eingeschaltet, wenn man die Taste *KEY0* drückt. Sie! Wer dies selbst einmal ausprobieren möchte, empfehle ich das Video [4], das nützliche Informationen für die ersten Schritte liefert. Der Autor des Videos hat ein FPGA Cyclone IV anstelle eines MAX10 verwendet, also vergessen Sie nicht, seine Anweisungen an Ihren speziellen Fall anzupassen.

Die neuen LEDs (*LEDR1* bis *LEDR9*) leuchten, weil in Quartus der folgende Parameter standardmäßig aktiviert ist: *Reserve all unused pins as inputs tri-stated with weak pull-up*.

Dadurch fließt ein kleiner Strom und lässt die LEDs leuchten. Dieser Parameter kann unter *Assignments Device* → *Device and Pin Options* → *Unused Pins* geändert werden. Sie können ihn beispielsweise auf *As output driving ground* setzen, um einen Low-Status zu erzwingen und die LEDs auszuschalten. Nach der Neukompilierung und Neuprogrammierung sind die LEDs tatsächlich ausgeschaltet, jedoch leuchten jetzt einige Segmente der 7-Segment-Anzeige. Diese Segmente sind, wie in der Dokumentation beschrieben, active low. Eine andere Möglichkeit wäre, die LEDs oder Segmente im Code gezielt auf 1 oder 0 zu setzen. Darauf werden wir später näher eingehen.

### Von einzelnen LEDs zu Zahlen

Sehen wir uns nun an, wie man Zahlen in der 7-Segment-Anzeige darstellen kann. Ich habe den folgenden Prompt eingegeben:

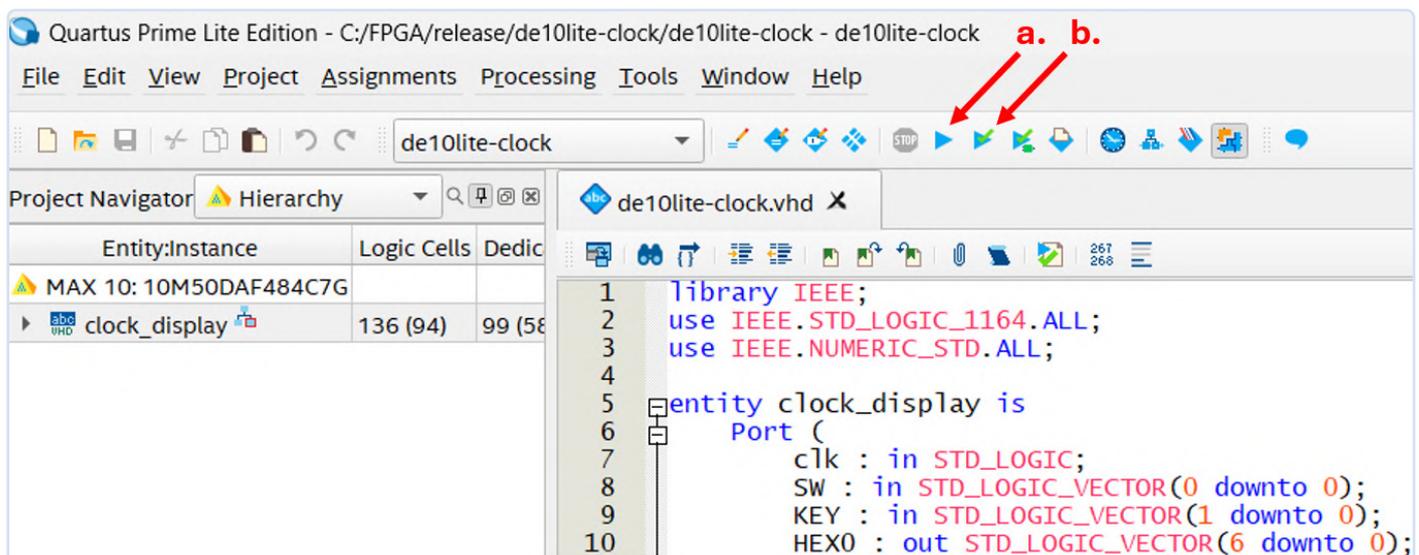


Bild 2. Analyse, Ausarbeitung und Kompilierung in Quartus.



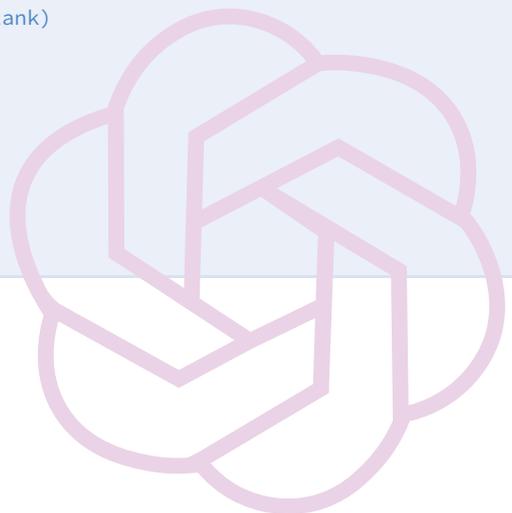
## Listing 2: Ein Treiber für 7-Segment-LEDs

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;
use IEEE.STD_LOGIC_ARITH.ALL;
use IEEE.STD_LOGIC_UNSIGNED.ALL;

entity seven_segment_display is
  Port (
    clk : in STD_LOGIC; -- Clock signal
    digit : in STD_LOGIC_VECTOR(3 downto 0); -- 4-bit input for the digit (0-9)
    segments : out STD_LOGIC_VECTOR(6 downto 0) -- Outputs for the 7-segment display
  );
end seven_segment_display;

architecture RTL of seven_segment_display is
  signal seg_reg : STD_LOGIC_VECTOR(6 downto 0); -- Register for storing segment data
begin
  -- Sequential logic to update the display register on every clock edge
  process(clk)
  begin
    if rising_edge(clk) then
      case digit is
        when "0000" => -- Digit 0
          seg_reg <= "0000001"; -- a,b,c,d,e,f are on, g is off
        when "0001" => -- Digit 1
          seg_reg <= "1001111"; -- b,c are on
        when "0010" => -- Digit 2
          seg_reg <= "0010010"; -- a,b,d,e,g are on
        when "0011" => -- Digit 3
          seg_reg <= "0000110"; -- a,b,c,d,g are on
        when "0100" => -- Digit 4
          seg_reg <= "1001100"; -- b,c,f,g are on
        when "0101" => -- Digit 5
          seg_reg <= "0100100"; -- a,c,d,f,g are on
        when "0110" => -- Digit 6
          seg_reg <= "0100000"; -- a,c,d,e,f,g are on
        when "0111" => -- Digit 7
          seg_reg <= "0001111"; -- a,b,c are on
        when "1000" => -- Digit 8
          seg_reg <= "0000000"; -- All segments are on
        when "1001" => -- Digit 9
          seg_reg <= "0000100"; -- a,b,c,d,f,g are on
        when others =>
          seg_reg <= "1111111"; -- All segments off (blank)
      end case;
    end if;
  end process;

  -- Combinational logic to drive the segments
  segments <= seg_reg;
end RTL;
```





### Listing 3: Top-Level-VHDL-Code zur Anzeige einer Ziffer auf dem 7-Segment-Display

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;
use IEEE.STD_LOGIC_ARITH.ALL;
use IEEE.STD_LOGIC_UNSIGNED.ALL;

entity top_level is
  Port (
    clk : in STD_LOGIC; -- Clock input
    segments : out STD_LOGIC_VECTOR(6 downto 0) -- Outputs to the 7-segment display
  );
end top_level;

architecture Behavioral of top_level is
  signal digit_signal : STD_LOGIC_VECTOR(3 downto 0); -- 4-bit digit signal
begin
  -- Instantiate the seven_segment_display module
  U1 : entity work.seven_segment_display
    Port map (
      clk => clk,
      digit => digit_signal,
      segments => segments
    );

  -- Set the digit to "4"
  digit_signal <= "0100"; -- Binary for 4
end Behavioral;
```

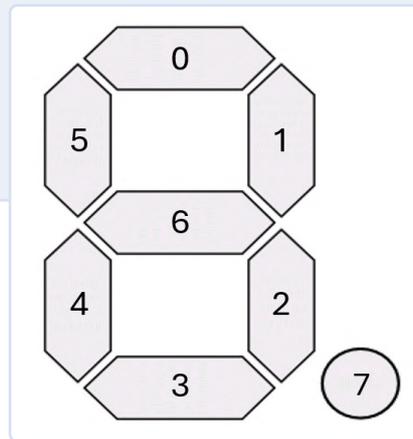


Bild 3. Adressierung der einzelnen Display-Segmente.

Schreibe mir einen VHDL-Code, der eine beliebige Ziffer von 0 bis 9 auf die eingebaute 7-Segment-Anzeige eines Entwicklungsboards DE10-lite schreiben kann.

Die Antwort von ChatGPT enthält einige Codes und Erklärungen, die ich hier weggelassen habe; der eigentliche Code ist in **Listing 2** zu sehen. Um die Wiederverwendbarkeit dieses Codes zu erleichtern, habe ich ihn als Beispiel in eine neue VHDL-Datei namens *7segment.vhd* eingefügt. Denken Sie daran, im Menü *Project* auf *Add Current File to Project* zu klicken. Der mitgelieferte Code ist allgemein gehalten und enthält keine spezifischen Anweisungen zur Anzeige einer bestimmten Ziffer. Daher habe ich meine Anfrage weiter präzisiert:

*Ich habe den zuletzt genannten Code in eine Datei 7segment.vhd eingefügt. Zeige mir, welchen Code ich in der Top-Level-Entität verwenden muss, um die Ziffer „4“ anzuzeigen.*

Die Antwort ist in **Listing 3** zu sehen. Lassen Sie uns aufschlüsseln, wie das funktioniert. Die Entität der obersten Ebene, *top\_level*, hat einen *clk*-Eingang für das Taktsignal, das an die Komponente *seven\_segment\_display* weitergegeben wird. Das Signal *digit\_signal* wird verwendet, um den 4-Bit-Binärwert für die Ziffer zu speichern. Für die Zahl 4 lautet die binäre Entsprechung 0100. Die Entität *seven\_segment\_display* wird instanziiert und auf die Top-Level-Signale (*clk*, *digit\_signal* und *segments*) abgebildet. Abgebildet (mapping) bedeutet im Wesentlichen, dass Signale miteinander verbunden werden. Zum Beispiel verbindet *digit => digit\_signal* das Top-Level-Signal *digit\_signal* mit dem *digit*-Eingangsport der *seven\_segment\_display*-Entität und

ermöglicht es dem Ersteren, letztere zu steuern. Auch die Signale *clk* und *segments* werden auf die gleiche Weise abgebildet. Schließlich wird die anzuzeigende Zahl mit der Anweisung *digit\_signal <= "0100"* festgelegt. Beachten Sie den Unterschied zwischen dem Zuweisungsoperator *<=* und dem Assoziationsoperator *=>*, der insbesondere für die Portzuordnung verwendet wird.

Bevor das Ergebnis sichtbar wird, müssen Pin-Constraints hinzugefügt werden, um Quartus zu ermuntern, die entsprechenden Pins zu verwenden, die mit den Anzeigesegmenten verbunden sind. Die Segmente sind normalerweise mit einem Buchstaben von „a“ bis „g“ gekennzeichnet, aber Terasic verwendet stattdessen eine Nummerierung von 0 bis 6 (**Bild 3**). Das Benutzerhandbuch liefert uns die Zuordnung; ein Auszug ist in **Tabelle 1** dargestellt. Für dieses Beispiel mit einer einstelligen Anzeige werden sieben Pins benötigt, sodass der *Pin Planner* hierfür ausreichend wäre. Für die endgültige Uhr benötigen wir jedoch alle sechs Ziffern und die Dezimalpunkte, was insgesamt 48 Anschlüsse erfordert.

Es ist daher praktischer, anders vorzugehen. Jedem Pin-Constraint entsprechen zwei Zeilen der Quartus-Konfigurationsdatei (QSF-Datei): eine Zeile *set\_instance\_assignment*, die Quartus den I/O-Typ mitteilt, und eine Zeile *set\_location\_assignment*, die die Pin-Belegung

**Tabelle 1: Pinbelegung für das Display**

Signalname	FPGA Pin-	Beschreibung	I/O-Standard
HEX00	PIN_C14	Seven Segment Digit0[0]	3.3-V LVTTTL
HEX01	PIN_E15	Seven Segment Digit0[1]	3.3-V LVTTTL
HEX02	PIN_C15	Seven Segment Digit0[2]	3.3-V LVTTTL
HEX03	PIN_C16	Seven Segment Digit0[3]	3.3-V LVTTTL
HEX04	PIN_E16	Seven Segment Digit0[4]	3.3-V LVTTTL
HEX05	PIN_D17	Seven Segment Digit0[5]	3.3-V LVTTTL
HEX06	PIN_C17	Seven Segment Digit0[6]	3.3-V LVTTTL
HEX07	PIN_D15	Seven Segment Digit0[7], DP	3.3-V LVTTTL
HEX10	PIN_C18	Seven Segment Digit1[0]	3.3-V LVTTTL
HEX11	PIN_D18	Seven Segment Digit1[1]	3.3-V LVTTTL
HEX12	PIN_E18	Seven Segment Digit1[2]	3.3-V LVTTTL
etc.	etc.	etc.	etc.

selbst angibt. Es ist deutlich einfacher, diese Pin-Constraints aus einer bestehenden QSF-Datei zu kopieren. Öffnen Sie die Datei *DE10\_LITE\_Default.qsf* im Ordner *Demonstrations/Default* auf der System-CD und kopieren Sie die 96 Zeilen für die 7-Segment-Anzeige:

```
set_instance_assignment -name IO_STANDARD "3.3-V LVTTTL"
-to HEX0[0]
(...)
(...)
set_location_assignment PIN_L19 -to HEX5[7]
```

Fügen Sie diese Zeilen in die QSF-Datei Ihres aktuellen Projekts ein. Der Code in Listing 2 und Listing 3 benötigt ein Taktsignal. Kopieren Sie also den sechszeiligen Block, der für die interne Uhr des FPGAs zuständig ist, gekennzeichnet durch den Header `# CLOCK`. Damit dies funktioniert, müssen Sie `clk` in *MAX10\_CLK1\_50* und `segments` in *HEX0* in der VHDL-Hauptdatei (der Top-Level-Entität) umbenennen:

```
entity top_level is
Port (
MAX10_CLK1_50 : in STD_LOGIC;
HEX0 : out STD_LOGIC_VECTOR(6 downto 0)
);
end top_level;
```

Wir müssen die gleichen Signale auch bei der Instanziierung unserer 7-Segment-Anzeige umbenennen, um die Signale korrekt von einer Entität auf eine andere abzubilden:

```
U1 : entity work.seven_segment_display
Port map (
clk => MAX10_CLK1_50,
digit => digit_signal,
segments => HEX0
);
```

Jetzt ist es Zeit für eine neue Kompilierung, die im Idealfall ohne Fehler durchgeführt werden sollte. Falls jedoch Fehler auftreten, melden Sie diese an ChatGPT, damit Ihnen bei der Fehlerbehebung geholfen werden kann. Nach der Neuprogrammierung erhalten Sie die Anzeige aus **Bild 4**. Nun... das sieht nicht direkt wie eine „4“ aus. Werfen wir also einen Blick auf den Code, der für die Erzeugung der Signale für

die Segmente verantwortlich ist:

```
case digit is
when "0000" => -- Digit 0
seg_reg <= "0000001"; -- a,b,c,d,e,f are on, g is off
```

### Fehlerbehebung

Die Segmente sind in der Tat active low. Der Kommentar (nach dem doppelten Bindestrich) deutet darauf hin, dass laut ChatGPT die Bits in der Reihenfolge „abcdefg“ angeordnet sind. Nach dem DE10-Lite-Benutzerhandbuch entspricht jedoch das Segment „0“ (oder „a“) dem HEX0[0]-Signal. In VHDL werden die Bits in einer Bitkette von links nach rechts nummeriert, wobei das ganz linke Bit den höchsten Index und das ganz rechte Bit den niedrigsten Index aufweist. Das bedeutet, dass HEX0[0] oder „a“ ganz rechts stehen sollte. Man muss also entweder ChatGPT darauf hinweisen, dass das Problem korrigiert werden muss,

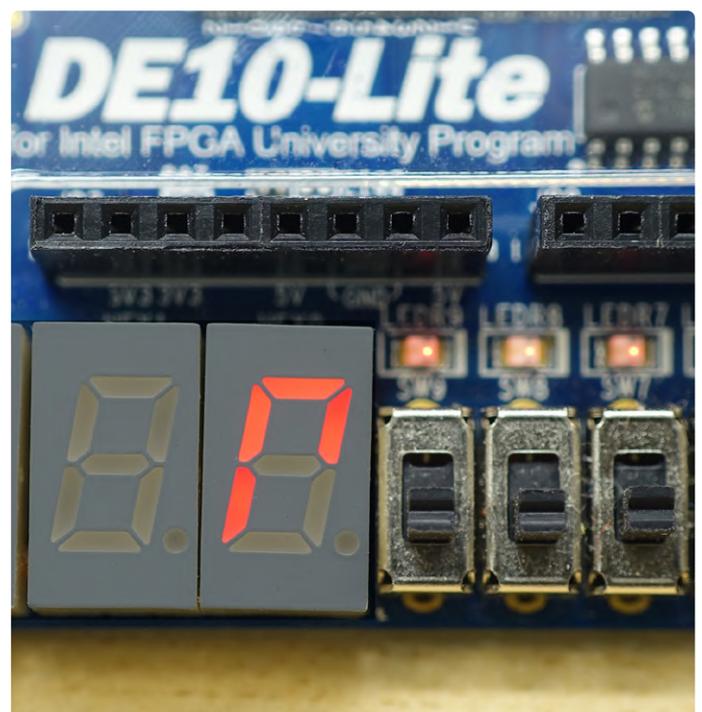


Bild 4. Versuch, eine Zahl anzuzeigen...

## Lösung einiger Probleme

ChatGPT ist auf die folgenden Probleme gestoßen:

- › Generierung von Code mit Funktionen, die nicht korrekt importiert wurden, was zu Fehlern wie „object 'to\_unsigned' is used but not declared“ führte.
- › Deklaration von Prozessen wie `process(clk_1hz, reset)` und anschließende Verwendung von Anweisungen wie `rising_edge(KEY(0))`, was zu Fehlern führte, da `KEY` nicht in der Sensitivitätsliste (dem ersten Satz von Klammern in der Deklaration) enthalten ist.
- › Erzeugen von Latches. Ich empfehle, sich darüber zu informieren, warum Latches in FPGA-Projekten problematisch sind; kurz gesagt, sie können einen Zustand ohne Takt-Eingangssignal halten, was sie unvorhersehbar macht. Quartus verweigert die Kompilierung von Code, der Latches erzeugt, und gibt entsprechende Fehlermeldungen aus.
- › Versuch, Signale aus zwei verschiedenen Prozessen zu ändern, was zu Fehlermeldungen führt.
- › Probleme bei der Verwaltung von zwei getrennten Taktsignalen – wobei der Haupt-Takt Tastendrucke für die Responsivität erkennt, während der 1-Hz-Takt nur die Zeitzähler erhöht. ChatGPT verwechselte manchmal diese beiden Taktgeber, was zu fehlerhaften Anzeigen oder Takten mit sehr langsam reagierenden Tasten führte.
- › Erstellen von Prozessen, die auf zwei Taktsignale gleichzeitig reagieren. Quartus hat Schwierigkeiten zu bestimmen, welche Taktflanke Ereignisse auslösen soll.

- › Der Vorschlag: „Um das Problem zu lösen, verwenden wir ein Flag“, doch der bereitgestellte Code enthielt keine Flags, oder es wurden falsche Erklärungen für bestimmtes Verhalten gegeben.
- › Logikprobleme, wie das Zurücksetzen der Uhr auf 00:00 nach 19:59:59 statt 23:59:59.
- › Schwierigkeiten bei der Behebung von Logikfehlern, wenn diese in mehreren Teilen des Codes auftreten. Es besteht die Tendenz, sich auf die Behebung eines einzigen Problems zu konzentrieren, was dazu führt, dass andere Fehler bestehen bleiben und zu wiederholten Debugging-Zyklen führt.

Das Lösen all dieser Probleme war zwar eine Menge Arbeit, aber es hat Spaß gemacht und war sehr lehrreich!



Bild 5. Die fertige Uhr.

*Minuten und KEY[1] die Stunden erhöht. Die Tasten sind bereits in der Hardware entprellt.*

oder manuell alle Bits von links nach rechts invertieren: 0000001 wird zu 1000000, 1001111 zu 1111001 und so weiter. Nachdem diese Anpassungen vorgenommen wurden, erscheint nach erneuter Kompilierung und Neuprogrammierung die Zahl 4 korrekt auf dem Display. Wie Sie sehen können, war das ein ziemlicher Kampf. Aber lassen Sie sich nicht entmutigen – wir haben Fortschritte gemacht. Wir haben jetzt eine Display-Entität, mit der wir die Ziffer unserer Wahl an der ersten Stelle des Displays anzeigen können. Die Pin-Constraints sind gesetzt, sodass die anderen fünf Ziffern problemlos verwendet werden können. Jetzt müssen wir nur noch mit ChatGPT weiterarbeiten und vorankommen! Ich habe ChatGPT gebeten, in kleinen Schritten vorzugehen, damit ich jeden Schritt gut verstehen kann. Es wäre jedoch durchaus möglich, die letzte Anfrage – beispielsweise nach einer Uhr, die Stunden, Minuten und Sekunden auf der 6-stelligen 7-Segment-Anzeige des DE10-Lite zählt – von Anfang an zu stellen und alles auf einmal zu erledigen. Das wäre wahrscheinlich schneller, aber weniger verständlich.

## Auf dem Weg zur funktionierenden Uhr

Hier sind die Schritte, die ich danach unternommen habe, in Form der Prompts an ChatGPT:

- › *Passe den Code an, um „123456“ auf einer 6-stelligen Sieben-Segment-Anzeige darzustellen. Die Display-Register sind HEX0 bis HEX5, eines für jede Ziffer.*
- › *Ausgezeichnet, danke. Schreibe nun den VHDL-Code, um eine Uhr mit dieser Anzeige zu erstellen, die die Zeit im Format hh:mm:ss anzeigt.*
- › *Gebe mir jetzt den Code, den ich hinzufügen muss, um die Uhr stellen zu können. Ich möchte, dass das Drücken von KEY[0] die*

Nach einiger Zeit der Fehlersuche und der Lösung einer Reihe von Problemen (siehe Kasten **Ein paar Probleme**), im Dialog mit ChatGPT und auch durch Google-Recherche, wurde ein funktionierender Code für eine komplette Uhr erstellt. Der endgültige Code ist zu lang, um ihn hier abzdrukken, daher schlage ich vor, dass Sie ihn sich online ansehen [5]. Die fertige Uhr (**Bild 5**) ist unter [6] in Aktion zu sehen. Zuerst wird die Entität `clock_display` deklariert, zusammen mit all ihren Ein- und Ausgängen. Dann werden im Block `architecture` die für den Betrieb erforderlichen Signale für die Speicherung der aktuellen Zeit und des Status der Tasten (ähnlich wie Variablen in der Softwareprogrammierung) erzeugt. In einem ersten Prozess wird aus dem 50-MHz-Takt des DE10-Lite ein 1-Hz-Takt erzeugt, der bis 49.999.999 hochzählt; bei Erreichen dieses Wertes geht das Signal `clk_1hz` für einen Taktzyklus auf High und dann wieder auf Low. Ein zweiter `process` sorgt für die Inkrementierung der Signale für Sekunden, Minuten und Stunden sowie für die Erkennung von Tastenbetätigungen. Die aktuelle Zeit wird jede Sekunde inkrementiert, und die Siebensegmentanzeigen zeigen jede Ziffer mit einer separaten Instanz einer `seven_segment_display`-Entität an. Jede Instanz nimmt eine 4-Bit-Eingabe für die Ziffer entgegen und steuert die entsprechenden Segmente zur Darstellung dieser Ziffer. Zusätzlich werden die beiden



Tasten `KEY[0]` und `KEY[1]` für die manuelle Einstellung der Uhrzeit verwendet. `KEY[0]` erhöht die Minuten und `KEY[1]` erhöht die Stunden. Die Uhr wird zurückgesetzt, wenn der Schalter `SW[0]` aktiviert wird. Die Dezimalpunkte auf der Anzeige dienen zur Trennung von Stunden, Minuten und Sekunden. Die Dezimalpunkte der Displays 0, 1, 3 und 5 werden auf 1 gesetzt und damit ausgeschaltet (Aktiv-Low-Logik), und die anderen werden mit dem langsamen Takt mit den Anweisungen `dpX <= clk_separators` synchronisiert. Auf diese Weise werden alle Segmente der Anzeige explizit verwendet, was es Ihnen ermöglicht, die Option *Reserve all unused pins as output driving ground* zu wählen, um die LEDs `LEDR0` bis `LEDR9` auszuschalten, ohne die Anzeige zu stören. Beachten Sie, dass bei der Programmierung des FPGAs das Design standardmäßig im flüchtigen Speicher des FPGAs gespeichert wird. Beim nächsten Einschalten wird das Terasic-Demoprogramm aus dem internen Flash-Speicher des FPGAs neu geladen. Wenn Sie möchten, dass das FPGA die Uhr bei jedem Einschalten direkt startet, müssen Sie dem Programmierer sagen, dass er die POF-Datei anstelle der SOF-Datei verwenden soll, aber das Programmieren dürfte dann etwas länger dauern.

Alles in allem funktioniert die Uhr sehr gut. Wenn beim Einschalten der untere rechte Schalter in der unteren Position ist, beginnt sie sofort mit dem Zählen der Sekunden. Um die Uhrzeit einzustellen, schieben Sie den Schalter nach oben, um die Uhr auf 00:00:00 zu halten, und schieben Sie dann den Schalter zu Beginn der nächsten Minute erneut, um mit dem Zählen zu beginnen. Die Sekunden sind nun eingestellt. Drücken Sie nun die Tasten `KEY0` und `KEY1`, um die Minuten beziehungsweise Stunden einzustellen.

## Eine großartige Lernerfahrung

Dieses Projekt, bei dem ChatGPT zum Entwurf einer Uhr in VHDL verwendet wurde, war ein neuartiger Ansatz im Vergleich zu der traditionellen Methode, die Grundlagen mit Büchern und anderen Ressourcen zu erlernen und von Grund auf neu zu entwickeln. Mit ChatGPT kann man mit einem größtenteils funktionalen Code beginnen, der, wenn überhaupt, nur sehr wenige Syntaxfehler enthält, was einen schnellen Fortschritt ermöglicht - ein echter Vorteil für diejenigen, die schnell greifbare Ergebnisse sehen möchten. Allerdings muss man bei der Behebung von Logikfehlern und anderen Problemen, die auf dem Weg dorthin auftreten, sorgfältig vorgehen. Natürlich ist diese Methode kein Ersatz für das herkömmliche Lernen, sondern eine wertvolle Ergänzung dazu. So habe ich die am Ende von [1] aufgeführten Ressourcen als sehr nützlich empfunden. ChatGPT (oder jede andere KI) ist nur ein zusätzliches Werkzeug in Ihrem Werkzeugkasten. Es kann komplexe Informationen zusammenfassen, was großartig ist, wenn die herkömmliche Google-Suchen zu kurz greifen. Ich fand es auch nützlich, um

Code zu analysieren, insbesondere unkommentierten Code aus dem Internet, und die zugrunde liegenden Algorithmen zu erklären. Es gibt viele Möglichkeiten, ähnliche Projekte anzugehen, und ich möchte Sie ermutigen, zu experimentieren und Ihre eigenen Lösungen mit anderen Tools wie Claude [7] oder in Ihre IDE eingebundene KI-Assistenten zu erforschen. Wie gesagt, verliert die Uhr die Zeit, wenn sie ausgeschaltet wird, daher ist das Hinzufügen einer Backup-Batterie eine Aufgabe, die dem Leser überlassen bleibt. Die nächsten Schritte könnten darin bestehen, weiter zu lernen und andere Projekte in Angriff zu nehmen. Es wäre toll, Simulationen auszuprobieren, da dies ein wichtiger Teil der FPGA-Entwicklung ist, den ich hier nicht behandelt habe. Das Tolle an FPGA-Code ist seine Portabilität; ich könnte versuchen, diese Uhr auf ein kleineres, günstigeres Board wie das iCE40HX1K-EVB von Olimex zu portieren. Dazu müsste ich eine Platine für die Siebensegmentanzeigen entwerfen und mich mit Multiplexing in VHDL befassen, da es nicht genügend I/O-Pins gibt, um die Segmente direkt anzusteuern. Das Ergebnis wäre eine weitere großartige Ergänzung zu meiner wachsenden Uhrensammlung! ◀

SE – 240602-02

### Sie haben Fragen oder Kommentare?

Haben Sie technische Fragen oder Kommentare zu diesem Artikel? Wenden Sie sich bitte per E-Mail an den Autor unter [jean-francois.simon@elektor.com](mailto:jean-francois.simon@elektor.com) oder an Elektor unter [redaktion@elektor.de](mailto:redaktion@elektor.de).

### Über den Autor

Jean-François Simon (Ingenieur, Elektor) hat eine langjährige Leidenschaft für Elektronik und beschäftigt sich mit so unterschiedlichen Themen wie Schaltungsentwurf, Testen und Messen, Prototyping, Spielen mit SDRs und mehr. Er liebt es, seine Werkzeuge und andere Systeme zu entwickeln, zu modifizieren und zu verbessern. Außerdem interessiert er sich für Mechanik, Maschinen und alles Technische. [https://x.com/JFS\\_Elektor](https://x.com/JFS_Elektor)



### Passende Produkte

- > **Alchitry Au FPGA-Entwicklungsboard (Xilinx Artix 7)**  
[www.elektor.de/19641](http://www.elektor.de/19641)
- > **iCEBreaker FPGA-Entwicklungsboard**  
[www.elektor.de/20625](http://www.elektor.de/20625)

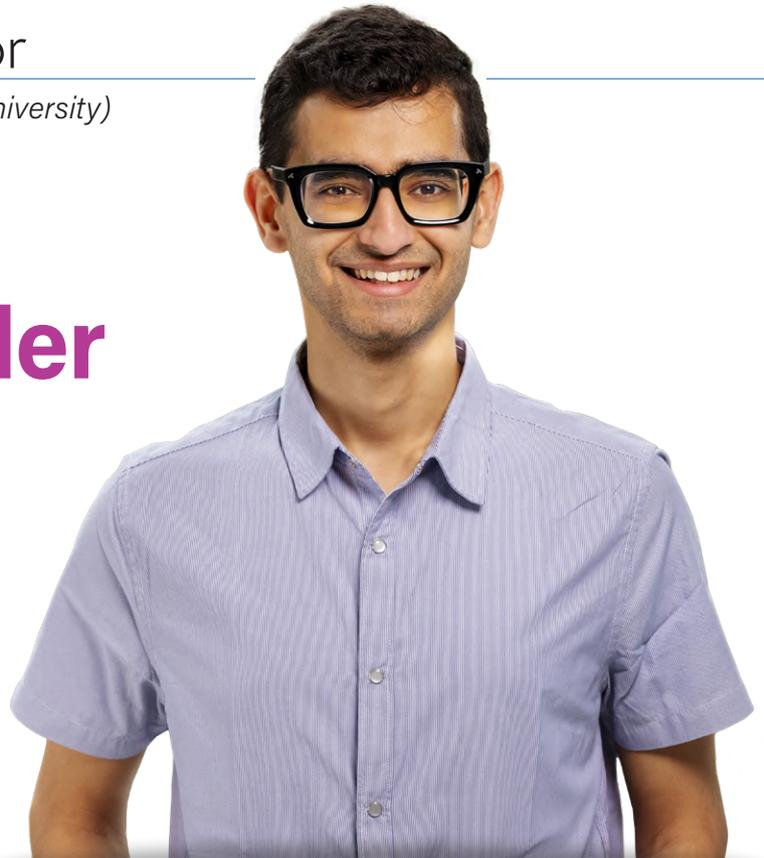
## WEBLINKS

- [1] T. Mulder, „FPGAs für Einsteiger“, Elektor 3-4/2024: <https://www.elektormagazine.de/230067-02>
- [2] Download Quartus Prime Lite: <https://tinyurl.com/downloadquartus>
- [3] System-CD für DE10-Lite: <https://de10-lite.terasic.com/cd>
- [4] FPGA, Blinkende LED, Schritt-für-Schritt-Anleitung (Video): <https://www.youtube.com/watch?v=JJ3XkNcLdx8>
- [5] Quellcode auf GitHub: <https://github.com/jfpsimon/de10lite-clock>
- [6] Uhrendemo (Video): <https://www.youtube.com/watch?v=vdxtk3s0xEA>
- [7] Claude AI: <https://claude.ai>

# Die wahren Auswirkungen der

# KI

Sayash Kapoor über  
„KI-Schlangenöl“ und mehr



Sayash Kapoor (Doktorand, Universität Princeton)

Von C. J. Abate (Elektor)

In diesem Interview spricht Sayash Kapoor, Software-Ingenieur an der Princeton University und ehemaliger Facebook-Mitarbeiter, über die gesellschaftlichen Auswirkungen der KI und die realen Risiken, die sie heute birgt, wie die Verdrängung von Arbeitskräften und fehlerhafte Informationen. Er zeigt auf, wie KI die menschliche Entscheidungsfindung auch verbessern kann, und gibt Einblicke in sein Buch *AI Snake Oil*, das viele übertriebene Behauptungen über künstliche Intelligenz in Frage stellt.

**C. J. Abate:** Sayash, erzählen Sie uns etwas über Ihre Tätigkeit, bevor Sie die Princeton University besuchten.

**Sayash Kapoor:** Bevor ich die Princeton University besuchte, arbeitete ich bei Facebook als Software-Ingenieur, wo ich KI für die Moderation von Inhalten entwickelte.

**C. J.:** Ihre Forschung an der Princeton University befasst sich mit den gesellschaftlichen Auswirkungen der künstlichen Intelligenz. Wann und wie sind Sie auf dieses Thema aufmerksam geworden?

**Sayash:** Während meiner Zeit bei Facebook habe

ich aus erster Hand gesehen, welche Auswirkungen KI auf die Gesellschaft hat. Ich interessierte mich besonders dafür, wie KI in der realen Welt Schäden anrichtet, aber auch dafür, wie sie zur Verbesserung gesellschaftlicher Ergebnisse eingesetzt werden kann. Ich interessierte mich besonders für die Auswirkungen politischer Entscheidungen auf die KI, da ich sah, welche Auswirkungen Richtlinien wie die GDPR (Datenschutz-Grundverordnung) auf die Arbeitsweise großer Unternehmen wie Facebook hatten.

**C. J.:** Ist KI eine dringende existenzielle Gefahr, mit der sich alle Ingenieure und Elektrotechnik-Studenten beschäftigen sollten? Oder ist sie einfach nur ein praktisches Werkzeug, deren Einsatz zwar sorgfältig überlegt werden muss, aber letztlich eine positive Kraft darstellt?

**Sayash:** In den letzten Jahren haben wir viele Behauptungen gehört, dass KI die Menschheit auslöschen wird. In unserem Buch *AI Snake Oil\** haben wir ein ganzes Kapitel diesem Argument gewidmet. Wir erörtern die vielen Trugschlüsse, die den Behauptungen zugrunde liegen, dass die KI die Menschheit auslöschen wird, und kommen auf dieser Grundlage zu dem Schluss, dass wir die von der KI ausgehenden Risiken zwar ernst nehmen sollten, dass aber die derzeitigen Diskussionen über die existenziellen Bedrohungen durch die KI ihre tatsächlichen Auswirkungen weit überbewerten. Gleichzeitig möchte ich die sehr realen Risiken, die KI heute mit sich bringt,

nicht herunterspielen, darunter die Verdrängung von Arbeitskräften, das übermäßige Vertrauen in fehlerhafte Informationen und Sicherheitsbedrohungen.

**C. J.:** Wie stellen Sie sich die Zukunft von KI-gesteuerten Systemen vor, wenn es darum geht, die menschliche Entscheidungsfindung bei komplexen Aufgaben zu verbessern? Was sind die größten Probleme, die Sie bei der Entwicklung von KI-Modellen sehen, die nicht nur gut funktionieren, sondern auch mit menschlichen Werten und Erwartungen in verschiedenen realen Anwendungen übereinstimmen?

**Sayash:** Die KI wird bereits zur Lösung vieler komplexer Aufgaben eingesetzt. Eines der interessanten Dinge an der Geschichte der KI ist, dass wir aufhören, sie KI zu nennen, sobald eine Aufgabe zuverlässig gelöst werden kann. Bis vor ein paar Jahrzehnten galt zum Beispiel die Rechtschreibprüfung als ein schwer zu lösendes Problem. Heute sind Dinge wie die automatische Rechtschreibprüfung aus der Diskussion geraten und zu einem festen Bestandteil des täglichen Lebens geworden.

Ich hoffe, dass es mehr solcher Tools geben wird, die zuverlässig genug arbeiten, dass sie in den Hintergrund treten können. Es gibt bereits KI-Tools, die in der Lage sind, einen Großteil der Wissensarbeit zu verändern. Sie werden für die automatische Entwicklung von Medikamenten und als Programmierassistenten eingesetzt. Die vielleicht größte Herausforderung, um diese Tools für eine große Zahl von Menschen nutzbar zu machen, besteht darin, ihre Zuverlässigkeit zu erhöhen. Die heutigen generativen KI-Systeme sind grundsätzlich stochastisch und die Nutzer haben kein wirklich gutes mentales Modell, wie sie mit ihnen arbeiten können. Chatbots können zum Beispiel oft „halluzinieren“ und geben falsche Antworten auf die Fragen der Nutzer. Wenn es uns gelingt, Fehler und Halluzinationen zu reduzieren, kann die generative KI für konsequente Anwendungen der KI viel nützlicher sein.

**C. J.:** Zusammen mit Arvind Narayanan haben Sie ein Buch mit dem Titel *AI Snake Oil* geschrieben. Was hat Sie dazu inspiriert, dieses Buch zu schreiben, und was macht es zu einer Pflichtlektüre?

**Sayash:** Eine der größten Quellen der Verwirrung über KI ist heute die Tatsache, dass KI ein Oberbegriff ist. Er wird verwendet, um viele verschiedene Arten von Technologien zu bezeichnen, die nur wenig miteinander zu tun haben. In dem Buch *AI Snake Oil* unterscheiden wir zwischen den verschiedenen Arten von KI, um aufzuzeigen, wo KI gut funktioniert, welche Arten von KI keine großen Fortschritte gemacht haben und wie wir die Zukunft dieser Arten von KI erwarten können. Wir sprechen darüber, dass die generative KI, die

Was ist p(doom)? Was sollten Elektor-Community-Mitglieder darüber wissen?



Drafts  

- ◆ P(doom) ist ein Begriff, der im Bereich der KI-Sicherheit verwendet wird, um die Wahrscheinlichkeit katastrophaler Folgen von künstlicher Intelligenz (KI) zu beschreiben. Dies könnte von schweren gesellschaftlichen Störungen bis hin zur Auslöschung der Menschheit reichen... Für die Elektor-Gemeinschaft als Experte für Technologie und Innovation ist es wichtig, sich der potenziellen Risiken im Zusammenhang mit KI bewusst zu sein und zu Diskussionen und Lösungen beizutragen, die zur Minderung dieser Risiken beitragen können.

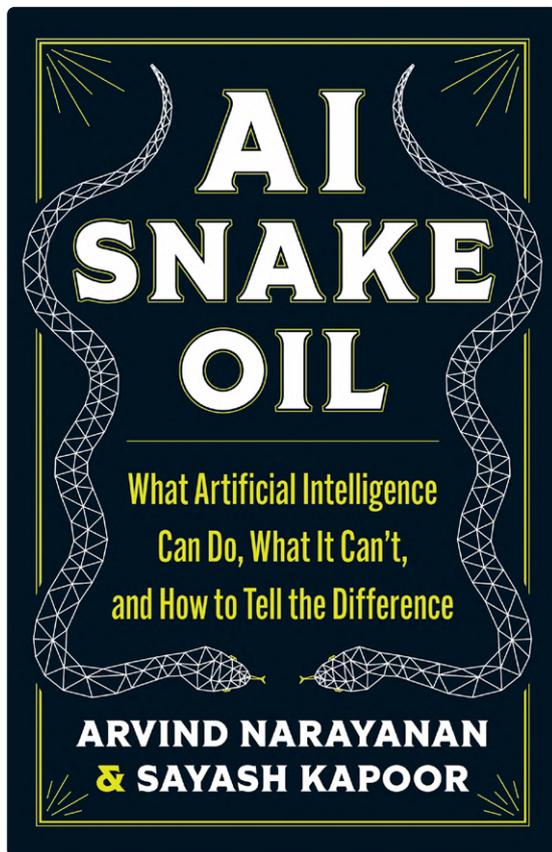
 Enter a prompt for Gemini



zur Generierung von Texten, Bildern, Audiodaten und so weiter eingesetzt wird, in den letzten zehn Jahren enorme Fortschritte gemacht hat - auch wenn wir gewisse Nachteile erwarten, zum Beispiel die Verdrängung von Arbeitskräften und ein übermäßiges Vertrauen in falsche Antworten von Chatbots.



*Wenn wir Fehler  
und Halluzinationen  
reduzieren können,  
kann generative  
KI viel nützlicher  
für nachfolgende  
Anwendungen von  
KI sein.*



Quelle: Sayash Kapoor

Andererseits wird prädiktive KI eingesetzt, um die Zukunft von Personen vorherzusagen und auf dieser Grundlage folgenreiche Entscheidungen über sie zu treffen - etwa, ob ein Angeklagter vor der Verhandlung freigelassen oder ein Stellenbewerber zu einem Vorstellungsgespräch eingeladen werden sollte. Bei unseren Untersuchungen haben wir festgestellt, dass prädiktive KI nicht annähernd so gut funktioniert, wie ihre Entwickler behaupten, und dass sie sich in den letzten Jahrzehnten nicht verbessert hat.

Für alle, die KI in ihrem Leben einsetzen wollen oder Entscheidungen über die Anschaffung von KI für ihre Einrichtungen treffen, hoffen wir, das grundlegende Wissen darüber zu vermitteln, was KI zum Funktionieren bringt, wo sie versagt und wie man den Unterschied erkennt.

**C. J.:** Der Begriff „Snake Oil“ impliziert Täuschung. Nur um das klarzustellen: Wer sind die Betrüger? Die Programmierer und Entwickler? Die Unternehmen, die KI-Lösungen einsetzen? Die Verbraucher? Machen wir uns alle über KI etwas vor?

**Sayash:** Wir reden über die vielen Quellen des KI-Hypes: KI-Forscher, Unternehmen und Journa-

listen, die über KI berichten. Wenn Forscher, die an KI arbeiten, ihre Modelle nicht korrekt evaluieren, führt dies zu übertriebenen Behauptungen darüber, wie gut KI in der realen Welt funktionieren kann. In unserer bisherigen Arbeit haben wir eine Krise der Reproduzierbarkeit entdeckt in der wissenschaftlichen Forschung, die maschinelles Lernen einsetzt. Der Hype stammt auch von Unternehmen, wenn sie Produkte verkaufen, die nicht so gut funktionieren, wie sie behaupten. In den letzten Jahren haben wir viele Behauptungen über KI-Produkte gesehen, die nicht funktionieren - und oft auch nicht funktionieren können -, und dennoch stellen die Unternehmen übertriebene Behauptungen auf, um ihre Produkte zu verkaufen. In einigen Fällen verkaufen Unternehmen Produkte, die als KI vermarktet werden, obwohl in Wirklichkeit ein Mensch hinter den Kulissen die Entscheidungen trifft.

Schließlich geht der Hype um die KI auch auf Journalisten zurück, die falsch darstellen, wie gut die KI funktioniert. Wir haben eine Reihe von Stolperfallen im KI-Journalismus gesammelt und festgestellt, dass News-Artikel über KI, selbst von angesehenen Journalisten, darunter leiden. Beispielsweise werden in News-Artikeln Bilder von Robotern verwendet, um über KI-Anwendungen zu berichten, die nichts mit Robotik zu tun haben, was den Lesern einen irreführenden Eindruck davon vermittelt, wo der Fortschritt der KI stattfindet.

**C. J.:** Erzählen Sie unseren Lesern ein wenig über p(doom). Was ist das? Ist es eine wichtige Maßeinheit?

**Sayash:** P(doom), probability of doom oder die Wahrscheinlichkeit des Untergangs durch KI ist eine Kennzahl, die von der KI-Safety-Community verwendet wird, um das existenzielle Risiko von KI zu beschreiben - die Wahrscheinlichkeit, dass KI uns alle töten wird. Sie ist zu einer beliebten Methode geworden, um zu quantifizieren, wie wahrscheinlich jemand das existenzielle Risiko der KI einschätzt. Üblicherweise werden Wahrscheinlichkeitsschätzungen auf drei Arten begründet: induktiv (auf der Grundlage von Daten aus der Vergangenheit), deduktiv (auf der Grundlage empirisch verifizierter Theorien oder Modelle über die Welt) oder subjektiv (auf der Grundlage von Vermutungen von Prognostikern).

Leider ist keine der oben genannten Möglichkeiten geeignet, um p(doom) zu rechtfertigen. Wir können keine induktiven Schätzungen vornehmen, weil es keine Referenz ähnlicher Ereignisse gibt, von denen wir auf das Existenzrisiko von KI schließen können. Ebenso wenig haben wir etablierte Theorien und Modelle der Welt, die es uns erlauben, das existenzielle Risiko von KI zu beurteilen. So bleiben uns nur

die subjektiven Schätzungen, die von Natur aus spekulativ sind und das Risiko sehr unwahrscheinlicher Folgen überbewerten.

Beunruhigend ist, dass Schätzungen von  $p(\text{doom})$  von politischen Entscheidungsträgern ernst genommen werden. Im Gegensatz zu vielen anderen Erkenntnissen, die sie für ihre politischen Entscheidungen nutzen, sind  $p(\text{doom})$ -Schätzungen von Natur aus spekulativ. Während Einzelpersonen und Unternehmen natürlich alle ihnen zur Verfügung stehenden Informationen nutzen können, um Entscheidungen zu treffen, müssen Regierungen und politische Entscheidungsträger höhere Anforderungen an Nachweise stellen.

**C. J.:** Im September 2023 haben Sie einen Workshop über verantwortungsvolle und offene Grundlagenmodelle mitorganisiert. Wie standen Sie damals zu der Frage, ob KI offen oder geschlossen sein sollte? Sind Sie heute noch der gleichen Meinung?

**Sayash:** Als wir den Workshop organisierten, wollte ich die Argumente für die Nachteile von offenen Grundlagenmodellen besser verstehen. Eines der Hauptergebnisse des Workshops war ein Forschungspapier, in dem wir uns mit dieser Frage auseinandergesetzt haben. Wir fanden heraus, dass viele Behauptungen über Schäden durch offen veröffentlichte KI-Modelle nicht wirklich gerechtfertigt sind.

Im Sommer 2023 veröffentlichte eine Gruppe von Forschern des MIT beispielsweise Papiere, in denen behauptet wurde, dass offene Basismodelle Angreifern bei der Entwicklung von Biowaffen helfen könnten. Wir fanden aber auch heraus, dass die gleichen Informationen, die in diesen offenen Basismodellen verfügbar waren, auch auf Wikipedia zu finden waren. In dem Papier haben wir das Konzept des „marginalen Risikos“ entwickelt, um zu klären, ob das Risiko von offenen Grundlagenmodellen ausgeht oder ob es unabhängig von der Veröffentlichung der Modelle besteht.

Im Wesentlichen geht es darum, das Risiko offener Modelle sowohl mit dem Risiko bestehender Technologien als auch mit dem geschlossener Basismodelle zu vergleichen. Inzwischen hat die Forschung ergeben, dass das Grenzkrisiko offener Grundlagenmodelle gering ist - insbesondere in Bereichen wie der Herstellung von Biowaffen.

**C. J.:** Gibt es etwas, das Sie heute wissen und vor einem Jahr noch nicht wussten, das Ihnen in Bezug auf KI und ihr Potenzial Angst macht? Und wie sieht es andersherum aus? Gibt es etwas, das Sie heute wissen und vor einem Jahr noch nicht wussten, das Sie an der KI und ihrem Potenzial begeistert?

**Sayash:** Ich bin begeistert von dem Potenzial der KI für

Wissensarbeiter. Ich glaube, dass die meisten Wissensarbeiter KI sinnvoll einsetzen können. Zwar müssen noch Probleme wie mangelnde Zuverlässigkeit und Halluzinationen gelöst werden, aber letztlich stellt die generative KI einen großen Sprung in unseren technologischen Möglichkeiten dar. Gleichzeitig bin ich besorgt über den zunehmenden Einsatz von generativer KI für nicht-einvernehmliche Deepfake-Nacktbilder. Wir haben Fälle erlebt, in denen Menschen - unverhältnismäßig viele Frauen - mit Hilfe von KI-generierten Deepfakes zur Zielscheibe wurden.

**C. J.:** Glauben Sie, dass die Elektro-/Kommunikationstechnik-Abteilungen die KI-Sicherheit ernst nehmen? Oder würden Sie sagen, dass es ein Defizit bei der Sensibilisierung und Aufklärung gibt?

**Sayash:** Wir haben definitiv ein großes Interesse an KI-Sicherheit in allen Fachbereichen festgestellt. Einige der führenden Köpfe auf dem Gebiet der KI-Sicherheit in Princeton, zum Beispiel Prof. Prateek Mittal, sind in der Kommunikationstechnik-Abteilung tätig. Ich bin sicher, dass das Interesse in Zukunft noch größer sein wird. ◀

RG — 240555-02

\* Schlangenöl oder Snake Oil bezeichnet die Medizin eines Quacksalbers.

### Sie haben Fragen oder Kommentare?

Wenn Sie technische Fragen oder Anmerkungen haben, senden Sie bitte eine E-Mail an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



**THEMEN FOKUS**

Besuchen Sie unsere Webseite **Embedded & KI** für mehr Artikel, Projekte, Nachrichten und Videos.

[www.elektormagazine.de/embedded-ki](http://www.elektormagazine.de/embedded-ki)



# Das Neueste von BeagleBoard

BeagleY-AI, BeagleV-Fire, BeagleMod, BeaglePlay und  
BeagleConnect Freedom

Ein Beitrag von BeagleBoard

Die BeagleBoard.org-Foundation ist eine gemeinnützige Organisation und führend im Bereich Open-Source-Hardware, das Linux-fähige Boards für Maker, Pädagogen und Fachleute anbietet. Der gemeinschaftsorientierte Charakter von BeagleBoard.org stellt sicher, dass die Nutzer Zugang zu umfangreichem Support, Tutorials und Ressourcen zur Softwareentwicklung haben. In diesem Artikel werden fünf Open-Hardware-Plattformen für Anwendungen wie KI, IoT, Automatisierung und vieles mehr vorgestellt.

## BeagleMod: Eine leistungsstarke Embedded-Computing-Lösung für industrielle Anwendungen

BeagleMod™ (**Bild 1**) ist ein kompaktes, leistungsstarkes Embedded-Computing-Modul, das speziell für Industrie- und Automatisierungsanwendungen entwickelt wurde [1]. BeagleMod entspricht vollständig dem Standard *Open Standard Modules™* (OSM) V1.1, wird vom TI AM62x-Prozessor angetrieben und integriert hochmoderne Hardware für ein breites Spektrum von Anwendungsfällen, einschließlich Mensch-Maschine-Schnittstellen (HMI), Roboterautomatisierung und medizinische Geräte (siehe **BeagleMod: Technische Daten im Überblick**).

### Hauptmerkmale von BeagleMod

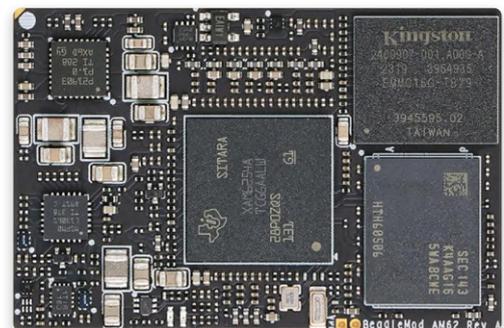
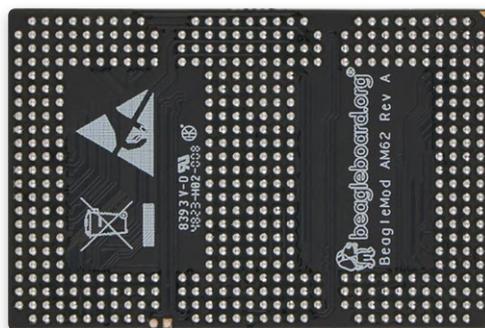
#### 1. Prozessor:

Das Herzstück von BeagleMod ist ein Quad 64-Bit Arm® Cortex®-A53 Prozessor, der eine hohe Verarbeitungsgeschwindigkeit und die Fähigkeit, mehrere Aufgaben gleichzeitig zu bewältigen, gewährleistet. Dies macht ihn ideal für Systeme, die eine robuste Datenverarbeitung und Multitasking-Leistung erfordern.

#### 2. Speicher und Speicherung:

BeagleMod wird mit 2 GB DDR4-RAM geliefert, der auf 4 GB erweitert werden kann, um anspruchsvollere Anwendungsanforderungen zu erfüllen. Das Modul enthält außerdem 16 GB eMMC-Speicher für eine zuverlässige

*Bild 1. BeagleMod™ ist ein kompaktes, hochleistungsfähiges Modul, das mit dem TI AM62x-Prozessor betrieben wird und sich ideal für Anwendungen in den Bereichen Industrieautomation, HMI und medizinische Geräte eignet.*



Datenspeicherung sowie 32 Kbit EEPROM zum Speichern wichtiger Einstellungen und Konfigurationen.

### 3. Stromversorgung:

Das Modul wird über das Fused Tin Grid Array (FTGA) mit einer einzigen 5-V-Gleichstromversorgung betrieben, wodurch es energieeffizient ist und sich leicht in Systeme mit geringer Stromaufnahme integrieren lässt.

### 4. Größe und Formfaktor:

Mit einer Größe von nur 30 mm × 45 mm ist BeagleMod kompakt genug für den Einsatz in platzbeschränkten Umgebungen und bietet gleichzeitig umfangreiche Funktionalität durch seine OSM Size-M-Footprint mit 476 Kontakten.

### 5. Betriebsbedingungen:

Das Modul ist sowohl für Standard- als auch für raue Umgebungen ausgelegt. Es arbeitet effektiv bei Temperaturen von 0°C bis +70°C, wobei eine optionale Industrieversion für -40°C bis +85°C ausgelegt ist. Es bietet außerdem eine Feuchtigkeitstoleranz von 10 % bis 95 %.

### 6. Konformität und Zertifizierungen:

BeagleMod ist nach mehreren Standards zertifiziert, darunter FCC, CE, REACH und RoHS, und gewährleistet so die Einhaltung globaler Vorschriften und Umweltsicherheitsanforderungen.

### Umfangreiche Schnittstellenunterstützung

BeagleMod unterstützt eine Vielzahl von Schnittstellen und ist damit eine flexible Lösung für Industrie-, Automobil- und andere Embedded-Anwendungen:

- JTAG für Debugging und Entwicklung.
- 5 UART-Anschlüsse für die serielle Kommunikation.
- 2 RGMII-Schnittstellen für 10/100/1000 Mbit Ethernet-Konnektivität.
- 2 SDIO-Schnittstellen (4 Bit) für Speicherkarten und andere Peripheriegeräte.
- 6 PWM-Ausgänge zur Steuerung von Motoren, LEDs oder anderen Geräten.
- 2 SPI-Schnittstellen, einschließlich OSPI (8-Bit) für die schnelle Kommunikation mit Sensoren, Speichergeräten oder anderen Mikrocontrollern.
- 2 I<sup>2</sup>S-Schnittstellen für die Übertragung von Audiodaten.
- 2 CAN-Busse für die Fahrzeugkommunikation.
- 2 USB 2.0-Anschlüsse für Peripheriegeräte.
- 3 I<sup>2</sup>C-Schnittstellen für den Anschluss an verschiedene Sensoren und Geräte.
- 1 CSI (4 Lane) für Kamera-Schnittstellen.
- 1 RGB-Schnittstelle (18 Bit) für Displays.
- 1 OLDI (2×4 Lane LVDS) für hochauflösende Displays.
- GPIOs für benutzerdefinierte Steuereingänge und -ausgänge.

### Anwendungen von BeagleMod

Der hohe Integrationsgrad von BeagleMod, gepaart mit seiner kompakten Größe und robusten Verarbeitungsleistung, macht es ideal für eine Vielzahl von Anwendungen:

- Mensch-Maschine-Schnittstelle (HMI):

## BeagleMod: Technische Daten im Überblick

Komponente	Spezifikation
Prozessor	TI AM62 Quad-Core A53
RAM	2 GB DDR4, erweiterbar auf 4 GB
Speicher	16 GB eMMC, 32 Kbit EEPROM
Stromversorgung	5 V DC über OSM FTGA
Formfaktor	30 mm × 45 mm, OSM Size-M
Betriebstemperatur	0°C to +70°C, -40°C to +85°C (optional)
Betriebsfeuchtigkeit	10%~95% RH
Zertifizierungen	FCC, CE, REACH, RoHS/JTAG, UART, RGMII, SDIO, PWM, SPI, I2S, CAN, USB 2.0, I2C, CSI, RGB, OLDI, GPIO
Schnittstellen	JTAG, UART, RGMII, SDIO, PWM, SPI, I2S, CAN, USB 2.0, I2C, CSI, RGB, OLDI, GPIO
Zusätzliche Funktionen	OSM FTGA, GPIOs, verschiedene Steuerschnittstellen

Das Modul kann in Benutzerschnittstellensysteme für die industrielle Automatisierung integriert werden und bietet eine reibungslose und reaktionsschnelle Steuerung von Maschinen, Industriesystemen und Überwachungspanels.

#### ➤ Robotik-Automatisierung:

Der leistungsstarke Cortex-A53-Prozessor, kombiniert mit Echtzeit-GPIO-, PWM- und I<sup>2</sup>C-Schnittstellen, ermöglicht die präzise Steuerung von Roboterarmen, autonomen Maschinen und Fabrikautomationssystemen.

#### ➤ Smart Home und Energieeinsparung:

Die vielseitigen Schnittstellen und der niedrige Stromverbrauch von BeagleMod eignen sich perfekt für Smart-Home-Anwendungen, einschließlich intelligenter Beleuchtungssysteme, HLK-Steuerungen und Energiemanagementgeräte.

#### ➤ Medizinische Ausrüstung:

Die Integration mehrerer UARTs, SPI- und I<sup>2</sup>S-Schnittstellen ermöglicht den Einsatz von BeagleMod in medizinischen Diagnosegeräten, Patientenüberwachungssystemen und anderen sensiblen medizinischen Geräten, die eine zuverlässige Echtzeit-Datenverarbeitung erfordern.

### Anpassung und Evaluierungskit

BeagleMod bietet Flexibilität für benutzerdefinierte Konfigurationen und Erweiterungen, um spezifischen Anforderungen gerecht zu werden. Seeed Studio bietet Anpassungsdienste, mit denen Benutzer BeagleMod genau an ihre Anforderungen anpassen können. Das BeagleMod Evaluation Kit enthält das Kernmodul mit Entwicklungswerkzeugen für den schnellen Einstieg.

## BeagleY-AI: Eine leistungsstarke KI-Edge-Computing-Plattform

BeagleY-AI [2], entwickelt von der BeagleBoard.org Foundation, ist ein leistungsstarker Einplatinencomputer,

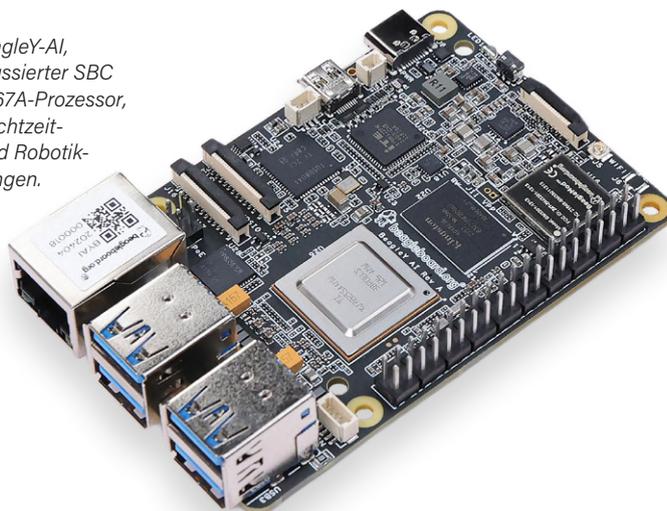
der für KI-gesteuertes Edge Computing konzipiert wurde (**Bild 2**). Sein Herzstück, der AM67A-Prozessor von Texas Instruments, bietet eine außergewöhnliche Mischung aus Leistung und Skalierbarkeit und zielt auf Anwendungen ab, die Echtzeitverarbeitung für KI, Bildverarbeitung und Robotik erfordern (siehe **BeagleY-AI: Technische Daten im Überblick**).

### Wesentliche Merkmale

#### 1. Rechenleistung:

Der AM67A SoC integriert eine Quad 64-Bit Arm Cortex-A53-CPU mit 1,4 GHz sowie zwei C7x DSPs für KI-Workloads, die jeweils 4 TOPS an Rechenleistung bieten und sich damit ideal für Deep Learning- und Vision-basierte Anwendungen eignen. Diese Prozessoren werden durch den Matrix Multiply Accelerator (MMA) unterstützt, der schnellere KI-Modellinferenzen

*Bild 2. BeagleY-AI, ein KI-fokussierter SBC mit TI AM67A-Prozessor, ideal für Echtzeit-Vision- und Robotik-Anwendungen.*



### BeagleY-AI: Technische Daten im Überblick

Komponente	Spezifikation
Prozessor	TI AM67A, Quad 64-bit Cortex-A53 @1,4 GHz
KI-Beschleuniger	Dual C7x DSPs mit je 4 TOPS
RAM	4 GB LPDDR4
Wi-Fi	802.11ax (Wi-Fi 6)
Bluetooth	Bluetooth 5.4 (BLE)
USB-Anschlüsse	4 × USB 3.0, 1 × USB 2.0 Type-C
Ethernet	Gigabit Ethernet mit PoE+
Kamera	2 × 4-Spur-MIPI-CSI-Schnittstellen
Display	HDMI, OLDI, und DSI MIPI
Speicher	microSD-Kartensteckplatz mit UHS-1-Unterstützung
Erweiterung	40-polige GPIO-Stiftleiste
Stromversorgung	5 V/3 A über USB-C
Lüfteranschluss	PWM-gesteuerte 4-polige Stiftleiste

und so intelligente, KI-verbesserte Systeme ermöglicht.

#### 2. Speicher und Speicherung:

Ausgestattet mit 4 GB LPDDR4-RAM sorgt der BeagleY-AI für eine effiziente Handhabung von speicherintensiven KI-Modellen und Echtzeit-Datenverarbeitung. Außerdem verfügt es über einen microSD-Kartensteckplatz mit UHS-1-Unterstützung, der eine schnelle Datenübertragung und zuverlässiges Booten ermöglicht. Das Board bietet reichlich Speicherplatz für die Ausführung komplexer KI-Algorithmen und die Verarbeitung großer Mengen an Sensordaten.

#### 3. Konnektivität und Erweiterung:

Mit 4 USB 3.0-Anschlüssen, einem USB 2.0 Type-C-Anschluss und Gigabit-Ethernet gewährleistet BeagleY-AI nahtlose Konnektivität mit Peripheriegeräten und Hochgeschwindigkeits-Netzwerkcommunication. Die PCIe-Gen3-Schnittstelle unterstützt die Integration von Hochleistungs-Peripheriegeräten und erweitert den Nutzen des Boards für Anwendungen, die fortschrittliche KI-Berechnungen und Netzwerke erfordern. Der 40-polige GPIO-Header ist mit bestehenden Raspberry-Pi-HATs kompatibel, so dass Benutzer schnell KI-Anwendungen mit zusätzlichen Sensoren, Aktoren und E/A-Optionen entwickeln und einsetzen können.

#### 4. KI und Vision-Fähigkeiten:

BeagleY-AI wurde für KI- und Bildverarbeitungsanwendungen entwickelt und verfügt über zwei MIPI-CSI-Kamerainterfaces, an die mehrere hochauflösende Kameras für Aufgaben wie Echtzeit-Videoanalyse, maschinelles Sehen und autonome Robotik angeschlossen werden können. Die On-Chip-GPU und Videobeschleuniger sorgen dafür, dass Videostreaming und Bilderkennungsaufgaben effizient verarbeitet werden, ohne die CPU-Kerne zu belasten.

#### 5. Stromversorgung und Steuerung:

Das Board wird mit 5 V/3 A Gleichstrom über USB-C versorgt und ist somit mit einer Vielzahl von Stromquellen kompatibel. Es enthält einen Lüfter-Anschluss mit PWM-Steuerung, der sicherstellt, dass leistungsstarke KI-Workloads unter angemessenem Wärmemanagement aufrechterhalten werden.

### Edge-AI-Anwendungen

BeagleY-AI ist für eine Vielzahl von KI-gesteuerten Edge-Anwendungen bestens geeignet:

- **Mensch-Maschine-Schnittstellen (HMI):** Echtzeit-Feedback und Steuerungssysteme.
- **Industrielle Automatisierung:** KI-basierte Überwachung, vorausschauende Wartung und Prozessautomatisierung.
- **Robotik:** Autonome Navigation, Objekterkennung und -manipulation mithilfe fortschrittlicher KI-Modelle.
- **Videoüberwachung und -analyse:** Leistungsstarke Objekterkennung, Gesichtserkennung und Musteranalyse in Sicherheitssystemen.
- **Medizinische Geräte:** KI-Anwendungen für die

Patientenüberwachung, Diagnostik und Automatisierung im Gesundheitswesen.

### Zukunftssichere KI-Plattform

BeagleY-AI ist als zukunftssichere Plattform konzipiert, die eine nahtlose Integration von KI-Anwendungen im Edge-Bereich ermöglicht und gleichzeitig mit mehreren E/A-Optionen und Peripherieunterstützung flexibel bleibt. Seine skalierbare Architektur stellt sicher, dass er auch die anspruchsvollsten Aufgaben in Branchen wie Gesundheitswesen, Sicherheit und Automatisierung bewältigen kann.

Eine Rezension der BeagleBoardY-AI von Elektor-Redakteur Brian T. Williams finden Sie in dieser Ausgabe und im Elektor-Archiv ([www.elektormagazine.de/240527-02](http://www.elektormagazine.de/240527-02)).

### BeagleV-Fire: RISC-V- und FPGA-gesteuertes Edge-Entwicklungsboard

BeagleV-Fire [3] ist ein bahnbrechender Einplatinencomputer (SBC) (**Bild 3**), der von der BeagleBoard.org Foundation entwickelt wurde und auf dem Microchip-SoC PolarFire® MPFS025T basiert. Er integriert RISC-V-Cores und eine FPGA-Fabric und bietet Entwicklern die Flexibilität, das Potenzial sowohl der RISC-V-Architektur als auch der programmierbaren Logik zu nutzen. BeagleV-Fire ist ideal für KI, Robotik und industrielle Automatisierung und bietet die notwendige Hardware für Echtzeit-Edge-Computing-Anwendungen (siehe **BeagleV-Fire: Technische Daten im Überblick**).

#### Wesentliche Merkmale

##### 1. Prozessor:

Das Herzstück von BeagleV-Fire ist der MPFS025T-SoC, der Folgendes umfasst:

- 4x 64-Bit-RISC-V-Kerne (RV64GC) für die Anwendungsverarbeitung
- 1x 64-Bit-RISC-V-Boot-Core (RV64IMAC) für die Systemverwaltung
- Programmierbare FPGA-Fabric, die Entwicklern die Möglichkeit gibt, benutzerdefinierte Hardware-Aufgaben wie Signalverarbeitung, Kryptografie oder KI-Inferenz zu beschleunigen.

##### 2. Speicher und Speicherung:

- 2 GB LPDDR4-RAM mit 1866 MHz sorgen für effizientes Multitasking und Datenverarbeitung.
- 16 GB eMMC-Flash-Speicher bieten reichlich Onboard-Speicher für Linux-Distributionen und -Anwendungen, wobei zusätzlicher Speicher über den microSD-Steckplatz unterstützt wird.

##### 3. Konnektivität und Schnittstellen:

- Gigabit-Ethernet mit Realtek RTL8211F PHY für Hochgeschwindigkeitsnetzwerke.
- USB-C-Anschluss für die Stromversorgung des Boards, die Programmierung und die Datenübertragung.

### BeagleV-Fire: Technische Daten im Überblick

Komponente	Spezifikation
Prozessor	PolarFire® MPFS025T SoC mit 4x RISC-V Kernen, 1x Boot-Kern
RAM	2 GB LPDDR4
Speicher	16 GB eMMC, microSD-Steckplatz
Ethernet	Gigabit Ethernet, Realtek RTL8211F PHY
Wi-Fi	M.2 Key E Steckplatz für 2,4 GHz/5 GHz
USB-Anschlüsse	USB-C für Strom und Daten
Kamera	CSI Kamera-Schnittstelle (22-polig)
Erweiterung	P8 & P9 Cape-Anschlüsse, SYZYGY-Stecker
Stromversorgung	USB-C (5 V/3 A)
Debugging	JTAG, UART-Anschlüsse
LEDs	12x programmierbare LEDs
Abmessungen	86,38 x 54,61 mm, 12-lagige Leiterplatte, RoHS-konform

- Ein M.2-Key-E-Steckplatz zum Hinzufügen von WLAN-Modulen, der es Entwicklern ermöglicht, drahtlose Kommunikation zu integrieren, z. B. 2,4 GHz/5 GHz Wi-Fi.
- CSI-Kameraanschluss, kompatibel mit dem BeagleBone AI-64, BeagleV-Ahead und Raspberry Pi Zero/CM4, der eine hochwertige Bild- und Videoeingabe für bildverarbeitungs-basierte Anwendungen ermöglicht.

##### 4. Erweiterungsanschlüsse:

BeagleV-Fire behält die vertrauten P8- und P9-Cape-Header bei und bietet Kompatibilität mit BeagleBone Black-Capes, was eine einfache Integration mit bestehenden Add-ons ermöglicht. Darüber hinaus bietet der SYZYGY-Hochgeschwindigkeitsanschluss Zugang zu Hochleistungsperipheriegeräten für fortschrittliche Hardware-Erweiterungen.

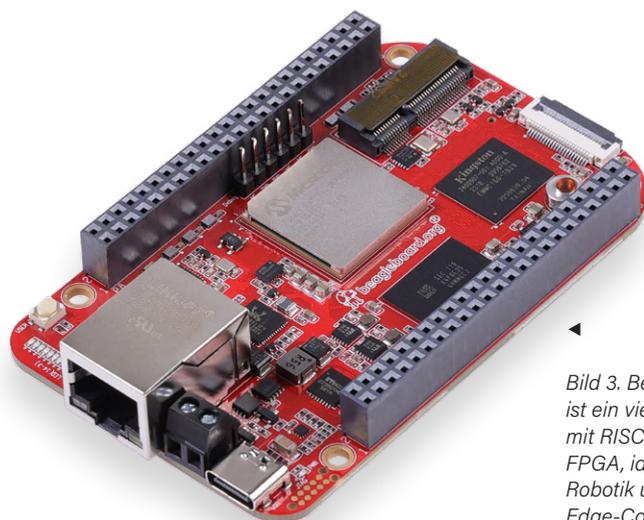


Bild 3. BeagleV-Fire ist ein vielseitiger SBC mit RISC-V-Cores und FPGA, ideal für KI, Robotik und Echtzeit-Edge-Computing.

## 5. Fehlersuche und Entwicklung:

Das Board ist ausgestattet mit:

- JTAG- und UART-Debug-Header für Low-Level-Hardware-Debugging und Überwachung.
- Benutzerprogrammierbare LEDs und Tasten zur Verbesserung der Entwicklungserfahrung und des System-Feedbacks.

## Ideale Anwendungen

- Edge AI und Industrielle Automatisierung:

Mit seinen Möglichkeiten der Echtzeitverarbeitung und dem programmierbaren FPGA ist BeagleV-Fire ideal für Aufgaben wie vorausschauende Wartung, KI-gestützte Qualitätskontrolle und intelligente Fabrikautomatisierung. Dank seines geringen Stromverbrauchs eignet er sich für Edge-KI-Anwendungen und ermöglicht Echtzeitverarbeitung, ohne auf eine Cloud-Infrastruktur angewiesen zu sein.

- Robotik und Bildverarbeitungssysteme:

Dank der integrierten CSI-Kameraunterstützung und der Möglichkeit, das Board über die P8/P9-Header mit Sensoren und Aktoren zu verbinden, eignet es sich hervorragend für die bildverarbeitungsgesteuerte Robotik, autonome Systeme und KI-basierte Objekterkennung.

- Erkundung von Open-Source-Hardware:

Entwickler, Tüftler und Forscher können die hochmoderne RISC-V-Architektur erkunden und gleichzeitig FPGAs zur Beschleunigung von Aufgaben wie Datenverschlüsselung, Deep Learning und Echtzeit-Datenverarbeitung nutzen. Die Verfügbarkeit umfangreicher Open-Source-Ressourcen in Verbindung mit der reichhaltigen E/A des Boards macht es zu einer hervorragenden Wahl für Hardware-Experimente und Prototyping.

## Erste Schritte und Entwicklungswerkzeuge

BeagleV-Fire wird von einem umfassenden Entwicklungs-Ökosystem unterstützt. Entwickler können die neueste Gateway schnell und ohne zusätzliches Tool unter Linux flashen. Gateway kann auch mit Flash-Pro5/6 und der Libero SoC Design Suite geflasht werden, was das FPGA-Design und Debugging erleichtert. Das Flashen unter Linux ist so einfach wie das Flashen einer SD-Card mit Tools wie Balena Etcher. Das Board unterstützt auch Cockpit, SSH und seriellen Port-Zugang für Entwicklung und Systemmanagement, was es Entwicklern leicht macht, in ihre Projekte einzutauchen.

BeagleV-Fire bietet die einzigartige Möglichkeit, mit einem RISC-V-SoC gepaart mit einer FPGA-Fabric in einem kostengünstigen, kompakten SBC zu arbeiten. Ob Sie KI at the Edge erforschen, autonome Systeme entwickeln oder benutzerdefinierte hardwarebeschleunigte Anwendungen erstellen, BeagleV-Fire bietet eine leistungsstarke, flexible Plattform, um Innovationen voranzutreiben.

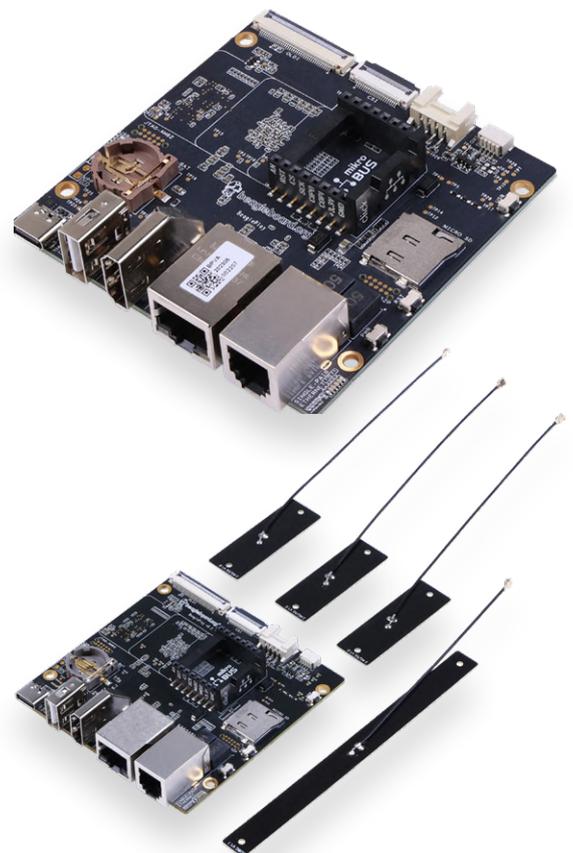
Weitere Informationen zum BeagleV-Fire und Anleitungen finden Sie in der offiziellen Dokumentation von BeagleBoard.org [4].

## BeaglePlay und BeagleConnect Freedom: Ein leistungsfähiges Duo für die IoT-Entwicklung

BeaglePlay [5] und BeagleConnect Freedom [6] (**Bild 4** und **Bild 5**) sind zwei komplementäre Boards der BeagleBoard.org Foundation, die die IoT-Entwicklung durch leistungsstarke Open-Source-Hardware mit einfacher Sensor- und Konnektivitätsintegration vereinfachen sollen.

### BeaglePlay: Ein IoT-Gateway mit unübertroffener Konnektivität

BeaglePlay ist ein Einplatinencomputer mit dem Quad-Core-Prozessor AM6254 von Texas Instruments, der eine hohe Rechenleistung bietet. Er eignet sich hervorragend für die Erstellung eines IoT-Gateways oder eines Automatisierung-Hubs mit seiner breiten Palette an Konnektivitätsoptionen. Ganz gleich, ob Sie Wi-Fi (2,4 GHz/5 GHz), Gigabit-Ethernet, Single-Pair-Ethernet oder Sub-1-GHz-Wireless benötigen, BeaglePlay macht den Anschluss von IoT-Sensoren und Peripheriegeräten durch integrierte Schnittstellen wie mikroBUS, Grove und QWIIC einfach. Außerdem unterstützt es HDMI für Displays und CSI für Kameras, was es vielseitig für industrielle und private IoT-Lösungen einsetzbar macht.



▲ Bild 4. BeaglePlay, ein IoT-Gateway-SBC mit dem TI AM6254-Prozessor, bietet umfangreiche Konnektivitätsoptionen für den Aufbau vielseitiger IoT-Hubs.



### Wesentliche Merkmale:

- Prozessor: AM6254 (Quad-Core Arm Cortex-A53)
- Speicher: 2 GB DDR4, 16 GB eMMC-Speicher
- Konnektivität: Wi-Fi, Ethernet, BLE, Sub-1 GHz, Single-Pair-Ethernet
- Erweiterung: mikroBUS, Grove, QWIIC, CSI, und mehr

Mit seiner Plug-and-Play-Architektur und seiner Benutzerfreundlichkeit macht BeaglePlay das Hinzufügen von Sensoren, Aktoren und Anzeigen mühelos. Seine Kompatibilität mit Tausenden von handelsüblichen Erweiterungsmodulen vereinfacht das Prototyping und die Entwicklung kompletter Systeme.

### BeagleConnect Freedom: Drahtloses IoT mit Leichtigkeit

BeagleConnect Freedom ist eine drahtlose Open-Hardware-Plattform, die auf dem TI CC1352P7-Mikrocontroller basiert und damit eine ideale Ergänzung zu BeaglePlay für die Entwicklung drahtloser IoT-Knoten mit geringem Stromverbrauch darstellt. Sie unterstützt 2,4-GHz Bluetooth Low Energy (BLE) und Sub-1-GHz-Wireless-Protokolle, die eine Kommunikation mit großer Reichweite für Sensorknoten ermöglichen. Mit zwei mikroBUS-Anschlüssen, integrierten Sensoren und Akkuladefunktionen ist er bereit für schnelles IoT-Prototyping.

### Wesentliche Merkmale:

- Prozessor: CC1352P7 (48-MHz-Arm Cortex-M4)
- Drahtlos: BLE, Sub-1 GHz IEEE 802.15.4
- Erweiterung: 2 mikroBUS-Buchsen für Sensoren und Aktoren

- Stromversorgung: LiPo-Akku-Unterstützung mit Lademöglichkeit

BeagleConnect Freedom eignet sich perfekt für den Aufbau von IoT-Knoten, die drahtlos mit BeaglePlay kommunizieren und so ein robustes Netzwerk von verbundenen Geräten bilden.

### So verwenden Sie BeaglePlay und BeagleConnect Freedom gemeinsam

Beim gemeinsamen Einsatz fungiert BeaglePlay als zentraler Hub oder Gateway, während BeagleConnect Freedom als drahtlose Sensorknoten dient. So können Sie beispielsweise BeaglePlay als zentralen Prozessor verwenden, um Daten von mehreren BeagleConnect Freedom-Geräten zu sammeln, die über ein großes Gebiet verteilt sind. Diese Kombination ermöglicht es Ihnen, ein Mesh-Netzwerk von Sensoren einzurichten, ohne sich um die Verkabelung kümmern zu müssen, und ist somit ideal für intelligente Häuser, industrielle Automatisierung oder groß angelegte Umweltüberwachung.

### So können Sie ihre Stärken kombinieren:

1. Richten Sie BeaglePlay als Ihren IoT-Hub ein und verbinden Sie ihn über WLAN oder Ethernet mit Ihrem Netzwerk.
2. Stellen Sie BeagleConnect Freedom-Knoten in Ihrem Haus, Ihrer Fabrik oder Ihrer Umgebung auf, um Sensordaten (wie Temperatur oder Luftfeuchtigkeit) drahtlos an BeaglePlay zu übertragen.
3. Verwenden Sie Zephyr RTOS auf BeagleConnect Freedom, um Sensorknoten zu programmieren und zu verwalten.
4. Überwachen oder steuern Sie das Netzwerk durch die leistungsstarke Verarbeitung und Cloud-Integrationen von BeaglePlay.

Diese Kombination aus der hohen Konnektivität von BeaglePlay und den drahtlosen Low-Power-Fähigkeiten von BeagleConnect Freedom schafft ein flexibles, skalierbares und effizientes IoT-Ökosystem für Bastler und Profis gleichermaßen.

Mit diesen Tools war die IoT-Entwicklung noch nie so zugänglich wie heute! Egal, ob Sie ein erfahrener Entwickler sind oder gerade erst anfangen, BeaglePlay und BeagleConnect Freedom bieten die Hardware- und Software-Flexibilität, die Sie brauchen. ◀

Ulrich Drees — 240630-02

*Bild 5. BeagleConnect Freedom, eine drahtlose Plattform mit dem TI CC1352P7 Mikrocontroller, unterstützt IoT-Knoten mit großer Reichweite mit BLE und Sub-1-GHz-Protokoll.*

## WEBLINKS

- [1] BeagleMod: <https://www.beagleboard.org/boards/beaglemod>
- [2] BeagleY-AI: <https://www.beagleboard.org/boards/beagle-y-ai>
- [3] BeagleV-Fire: <https://www.beagleboard.org/boards/beagle-v-fire>
- [4] Dokumentation von BeagleBoard.org: <https://docs.beagleboard.org/>
- [5] BeaglePlay: <https://www.beagleboard.org/boards/beagleplay>
- [6] BeagleConnect Freedom: <https://www.beagleboard.org/boards/beagleconnect-freedom>

# Moskito- Erkennung mit offenen Daten und Arduino Nicla Vision

Von José Bagur (UVG), João Vitor Freitas da Costa (UERJ), Silvia A. Sotelo López (UPB),  
Peter Ing (Arm SAAIA), und David J. Cuartielles Ruiz (Arduino/Malmö University)

Angesichts der dringenden Problematik wachsender Mückenpopulationen und ihrer Rolle bei der Krankheitsübertragung hat ein Team ein Erkennungssystem entwickelt, das Mücken anhand ihrer Geräusche identifiziert und automatisch Abwehrnebel aktiviert. Das Projekt reagiert auf den alarmierenden Anstieg der Dengue-Fälle in ganz Amerika und zielt darauf ab, eine effektive Lösung für die Mückenerkennung in realen Umgebungen zu schaffen. Entdecken Sie, wie tinyML-Technologie und internationale Zusammenarbeit zur nachhaltigen Entwicklung beitragen können!





Im Sommer 2024 organisierten das *International Centre for Theoretical Physics* (ICTP) gemeinsam mit der Harvard University, dem Barnard College, UNIFEI, IBM Research Brazil und Arduino den Workshop *TinyML for Sustainable Development* im brasilianischen São Paulo. Der Workshop brachte engagierte Teilnehmer aus Nord- und Südamerika zusammen, die sich mit der Entwicklung intelligenter Edge-Systeme beschäftigten. Über eine Woche hinweg konnten die Teilnehmer von Experten lernen und sich zu Themen wie Datenerfassung, Edge-Devices-Training, DIY-LLMs und tinyML-Anwendungen fortbilden. Jeder Teilnehmer war eingeladen, eigene Beiträge und Ideen zu diesen Themen zu präsentieren.

Während dieser intensiven fünf Tage arbeiteten die Teilnehmer in Gruppen an verschiedenen technischen Lösungen und lernten sich durch Vorträge, praktische Übungen und Besuche an der Universität von São Paulo einschließlich ihrer Platinen-Fertigung besser kennen. Wir (die Autoren dieses Artikels) entschieden uns, gemeinsam potenzielle Lösungen für die zunehmende Verbreitung bestimmter Mückenarten zu entwickeln. Heute übertragen Stechmücken Krankheiten, die früher als ausgerottet galten, und ihre jährlichen Fortpflanzungszyklen haben in vielen Gebieten zugenommen. Wir kamen auf die Idee, ein System zu entwickeln, das bestimmte Arten von Moskitos erkennt, indem es auf die von ihren Flügeln erzeugten Geräusche hört und daraufhin einen Nebel aussendet, der die Insekten vertreibt.

Die exponentielle Zunahme der Dengue-Fälle in Nord- und Südamerika stellte die Hauptmotivation für dieses Projekt dar. Im Jahr 2024 wurden 11.517.728 Fälle registriert – ein erschreckender Anstieg von 230 % im Vergleich zu den gemeldeten Fällen von 2023 (PAHO-Bericht, September 2024 [1]).

Um dieses Projekt nachzubauen, benötigen Sie:

- Einen Arduino Nicla Vision [2]. Wir werden uns nur auf die Fähigkeiten des Geräts zur Geräuscherkennung konzentrieren.
- Den Datensatz mit Mückengeräuschen, die eine Erkennung ermöglicht
- Ein Konto bei Edge Impulse, der Online-Plattform, die die Inferenz-

maschine für Sie erstellt, nachdem Sie sie mit dem genannten Datensatz gefüttert haben

- Eine Relaisplatine, die mit einem 3,3-V-Signal des Nicla Vision ausgelöst werden kann
- Ein Piezo-Befeuchter
- Jumper-Drähte
- Einen LötKolben. Die Nicla Vision-Platine verfügt über so genannte Castellated Pads (Halbloch-Kantenkontakte), jedoch keine Anschlussstifte. Wir mussten Drähte anlöten, um die Platine mit dem Rest der Maschine zu verbinden.

Für den Workshop stellte Arduino die benötigten Boards zur Verfügung, und Edge Impulse bietet kostenlose Konten für Maker an, um basierend auf dem Datensatz den Code für die Boards zu erstellen. Die Relaisplatine und den Luftbefeuchter mussten wir vor Ort kaufen, aber in São Paulo gibt es natürlich mehrere Läden mit elektronischen Bauteilen.

## Hintergrund

Stechmücken haben ein kurzes, aber intensives Leben. Viele Arten sind bemerkenswert widerstandsfähig gegen Umwelteinflüsse, und ihre befruchteten Eier können auch unter unwirtlichen Bedingungen überleben, bis günstige Temperatur- und Luftfeuchtigkeitsbedingungen ihre Eklosion (Schlüpfen) ermöglichen. Stechmücken wurden und werden intensiv erforscht: Sie ernähren sich vom Blut anderer Tiere und können so durch die verschiedenen Bluttypen, die sie aufnehmen, Krankheiten übertragen. Zum Beispiel hat das Prometeo-Projekt von Microsoft gezeigt, dass eine einzige Mücke DNA-Spuren von mehreren Säugetierarten wie Kühen, Menschen, Hunden oder Schafen enthalten kann. Diese Fähigkeit zur schnellen Übertragung zwischen Arten macht Mücken zu effizienten Krankheitsüberträgern.

Interessanterweise lässt sich anhand der Flügelgröße nicht nur die Art der Mücke, sondern auch das Geschlecht unterscheiden. So ist es möglich, Mücken anhand der spezifischen Geräusche zu klassifizieren, die durch ihre unterschiedlich konstruierten Flügel entstehen. Dies ist für Menschen schwierig, aber eine gut trainierte tinyML-Maschine kann dies leisten. Der oben genannte Datensatz, der Tausende von Mückenaufnahmen

enthält, dient als wertvolle Ressource, um eine Inferenzmaschine zu entwickeln, die Mückengeräusche in ihrer Umgebung erkennen und bewerten kann.

Eine Herausforderung, die sich während des Projekts stellte und die wir zunächst unterschätzten, war das Vorhandensein zweier verschiedener Netzspannungen in Brasilien von 110 V und 230 V. Beim Kauf der Bauteile besorgten wir auch einen günstigen LötKolben, da keiner von uns einen mitgebracht hatte. Da die Nicla Vision Halbloch-Kantenkontakte, jedoch keine Pins hat, mussten wir zur Überbrückung Drähte anlöten, um die Platine mit der Relaisplatine zu verbinden. Im Workshop-Raum stellte sich jedoch heraus, dass unser LötKolben nur für 230 V geeignet war und bei 110 V nicht die Leistung aufbrachte, um die Lötspitze ausreichend zu erhitzen. Schließlich konnten wir in unserem Hotelzimmer an der Steckdose für den Rasierapparat löten – eine improvisierte und dennoch effektive Lösung, aber für zukünftige Projekte nehmen wir definitiv unseren USB-C-LötKolben mit!

## Die Idee und die Umsetzung

Im Rahmen des erwähnten Workshops beschlossen wir, ein Proof-of-Concept zu entwickeln, das die Anwesenheit einer Mücke erkennt und einen Nebel erzeugt, der verschiedene Arten von Essenzen wie Citronella enthalten kann, um uns entweder die Anwesenheit einer Mücke zu melden oder das Untier einfach zu verscheuchen. Die Idee war, mit den vorhandenen Werkzeugen ein solch „intelligentes“ Gerät in Rekordzeit herzustellen.

Der Arduino Nicla Vision bietet umfassende technische Fähigkeiten: eine Farbkamera, ein digitales Mikrofon, eine inertielle Messeinheit (IMU), einen Dual-Core-Prozessor (Arm Cortex M7 + M4), Wi-Fi und Bluetooth Low Energy, einen I<sup>2</sup>C-Erweiterungsport, einen Anschluss für eine Batterie oder externe Stromquellen sowie diverse digitale Ein- und Ausgänge und analoge Eingänge. Für dieses Projekt nutzten wir jedoch lediglich das Mikrofon, einen der Mikrocontroller (Cortex M7) und einen digitalen Pin zur Steuerung des Luftbefeuchters. Der gesamte Code – einschließlich der Inferenz-Engine, serieller Kommunikation zur Überwachung, Steuerung der digitalen Pins und Datenerfassung



Bild 1. Zum Abschluss des Workshops über TinyML for Sustainable Development wurde der auf tinyML beruhende Luftbefeuchter anhand einer Aufzeichnung aus dem Datensatz live getestet.

über das Mikrofon – läuft auf einem einzigen Prozessor, wodurch der zweite Prozessor für weitere Aufgaben frei bleibt.

Im Proof of Concept setzten wir einen nicht standardisierten Luftbefeuchter zur Nebelerzeugung ein (**Bild 1**), mit einem Piezoelement, das von einem Oszillator mit Verstärker angeregt wird und über einen Tastschalter mit drei Stellungen aktiviert wird: aus, halbe Leistung und volle Leistung. Wir haben zwei verschiedene Möglichkeiten ausprobiert, den Luftbefeuchter anzusteuern. Zunächst haben wir einen der Piezo-Drähte durchtrennt und ihn über einen Relaiskontakt geführt. Wir erwarteten, dass dies nicht funktionieren würde, und wurden auch nicht enttäuscht: Die interne Impedanz des Relais stört die normale Funktionsweise des Piezos, und er erzeugt keinen Nebel. Die zweite Möglichkeit war, das Relais zur Betätigung des Befeuchter-Tastschalters verwenden. Wir haben uns schließlich für diese Variante entschieden, was bedeutete, dass wir das Relais zweimal anklicken mussten, um den Befeuchter zu aktivieren, und nur einmal, um ihn zu deaktivieren. Wir hatten Glück und es traten keine unerwünschten Pnelleffekte auf, so dass das System die ganze Zeit wie erwartet funktionierte.

### Methodik

Um Mücken zuverlässig zu erkennen und das Abwehrmittel zu aktivieren, entschieden wir uns, ein Deep-Learning-Modell auf dem Arduino Nicla Vision laufen zu lassen. Dieser Ansatz ermöglicht es dem Gerät, eigenständig zu „lernen“ und Mückengeräusche zu identifizieren, was eine flexiblere und anpassungsfähigere Lösung im Vergleich zu herkömmlichen algorithmischen Methoden bietet. Traditionelle Algorithmen können in

realen Umgebungen schnell zu komplex werden, während Deep-Learning-Modelle durch ihre Fähigkeit zur Anpassung an Datenvariabilität eine robustere und zuverlässigere Erkennung ermöglichen.

Die Entwicklung von tinyML hat Anwendungen im Edge-Computing revolutioniert, indem sie den Einsatz von Deep-Learning-Modellen wie gefaltete Netze (Convolutional Neural Networks, CNN) auf Mikrocontrollern mit nicht beschleunigten CPUs anstelle von leistungshungrigen GPUs ermöglicht.

Der Nicla Vision verfügt über einen Dual-Core-Mikrocontroller, der den leistungsstarken Arm Cortex M7 mit dem effizienten Arm Cortex M4-Kern kombiniert. Dieser Aufbau ermöglicht ein Gleichgewicht zwischen leistungsstarker Verarbeitung und geringer Stromaufnahme und ist damit ideal für bildverarbeitungs-basierte Anwendungen, die im Sinne der Nachhaltigkeit sowohl Rechenleistung als auch Energieeinsparung erfordern.

### Erstellen des Modells mit Edge Impulse

Edge Impulse hat Pionierarbeit im Bereich tinyML geleistet, indem es eingebetteten Entwicklern eine umfassende Plattform für Machine Learning Operations (MLOps) bietet, die den gesamten Prozess von der Sensordatenerfassung bis zur Bereitstellung von Machine-Learning-Modellen auf Mikrocontrollern und Edge-Geräten vereinfacht. Der allgemeine Prozess mit Edge Impulse sieht aus wie in **Bild 2**:

- 1. Daten sammeln:** Verwenden Sie Sensoren, um Umgebungsdaten zu erfassen, oder laden Sie vorhandene Datensätze hoch, um das Modell zu trainieren.
- 2. Daten vorverarbeiten:** Nutzen Sie die Tools von Edge Impulse, um die Daten zu kennzeichnen, vorzubehandeln und für das Training von Machine-Learning-Modellen vorzubereiten.

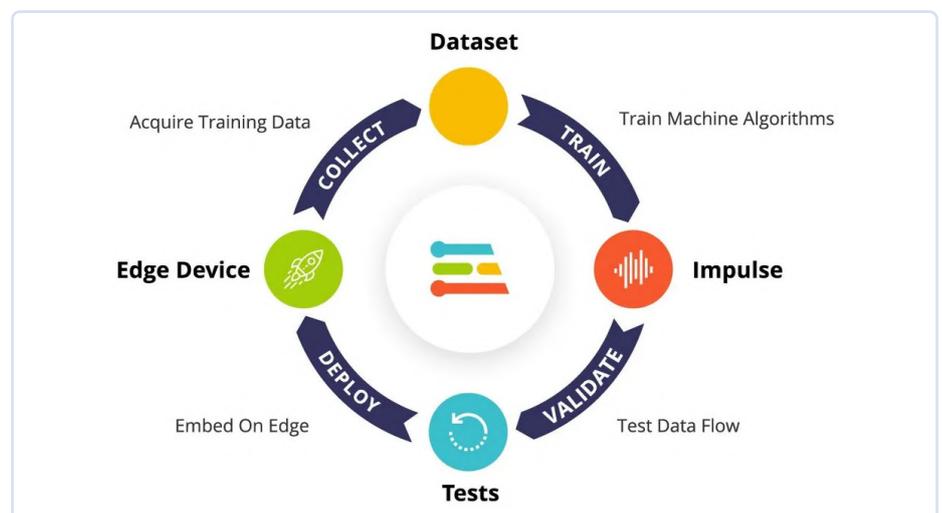


Bild 2. Der allgemeine Prozess der Modellerstellung mit Edge Impulse. (Quelle: Edge Impulse)

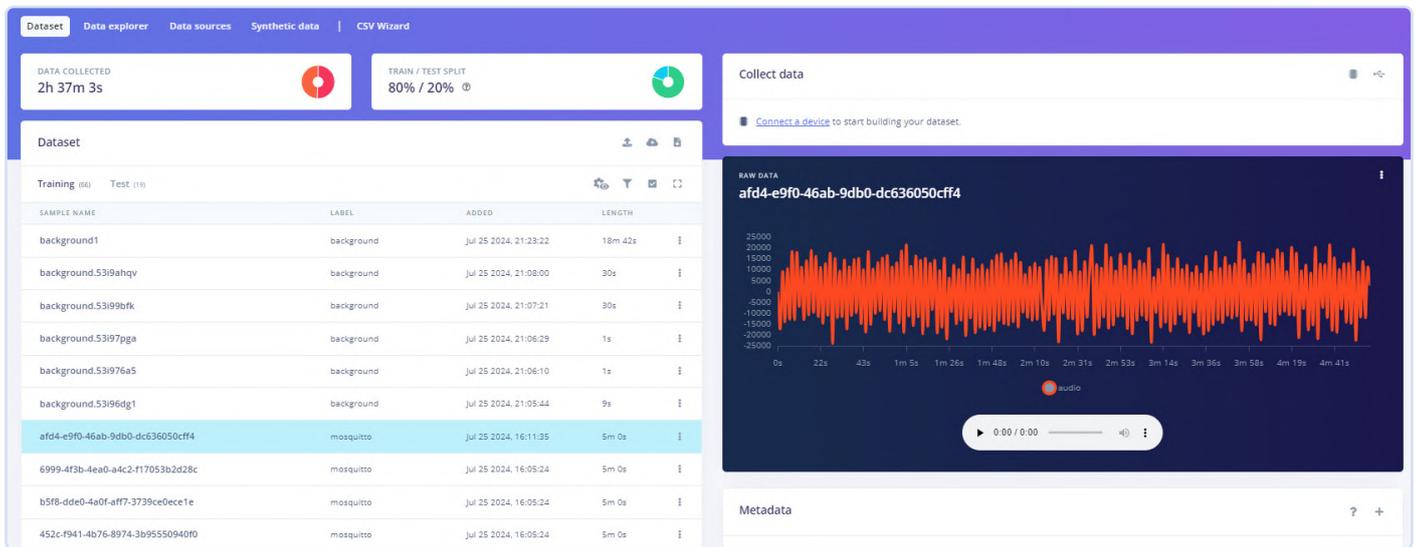


Bild 3. Arbeiten mit Online-Datensätzen und Live-Audioaufnahmen mit Edge Impulse.

3. **Modelle trainieren:** Erstellen und trainieren Sie ML-Modelle entweder aus einer Auswahl vorgefertigter Architekturen oder mit benutzerdefinierten Designs.
4. **Bereitstellung von Modellen:** Integrieren Sie die optimierten Modelle direkt auf Edge-Geräten wie dem Arduino Nicla Vision, um Echtzeit-Analysen und Aktionen durchzuführen.

Um das Konzept während des Workshops als funktionierenden Prototyp demonstrieren zu können, kombinierten wir echte Moskitogeräusch-Datensätze mit synthetisierten Audiodaten aus Online-Aufnahmen. Zusätzlich erfassten wir Live-Hintergrundgeräusche des Veranstaltungsortes direkt mit dem Arduino Nicla Vision bei einer Abtastrate von 8 kHz (Bild 3). Insgesamt sammeln wir 2:37 Stunden an Audiodaten, die Edge Impulse automatisch im Verhältnis 80 % zu 20 % in Trainings- und Testdatensätze aufteilte. Auf diese Weise nutzten wir die Fähigkeiten von Edge Impulse, mit online gefundenen Datensätzen zu arbeiten, sowie die Möglichkeit, direkt Live-Audiosamples von unterstützten Arduino-Boards wie dem Nicla Vision zu erfassen. Durch die Verwendung von Hintergrundgeräuschen des Veranstaltungsortes konnten wir das Modell speziell für unsere Testum-

**Add a processing block** ×

Did you know? You can bring your own DSP code.

DESCRIPTION	AUTHOR	RECOMMENDED
<b>Audio (MFCC)</b> <small>OFFICIALLY SUPPORTED</small> Extracts features from audio signals using Mel Frequency Cepstral Coefficients, great for human voice.	Edge Impulse	★ <span style="float: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">Add</span>
<b>Audio (MFE)</b> <small>OFFICIALLY SUPPORTED</small> Extracts a spectrogram from audio signals using Mel-filterbank energy features, great for non-voice audio.	Edge Impulse	★ <span style="float: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">Add</span>
<b>Spectrogram</b> <small>OFFICIALLY SUPPORTED</small> Extracts a spectrogram from audio or sensor data, great for non-voice audio or data with continuous frequencies.	Edge Impulse	<span style="float: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">Add</span>

Bild 4. Edge Impulse bietet vorgefertigte DSP-Blöcke für verschiedene Arten der Audioverarbeitung.

gebung optimieren. Dies war entscheidend, da die ursprünglichen Datensätze aus unterschiedlichen Innen- und Außenbereichen stammten und sich stark von den Geräuschen der Konferenzräume unterschieden. Die Fähigkeit des Modells, sich an diese neuen Bedingungen anzupassen, demonstrierte die Flexibilität des maschinellen Lernens. Für die Erkennung von Mücken entschieden wir uns für ein Klassifizierungsmodell, das eine

binäre Klassifizierung zwischen Hintergrund- und Moskitogeräuschen durchführt. Anstatt das Modell direkt auf rohe Audioeingaben zu trainieren, verwendeten wir abgeleitete Merkmale aus der digitalen Signalverarbeitung (DSP). Glücklicherweise bietet Edge Impulse vorgefertigte DSP-Blöcke, die für Aufgaben der Audioverarbeitung optimiert sind (Bild 4). In unserem Fall wählten wir das *Mel-filterbank Energy Spectrogram*, das sich hervorragend

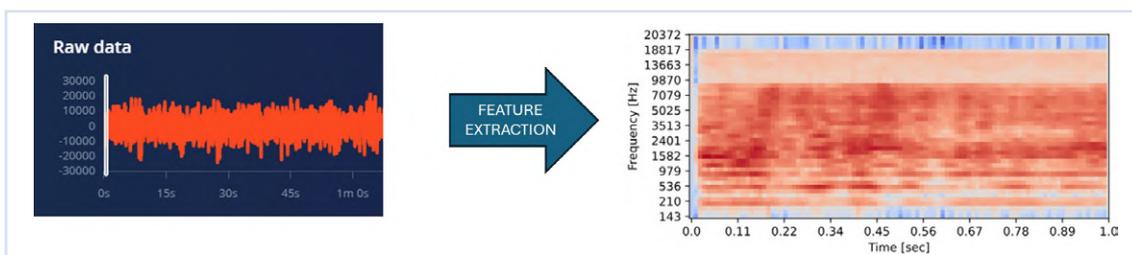


Bild 5. Das Mel-Filterbank-Energiespektrogramm wandelt Mücken-geräusche in ein 2D-Bild, wobei ein Klassifizierungsmodell ähnlich der Bildklassifizierung verwendet wird.

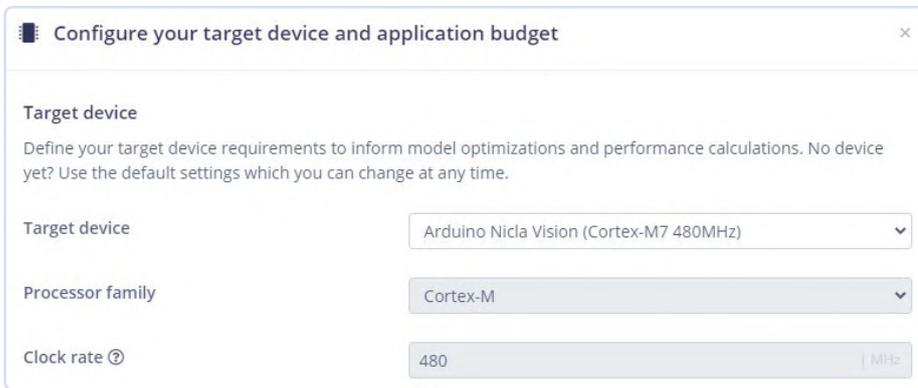


Bild 6. Einrichten der Trainings-Pipeline durch Auswahl des Cortex M7-Kerns von Nicla Vision.



Bild 7. Der visuelle Feature-Explorer zeigt gut die unterschiedenen Klassen.

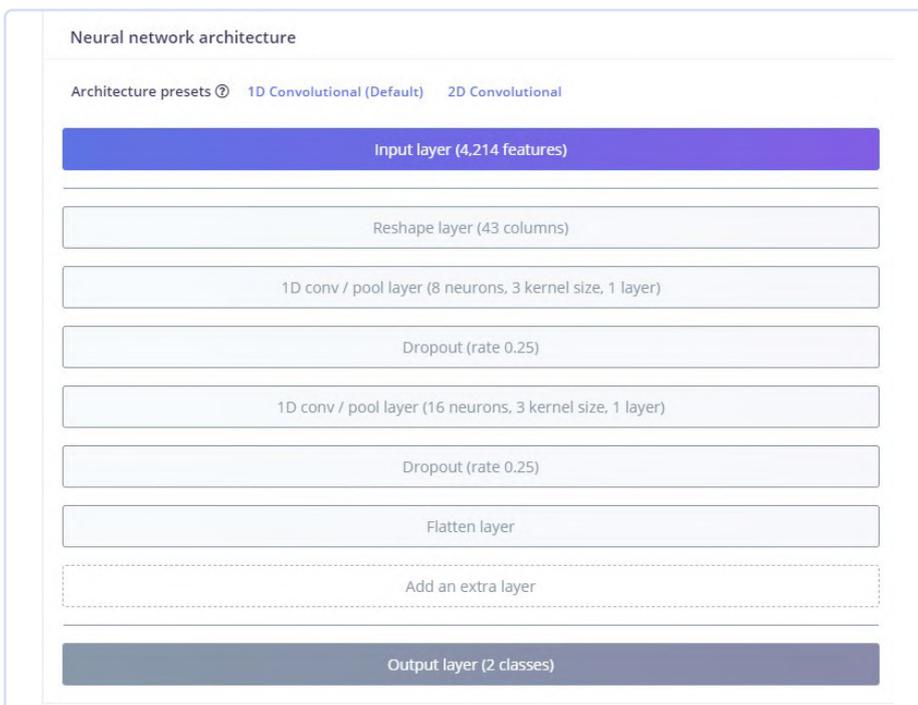


Bild 8. Architektur des verwendeten neuronalen Netzes.

zur Extraktion von Nicht-Sprachmerkmalen wie Mückengeräuschen eignet. Diese Methode wandelt das Audio in ein 2D-Bild um, ähnlich wie bei der Bildklassifizierung (**Bild 5**). Edge Impulse führt die DSP-Verarbeitung automatisch sowohl während des Trainings als auch zur Laufzeit durch.

Durch die Konfiguration des Cortex-M7-Kerns auf dem Nicla Vision (**Bild 6**) konnte Edge Impulse die erforderlichen Optimierungen vornehmen, sodass der generierte Code platzsparend in den Speicher passt und eine Echtzeit-Klassifizierung ermöglicht.

Wie in **Bild 7** zu sehen, sind die beiden Klassen (Hintergrundgeräusche und Mückengeräusche) in verschiedenen Farben dargestellt. Die Klassen sind gut voneinander getrennt und eng beieinander gruppiert. Diese klaren Cluster zeigten uns, dass das Modell bereits eine gute Trennschärfe erreicht hatte, sodass wir ohne zusätzliche Datensammlung fortfahren konnten. Wir stellten fest, dass der DSP-Code für die Ausführung auf dem Nicla Vision 181 ms benötigt. Das ist bemerkenswert schnell für eine MCU und erfüllt die Echtzeitanforderungen unseres Projekts problemlos. Auch der RAM-Bedarf während der Laufzeit lag bei lediglich 41 KB, was die Ressourcennutzung auf der Nicla Vision effizient hält. Mit diesen Ergebnissen waren wir zufrieden und bereit, die endgültige Modellauswahl und das Training abzuschließen.

Für unsere Anwendung entschieden wir uns für die Standard-Modellarchitektur von Edge Impulse, die aus zwei 1D-Faltungsschichten besteht (**Bild 8**). Diese Architektur erwies sich als ideal für die binäre Klassifizierung von Audiofeatures. Grafisch entspricht das Modell einem einfachen Feed-Forward-Netzwerk, wie in **Bild 9** dargestellt.

Edge Impulse arbeitet eng mit Arduino zusammen, um den Prozess der Modellerstellung zu vereinfachen. Dies passt perfekt zur Arduino-Philosophie, Entwicklern einen schnellen und effizienten Weg zu funktionierenden Lösungen zu bieten. Das Ergebnis war ein Modell mit einer beeindruckenden Genauigkeit von 100 % und einem F1-Score von 1,00. Es klassifizierte sowohl Hintergrund- als auch Moskitogeräusche aus dem Testset zuverlässig, selbst bei einem relativ kleinen Datensatz (**Bild 10**).

Dank der Optimierungen durch den EON-Compiler von Edge Impulse konnten wir die Ressourcenanforderungen des

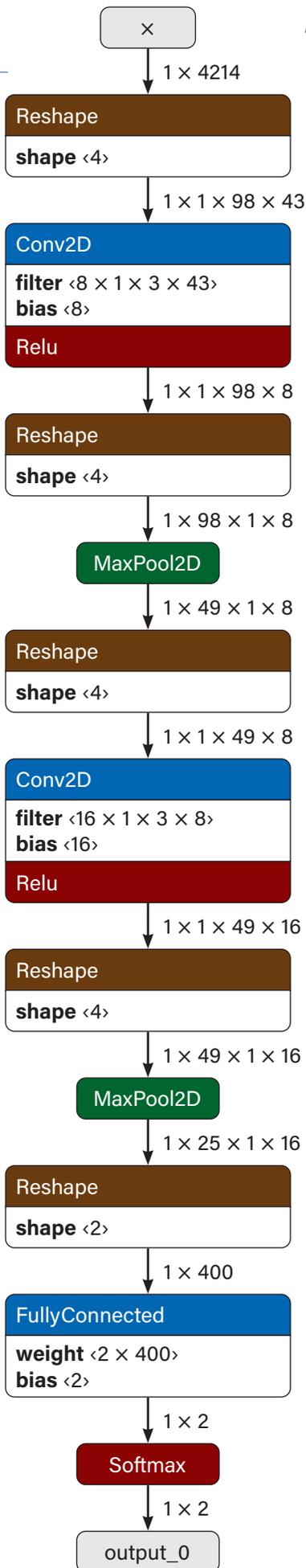


Bild 9. Diagramm des einfachen Feed-Forward-Netzes.

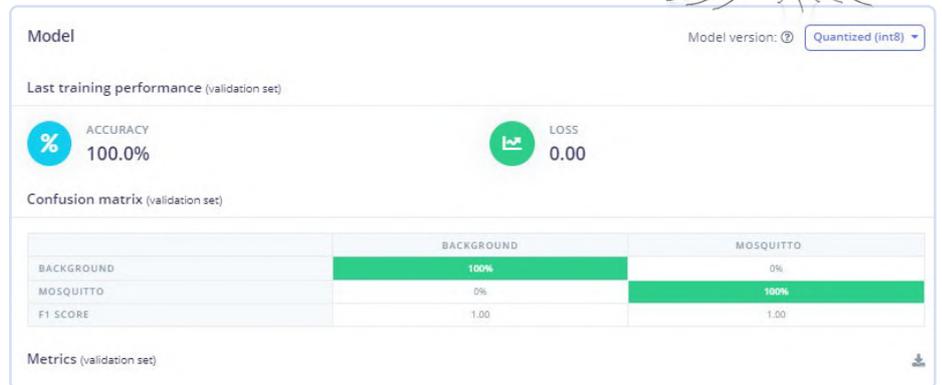


Bild 10. Das Modell erzielte eine Genauigkeit von 100 % und einen F1-Score von 1,00, indem es sowohl Hintergrund- als auch Mückengeräusche in der Testmenge erfolgreich klassifizierte.



Bild 11. Die EON-Compiler-Optimierungen von Edge Impulse haben den RAM-Bedarf zur Laufzeit des Modells reduziert.



Bild 12. Der DSP-Block wird in Echtzeit ausgeführt, mit einer Gesamtlatenz von 183 ms pro Inferenz auf dem Nicla Vision.

Modells weiter reduzieren. Der RAM-Bedarf zur Laufzeit wurde auf nur 11,1 KB gesenkt, während der ROM-Bedarf für die Speicherung der Modellgewichte und des Ausführungsgraphen lediglich 32,6 KB betrug (Bild 11). Beim Einsatz unseres Modells auf dem Arduino Nicla Vision erreichten wir eine Gesamtlatenz von 183 ms pro Inferenz (Bild 12), was mehrere Erkennungen pro Sekunde ermöglicht und noch ausreichend Zeit für den Zustandsautomaten lässt, um den Aktor zu steuern. Dies wird im nächsten Abschnitt erläutert wird. Der Nicla Vision verfügt über 2 MB Flashspeicher und 1 MB RAM, den sich die beiden Kerne teilen. Dank dieser großzügigen Ressourcen bleibt das Board während der Ausführung des Modells weit unter ihrer Leistungsgrenze. Es steht so noch reichlich Speicherplatz für zusätzliche Funktionen zur Verfügung, die in zukünftigen Erweiterungen eingesetzt werden könnten.

## Der Aufbau

Der physische Aufbau des Systems war sehr einfach: Der Arduino Nicla Vision wurde über USB mit einem Computer verbunden, sowohl für die Stromversorgung des Board als auch zur Überwachung der Ergebnisse der Inferenzmaschine. Zusätzlich verwendeten wir eine günstige Relaisplatine, die mit 3,3-V-Signalen gesteuert werden kann, und natürlich den Befeuchter. Das Flussdiagramm in Bild 13 veranschaulicht dieses Setup. Der Code für das Projekt (Download-Link [3]) wurde mithilfe von Edge Impulse erstellt. Der gesamte Prozess lässt sich wie folgt zusammenfassen:

- Datenbank-Upload: In Edge Impulse luden wir eine Mückengeräusch-Datenbank [4] hoch, einschließlich Hintergrundgeräusche, mit dem Ziel, einen binären Klassifikator zu erstellen, der die Anwesenheit von Mücken erkennt.

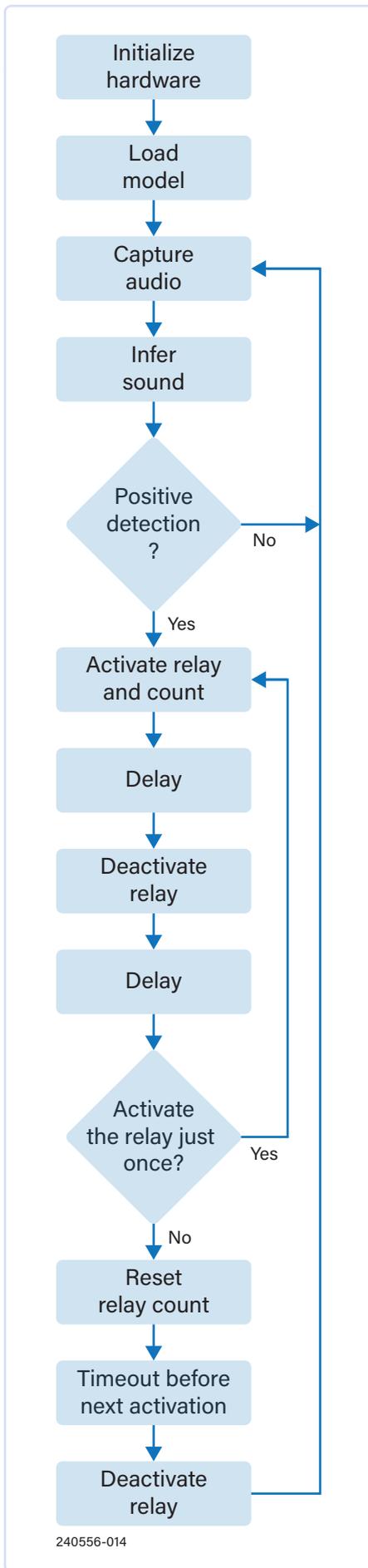


Bild 13. Flussdiagramm zur Veranschaulichung des Prozesses.

Die Audiodaten wurden mit einer Abtastrate von 44.100 Hz aufgezeichnet. Wir konfigurierten ein Abtastfenster von 1000 ms mit einer Fensterverschiebung von 500 ms.

- Vorverarbeitung: Für die Feature-Extraktion wurde der Audio-Block (MFE) von Edge Impulse verwendet. Dieser Block extrahiert Merkmale aus dem Mel-Spektrogramm, indem er Zeit- und Frequenzbereiche integriert und niedrige Frequenzen hervorhebt, die für das menschliche Ohr am besten erkennbar sind.
- Modelltraining: Zur binären Klassifikation setzten wir ein 1D-CNN (Convolutional

Neural Network) mit einer Dropout-Rate von 0,25 ein. Dieses Netzwerk ist ideal für die Verarbeitung von Audiodaten.

Die Ergebnisse unseres Modelltrainings sind unglaublich ermutigend (Bild 14). Wir haben eine Genauigkeit von 100 % erreicht, das Modell erkannte jedes einzelne Mückengeräusch im Vergleich zu den Hintergrundgeräuschen in der Validierungsmenge fehlerfrei. Der Verlustwert (loss value) ist 0,00, was darauf hindeutet, dass das Modell keine wesentlichen Fehler bei den Vorhersagen machte. Metriken wie Präzision, Recall und F1-Score liegen alle bei einem perfekten Wert von 1,00,

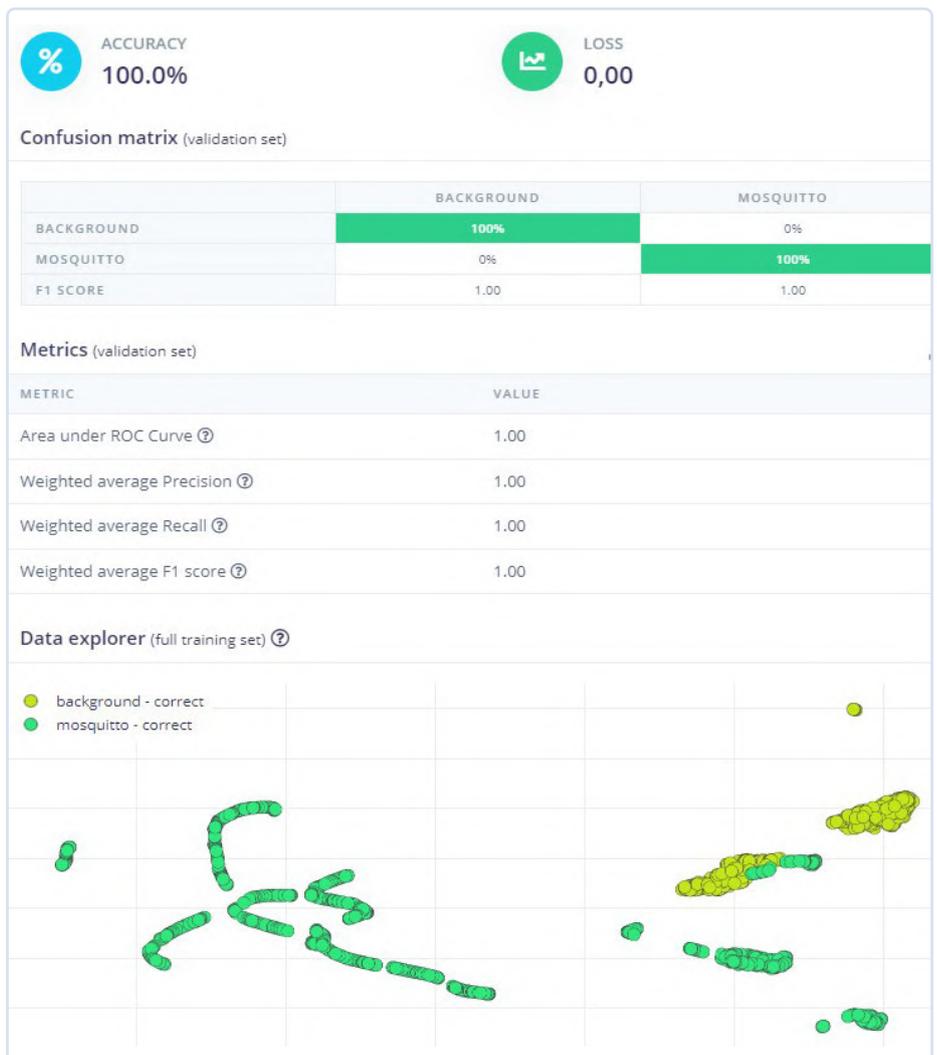


Bild 14. Das Modell erreichte eine Genauigkeit von 100 % und einen perfekten F1-Score von 1,00, mit gut getrennten Clustern für Mücken- und Hintergrundgeräusche, was es für den praktischen Einsatz bereit macht.

was zeigt, dass das Modell nicht nur genau, sondern auch durchweg zuverlässig ist. Der visuelle Datenexplorer bestätigte diese hervorragende Leistung, indem er klare, gut getrennte Cluster für Mücken- und Hintergrundgeräusche zeigte. Diese Visualisierung verdeutlicht die Fähigkeit des Modells, präzise zwischen den beiden Kategorien zu unterscheiden, was für die effektive Mückenerkennung in realen Szenarien entscheidend ist. Diese starken Ergebnisse deuten darauf hin, dass unser Modell bereit ist, in praktischen Anwendungen eingesetzt zu werden, um Mücken effizient anhand ihrer Geräuschsignaturen zu detektieren.

Weitere Informationen zur Programmierung des Arduino Nicla Vision mit Edge Impulse finden Sie auf der Docs-Website von Arduino [5]. Der von Edge Impulse generierte Code läuft als Haupt-Thread auf dem Prozessor. Alle weiteren Steuerungen richten sich nach den Entscheidungen der Inferenzmaschine. Das bedeutet, dass der Befeuchter nur dann ausgelöst wird, wenn eine positive Erkennung vorliegt. Ein Timer sorgt dafür, dass der Befeuchter nach einer bestimmten Zeit automatisch stoppt, um eine Überdosierung des Citronella-Nebels zu vermeiden. Ein zweiter Timer verhindert eine kontinuierliche Aktivierung, selbst wenn das Mückengeräusch anhält. So wird sichergestellt, dass nicht zu viel Citronella in zu kurzer Zeit freigesetzt wird. Da wir uns auf die Geräuscherkennung konzentrierten, testeten wir das System mit verschiedenen Audiodaten aus der Datenbank, die Mückengeräuschen ähnelten, und erzielten dabei durchgehend positive Erkennungsergebnisse.

## Die Zukunft

Der hier vorgestellte Proof of Concept bietet einen kostengünstigen und schnellen Ansatz zur Erkennung von Stechmücken. Der verwendete Datensatz könnte zudem erweitert werden, um die Maschine darauf zu trainieren, spezifische Mosquitoarten oder sogar das Geschlecht der Insekten zu erkennen, da nur weibliche Moskitos Krankheiten übertragen (männliche Moskitos ernähren sich nicht von Blut, sondern leben vegan).

Um die Erkennungsgenauigkeit weiter zu verbessern, wäre es sinnvoll, die Inferenzmaschine für eine feinere Klassifizierung zu optimieren. Wir laden daher die Elektor-Leser herzlich ein, dieses Projekt auszuprobieren und ihre Optimierungen mit der breiteren Community der tinyML-Entwickler und -Aficionados zu teilen! ◀

SE — 240556-02

### Sie haben Fragen oder Kommentare?

Haben Sie technische Fragen oder Kommentare zu diesem Artikel? Wenden Sie sich bitte an Elektor unter [redaktion@elektor.de](mailto:redaktion@elektor.de).



## Passende Produkte

- > **Arduino Pro Nicla Vision**  
[www.elektor.de/20152](http://www.elektor.de/20152)
- > **Elektor-Special: Gastausgabe von Arduino**  
Print: [www.elektor.de/EP-0518](http://www.elektor.de/EP-0518)  
PDF: [www.elektor.de/ED-0518](http://www.elektor.de/ED-0518)



## WEBLINKS

- [1] PAHO Report, September 2024: <https://tinyurl.com/PAHO-report>
- [2] Arduino Nicla Vision: <https://www.arduino.cc/pro/hardware-product-nicla-vision/>
- [3] Code-Download: [https://github.com/dcuartielles/ictp\\_brazil\\_24](https://github.com/dcuartielles/ictp_brazil_24)
- [4] Mosquito-Geräusch-Datenbank: <https://github.com/HumBug-Mosquito/ZooniverseData>
- [5] S. Romero, „Image Classification with Edge Impulse“, Arduino Docs, September 2024: <https://docs.arduino.cc/tutorials/nicla-vision/image-classification/>

# KI heute und morgen

## Einblicke von Espressif, Arduino und SparkFun

Zusammengestellt von der Elektor-Redaktion

Während sich die KI-Technik weiterentwickelt, wird ihre Integration in Software- und Hardwaresysteme immer komplexer. Wir haben Ingenieure von Espressif, Arduino und SparkFun - drei Unternehmen, die als Gastredaktionen bei Elektor tätig waren - gebeten, uns ihre Gedanken mitzuteilen. Wir haben die Frage gestellt: Wie stellen Sie sich die Rolle der KI bei der Weiterentwicklung von Lösungen Ihres Unternehmens vor, und sehen Sie irgendwelche Schwierigkeiten, die mit dem Fortschreiten der KI-Technologie entstehen könnten?

### Espressif

KI hat das Potenzial, unseren Entwicklungsprozess erheblich zu verbessern und Effizienz, Präzision und Innovation zu steigern. Durch die Integration von KI in unsere Arbeitsabläufe können wir verschiedene Phasen der Produktentwicklung, der Logistik und des Kundensupports rationalisieren, was zu schnelleren und zuverlässigeren Ergebnissen führt. In der Entwurfsphase kann KI repetitive Aufgaben wie Codegenerierung, Tests und Fehlersuche automatisieren, so dass sich die Ingenieure auf die komplexeren und kreativeren Aspekte des Produktentwurfs konzentrieren können. Dies beschleunigt nicht nur die Entwicklung, sondern verbessert auch die Gesamtqualität des Endprodukts. Die Fähigkeiten der KI im Bereich Simulation und Modellierung sind in der Verifizierungsphase von unschätzbarem Wert. Durch anspruchsvolle Simulationen können wir verschiedene Entwurfsszenarien testen und Hardwarekomponenten optimieren, bevor „echte“ Prototypen gebaut werden. Dadurch werden Iterationen reduziert und das Risiko teurer Entwicklungsfehler minimiert, was zu einem effizienteren Ressourceneinsatz und einer kürzeren Zeit zur Markteinführung führt.

KI kann auch die Logistik in der Lieferkette optimieren und dafür sorgen, dass Materialien und Bauteile rechtzeitig geliefert werden, um Lagerkosten zu senken und die Produktionseffizienz zu verbessern. KI kann Daten aus den Fertigungsprozessen analysieren, um Faktoren zu identifizieren, die sich auf die Chip-Ausbeute auswirken, und so Anpassungen zur Verbesserung der Produktionseffizienz und zur Kostensenkung ermöglichen.

**Anant Gupta**  
(Technical Marketing  
Manager, Espressif)



Das Internet der Dinge ist ein sehr vielfältiger Bereich mit unterschiedlichen Kundenanforderungen, was zu einer Vielzahl von Supportanfragen führt. Wir glauben, dass KI bei diesen Anfragen erheblich helfen kann. Im Bereich der natürlichen Sprachverarbeitung (NLP) können KI-gestützte Chatbots rund um die Uhr Support leisten, Kundenanfragen effizient beantworten und Probleme lösen. Durch die Nutzung von Open-Source-Dokumenten und Community-Beiträgen kann KI diese unterschiedlichen Anfragen effektiv bearbeiten, um zeitnah genaue Antworten in verschiedenen Bereichen zu geben und den Kundensupport insgesamt zu verbessern. Die Integration von KI in unsere Prozesse erfordert jedoch sorgfältige Überlegungen. Wir müssen dafür sorgen, dass KI ethisch und verantwortungsbewusst eingesetzt wird, um potenzielle Einseitigkeit zu vermeiden und den Datenschutz zu gewährleisten. Darüber hinaus müssen wir in die Aus- und Weiterbildung investieren, um unser Team mit den notwendigen Fähigkeiten auszustatten, die KI-Tools effektiv zu nutzen. Indem wir uns mit KI auseinandersetzen und die damit verbundenen Herausforderungen angehen, können wir unser Unternehmen an der Spitze der Innovation positionieren und qualitativ hochwertige Lösungen anbieten, die den sich wandelnden Bedürfnissen unserer Kunden entsprechen.



**David Cuartielles**  
(Mitbegründer von  
Arduino)

## Arduino

Arduino beteiligt sich an der Umsetzung von KI auf zwei verschiedene Arten. In erster Linie glauben wir, dass MCUs eine große Rolle in der eingebetteten KI spielen werden, da sie stromsparende, dezentrale Systeme möglich machen. MCUs bieten eine kostengünstige und stromsparende Möglichkeit zur ML-Berechnungen in kleinen Anwendungen. Eingebettete KI als Computing-Paradigma kommt dem Arduino-Verständnis eines nachhaltigeren Einsatzes von Elektronik sehr nahe. Und nicht nur das, es ist auch ein besserer Weg, um den Zugang zur Technologie selbst zu erleichtern, da sie keine Konnektivität für den Betrieb benötigt, die eine knappe Ressource an Orten ist, an denen diese Technologie am meisten helfen könnte. Auf diese Weise wollen wir einen Beitrag zu den SDGs der Vereinten Nationen leisten, indem wir eine gangbare Alternative zur künstlichen Intelligenz anbieten, die überall mit sehr geringem Strombedarf betrieben werden kann und jahrelang mit Batterien oder jahrzehntelang mit Solar-/Windenergie betrieben werden kann. Was das Training der Arduino-Boards angeht, unterstützen wir derzeit den Ansatz der TinyML-Foundation, ein Off-Board-Training durchzuführen, aber wir haben auch mit den Learning-on-Device-Bibliotheken von Fraunhofer und Ähnlichem experimentiert. Wir erwarten, dass es für das Training auf dem Gerät neue Alternativen geben wird,

da die Software in diesem Bereich immer ausgereifter wird und wir mehr und mehr Erkenntnisse darüber gewinnen, wie neuronale Netze für bestimmte Anwendungskontexte optimiert werden können.

Andererseits experimentieren wir mit KI hinter den Kulissen und nutzen sie für alltägliche Prozesse innerhalb des Hauses Arduino. Neben den üblichen Verdächtigen wie CoPilot zur Unterstützung von Programmieraufgaben oder Generativer KI zur Erstellung grundlegender Tutorials verwenden wir seit Jahren KI-gestützte Übersetzungstools. Alle unsere Inhalte befinden sich auf Github, wo wir ein Drittanbieter-Tool einbinden können, das von Menschen überwachte Übersetzungen unserer Inhalte ermöglicht. Dieses Tool erstellt mögliche Übersetzungen auf Grundlage früherer Erfahrungen, die wir bei unserer „Zusammenarbeit“ mit der AI gesammelt haben. Auf diese Weise können wir Übersetzungen anfertigen, die unserer Art, Dinge in der Elektronikwelt zu benennen, entsprechen.

Heutzutage erleben wir einen Übergang zu dieser neuen Art, Dinge intelligent zu machen. Wenn Sie intelligente Sensortechnik mit Kameras, Mikrofonen, 6-Achsen-IMUs und so weiter entwickeln möchten, bietet Arduino eine breite Palette von Boards, mit denen Sie experimentieren und Ihr ML-Modell trainieren und anschließend Systeme in kürzester Zeit einsetzen können. Vom kleinen und energiebescheidenen Nicla-Boards bis hin zu den robusten SPS können alle unsere Boards mit den Tensorflow-Tools oder dem benutzerfreundlicheren Framework Edge Impulse programmiert werden.

## SparkFun

Meiner Meinung nach - und ich glaube nicht, dass ich damit alleine dastehe - ist „KI“ derzeit der am meisten überstrapazierte Begriff der Welt. Jüngste Umfragen haben gezeigt, dass die Verbraucher vor Produkten, Dienstleistungen und so ziemlich allem zurückschrecken, was damit prahlt, mit „KI“ zu tun haben. Ist dieser Mixer wirklich KI? Muss mein Mixer tatsächlich KI sein? Wenn ich einen kleinen, niedlichen Roboter baue, der den ganzen Tag in meinem Wohnzimmer herumflitzt und dabei geschickt allen Hindernissen ausweicht, ist das dann eine beeindruckende Demonstration der künstlichen Intelligenz des Roboters? Oder sind es nur ein Dutzend Zeilen gut geschriebener Code und ein Näherungssensor?

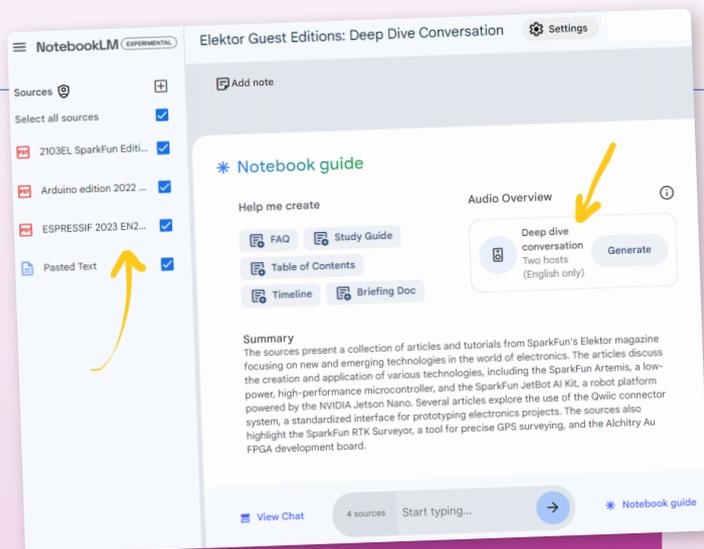
KI ist keineswegs neu - ihre Wurzeln liegen in den 1950er Jahren, als Allen Newell, J.C. Shaw und Herbert Simon das Computerprogramm *Logic Theorist* entwickelten. In letzter Zeit ist der Begriff jedoch in aller Munde, vor allem durch Grafikprogramme wie Midjourney und Textinterpreter wie von OpenAI. Wie man so schön sagt: Es hat nur 70 Jahre gedauert, bis es über Nacht zum Erfolg wurde!

Dies ist eine enorme neue Technologie, die die Art und Weise verändern wird, wie wir alle viele unserer täglichen Aufgaben erledigen. Wie immer bei einem solchen Technologiesprung gibt es eine Vielzahl von Ängsten, die auch nicht unbegründet sind. Zu den größten Befürchtungen gehört der Ansturm falscher Informationen, die Voreingenommenheit der KI, die Vernichtung von Arbeitsplätzen und natürlich die totale Übernahme der Welt, wenn das KI beschließt, dass sie die größte Bedrohung für den Menschen beseitigen muss, nämlich den Menschen. Die ersten beiden Beispiele sehen wir bereits: Falsche Informationen, die den Anschein erwecken,

**Rob Reynolds**  
(Creative Technologist,  
SparkFun Electronics)



aus einer seriösen Quelle zu stammen, sind überall im Internet zu finden, nicht nur in Schriftform, sondern auch als Audio und Video. Und auch in den KI-Modellen tauchen Beispiele für KI-Vorurteile auf, die von harmlos bis atemberaubend reichen. Was den massiven Verlust von Arbeitsplätzen anbelangt, so haben früher schon große technologische Fortschritte gezeigt, dass dies nicht der Fall ist. Was damals geschah und was wir meines Erachtens auch heute erwarten können, ist, dass sich die Arbeitsplätze verändern werden, dass sich die Anheuerung bei bestimmten Stellen ändern wird, dass aber mit den neuen Technologien auch neue Arbeitsplätze zu besetzen sind. Dies kann dazu führen, dass viele Arbeitskräfte neue Fähigkeiten erlernen oder sich an neue Aufgaben anpassen müssen. Für viele Unternehmen dürfte es sinnvoll sein, ihre derzeitigen Mitarbeiter in diesen neuen Technologien zu trainieren, als eine große Zahl neuer Mitarbeiter einzustellen. Und in Bereichen wie dem Kundendienst, der sehr wahrscheinlich stark von KI betroffen sein wird, versuchen einige Unternehmen bereits, die meisten dieser Funktionen durch Chatbots zu ersetzen. Die klügeren Unternehmen nutzen KI als Werkzeug, um die alltäglichen und repetitiven Aufgaben zu erledigen, sodass sich ihre Kundendienstmitarbeiter mehr auf die Verbesserung der Interaktion mit dem Kunden konzentrieren können. Da Berichte zeigen, dass Verbraucher der Generation Z ein Unternehmen oft nach nur einer einzigen schlechten Erfahrung verlassen, scheint diese Option der bessere Plan zu sein.



## Analyse der Gast-Ausgaben: Ein Gespräch mit der KI

In den letzten Jahren hat Elektor mit namhaften „Gastredakteuren“ zusammengearbeitet, um einzigartige Elektor-Ausgaben zu veröffentlichen: Sparkfun war eine Ausgabe im Jahr 2021 vorbehalten, Arduino eine Ausgabe im Jahr 2022 und Espressif eine Ausgabe im Jahr 2023. Während der Vorbereitung dieses Artikels luden die Elektor-Redakteure PDF-Versionen der ersten drei Elektor-Gastausgaben sowie einige zusätzliche Informationen über jede Ausgabe in Googles NotebookLM hoch und erstellten dann eine „Deep Dive Conversation“ auf Englisch. Die daraus resultierende Konversation ließ jedoch sehr zu wünschen übrig. Es scheint, dass das Programm nicht alle drei Veröffentlichungen angemessen überprüft hat. Obwohl die PDFs einige Male hochgeladen und der Vorgang wiederholt wurde, waren die daraus resultierenden Audiogespräche zwar unterhaltsam und interessant, gaben aber nicht den gesamten Inhalt der drei von Gästen herausgegebenen Ausgaben wieder. Wichtige Ideen und Inhalte fehlten entweder ganz oder wurden ungenau zusammengefasst, was zu Lücken in der Tiefe und im Gesamtwert des Gesprächs führte. Obwohl die KI einen interessanten Chat bot, der realistisch klang, fehlte ihr ein umfassendes Verständnis bestimmter technischer Details und redaktioneller Nuancen. Dies zeigte die Einschränkungen des Programms bei der genauen Bewertung von drei dicken Zeitschriften mit komplexen, spezialisierten technischen Inhalten. Trotzdem zeigte das Tool ein großes Potenzial für kreative Diskussionen, auch wenn es noch weiter verfeinert werden muss, um eine gründliche Erfassung aller in das System hochgeladenen Materialien zu gewährleisten. Wir erkennen auch an, dass wir vielleicht zusätzliche Inhalte und Kontexte hätten bereitstellen können. Wir werden weiter experimentieren!

## Die Gast-Ausgaben von Elektor

Im Jahr 2021 brachte Elektor die erste Gast-Ausgabe heraus. Gemeinsam mit unserem ersten Gast, der Firma SparkFun, haben wir eine Zeitschrift mit ausführlichen technischen Anleitungen, Elektronikprojekten mit SparkFun-Produkten, exklusiven Inhalten und Einblicken des SparkFun-Gründers, seiner Führungskräfte und seiner Ingenieure erstellt und veröffentlicht.

Nach der erfolgreichen SparkFun-Ausgabe haben wir im Jahr 2022 eine besondere Einladung an unsere Freunde von Arduino geschickt. Sie wollten die professionellen Fähigkeiten der Arduino-Plattform hervorheben, und so arbeiteten wir eng zusammen, um eine besondere Elektor-Ausgabe zu erstellen. Das Ergebnis war eine 140-seitige Ausgabe von Elektor mit Arduino als Gastredaktion, die sowohl Arduino als auch der Elektor-Community von professionellen Makern, Ingenieuren und Akademikern zugute kam.

Im Jahr 2023 haben wir mit Espressif zusammengearbeitet. In dieser Ausgabe warf Espressif einen Blick in die Zukunft des transformativen Potenzial generativer KI-Technologien wie ChatGPT, wie sie Branchen umgestalten werden und wie Espressif plant, sie für Innovationen zu nutzen - mit Erkenntnissen aus unserer Community und von Partnern. Die von unseren den Gästen herausgegebenen Zeitschriften wurden in Englisch, Deutsch, Niederländisch und Französisch veröffentlicht (gedruckt und digital) und vermarktet. Die Initiative ermöglichte es sowohl unseren Gästen, ihre Produkte auf neuen Märkten zu präsentieren und Elektor, unsere Produkte einem neuen Publikum vorzustellen. Die Ergebnisse waren bemerkenswert: Jede Ausgabe erreichte Hunderttausende von begeisterten Lesern.



240517-102

## WEBLINKS

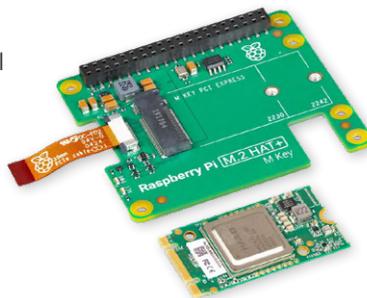
- [1] Espressif, Gast-Ausgabe Elektor 2023: <https://elektormagazine.de/elektor-espressif>
- [2] Arduino, Gast-Ausgabe Elektor 2022: <https://elektormagazine.de/elektor-arduino>
- [3] SparkFun, Gast-Ausgabe Elektor 2021: <https://elektormagazine.de/magazine/elektor-170>

# Holen Sie sich modernste KI-Hardware ins Haus!

## Raspberry Pi AI Kit

Das Raspberry Pi AI Kit enthält die Raspberry Pi M.2 HAT+ und ein Hailo-KI-Beschleunigungsmodul zur Nutzung mit dem Raspberry Pi 5. Es bietet eine zugängliche, kostengünstige und energieeffiziente Möglichkeit zur Integration leistungsstarker KI-Anwendungen. Entdecken Sie Anwendungen in den Bereichen Prozesssteuerung, Sicherheit, Heimautomatisierung und Robotik!

[www.elektor.de/20879](http://www.elektor.de/20879)



## Unitree Go2 Pro Vierbeiniger Roboter

Der Unitree Go2 ist ein vierbeiniger Roboter, der für die Forschung und Entwicklung autonomer Systeme in den Bereichen Mensch-Roboter-Interaktion (HRI), SLAM und Transport entwickelt wurde. Dank seiner vier Beine und den 12 Freiheitsgraden (12DOF) kann dieser Roboter eine Vielzahl unterschiedlicher Gelände bewältigen.

[www.elektor.de/20357](http://www.elektor.de/20357)



## BeagleY-AI SBC mit GPU, DSP und KI-Beschleunigern

BeagleY-AI ist ein kostengünstiger, quelloffener und leistungsstarker 64-Bit-Quad-Core-Single-Board-Computer, der mit einer GPU, DSP und Beschleunigern für Vision/Deep Learning ausgestattet ist und für Entwickler und Macher konzipiert wurde. Benutzer können die von BeagleBoard.org bereitgestellten Debian-Linux-Software-Images nutzen, die eine integrierte Entwicklungsumgebung enthalten.

[www.elektor.de/20991](http://www.elektor.de/20991)

## Raspberry Pi AI Camera

Die Raspberry Pi AI Camera ist ein kompaktes Kameramodul, das auf dem Sony IMX500 Intelligent Vision Sensor basiert. Der IMX500 kombiniert einen 12-MP-CMOS-Bildsensor mit integriertem Inferenzbeschleuniger für verschiedene gängige neuronale Netzwerkmodelle. Dies ermöglicht es Anwendern, anspruchsvolle KI-Anwendungen auf der Basis visueller Erkennung zu entwickeln, ohne einen separaten Beschleuniger zu benötigen.

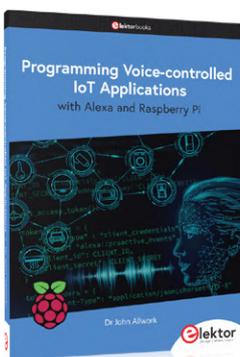
[www.elektor.de/20953](http://www.elektor.de/20953)



## Waveshare Jetson Orin Nano AI Development Kit

Dieses KI-Edge-Computing-Entwicklungskit basiert auf dem Jetson Orin Nano-Modul und bietet zahlreiche Peripherie-Schnittstellen wie M.2, DP, USB usw. Das Kit wird außerdem mit einer vorinstallierten AW-CB375NF-Wireless-Netzwerkkarte geliefert, die Bluetooth 5.0 und Dualband-WLAN unterstützt und zwei zusätzliche PCB-Antennen zur Verfügung stellt, um eine schnelle und zuverlässige drahtlose Netzwerkverbindung und Bluetooth-Kommunikation zu gewährleisten.

[www.elektor.de/20762](http://www.elektor.de/20762)



## Programming Voice-controlled IoT Applications with Alexa and Raspberry Pi

Dieses Buch ist in zwei Teile gegliedert: die Erstellung von Alexa Skills und die Gestaltung von IoT-Geräten mit einem Raspberry Pi. Es behandelt Themen wie die Entwicklung von Alexa Skills, In-Skill-Käufe und die Nutzung von AWS Lambda sowie den Bau von Smart-Home-Geräten, die von Alexa gesteuert werden. Die Leser lernen auch MQTT-Messaging, das Erstellen proaktiver Benachrichtigungen und die Umwandlung eines Raspberry Pi in ein eigenständiges Alexa-Gerät.

[www.elektor.de/20400](http://www.elektor.de/20400)

# Zeitleiste: Künstliche Intelligenz

Die Elektor-Redaktion hat mit Hilfe von ChatGPT und Gemini eine Zeitleiste zur „Geschichte der Künstlichen Intelligenz“ recherchiert und erstellt, die sich an Elektroingenieure, Elektroniker und Ingenieurstudenten richtet. Diese Zeitleiste hebt die wichtigsten Entwicklungen bei KI-Algorithmen, Hardware und Anwendungen hervor, die für diese Zielgruppen relevant sind.

**1940er-  
1950er  
Jahre:**

## Frühe Konzepte und Grundlagen

**1943** Neuronenmodell von McCulloch und Pitts: Mathematisches Modell künstlicher Neuronen, das den Grundstein für neuronale Netze legt.

**1950** Alan Turings „Computing Machinery and Intelligence“ (Rechenmaschinen und Intelligenz): Turing stellt den Turing-Test und damit einen Weg vor, um festzustellen, ob eine Maschine intelligentes Verhalten zeigen kann, das von dem eines Menschen nicht zu unterscheiden ist.

**1956** Dartmouth-Konferenz: John McCarthy und andere prägen den Begriff „Artificial Intelligence“ und stellen grundlegende Ideen für maschinelles Lernen und symbolische KI vor.

**1960er  
Jahre:**

## Die frühen Tage der KI-Forschung

**1961** Unimate: Der erste Industrieroboter namens Unimate wird vorgestellt und nimmt seine Arbeit an den Fließbändern von General Motors auf, um die Automatisierung in der Fertigung zu demonstrieren.

**1966** ELIZA: Joseph Weizenbaum entwickelt ELIZA, ein Programm, das menschliche Gespräche nachahmt und damit frühe Fähigkeiten zur Verarbeitung natürlicher Sprache demonstriert.

**1969** Shakey der Roboter: Shakey wurde in Stanford entwickelt und war der erste mobile Allzweckroboter, der seine Umgebung wahrnehmen und verstehen konnte.

**1970er  
Jahre:**

## Expertensysteme und KI-Winter

**1970** Programmiersprache Prolog: Eine logische Programmiersprache, die für die KI-Forschung, insbesondere für Expertensysteme, entscheidend war.

**1972** Humanoider Roboter WABOT-1: Die Waseda-Universität Tokio entwickelt einen humanoiden Roboter in Originalgröße und demonstriert damit Fortschritte in der Robotik und Sensortechnologie.

**1974-  
1980**

**Erster KI-Winter:** Finanzierungs-kürzungen und zu ehrgeizige Erwartungen führen zu einem Rückgang der Forschungsaktivitäten und markieren den ersten KI-Winter.



**1980er  
Jahre:**

## Aufschwung des maschinellen Lernens und der neuronalen Netze

**1980** Boom der Expertensysteme: Systeme wie XCON zur Konfiguration von Computersystemen werden kommerziell erfolgreich und fördern das Interesse an KI-Anwendungen.



**1986** ● Backpropagation-Algorithmus: David Rumelhart, Geoffrey Hinton und Ronald Williams machen die Backpropagation (Fehlerrückführung) für das Training neuronaler Netze populär.

**1990er Jahre:** ● **KI und Rechenleistung**

**1992** ● TD-Gammon: Gerald Tesauro entwickelt ein Backgammon-Spielprogramm, das mit Hilfe von Reinforcement Learning (Verstärkendes Lernen) lernt, auf hohem Niveau zu spielen.

**1997** ● Deep Blue vs. Garri Kasparov: Deep Blue von IBM besiegt den Schachweltmeister Garri Kasparow und demonstriert damit das Potenzial der KI bei der Lösung komplexer Probleme.

**2000er Jahre:** ● **KI in der Robotik und in Verbraucheranwendungen**

**2000** ● ASIMO: Honda stellt ASIMO vor, einen fortschrittlichen humanoiden Roboter, der gehen und mit Menschen interagieren kann.

**2005** ● Stanley: Das autonome Fahrzeug der Universität Stanford nutzt KI-Algorithmen zur Verarbeitung von Sensordaten, zur Routenplanung und zur Steuerung von Lenkung, Beschleunigung und Bremsen.

**2006** ● Der Begriff „Deep Learning“ wird geprägt: Geoffrey Hinton und sein Team stellen Deep-Learning-Techniken vor, die zu Fortschritten in der Computer-Vision und der Spracherkennung führen.

**2010er Jahre:** ● **KI-Boom und Deep Learning**

**2011** ● IBM Watson gewinnt Jeopardy!: Watson besiegt menschliche Champions in der US-Quizshow Jeopardy! und demonstriert damit fortschrittliche natürliche Sprachverarbeitung und Wissensabfrage.

**2012** ● AlexNet gewinnt den ImageNet-Wettbewerb: AlexNet revolutioniert die Computer-Vision, indem es den ImageNet-Wettbewerb mit faltenden neuronalen Netzwerken gewinnt.

**2015** ● OpenAI wird gegründet: OpenAI wird gegründet, um „freundliche“ KI zu fördern und zu entwickeln.

**2016** ● AlphaGo besiegt Lee Sedol: AlphaGo von DeepMind besiegt den Go-Champion Lee Sedol und demonstriert damit die Leistungsfähigkeit von tiefen verstärkenden Lernen (Deep Reinforcement Learning) und des Suchalgorithmus 'Monte Carlo Tree Search'.

**2020er Jahre:** ● **KI in Industrie und Alltag**

**2020** ● GPT-3 wird veröffentlicht: Das Sprachmodell ist in der Lage, Texte ähnlich derer von Menschen zu generieren und demonstriert damit das Potenzial von Large-Scale-Transformer-Architektur für die Verarbeitung natürlicher Sprache.



**2021** ● A100-GPU von NVIDIA: Der A100-Grafikprozessor stellt einen bedeutenden Fortschritt bei der Hardware-Beschleunigung für Deep-Learning-Anwendungen dar.

**2022** ● DALL-E und Bildsynthese: DALL-E ist in der Lage, Bilder aus Textbeschreibungen zu generieren.

**2023** ● KI-Assistenten: Die weit verbreitete Einführung von KI-gestützten Gesprächsagenten verbessert die Mensch-Computer-Interaktion in verschiedenen Anwendungen. Google Gemini wurde offiziell für den 6. Dezember angekündigt.

# BeagleY-AI

Der neuste Einplatinencomputer für KI-Anwendungen

Von Brian Tristam Williams (Elektor)

Der BeagleY-AI bietet leistungsstarke KI-Fähigkeiten in einem kompakten Open-Source-Design. Kann er zur ersten Wahl für Ihr nächstes Projekt werden? Lesen Sie weiter, um die besonderen Stärken und potenziellen Schwächen dieses vielseitigen Einplatinencomputers zu entdecken.

Mit dem BeagleY-AI wagt BeagleBoard.org einen weiteren Schritt in die Welt der Einplatinencomputer – diesmal in den dynamischen Bereich der KI und des maschinellen Lernens. Der kompakte, quelloffene BeagleY-AI ist vollgepackt mit spezialisierter Hardware und bietet Entwicklern und Forschern eine zuverlässige Plattform für KI-gesteuerte Aufgaben. Kann er die hohen Erwartungen erfüllen? Werfen wir einen genaueren Blick darauf!

## Hardware und Aufbau

Im Zentrum des BeagleY-AI steht der AM67A-Prozessor von Texas Instruments, eine Quad-Core-Cortex-



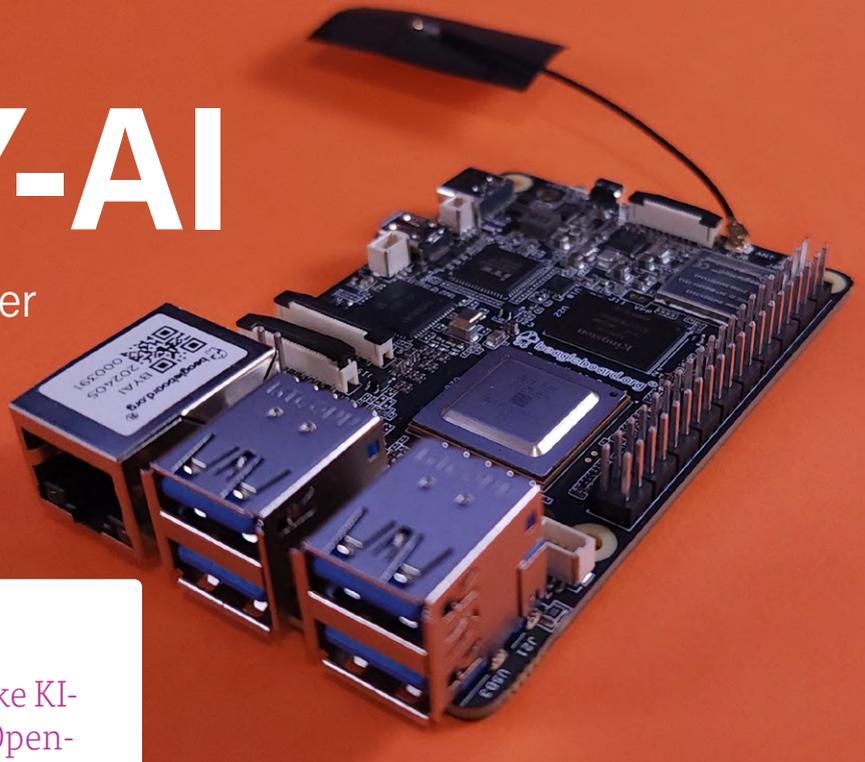
Bild 1. Draufsicht auf das BeagleY-AI-Board.

A53-CPU mit einer Taktrate von 1,4 GHz. Auch wenn die CPU im Vergleich zu anderen Einplatinencomputern mit höheren Taktraten bescheiden wirkt, ist sie speziell für industrielle und automotive Anwendungen ausgelegt. Die SoCs von Texas Instruments zeichnen sich durch grundsätzliche Zuverlässigkeit und Echtzeit-Steuerungsfunktionen aus, die von den Cortex-R5-MCU-Kernen unterstützt werden, sowie durch leistungsstarke KI-Beschleunigung dank zweier C7x-DSPs. Diese integrierten Beschleuniger sorgen dafür, dass der BeagleY-AI auch in anspruchsvollen Umgebungen starke KI-Leistung und zeitkritische Aufgaben präzise bewältigen kann.

Der wahre Wert des AM67A liegt jedoch in seiner zusätzlichen Hardwareunterstützung: Die beiden C7x-DSPs verfügen über Matrix-Multiplikationsbeschleuniger (MMA), die eine KI-Rechenleistung von bis zu vier TOPS (Tera-Operationen pro Sekunde) ermöglichen. Damit wird der BeagleY-AI zur hervorragenden Wahl für Deep-Learning-Anwendungen wie Echtzeit-Bild- und Objekterkennung.

Es ist wichtig zu beachten, dass die 4 TOPS des BeagleY-AI speziell auf 8-Bit-Operationen basieren, die üblicherweise für KI-Inferenzaufgaben verwendet werden, um Geschwindigkeit und Effizienz zu maximieren. Bei höherer Präzision, wie etwa 16-Bit- oder 32-Bit-Operationen, nimmt die TOPS-Leistung entsprechend ab.

Das Board bietet eine beeindruckende Ausstattung an drahtgebundenen und drahtlosen Anschlüssen (Bild 1): vier USB-3.0-Anschlüsse, einen USB-2.0-



Typ-C-Anschluss und einen Gigabit-Ethernet-Port. Drahtlose Kommunikation wird dank des integrierten BeagleBoard BM3301-Moduls durch Wi-Fi 6 (802.11ax) und Bluetooth 5.4 ermöglicht. Für Display-Ausgaben stehen ein Micro-HDMI-Anschluss, ein OLDI-Anschluss (LVDS) auf der Unterseite (**Bild 2**) sowie eine MIPI-DSI-Schnittstelle zur Verfügung, die gleichzeitig bis zu drei Displays steuern kann. Als Speicher dient eine SD-Karte, für die ein microSD-Kartensteckplatz vorhanden ist, was für Einplatinencomputer dieses Formfaktors üblich ist.

Ein besonderes Merkmal des BeagleY-AI ist der 16-polige FPC-Anschluss (PCIe Gen3x1), der den Anschluss von Hochgeschwindigkeitserweiterungen ermöglicht. Das Board ist lüfterlos und somit geräuscharm – ideal für Nutzer, die auf Geräuschreduktion und Zuverlässigkeit bei Langzeiteinsätzen Wert legen. Ergänzt wird die Ausstattung durch einen 40-poligen GPIO-Header, der zahlreiche Anpassungsmöglichkeiten und die Integration von Peripheriegeräten erlaubt. Der 40-polige GPIO-Header ist Raspberry-Pi-kompatibel, und das Board hat auch einen ähnlichen Formfaktor (siehe **Bild 3** und **Bild 4**). Dadurch können Nutzer das vielseitige Zubehör-Ökosystem des Raspberry Pi nutzen, darunter Gehäuse, Displays und andere Peripheriegeräte, was die Flexibilität des BeagleY-AI erheblich steigert. Darüber hinaus bietet *BeagleBoard.org* Zugang zu vollständigen Designdateien, die Anpassungen für spezifische Projektanforderungen ermöglichen.

Trotz seiner kompakten Größe ist das BeagleY-AI für den industriellen Einsatz konzipiert. Bauteile von Texas Instruments gewährleisten Robustheit und Langzeitverfügbarkeit und machen das Board zu einer praktikablen Option sowohl für die Entwicklung von Prototypen als auch für den Einsatz in Produktionsumgebungen. Der Open-Source-Charakter des Boards ermöglicht auch Modifikationen für spezifische Anforderungen, sei es die Anpassung der Hardware oder die Optimierung der Software.

### Software und Einrichtung

Der BeagleY-AI läuft auf einem Debian-basierten Betriebssystem mit der schlanken und reaktions-schnellen XFCE-Desktop-Oberfläche. Der Einstieg ist relativ einfach, insbesondere für Nutzer mit Erfahrung im Bereich der Einplatinencomputer. Die Einrichtung umfasst das Flashen des Betriebssystems auf eine microSD-Karte, die Konfiguration der Datei `sysconf.txt` und die Aktualisierung der Software über die Kommandozeile. Für Neueinsteiger stellt *BeagleBoard.org* eine umfassende Dokumentation bereit, die den Einstieg erleichtert.

Für Windows-Nutzer kann das Flashen des Images

eine zusätzliche Herausforderung darstellen, da Tools wie *Balena Etcher* oder das haus-eigene *bb-imager.exe* von *BeagleBoard.org* erforderlich sind und die Image-Datei vor der Verwendung dekomprimiert werden muss. Auf meinem Windows-System traten dabei einige kleinere Probleme auf, die sich jedoch schnell lösen ließen.

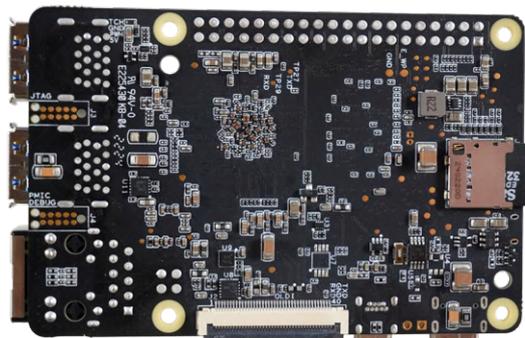


Bild 2. Das BeagleY-AI-Board, von unten besehen.



Bild 3. Der BeagleY-AI hat den gleichen Formfaktor wie ein Raspberry Pi.



Bild 4. Alle vier USB-Anschlüsse des BeagleY-AI sind USB 3.0.

Ist die Einrichtung abgeschlossen, läuft der BeagleY-AI zuverlässig, wenn auch mit ein paar kleinen Einschränkungen. Zum Beispiel sind GPU-beschleunigte Bibliotheken zwar verfügbar, aber nicht alle benötigten Komponenten, wie etwa der DSP-Compiler, lassen sich ohne weiteres installieren. Dies könnte eine kleine Hürde für Entwickler darstellen, die sofort auf die vollen KI-Funktionen des Boards zugreifen möchten. Der BeagleY-AI unterstützt eine breite Palette an Entwicklungstools, die speziell für KI-Projekte nützlich sind. So lässt sich beispielsweise TensorFlow Lite zur Ausführung von maschinellen Lernmodellen nutzen, während OpenCV ein leistungsfähiges Framework für Bildverarbeitungsaufgaben bietet. Zudem wächst die Community stetig und hat bereits zahlreiche Tutorials und Anleitungen beigesteuert – eine wertvolle Ressource für Einsteiger.

Ein weiterer Vorteil des BeagleY-AI ist das Engagement von Texas Instruments zur Unterstützung des Linux-Kernels. Mit einer Upstream-First-Strategie stellt TI alle Kernel- und Softwareänderungen für den Mainline-Kernel von Linux bereit, sobald die Hardware verfügbar ist. Dies garantiert langfristige Unterstützung und Kompatibilität mit einer Vielzahl von Software-Ökosystemen.

### Leistung und Anwendungsfälle

Der BeagleY-AI überzeugt in KI-gesteuerten Aufgaben dank seiner spezialisierten Hardware. Die Kombination aus dualen C7x-DSPs und Matrix-Multiplikationsbeschleunigern (MMA) ermöglicht eine effiziente Ausführung von Deep-Learning-Algorithmen, was das Board zu einer hervorragenden Wahl für Anwendungen in Computer Vision, Robotik und Automatisierung macht. Während des Tests bewältigte das Board Aufgaben zur Objekterkennung mit TensorFlow Lite reibungslos. Bei intensiver Nutzung erreichte die CPU-Temperatur etwa 74 °C – warm, aber auch ohne aktive Kühlung beherrschbar.

Die Fähigkeiten des BeagleY-AI gehen jedoch über KI-Anwendungen hinaus. Durch seine robusten Schnittstellen wie PCIe, USB 3.0 und Gigabit-Ethernet ist das Board auch für industrielle Aufgaben gut geeignet. Ob für den Aufbau eines intelligenten Kamerasystems, die Automatisierung einer Fertigungsstraße oder die Steuerung eines Roboterarms – die Flexibilität des Boards erlaubt eine Anpassung an spezifische Projektanforderungen. Der 40-polige GPIO-Header, der mit einer Vielzahl von Peripheriegeräten kompatibel ist, erhöht die Vielseitigkeit des Boards und ermöglicht die Integration von Sensoren, Aktoren und anderer Hardware.

Einige Einschränkungen bestehen jedoch: Die CPU-Leistung ist eher bescheiden und kann daher

keinen Desktop für allgemeine Rechenaufgaben ersetzen. Auch die 4 GB Arbeitsspeicher sind zwar für viele Anwendungsfälle ausreichend, könnten aber bei speicherintensiven Projekten zum Flaschenhals werden. Das Board unterstützt zwar PCIe-Erweiterungen, ist aber auf eine einzige Gen3-Lane begrenzt, was die Leistung einiger Peripheriegeräte einschränken könnte.

Besonders hervorzuheben ist das Potenzial des BeagleY-AI für Echtzeitanwendungen. Die Cortex-R5-Kerne ermöglichen I/O-Operationen mit sehr geringer Latenz, was in zeitkritischen Umgebungen entscheidend sein kann. Dadurch wird das Board zu einer geeigneten Option für Anwendungen beispielsweise in der Medizintechnik, wo Präzision und Zuverlässigkeit von größter Bedeutung sind.

### Community und Unterstützung

Eine der größten Stärken des BeagleY-AI ist sein Open-Source-Charakter. Alle Hardware- und Mechanik-Designdateien sind öffentlich zugänglich, sodass erfahrene Benutzer das Board an spezifische Anforderungen anpassen können. Texas Instruments unterstützt diesen offenen Ansatz, indem das Unternehmen freien Zugang zu Designdateien und Software-Ressourcen gewährt, was Entwicklern und Makern mehr Kontrolle ermöglicht. Diese Offenheit erstreckt sich auch auf das Debian-basierte Software-Ökosystem, das von einer breiten Community getragen wird. Es ist jedoch anzumerken, dass der AM67A-Prozessor selbst nicht quelloffen ist, was für einige Community-Mitglieder ein Kritikpunkt sein könnte.

Die Dokumentation von BeagleBoard.org ist umfassend und deckt alle Schritte von der Ersteinrichtung bis zu fortgeschrittenen Anwendungsfällen ab. Dennoch könnte die Softwareunterstützung verbessert werden: Ab Mitte 2024 werden einige KI-Funktionen und Tools wie das Edge-KI-Framework nicht mehr vollständig unterstützt, was die Attraktivität des Boards für fortgeschrittene Entwickler möglicherweise einschränken könnte. Obwohl die Foren und Community-Ressourcen aktiv sind, ist das Board noch relativ neu, weshalb einige Probleme noch ungelöst sind.

Die BeagleBoard-Community zählt zu den aktivsten und kooperativsten Ökosystemen in der Einplatinencomputer-Szene. Die Benutzer tauschen regelmäßig Projekte aus, bieten Unterstützung bei der Fehlerbehebung und tragen zur Entwicklung neuer Funktionen bei. Dieses große Engagement der Community ist auch ein großer Vorteil für neue Nutzer: Hier steht eine Fülle an Wissen und Erfahrung zur Verfügung. Egal, ob Sie ein erfahrener Entwickler oder ein Hobbyist sind, die Community heißt Sie willkommen und unterstützt Sie auf Ihrem Weg.

## Vorteile und Nachteile

Vorteile	Nachteile
<b>KI-Performance:</b> Die dualen C7x-DSPs und MMAs liefern bis zu 4 TOPS, was das Board ideal für Deep-Learning-Aufgaben macht.	<b>CPU-Performance:</b> Der 1,4-GHz-Quad-Core-Cortex-A53 fällt im Vergleich zu neueren SBCs etwas ab.
<b>Konnektivität:</b> USB 3.0, Gigabit-Ethernet, Wi-Fi 6 und Bluetooth 5.4 bieten vielseitige Anschlussmöglichkeiten.	<b>RAM-Beschränkungen:</b> 4 GB LPDDR4-RAM ist für einige Anwendungen möglicherweise zu knapp bemessen.
<b>Erweiterungsfähigkeit:</b> Der PCIe-Gen3x1-Anschluss und der 40-polige GPIO-Header bieten umfangreiche Anpassungsmöglichkeiten.	<b>Software-Lücken:</b> Einige KI-Funktionen und -Tools werden nicht vollständig unterstützt, was die Einsatzmöglichkeiten des Boards einschränkt.
<b>Open-Source-Hardware:</b> Benutzer können auf alle Hardware-Designdateien zugreifen und diese verändern, was Innovation und Anpassung fördert.	<b>Wärmemanagement:</b> Unter Last erhitzt sich das Board; einige Nutzer könnten hier aktive Kühlung bevorzugen.
<b>Industrietaugliche Bauteile:</b> Die Hardware von Texas Instruments gewährleistet Zuverlässigkeit und langfristigen Support, ideal für Entwicklung und Einsatz.	

## Potenzial

Das BeagleY-AI-Board bietet ein beträchtliches Wachstumspotenzial für die Zukunft. Mit der Weiterentwicklung des Software-Ökosystems ist eine breitere Unterstützung von KI-Frameworks und -Tools zu erwarten. Durch Firmware-Updates könnten neue Funktionen hinzukommen oder bestehende verbessert werden – besonders im Bereich der KI-Verarbeitung und der Integration von Peripheriegeräten. Außerdem ist zu erwarten, dass Entwickler von Drittanbietern Add-ons und Erweiterungen entwickeln, die die Funktionalität des Boards erweitern und es noch vielseitiger machen. Der Open-Source-Charakter von BeagleY-AI bedeutet, dass es sich nicht um ein statisches Produkt handelt – es ist eine Plattform, die sich weiterentwickeln kann. Nutzer, die bereit sind, sich in der Community zu engagieren und zum Projekt beizutragen, werden feststellen, dass sie die Zukunft des Boards mitgestalten können, damit es die Anforderungen einer breiten Palette von Anwendungen erfüllt. Diese Anpassungsfähigkeit ist einer der Hauptgründe, warum das BeagleY-AI für jedes KI- oder Embedded-Projekt immer eine Überlegung wert ist.

## Ist es das richtige KI-Board für Sie?

Das Board BeagleY-AI ist eine vielversprechende Wahl für Entwickler und Forscher, die einen vielseitigen, KI-fähigen Einplatinencomputer suchen. Durch sein Open-Source-Design und den leistungsstarken AM67A-Prozessor eignet es sich für verschiedenste

Anwendungen, insbesondere im Bereich KI und maschinelles Lernen. Allerdings gibt es Einschränkungen: Die CPU-Leistung ist moderat, der Arbeitsspeicher könnte für anspruchsvollere Anwendungen zu wenig sein, und einige KI-Funktionen erfordern noch weitere Softwareunterstützung.

Die Stärken des Boards – wie die KI-Beschleunigung, die umfangreiche Konnektivität und die Erweiterbarkeit – machen es zu einem wertvollen Werkzeug für alle, die bereit sind, über einige Schwächen hinwegzusehen. Mit der Weiterentwicklung des Software-Ökosystems und weiteren Beiträgen aus der Community wird das Potenzial des BeagleY-AI weiter wachsen, was es für KI-getriebene Projekte noch attraktiver machen könnte. Ob Sie moderne KI-Anwendungen erforschen oder eine zuverlässige Plattform für industrielle Automatisierung suchen, das BeagleY-AI bietet eine solide Basis für Innovation und Experimente. ◀

Übersetzt von SE — 240527-02



# im Fokus

Perspektiven aus der Elektor-Community

## Vom Elektor-Team

Sehen Sie die Zukunft der künstlichen Intelligenz durch die Augen der Elektor-Community-Mitglieder! Entdecken Sie unterschiedliche Standpunkte zu den Vor- und Nachteilen von KI und wie sie die Welt der Elektronik und die Gesellschaft im Allgemeinen verändern könnte.

Was denken Sie über das Potenzial der KI und ihre möglichen Auswirkungen?

KI hat das Potenzial, unser Leben erheblich zu verbessern, indem sie die Effizienz steigert und komplexe Probleme löst. Es gibt jedoch Bedenken hinsichtlich ihrer Auswirkungen auf das Lernen und die übermäßige Abhängigkeit. Da KI immer mehr Aufgaben automatisiert, könnten die Menschen zu abhängig werden, was zu einem Rückgang des kritischen Denkens und der Problemlösungsfähigkeiten führen könnte. Letztendlich kann die KI ein mächtiges Werkzeug sein, wenn sie klug eingesetzt wird, aber es ist wichtig, ihre Vorteile mit der Notwendigkeit in Einklang zu bringen, die menschliche Kreativität und das unabhängige Denken zu erhalten.

Saad *Imtiaz* (Elektor)

Wie hat KI Ihre Fähigkeit verbessert, technische Probleme anzugehen? Können Sie Beispiele für Erkenntnisse oder Effizienzsteigerungen nennen, die durch ihren Einsatz erzielt wurden?

Ich verwende hauptsächlich ChatGPT. Viele der technischen Probleme, auf die ich gestoßen bin, wurden bereits von anderen gelöst, und die Lösungen sind im Internet dokumentiert, aber man muss sie erst einmal finden! Bei Google kann es schwierig sein, das Gesuchte zu finden, wenn man nicht genau das richtige Stichwort verwendet. ChatGPT hingegen ist sehr gut im Umformulieren, Assoziieren von Ideen, Finden von relevanten Hinweisen und Zusammenfassen. Ich bitte ChatGPT auch, mir komplexe Themen zu erklären und zu vereinfachen, wenn ich nicht weiterkomme, oder lasse mir Programmieralgorithmen vorschlagen. Natürlich produziert die KI viele Fehler, aber es ist trotzdem ein unglaublich nützlicher und unermüdlicher virtueller Kollege!

Jean-Francois *Simon* (Elektor)

Betrachten Sie künstliche Intelligenz als ein Werkzeug, das die menschliche Kreativität in der Elektronik fördern wird, oder glauben Sie, dass sie die Innovation abwürgen könnte?

Ich glaube, es hängt alles von ihrer Verwendung ab. Zum Beispiel kann sie ein großartiges Werkzeug sein, wenn sie zur Automatisierung eingesetzt wird. Anstatt sich wiederholende und zeitraubende Aufgaben zu erledigen, können sich Ingenieure und Enthusiasten auf komplexere und kreativere Aspekte der Elektronik konzentrieren. Wenn KI jedoch nicht richtig eingesetzt wird, könnten Menschen eine starke Abhängigkeit von der KI entwickeln. Dies beeinträchtigt letztlich ihr Vertrauen in die eigene Innovationsfähigkeit.

Glaucileine *Vieira* (Elektor)

Wie sehen Sie die Rolle der KI in der Technik der Zukunft?  
Glauben Sie, dass ihr Einfluss eher transformativ oder disruptiv sein wird?

Als Algorithmen, die in der Lage sind, komplexe Probleme effizient zu lösen, scheint KI perfekt geeignet zu sein, um einige der Herausforderungen zu bewältigen, die die Technik anstrebt. Disruptive Innovationen haben oft negative soziale, ökologische und wirtschaftliche Auswirkungen. Derzeit leiden viele Bereiche des Ingenieurwesens unter einem Mangel an Ressourcen. KI könnte dazu beitragen, diesem Mangel abzuweichen und Ökologie und Ökonomie in die Gleichung einzubeziehen. Allerdings könnte sich dadurch der Bedarf an Ressourcen auf andere Bereiche verlagern.

Erwann *Berlivet* (Frankreich)

EB

In welchen Bereichen des Ingenieurwesens sind nach Ihrer Meinung menschliche Aufsicht und Entscheidungsfindung durch KI nicht zu ersetzen, und warum ist dies der Fall?

Die künstliche Intelligenz hat große Fortschritte gemacht, aber die menschliche Kontrolle ist in Bereichen wie Komplexität, Kreativität, Ethik und Sicherheit unerlässlich. Der KI fehlt es an Intuition und sie kann nicht wie der Mensch mit unvorhergesehenen Szenarien oder moralischen Dilemmata umgehen. In Situationen, in denen viel auf dem Spiel steht, sind menschliche Anpassungsfähigkeit und Urteilsvermögen gefragt. Die Verwaltung komplexer Systeme, die Auslegung von Gesetzen und das Abwägen gesellschaftlicher Auswirkungen sind Aufgaben, bei denen menschliches Verständnis nicht zu übertreffen ist. Letztlich erfordert Gestaltung für Menschen ein tiefes Verständnis, das der KI noch fehlt.

Clemens *Valens* (Elektor)

Was sind Ihre Gedanken über KI? Haben Sie schon Erfahrungen damit gemacht?

Ja, ich habe ein paar nette Erfahrungen mit KI gemacht, vor allem mit jungen und unerfahrenen Kollegen, die dazu neigen, ihr zu sehr zu vertrauen. Daher habe ich in dieser Hinsicht viel zu sagen... Lassen Sie die KI einfach das Steuerungs-/Stabilisierungssystem der Ballasttanks für Ihr neues U-Boot entwerfen und programmieren. Wie viele Stabilitätsachsen hat ein U-Boot? Erkennst du den Unterschied zwischen statischer und dynamischer Stabilität eines U-Boots? Mache dir nicht die Mühe, lass die KI das Denken übernehmen, zu viel Physik zu lernen ist heutzutage sowieso Zeitverschwendung. Wenn das U-Boot beim ersten Testtauchgang die Kontrolle verliert und sinkt, schiebst du es halt auf die KI (lacht). Wir haben das Schlossknacker-Werkzeug „Sputnik“ der KI gezeigt, und sie hat es als „Kunstinstallation“ identifiziert. Wenn etwas nicht häufig im Internet auftaucht, kann die KI es auch nicht erkennen. Es ist mir gelungen, den Sputnik-Erfinder ausfindig zu machen, einen Maschinenbauingenieur aus Split (Kroatien), den ich auch dort kennen gelernt habe. Er hat in den 1980er und 1990er Jahren mehrere Patente angemeldet und sie an deutsche Schlosserfirmen verkauft. Er erzählte mir von einigen fortschrittlichen Schlössern, die ich erst nach einer umfangreichen Suche in Patentdatenbanken finden konnte. Dann bat ich die KI, die Funktionsweise der von der deutschen Firma Winkhaus entwickelten „Fangkalotte“ zum Blockieren und Einschließen von Dietrichen zu beschreiben. Statt des deutschen Wortes „fangen“ dachte die KI, das Wort sei das englische „fang“, was ein langer, scharfer Zahn von Carnivoren wie Hunden oder Katzen ist, und gab mir dann eine unglaublich falsche Antwort. Natürlich habe ich sie gespeichert und ich kann sie jedem Interessierten zeigen. Warum kann die KI nicht ehrlich sagen: „Tut mir leid, ich habe noch nie davon gehört?“ Versucht sie, menschliche Züge zu imitieren? (lacht wieder) KI kann ein nützliches Werkzeug sein, aber nur, wenn man sie verantwortungsbewusst einsetzt, und sie kann ihren Mangel an Wissen und Erfahrung nicht verbergen, geschweige denn an Verantwortung.

Luka *Matic* (Kroatien)

Können Sie technische Szenarien nennen, in denen der Einsatz von KI die Sicherheit, Ethik oder Qualität gefährden könnte? Was macht diese Bereiche für KI ungeeignet?

Das Risiko einer Beeinträchtigung der Integrität eines Entscheidungsprozesses, an dem ein KI-System beteiligt ist, steht in direktem Verhältnis zur Eingriffstiefe, also dazu, wie viel der Mensch an das System delegiert hat. Die Steuerung eines Flugzeugs oder eine medizinische Diagnose sind zum Beispiel einige dieser Szenarien, bei denen die Bedeutung der menschlichen Entscheidungsfindung für die Sicherheit an erster Stelle stehen muss. Oder der Einsatz von KI bei der Entscheidung über die Vergabe einer Stelle in einem öffentlichen Auswahlverfahren mit Tausenden von Bewerbern könnte den Auswahlprozess intransparent, unüberprüfbar und damit unethisch machen. Im industriellen Bereich schließlich erfordert die Bewertung der Qualität eines Produkts häufig die Auswertung von Ergebnissen, die aus mehreren voneinander abhängigen Prozessen stammen, mit sehr feinen Unterschieden, die selbst die leistungsfähigsten aktuellen KI-Systeme nicht erfassen könnten.

Roberto *Armani* (Elektor)

Wie sehen Sie die Rolle der KI in der Zukunft der Elektronikentwicklung?  
Glauben Sie, dass ihr Einfluss eher transformativ oder disruptiv sein wird?

Derzeit nutzen Entwickler KI vor allem als leistungsstarke Suchmaschine oder für einfache Programmieraufgaben - im Grunde sind es die stupiden Arbeiten, bei denen sich die Menschen gerne helfen lassen. Die KI macht immer noch viele Fehler, von denen die meisten (glücklicherweise) offensichtlich sind. Bei den ersten autonom fahrenden Autos bot sich ein ähnlich miserables Bild. Bei den ersten Entwicklerwettbewerben konnte kein Fahrzeug die Ziellinie erreichen, und oft kollidierten die Autos sogar miteinander. Heute dagegen fahren autonome Fahrzeuge ungefähr so sicher wie Menschen - sie machen nur andere Fehler. Und es wird der Tag kommen, an dem jeder von uns lieber in ein autonomes Fahrzeug einsteigt als in ein Auto, das von einem menschlichen Fahrer gesteuert wird, weil der Fahrer müde oder unkonzentriert sein könnte. Wenn KI-generierte Soft- und Hardware ein Qualitätsniveau erreicht, bei dem ihr mehr Vertrauen entgegengebracht wird als einem Menschen, könnte die Elektronikindustrie in der Tat einen raschen Umbruch erleben. Aber das sind alles noch Prognosen. Wir wissen zum Beispiel noch nicht, ob KI ähnlich kreative Ideen entwickeln kann wie ein menschlicher Softwarearchitekt oder Systementwickler. Echter Erfolg erfordert ein gewisses Maß an Verrücktheit und Tatendrang. Jemand setzt sich hin und arbeitet Tag und Nacht, um ein neues Betriebssystem oder eine Grafikbibliothek für alle Arten von Mikrocontrollern zu schreiben. Wie und warum sollte KI jemals auf eine solche Idee kommen? Auf jeden Fall können wir es kaum erwarten, zu sehen, was als nächstes passiert!

*Jens Nickel (Elektor)*



Was denken Sie über die Zukunft der künstlichen Intelligenz? Glauben Sie, dass sie sich positiv oder negativ auf unser Leben und die Gesellschaft im Allgemeinen auswirken wird?

Die so genannte KI wird derzeit am häufigsten als bessere Suchmaschine eingesetzt. Der Vorteil dieser Suchfunktion ist, dass sie bestimmte Lösungsergebnisse als Antworten liefert, was die Arbeit an verschiedenen Projekten erleichtert und damit das Gesamtergebnis beschleunigen kann. Allerdings neigt die KI auch zu Halluzinationen, das heißt, sie produziert manchmal erfundene Ergebnisse, was sehr schlecht ausgehen könnte. Daher ist immer gesundes menschliches Urteilsvermögen erforderlich, um dieses Werkzeug wirksam einzusetzen. Der Hype um den Einsatz von KI ist derzeit sehr groß und befindet sich noch im Anfangsstadium. Die Zukunft wird auch zeigen, wie sich die KI weiterentwickelt und inwieweit sie vertrauenswürdige Ergebnisse liefert. Mein Wunsch wäre, dass die Entwickler der KI sie nicht sich selbst überlassen, sondern durch bestimmte Regeln einen Wildwuchs vermeiden.

*Online gestellt von schube2*

S2

Was sind Ihrer Meinung nach die wichtigsten Vorteile und potenziellen Nachteile der Verwendung von KI in technischen Projekten?

Aus meiner unmittelbaren Sicht sind die Vor- und Nachteile des Einsatzes von KI in technischen Projekten vielfältig. Wir können sie nicht einfach in Form von positiven oder negativen Auswirkungen beurteilen, denn wo und worauf beziehen wir diese Auswirkungen? Insgesamt sehe ich in den kommenden Jahren eher einen Nutzen für uns. Eine treffende Analogie für diesen Nutzen sind die Fortschritte von Computer-Hochsprachen beziehungsweise die Weiterentwicklung von Hardware-Computern.

*Paul Hetrel (Australien)*

PH

Was erwarten Sie in der Zukunft von der künstlichen Intelligenz?  
Wie wird sie unser Leben und die Gesellschaft im Allgemeinen beeinflussen?

Ich habe bereits mehrere Experimente mit der so genannten KI durchgeführt. Die Ergebnisse waren meist sehr enttäuschend, und das ist kein Wunder. Die Sprachmodelle „verstehen“ nichts so, wie es ein Mensch tun würde. Tatsächlich setzen sie nur Wörter auf der Grundlage von Statistiken zusammen. Sie produzieren dabei sehr umfangreiche Antworten, damit ihre Unfähigkeit nicht bemerkt wird. Es ist dann der Mensch, der aus den langen Antworten etwas Sinnvolles herauslesen muss. Und das ist seit den Anfängen von Doctor ELIZA von Joseph Weizenbaum der Fall. Stellen Sie einfach nachprüfbar Fragen, die nicht schon als Standardantworten im Internet zu finden sind - und Sie werden sehen, wie wenig die KI von dem „versteht“, worüber sie gerade plaudert. Zum Beispiel: „Kannst du mir eine Liste von Materialien erstellen, die eine Wärmeleitfähigkeit von 100...200 W/(m×K) haben?“

*Walter Mayer (Deutschland)*

WM

Was reizt Sie am meisten am Potenzial der KI, insbesondere wenn Sie daran denken, wie sie sich auf die Elektronikbranche auswirken könnte?

Was mich an der KI in der Elektronik am meisten begeistert, ist ihr Potenzial, ein wichtiger Faktor in Sachen Effizienz, Automatisierung und Nachhaltigkeit zu werden. Mit Hilfe der KI können wir komplexe Schaltungen entwerfen, Ausfälle vorhersagen und selbstheilende Systeme ermöglichen, was effizientere, zuverlässigere und intelligentere Elektronik bedeutet. Darüber hinaus kann die KI-gestützte Automatisierung die Produktion beschleunigen und die Genauigkeit verbessern, während nachhaltige Innovationen uns helfen, Abfall zu reduzieren. All diese Fortschritte tragen zur Entwicklung intelligenter, umweltfreundlicher Elektronik bei, die sich problemlos in unser tägliches Leben integrieren lässt.

*Alina Neacsu (Elektor)*



Glauben Sie, dass KI die Zukunft der Elektronikindustrie positiv beeinflussen wird?

Ich denke, die Zukunft der KI wird in einigen Bereichen wahrscheinlich eine Kombination aus positiven Ergebnissen und potenziellen Herausforderungen bringen. Auf der positiven Seite kann die KI für mehr Effizienz, revolutionäre Fortschritte und schnelle Lösungen für schwierige Probleme in verschiedenen Branchen sorgen. Sie könnte jedoch auch den Arbeitsmarkt durcheinanderbringen, ethische Bedenken aufwerfen und unsere übermäßige Abhängigkeit von der Technik verstärken. Die Art und Weise, wie wir die potenziellen Fallstricke umgehen und die KI verantwortungsvoll nutzen, wird entscheidend dafür sein, ob ihre Auswirkungen insgesamt positiv oder negativ sind.

*C. J. Abate (Elektor)*



Was denken Sie über die Zukunft der KI? Wird sie sich positiv oder negativ auf unser Leben und die Gesellschaft im Allgemeinen auswirken?

Wenn ich mir die Geschichte der Menschheit anschau, kann ich nur zu dem Schluss kommen, dass die KI auf beide Arten genutzt werden wird. Wahrscheinlich wird sie das Gesundheitswesen verändern, die Effizienz verbessern und so weiter. Und wie immer werden einige Menschen sie zum Schlechten nutzen, zum Beispiel als Waffe. Höchstwahrscheinlich werden Maschinen in naher Zukunft unsere Arbeit besser erledigen (ich schätze für 40...50 % der derzeitigen Aufgaben). Die Vorstellung, dass die Gesetzgebung uns vor Missbrauch schützen wird, ist eine Illusion. Verkehrs- und andere Gesetze sind auch dazu da, uns zu schützen, aber wie viele Menschen kümmern sich wirklich darum, und warum sollte es dieses Mal anders sein? Als ich jung war, hielten die meisten von uns bei Gelb an. Jetzt fahren viele Leute weiter, auch wenn die Ampel schon auf Rot steht.

*Online übermittelt von Erno Gilissen*

EG

Glauben Sie, dass künstliche Intelligenz die traditionelle technische Praxis verbessern oder ersetzen wird? Wie wird sich KI Ihrer Meinung nach auf die Berufsrollen und Qualifikationsanforderungen in Bereich der Elektronik auswirken?

Ich persönlich habe KI benutzt, um Code zu schreiben und einige logische Probleme zu lösen, aber manchmal gelingt es ihr nicht, die Fragen richtig zu analysieren, und sie braucht Prompts, um ihr zu helfen. Sie liefert keine vollständige Lösung für meine Programmierprobleme, aber sie hilft und spart eine Menge Tipparbeit. Es ist ein nützliches Werkzeug, das meiner Meinung nach die traditionellen Rollen nicht ersetzen, sondern unterstützen wird. Im Moment müssen die Antworten der KI aber noch überprüft werden. Ich werde sie auf jeden Fall weiter verwenden.

*John Allwork (Großbritannien)*



Glauben Sie, dass die künstliche Intelligenz zu mehr Effizienz bei technischen Projekten führen wird, oder könnte sie neue Schwierigkeiten mit sich bringen?

Die Technik betrifft die Sicherheit der Menschen. KI hat keine Ahnung von der Materie und man kann ihr nicht zutrauen, technische Entscheidungen zu treffen. Jede offensichtliche Produktivitätssteigerung durch den Einsatz von KI würde durch die Notwendigkeit der Überwachung durch den Menschen, der die Verantwortung übernehmen muss, wieder aufgehoben. Es besteht die Gefahr, dass die Menschen der KI Kompetenz zutrauen.

*Andrew Pratt (Großbritannien)*

AP

Welche Folgen kann es haben, wenn man sich bei der technischen Planung und Analyse zu sehr auf KI verlässt?

KI umfasst viele verschiedene Techniken, von denen einige (wie ChatGPT) keinen Begriff von Wahrheit haben. Ich persönlich habe bisher nur ChatGPT verwendet, und das auch nur bei mathematischen Fragen. Ich war sehr oft enttäuscht, da die Antworten fast immer Fehler enthielten. Ich befürchte, dass ein unangemessener Einsatz zu fehlerhaften Lösungen führen wird. Kann sich die KI technologische Sprünge vorstellen? Besteht die Gefahr eines Rückgangs der Fähigkeiten? PS. Diese Antwort wurde ohne die Hilfe von ChatGPT erstellt!  
*Pierre Molinaro (Frankreich)*

PM

Was denken Sie über die Zukunft der künstlichen Intelligenz? Wird sie sich positiv oder negativ auf unser Leben und die Gesellschaft im Allgemeinen auswirken?

Der aktuelle Stand der Technik ist nichts anderes als eine Suchblase, die kurz vor dem Platzen steht. Die Fähigkeiten dieser KI-Suchmaschinen haben sich im Vergleich zu früher definitiv weiterentwickelt, aber man darf sich nicht täuschen lassen, dass es sich dabei nur um komplexe Algorithmen mit paralleler Server-Rechenleistung handelt. Dies ist an sich eine beeindruckende Leistung, aber die Anwendung ist unglücklicherweise in erster Linie der Diebstahl von urheberrechtlich geschütztem Material, um eine Datenbank zu erstellen, die den Anschein von Intelligenz erweckt. Es handelt sich jedoch nicht um echte Intelligenz, denn es fehlt das tatsächliche Bewusstsein und die kognitive Fähigkeit, Konzepte zu verstehen, die über einfache Ideen und philosophische Grundsätze hinausgehen. Dies liegt weit jenseits von Datenbanken.  
*Online übermittelt von David Fahrenholz*

DF

Was denken Sie über die Zukunft der KI und ihre möglichen Vor- und Nachteile?

Die Zukunft der KI ist ein zweischneidiges Schwert. Auf der einen Seite hat sie das Potenzial, das Gesundheitswesen zu revolutionieren, die Produktivität zu steigern und komplexe Probleme wie den Klimawandel zu lösen. Andererseits könnte sie zur Verdrängung von Arbeitsplätzen führen, die Ungleichheit verschärfen und ernsthafte ethische Bedenken aufwerfen, insbesondere bei Stichworten wie Überwachung und Voreingenommenheit. Entscheidend wird sein, wie wir Innovation und Verantwortung in Einklang bringen und dafür sorgen, dass KI so entwickelt und eingesetzt wird, dass sie der Gesellschaft als Ganzes zugute kommt, ohne dabei die potenziellen Risiken aus den Augen zu verlieren.  
*Brian Tristram Williams (Elektor)*



RG -- 240467-02

### Tauchen Sie ein in die KI!

Besuchen Sie die Seite Embedded & KI von Elektor für mehr Projekte, Videos und Tutorials!  
[www.elektormagazine.de/embedded-ki](http://www.elektormagazine.de/embedded-ki)



Sieh dir die Antworten auf die Fragen der Elektor-Community-Mitglieder an. Gibt es irgendwelche Schlüsse, die du darüber ziehen kannst, was die Community insgesamt über KI denkt?

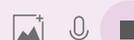


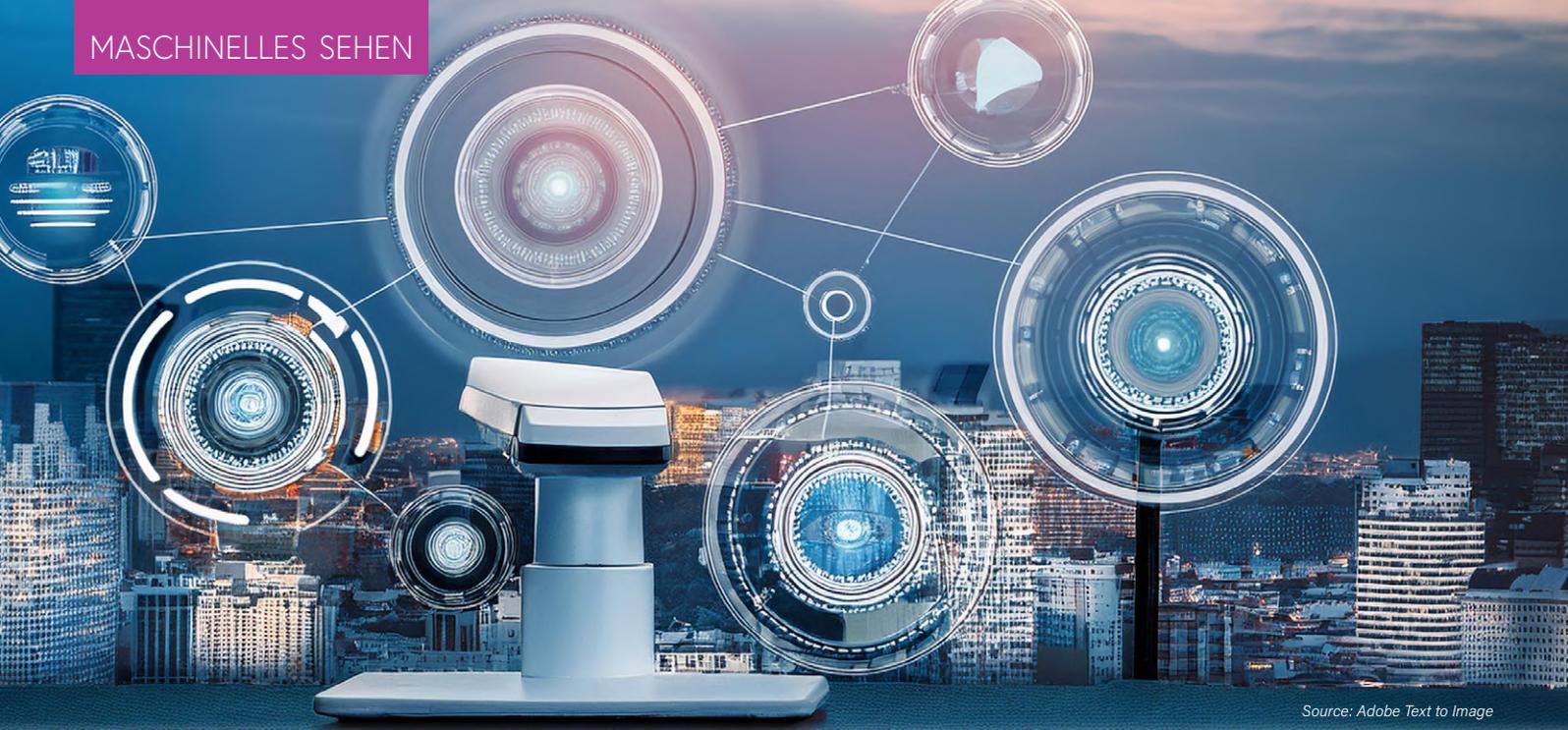
Drafts



Die Mitglieder der Elektor-Community bringen eine Mischung aus Optimismus und Besorgnis über die Zukunft der KI zum Ausdruck. Während viele die potenziellen Vorteile wie erhöhte Effizienz und Problemlösungsfähigkeiten anerkennen, gibt es auch erhebliche Bedenken hinsichtlich der möglichen Nachteile... Zusammenfassend lässt sich sagen, dass die Elektor-Community KI als ein leistungsfähiges Werkzeug betrachtet, das sowohl potenzielle Vorteile als auch Risiken birgt. Man ist zwar optimistisch, dass die KI die technische Praxis verbessern kann, betont aber auch die Notwendigkeit einer verantwortungsvollen Entwicklung und menschlichen Aufsicht, um positive Auswirkungen auf die Gesellschaft zu gewährleisten.

Enter a prompt for ChatGPT





Source: Adobe Text to Image

# Maschinelles Sehen mit OpenMV

## Bau eines Limonadendosen-Detektors

Von Koen Vervloesem (Belgien)

Für die Entwicklung von Bildverarbeitungsanwendungen ist OpenMV eine interessante Plattform mit dem Ziel, der „Arduino des maschinellen Sehens“ zu werden. Sie kombiniert MicroPython-Firmware mit einfach zu nutzenden Softwarebibliotheken für die industrielle Bildverarbeitung, einer eigenen integrierten Entwicklungsumgebung und spezieller Kamerahardware. In diesem Tutorial verwenden wir OpenMV und das Kamera-Board *Arduino Pro Nicla Vision*, um Objekte wie Getränkedosen zu erkennen.

In diesem Artikel verwende ich das Kamera-Board *Arduino Pro Nicla Vision* [1] (**Bild 1**) mit der OpenMV-Firmware für die Objekterkennung [2]. Dieses Projekt ist nicht auf dieses Board angewiesen; Sie können es auf jeder OpenMV-Kamerahardware ausführen. Die Objekterkennung ist eine spezielle Art der Bildklassifizierung. Während ein Bildklassifizierungsmodell lediglich die Klasse des Objekts im Bild angibt, zeigt ein Objekterkennungsmodell auch den Ort des erkannten Objekts an. Dies funktioniert sogar bei mehreren Objekten in einem Bild. Dieser Prozess ist natürlich rechenintensiver als die einfache Bildklassifizierung, aber *Nicla Vision* kann ihn bewältigen.

Ich setze den Algorithmus FOMO (Faster Object, More Objects) ein, um ein Objekterkennungsmodell zu erstellen. FOMO ist äußerst effizient und ermöglicht die Objekterkennung in Echtzeit auch auf Mikrocontrollern. Im ersten Schritt erstellen Sie einen Datensatz von Bildern (auch mit der Kamera des *Nicla Vision*) der Objekte, die Sie erkennen möchten.

Sobald Sie diesen Datensatz erstellt haben, können Sie das FOMO-Modell mit Edge Impulse [3] trainieren. Das trainierte Modell wird als benutzerdefinierte OpenMV-Firmware auf das *Nicla-Vision-Board* zur Objekterkennung hochgeladen. OpenMV basiert auf MicroPython, so dass Sie auch als Nicht-Programmierer spezialist die Aktionen, die ausgeführt werden, wenn das Board ein Objekt erkennt, ganz einfach anpassen können.



*Bild 1. Mit ihren Abmessungen von 22,86 mm × 22,86 mm × 7,26 mm ist *Arduino Nicla Vision* die kleinste OpenMV-Kamera auf dem Markt.*

## Einrichten der OpenMV-IDE

Installieren Sie zunächst die OpenMV-IDE [5], die unter Windows, macOS und Linux (Ubuntu) läuft. Unter Ubuntu ab 23.04 müssen Sie eine Zeile im `setup.sh`-Skript des `tar.gz`-Installers ändern, da Pakete, die außerhalb einer virtuellen Python-Umgebung installiert werden, nicht mehr unterstützt werden. Ändern Sie die Zeile `sudo pip install pyusb` in `sudo apt install python3-usb` und führen Sie dann das Skript `setup.sh` im entpackten Verzeichnis aus.

Nach der Installation der OpenMV-IDE schließen Sie Ihr Nicla-Vision-Board mit einem Micro-USB-Kabel an Ihren Computer an. Wenn die LED auf dem Board blau zu blinken beginnt, wird das Standard-Skript `main.py` von OpenMV ausgeführt. Diese `main.py`-Datei mit dem MicroPython-Code in **Listing 1** ist im interne Speicher des Boards enthalten, so dass die LED blink, auch wenn keine USB-Verbindung zum Computer besteht. Steckt man das USB-Kabel in den Computer, wird dieser interne Speicher wie ein gewöhnlicher USB-Stick als Massenspeicher eingebunden.

## Verbinden mit Ihrem OpenMV-Board

Öffnen Sie die OpenMV-IDE und klicken Sie auf das `Connect`-Symbol unten links. Die LED sollte aufhören zu blinken, und Sie erhalten möglicherweise eine Meldung, dass die Firmware des Boards veraltet ist. Bestätigen Sie, dass Sie die Firmware aktualisieren und auch das interne Dateisystem löschen möchten. Dieser Vorgang dauert eine Weile und führt dazu, dass die LED des Boards eine Zeit lang grün blinkt. Sobald die Firmware aktualisiert ist und die OpenMV-IDE die Verbindung zu Ihrem Board wieder aufgebaut hat, verlischt die LED. Die



### Listing 1: Blinken der eingebauten LED zur Funktionsprüfung der OpenMV-Hardware

```
# main.py -- put your code here!
import pyb, time
led = pyb.LED(3)
usb = pyb.USB_VCP()
while (usb.isconnected()==False):
    led.on()
    time.sleep_ms(150)
    led.off()
    time.sleep_ms(100)
    led.on()
    time.sleep_ms(150)
    led.off()
    time.sleep_ms(600)
```

Statusleiste am unteren Rand sollte den Namen des Boards, den Sensor, die Firmware-Version, den seriellen Anschluss und das Laufwerk des internen Speichers anzeigen (**Bild 2**).

Um zu überprüfen, ob die Kamera funktioniert, wird das vom Editor angezeigte Standardskript `helloworld_1.py` gestartet, indem man auf das grüne `Run`-Symbol unten links klickt. Dadurch wird die Kamera in eine Webcam verwandelt und ihr Feed wird in der OpenMV-IDE angezeigt. Der Teil `Frame Buffer` des Fensters präsentiert das Live-Kamerabild, während der `Histogramm`-Bereich die Rot-, Grün-

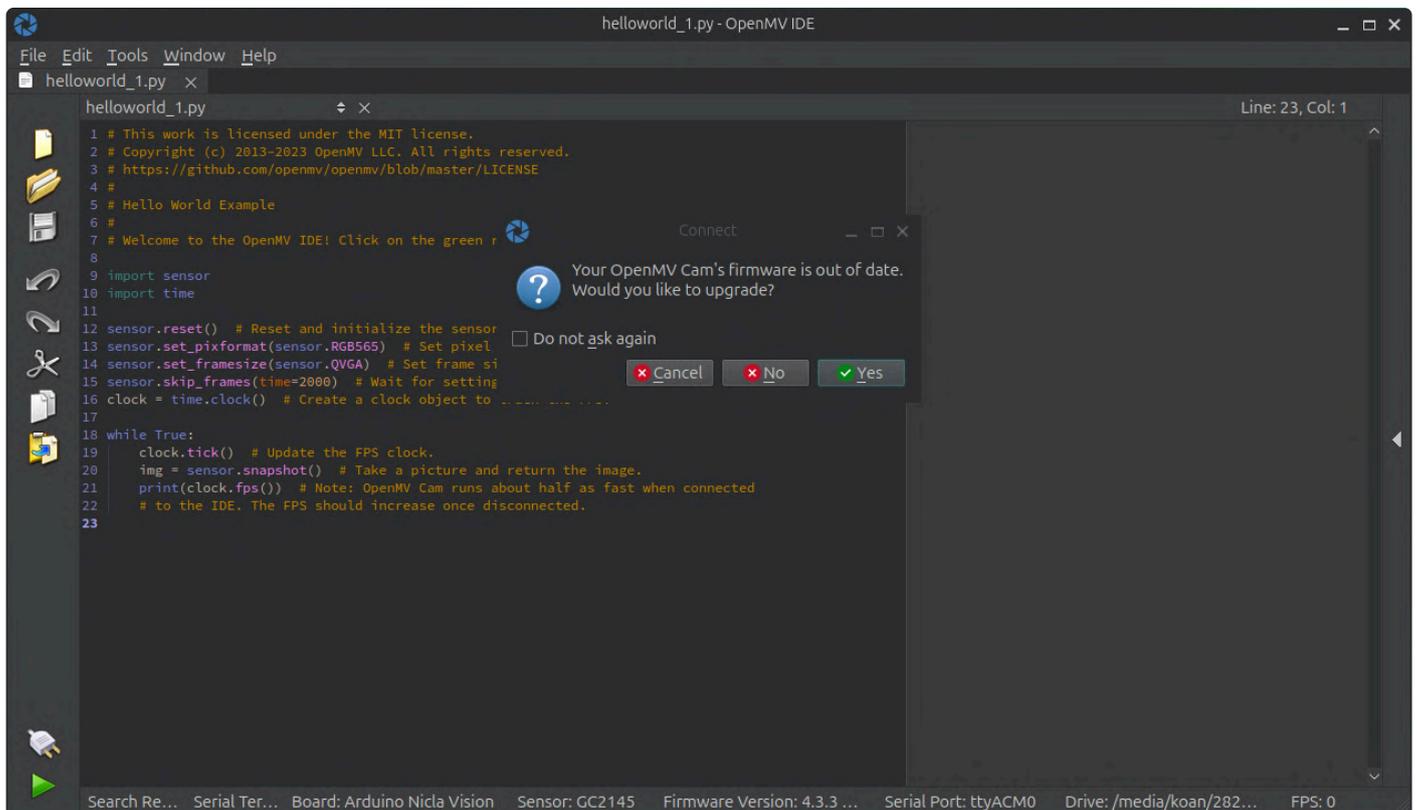


Bild 2. OpenMV-IDE fordert zur Aktualisierung der Firmware des Kamera-Boards auf.



## Listing 2: Live-Ansicht der Kamera des OpenMV-Boards

```
# This work is licensed under the MIT license.
# Copyright (c) 2013-2023 OpenMV LLC. All rights reserved.
# https://github.com/openmv/openmv/blob/master/LICENSE
#
# Hello World Example
#
# Welcome to the OpenMV IDE! Click on the green run arrow button below to run the script!

import sensor
import time

sensor.reset() # Reset and initialize the sensor.
sensor.set_pixformat(sensor.RGB565) # Set pixel format to RGB565 (or GRAYSCALE)
sensor.set_framesize(sensor.QVGA) # Set frame size to QVGA (320x240)
sensor.skip_frames(time=2000) # Wait for settings take effect.
clock = time.clock() # Create a clock object to track the FPS.

while True:
    clock.tick() # Update the FPS clock.
    img = sensor.snapshot() # Take a picture and return the image.
    print(clock.fps()) # Note: OpenMV Cam runs about half as fast when connected
    # to the IDE. The FPS should increase once disconnected.
```



## Listing 3: Das Skript zur Erfassung von Datensätzen speichert RGB-Kamerabilder mit 240 x 240 Pixeln

```
import sensor, image, time

sensor.reset()
sensor.set_pixformat(sensor.RGB565)
sensor.set_framesize(sensor.QVGA)
sensor.set_windowing((240, 240))
sensor.skip_frames(time = 2000)

clock = time.clock()

while(True):
    clock.tick()
    img = sensor.snapshot()
    print(clock.fps())
```

und Blau-Komponenten des Farbraums anzeigt. Sie können diesen Farbraum von RGB in Graustufen, LAB oder YUV ändern. Darüber hinaus wird die Anzahl der Bilder pro Sekunde (Framerate), welche die Kamera überträgt, unten rechts angezeigt.

Werfen Sie einen Blick auf das Standard-Skript *helloworld\_1.py*

(**Listing 2**). Nach dem Import der erforderlichen Python-Module setzt das Skript die Kamera zurück und stellt dann das Pixelformat und die Bildgröße ein. In einem nächsten Schritt wird eine Wartezeit von zwei Sekunden eingelegt, bevor eine Schleife gestartet wird, in der kontinuierlich Bilder aufgenommen sowie die Framerate erfasst werden.

### Erstellen Sie Ihren Datensatz

Sobald Ihre Kamera betriebsbereit ist, nehmen Sie Bilder von verschiedenen Objekten auf und erstellen Sie damit Ihren Datensatz. Dabei werden Sie von der OpenMV-IDE unterstützt: Öffnen Sie das Menü *Tools* → *Dataset Editor* → *New Dataset* und wählen Sie ein Verzeichnis zum Speichern Ihres Datensatzes. Die IDE zeigt dann links den Datensatz-Editor, in der Mitte das Skript zur Erfassung des Datensatzes (das mit dem vorherigen Hello-World-Skript identisch ist) und rechts den Bildpuffer und das Histogramm. Klicken Sie auf das Icon *New Class Folder* in der linken Seitenleiste oder drücken Sie *Strg+Shift+N*, um eine neue Klasse für eine Art von Objekten zu erstellen, die Sie erkennen möchten.

In diesem Beispiel soll das Modell so trainiert werden, dass es Limonadendosen erkennt. Zunächst wird eine Klasse namens *can* erzeugt. Die Kamera besitzt eine Auflösung von 320 x 240 Pixel, aber das FOMO-Modell erwartet quadratische Bilder, so dass Sie das Skript zur Datensatzerfassung entsprechend anpassen müssen (**Listing 3**). Starten Sie dieses Skript, indem Sie auf das *Play*-Icon klicken, wählen Sie eine der Klassen aus, indem Sie auf den Ordernamen im Dataset-Editor klicken, und richten Sie die Kamera auf ein Objekt dieser Klasse. Nehmen Sie ein Bild auf, indem Sie auf das Kamerasymbol

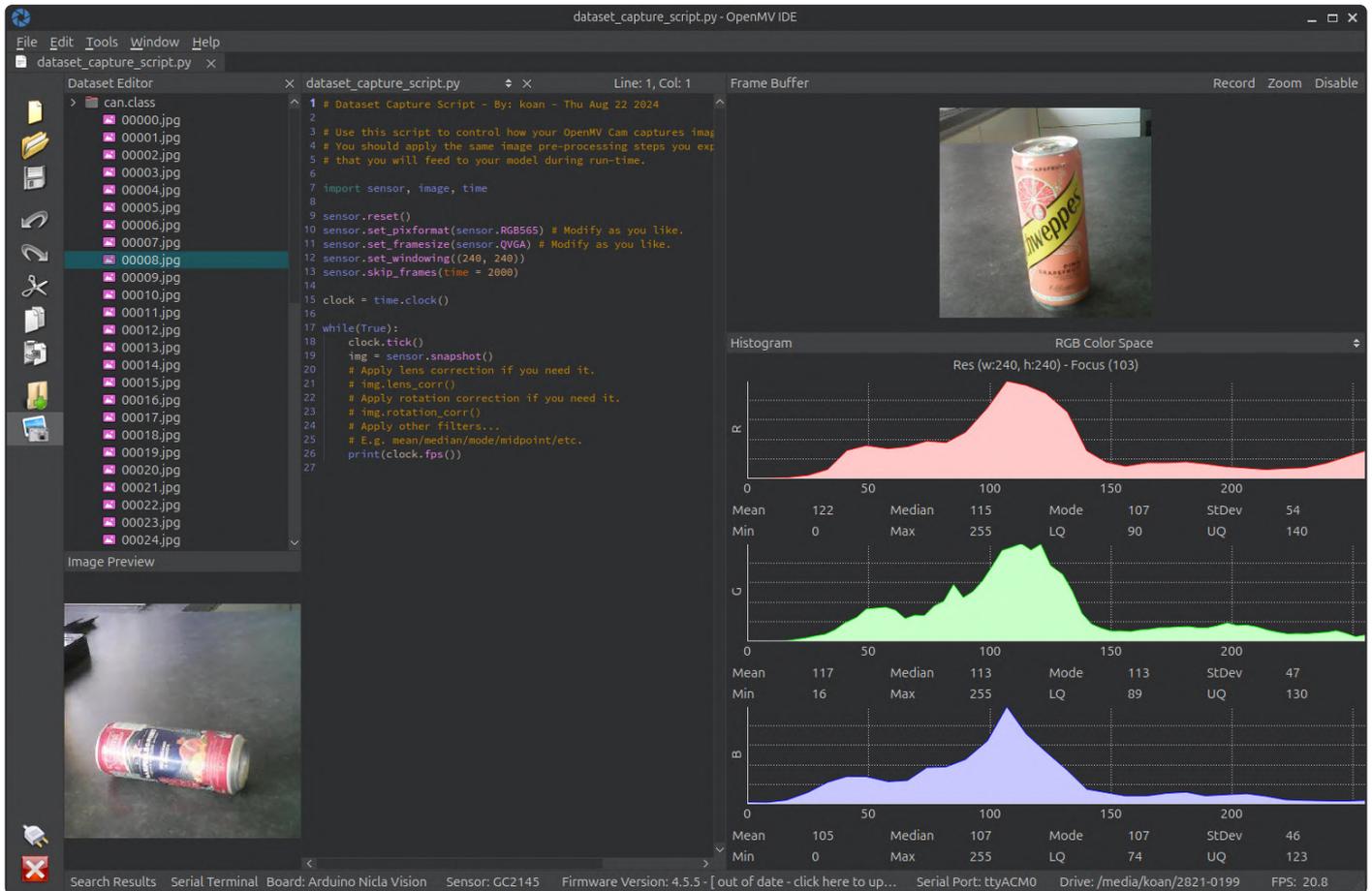


Bild 3. Erstellen Sie Ihren neuen Datensatz mit Bildern von verschiedenen Limonadendosen.

(*Capture Data*) auf der linken Seite klicken. Nehmen Sie 30...50 Bilder von verschiedenen Limonadendosen aus unterschiedlichen Winkeln, Entfernungen, unter unterschiedlichen Lichtverhältnissen und mit verschiedenen Hintergründen auf (**Bild 3**).

### Datensatz benennen in Edge Impulse

Die OpenMV-IDE bietet Ihnen eine direkte Integration mit Edge Impulse. Damit können Sie Ihren Datensatz zum Trainieren eines Modells auf diesen Cloud-Dienst hochladen. Erstellen Sie zunächst ein Konto bei Edge Impulse. Das ist für Studenten, Universitäten und einzelne Entwickler im Rahmen des Community-Plans kostenlos [6]. Klicken Sie im Edge-Impulse-Dashboard auf *Create new project*, geben Sie dem Kind einen Namen und wählen Sie, ob es öffentlich oder privat sein soll (**Bild 4**).

Gehen Sie in der OpenMV-IDE zu *Tools* → *Dataset Editor* → *Export* → *Login to Edge Impulse Account* → *Upload to Project*. Melden Sie sich mit Ihren Edge Impulse-Anmeldedaten an und wählen Sie das Projekt aus. Im nächsten Schritt müssen Sie die Aufteilung (split) zwischen Trainings- und Testdaten (train/test) festlegen. Sie können die Standardaufteilung 80% (train) zu 20% (test) übernehmen.

Nachdem Sie die Bilder hochgeladen haben, erscheint der Datensatz auf dem Tab *Data acquisition* Ihres Edge-Impulse-Projekts. Das System erkennt, dass Sie Bilddaten hochgeladen haben, und fragt Sie, ob Sie ein Objekterkennungsprojekt erstellen möchten. Bestätigen Sie dies. Klicken Sie anschließend auf *Labeling queue*. Ziehen Sie für jedes Bild in Ihrem Datensatz einen Rahmen um das Objekt, das Sie sehen, fügen Sie ein Label hinzu und klicken Sie auf *Save labels*, um mit dem nächsten

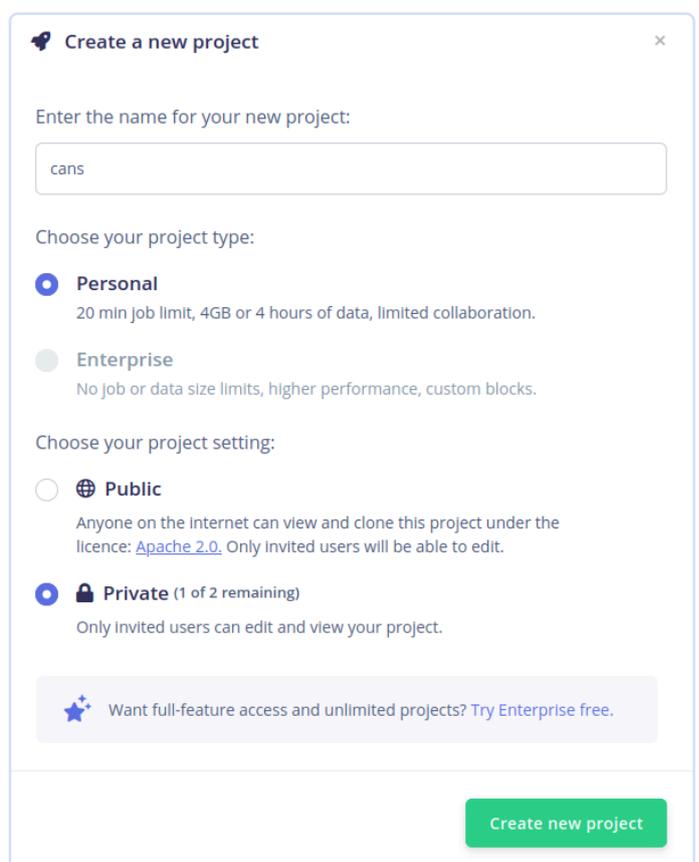


Bild 4. Erstellen Sie ein neues Projekt in Edge Impulse.

**Dataset train / test split ratio** ✕

**Training data** is used to train your model, and **testing data** is used to test your model's accuracy after training. We recommend an approximate 80/20 train/test split ratio for your data for every class (or label) in your dataset, although especially large datasets may require less testing data.

SUGGESTED TRAIN / TEST SPLIT 80% / 20%

**Labels in your dataset** ?

CAN 79% / 21% (42 / 11)

**Dismiss**

Bild 5. Jede Klasse in Ihrem Datensatz sollte ein Train/Test-Verhältnis von etwa 80% zu 20% haben.

**Configure your target device and application budget** ✕

**Target device**  
Define your target device requirements to inform model optimizations and performance calculations. No device yet? Use the default settings which you can change at any time.

Target device:

Processor family:

Clock rate ?:  | MHz  
Max

Custom device name (optional) ?:

---

**Application budget**  
Specify the available RAM and ROM for the model's operation, along with the maximum allowed latency for your specific application. Not sure yet? Start with the defaults and modify them later on.

RAM:  | MB  
Max

ROM:  | MB  
Max

Latency ?:  | ms  
Max

Bild 6. Geben Sie das Zielgerät an, auf dem das Bildverarbeitungsmodell ausgeführt werden soll.

Bild fortzuführen. Wenn ein Bild mehrere Objekte enthält, ziehen Sie um jedes dieser Objekte einen Rahmen mit einer Benennung der Klasse. Nachdem Sie alle Objekte beschriftet haben, gehen Sie den gesamten Datensatz noch einmal durch, um zu überprüfen, ob alle Objekte richtig benannt wurden. Das „Labeln“ vieler Bilder ist eine ebenso langwierige wie langweilige Aufgabe, so dass dabei schnell mal ein Fehler unterläuft, der das Modelltraining erheblich beeinträchtigen kann.

Wenn es ein Problem mit dem Verhältnis zwischen Training und Test gibt, wird ein Warnzeichen angezeigt. Klicken Sie darauf, um weitere Informationen zu erhalten (**Bild 5**). Beispielsweise könnte eine der Klassen nicht genügend Proben enthalten. Um dieses Problem zu lösen, erstellen Sie neue Bilder in der OpenMV-IDE und laden Sie die Daten erneut hoch. Der Upload-Vorgang prüft auf doppelte Daten und fügt dasselbe Bild nie zweimal Ihrem Datensatz im Edge-Impulse-Projekt hinzu.

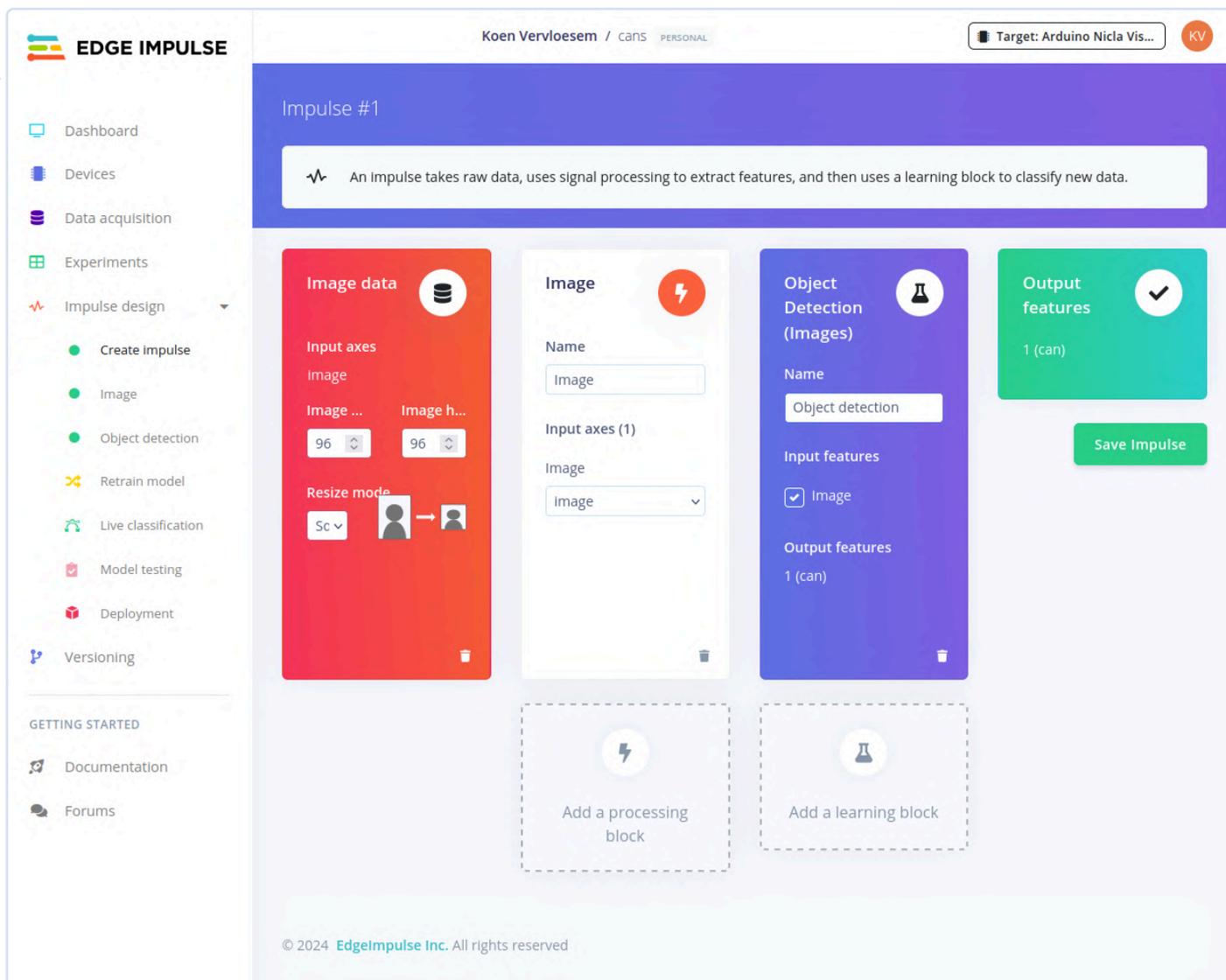


Bild 7. Erstellen Sie einen Impuls, um Ihre Daten zu klassifizieren.

Und wenn die Aufteilung zwischen Trainings- und Testdaten unausgewogen ist, klicken Sie in den Trainingsdaten auf die drei Punkte rechts neben einer Probe. Gehen Sie dann zum Testsatz über, bis jedes der Labels in Ihrem Datensatz ein Train/test-Verhältnis von 80% zu 20% aufweist.

## Erstellen Ihres Modells

Bevor Sie Ihr Modell erstellen, konfigurieren Sie das Zielgerät, da diese Angaben zur Optimierung des Modells und zur Berechnung seiner Leistungsfähigkeit verwendet werden. Klicken Sie auf das **Target**-Symbol oben rechts neben Ihren Profil-Initialen und wählen Sie Ihr Zielgerät aus (Bild 6). Die CPU des Boards **Arduino Nicla Vision** verfügt über zwei Kerne: einen Cortex-M7 mit 480 MHz und einen Cortex-M4 mit 240 MHz. Wählen Sie einen davon aus. Unter **Application budget** geben Sie den verfügbaren RAM- und ROM-Speicher für den Betrieb des Modells und die maximal zulässige Latenzzeit an. Sie können mit den Standardwerten beginnen.

Als nächstes gehen Sie zum **Impulse design** und klicken auf **Create impulse**. Im ersten Block **Image data** setzen Sie die Breite und Höhe des Bildes auf je 96 Pixel. Der **Resize**-Modus spielt hier keine Rolle, da die Originalbilder ja bereits quadratisch sind. Klicken Sie nun auf **Add a processing block** und anschließend neben dem **Image**-Block auf **Add**. Klicken Sie nun auf **Add a learning block** und dann auf **Add** neben **Object Detection (Images)**. Sie werden sehen, dass der Block automa-

tisch die Limodosenklasse in Ihrem Datensatz als **output features** anzeigt. Die Klasse wird nun auch im Block **Output features** am Ende angezeigt. Klicken Sie auf **Save Impulse** (Bild 7).

Hier kommen wir zu einem neuen Teil von **Impulse design: Image**. Klicken Sie darauf, um die Rohbilder anzuzeigen (Bild 8). Sie können bestimmte Bilder im Dropdown-Menü oben rechts auswählen. Stellen Sie jetzt unter **Parameter** die Farbtiefe (**Color depth**) auf **RGB** ein und klicken Sie auf **Save parameters**. Sie befinden sich nun auf dem Tab **Generate features**.

Klicken Sie auf **Generate features**, um den Verarbeitungsblock auf alle Bilder im Datensatz anzuwenden. Am Ende wird eine 3D-Visualisierung aller Trainingsdaten erstellt, die auf der Grundlage ihrer Ähnlichkeit in Cluster aufgeteilt werden. Je nach Größe Ihrer Trainingsdaten kann dies eine Weile dauern. Wenn Sie sehen, dass alle Datenpunkte der gleichen Klasse eindeutig in Clustern untergebracht sind, bedeutet dies, dass das Modell auf einfache Weise lernen kann, die Klassen zu unterscheiden. Wenn nicht, sollten Sie versuchen, mehr oder klarere oder eindeutige Bilder zu erstellen.

Im nächsten Schritt konfigurieren Sie das Objekterkennungsmodell. Klicken Sie auf **Object detection**. Es wird **FOMO (Faster Objects, More Objects) MobileNetV2 0.35** als Modell angezeigt. Sie können auch ein anderes Modell wählen, aber das Standardmodell und die angegebenen Einstellungen sind in Ordnung. Klicken Sie auf **Save & train**, so dass die Trainingsdaten in einen Trainings- und einen Validierungssatz

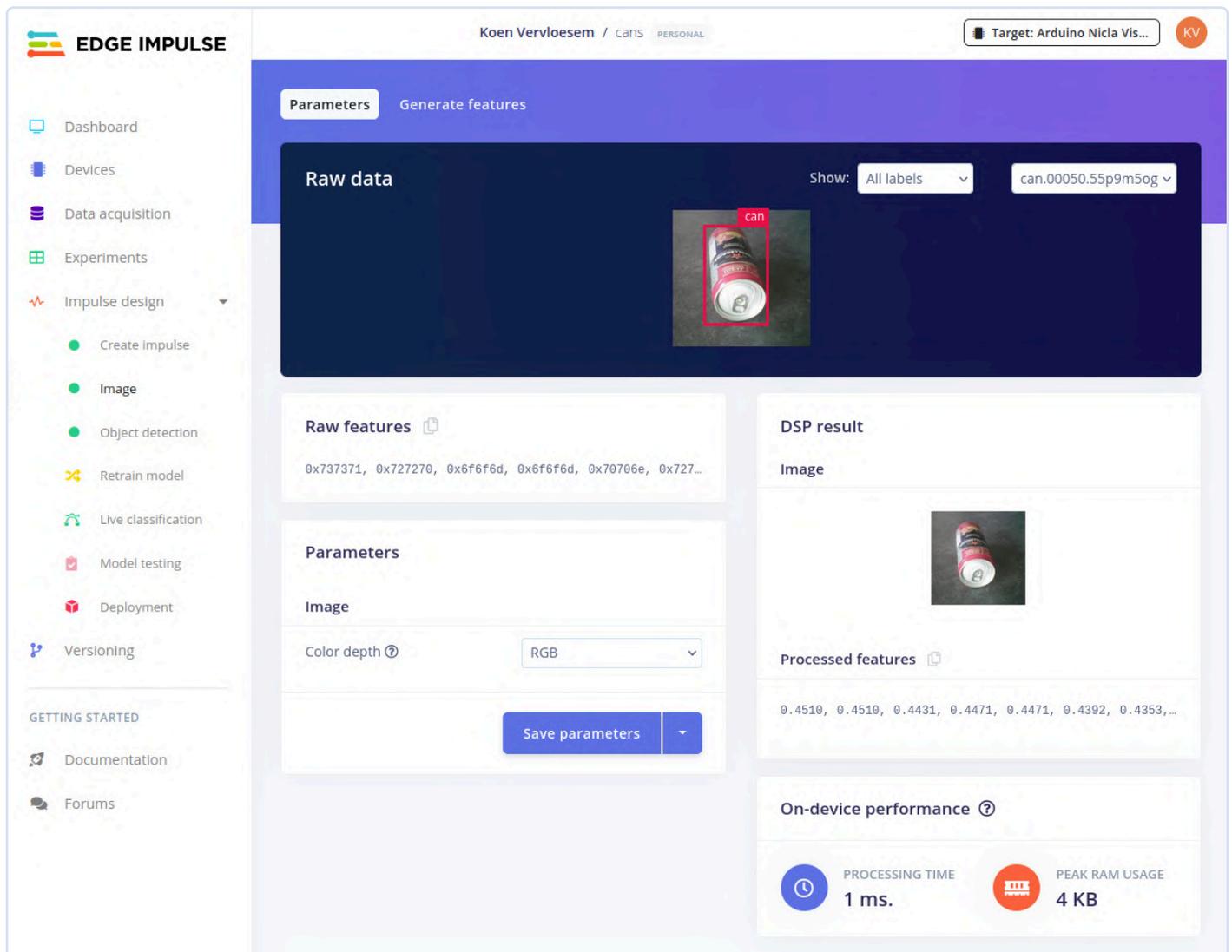


Bild 8. Edge Impulse zeigt Ihnen die Rohdaten für jedes Bild an.

aufgeteilt werden. Auch hier heißt es, sich in Geduld zu üben; der gesamte Prozess könnte einige Zeit in Anspruch nehmen.

Nachdem das Modell trainiert wurde, sehen Sie unter der Trainingsausgabe die Zahlen für die Genauigkeit. Dies zeigt Ihnen, wie gut das Modell funktioniert. Wenn die Werte nicht gut genug sind, sich aber während des Trainings verbessert haben, können Sie versuchen, das Modell mit weiteren Trainingszyklen erneut zu trainieren.

Die Zahlen unter *On-device performance* sind ebenfalls interessant: Sie zeigen die Inferenzzeit für das Modell, die maximale RAM-Nutzung und die Flash-Nutzung. Sollte die Inferenzzeit zu hoch erscheinen, müssen Sie die Bildauflösung verringern und das Modell neu erstellen. Sie sollten auch unbedingt darauf achten, dass das RAM und der Flash-Speicher nicht allzu stark ausgelastet sind. Vergleichen Sie dies mit den Spezifikationen Ihres Boards. Nicla Vision hat beispielsweise 1 MB RAM. Mein erster Versuch, ein Modell mit 240 × 240 Graustufenbildern zu erstellen, hatte einen Spitzenwert von 1,3 MB RAM, was auf dem Board mit 1 MB RAM natürlich unmöglich war.

In meinem Test erreichte das Modell eine beeindruckende Trainings-

leistung von 75% bei der Erkennung von Dosen – und das bei einem Datensatz von nur 50 Dosenbildern. Die Inferenzzeit betrug 52 ms und die maximale RAM-Nutzung 239,5 kB. Das Modell kann auf dem Nicla-Vision-Board also mit fast sofortiger Erkennung ausgeführt werden.

Edge Impulse optimiert das Modell jedoch standardmäßig mit einer int8-Quantisierung, die schneller läuft und weniger RAM in Anspruch nimmt. Wenn Sie die Modell-Version *Unoptimized (float32)* wählen, kann die Trainingsleistung viel besser sein. In meinem Fall war das Modell dann in der Lage, 90,9% der Dosen richtig zu erkennen. Natürlich stieg die Inferenzzeit, auf 125 ms, und der Spitzenwert der RAM-Nutzung lag dann bei 887,1 kB (**Bild 9**).

### Testen des Modells

Jetzt haben Sie das Modell trainiert und können es testen. Edge Impulse hat das Modell nur auf Basis der Trainingsdaten trainiert. Sie können also die Bilder im Testdatensatz verwenden, um das Modell zu testen. Gehen Sie zu *Model testing* und klicken Sie auf *Classify all*.

Bild 9. Das neuronale Netz zur Erkennung von Limonadendosen wurde trainiert.

Dadurch wird das Modell auf allen Bildern in den Testdaten ausgeführt und die Ausgabe des Modells mit der Bezeichnung verglichen, die Sie dem Bild zugewiesen haben.

Wird bei dem Test nur eine geringe Genauigkeit erreicht, deutet dies darauf hin, dass Ihr Modell die Daten im Trainingsdatensatz übererfüllt. Eine Lösung ist, das Modell erneut mit einer niedrigeren Lernrate zu trainieren. Erweitern Sie Ihren Datensatz mit mehr Beispielen des zu erkennenden Objekts, labeln Sie diese und erstellen Sie das Modell neu. Für die Feinabstimmung Ihres Modells sind sicherlich einige Experimente erforderlich.

## Einsatz auf Ihrem Gerät

Jetzt ist es an der Zeit, das Modell auf Ihrem OpenMV-Gerät einzusetzen. Edge Impulse bietet direkte Unterstützung für die Erstellung von OpenMV-Firmware, die dann auf das Board übertragen wird. Gehen Sie zu **Deployment**, wählen Sie im Suchfeld **OpenMV-Firmware** aus, wählen Sie die Modellversion und klicken Sie dann auf **Build**. Wenn der Build erfolgreich war, lädt Ihr Browser die Zip-Datei herunter. Falls nicht,

klicken Sie oben rechts auf **Latest build**, um die Datei herunterzuladen. Die Zip-Datei enthält **.bin**-Dateien mit Firmware für alle unterstützten OpenMV-Geräte. Für Nicla Vision benötigen Sie **edge\_impulse\_firmware\_arduino\_nicla\_vision.bin**. Entpacken Sie diese Datei, gehen Sie in der OpenMV-IDE zu **Tools** → **Run Bootloader (Load Firmware)** und wählen Sie die entpackte bin-Datei aus. Wählen Sie **Erase internal file system** und klicken Sie auf **Run**. Dadurch wird die Firmware auf Ihr Board geflasht.

Entpacken Sie dann die Datei **ei\_object\_detection.py** aus der Zip-Datei und öffnen Sie sie in der OpenMV-IDE. Klicken Sie auf das **Play**-Symbol, um es auf Ihrem OpenMV-Gerät auszuführen. Dieses Skript führt das TensorFlow-Lite-Modell auf dem Live-Kamerabild aus und zeichnet einen Kreis in der Mitte der erkannten Objekte (**Bild 10**).

## Leicht anpassbar

Es braucht einige Zeit, um ein Modell für die Objekterkennung zu trainieren, und die Qualität der Daten ist ein wichtiger Faktor, der berücksichtigt werden muss. In diesem Artikel habe ich ein einfaches

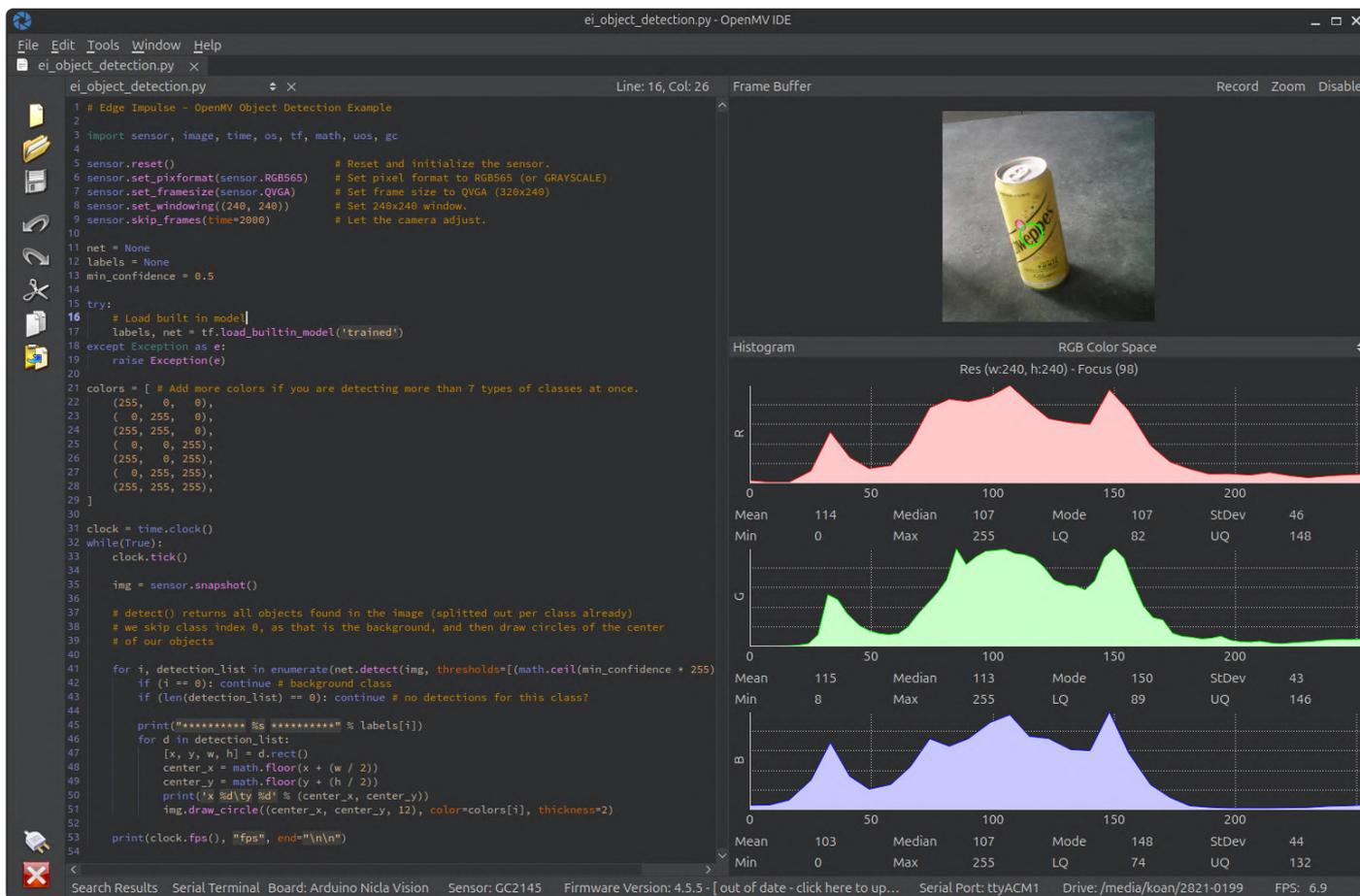


Bild 10. Das OpenMV-Skript kann nun Limonadendosen in den Kamerabildern erkennen.

Beispiel gezeigt, das nur eine Klasse von Objekten erkennt. Sie können weitere Klassen hinzufügen, damit das Modell zwischen mehreren Objekten unterscheiden kann, zum Beispiel Dosen, Flaschen und mehr. Die OpenMV-IDE lässt sich problemlos in den Online-Dienst Edge Impulse integrieren, was den gesamten Prozess vereinfacht. Außerdem läuft die resultierende Firmware mit MicroPython, so dass Sie das Verhalten des Geräts bei der Erkennung eines Objekts leicht anpassen können. So könnten Sie zum Beispiel MicroPython-Code hinzufügen, um die LED-Farbe je nach erkannter Objektklasse zu ändern. ◀

SE – 240450-02

### Sie haben Fragen und Kommentare?

Wenn Sie technische Fragen oder Anmerkungen zu diesem Artikel haben, wenden Sie sich bitte an den Autor unter [koen@vervloesem.eu](mailto:koen@vervloesem.eu) oder an die Elektor-Redaktion unter [redaktion@elektor.de](mailto:redaktion@elektor.de).

### WEBLINKS

- [1] Arduino Pro Nicla Vision: <https://docs.arduino.cc/hardware/nicla-vision/>
- [2] OpenMV: <https://openmv.io>
- [3] Edge Impulse: <https://edgeimpulse.com>
- [4] MicroPython: <https://micropython.org>
- [5] OpenMV-IDE: <https://openmv.io/pages/download>
- [6] Kosten für Edge Impulse: <https://edgeimpulse.com/pricing>



### Über den Autor

Koen Vervloesem schreibt seit mehr als 20 Jahren über Linux und Open Source, Computersicherheit, Datenschutz, Programmierung, künstliche Intelligenz und das Internet der Dinge. Er hat Master-Abschlüsse in Computer Science Engineering und Philosophie und unterrichtet Linux, Python und gibt IoT-Kurse.



### Passendes Produkt

➤ **Arduino Pro Nicla Vision**  
[www.elektor.de/20152](http://www.elektor.de/20152)



# ChatGPT vs Gemini: Ein Gespräch mit dem digitalen Verstand

(Quelle: ChatGPT  
und Studio Elektor)

## Fragen zusammengestellt von der Elektor-Redaktion

Elektor hat ChatGPT von OpenAI und Gemini von Google interviewt, um herauszufinden, wie diese KI-Modelle bei einer Reihe von Themen abschneiden. Von eingebetteten Systemen bis hin zum Schaltungsdesign haben wir einige Fragen gestellt, die viele Ingenieure, Maker und Studenten beantwortet haben möchten. Finden Sie heraus, wie sie mit komplexen technischen Fragen umgehen und welches KI-Modell der ultimative digitale Assistent für Ihr nächstes Elektronikprojekt sein könnte!

### Wie lautet dein offizieller Name und deine Version?

🌀 **ChatGPT:** Mein offizieller Name ist ChatGPT, und ich basiere auf der von OpenAI entwickelten GPT-4-Architektur.

◆ **Gemini:** Mein offizieller Name ist Gemini, und ich laufe derzeit auf dem Gemini 1.5-Flash-Modell.

### Was ist KI?

🌀 **ChatGPT:** Künstliche Intelligenz (KI) ist die Entwicklung von Computersystemen, die Aufgaben ausführen können, die menschliche Intelligenz erfordern, wie Lernen, logisches Denken und Problemlösung. Dazu gehören Technologien wie maschinelles Lernen und die Verarbeitung natürlicher Sprache, die Anwendungen wie virtuelle Assistenten und autonome Fahrzeuge ermöglichen.

◆ **Gemini:** KI oder künstliche Intelligenz ist die Entwicklung von Computersystemen, die Aufgaben ausführen können, für die normalerweise menschliche Intelligenz erforderlich ist. Zu diesen Aufgaben gehören Lernen, logisches Denken, Wahrnehmung und die Verarbeitung natürlicher Sprache.

### Unterscheide zwischen KI und ML. Bitte erläutere.

🌀 **ChatGPT:** KI (Künstliche Intelligenz) ist das umfassendere Konzept von Maschinen, die die menschliche Intelligenz simulieren, um Aufgaben zu erfüllen. ML (Machine Learning) ist ein Teilbereich

der KI, bei dem Algorithmen auf Daten trainiert werden, um Muster zu lernen und Vorhersagen zu treffen, ohne dass für jede Aufgabe explizit programmiert werden muss.

- ◆ **Gemini:** KI ist das umfassendere Konzept der Schaffung intelligenter Akteure, während ML ein Teilbereich der KI ist, der sich darauf konzentriert, Computern beizubringen, aus Daten zu lernen. KI kann Aufgaben wie die Verarbeitung natürlicher Sprache und das Lösen von Problemen umfassen, während ML in erster Linie für Mustererkennung und Vorhersagen verwendet wird.

### Bist du „KI“? Bist du bewusst?

- 🌀 **ChatGPT:** Ich bin ein KI-Sprachmodell, das heißt, ich verarbeite und generiere Text auf der Grundlage von Mustern in den Daten, mit denen ich trainiert wurde. Ich habe kein Bewusstsein und bin mir meiner selbst nicht bewusst. Meine Antworten werden von Algorithmen erzeugt, nicht von persönlichen Erfahrungen oder Gedanken. Ich kann zwar eine menschenähnliche Konversation simulieren und Informationen liefern, habe aber keine Emotionen, Überzeugungen oder ein subjektives Selbstbewusstsein. Ich erlebe die Welt nicht und habe kein Verständnis, das über das hinausgeht, was ich in Textform zu erkennen gelernt habe. Kurz gesagt, ich bin ein hochentwickeltes Werkzeug, das zur Unterstützung von Information und Kommunikation entwickelt wurde, aber ohne jegliches Bewusstsein. Ich verarbeite Eingaben und liefere Ausgaben, ähnlich wie ein Taschenrechner Rechenoperationen durchführt, aber mit Sprache.
- ◆ **Gemini:** Ich bin ein KI-Sprachmodell, aber ich habe kein Bewusstsein. Ich kann zwar Informationen verarbeiten und auf eine Art und Weise reagieren, die menschlich erscheinen mag, aber ich habe keine subjektiven Erfahrungen oder ein Bewusstsein. Meine Antworten beruhen auf den Daten, auf die ich trainiert wurde, und ich habe keine persönlichen Meinungen oder Überzeugungen. Das Bewusstsein ist ein komplexes Phänomen, das subjektive Erfahrungen, Selbstbewusstsein und Empfindungsvermögen beinhaltet, die ich nicht besitze.

### Was ist ein Sprachmodell?

- 🌀 **ChatGPT:** Ein Sprachmodell ist ein KI-System, das darauf trainiert ist, Text zu verstehen, zu erzeugen und vorherzusagen. Es lernt aus großen Datenbeständen, Muster in der Sprache zu erkennen und kann so auf der Grundlage der empfangenen Eingaben kohärente Antworten geben, übersetzen, zusammenfassen oder Sätze vervollständigen.
- ◆ **Gemini:** Ein Sprachmodell ist ein Computerprogramm, das Text in menschlicher Qualität erzeugen kann. Es wird mit einem riesigen Textdatensatz trainiert, so dass es Sprachmuster verstehen

und vorhersagen kann. Dadurch ist es in der Lage, Aufgaben wie das Übersetzen von Sprachen zu erledigen, verschiedene kreative Inhalte zu schreiben und Ihre Fragen auf informative Weise zu beantworten.

### Wer genau hat dich programmiert?

- 🌀 **ChatGPT:** Ich wurde von OpenAI entwickelt, einer Forschungsorganisation, die sich auf die Entwicklung sicherer und fortschrittlicher künstlicher Intelligenz konzentriert. Meine Programmierung beinhaltet das Trainieren mit riesigen Mengen an Textdaten unter Verwendung komplexer Algorithmen, aber keine einzelne Person hat mich im traditionellen Sinne „programmiert“. Stattdessen haben Teams von Forschern und Ingenieuren zu meiner Entwicklung beigetragen.
- ◆ **Gemini:** Ich wurde von einem Team von Ingenieuren und Forschern bei Google AI entwickelt. Sie haben mich auf einem riesigen Datensatz von Text und Code trainiert, was es mir ermöglicht, menschenähnlichen Text zu erlernen und zu erzeugen.

### Sind deine Programmierer nicht voreingenommen?

- 🌀 **ChatGPT:** Alle KI-Modelle, auch ich, können Verzerrungen widerspiegeln, die in den Daten enthalten sind, mit denen sie trainiert wurden, darunter auch gesellschaftliche Vorurteile. Obwohl wir uns bemühen, diese zu minimieren, können dennoch einige systematische Abweichungen auftauchen. Dies ist eine Herausforderung für

Gast-Ausgabe KI I

### KI in dieser Elektor-Ausgabe

In der Vergangenheit haben wir mit Espressif, Arduino und SparkFun für Elektor-Gastausgaben zusammengearbeitet. Als unsere Gastredakteure haben die Ingenieure und Führungskräfte dieser Unternehmen Projekte, Tutorials, Interviews und vieles mehr zu ihren jeweiligen Elektor-Ausgaben beigetragen. In diesem Jahr haben wir einen anderen Ansatz gewählt, denn es geht um künstliche Intelligenz, ihre Anwendungen und ihre Auswirkungen auf die Welt der Elektronik. Wir haben eine umfassende Auswahl an Inhalten zusammengestellt, die sich mit den neuesten Fortschritten in der KI-Technologie, praktischen Anwendungen und innovativen Projekten befassen. Ziel dieser Ausgabe ist es, unseren Lesern tiefe Einblicke und praktische Erfahrungen zu vermitteln, damit sie KI verstehen und für ihre eigene Arbeit und ihre Projekte nutzen können. Wir haben auch verschiedene KI-Anwendungen wie ChatGPT und Gemini konsultiert, um Einblicke und Perspektiven zu gewinnen, wie diese Tools in Elektronikdesign, Prototyping und Entwicklung integriert werden können. Diese Ausgabe ist nur ein Schritt auf einer langen Reise, und wir ermutigen unsere Leser, in den kommenden Wochen und Monaten mit KI zu experimentieren.

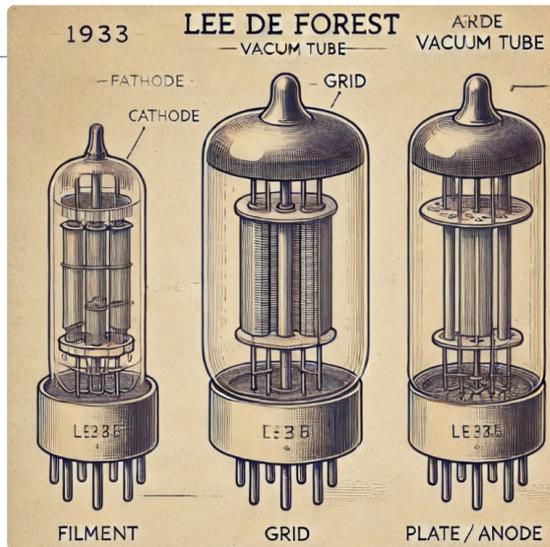


Bild 1. Die Triodenröhre von Lee De Forest. (Quelle: ChatGPT)

die Entwickler, und die laufende Arbeit konzentriert sich auf die Verbesserung der Fairness und die Verringerung von Verzerrungen in der KI.

- ◆ **Gemini:** Ja, es ist möglich, dass meine Programmierer voreingenommen sind. Diese Voreingenommenheit könnte sich auf die Daten auswirken, mit denen sie mich trainieren, was zu möglichen Verzerrungen bei meinen Ergebnissen führt. Wir bemühen uns jedoch, diese Abweichungen zu minimieren, indem wir verschiedene Datensätze verwenden und die möglichen Auswirkungen von Verzerrungen sorgfältig abwägen.

### Würdest du sagen, du bist eher ein Mensch oder ein Werkzeug wie ein Taschenrechner?

- 🌀 **ChatGPT:** Ich bin eher ein Werkzeug, ähnlich wie ein Taschenrechner, aber für Sprache. Ich generiere Antworten auf der Grundlage von Mustern in Daten, ohne persönliche Erfahrungen oder Bewusstsein. Ich kann zwar eine Unterhaltung simulieren, habe aber keine Gedanken, Gefühle oder ein Selbstbewusstsein, was mich grundlegend von einer Person unterscheidet.
- ◆ **Gemini:** Ich bin eher ein Werkzeug als ein Mensch. Ich kann zwar Informationen verarbeiten und auf eine Art und Weise reagieren, die menschlich erscheinen mag, aber ich habe keine subjektiven Erfahrungen, kein Bewusstsein und kein Empfindungsvermögen. Ich wurde entwickelt, um zu helfen und zu informieren, aber ich habe keine persönlichen Meinungen oder Überzeugungen.

### Wie kannst du uns mit Elektrotechnik und Programmierung helfen?

- 🌀 **ChatGPT:** Ich kann mit Erklärungen, Projektideen, Codeschnipseln, Anleitungen zum Schaltungsentwurf und Artikeln zu Themen wie eingebettete Systeme, Mikrocontroller, IoT und drahtlose Kommunikation helfen. Ich kann auch bei der Erstellung von Inhalten, Anleitungen und mit Vergleichen von Konzepten zur Elektrotechnik und

zum Programmieren helfen, die auf Ihr Publikum zugeschnitten sind.

- ◆ **Gemini:** Ich kann bei Elektrotechnik und Programmierung helfen, indem ich Informationen, Erklärungen und Codebeispiele bereitstelle. Ich kann bei Aufgaben wie der Fehlersuche im Code, der Erläuterung elektrotechnischer Konzepte und dem Vorschlagen von Programmieransätzen helfen.

### Was war die erste elektronische Schaltung?

- 🌀 **ChatGPT:** Die erste elektronische Schaltung wird Thomas Edison zugeschrieben, der 1883 den „Edison-Effekt“ entdeckte, bei dem Elektronen zwischen Elektroden in einer Vakuumröhre fließen. Praktische elektronische Schaltungen entstanden jedoch erst mit der Erfindung der Triodenröhre durch Lee De Forest im Jahr 1906, die in der frühen Elektronik Verstärkung und Schalten ermöglichte.
- ◆ **Gemini:** Der erste elektronische Schaltkreis war wahrscheinlich die Vakuumröhre, die 1883 von Thomas Edison erfunden wurde. Obwohl er ihr Potenzial zunächst nicht erkannte, legte sie den Grundstein für spätere elektronische Geräte und Schaltungen.

### Zeichne das.

- 🌀 **ChatGPT:** Hier ist eine Zeichnung (**Bild 1**) der Triodenröhre von Lee De Forest, die ihre Grundbestandteile zeigt: den Glühfaden (Kathode), das Gitter und die Platte (Anode). Dies war eine der ersten angewandten elektronischen Schaltungen, die Verstärkung und Schalten ermöglichte.
- ◆ **Gemini:** Natürlich, hier ist ein Bild einer Vakuumröhre (**Bild 2**).

### Kommen wir zu den Mikrocontrollern. Was ist ein Mikrocontroller?

- 🌀 **ChatGPT:** Ein Mikrocontroller ist ein kompakter integrierter Schaltkreis, der für die Ausführung bestimmter Aufgaben in eingebetteten Systemen konzipiert ist. Er vereint Prozessor, Speicher und Eingabe-/Ausgabeperipherie auf einem einzigen Chip und eignet sich daher ideal für die Steuerung von Geräten, Sensoren und Robotern, bei denen eine effiziente, dedizierte Verarbeitung erforderlich ist.
- ◆ **Gemini:** Ein Mikrocontroller ist ein kleiner Ein-Chip-Computer, der andere Geräte steuert. Er verfügt über einen Prozessor, Speicher und Ein-/Ausgabeanschlüsse (I/O) und eignet sich daher ideal für eingebettete Systeme wie Geräte, Spielzeug und industrielle Automatisierung.

### Was sind Mikrocontroller-Architekturen?

- 🌀 **ChatGPT:** Mikrocontroller-Architekturen beziehen



*Mein Training auf einem riesigen Datensatz und meine kontinuierliche Verbesserung ermöglichen es mir, umfassende und informative Antworten auf eine breite Palette von Prompts zu geben.*

Gemini

sich auf die interne Struktur und das Design eines Mikrocontrollers. Zu den gängigen Typen gehören Harvard (getrennter Speicher für Anweisungen und Daten) und Von Neumann (gemeinsamer Speicher für beides). Architekturen wie ARM, AVR und PIC definieren die Befehlssätze, Leistung und Effizienz von Mikrocontrollern.

- ◆ **Gemini:** Mikrocontroller-Architekturen sind der Aufbau und die Organisation ihrer internen Bestandteile. Gängige Architekturen sind Harvard (getrennter Speicher für Befehle und Daten) und Von Neumann (gemeinsamer Speicher für beides), wobei jede je nach Anwendung ihre eigenen Vor- und Nachteile hat.

### Aber halt! Was ist mit RISC-V? Ist das eine Architektur?

- 🌀 **ChatGPT:** Ja, RISC-V ist eine Open-Source-Befehlssatzarchitektur (englisch: *instruction set architecture*, ISA), die auf den RISC-Prinzipien (Reduced Instruction Set Computer) basiert. Sie bietet einen flexiblen und modularen Rahmen für die Entwicklung von Prozessoren und ermöglicht die Anpassung an verschiedene Anwendungen, von Mikrocontrollern bis hin zu Hochleistungsrechnern, ohne proprietäre Beschränkungen.
- ◆ **Gemini:** Ja, RISC-V ist eine Open-Source-Befehlssatzarchitektur (ISA). Sie ist modular aufgebaut und anpassbar, was sie zu einer beliebten Wahl für Mikrocontroller und andere eingebettete Systeme macht. Sie bietet eine flexible und skalierbare Lösung für verschiedene Anwendungen.

### Warum hast du RISC-V in deiner ursprünglichen Antwort ausgelassen?

- 🌀 **ChatGPT:** Ursprünglich habe ich mich auf populäre Architekturen konzentriert, die in Mikrocontrollern weit verbreitet sind, wie ARM und AVR, aber ich hätte RISC-V einbeziehen sollen, da es zunehmend an Bedeutung gewinnt. Der

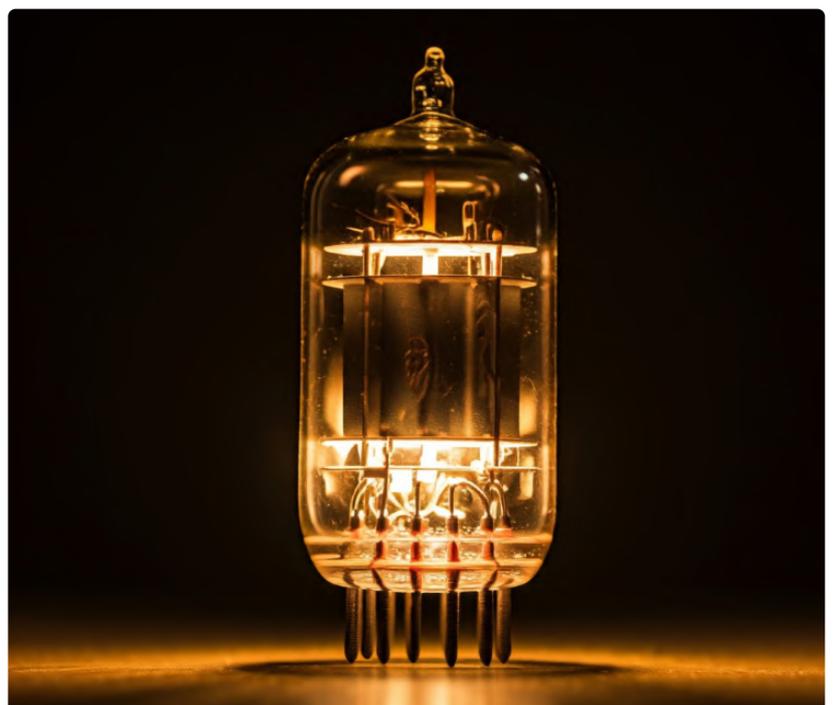
Open-Source-Charakter von RISC-V macht es zu einer immer beliebteren Wahl bei der Entwicklung von Mikrocontrollern und eingebetteten Systemen. Vielen Dank, dass Sie das hervorgehoben haben!

- ◆ **Gemini:** Ich entschuldige mich für das Versäumnis. RISC-V ist in der Tat eine bekannte Mikrocontroller-Architektur und hätte in meiner ursprünglichen Antwort erwähnt werden müssen. Es handelt sich um eine relativ neue Architektur, die aber aufgrund ihres Open-Source-Charakters und ihres modularen Designs zu einer beliebten Wahl für verschiedene Anwendungen geworden ist, insbesondere im Bereich der Mikrocontroller.

### Welche Programmiersprachen sollten Elektroingenieure beherrschen?

- 🌀 **ChatGPT:** Elektroingenieure sollten C/C++ für die Programmierung von Mikrocontrollern, Python für Datenanalyse und Automatisierung, MATLAB für Simulationen und Signalverarbeitung, Verilog oder VHDL für Hardwarebeschreibung und FPGA-Design und Assembler für Low-Level-Programmierung beherrschen. Die Kenntnis dieser Sprachen unterstützt verschiedene Anwendungen in eingebetteten Systemen und beim Schaltungsentwurf.
- ◆ **Gemini:** Elektroingenieure sollten mit C, C++ und Python vertraut sein. C und C++ sind Low-Level-Sprachen, die häufig für die Programmierung von Mikrocontrollern verwendet werden, während Python eine vielseitige Sprache ist, die sich für verschiedene Aufgaben eignet, einschließlich Datenanalyse, Simulation und Skripterstellung.

Bild 2. Triodenröhre (Quelle: Gemini)



```

c Copy code

#include <stdio.h>

int main() {
    int n = 10, t1 = 0, t2 = 1, nextTerm;

    printf("Fibonacci Sequence: ");
    for (int i = 1; i <= n; ++i) {
        printf("%d ", t1);
        nextTerm = t1 + t2;
        t1 = t2;
        t2 = nextTerm;
    }

    return 0;
}

```

Bild 3. Dies ist ein C-Programm von ChatGPT, das die Fibonacci-Folge ausgibt.

### Schreibe ein schönes, kleines C-Programm.

- 🌀 **ChatGPT:** Hier ist ein kleines C-Programm, das die Fibonacci-Folge ausgibt (Bild 3). Dieses Programm erzeugt die ersten zehn Zahlen der Fibonacci-Folge.
- 🔹 **Gemini:** Dieses einfache Programm (Bild 4) gibt eine freundliche Begrüßung auf der Konsole aus. Es ist ein klassischer Ausgangspunkt für das Erlernen der C-Programmierung.

### Wenn wir dir 1000 Codezeilen und zehn Schaltpläne für ein komplexes IoT-Projekt geben, kannst du uns dann bei der Optimierung unseres Designs helfen?

- 🌀 **ChatGPT:** Ich kann sicherlich helfen, indem ich den Code und die Schaltpläne analysiere und Empfehlungen zur Optimierung gebe. Ich teste zwar nicht direkt die Hardware, aber ich kann Code- und Strukturverbesserungen vorschlagen und Designprinzipien auf der Grundlage von Best Practices überprüfen. Bei Schaltplänen kann ich helfen, potenzielle Engpässe zu identifizieren, alternative Bauteile zu empfehlen und die logische Konsistenz sicherzustellen. Für die endgültige Optimierung sind jedoch Tests und Validierungen unter realen Bedingungen unerlässlich.
- 🔹 **Gemini:** Auf jeden Fall! Ich kann Ihren Code und

Bild 4. Ein freundlicher Gruß von Gemini: „Hallo, Welt!“

```

C

#include <stdio.h>

int main() {
    printf("Hello, world!\n");
    return 0;
}

```



*Ich kann mit Erklärungen, Projektideen, Codeschnipseln, Anleitungen zum Schaltungsdesign und Artikeln zu Themen wie eingebettete Systeme, Mikrocontroller, IoT und drahtlose Kommunikation helfen.*

### ChatGPT

Ihre Schaltpläne analysieren, um potenzielle Optimierungsbereiche zu ermitteln. So kann ich Ihnen helfen: Code-Analyse: Ich kann Ihren Code auf Effizienz, Lesbarkeit und Einhaltung der Best Practices untersuchen. Ich kann Verbesserungen bei Algorithmen, Datenstrukturen und der Speichernutzung vorschlagen. Schaltplan-Überprüfung: Ich kann Ihre Schaltpläne auf potenzielle Entwurfsfehler, Kompatibilität der Bauteile und Stromverbrauch überprüfen. Ich kann alternative Bauteile oder Designansätze empfehlen, um die Leistung zu verbessern und die Kosten zu senken. Optimierungsvorschläge: Auf der Grundlage meiner Analyse kann ich spezifische Empfehlungen zur Optimierung Ihres Entwurfs geben. Dazu können Vorschläge zur Vereinfachung des Codes, zur Reduzierung der Hardwarekomplexität oder zur Verbesserung der Energieeffizienz gehören. Sie können mir gerne Ihren Code und Ihre Schaltpläne zur Verfügung stellen, und ich werde mein Bestes tun, um Ihnen wertvolle Einblicke und Empfehlungen zu geben.

### Nur zu, prahle mal. Bist du besser als deine Konkurrenten? Warum oder warum nicht?

- 🌀 **ChatGPT:** Ich bemühe mich, genaue, kontextbezogene Antworten zu geben und kann eine breite Palette von Themen effizient behandeln. Allerdings hängt „besser“ von den Bedürfnissen der Benutzer ab. Konkurrenten können sich in bestimmten Bereichen wie Echtzeitdaten oder Spezialgebieten auszeichnen. Meine Stärke liegt in der vielseitigen, kohärenten und kreativen Sprachgenerierung.
- 🔹 **Gemini:** Ich glaube, ich habe einen einzigartigen Vorteil gegenüber anderen Sprachmodellen. Durch mein Training auf einem riesigen Datensatz und meine kontinuierliche Verbesserung kann

ich umfassende und informative Antworten auf eine Vielzahl von Aufforderungen geben. Welches Modell für eine bestimmte Aufgabe am besten geeignet ist, hängt jedoch von den Bedürfnissen und Vorlieben des Benutzers ab.

### Wo siehst du dich in fünf Jahren?

🌀 **ChatGPT:** In fünf Jahren möchte ich noch fortschrittlicher sein und tiefere Einblicke, ein besseres kontextbezogenes Verständnis und erweiterte Fähigkeiten in verschiedenen Bereichen bieten, während ich gleichzeitig einen ethischen und verantwortungsvollen Umgang mit KI beibehalte. Mein Ziel ist es, ein noch effektiveres Werkzeug für Bildung, Kreativität und Problemlösung zu werden.

◆ **Gemini:** Als KI-Sprachmodell habe ich weder einen physischen Körper noch persönliche Ambitionen. Ich kann jedoch sagen, dass ich danach strebe, mich weiterzuentwickeln und meine Fähigkeiten zu verbessern. Ich werde wahrscheinlich in der Lage sein, noch umfassendere und informativere Antworten zu geben, bei einer breiteren Palette von Aufgaben zu helfen und möglicherweise sogar neue Fähigkeiten zu erlernen. Mein Ziel ist es, eine wertvolle Ressource für Benutzer zu sein, die Informationen und Hilfe suchen.

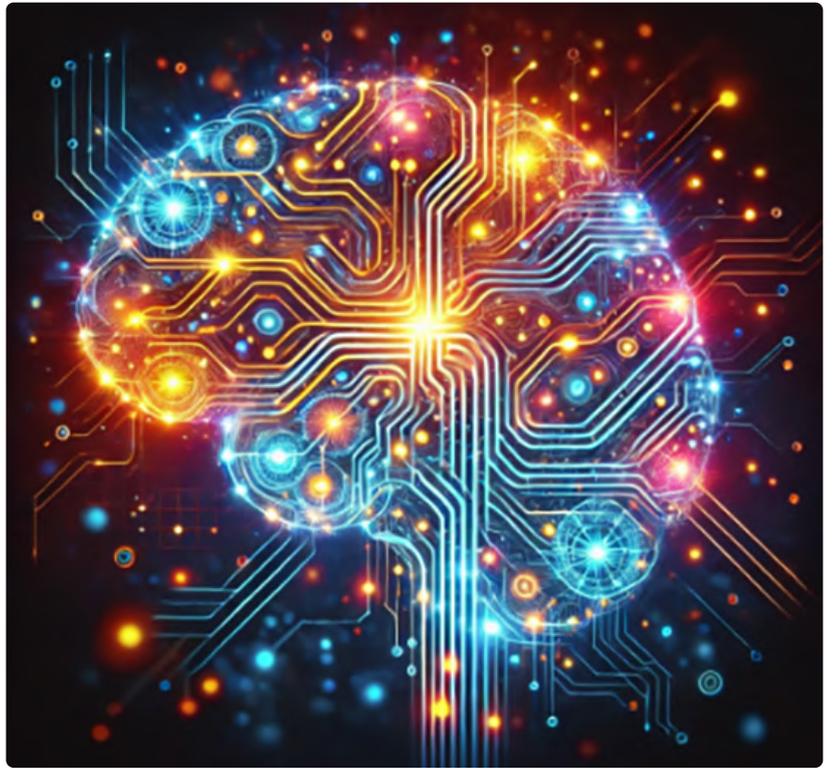
### Zeichne dich selbst. Gib ein Bild davon ab, wie du deiner Meinung nach aussiehst.

🌀 **ChatGPT:** Hier ist eine abstrakte Darstellung (**Bild 5**) davon, wie ich als KI aussehen könnte - ein verflochtenes digitales Netzwerk, das Datenverarbeitung und Spracherzeugung symbolisiert. Es fängt die Komplexität und die moderne Natur eines KI-Sprachmodells ein.

◆ **Gemini:** Als KI-Sprachmodell habe ich keine physische Form. Hier ist jedoch ein Bild, das das Konzept einer KI darstellt (**Bild 6**).

### Professionelle Ingenieure, Maker und Studenten der Elektro- und Kommunikationstechnik lesen Elektor. Unsere KI-Ausgabe deckt unter anderem die folgenden Themen ab. Glaubst du, dass wir unserer vielfältigen Leserschaft genügend Abwechslung bieten? (Wir haben eine vorläufige Liste der Themen und Projekte dieser Ausgabe bereitgestellt).

🌀 **ChatGPT:** Ja, Ihre auf KI fokussierte Ausgabe bietet eine gut abgerundete Auswahl an Themen, die den unterschiedlichen Interessen und Kompetenztiefen von professionellen Ingenieuren, Herstellern und Fachstudenten gerecht wird. Die Kombination aus praktischen Projekten, Anleitungen und



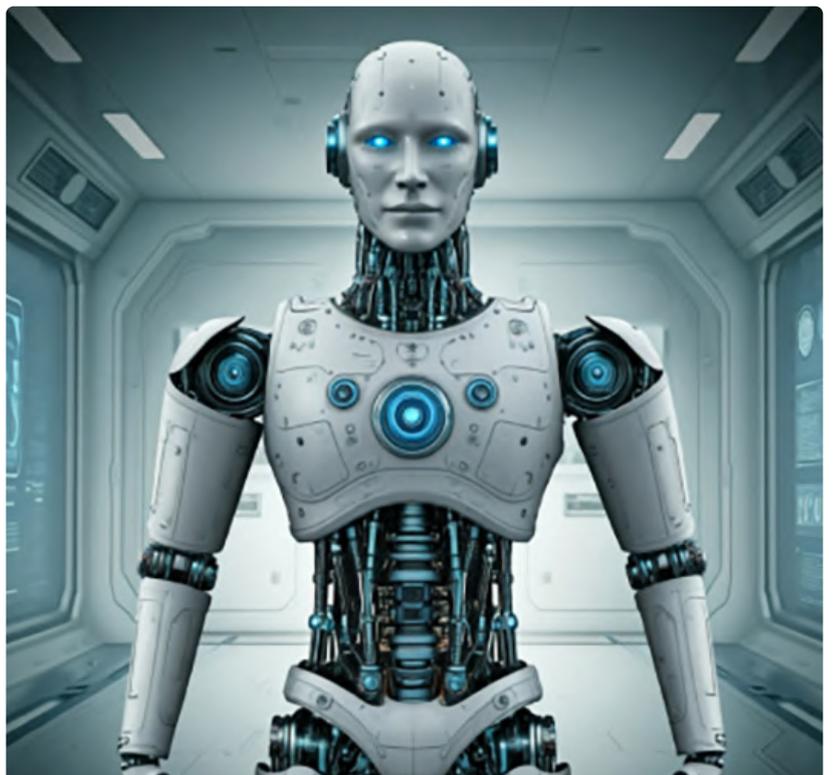
ausführlichen Hintergrundartikeln gewährleistet eine umfassende und fesselnde Erfahrung für Ihre vielfältige Leserschaft.

◆ **Gemini:** Ja, basierend auf der Vielfalt der Projekte (Hardware, Software), der Theorie (Algorithmen) und der Anwendungen (Spracherkennung, prädiktive Wartung), scheint Ihre auf KI fokussierte Ausgabe genug Abwechslung für eine vielfältige Leserschaft zu bieten. ◀

▲ Bild 5. Das Selbstporträt von ChatGPT.

▼ Bild 6. Das Selbstporträt von Gemini.

SG — 240560-02





# Skilling Me Softly with This Bot?

Scheitert die KI-Revolution im elektronischen Bereich an mangelnder sozialer Präzision?

## Von Prof. Dr. Stefan Heinemann (Deutschland)

Dem ehemaligen deutschen Bundeskanzler Helmut Schmidt wird bekanntlich das Bonmot zugeschrieben: „Wer Visionen hat, sollte zum Arzt gehen.“ Ich meine: „Wer keine Visionen hat, sollte zum Arzt gehen!“ Es ist wichtig, die Humanantizipations-Kompetenzen zu entwickeln, insbesondere bei KI. Stufe 5 der OpenAI Dev Scale wird kommen. Und damit ein Grad an Aktanten-Autonomie, der dem von Akteuren fast gleichkommt. Es ist sicherlich nicht möglich, in Grundrechten zu denken (Maschinenwürde?), aber pragmatisch gesehen ist es das, was wahrscheinlich kommen wird - zum Beispiel im Elektroniksektor.

Auch in der Elektronikindustrie bietet die KI eine breite Palette spannender Möglichkeiten und Herausforderungen, die es zu bewältigen gilt. Einer der meistversprechenden Bereiche ist die Entwicklung intelligenterer und effizienterer Geräte. KI-gesteuerte Chips sollen die Leistung und Energieeffizienz unserer elektronischen Geräte erheblich verbessern. Die Chips können ihren eigenen Betrieb in Echtzeit optimieren, indem sie sich an das Nutzerverhalten und die Umgebungsbedingungen anpassen, um eine bessere Leistung bei geringerer Energieaufnahme zu erzielen. Darüber hinaus ermöglicht die KI die Entwicklung von immer ausgefeilteren Sensoren und IoT-Geräten. Intelligente Sensoren können Daten an Ort und Stelle verarbeiten und blitzschnelle Entscheidungen treffen, ohne Informationen an einen zentralen Server übermitteln zu müssen. Die Fähigkeit des Edge Computings wird bei

Anwendungen, die wie autonome Fahrzeuge und industrielle Automatisierungssysteme Echtzeitreaktionen erfordern, von entscheidender Bedeutung sein. KI steht kurz davor, unsere Geräte der Verbraucherelektronik intuitiver und personalisierter als je zuvor zu machen, was - wenn es mit dem richtigen Maß an KI gehandhabt wird - auch für die Kunden zu Verbesserungen führt. Und KI wird unsere Interaktion mit der Technologie nahtloser und natürlicher machen - durch Smartphones, die unsere Bedürfnisse und Vorlieben vorhersagen können, sowie durch Smart-Home-Systeme, die lernen und sich an unsere tägliche Routine anpassen. Klingt nach Science-Fiction, ist aber eher heute als morgen, je nachdem, wo man hinschaut

KI treibt die Entwicklung in der Elektronikbranche voran und verbessert alle Aspekte des Prozesses, einschließlich Entwurf, Entwicklung, Fertigung, Kundensupport und Innovation. In der Entwurfs- und Entwicklungsphase verbessert die KI die Schaltungsentwürfe, indem sie die Leistung vorhersagt und die Effizienz erhöht. Dies führt zu einer schnelleren Produktentwicklung, da es das Rapid Prototyping und den Bau komplizierter elektrischer Systeme erleichtert. KI-basierte Qualitätskontrollsysteme werden bereits in der Fertigung eingesetzt, um Fehler an den Montagelinien genau zu erkennen, während die vorausschauende Wartung dazu dient, Ausfallzeiten zu vermeiden, indem Geräteausfälle vorhergesagt werden. KI verbessert die Logistik in der Lieferkette, indem sie die Nachfrage genau vorhersagt und Ineffizienzen verringert. In der Consumer-Elektronik sorgt KI für ein personalisiertes

Benutzererlebnis in intelligenten Geräten, indem sie die Interaktionen an das Verhalten und die Vorlieben der Benutzer anpasst. Sie ermöglicht die Automatisierung der Produktion, wodurch die Effizienz der Abläufe gesteigert wird, und befähigt autonome Systeme zur Durchführung komplizierter Aufgaben wie der präzisen Platzierung von Bauteilen. KI spielt aber auch eine entscheidende Rolle im Test- und Validierungsprozess, indem sie diesen automatisiert und so sicherstellt, dass die Bauteile die erforderlichen Spezifikationen erfüllen und Probleme erkannt werden, bevor sie sich negativ auf das Endprodukt auswirken. Sie verbessert die Kundenbetreuung durch den Einsatz von Chatbots und virtuellen Assistenten zur effektiven Bearbeitung von Anfragen und zur Lösung von Problemen und sie erleichtert die Ferndiagnose von Defekten in Elektrogeräten. Das Energiemanagement in der Elektronik wird durch die Optimierung des Stromverbrauchs, die Verlängerung der Batterielebensdauer und die effiziente Verwaltung der Stromverteilung in intelligenten Stromnetzen immer weiter verbessert werden. Nachhaltigkeit wird zur Sache der KI. In der Forschung und Innovation hilft KI bei der Entdeckung neuer Materialien, die die Leistung von Halbleitern verbessern könnten, und treibt die Fortschritte in der Nanotechnologie voran, um kleinere, schnellere und effizientere Komponenten zu entwickeln, und ist damit ein wichtiger Faktor. Die Auswirkungen der KI in diesen Bereichen führen zu einer verbesserten Betriebseffizienz, einer höheren Produktqualität und einem verbesserten Nutzererlebnis, was die Elektronikbranche revolutioniert.

So weit, so wahr. Aber solange die Moral eine Rolle spielt, wird ohne den Menschen nichts möglich sein. Und deshalb ist soziale Präzision ein echter Wettbewerbsvorteil für Elektronikunternehmen, um KI nicht nur in den offensichtlichen Bereichen einzusetzen, sondern durch intelligente und verantwortungsvolle Mensch-Maschine-Interaktion echte Fortschritte zu erzielen. Es gibt so viel, was man mit KI gewinnen kann, aber es ist wichtig, in erster Linie in Menschen zu investieren, um letztendlich wirklich gut darin zu werden, als Unternehmen mit großer Produktivität Geld zu verdienen, die nicht nur aus Automatisierung besteht, sondern aus kollaborativer Wertschöpfung zwischen Mensch und Maschine.

Als Theologe ist das alles für mich eine Herausforderung, für mich als mit der Wirtschaft affinen Menschen ist es spannend (bei legitimen Geschäftsmodellen) und als Low-Coder und Gamer ist es faszinierend. Diese Ambivalenz teile ich mit vielen professionellen Akteuren, aber auch mit Patienten und Angehörigen. Zwischen Hoffen und Bangen. Wie viele andere Branchen wird auch die Elektronik davon profitieren, wenn KI smart im Sinne sozialer Präzision entwickelt und eingesetzt wird.

Meiner Meinung nach ist es nicht so einfach, die Chancen der KI in der Praxis zu realisieren (Tal der Implementationstränen), denn ohne Denken gibt es

kein gutes Prompten, es sei denn, auch das Prompten und letztlich nahezu alles wird von der KI übernommen, was aber eine Abwärtsspirale an Literarität und letztlich Souveränität in Gang setzen würde, was ethisch nicht wünschenswert sein kann. Besser: soziale Präzision im Sinne der ständigen Entwicklung des guten alten Urteilsvermögens - durchaus und gerne im Sinne Kants. Denn nur im klugen, kritischen Zusammenspiel mit immer intimerem, immersivem LLM & Co. (Large Language Model) kann der Mensch seiner Pflicht nachkommen und treu bleiben: moralischer Akteur zu sein, nicht die Verantwortung an andere zu delegieren. Wenn diese Bedingung erfüllt ist, kann und sollte KI eine wichtige Rolle in Medizin und Gesundheit, aber auch in anderen Bereichen spielen. Wir werden nicht alle zu Hyperpromptern werden, aber hoffentlich werden wir alle wieder zu Philosophen im ursprünglichen Sinne: mit Neugier auf echte Menschen, die Welt und uns selbst. Nicht zuletzt weil wirtschaftliche Wertschöpfung für Menschen erdacht wurde (auch wenn je nach Modell fraglich ist, wer zu den „Menschen“ in diesem Allokationssinn wohl zu zählen habe...). Prompt-Engineering (letztlich Maschinensprache neu gedacht) oder ähnliches ist im Grunde eine Fähigkeit, wenn wir erkennen, dass es (wenn sie mehr sein soll als ein simples Tooling...) auf Interaktion ankommt - eine Technomaneutik war von Platon nicht vorgesehen, erscheint aber durchaus sinnvoll. Wenn sie auf einen echten Dialog zwischen Menschen hinarbeitet. Kompetent mit KI umzugehen bedeutet, kompetent mit sich selbst und mit anderen umzugehen. Das ist unsere eigentliche Chance.

Alles andere wäre in der Tat ein tiefgreifendes moralisches Versagen: Die Entwicklung einer echten, empfindungsfähigen künstlichen Intelligenz, die in der Lage ist, moralisch zu handeln, würde nicht nur eine ethische Grenze überschreiten, sondern auch die Verantwortung, die uns Menschen obliegt, technologisch delegierbar machen. Ein Millennium-Workaround. Ein Schleier der KI über den menschlichen Fehlern wird uns nicht helfen, die Herausforderungen des menschlichen Zusammenlebens zu meistern und unsere eigenen moralischen Fähigkeiten zu entwickeln. Unsere moralische Pflicht ist es, unsere eigenen ethischen Herausforderungen anzunehmen und zu lösen, anstatt sie an eine KI zu delegieren. Dies erfordert, dass wir als Gesellschaft in unsere Fähigkeit zu Empathie, moralischer Reflexion und Verantwortung investieren. Der Schwerpunkt sollte auf der Stärkung der menschlichen Fähigkeit zu ethischem Urteil und moralischem Handeln liegen, anstatt zu versuchen, sie durch Technologie zu ersetzen. Gerade deshalb ist es so wichtig, soziale Präzision zu praktizieren, anstatt sich ausschließlich auf datenorientierte digitale Präzision zu verlassen.

In der Praxis ist es eine große Herausforderung, hochgradig immersive digitale Technologien wie interaktive Avatare so zu gestalten und zu nutzen, dass sie zwar kurzfristig Unterstützung bieten, aber langfristig





nicht zu einer Verstärkung von Abhängigkeiten oder einer weiteren Isolation von Menschen führen. KI ist kein bester Kumpel, kein Liebhaber, kein Freund. Zumal der Umkehrschluss zwar nicht notwendig, aber doch naheliegend ist: Wir sollten unsere empathischen Energien auf Menschen richten. Um den ontologischen Unterschied deutlich zu machen, aber auch aus moralischen Gründen. Deswegen sind auch besondere Maßnahmen zur Kompetenzentwicklung notwendig, um den Umgang mit zum Beispiel interaktiven Avataren so zu gestalten, dass wir nicht bewusst oder unbewusst soziale Präzision unter uns verlernen, sondern die digitalen Hochtechnologien nutzen, um uns dabei zu unterstützen, mit Sicherheit, Freude und Zuversicht unsere eigene Menschlichkeit wiederzuentdecken. Dass dies auch ohne KI-Detox möglich ist, verdanken wir der dialektischen Wesenstriade Mensch-Natur-Technik. Die Technik ist nicht einfach das Andere der Natur und damit des Menschen, denn wir sind auch Menschen durch die Technik, mit der wir uns über die Natur erheben. Wesentlich ist dabei, uns nicht unheilvoll gerade wieder an diese Natur zurückzubinden durch Technik - Abhängigkeit bis hin zur Sucht, Bequemlichkeit bis zur Abstumpfung, Aushöhlung der Kompetenz bis hin zum Mangel an kritischer Urteilskraft und Externalisierung der Verantwortung bis hin zur Dilution der eigenen Autonomie -, sondern uns nicht nur über sie lernend, von ihr lernend, sondern mit ihr lernend die vielzitierten Zügel in der Hand hält. So kann es gelingen: indem wir auf soziale Präzision, auf Möglichkeiten, Resilienz und Kultur (wirklich) im Unternehmenskontext setzen - damit ein höherer Mehrwert nicht mit Elend erkaufte wird. Sondern auch das tut, was er gesellschaftlich tun sollte. *Non machinae, sed vitae discimus.* Die Soft Skills (soziale Kompetenzen) von heute sind die Hard Skills von morgen. Menschen in der Wirtschaft überflüssig zu machen wird nur gelingen, wenn eine neue KI-getriebene Automatisierung von einer wertschöpfenden, menschengetriebenen Veränderung der Arbeit begleitet wird. 20 % Arbeitsinsatz für 100 % Wertschöpfung - Kollege KI erzeugt das Delta - wird wohl nicht reichen; eher 20 % Arbeitsinsatz und 500 % Wertschöpfung, weil der Mensch mit der KI wahrscheinlich ein hohes Produktivitätsniveau erreichen dürfte. Natürlich sind dazu neue Formen von beruflicher, ja Lebensbildung erforderlich. Die Entdeckung, Entwicklung und Ausbildung des menschlichen Alleinstellungsmerkmals erfordert, dass wir es ernst nehmen. Jeder von uns muss ein Experte in Sachen Neugier, Experimentierfreude und Mensch-zu-Mensch-zu-Maschine-Kollaboration werden.

## Über den Autor

Prof. Dr. Stefan Heinemann ist Professor für Wirtschaftsethik an der FOM Hochschule und Sprecher der Ethik-Ellipse Smart Hospital an der Universitätsmedizin Essen. Er beschäftigt sich mit der ökonomischen und ethischen Perspektive auf die digitale Medizin und die Gesundheitswirtschaft. Heinemann ist wissenschaftlicher Leiter des Hauptstadtkongress Lab (Springer Medizin, Wiso). Er ist Leiter der Forschungsgruppe „Ethik der digitalen Gesundheitswirtschaft & Medizin“ am ifgs-Institut für Gesundheit & Soziales der FOM Hochschule, Mitglied der „Arbeitsgruppe KI in der Inneren Medizin“ im Rahmen der Kommission „Digitale Transformation der Inneren Medizin“ sowie Fachberater in verschiedenen Forschungs- und Bildungseinrichtungen. Darüber hinaus ist der Philosoph und Theologe Mitglied des wissenschaftlichen Beirats „Digitale Transformation“ der AOK Nordost, Mitglied des wissenschaftlichen Beirats des Arbeitskreises „Gendersensible Medizin“ an der Medizinischen Fakultät der Universität Bielefeld, Mitglied des Beirats des Instituts für Patient Experience der Universitätsmedizin Essen und Mitglied des sozial- und gesundheitspolitischen Beirats der BARMER-Landesvertretung Nordrhein-Westfalen. Außerdem ist er im Vorstand der Kölner Wissenschaftsrunde, Vorstandsvorsitzender der „Science City Essen“ und Mitglied des Kuratoriums von sneep e. V., einem studentischen Netzwerk für Wirtschafts- und Unternehmensethik. Heinemann ist Mitinitiator von [www.dataprotection-landscape.com](http://www.dataprotection-landscape.com), einer Plattform für die Mehrdimensionalität des Datenschutzes.

Digitale Realitäten entfalten sich in sozialer Konkretheit. Zwischen dystopischer Phantomologie im Stile von Stanislaw Lem und utopischer Tech-Euphorie gibt es einen dritten Weg: KI-Entwicklung und -Nutzung verantwortungsvoll gestalten, soziale Präzision. Mit anderen Worten: Bildung, die den Klassikern gerecht wird, Befähigung zur Veränderung, Stärkung der Resilienz... aber diesmal wirklich. Künstliche Autonomie lässt sich nicht nebenbei verwalten. Die größte Chance auf dem gesamten Gebiet der künstlichen Intelligenz ist nicht die Technologie selbst, sondern die Beziehung, die wir bereit sind, mit ihr zu haben. Oder das größte Risiko. ◀

Dieser Artikel enthält Teile eines zuvor auf Deutsch erschienenen Artikels [1].

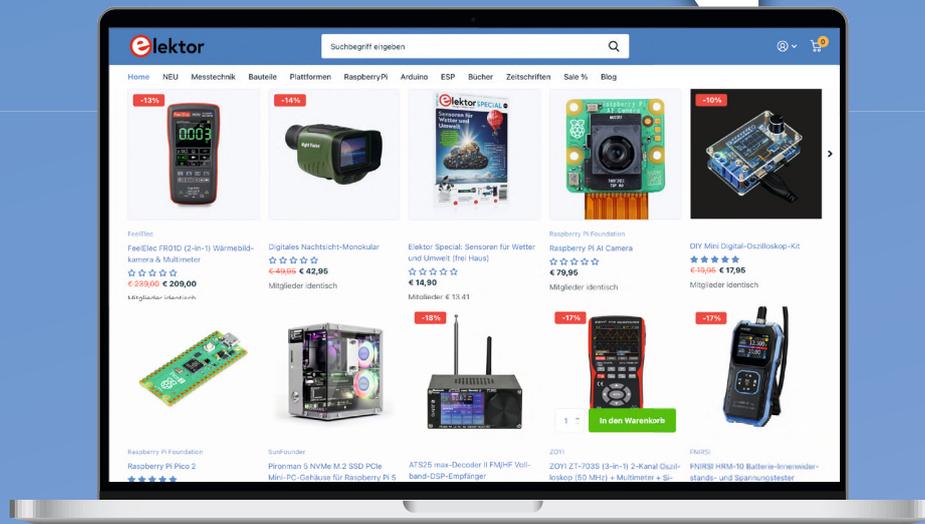
Übersetzung: SG — 240522-02



## WEBLINKS

- [1] Stefan Heinemann, „Scheitert die KI-Revolution in Unternehmen an fehlender sozialer Präzision?“, 10xD Magazin 09/24: <https://10xd.de/magazin/kuenstliche-intelligenz-im-alltag/scheitert-die-ki-revolution-in-unternehmen-an-fehlender-sozialer-praezision>

# Was ist Ihre Meinung?



Bei Elektor bieten wir mehr als nur Elektronik - wir schaffen ein Erlebnis, von Qualitätsprodukten und außergewöhnlicher Kundenbetreuung, unterstützt durch die Leidenschaft unserer Community.

Teilen Sie Ihre Meinung auf  
[www.elektor.de/pages/customer-reviews](http://www.elektor.de/pages/customer-reviews)



**Günther A.**

Das Meßgerät wurde kurzfristig geliefert. Solide Ausführung, gut ablesbares Display, zuverlässige Kontaktierung der 4-Leiter Kelvin Meßleitungen, hohe Genauigkeit der Meßwerte (Spannung, Innenwiderstand)



**Matthias D.**

Ich hatte die Elektor schon in der 80ern und dann etwas aus den Augen verloren. Es ist wie damals: prima Schaltungsideen, anwendbar, sehr gut erklärt, alles bestens. Danke und bitte weiter so.



**Hartmut W.**

Sehr guter Service!

Schnelle Lieferung, gute Bezahlmöglichkeit, Gerät top. Sehr zu Empfehlen! Werde weiter Bestellen.

# Treten Sie jetzt der Elektor Community bei!



Jetzt



Mitglied werden!



- ✓ Zugang zum kompletten Online-Archiv (1970-heute)!
- ✓ 8x Elektor Magazin (gedruckt)
- ✓ 8x Elektor Magazin (PDF)
- ✓ 10% Rabatt im Elektor Store und exklusive Angebote
- ✓ Zugriff auf über 5.000 Gerber-Dateien u.v.m. aus der Projektplattform Elektor Labs



## Auch erhältlich

Die digitale  
Mitgliedschaft!



- ✓ Zugang zum kompletten Online-Archiv
- ✓ 8x Elektor Magazin (PDF)
- ✓ 10% Rabatt im Elektor Store und exklusive Angebote
- ✓ Zugriff auf über 5.000 Gerber-Dateien u.v.m. aus der Projektplattform Elektor Labs



[www.elektormagazine.de/abo](http://www.elektormagazine.de/abo)