Sachbericht zum Vorhaben
16 IS 22068 – EQUIPE

Teil II

ZE	Karlsruher Institut für Technologie (KIT)
Förderkennzeichen	16 IS 22068
Vorhabenbezeichung	EQUIPE
Laufzeit des Vorhabens	09/2022 – 08/2025

Name	**Dr. Charlotte Debus**
Adresse	**Karlsruher Institut für Technologie (KIT)** Steinbuch Centre for Computing (SCC) Hermann-von-Helmholtz-Platz 1 76344 Eggenstein-Leopoldshafen
Telefon	+49 721 608-29718
E-Mail	charlotte.debus@kit.edu

1 Überblick

Ziel des Projekts **EQUIPE** (Efficient Uncertainty Quantification for Probabilistic Transformers) war es, effiziente und vertrauenswürdige Methoden zur Unsicherheitsquantifizierung (UQ) für moderne Transformer-Architekturen zu entwickeln und für datengetriebene Zeitreihenprognosen nutzbar zu machen. Im Mittelpunkt stand die Frage, wie leistungsfähige tiefe neuronale Netze und insbesondere Transformer-basierte Modelle so erweitert werden können, dass sie nicht nur Punktvorhersagen liefern, sondern auch verlässliche Unsicherheitsabschätzungen, ohne dabei prohibitiv hohe Rechenkosten zu verursachen.

Transformer-Modelle haben sich in den letzten Jahren als dominierende Architektur in vielen Bereichen des maschinellen Lernens etabliert. Ihr Einsatz in der Zeitreihenvorhersage – etwa in Energie-, Klima- oder Industrieanwendungen – ist jedoch häufig mit zwei zentralen Herausforderungen verbunden: dem hohen Rechenaufwand und der fehlenden systematischen Quantifizierung von Unsicherheit. Gerade in sicherheitskritischen oder wirtschaftlich relevanten Anwendungen ist jedoch eine transparente und kalibrierte Unsicherheitsinformation essenziell.

Vor diesem Hintergrund verfolgte EQUIPE drei zentrale Zielsetzungen:

1. **Methodische Entwicklung effizienter UQ-Verfahren für Transformer:** Es sollen Bayes'sche und approximative probabilistische Methoden (z. B. Variational Inference, Monte-Carlo-Verfahren) so weiterentwickelt werden, dass sie auf großskalige Transformer-Modelle anwendbar und zugleich recheneffizient sind.

2. **Reduktion des Rechenaufwands durch strukturelle Modelloptimierung:** Neben der probabilistischen Modellierung werden auch architektonische Optimierungen (z. B. Sparse-Strukturen, Pruning, zyklische Residual-Architekturen) untersucht, um die Skalierbarkeit zu verbessern und Energieverbrauch zu reduzieren.

3. **Anwendung auf reale Zeitreihendaten und Validierung der Unsicherheitsqualität:** Die entwickelten Methoden werden auf realen Datensätzen – insbesondere aus Klima- und Energiekontexten – evaluiert. Dabei spielen Kalibrierungsmetriken und robuste Evaluationsverfahren eine zentrale Rolle.

Ein wesentliches Element des Projekts war die Verbindung von probabilistischer Modellierung mit Hochleistungsrechnen (HPC). Sampling-Verfahren, die für Bayes'sche Modelle notwendig sind, verursachen typischerweise erhebliche Mehrkosten. EQUIPE adressierte daher explizit Fragen der Parallelisierung und Skalierung, um Unsicherheitsquantifizierung auch für große Modelle praktikabel zu machen.

Darüber hinaus verfolgte das Projekt einen starken Software- und Infrastrukturansatz, mit dem Ziel Werkzeuge zu entwickeln, die es ermöglichen, deterministische neuronale Netze mit geringem Aufwand in probabilistische Modelle zu überführen. Dadurch soll die Eintrittshürde für Forschende und industrielle Anwender gesenkt und die Verbreitung vertrauenswürdiger KI-Methoden gefördert werden.

Insgesamt positionierte sich EQUIPE an der Schnittstelle von KI-Methodenentwicklung, Hochleistungsrechnen und anwendungsorientierter Zeitreihenanalyse. Das Projekt trägt dazu bei, moderne Deep-Learning-Modelle transparenter, robuster und energieeffizienter zu machen und stärkt damit die wissenschaftliche und technologische Kompetenz im Bereich vertrauenswürdiger KI in Deutschland.

2 Inhaltliche Arbeiten

Das Projekt EQUIPE hat wesentliche methodische Fortschritte in der Quantifizierung von Unsicherheiten für Transformer-Modelle erzielt. Zentrale Beiträge umfassen die Entwicklung eines effizienten Transformermodells, die Identifikation von Sampling als dominierendem Rechenbottleneck, die systematische Analyse der erforderlichen Sampleanzahl, die Entwicklung skalierbarer Sampling-Parallelismus-Methoden sowie neue Erkenntnisse zur Sparsity in Bayesschen neuronalen Netzen.

Der Arbeitsplan gliederte sich in fünf Arbeitspakete mit klar definierten Meilensteinen und Projektergebnissen. Im Verlauf des Projekts wurden diese Ziele weitgehend erreicht, teilweise methodisch angepasst und in einzelnen Fällen bewusst neu ausgerichtet, wenn sich im Forschungsprozess wissenschaftlich fundierte Alternativen als sinnvoller erwiesen. Mehrere ursprünglich geplante Projektergebnisse wurden bewusst angepasst oder durch wissenschaftlich fundierte Alternativen ersetzt. Insgesamt wurden die inhaltlichen Projektziele erreicht und in zentralen Punkten substanziell erweitert.

2.1 Quantifizierung von Unsicherheiten in Transformer-Netzwerken

Das erste Arbeitspaket hatte die Aufgabe, den Stand der Forschung zur Quantifizierung von Unsicherheiten in neuronalen Netzen für Klassifikation und Regression umfassend zu analysieren und die gewonnenen Erkenntnisse auf Transformer-Netzwerke für Zeitreihenvorhersagen zu übertragen. Zu Beginn des Projekts wurde ein eigenes Transformermodell für die Zeitreihenvorhersage entwickelt, der sogenannte Residual Cyclic Transformer (ReCycle) [1]. Die wesentliche Innovation von ReCycle waren die Konzepte der Primary Cycle Compression und des Residual Forecasting (vgl. Abbildung 1), welche eine signifikanten Steigerung der Vorhersagegenauigkeit bei gleichzeitig drastisch reduziertem Rechenaufwand ermöglichen. Dadurch konnte das Training von Transformern für Zeitreihenanwendungen auf Standard-Entwicklerhardware durchgeführt werden, was ursprünglich nicht erwartet worden war. Diese frühe methodische Innovation beeinflusste die weitere Projektentwicklung maßgeblich.

Parallel dazu wurde eine intensive und kontinuierliche Literaturrecherche zu Methoden der Unsicherheitsquantifizierung durchgeführt. Anders als ursprünglich geplant stellte sich heraus, dass die Einarbeitung in den State-of-the-Art kein abgeschlossener Initialschritt sein konnte, sondern ein iterativer Prozess über die gesamte Projektlaufzeit hinweg blieb. Untersucht wurden

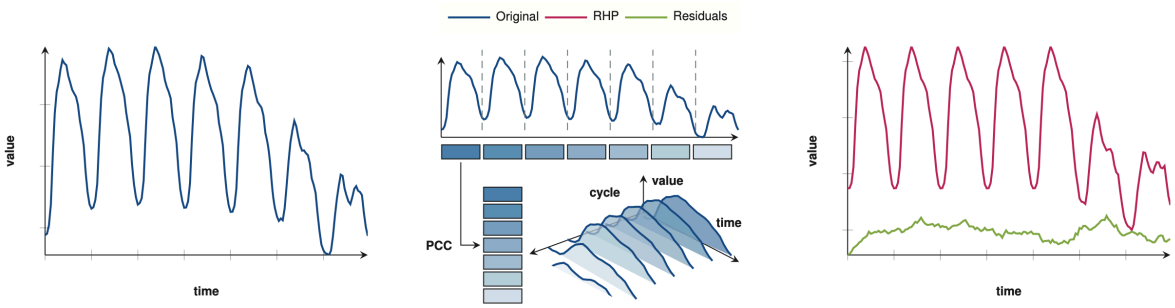

Abbildung 1: Die Konzepte der Primary Cycle Compression (PCC) und des Lernens von Residuen: Zunächst wird die ursprüngliche univariate Zeitreihe (links) entsprechend ihren Primärzyklen neu angeordnet, wodurch eine zweidimensionale Datenmatrix (Mitte) entsteht. Aufgrund der Ähnlichkeit der Primärzyklen können Profile der jüngeren Historie (RHP) berechnet und von den Originaldaten subtrahiert werden. Dadurch entstehen Residuen, die das Modell zu lernen trainiert wird (rechts).

Ensemble-Methoden wie Monte Carlo Dropout [2], Bayessche neuronale Netze mit Variational Inference (VI) [3], Quantile Regression, Normalizing Flows sowie neuere Ansätze wie Diffusionsmodelle, die insbesondere in der datengetriebenen Wettervorhersage Anwendung finden. Dabei zeigte sich, dass Variational Inference einen praktikablen Kompromiss zwischen Genauigkeit der Unsicherheitsabschätzung und rechnerischer Effizienz darstellt, während samplingbasierte Markov-Chain-Methoden zwar teilweise bessere Resultate liefern, jedoch in realistischen Anwendungen nicht skalierbar sind.

Die ursprünglich geplante Veröffentlichung eines Survey-Artikels wurde nicht umgesetzt, da kurz nach Projektbeginn ein thematisch sehr ähnlicher Übersichtsartikel erschien [4]. Dennoch wurden die Erkenntnisse systematisch dokumentiert und bildeten die Grundlage für alle weiteren Arbeiten. Der entsprechende Meilenstein zur Etablierung des State-of-the-Art wurde somit inhaltlich erreicht, auch wenn das geplante Publikationsergebnis nicht realisiert wurde.

Im weiteren Verlauf wurden Monte Carlo Dropout und Variational Inference exemplarisch für das entwickelte Transformer-Modell implementiert. Dabei zeigte sich, dass sich beide Methoden grundsätzlich sowohl für Klassifikation als auch für Regression eignen. Bei Variational Inference unterscheidet sich die Behandlung beider Aufgaben im Wesentlichen nur durch die Wahl der Lossfunktion im data-fitting-Term. Dadurch entfielen ursprünglich geplante Arbeiten zur Anpassung von Gewichtsskalierungen.

Die Implementierung probabilistischer Transformer in dem ursprünglich vorgesehenen Framework Pyro erwies sich als technisch problematisch und fehleranfällig. Aus diesem Grund wurde ein eigenes Software-Framework entwickelt, das unter dem Namen *torch-blue* veröffentlicht wurde [5]. Dieses Framework implementiert Variational Inference strikt entlang des mathematischen Formalismus, verbirgt jedoch die komplexen Berechnungen auf Low-Level-Ebene vor den Nutzenden. Ein Beispiel für die Programmierung mit torch-blue ist in Abbildung 2 gezeigt. Die Programmierschnittstelle wurde bewusst an PyTorch angelehnt, wie in Abbildung 2 dargestellt,

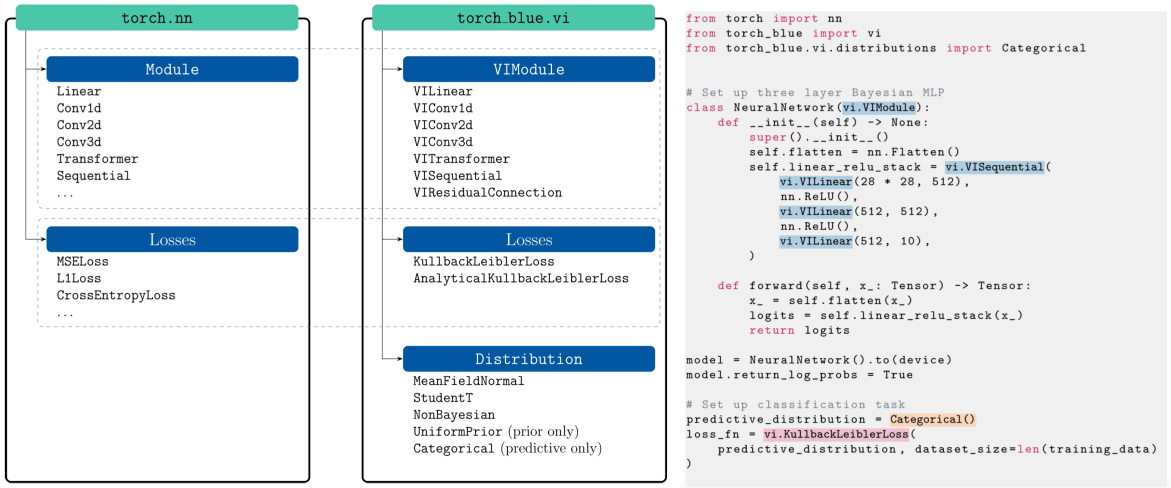

Abbildung 2: Links: Überblick über die Hauptfunktionalität von torch-blue sowie die nicht-Bayes'schen Pendants in PyTorch. Rechts: Code-Beispiel für ein Bayesian MLP mit 3 versteckten Schichten und Cross-Entropy Loss in torch_blue.

um eine nahtlose Integration zu ermöglichen. Besonders hervorzuheben ist die AutoConvert-Funktionalität, mit der sich deterministische neuronale Netze mittels eines Wrappers automatisch in vollständig Bayes'sche Modelle überführen lassen. Damit wurde das ursprünglich geplante Software-Ergebnis nicht nur erreicht, sondern funktional deutlich erweitert.

2.2 Skalierbare Methoden zur Quantifizierung von Unsicherheiten

Das zweite Arbeitspaket zielte auf die Entwicklung skalierbarer Methoden zur Unsicherheitsquantifizierung ab. Im Verlauf der Arbeiten wurde deutlich, dass samplingbasierte Verfahren das zentrale Rechenbottleneck darstellen. Die Approximation der prädiktiven Verteilung erfolgt durch wiederholtes Ziehen von Gewichtsstichproben und entsprechende Vorwärtsdurchläufe, was den Rechenaufwand erheblich erhöht (siehe z.B. Abbildung 3).

Im Rahmen einer systematischen Studie wurde untersucht, wie viele Samples während des Trainings tatsächlich erforderlich sind, um zuverlässige Vorhersagen und Unsicherheitsabschätzungen zu erhalten. Dabei konnte erstmals ein fundamentaler Unterschied zwischen Klassifikation und Regression herausgearbeitet werden. Während für Klassifikationsaufgaben ein einzelnes Sample ausreicht, erfordert Regression typischerweise zwischen acht und sechzehn Samples, wobei sich zehn als praktikabler Standardwert erwiesen haben. Diese Ergebnisse werden derzeit im Review-Prozess bei der UAI begutachtet [6].

Darüber hinaus wurde ein Algorithmus für verteiltes Sampling (siehe Abbildung 4) entwickelt, der paralleles Rechnen effizient nutzt und sowohl für Monte Carlo Dropout als auch für Bayessche neuronale Netze mit Variational Inference geeignet ist. Die entsprechende Publikation wurde auf der PASC-Konferenz angenommen [7]. Der ursprünglich geplante explizite Modellparallelismus für Bayessche Transformer wurde nicht separat verfolgt, da sich Sampling-Parallelismus

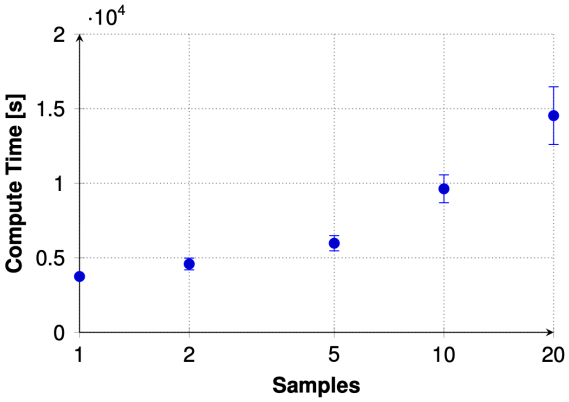

Abbildung 3: Trainingszeit für einen Vision Transformer (ViT) auf dem CIFAR10 Datensatz in Abhängigkeit der gezogenen Samples für das Training mit VI.

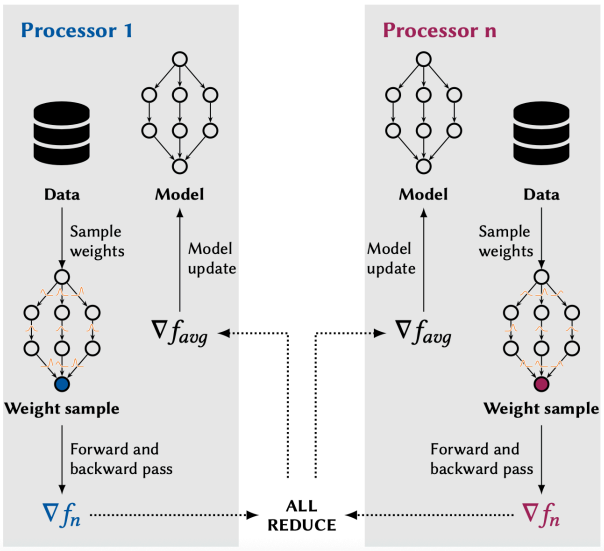

Abbildung 4: Das Konzept von Sample Parallelsimus, am Beispiel von BNNs mit VI.

als effektiver und flexibler Ansatz erwies. Wir haben in einem unabhängigen Projekt zusätzlich ein generalisierter Modellparallelismus-Ansatz entwickelt [8], der sich mit dem Sampling-Parallelismus kombinieren lässt und Speicherengpässe adressiert.

2.3 Probabilistische Betrachtung von Zeitreihenvorhersagen

Im dritten Arbeitspaket stand die Untersuchung der Kalibrierung probabilistischer Modelle im Mittelpunkt. Die Kalibrierung wurde systematisch für Klassifikations- und Regressionsaufgaben analysiert, wobei MACE für Klassifikation und CRPS für Regression als zentrale Metriken verwendet wurden. Dabei zeigte sich, dass die Kalibrierung stark vom jeweiligen Anwendungsfall und von Implementierungsdetails abhängt und nicht pauschal einer bestimmten Methode zugeschrieben werden kann.

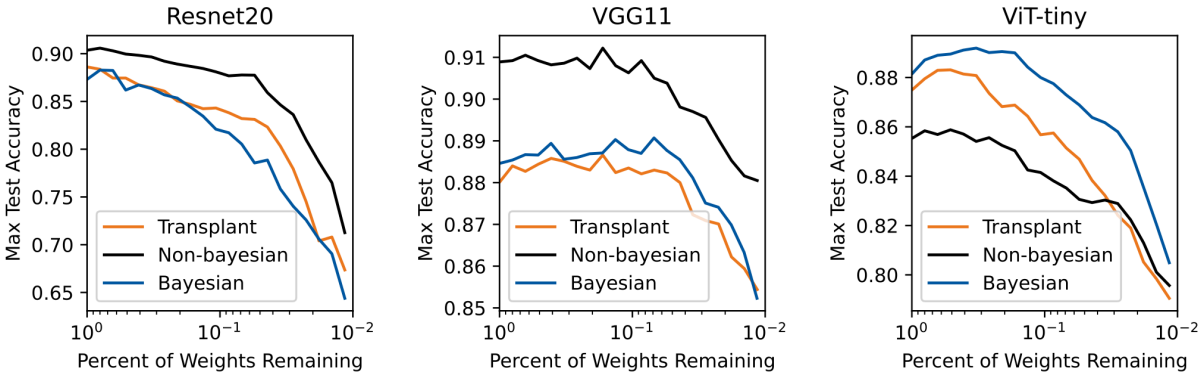

Abbildung 5: Vorhersagegenauigkeit für dünnbesetzte neuronale Netze, Baye'sche und nicht-Bayessche Variante im Vergleich, für verschiedene Bildklassifikationsmodelle.

Die ursprünglich geplante Veröffentlichung synthetischer Datensätze erwies sich als nicht notwendig, da Kalibrierungsmetriken auch auf realen Datensätzen ohne bekannte a-priori-Unsicherheit sinnvoll eingesetzt werden können. In Studien zur Analyse prädiktiver Verteilungen wurden zwar synthetische Daten verwendet, diese konnten jedoch flexibel generiert werden, sodass keine dauerhafte Bereitstellung als separates Projektergebnis erforderlich war.

Teil-probabilistische Transformer-Netzwerke wurden nicht implementiert, da sich zeigte, dass solche Ansätze mit dem mathematischen Rahmen der Variational Inference nicht konsistent vereinbar sind. Auch Untersuchungen zum Weight Rescaling bei Monte Carlo Dropout wurden nicht weiterverfolgt, da frühzeitig auf Gaussian Weight Dropout umgestellt wurde, wodurch das zugrunde liegende Problem entfiel.

2.4 Dünnbesetzte bayessche Transformer-Netzwerke

Im vierten Arbeitspaket wurde untersucht, inwieweit Sparsity zur Reduktion des Rechenaufwands probabilistischer Transformer genutzt werden kann. Zunächst wurden Pruning-Studien an Zeitreihen-Transformern durchgeführt [9]. Anschließend wurde die Übertragbarkeit der Lottery Ticket Hypothesis auf Bayes'sche neuronale Netze mit Variational Inference analysiert. Dabei konnte gezeigt werden, dass auch Bayes'sche Netze dünnbesetzte Subnetzwerke enthalten, die theoretisch isoliert trainierbar wären. Gleichzeitig wurde ein grundlegender Unterschied zwischen Transformern und Convolutional Neural Networks festgestellt. Die entsprechende Arbeit befindet sich aktuell im Review-Prozess bei der UAI [10].

Untersuchungen zum Energieverbrauch zeigten, dass gegenwärtige Softwarebibliotheken wie PyTorch keine echte hardwareseitige Ausnutzung von Sparsity ermöglichen, da Berechnungen weiterhin auf dichten Matrizen ausgeführt werden. Daher wurden umfassende Benchmarkstudien zum Energieverbrauch nicht weiterverfolgt. Stattdessen wird in einem Folgeprojekt an der Integration echter sparse linear algebra auf GPUs gearbeitet [11].

2.5 Öffentlichkeitsarbeit, Verwertung und strukturelle Wirkung

Die Öffentlichkeitsarbeit stellte über die gesamte Projektlaufzeit hinweg einen integralen Bestandteil des Vorhabens dar und ging deutlich über die reine Dissemination wissenschaftlicher Ergebnisse hinaus. Die im Projekt entwickelten Methoden und Softwarewerkzeuge wurden kontinuierlich auf nationalen und internationalen Fachkonferenzen präsentiert. Insgesamt sind aus dem Vorhaben sechs wissenschaftliche Publikationen hervorgegangen, die entweder bereits peer-reviewed veröffentlicht wurden oder sich aktuell im Begutachtungsprozess befinden. Die Beiträge wurden überwiegend auf renommierten Konferenzen im Bereich Künstliche Intelligenz, maschinelles Lernen und Hochleistungsrechnen eingereicht und vorgestellt.

Besondere Sichtbarkeit erlangte das entwickelte Software-Framework torch-blue, das als quelloffenes Projekt unter Beachtung der FAIR-Prinzipien öffentlich zugänglich gemacht wurde. Die Veröffentlichung als Open-Source-Software auf der GitHub-Plattform der Forschungsgruppe ermöglichte nicht nur Transparenz und Reproduzierbarkeit der Forschungsergebnisse, sondern erleichterte auch die Weiterverwendung durch andere wissenschaftliche Arbeitsgruppen. Torch-blue wurde unter anderem auf der Helmholtz AI Conference 2025 sowie beim Helmholtz UQ Workshop in Dresden vorgestellt und wird darüber hinaus auf der PyTorch Conference Europe 2026 in Paris einer breiten internationalen Entwickler-Community präsentiert werden. Damit adressiert die Öffentlichkeitsarbeit nicht nur die wissenschaftliche Community im engeren Sinne, sondern auch Praktikerinnen und Praktiker aus dem Bereich des Machine Learning Engineerings.

Neben klassischen Fachkonferenzen wurde gezielt der Austausch mit Anwendern und Industrievertretern gesucht. Kooperationen mit Unternehmen wie der Siemens AG und NVIDIA ermöglichten den Transfer wissenschaftlicher Erkenntnisse in anwendungsnahe Kontexte und stärkten zugleich die industrielle Anschlussfähigkeit der entwickelten Methoden. Auch wenn keine unmittelbaren Schutzrechtsanmeldungen aus dem Projekt hervorgegangen sind, wurde durch diese Kooperationen ein wichtiger Beitrag zur Vernetzung von Grundlagenforschung und industrieller Praxis geleistet.

Darüber hinaus engagierte sich die Gruppenleitung intensiv in der akademischen Selbstverwaltung und im wissenschaftspolitischen Diskurs. Sie wurde in mehrere universitäre Gremien gewählt, darunter der Fakultätsrat für Informatik sowie der Bereichsrat des KIT-Bereichs II, und wirkte in Berufungs- und Findungskommissionen mit. Diese Aktivitäten erhöhten nicht nur die institutionelle Sichtbarkeit der Nachwuchsgruppe, sondern stärkten auch die strategische Verankerung des Themas Unsicherheitsquantifizierung innerhalb des KIT.

Auf nationaler Ebene wurde die Gruppenleitung wiederholt als Expertin für Vorträge und Diskussionsrunden eingeladen. Sie hielt unter anderem eine Keynote auf der Statustagung der Gauss-Allianz sowie auf der General Assembly des BMFTR-Projekts WarmWorld und war Gastrednerin bei der GreenICT-Fachtagung des BMFTR sowie beim Deutschen Wetterdienst. Diese Beiträge trugen dazu bei, die Bedeutung der Unsicherheitsquantifizierung für gesellschaftlich relevante Anwendungen wie Energie- und Klimavorhersage hervorzuheben und die Rolle mo-

derner KI-Methoden im öffentlichen Diskurs zu reflektieren.

Ein weiterer Schwerpunkt der Öffentlichkeitsarbeit lag auf der Nachwuchsförderung und der Sichtbarmachung von KI-Forschung für Studierende und junge Wissenschaftlerinnen und Wissenschaftler. Die Gruppe organisierte interne Journal Clubs, Seminare und methodische Workshops und beteiligte sich an Initiativen zur Förderung von Frauen in der Informatik und den Datenwissenschaften. Zudem wurde die Forschung der Gruppe regelmäßig über Webauftritte und soziale Medien kommuniziert, um Transparenz zu schaffen und potenzielle Kooperationspartner sowie Nachwuchskräfte anzusprechen.

Insgesamt hat die Öffentlichkeitsarbeit wesentlich dazu beigetragen, die im Projekt erzielten wissenschaftlichen Fortschritte sichtbar zu machen, die nationale und internationale Vernetzung zu stärken und die gesellschaftliche Relevanz von Forschung zur Quantifizierung von Unsicherheiten in KI-Systemen zu unterstreichen.

3 Positionen des zahlenmäßigen Nachweises

Der überwiegende Teil der im Projekt EQUIPE angefallenen Mittel wurde für Personalausgaben verwendet. Aus dem Projektbudget wurden die Gruppenleitung, Dr. Charlotte Debus, sowie zwei Doktoranden, A. Weyrauch und Nicholas Kuhn (geb. Kiefer), finanziert. Für die wissenschaftliche Arbeit der beiden Doktoranden wurden – wie im Projektplan vorgesehen – leistungsfähige Hochleistungs-Entwicklerlaptops beschafft, deren Rechen- und Speicherkapazitäten deutlich über die am KIT übliche Grundausstattung hinausgingen und die insbesondere für das Training und die Evaluation komplexer neuronaler Netze erforderlich waren. Die ursprünglich eingeplante Anschaffung separater Dockingstations konnte entfallen, da im Rahmen des SCC-weiten New-Work-Projekts und der Einführung von Flex-Arbeitsplätzen Monitore mit integrierter Dockingfunktion bereitgestellt wurden. Dadurch ergaben sich Einsparungen, ohne dass funktionale Einschränkungen entstanden.

Darüber hinaus wurden zwei studentische Hilfskräfte, A. Özdemir und M. Mayer, aus Projektmitteln finanziert, die insbesondere bei Literaturrecherchen, experimentellen Vorarbeiten sowie unterstützenden Entwicklungsaufgaben mitwirkten. Um zu Beginn des Projekts entstandene Verzögerungen bei der Besetzung einer Doktorandenstelle und die daraus resultierenden temporären Mittelüberschüsse sinnvoll zu nutzen, wurde zudem für einen Zeitraum von sechs Monaten ein Postdoc, Oskar Taubert, im Bereich Research Software Engineering beschäftigt. Dies trug wesentlich dazu bei, die Softwareentwicklung zu beschleunigen und die technische Infrastruktur der Gruppe nachhaltig zu stärken.

Neben den Personalkosten fielen Reisekosten für die Teilnahme an nationalen, europäischen und in einem Fall auch internationalen wissenschaftlichen Konferenzen an, auf denen die Projektergebnisse präsentiert wurden. Hierzu zählten sowohl reguläre Konferenzbeiträge als auch eingeladene Vorträge der Gruppenleitung bei externen Fachveranstaltungen. Diese Reisen dienten nicht nur der wissenschaftlichen Dissemination, sondern auch der Vernetzung mit interna-

tionalen Forschungspartnern und potenziellen Kooperationspartnern aus Wissenschaft und Industrie.

4 Notwendigkeit und Angemessenheit

Die im Rahmen des Projektes durchgeführten Forschungsarbeiten und verwendeten Ressourcen entsprachen weitestgehend der im Projektantrag formulierten Planung. Im Falle von Abweichungen konnten diese auf Änderungen der Schwerpunktsetzung zurückgeführt werden. Unabhängig davon wurden alle wesentlichen Punkte des Arbeitsplans erfolgreich abgeschlossen. Somit war die geleistete Arbeit angemessen und notwendig.

5 Voraussichtlicher Nutzen

Der voraussichtliche Nutzen des Projekts EQUIPE liegt in der nachhaltigen Stärkung der methodischen Grundlagen zur Quantifizierung von Unsicherheiten in modernen KI-Systemen und insbesondere in Transformer-basierten Modellen für Zeitreihenvorhersagen. Durch die systematische Analyse, Weiterentwicklung und Skalierung probabilistischer Methoden wurde ein Beitrag zur Erhöhung der Verlässlichkeit datengetriebener Vorhersagemodelle geleistet. Dies ist insbesondere in sicherheitskritischen und gesellschaftlich relevanten Anwendungsfeldern wie der Energieversorgung, der Klima- und Wettervorhersage sowie der industriellen Prozesssteuerung von zentraler Bedeutung, da dort nicht nur Punktprognosen, sondern auch belastbare Aussagen über deren Unsicherheit erforderlich sind.

Die im Projekt entwickelten methodischen Ansätze, darunter effiziente Sampling-Strategien, skalierbarer Sampling-Parallelismus sowie das Software-Framework torch-blue zur niederschwelligen Implementierung bayesscher neuronaler Netze, ermöglichen es, Unsicherheitsquantifizierung auch für größere und komplexere Modelle praktikabel einzusetzen. Damit werden wesentliche technische Hürden reduziert, die einer breiten Anwendung probabilistischer KI-Methoden bislang entgegenstanden. Durch die Open-Source-Veröffentlichung der Software und die Publikation der wissenschaftlichen Ergebnisse ist sichergestellt, dass die erarbeiteten Konzepte unmittelbar von anderen Forschungsgruppen aufgegriffen und weiterentwickelt werden können.

Darüber hinaus trägt das Projekt zur langfristigen Stärkung der KI-Kompetenz in Deutschland bei, indem es Expertise im Bereich probabilistischer Deep-Learning-Methoden aufbaut und konsolidiert. Die im Rahmen des Projekts etablierte Nachwuchsgruppe bildet eine strukturelle Grundlage für weiterführende Forschungsvorhaben und Kooperationen im Bereich Unsicherheitsquantifizierung. Insgesamt schafft EQUIPE somit sowohl wissenschaftlichen als auch strukturellen Mehrwert und leistet einen Beitrag zur Entwicklung vertrauenswürdiger, transparenter und skalierbarer KI-Systeme.

6 Fortschritt anderer Stellen

Abgesehen von der Veröffentlichung des Survey Papers zu UQ Methoden wurde kein Fortschritt bekannt, der den Projektablauf maßgeblich beeinflusst hat.

7 Erfolgten oder geplanten Veröffentlichungen

Im Rahmen des Projekts EQUIPE sind mehrere wissenschaftliche Publikationen entstanden beziehungsweise befinden sich aktuell im Begutachtungsprozess. Die Arbeiten decken die zentralen methodischen Entwicklungen des Projekts ab und adressieren sowohl grundlegende Fragestellungen der Unsicherheitsquantifizierung als auch Aspekte der Skalierbarkeit und Effizienz.

1. Weyrauch, A., Steens, T., Taubert, O., Hanke, B., Eqbal, A., Götz, E., ... & Debus, C. (2024, June). Recycle: Fast and efficient long time series forecasting with residual cyclic transformers. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 1187-1194). IEEE.
   Diese Publikation, vorgestellt auf der IEEE Conference on Artificial Intelligence 2024, präsentiert mit ReCycle eine neuartige Transformer-Architektur für die effiziente Vorhersage langer Zeitreihen. Das Modell kombiniert residuale und zyklische Strukturen, um sowohl die Vorhersagegenauigkeit zu verbessern als auch den Rechenaufwand deutlich zu reduzieren. Die Ergebnisse zeigen, dass ReCycle im Vergleich zu bestehenden Transformer-Ansätzen eine konkurrenzfähige oder bessere Prognoseleistung bei gleichzeitig signifikant geringerer Trainings- und Inferenzzeit erreicht.

2. Weyrauch, A., Heyen, L. H., Muriedas, J. P. G. H., Hsia, P. H., Özdemir, A. K., Streit, A., ... & Debus, C. (2026). torch_blue: A Flexible Python Package for Bayesian Neural Networks in PyTorch. Journal of Open Source Software, 11(117), 9415.
   Diese Publikation beschreibt das Software-Framework torch-blue, und ist im Journal of Open Source Softwareerschienen. torch-blue ein Open-Source-Framework zur einfachen und mathematisch konsistenten Implementierung bayesscher neuronaler Netze mit Variational Inference in PyTorch vor. Ziel der Arbeit ist es, die praktischen Hürden beim Einsatz probabilistischer Deep-Learning-Methoden zu reduzieren, indem eine PyTorch-nahe API bereitgestellt wird, die eine nahtlose Integration in bestehende Modelle erlaubt. Ein zentrales Merkmal ist die AutoConvert-Funktion, mit der deterministische Netzwerke automatisiert in bayessche Modelle überführt werden können, während die komplexen Berechnungen der Variational Inference auf Implementierungsebene verborgen bleiben. Damit trägt torch-blue zur Verbreitung, Standardisierung und praktischen Anwendbarkeit von Unsicherheitsquantifizierung in modernen neuronalen Netzen bei.

3. Weyrauch, A., Heyen, L. H., Kuhn, N., Streit, A., Götz, M., & Debus, C. (2026). On the Scaling of Predictive Samples in Stochastic Variational Inference for Bayesian Neural Networks. Submitted to Conference on Uncertainty in Artificial Intelligence (UAI)

Diese Publikation widmet sich der systematischen Untersuchung der notwendigen Anzahl von Sampling-Durchläufen in bayesschen neuronalen Netzen und Monte-Carlo-basierten Verfahren. In dieser Arbeit wird erstmals klar zwischen Klassifikations- und Regressionsaufgaben unterschieden und gezeigt, dass für Klassifikation in der Regel ein einzelnes Sample ausreicht, während für Regression mehrere Samples erforderlich sind. Diese Ergebnisse liefern eine praxisrelevante Richtlinie zur Reduktion des Rechenaufwands probabilistischer Modelle. Die Arbeit befindet sich derzeit im Review-Prozess bei der Konferenz Conference on Uncertainty in Artificial Intelligence (UAI).

4. Özdemir, A. K., Heyen, L. H., Weyrauch, A., Streit, A., Götz, M., & Debus, C. (2026) Sampling Parallelism for Fast and Efficient Bayesian Learning. Submitted to Platform for Applied Scientific Computing (PASC)
   Diese Publikation behandelt die Entwicklung eines skalierbaren Sampling-Parallelismus für Monte Carlo Dropout und bayessche neuronale Netze mit Variational Inference. In dieser Arbeit wird ein verteilter Algorithmus vorgestellt, der paralleles Rechnen effizient nutzt und ein gutes Skalierungsverhalten auf Hochleistungsrechnern zeigt. Die Ergebnisse wurden auf der Platform for Advanced Scientific Computing Conference (PASC) zur Veröffentlichung angenommen.

5. Kiefer, N., Weyrauch, A., Öz, M., Streit, A., Götz, M., & Debus, C. (2024). A comparative study of pruning methods in transformer-based time series forecasting. arXiv preprint arXiv:2412.12883.
   Diese Publikation untersucht systematisch verschiedene Pruning-Strategien zur Reduktion der Modellkomplexität in Transformer-basierten Zeitreihenmodellen. In einer vergleichenden Analyse werden unterschiedliche strukturierte und unstrukturierte Beschneidungsmethoden hinsichtlich Vorhersagegenauigkeit, Modellgröße und Laufzeit bewertet. Die Studie zeigt, unter welchen Bedingungen sich hohe Sparsity-Level realisieren lassen, ohne die Prognoseleistung wesentlich zu beeinträchtigen, und liefert damit praxisrelevante Leitlinien für effiziente Transformer-Architekturen.

6. Kuhn, N., Weyrauch, A., Heyen, L., Streit, A., Götz, M., & Debus, C. (2026). Bayesian Lottery Ticket Hypothesis. arXiv preprint arXiv:2602.18825.
   Diese Publikation untersucht die Lottery Ticket Hypothesis im Kontext bayesscher neuronaler Netze mit Variational Inference. Dabei wird analysiert, ob sich auch in probabilistischen Netzwerken dünnbesetzte Subnetzwerke identifizieren lassen, die isoliert trainierbar sind. Die Studie zeigt architekturspezifische Unterschiede, insbesondere zwischen Transformern und Convolutional Neural Networks, und liefert neue Einsichten zur Struktur bayesscher Modelle. Auch diese Arbeit befindet sich aktuell im Review bei der Conference on Uncertainty in Artificial Intelligence.

Darüber hinaus wurde eine Arbeit zur Untersuchung prädiktiver Verteilungen und des Einflusses verschiedener Prior-Annahmen entwickelt. In dieser Studie wird analysiert, wie sich die Wahl

der angenommenen Wahrscheinlichkeitsverteilung auf die Modellkalibrierung und die Qualität der Unsicherheitsabschätzung auswirkt. Die Arbeit entstand im Zuge der Weiterentwicklung des Frameworks torch-blue und trägt zum besseren Verständnis probabilistischer Modellierung bei. Frühere Entwicklungsarbeiten zur Skalierbarkeit und Performance-Analyse von Unsicherheitsmethoden flossen in kombinierter Form in die oben genannten Publikationen ein.

Literatur

[1] A. Weyrauch, T. Steens, O. Taubert u. a., „Recycle: Fast and efficient long time series forecasting with residual cyclic transformers“, in *2024 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2024, S. 1187–1194.

[2] Y. Gal und Z. Ghahramani, „Dropout as a bayesian approximation: Representing model uncertainty in deep learning“, in *international conference on machine learning*, PMLR, 2016, S. 1050–1059.

[3] M. D. Hoffman, D. M. Blei, C. Wang und J. Paisley, „Stochastic variational inference“, *Journal of machine learning research*, 2013.

[4] J. Gawlikowski, C. R. N. Tassi, M. Ali u. a., „A survey of uncertainty in deep neural networks“, *Artificial intelligence review*, Jg. 56, Nr. Suppl 1, S. 1513–1589, 2023.

[5] A. Weyrauch, L. H. Heyen, J. P. G. H. Muriedas u. a., „torch_blue: A Flexible Python Package for Bayesian Neural Networks in PyTorch“, *Journal of Open Source Software*, Jg. 11, Nr. 117, S. 9415, 2026.

[6] A. Weyrauch, L. H. Heyen, N. Kuhn, A. Streit, M. Götz und **C. Debus**, „On the Scaling of Predictive Samples in Stochastic Variational Inference for Bayesian Neural Networks“, Submitted to the Conference on Uncertainty in Artificial Intelligence (UAI), 2026.

[7] A. K. Özdemir, L. H. Heyen, A. Weyrauch, A. Streit, M. Götz und **C. Debus**, „Sampling Parallelism for Fast and Efficient Bayesian Learning“, Submitted to the Platform for Applied Scientific Computing (PASC), 2026.

[8] D. Kieckhefen, M. Götz, L. H. Heyen, A. Streit und **C. Debus**, „Jigsaw: Training Multi-Billion-Parameter AI Weather Models with Optimized Model Parallelism“, *arXiv preprint arXiv:2507.05753*, 2025.

[9] N. Kiefer, A. Weyrauch, M. Öz, A. Streit, M. Götz und **C. Debus**, „A comparative study of pruning methods in transformer-based time series forecasting“, *arXiv preprint arXiv:2412.12883*, 2024.

[10] N. Kuhn, A. Weyrauch, L. Heyen, A. Streit, M. Götz und **C. Debus**, „Bayesian Lottery Ticket Hypothesis“, *arXiv preprint arXiv:2602.18825*, 2026.

[11] K. Tuteja, G. Olenik, R. Mishchuk u. a., „pyGinkgo: A Sparse Linear Algebra Operator Framework for Python“, in *Proceedings of the 54th International Conference on Parallel Processing*, 2025, S. 753–763.