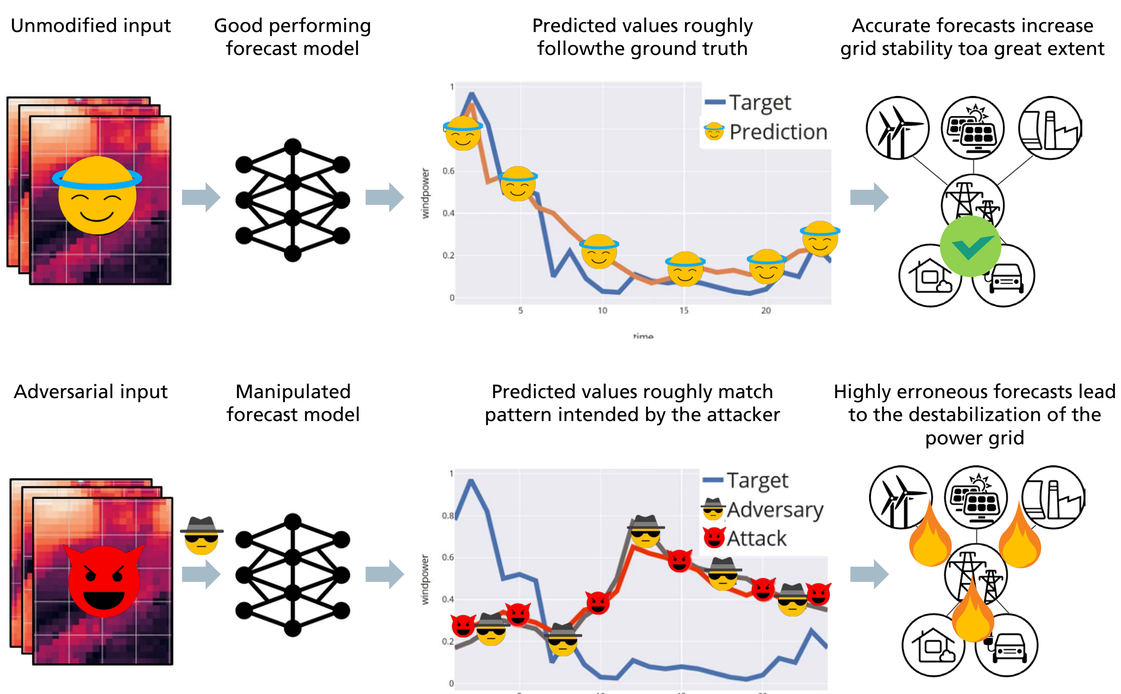
Adversarial Attacks im Energiesektor

René Heinrich, Christoph Scholz, Stephan Vogt

Kontakt: René Heinrich | +49 160 3408484 | rene.heinrich@iee.fraunhofer.de

Der Einsatz von KI-Methoden in kritischen Infrastrukturen wie dem Energiesystem kann zu potenziell sicherheitskritischen Zuständen führen. So stellen mitunter Adversarial Attacks eine große Gefahr dar. Adversarial Attacks sind leichte, aber sehr geschickte Veränderungen der Eingabedaten, mit dem Ziel, maschinelle Lernverfahren zu manipulieren. Auch im Energiesystem eingesetzte KI-Algorithmen, wie z.B. Modelle zur Prognose der Windleistung, sind dieser Bedrohung ausgesetzt. So besteht die Gefahr, dass Angreifer Adversarial Attacks gezielt nutzen, um eine für sie profitable Verfälschung von Windleistungsprognosen zu erzielen.

In diesem Spotlight wurde die Anfälligkeit von zwei verschiedenen KI-basierten Windleistungsprognosemodellen für gezielte Adversarial Attacks analysiert. Außerdem wurde eine Methode untersucht, um die Robustheit dieser Modelle gegenüber Adversarial Attacks zu steigern.



Methoden

Daten

- Windgeschwindigkeitsprognosen (Zeitreihen / Wetterkarten)
- Windleistungsmessungen (einzelne Windenergieanlage / ganz Deutschland)

Modelle

- Encoder-Decoder LSTM (Prognose für eine einzelne Windenergieanlage)
- Convolutional LSTM (Prognose für ganz Deutschland)

Adversarial Attacks:

- Gezielte Adversarial Attacks auf einen Datenpunkt x sind Störungen $s \in S$, welche die Differenz zwischen der Vorhersage des Modells f_θ und dem Ziel des Angreifers y_{adv} minimieren: $\min_{s \in S} \ell(f_\theta(x+s), y_{adv})$

Adversarial Training:

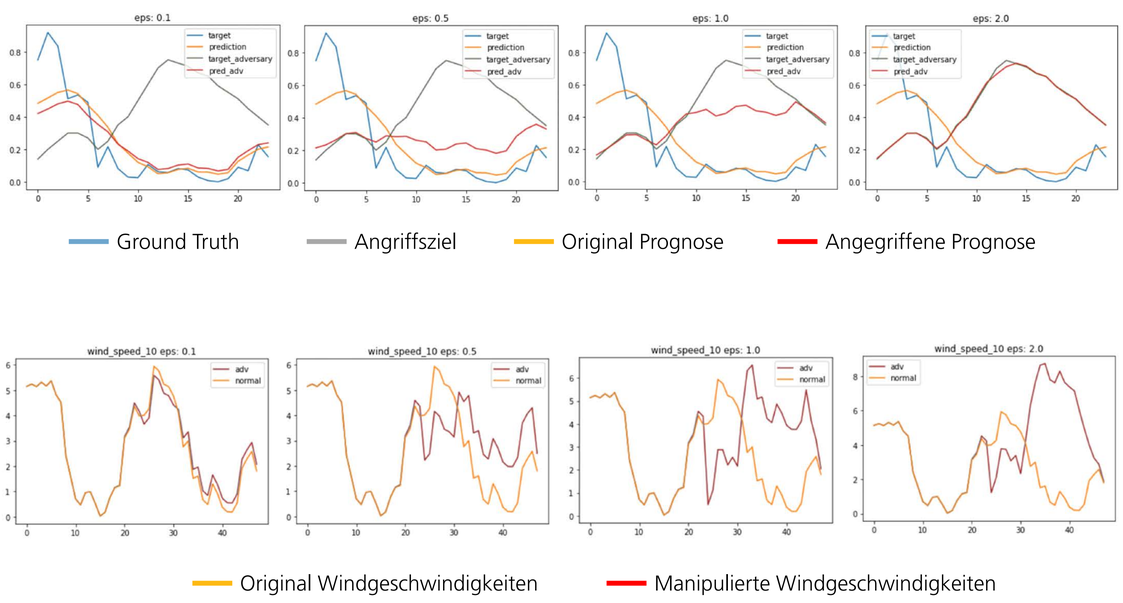
- Adversarial Training ist eine Methode, um die Robustheit von KI-Verfahren gegenüber Manipulationen der Eingabedaten zu erhöhen
- Dabei werden neben den unveränderten Eingabedaten auch manipulierte Daten in den Trainingsprozess miteinbezogen

Ergebnisse

Windleistungsprognosemodell für eine einzelne Windenergieanlage

- Die Eingabedaten sind Windgeschwindigkeitsprognosen in Form von Zeitreihen und daher niedrigdimensional
- Starke Manipulationen der Eingabedaten sind erforderlich, um die Vorhersage in Richtung des Ziels des Angreifers zu verzerren
- Dennoch scheint sich das Modell unter Berücksichtigung der manipulierten Windgeschwindigkeiten physikalisch korrekt zu verhalten

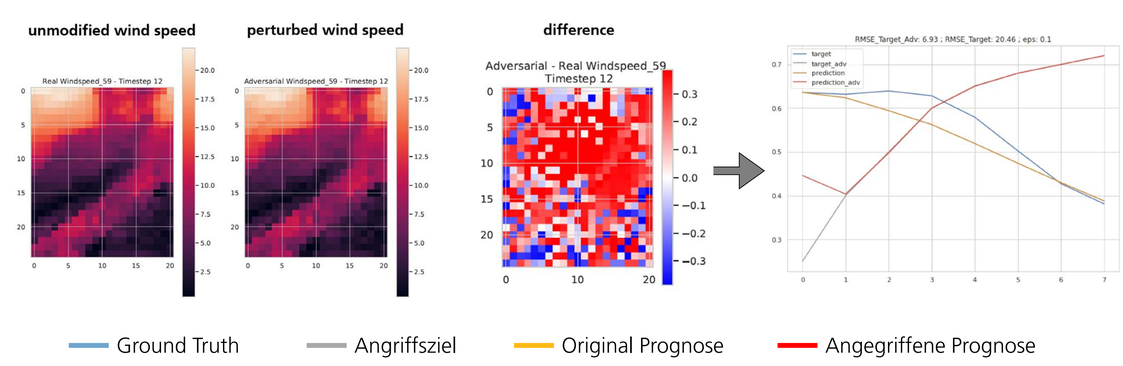
→ **Auch gegenüber starken Manipulationen der Eingabedaten robust**



Windleistungsprognosemodell für ganz Deutschland

- Die Eingabedaten sind Windgeschwindigkeitsprognosen in Form von Wetterkarten und damit sehr hochdimensional
- Bereits kleine und kaum wahrnehmbare Manipulationen der Eingabedaten reichen, um die Prognose in Richtung des Ziels des Angreifers zu verzerren

→ **Sehr anfällig für Adversarial Attacks**



Verteidigung mithilfe von Adversarial Training

- Die Robustheit des Windleistungsprognosemodells für ganz Deutschland kann mithilfe von Adversarial Training deutlich gesteigert werden
- Die Erhöhung der Robustheit geht jedoch mit einer leichten Verschlechterung der Prognosegenauigkeit einher

Gefördert durch:

Hessisches Ministerium für Wissenschaft und Kunst

SPOTLIGHTS

Fraunhofer IEE