

Prüfungsleistung

Allgemeine Informationen:

- Bearbeitung der Prüfungsleistung in Zweierteams
- Späteste Abgabe der Prüfungsleistung: **So., 02.02.2025, 23:59 Uhr**
- Abschlusspräsentation im Prüfungszeitraum (individuell vereinbarter Termin)
- Sie müssen sich zu dieser Prüfungsleistung eigenständig beim Prüfungsamt anmelden!

Inhaltsverzeichnis:

1. Beschreibung, *Seite 2*
2. Einreichung eines Modells, *Seite 3*
3. Erklärung des Datensatzes, *Seite 4*
4. Aufgabenstellung, *Seite 4*
5. Hinweise und Tipps, *Seite 4*
6. Informationen zur Abschlusspräsentation, *Seite 5*
7. Hinweise zur Bewertung, *Seite 5*
8. Hinweise zum Einsatz von KI-basierten Hilfsmitteln wie ChatGPT oder Copilot, *Seite 5*
9. Informationen zur Abgabe, *Seite 6*

Beschreibung:

Sie nehmen als Team an einer fiktiven Machine Learning Competition teil (ähnlich wie Hackathons). Ihre Aufgabe besteht darin, das bestmögliche Modell für einen vorgegeben Datensatz zu entwickeln. In der Competition gibt es drei verschiedene Kategorien, an denen Sie teilnehmen können. Sie müssen an mindestens zwei teilnehmen, dürfen aber auch an allen drei teilnehmen. Eingereichte Modelle werden je nach Kategorie mit anderen Kriterien bewertet. Am Ende gibt es separate Platzierungen für jede Kategorie und eine Gesamtplatzierung (über alle Kategorien hinweg). Können Sie einen Sieg in der Gesamtbewertung erringen?

- **Kategorie 1: Manual Feature Extraction**

- Manuelle Extraktion geeigneter Merkmale und dadurch Überführung der Daten in ein tabellarisches Format. Modellierung erfolgt darauf aufbauend mit geeigneten Modellen für tabellarische Daten.
- Berechnung des Scores: $1 * Accuracy_{mean} - 0.5 * Accuracy_{Std} - 0.1 * \log(n_{features})$
- Bestmöglicher Score: 1,00

- **Kategorie 2: Deep Learning**

- Verwendung von Deep Learning Ansätzen zur Modellierung. Keine manuelle Extrahierung von Merkmalen.
- Berechnung des Scores: $1 * Accuracy_{mean} - 0.5 * Accuracy_{Std} - 0.1 * \log(n_{sensors})$
- Bestmöglicher Score: 1,00

- **Kategorie 3: Computer Vision:**

- Umwandlung der Daten in Bilddaten und anschließend Verwendung von Computer Vision Ansätzen zur Modellierung.
- Berechnung des Scores: $1 * Accuracy_{mean} - 0.5 * Accuracy_{Std} - 0.1 * \log(width * height * channels)$
- Bestmöglicher Score: 1,00

- **Hinweis:** In Kategorie 2 & 3 können Sie Bonuspunkte sammeln (siehe Aufgabenstellung).

Einreichung eines Modells:

Um eine Vergleichbarkeit zu gewährleisten, soll folgender train-test-split verwendet werden. Bitte beachten Sie das stratify-Argument:

```
# Train-Test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y_encoded, test_size=0.2, random_state=RANDOM_STATE, stratify=y_encoded
)
```

Um die Robustheit der trainierten Modelle zu garantieren, wird eine dreifache Kreuzvalidierung gefordert, d.h. das Modell muss dreimal trainiert werden, wobei jedes Mal ein anderer random_state verwendet wird. **Folgende random_states sollen verwendet werden: 27, 6728, 49122**

Vorgehensweise, um ein Modell in einem der drei Scoreboards zu platzieren:

- Modell mit den vorgegebenen Train-Test-Sets dreimal trainieren.
- Erreichte Accuracy für jeden Split aufzeichnen, z.B. mit classification_report von scikit-learn:

```
print(classification_report(y_test, preds, zero_division=0.0))
```

	precision	recall	f1-score	support
0	0.99	0.94	0.96	72
1	0.82	0.78	0.80	72
2	0.57	0.75	0.65	72
3	0.96	0.90	0.93	225
accuracy			0.86	441
macro avg	0.84	0.84	0.84	441
weighted avg	0.88	0.86	0.87	441

- Mittelwert und Standardabweichung der Accuracy berechnen.
- Erreichtes Ergebnis zusammen mit weiteren Informationen (z.B. Anzahl verwendeter Merkmale, verwendetes Modell) an Christian Seidler übermitteln. Ihr Ergebnis wird dann zeitnah im öffentlichen Scoreboard eingetragen.

Sie dürfen beliebig viele Modelle während der Competition einreichen. Für die Berechnung der Endplatzierung wird nur das beste Modell des Teams verwendet.

Tipp: Überlegen Sie sich frühzeitig, wie Sie die Rechenzeit Ihrer Modelle optimieren können, da Sie jedes Modell dreimal trainieren müssen. Speicher Sie Ergebnisse wichtiger Zwischenschritte Ihrer ML-Pipeline (z.B. extrahierte Merkmale) ab, um Rechenzeit zu sparen.

Modelle können bis zum Abgabende der Prüfungsleistung eingereicht werden.

Erklärung des Datensatzes:

Für die Prüfungsleistung wird folgender Datensatz verwendet: Condition Monitoring of Hydraulic Systems. Der Datensatz besteht aus 2205 Messungen, 17 Sensoren (zwischen 1 und 100 Hz) und 5 verschiedenen Zielgrößen. Jede Messung umfasst 60 Sekunden Daten. Es wurden die Zustände verschiedener hydraulischer Komponenten (Kühler, Ventil, Pumpe, Akku) variiert.

Wichtig: Es soll nur die **Valve Condition** als Zielgröße verwendet werden! Diese besteht aus 4 verschiedenen Zuständen: 100 (= optimal), 90 (= leichte Verzögerung), 80 (= starke Verzögerung), 73 (= kurz vor Versagen der Komponente).

Downloadlink: <https://archive.ics.uci.edu/dataset/447/condition+monitoring+of+hydraulic+systems>

Aufgabenstellung:

- **Data Exploration:** Visualisieren Sie den Datensatz, um ein Gefühl für die vorliegenden Daten zu entwickeln. Welche Schlüsse ziehen Sie aus Ihren Visualisierungen?
- **Signal Preprocessing:** Verarbeiten Sie die Signale gemäß Ihren Beobachtungen während der Data Exploration für die nachfolgende Modellierung vor.
Hinweis: Je nach verwendetem Modell sind hier andere Schritte notwendig.
- **Modellierung:** Implementieren, Evaluieren und Optimieren Sie verschiedene (Klassifikations-) Modelle für mindestens zwei der drei Kategorien zur Vorhersage der Valve Condition. Sie dürfen hierfür jegliches öffentliches Python-Paket verwenden.
Wichtig: Die Anwendung von Optimierungstechniken darf nicht auf dem Testset erfolgen. Erstellen Sie hierfür ein extra Validierungsdatenset aus den Trainingsdaten.
- **Bonuspunkte:** Bonuspunkte fließen in die Gesamtbewertung des Projekts ein. Sie können Bonuspunkte in den Kategorien 2 und 3 erlangen:
 - Kategorie 2: Implementieren und verwenden Sie mindestens zwei der folgenden State-of-the-Art Modelle: [TimesNet](#), [TimeMixer](#), [Non-stationary Transformer](#), [Informer](#)
 - Kategorie 3: Neben Spektrogrammen existieren noch weitere Möglichkeiten, Zeitreihen in Bilder zu überführen. Überführen Sie die Daten mit mindestens zwei der folgenden Ansätze in Bilddaten und verwenden Sie diese für die Modellierung: [Recurrence Plot](#), [Markov transition field](#), [Gramian angular summation/difference field](#)

Hinweise und Tipps:

- Dokumentieren Sie Ihre Beobachtungen und Entscheidungen in einer PowerPoint fortlaufend. Diese Informationen können Sie dann für die Abschlusspräsentation wiederverwenden.
- Sie dürfen mich bei Fragen / Unklarheiten jederzeit kontaktieren. Außerdem dürfen Sie mich um Tipps bitten.

Informationen zur Abschlusspräsentation:

Sie müssen als Gruppe eine Abschlusspräsentation vorbereiten. Die Abschlusspräsentation erfolgt im Prüfungszeitraum, der Termin wird nach Absprache mit den einzelnen Gruppen vereinbart. Die Dauer der Abschlusspräsentation beträgt insgesamt 30 Minuten (mit Rückfragen) und besteht aus 2 verschiedenen Bestandteilen. Die 15-minütige Präsentation soll Ihre erarbeiteten Ergebnisse darstellen, z.B. verwendete Modelle, erreichte Ergebnisse, Lessons Learned. Sie können für die Präsentation annehmen, dass das Plenum gute Kenntnisse im Bereich KI aufweist.

Der Ablauf der Abschlusspräsentation ist wie folgt:

- Teil 1 (15 Minuten): Präsentation der Ergebnisse durch die Gruppe.
- Teil 2 (15 Minuten): Fragen zum abgegebenen Code, den vorgestellten Präsentationsinhalten und zum Kursinhalt.

Hinweise zur Bewertung:

Die Bewertung erfolgt anhand von mehreren Kriterien, u.a.:

- Abschlusspräsentation
- Eindruck des abgegebenen Codes (z.B. Bugfreiheit, Übersicht, Aufbau, Nachvollziehbarkeit)
- Anzahl und Varianz der implementierten Modellierungsansätze
- Korrektheit der einzelnen Implementierungen
- Erreichte Modellierungsergebnisse

Hinweise zum Einsatz von KI-basierten Hilfsmitteln wie ChatGPT oder Copilot:

Die Verwendung von KI-basierten Hilfsmitteln ist in diesem Kurs grundsätzlich gestattet. Allerdings wird ein **transparenter** und **verantwortungsvoller** Einsatz dieser Tools erwartet:

- **Transparent** bedeutet, dass eine Angabe erfolgt, dass der Paragraf, der Codeblock, etc. unter Verwendung eines KI-Tools entstanden ist. Dies ist ähnlich zu der Quellenangabe für einen Nachweis einer externen Quelle und wird von den vielen KI-Tools, wie beispielsweise ChatGPT ausdrücklich verlangt (vgl. <https://openai.com/policies/sharing-publication-policy/>). Falls keine Quelle für die Ausgabe aus einem KI-Tool erfolgt, liegt ein Plagiat vor. Dies gilt auch für Codeblöcke, die unverändert übernommen werden! Bei Umformulierung von eigens erstellten Sätzen, oder Optimierung von eigens erstelltem Code durch ein KI-Tool muss keine Angabe erfolgen. Auch bei einer Umformulierung oder umfangreichen Abänderung des generierten Codes muss keine Angabe erfolgen. **Zum Nachweis muss ein eigenes Dokument erstellt werden, in dem ein Screenshot des Prompts mit der Antwort des KI-Tools hinterlegt ist. Außerdem muss jeder Nachweis eine eigene, eindeutige ID haben (ähnlich wie traditionelle Quellenangaben), welche im Code bzw. im eigenen Text referenziert wird.**
- **Verantwortungsvoll** bedeutet, dass die Ausgaben des KI-Tools auf Vollständigkeit und Korrektheit geprüft werden. Sie übernehmen mit Ihrem Namen die Verantwortung für die Aussagen. Falsche oder unvollständige Aussagen von Seiten des KI-Tools müssen durch Sie korrigiert bzw. ergänzt werden.

Abgabe:

Bitte laden Sie folgende Dokumente gruppenweise auf Ilias hoch:

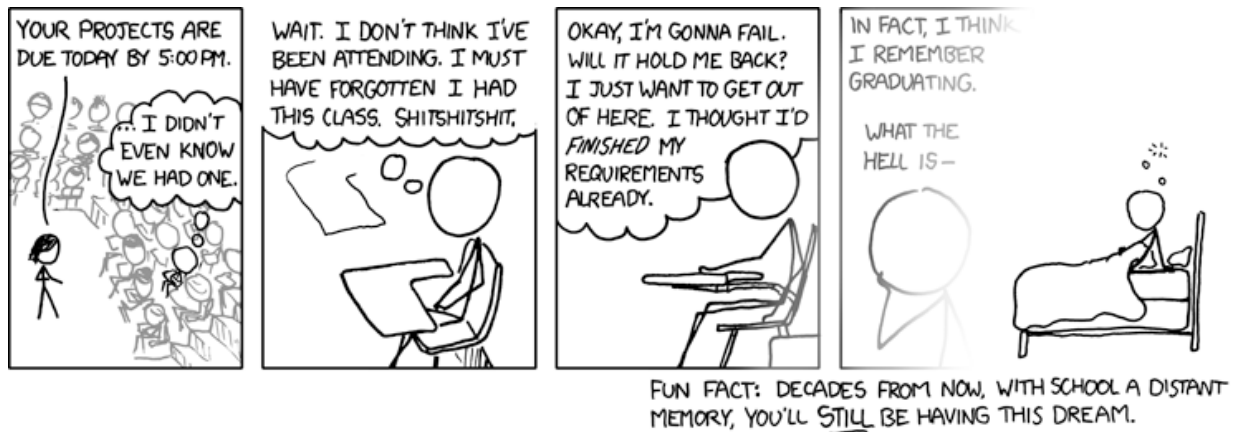
- Abschlusspräsentation, als .pdf-Datei.
- Ihren Code, als .zip-Datei
- Falls verwendet: Liste mit den verwendeten KI-Prompts, als .pdf-Datei.

Späteste Abgabe der Prüfungsleistung: **So., 02.02.2025, 23:59 Uhr**

Bei Fragen / Unklarheiten stehe ich per E-Mail zur Verfügung: seidler@hs-albsig.de



Gutes Gelingen!



Quelle: <https://xkcd.com/557/>