FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Mid-Session: Interpretable Machine Learning

Presentations: September 9 & 10, 2025
Final Report Deadline: September 22, 2025 (strict deadline)

Overview

This document summarizes expectations for your seminar presentation and final report.
Each student must:

- Choose an application domain (e.g., healthcare, finance, justice).
- Select one interpretable model and one black-box model.
- Apply both local and global post hoc interpretability methods.
- Deliver:
  - A 20-minute presentation + 10 minutes for discussion (50% of final grade).
  - A final report of 6–8 one-column pages in LNCS format (50% of final grade).

Presentation Guidelines

- Template:
  `https://www.intern.fau.de/kommunikation-und-marke/vorlagen/praesentationsvorlagen-powerpoint`
- Time: 20 minutes talk + 10 minutes for questions.
- Focus on explaining and comparing interpretability strategies.
- Include critical reflections on strengths, limitations, and trade-offs.
- Practice to ensure clarity and timing.

Final Report Guidelines

The final report must follow the Springer LNCS format:

- Template:
  `https://www.overleaf.com/latex/templates/springer-lecture-notes-in-computer-science/kzwwpvhwnvfj`
- Citation style: Use the default LNCS bibliography style.
- Report length: 6–7 one-column pages.
- Your report should include the following sections:

1. Introduction

- Define your task and motivation.
- Mention why interpretability is relevant in this domain.

2. Data and Preprocessing

- Describe your dataset (origin, size, features, target variable).
- Include Exploratory Data Analysis (EDA) to support your understanding of the data. This may include, but is not limited to:
  - Summary statistics
  - Missing data analysis
  - Feature distributions
  - Correlations or relationships between variables
- You are encouraged to select EDA tasks that help justify your modeling decisions and interpretability analysis later in the report.
- If you apply feature transformations or engineering, ensure that the resulting features remain interpretable (i.e., meaningful to humans).

3. Models

- Briefly describe the interpretable and black-box models used.
- Report on performance using standard metrics.
- Justify model choice.

4. Interpretability Methods

- Apply local methods (e.g., SHAP, LIME) and explain the predictions of individual instances.
- Apply global methods (e.g., SHAP bar/summary plots, permutation importance) to identify overall feature importance.
- Use visualizations effectively.

5. Discussion

- Critically assess how well the interpretability methods worked.
- Discuss trade-offs between performance and interpretability.
- Reflect on limitations and possible improvements.

6. Conclusion

- Summarize key findings.
- Include possible future work.

Structure and Expectations

Your project should follow a clear, hypothesis-driven structure. Based on your understanding of interpretability methods (e.g., from class content or the kickoff session), you should formulate an initial hypothesis to guide your experiments.

- **Example Hypotheses:**
  - SHAP provides more robust and consistent local explanations than LIME.
  - Permutation importance better captures global feature relevance than SHAP.
  - Neural networks outperform random forests in predictive accuracy on the selected dataset.
  - Interpretable models (e.g., logistic regression, decision trees) lead to similar conclusions as black-box models but are easier to justify.
- Your experimental design, data preprocessing, model selection, and interpretability evaluation should aim to test this hypothesis.
- Reflect on whether the results support or contradict your hypothesis, and discuss possible reasons (a well-justified rejection can be equally valuable).
- If your hypothesis evolves during the project, explain how and why during you presentation.
- Students are welcome to use interpretability methods not explicitly covered in the seminar (e.g., Deep SHAP) if they are meaningfully compared against at least one method discussed during the seminar.

This approach ensures that your report is not just a technical summary, but a coherent narrative demonstrating analytical thinking.

Project Title

Below are some example project titles that reflect the goals of the seminar. These can help guide your thinking when choosing a focus and framing your report.

- *LIME vs. SHAP: A Comparative Study on Local Interpretability for Loan Approval Predictions*
- *When Simplicity Wins: Interpretable Models vs. Deep Learning in Predicting Patient Readmission*
- *From Predictions to Explanations: Permutation Importance and SHAP on Financial Data*
- *The Cost of Complexity: Comparing Random Forests and Logistic Regression in Explaining Model Decisions*
- *Trustworthy AI in Practice: Explaining Misclassifications with LIME and SHAP*
- *Beyond Accuracy: Exploring Interpretability-Performance Trade-offs*
- *Explain First, Predict Later: A Design-first Approach to Transparent Loan Classification*
- *Understanding the Disagreement Between SHAP and Permutation Importance*

Reminders

- Features must be interpretable; avoid black-box feature engineering.
- Reports must be individual, and presentations must reflect your own work.
- The final deadline for submission is **September 22, 2025**. Late submissions will not be accepted.