MLPC 2025 Task 3: Classification Experiments

Tara Jadidi, Florian Schmid, Paul Primus

April 2025

Context

Kepler Intelligent Audio Labs (KIAL), a soon-to-be-founded innovative AI startup, aims to collect a general-purpose dataset with strong temporal annotations to train sound event detection systems for their customer (see Figure 1 for a schematic illustration of sound event detection).

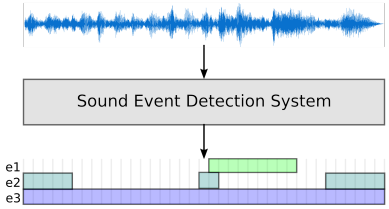

Figure 1: **Sound Event Detection (SED)** systems take an audio recording as input and predict acoustic events of interest $(e1, e2, e3)$ *including* their temporal onsets and offsets.

To make this data set reusable for many potential future customers, KIAL decides to annotate sounds with free text annotations instead of relying on a fixed set of categories (see Figure 2). For a detailed description of the project, look at the project description on our Moodle page.

Figure 2: *Left:* Typically, SED data sets are only annotated for a fixed set of labels $(e1, e2, e3)$. *Right:* Instead of relying on a fixed set of labels, KIAL decided to use arbitrary free text descriptions. Text annotations can be mapped to labels of interest $(e1, e2, e3)$ by matching synonyms, using word embedding models, or leveraging LLMs.

KIAL has divided the project into these five stages:

1. **Data Collection:** Collect a large dataset of audio recordings, containing a large variety of potential target sounds and events. (✓)
2. **Data Annotation:** Once the data is gathered, human annotators carefully annotate sounds in the recordings with textual descriptions and temporal onsets and offsets. (✓)
3. **Data Analysis:** The dataset undergoes a thorough examination through exploratory data analysis, allowing for a deeper understanding of its characteristics and potential challenges before further processing. (✓)
4. **Model Training & Selection:** KIAL's R&D team needs to demonstrate the usefulness of the new approach to the managers by *training machine learning models* for one or multiple fictitious customers (each customer is represented by a fixed set of labels). (← we are here)

5. **Challenge:** A customer requests a sound event detection system for a fixed set of labels. Provide predictions on a hidden test set to prove to the customer that the developed system has a high detection performance and win an (also fictitious) contract.

KIAL has already collected the dataset by scraping publicly available audio recordings on the web (step 1) and annotated the recordings with textual annotations (step 2). Additionally, they conducted a thorough analysis of the dataset (step 3), laying the groundwork for subsequent model training.

Task Outline: Classification Experiments (40 points)

Deadline: Thursday, May 22nd, 23:59
Team Restructering Deadline: Friday, May 2nd, 23:59
Submission: Submit your team's report and slides via Moodle. Only **one** submission per team is allowed. The final submission via Moodle counts.
Group Work: Coordinate early with your team to distribute tasks, set a schedule, and define deadlines. *Check in regularly* and plan buffer time to assist teammates if needed. Inform us promptly if a team member is unresponsive.

This task is mandatory. Content and formal requirements are outlined below.

1 Overview

For this task, we provide labels for some classes that KIAL's customers might be interested in. Your task is to perform systematic classification experiments in your team using the updated data and the discretized class labels.

- Focus on predicting the absence or presence of classes of interest for 120 ms audio frames.
- Decide on an appropriate data split and evaluation criterion to perform model selection and estimate the final performance.
- Apply at least three different learning algorithms from different major groups: Support Vector Machines, Nearest Neighbor Classifiers, Decision Trees, Neural Networks, etc.
- For each algorithm, systematically evaluate different hyperparameter settings, especially those that control the algorithm's overfitting behavior. Analyse and document how the hyperparameters affect whether overfitting occurs (and to what extent it occurs) and how they affect classification performance.
- Visualize the classifier's predictions for two interesting audio files and discuss both strengths and weaknesses. Reflect on potential postprocessing steps (e.g., smoothing or thresholding) that could improve the temporal consistency and overall quality of the predictions.
- If computational resources are limited, you may reduce the size of your experiments by subsampling data points, selecting a subset of features, or focusing on a smaller set of classes.

2 Report (max. 37 points)

For the first part of this assignment, you will have to write a report based on the template provided on Moodle. Your report needs to discuss *all of the following points*:

1. **Labeling Function**: For your analysis, you may focus on a subset of the 54 classes provided.
   (a) Assess how accurately the applied labeling functions capture the intended classes. Do the mapped classes correspond well to the free-text annotations? Are the labeled events clearly audible within the indicated time regions?

(b) Which audio features appear most useful for distinguishing between the classes of interest? *(Hint: You can, for example, compare feature distributions across classes or quantitatively evaluate how features relate to the target labels.)*

(c) How well do the chosen audio features group according to the discretized class labels? Do samples of the same class form tight clusters?

2. **Data Split**:

   (a) Describe how you split the data for model selection and performance evaluation. *Remember: you will need to train your model, tune hyperparameters, and estimate final performance.*

   (b) Are there any potential factors that could cause information leakage across the data splits if they are not carefully designed? If yes, how did you address these risks?

   (c) Describe how you obtained unbiased final performance estimates for your models.

3. **Audio Features**:

   (a) Which subset of audio features did you select for your final classifier? Describe the selection process and the criteria you used to make your choice.

   (b) Did you apply any preprocessing to the audio features? If so, explain which techniques you used and why they were necessary.

4. **Evaluation**:

   (a) Which evaluation criterion did you choose to compare hyperparameter settings and algorithms, and why?

   (b) What is the baseline performance? What could be the best possible performance?

5. **Experiments**:

   (a) For at least three different classifiers, systematically vary the most important hyperparameters and answer the following questions for each of them:

      i. How does classification performance change with varying hyperparameter values? Visualize the change in performance.

      ii. (To what extent) Does overfitting or underfitting occur, and what does it depend on?

   (b) After selecting appropriate hyperparameters, compare the final performance estimate of the three classifiers.

6. **Analysing Predictions**: Find two interesting audio files that have not been used for training and qualitatively evaluate your classifier's predictions.

   (a) Use the spectrogram and the sequence of predictions to visualize the classifier output.

   (b) Listen to the audios and inspect the corresponding predictions of the classifier. How well does the classifier recognize the classes?

   (c) What are particular problematic conditions that cause the classifier to mispredict classes? Can you think of simple postprocessing steps that might help improve the predictions?

The report **must not exceed a page limit of 7 pages**, of which **in total at most 5 pages should be text**.

3 Statement of Contributions

In addition to addressing these questions above, **add a statement of the contributions of all team members** as indicated in the template.

4 Slide Set (3 points)

The complementary slide set should present the results from your written report in a concise manner. More precisely, you will have to answer the questions corresponding to one of the sub-topics outlined in the previous section. The specific topic is determined based on the first letter of your group name, i.e. A for Team Aberrant, or B for Team Bed. To find your topic, determine the according letter, and find your topic in the following list:

First letter of group name	Topic
A, C, E, M, Q	Analysing Predictions
B, F, I, L, N, P	Experiments
D, G, J, R, T, U, W	Data Split & Evaluation
H, K, O, S, V, Y, Z	Labeling Functions & Audio Features

Table 1: Assignment of topics for the slides based on the first letter our your group name.

The **upper limit for the number of slides you should prepare is 4** (excluding an additional title slide that should contain your group name and the member names).

5 Dataset

The dataset download links are available on Moodle. Download the updated version of the dataset for Task 3, provided in Section *Task 3: Classification* on Moodle. Please refer to the slide decks for information on the audio features, the class labels, and the file formats. Have a look at the slide deck of our first tutorial session to see how the dataset can be loaded.

6 Grading

Each of the six topics given in the task outline above will be evaluated according to the following criteria:

- Thoroughness & Completeness: Have you thought about the problem, and answered every question?
- Clarity: Are the ideas, features, algorithms, and results described clearly? Based on your descriptions, could the reader reconstruct your experiments?
- Presentation: Did you select an appropriate way of communicating your results, e.g., did you use meaningful plots where helpful?
- Correctness: Is the proposed procedure/experiment sound, correct?
- Punctuality: The reports must be submitted in time. Any delay will result in reduced grades. Specifically, submitting on the day after the deadline will deduct 1/3 of the points, submitting on the second day after the deadline will deduct 2/3 of the points; submissions on the third day after the deadline or later will be rejected.

You will be awarded all points for the slide set if it addresses the assigned topic, is within the slide limit, and is submitted before the deadline.

7 Summary

- Completing Task 3 is a requirement to pass this course.
- Look at the given questions and answer all of them appropriately in a written report. Make sure to use the report template provided to you via Moodle. Adhere to the given page limit (max. 7 pages where at most 5 pages can be text) and include a statement about the contributions of all team members.

- Create a set of slides tackling the questions of one of the topics. The topic is determined by the first letter of your group name. Make sure to adhere to the slide limit for this step as well (max. 4 + 1 title slide).
- Upload the written report as well as your slides to Moodle before the deadline.
- You will get a maximum number of 37 points for your written report, and 3 points for the slide set.