# Project: ESOF5052 – Big Data Machine Learning

**Objective:**

- This project introduces the students to a real-world big data analytical problem.
- The students will use machine learning or deep learning algorithms to solve a practical classification, regression, or clustering problem and learn the different characteristics of the various models and their performances.
- Setting a foundation for the graduate student's thesis or research project.

**Project groups:**

- The thesis-based graduate students are encouraged to work individually. Topic must be aligned with the student's thesis. Discuss the project topic with the course instructor before start working on it.
  Note: Thesis-based may work in a group of two (1 course-based + 1 thesis-based), with the approval of the course instructor.
- The course-based graduate students can work in a group of three (at max).

**Topic for course-based MSc. students:** Electroencephalogram (EEG)-based emotion recognition.

Note: the students can also propose a project to solve either a descriptive, predictive, or prescriptive problem based on their interest. However, they must get the approval from the instructor.

**Source:**

- Data: Use SEED-IV EEG dataset containing samples for 4 different emotions. You can also use the following URL to access the dataset - https://bcmi.sjtu.edu.cn/home/seed/index.html

- You should carefully select the training and testing samples mentioned in the existing works. For instance, these papers - 1906.01704v1.pdf (arxiv.org) - Section III: Experiments and IEEE Xplore Full-Text PDF - Section IV: Methods explicitly describe the training and testing samples.

  1. **Understand how SEED-IV samples are collected:** SEED-IV dataset also contains 15 subjects, and each subject has **3 sessions**. But it includes four emotion types, and each emotion has 6 film clips. Thus, there is a total of **24 trials**, and each trial has 12-64 samples for one session of each subject. Then there is a total of about 830 samples in one session.

2. **Creating mutually exclusive training and test datasets:** To make fair comparisons, we must follow the way existing works handle the dataset split. The creators of the SEED-IV dataset conduct two types of experiments:

   Experiment A: Out of the 24 trials used in sample collection, the first 16 trials are the training data, and the last 8 trials are the test data.

   Experiment B: Out of the 3 sessions used in sample collection, the data of one session are used as the training set, and the data of another session are used as the test set.

- Public sample code and publications on the same topic can be found on the following website: leader board with papers and code. You can also use the following URL to access them - https://paperswithcode.com/dataset/seed-1

**Deliverables:**

1- A better data-driven machine learning or deep learning model(s) compared to the existing solutions in the publications. If your machine learning model has better performance than the reported results in the research articles, the project will be awarded a maximum of 5 bonus points with respect to the percentage of improvement.
2- Each group will share a folder on Google Drive to show their implementation (code and executed results). The folder must contain a very detailed documented code. A code without documentation **WILL NOT be accepted**. The code and intermediate reports (see next page) must also be uploaded to the D2L.
Please refer to "coding-and-report-writing-tips.pdf", for proper coding and documentation.
3- A well-written formal project report with appropriate flowcharts, block diagrams, data visualizations, experimental results, tables, figures, graphs, etc. (the report is to be written in IEEE - Manuscript Templates for Conference Proceedings)

**Infrastructure:**

- A desktop or a laptop with quad-core, W10, and 8GB memory, will be sufficient for this project.
- It is recommended to use the Google CoLab cloud environment.

# Project: ESOF5052 – Big Data Machine Learning

**Prerequisite Programming Skills:**

- Sufficient knowledge of Python, R, or MATLAB is required. However, students are encouraged to learn and use R for this project. Adopting a new programming paradigm is always good for career choices.

**Project Assessment Rubric:**

| Graduate Attribute 3 - Investigation | |
|---|---|
| **Aspect** | **Data analysis and synthesis of information: functional requirements** |
| 1 | Most functional requirements were missing, incomplete, and extremely inaccurate. |
| **2** | Some basic functional requirements were modeled, with however several inaccuracies.<br>The developed model will not be operational. |
| **3** | Most functional requirements were modeled accurately.<br>The content shows a basic understanding of key ideas.<br>The developed model will be operational with some limitations. |
| **4** | Accurate identification and modeling of functional requirements. |
| **5** | Uses in-depth analysis of the available algorithms to design and develop an innovative solution. The solution eclipses the performances of existing models. |

**Detailed Deliverables with Expected Due Dates:**

| Activity | Due by the end of | 40% total |
|---|---|---|
| Project Proposal, including group detail and a shared folder (Google Drive setup) | 2nd Week | - |
| Code: 25% of the code completion<br>Report: Add-ons to the Stage#1 (proposal) with a strong Introduction, and Literature review (minimum of ten peer-reviewed articles must be considered). | 5th week (In-class investigation) | Project - 6%<br>Report – 3% |
| Code: 50% of the code completion<br>Report: Add-ons to the previous Stage#2 report.<br>- A *data-centric* approach focuses on the early stages of the ML model building, emphasizing improving *data quality* to achieve the desired robust solution. | 8th week | Project - 6%<br>Report - 3% |

| | | |
|---|---|---|
| Therefore, you should show greater data understanding (including analysis of bias) via exploratory data analysis (EDA), data validation and cleaning/ preprocessing, data augmentation, and dataset curation (formation of mutually exclusive training, validation, and tests).<br>- Use graphical/ statistical approaches to understand the samples and their distribution with respect to datasets and classes.<br>- Describe the proposed method using block diagrams and data/operation flows. | | |
| Code: 75% of the code completion<br>Report: Add-ons to the previous Stage#3 report - Results and Discussion | 10th week | Project - 6%<br>Report - 3% |
| Code: 100% of the code completion<br>Report: Add-ons to the previous Stage#4 report - Conclusion, Limitation, Future direction, and References<br>Presentation: A formal presentation | 12th week | Project - 6%<br>Report - 3%<br>Presentation: 4% |

**Important Notes:**

- Students must upload/share the deliverables by the deadline. Students can still upload their deliverables **within 48 hours from the deadline** for a **penalty of 10%** of the marks associated with the deliverable item. **After 48hrs a zero mark** will be assigned to the group without submission.
- The students are encouraged to contact the instructors by email and book an appointment to discuss the aspects of the projects and any other concerns.

**"Make this project your own such that you can be proud of it for a long time."**

**Good Luck!**