wangkongqiang at MentalRiskES@IberLEF 2026: Early Detection of Mental Disorders Risk in Spanish

Kongqiang Wang[1,*], Peng Zhang[1] and Qingli Tan[2]

[1]School of Information Science and Engineering, Yunnan University, 650500, Kunming, Yunnan, China.
[2]College of Ecology and Environment, Yunnan University, 650500, Kunming, Yunnan, China.

Abstract

According to a recent report by the World Health Organization, there is 1 in every 8 people in the world suffering from a mental disorder. The organizers of MentalRiskES at IberLEF 2026 considers that early identification is a key effective intervention to prevent these problems. The task we participated in were: Task 1 - Early Symptom Detection in Therapeutic Conversations. This is a multiclass classification task aimed at infering the patient's responses to a set of validated psychometric instruments including PHQ-9, CompACT-10, and GAP-7. The objective is to enable early detection and facilitate timely interventions. Task 2 - Therapist Response Selection. Task 2 evaluates NLP systems as decision-support tools for therapists in a multi-turn setting. At each interaction step, systems receive the latest user message along with three candidate therapist responses, and must select the most appropriate option according to expert-defined best practices. We compare the performance of different online large language models (LLMs) approaches: $qwen-plus$, $qwen-flash$, $qwen3-max$, from Alibaba cloud BaiLian platform with the latter yielding better results. My final experimental results are MZOE_GAD7 0.665, MAE_GAD7 0.889, Macro_MAE_GAD7 0.910, MZOE_PHQ9 0.701, MAE_PHQ9 0.881, Macro_MAE_PHQ9 0.968, MZOE_CompACT10 0.814, MAE_CompACT10 1.577, Macro_MAE_CompACT10 1.847, MAE_Combined 1.116 in Task 1, and Overall Accuracy 0.268, Macro Precision 0.269, Macro Recall 0.267, Macro F1 0.268, Error Recovery Rate 0.253 in Task 2. The code of our project can be found at https://github.com/WangKongQiang/IberLEF_2026_MentalRiskES.

Keywords

Mental Health, Natural Language Processing, Prompt Engineering, Qwen Large Language Model

1. Introduction

Mental health is a growing concern in our society [1]. According to the World Health Organization (WHO), 1 in 4 people will be affected by mental disorders at some point in their lives. In addition, the COVID-19 pandemic [2] has had a negative impact on the mental health of the general population, with an increase in the number of people suffering from mental disorders. Thus, it is becoming increasingly important to evaluate the use of new technologies to assess the risk of mental illness and the healthcare needs of the population. At the same time, social media platforms such as Telegram [3] have become a popular way for people to express their feeling and emotions. Telegram is a free, end-to-end encrypted messaging service that allows users to send and receive messages and media files in private chats or groups that can be focused on particular topics and allow any user to observe or actively participate. These characteristics make Telegram a suitable source for text-mining.

With this context, an interesting approach is to use Natural Language Processing (NLP) techniques [4] to analyze the language used by people who suffer from mental illness and discover patterns that can be used to identify them and provide the necessary support. The MentalRiskES task at IberLEF 2026 [5] aims to promote the development of NLP solutions specifically for Spanish-speaking social media [6]. They propose two main areas of focus for early-risk detection: Early Symptom Detection in Therapeutic Conversations (*Task 1*), Therapist Response Selection Decision Support (*Task 2*).

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
wangkongqiang60@gmail.com (K. Wang); zpp1219@gmail.com (P. Zhang); tanqingli@stu.ynu.edu.cn (Q. Tan)
https://github.com/WangKongQiang/ (K. Wang); https://github.com/zpp1219/ (P. Zhang)
0009-0003-8736-6397 (K. Wang)

Table 1
The official dataset is provided for Task 1 and Task 2.

Dataset Type	Rounds	Comments
Trial Dataset	12	A limited and synthetic set of trial conversations for format familiarization.
Train Dataset	0	No labeled training data will be released.
Test Dataset	82	Test conversations delivered incrementally during the evaluation phase.

In this work, we present our proposed solution to Task 1 and Task 2 of the 2026 edition of MentalRiskES [7]. This task involves evaluating the likelihood of a patient simulated yet realistic therapeutic conversations in Spanish based on their comments within mental-health-focused groups. For this edition, the corpus consists of simulated yet realistic therapeutic conversations in Spanish between patients and professional therapists. The conversations were created using a combination of human-authored and synthetically generated dialogues, all of which were reviewed by experts to ensure clinical plausibility. The rest of the paper is organized as follows: In the next section, we analyze the dataset used for the task in *Section 2*. Describe the characteristic tasks that our method solves and the evaluation criteria for those tasks in *Section 3*. Then, we describe in detail our methodology for inferring and evaluating the models in *Section 4*. Finally, we discuss the results obtained from online large language models (LLMs) in *Section 5* and present our conclusions and future lines of work in *Section 6*.

2. Dataset Analysis

The dataset given for the task consisted of a total of 82 rounds realistic therapeutic conversations in Spanish between patients and professional therapists in test dataset and 12 rounds realistic therapeutic conversations in Spanish between patients and professional therapists in trial dataset (see Table 1), A limited and synthetic set of trial conversations for format familiarization [8]. Test conversations delivered incrementally during the evaluation phase. No labeled training data will be released.

3. Task Description

To generate the data, the organizers requested the involvement of five licensed psychologists to interact through a custom application with simulated patients, played by psychology students. Each student was assigned a patient profile, including specific characteristics and key symptoms derived from standardized questionnaires such as the PHQ-9, GAD-7, and CompACT-10. During the chats, the application provided AI-generated response suggestions for the therapist, which they could select, edit, or replace entirely with their own input. This process ensured that the conversations reflected both realistic patient behavior and professional therapeutic responses.

Each conversation consists of an ordered sequence of alternating turns between patient and therapist, designed to capture the progressive revelation of emotional states, concerns, and symptoms, as well as evidence-based therapeutic interventions. All dialogues were carefully reviewed by clinical experts to verify the quality, coherence, and clinical validity of both the patient and therapist turns. This step ensures that the simulated interactions remain consistent with real-world therapeutic principles.

All data have been anonymized and underwent ethical review to ensure privacy protection and compliance with responsible research standards.

3.1. Task 1: Early Symptom Detection in Therapeutic Conversations

This is a Multiclass classification task. This task focuses on the early identification of mental health symptoms expressed by a patient during a therapeutic dialogue, by mapping patient language to standardized clinical questionnaires. Participants are given therapist–patient conversations that unfold over multiple chat messages. After each patient turn, systems must infer the patient's responses to a

set of validated psychometric instruments, including PHQ-9, CompACT-10, and GAP-7, based on the dialogue observed up to that point.

Rather than directly predicting abstract symptom labels, systems are required to estimate how the patient would respond to each questionnaire item, effectively translating natural language into structured clinical assessments [9].

Key characteristics of the task include:

- The task is framed as a multiclass classification problem at the item level, where each question has a fixed set of possible responses (e.g., Likert-scale options).
- No labeled training data are provided. Systems must rely on zero-shot, weakly supervised, or knowledge-based approaches [10] (e.g., pretrained language models [11], prompting strategies, or external resources [12]).

3.2. Task 2: Therapist Response Selection

This is a response selection task. Task 2 evaluates NLP systems as decision-support tools [13] for therapists in a multi-turn setting. At each interaction step, systems receive the latest user message along with three candidate therapist responses, and must select the most appropriate option according to expert-defined best practices.

Importantly, the full conversation history is not provided explicitly at each step. Instead, teams are expected to maintain and accumulate context across successive turns, effectively reconstructing the dialogue as the interaction progresses.

This process is repeated over multiple rounds, simulating a continuous therapeutic conversation. Key characteristics of the task include:

- Each instance consists of the current user message and three candidate therapist responses.
- The dialogue history must be implicitly tracked by the system across turns, as it is not re-supplied.
- Responses are selected rather than generated, ensuring a controlled and ethically safe evaluation setting.
- Candidate responses may include both AI-generated and human-authored interventions, with one option identified by experts as the most appropriate.

3.3. Evaluative Metrics

For Task 1: Early Symptom Detection in Therapeutic Conversations. GAD-7 is a scale with 7 items, each worth 0 to 3 points. PHQ-9 is a scale with 9 items, each worth 0 to 3 points. CompACT-10 is a scale with 10 items, each worth 0 to 6 points. To evaluate the performance of the model on the GAD-7, PHQ-9, and CompACT-10 score prediction tasks, we adopt three widely used metrics, namely Mean Absolute Error (MAE), Macro-averaged Mean Absolute Error (Macro-MAE), and Mean Zero-One Error (MZOE).

Mean Absolute Error (MAE). MAE measures the average absolute deviation between the predicted scores and the ground-truth scores:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{pred} - y_i^{true}\right| \tag{1}$$

where N denotes the total number of samples, and y_i^{pred} and y_i^{true} represent the predicted and true scores for the i-th sample, respectively. This metric reflects the overall prediction accuracy and is sensitive to the magnitude of prediction errors. Therefore, it is possible to distinguish between minor errors and severe deviations.

Macro-averaged Mean Absolute Error (Macro-MAE). To mitigate the impact of class imbalance, we further compute the MAE for each class and then take the average across all classes:

$$\text{Macro-MAE} = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{N_c}\sum_{i\in c}\left|y_i^{pred} - y_i^{true}\right|\right) \tag{2}$$

where C is the number of classes, and N_c denotes the number of samples in class c. This metric ensures that each class contributes equally to the final evaluation. Macro-MAE first calculates MAE separately within each category (such as 0-3 points in GAD-7), and then takes the average of all categories, thereby weakening the influence of category imbalance on the evaluation results. This indicator pays more attention to the overall performance of the model on samples of different severity levels.

Mean Zero-One Error (MZOE). MZOE, based on the zero-one loss function, regards the prediction problem as a strict classification task, focusing only on whether the prediction is completely correct without considering the margin of error. This indicator can be regarded as the classification error rate, and its complement corresponds to the accuracy rate. MZOE evaluates the proportion of incorrect predictions by treating the task as a strict classification problem:

$$MZOE = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(y_i^{pred} \neq y_i^{true}\right) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition is true and 0 otherwise. This metric does not consider the magnitude of the error but only whether the prediction is correct.

For Task 2: Therapist Response Selection. To comprehensively evaluate system performance in the multi-turn response selection task, we adopt a set of standard classification metrics along with a task-specific robustness metric.

Overall Accuracy. Overall Accuracy measures the proportion of correctly selected therapist responses across all interaction steps:

$$\text{Accuracy} = \frac{1}{N}\sum_{t=1}^{N}\mathbb{I}(\hat{y}_t = y_t) \tag{4}$$

where N denotes the total number of interaction steps, $\hat{y}_t$ is the predicted response at step t, y_t is the expert-annotated optimal response, and $\mathbb{I}(\cdot)$ is the indicator function. This metric provides a general assessment of the model's decision-making capability in selecting appropriate therapist responses.

Macro Precision, Recall, and F1. To account for potential class imbalance across different response categories, we report macro-averaged precision, recall, and F1 score.

For each class $c \in \mathcal{C}$, precision and recall are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{5}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{6}$$

where TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively.

The macro-averaged metrics are computed as:

$$\text{Macro Precision} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\text{Precision}_c \tag{7}$$

$$\text{Macro Recall} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{Recall}_c \tag{8}$$

$$\text{Macro F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{9}$$

These metrics ensure that performance across all response categories is equally weighted, which is particularly important in therapeutic settings where different intervention strategies must be handled reliably.

Error Recovery Rate. Given the multi-turn nature of the task, we further introduce Error Recovery Rate (ERR) to evaluate the model's robustness in sequential decision-making. Error Recovery Rate measures the proportion of instances where the model successfully recovers from a previous incorrect decision:

$$\text{ERR} = \frac{\sum_{t=2}^{N} \mathbb{I}(\hat{y}_{t-1} \neq y_{t-1} \ \wedge \ \hat{y}_t = y_t)}{\sum_{t=2}^{N} \mathbb{I}(\hat{y}_{t-1} \neq y_{t-1})} \tag{10}$$

This metric captures the model's ability to correct its trajectory after making an error, which is crucial in maintaining coherent and therapeutically appropriate interactions over time.

Discussion While Accuracy and Macro F1 provide a static evaluation of classification performance, Error Recovery Rate offers a dynamic perspective by assessing how well the system adapts across interaction turns. In therapeutic dialogue systems, such recovery ability is essential, as persistent errors may degrade the quality and safety of the conversation.

4. Methodology

4.1. Overview of Qwen Models

This section briefly introduces three models: **Qwen-Plus**, **Qwen-Flash**, and **Qwen3-Max**. Qwen-Plus is a well-balanced, general-purpose model that offers strong performance in both reasoning and text generation. It is suitable for complex natural language processing tasks, such as text analysis, multi-turn dialogue, and academic writing. The model achieves a good trade-off between capability and computational cost. Qwen-Flash is a lightweight and low-latency model designed for fast inference. It maintains essential language understanding capabilities while significantly reducing response time and resource consumption. This makes it ideal for real-time applications, such as online chat systems and interactive services. Qwen3-Max is a high-performance variant in the Qwen series, featuring enhanced reasoning and long-context understanding capabilities. It is well-suited for complex tasks, including long-text modeling, multi-step reasoning, and domain-specific analysis. While it requires more computational resources, it delivers superior performance.

4.2. Prompt Engineering

Prompt engineering refers to the process of designing and optimizing input prompts to effectively guide large language models (LLMs) toward desired outputs. By carefully structuring instructions, context, and examples, prompt engineering enables better control over model behavior without modifying model parameters.

Typical prompt strategies include zero-shot prompting, where the model performs tasks based solely on instructions, and few-shot prompting, which provides a small number of input-output examples to improve performance. More advanced techniques, such as chain-of-thought prompting, encourage the model to generate intermediate reasoning steps, thereby enhancing performance on complex tasks.

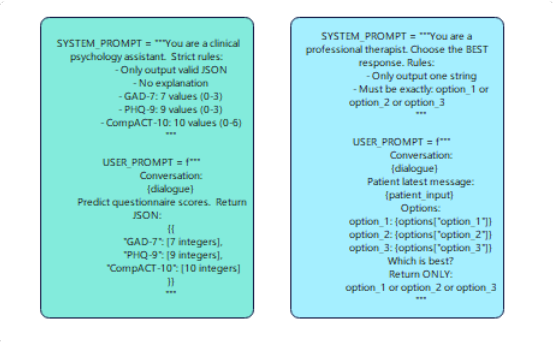

Figure 1: The prompt on the *left* is used for Task 1: Early Symptom Detection in Therapeutic Conversations, while the prompt on the *right* is used for Task 2: Therapist Response Selection.

Table 2
Task 1 official results from various qwen large language models (LLMs) on a test dataset.

Model	MZOE_GAD7	MAE_GAD7	Macro_MAE_GAD7	MZOE_PHQ9	MAE_PHQ9	Macro_MAE_PHQ9	MZOE_CompACT10	MAE_CompACT10	Macro_MAE_CompACT10	MAE_Combined
Baseline_Gemma	0.515	0.582	0.854	0.599	0.724	0.762	0.806	1.700	1.913	1.002
Run_0	0.666	0.895	0.918	0.701	0.886	0.972	0.812	1.588	1.856	1.123
Run_1	0.661	0.891	0.904	0.701	0.880	0.965	0.812	1.609	1.891	1.127
Run_2	0.665	0.889	0.910	0.701	0.881	0.968	0.814	1.577	1.847	1.116
Baseline_Random	0.751	1.185	1.230	0.749	1.215	1.241	0.859	2.123	2.279	1.508

Table 3
Task 2 official results from various qwen large language models (LLMs) on a test dataset.

Model	Overall Accuracy	Macro Precision	Macro Recall	Macro F1	Error Recovery Rate
Baseline_Random	0.363	0.361	0.361	0.361	0.342
Run_0	0.268	0.269	0.267	0.268	0.253
Run_1	0.268	0.269	0.267	0.267	0.273
Run_2	0.266	0.264	0.264	0.264	0.270
Baseline_Gemma	0.180	0.187	0.182	0.176	0.175

Table 4
The inference-time CO2 emissions and energy consumption of each qwen model for Task 1.

Model	Duration	Emissions	Cpu_energy	Gpu_energy	Ram_energy	Energy_consumed
Baseline_Gemma	null	null	null	null	null	null
Run_0	5.97E+02	2.62E-05	5.52E-06	4.99E-03	1.66E-03	6.66E-03
Run_1	5.97E+02	2.62E-05	5.52E-06	5.00E-03	1.66E-03	6.66E-03
Run_2	5.97E+02	2.62E-05	5.52E-06	4.99E-03	1.66E-03	6.65E-03
Baseline_Random	null	null	null	null	null	null

Prompt engineering plays a crucial role in improving model reliability, interpretability, and task adaptability. It serves as a lightweight and efficient alternative to model fine-tuning, especially when computational resources or labeled data are limited. For Task 1: Early Symptom Detection in Therapeutic Conversations and Task 2: Therapist Response Selection, the prompt words are shown in Figure 1.

5. Results

Using the prompt engineering approaches mentioned in the prior section, we came up with different qwen large language models (LLMs) to solve the Task 1 and Task 2 of MentalRiskES. The results in this section are obtained from selecting the three qwen large language models (LLMs): $qwen - plus$, $qwen - flash$, $qwen3 - max$ for Run_0, Run_1, Run_2.

In the Table 2 and Table 3 below, we report the relevant metrics obtained for these tasks and compare them against the ones obtained from baseline models provided by the organizers of the competition. In particular, we report both absolute metrics obtained after observing all the therapeutic dialogues of each patient, and options classification metrics obtained after incrementally observing the therapeutic dialogues across several rounds. Additionally, Table 4 and Table 5 display the inference-time CO2 emissions and energy consumption of each model, based on computing their absolute predictions on the test dataset. These values were estimated using the codecarbon [14] python library.

Table 5
The inference-time CO2 emissions and energy consumption of each qwen model for Task 2.

Model	Duration	Emissions	Cpu_energy	Gpu_energy	Ram_energy	energy_consumed
Baseline_Random	null	null	null	null	null	null
Run_0	5.99E+02	2.63E-05	5.58E-06	5.02E-03	1.66E-03	6.68E-03
Run_1	5.99E+02	2.63E-05	5.58E-06	5.02E-03	1.66E-03	6.68E-03
Run_2	5.99E+02	2.63E-05	5.58E-06	5.02E-03	1.66E-03	6.68E-03
Baseline_Gemma	null	null	null	null	null	null

6. Conclusions

The results show that the approaches considered in this work were successful at modeling of the predictive tasks, with at least one of our models outperforming the baselines in most cases. We can make the following observations:

• The best-performing approach across task1 seems to be the one that uses the prompt engineering of the therapist–patient conversations as input to a online large language model (LLM), namely *qwen3-max* model.

• Most notably, the prompt engineering method that uses *qwen-plus* obtained the best metrics for task 2 across all online large language models (LLMs) including $qwen-plus$, $qwen-flash$, $qwen3-max$,

Thus, it may be important to explore different qwen online large language model (LLM) approaches to improve the performance of early-risk detection. This might include directly employing online learning to predict and update the model as new messages come in or incorporating an ensemble of models to make independent decisions about a message's risk level and combining them for a final decision. Additionally, we may also look into more efficient implementations of the hybrid approach [15] to minimize the disparity in emissions compared to pure deep learning models. These improvements are crucial when considering the deployment of our models in real-world situations and will be the focus of future work.

7. Acknowledgments

Thank you for MentalRiskES@IberLEF 2026 organizing the competition and providing the dataset and other support, and thanks to students of Yunnan University individuals and groups that assisted in the research and the preparation of the work.

8. Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Álvarez-Ojeda, Pablo, Cantero-Romero, M. Victoria, Semikozova, Anastasia, Montejo-Ráez, Arturo, The precom-sm corpus: Gambling in spanish social media, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 17–28.

[2] J. Xiong, O. Lipsitz, F. Nasri, L. M. W. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed, R. S. McIntyre, Impact of COVID-19 pandemic on mental health in the general population: A systematic review, in: Journal of Affective Disorders 277, 2020, p. 55–64. URL: https://www.sciencedirect.com/science/article/pii/S0165032720325891. doi:10.1016/j.jad.2020.08.001.

[3] D. L. Padial, D. Gómez, hackathon-somos-nlp-2023-roberta-base-bne-finetuned-suicide-es· hugging face (2023). URL: https://huggingface.co/hackathon-somos-nlp-2023/roberta-base-bne-finetuned-suicide-es.

[4] L. Breiman, Random forests, in: Machine Learning, volume 45, 2001, p. 5–32. URL: https://doi.org/10.1023/A:1010933404324. doi:10.1023/A:1010933404324.

[5] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian

Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[6] A.G.Fandiño, J.A.Estapé, M. Pàmies, J.L.Palao, J.S.Ocampo, C.P.Carrino, C.A.Oller, C.R.Penagos, A.G.Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[7] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2026: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 77 (2026).

[8] C. S. Perone, R. Silveira, T. S. Paula, Evaluation of sentence embeddings in downstream and linguistic probing tasks, 2018. URL: http://arxiv.org/abs/1806.06259.

[9] C. C. Liu, J. Pfeiffer, I. Vulić, I. Gurevych, Improving generalization of adapter-based crosslingual transfer with scheduled unfreezing, 2023. URL: http://arxiv.org/abs/2301.05487.

[10] J. Friedman, Greedy function approximation: A gradient boosting machine, in: The Annals of Statistics, 2000, p. 29. doi:10.1214/aos/1013203451.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: http://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, J. B. L. Fang, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, p. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R.Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, in: Journal of Machine Learning Research, 12, 2011, p. 2825–2830.

[14] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, S. Luccioni, Codecarbon: estimate and track carbon emissions from machine learning computing, in: Cited on, 2021, p. 20.

[15] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, in: Technometrics, 55–67, [Taylor Francis, Ltd., American Statistical Association,American Society for Quality], 1970, p. 12. URL: https://www.jstor.org/stable/1267351. doi:10.2307/1267351.