Does Complexity Pay Off?
Applying Advanced Algorithms to Depression Detection on the GLOBEM Dataset

Sebastian Cavada, Alvaro Berobide, Yevheniia Kryklyvets
Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
e-mail: {name.surname}@mbzuai.ac.ae

***Abstract*—This manuscript evaluates the performance of state-of-the-art time series analysis algorithms for depression detection on the Generalization of LOngitudinal BEhavior Modeling (GLOBEM) dataset. We assess *Time-Series Mixer (TSMixer)*, *Crossformer*, *Gated Recurrent Unit (GRU)*, *Convolutional Neural Network with Long Short-Term Memory (CNN_LSTM)* and introduce a novel self-developed algorithm with the goal of increasing accuracy over the original *Reorder*. While these models demonstrate robust out-of-domain generalization, they fail to surpass the accuracy of the baseline *Reorder* algorithm, which was specifically developed for in-domain analysis by the GLOBEM team. Our findings reveal consistently low performance across all models, suggesting limitations inherent in the dataset rather than the algorithms themselves. We hypothesize that the dataset's absence of critical variables and insufficient granularity likely limits model convergence. This hypothesis is supported by similar studies that achieved higher accuracy using more frequent data points with similar architecture approaches. Based on these insights, we suggest that future studies might benefit from incorporating more granular sensor measurements and more sophisticated data types, such as, but not limited to, Heart Rate Variability (HRV).**

Keywords-Depression Detection; Time-Series Analysis; Deep Learning; Domain Generalization; Mental Health.

I. Introduction

It is estimated that 3.8% of the global population suffers from clinical depression condition. This mood disorder affects over 280 million people, ranking it among the leading causes of disability [1]. Despite its prevalence, this condition remains challenging to diagnose and treat effectively, often due to delayed detection. Traditional diagnostic methods, relying on subjective assessments, can miss early warning signs. This underscores the need for objective, data-driven approaches to enable earlier and more accurate diagnosis [2] by building applications that will allow for self-monitoring and alerting when professional assistance is required.

In particular, recent advancements in wearable hardware have enabled continuous monitoring of human physiological data, including heart rate, oxygen levels, and movement patterns. This wave of technology sparked interest in the deep learning community to leverage this temporal information to develop automated methods for depression detection [3][4][5]. Despite these innovations, the efficacy of such approaches remains limited, with results often being minimally informative and thus remaining a subject of ongoing research and debate [3].

To the best of our knowledge, this study represents the first evaluation of state-of-the-art time series analysis algorithms for depression detection tasks using the GLOBEM dataset [3]. We examine various advanced models, including *TSMixer* [6], *Crossformer*[7], *GRU* [8], *CNN_LSTM* [9]. Additionally, we introduce a novel algorithm that enhances the baseline *Reorder* [3] with LSTM capabilities. The aim for the new model is to improve the current *Reorder* algorithm. By adapting all these models, we aim to give a snapshot of the current state of depression detection algorithms and emphasize a critical finding: the key to improvement may lie in better data rather than more complex algorithms.

The remainder of the paper is organized as follows: In Section II, we present a review of related work in the field of automated depression detection. Section III details the methodology of our study, including the dataset used and the algorithms evaluated. Section IV presents our results. In Section V, we discuss the implications of our results and the limitations of current approaches. Section VI offers the conclusion and directions for future research.

II. Related Work

Our research focuses on the application of Artificial Intelligence (AI) to address critical health issues like depression, leveraging multi-year longitudinal data. The GLOBEM dataset [3] stands out as a pioneering dataset culled from a comprehensive multi-year data collection study, capturing a broad spectrum of data from 497 unique participants, totaling 705 person-years.

In the field of Time-Series Forecasting (TSF), transformers have revolutionized sequence modeling with their unparalleled performance across domains [10]. However, their application in TSF, especially for long-term forecasting, has yielded mixed results. Some studies have highlighted limitations [11], while others suggest that transformers may still hold untapped potential in this area [12]. The all-Multi-Layer Perceptron (MLP) architecture, initially conceived for Computer Vision [13], has been repurposed for TSF through the *TSMixer* work [14], enabling the handling of multivariate data and highlighting the adaptability to large datasets and complex real-world scenarios Recurrent Neural Networks (RNNs) [15] and their variants [8] have long been the standard approach for time series forecasting. Their ability to handle sequential data has made them particularly useful for multivariate time series prediction over many years.

Domain generalization in time-series prediction encompasses various related works aimed at developing models capable of performing well across different domains without the need for domain-specific training data [16][17][18]. These

ISBN: 978-1-68558-204-3

methodologies address the challenge of domain shift, enabling models to generalize effectively across diverse domains.

III. Methods

This section provides a comprehensive overview of our methodology, covering four key areas: Dataset description, algorithms, experimental setup, and implementation details.

A. Dataset Description

The GLOBEM dataset spans four years and includes data from 705 person-years [19][20]. It consists of two primary data types: survey data and passive mobile sensing data.

Survey data, collected periodically throughout the study, includes metrics from the Beck Depression Inventory-II (BDI-II) and the Patient Health Questionnaire-4 (PHQ-4), which serve as ground truth for depression and anxiety. This data is critical for the binary classification of mental health states (if the pathology is present or not), providing insights into the severity of symptoms across a diverse population.

Passive mobile sensing data gathered via a dedicated app on iOS and Android devices and Fitbit wearable tracks location, phone usage, physical activity, and other behaviors in real-time. This extensive data set, with more than 1000 distinct features from phone usage alone (extracted and standardized by the Reproducible Analysis Pipeline for Data Streams Open Source platform [21]), is crucial to analyzing daily routines and behaviors, offering a comprehensive view of the impacts of lifestyles on mental health. Given the high dimensionality of the raw data, a rigorous feature selection and data preparation process was implemented. This process aimed to distill the most impactful insights while managing computational complexity. The final prediction model utilizes a subset of 54 key features selected for their relevance and predictive power. Data is structured in batches, each representing a 4-week (28-day) period. This temporal structure allows for analyzing both short-term fluctuations and longer-term behavioral patterns, enhancing the dataset's utility for depression detection tasks.

B. Algorithms

1) Reorder - the baseline algorithm: The *Reorder* algorithm is a multi-task learning model that uses the continuity of behavior trajectories to enhance domain generalization in behavioral models; the details are shown in Figure 1. Its primary goal is maintaining time continuity while addressing a principal classification task. It optimizes two distinct losses simultaneously: the binary cross-entropy loss, based on the ground truth, and a second loss from a self-supervised task. This task involves predicting the position of segments randomly shuffled in a subset of all possible permutations. This self-supervised task act as a regularizer, encouraging the network to understand the temporal dimension and improve generalization to the main task.

2) TSMixer - All-MLP Architecture: The *TSMixer* architecture, part of the ALL-MLP family, is chosen for its ability to efficiently handle multivariate time-series data through MLPs. This model simplifies complex pattern recognition across time and feature dimensions, making it suitable for the computational demands of depression diagnosis prediction [6].

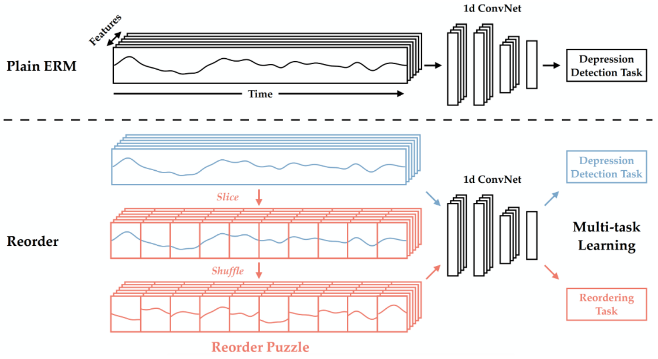

Figure 1. The *Reorder* architecture, image from the original paper [3].

3) Crossformer - Transformer Based Model: We utilize the *Crossformer* architecture due to its advanced capacity for handling long-term dependencies and high-dimensional data. Its hierarchical integration of features allows for a nuanced understanding of time-series patterns, crucial for accurate predictions in time-series data [7].

4) Utilization of Models from the RNN Family: RNNs, including *GRU* and *LSTM* variants, are employed for their unique ability to maintain a memory of past meaningful information, enabling effective modeling of time-dependent data. This characteristic is particularly beneficial for tracking the progression of depressive symptoms over time [8].

5) Reorder + CNN_LSTM: Our top-performing model, as shown in Table I, is a novel algorithm that we have named *Reorder + CNN_LSTM*. This algorithm combines the strengths of the original Reorder model [3] with the capabilities of a CNN_LSTM architecture. The LSTM module was added particularly to capture long-term dependencies in the sequence, enhancing the model's ability to recognize patterns over extended periods.

This hybrid approach used three times as many parameters as the original *Reorder* but allowed us to leverage the benefits of each individual component:

- Reorder: Effective temporal data handling
- CNN: Spatial feature extraction
- LSTM: Long-term dependency learning and improved generalization

The 32,138 parameters result from merging the different models: Reorder and CNN_LSTM. Some parameters are shared among various modules, such as the initial and final layers, therefore the total number of parameters doesn't exactly add up to the individual number of parameters of each model.

We report parameter count as a key indicator of model complexity, especially relevant in resource-constrained environments.

C. Experimental Setup

In this research, we adhered to the experimental framework presented in the GLOBEM paper [3] to ensure the comparability

TABLE I

ALL RESULTS ARE IN DESCENDING ORDER, OUR METHODS IN DIFFERENT COLORS, RESULTS ARE IN BALANCED ACCURACY. THE STANDARD DEVIATION IS CALCULATED ON THE NUMBER OF RUNS BETWEEN THE DATASETS. **The number of parameters takes into account only trainable parameters - The comma is used to separate thousands, while the point is used for decimals.*

Model	**Number of Parameters***	**Results**		
		Single Dataset	Leave one out	Pre/Post Covid
Reorder + CNN-LSTM	32,138	0.629 ± 0.045	0.542 ± 0.009	0.530 ± 0.001
Reorder	**10,162**	**0.626±0.063**	**0.548±0.030**	**0.513±0.009**
CNN-LSTM	**24,378**	**0.601±0.026**	**0.513±0.009**	**0.507±0.004**
GRU	**62,226**	**0.591±0.034**	**0.516±0.011**	**0.502±0.001**
Crossformer	**131,527**	**0.590±0.001**	**0.503±0.003**	**0.516±0.002**
ERM-Transformer	12,354	0.584±0.013	0.509±0.008	0.512±0.016
IRM	2,698	0.573±0.016	0.506±0.006	0.499±0.000
ERM-1dCNN	2,698	0.568±0.006	0.510±0.008	0.514±0.006
ERM-Mixup	2,698	0.568±0.006	0.501±0.008	0.507±0.004
ERM-LSTM	22,186	0.565±0.019	0.512±0.006	0.512±0.003
TSMixer	**43,429**	**0.543±0.035**	**0.521±0.006**	**0.499±0.000**
CSD-D	2,839	0.562±0.022	0.521±0.002	0.512±0.006
Siamese Network	2,664	0.545±0.025	0.509±0.010	0.515±0.002
CSD-P	2,875	0.542±0.010	0.511±0.006	0.516±0.000
ERM-2dCNN	12,994	0.533±0.013	0.510±0.006	0.504±0.006
DANN-D	3,281	0.526±0.016	0.514±0.004	0.514±0.000
MLDG-D	2,698	0.522±0.013	0.511±0.006	0.495±0.004
MLDG-P	2,698	0.508±0.011	0.510±0.003	0.500±0.003
MASF-D	2,970	0.505±0.006	0.505±0.001	0.504±0.007
DANN-P	3,578	0.502±0.002	0.500±0.000	0.500±0.000
MASF-P	2,970	0.495±0.007	0.505±0.004	0.509±0.011

of our results. Our experiments were designed to evaluate the performance of algorithms in three distinct scenarios:

1) **Single Dataset** This method divides the data for each participant within a dataset, using the first 80% for training and the remaining 20% for testing. This setup assesses the model's predictive capability using past data to forecast future outcomes.
2) **Leave-One-Dataset-Out:** In this cross-dataset approach, the model is trained on three datasets and tested on the fourth. This configuration evaluates the model's generalizability across different datasets.
3) **Pre/Post-COVID Analysis:** This setup aims to discern the impact of the COVID-19 pandemic on model performance. It involves training on datasets INS-1 (Data Set year 1) and INS-2 (pre-COVID) and testing on INS-3 and INS-4 (post-COVID), with a subsequent reversal of training and testing datasets to examine the effects comprehensively. The different particularities of the dataset are explained in the following "Dataset Description" section.

D. Implementation details

All computational experiments were conducted on a high-performance workstation equipped with a GPU 4090 and 44GB of RAM, using TensorFlow and Keras for model implementation. We adopted the Adam optimizer with an initial learning rate of 0.001, adjusted by cosine annealing with a decay rate of 0.95 and a step size of 20. The models were trained for up to 200 epochs with early stopping based on the best validation loss, allowing a minor degree of data leakage, as noted in the original paper. Consistent with established protocols to ensure a fair comparison with previous studies, we used balanced accuracy as our main evaluation metric [3]. This metric, calculated as $\frac{1}{2}\left(Sensitivity + Specificity\right)$, where $Sensitivity = \frac{TruePositive}{TruePositive+FalseNegative}$ and $Specificity = \frac{TrueNegative}{TrueNegative+FalsePositive}$, is particularly effective in contexts with class imbalances. Using balanced accuracy allows us to accurately assess and compare the performance of our proposed approach against existing methods, providing a robust measure of effectiveness across diverse models and datasets.

IV. RESULTS

Table I reports the balanced accuracy for all methods, ordered by performance on the single dataset. Our experiments are highlighted in different colors, while results from the original paper are in black.

Three of the four State-Of-The-Art (SOTA) models implemented in this study outperform nearly all the models discussed in the original paper, except for their top model, *Reorder*. The gap between *Reorder* and the best-adapted model is only 2%, with more recent models lagging behind despite having at least twice as many parameters. This decreased return in performance is observed across all three tasks, with Crossformers being one exception noted below.

Our novel model, *Reorder + CNN_LSTM*, achieved the highest performance in Table I. It showed a slight increase in accuracy (0.3%) on the single dataset; conversely, it improved accuracy on the Pre/Post Covid dataset by a non-trivial 2%. It also showed negligible lowered performance in the leave-one-out dataset. These improvements come with a cost of three times more parameters, as mentioned before.

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-204-3

The Pre/Post COVID dataset proved to be the most challenging task, likely due to lifestyle disruptions in individuals. Higher accuracy on this dataset indicates better model robustness to strong shifts in the test domain. Notably, the *Crossformer* model surpassed the baseline in this task by nearly 1%, but at the cost of having almost 10 times as many parameters.

The fourth SOTA model analyzed, *TSMixer*, significantly underperformed in all tasks, lagging behind both other SOTA models and older deep learning approaches despite requiring a substantial increase in the number of parameters.

V. Discussion

Our comprehensive evaluation of SOTA algorithms and original deep learning methods for depression detection using wearable data has revealed several important insights. Across all methods, we observed consistently low accuracies, a finding that aligns with Xu et al. [3], who noted that "Current cross-dataset generalizability algorithms are still far from satisfactory for real-life deployment." This persistent challenge suggests that despite the variety of algorithms employed, the data itself might lack sufficiently informative values for reliable depression detection.

The limitations of the dataset, as acknowledged by its original authors, are particularly noteworthy. The absence of certain sensor signals, such as Heart Rate Variability and Saturation of peripheral Oxygen (SpO2) measures, may be critical missing variables needed to increase accuracy and more reliably detect depression [3][22]. This observation is further supported by research on more granular data, such as minute-per-minute Heart Rate Variability, which has achieved higher accuracy rates of around 71% in similar settings [22].

Our results also indicate that increased model complexity does not necessarily translate to improved performance. The novel *Reorder + CNN_LSTM* algorithm demonstrated only considerable improvements over the original *Reorder* in one out of three tasks, raising questions about the cost-benefit ratio of increased model complexity. Similarly, the poor performance of the TSMixer model, despite its increased parameter count, suggests that its linear nature may not adequately capture the intricacies of this particular time series multivariate distribution.

VI. Conclusion and Future Work

In conclusion, our research adapted new state-of-the-art time series analysis algorithms, specifically *TSMixer*, *Crossformer*, *GRU*, and *CNN_LSTM* for depression detection on the GLOBEM dataset. While these algorithms exhibited robust out-of-domain generalizability with balanced accuracy on par with specialized architectures, they did not surpass the baseline *Reorder*. We also introduced a novel variant of the *Reorder* algorithm, which improved performance, especially on the Covid cross-dataset. Nevertheless, the baseline *Reorder* still maintains superior computational Pareto efficiency, offering the best accuracy-to-parameter ratio.

Our findings indicate that larger, more complex models perform no better than their simpler counterparts in this specific task, similar to results in other studies where *simpler traditional methods* like tree-based models outperformed complex deep models on tabular data [23]. Furthermore, the results remain close to a non-informative baseline, and we suggest that the current dataset may have insufficient variables for reliable depression detection. Future studies might benefit from incorporating both more granular measurements as well as additional data types such as HRV and SpO2. Furthermore, new sensors, such as the electrocardiogram (ECG) from the Apple Watch, will become available as new devices are released on the market enabling new research.

References

[1] World Health Organization, "Depressive disorder (depression)", *World Health Organization: WHO*, Mar. 2023.

[2] S. Saeb *et al.*, "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study", eng, *Journal of Medical Internet Research*, vol. 17, no. 7, e4273, Jul. 2015, ISSN: 1438-8871. DOI: 10.2196/jmir.4273.

[3] X. Xu *et al.*, *GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization*, arXiv:2211.02733 [cs], Mar. 2023.

[4] P. Chikersal *et al.*, "Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection", *ACM Transactions on Computer-Human Interaction*, vol. 28, no. 1, pp. 1–41, Jan. 2021, ISSN: 1073-0516. DOI: 10.1145/3422821.

[5] R. Wang *et al.*, "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones", en, in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle Washington: ACM, Sep. 2014, pp. 3–14, ISBN: 978-1-4503-2968-2. DOI: 10.1145/2632048.2632054.

[6] S.-A. Chen *et al.*, *TSMixer: An All-MLP Architecture for Time Series Forecasting*, arXiv:2303.06053 [cs], Sep. 2023.

[7] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting", in *The Eleventh International Conference on Learning Representations*, 2023.

[8] Y. Zhang, R. Wu, S. M. Dascalu, and F. C. Harris, "A novel extreme adaptive gru for multivariate time series forecasting", *Scientific Reports*, vol. 14, no. 1, p. 2991, Feb. 2024. DOI: 10.1038/s41598-024-53460-y.

[9] H. Widiputra, A. Mailangkay, and E. Gautama, "Multivariate cnn-lstm model for multiple parallel financial time-series prediction", *Complexity*, vol. 2021, p. 14, Oct. 2021. DOI: 10.1155/2021/9903518.

[10] Q. Wen *et al.*, *Transformers in Time Series: A Survey*, arXiv:2202.07125 [cs, eess, stat], May 2023.

[11] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 11 121–11 128.

[12] *Yes, Transformers are Effective for Time Series Forecasting (+ Autoformer)*, [Online; accessed 1. May 2024], May 2024.

[13] I. O. Tolstikhin *et al.*, "Mlp-mixer: An all-mlp architecture for vision", *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.

[14] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, *Large Language Models Are Zero-Shot Time Series Forecasters*, arXiv:2310.07820 [cs], Oct. 2023. DOI: 10.48550/arXiv.2310.07820.

[15] R. M. Schmidt, *Recurrent neural networks (rnns): A gentle introduction and overview*, 2019. arXiv: 1912.05911 [cs.LG].

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[17] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[18] I. Gulrajani and D. Lopez-Paz, *In search of lost domain generalization*, 2020. arXiv: 2007.01434 [cs.LG].

[19] I. B. Weiner and W. E. Craighead, *The corsini encyclopedia of psychology, volume 4*. John Wiley & Sons, 2010, vol. 4.

[20] J. Stanhope, “Patient Health Questionnaire-4”, *Occupational Medicine*, vol. 66, no. 9, pp. 760–761, Dec. 2016, ISSN: 0962-7480. DOI: 10.1093/occmed/kqw165.

[21] *RAPIDS, https://www.rapids.science/1.10*, [Online; accessed 22. Sep. 2024], Sep. 2024.

[22] L. V. Coutts, D. Plans, A. W. Brown, and J. Collomosse, “Deep learning with wearable based heart rate variability for prediction of mental and general health”, *Journal of Biomedical Informatics*, vol. 112, p. 103 610, 2020.

[23] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?”, *Advances in neural information processing systems*, vol. 35, pp. 507–520, 2022.