StressLLM: Large Language Models for Stress Prediction via Wearable Sensor Data

Bishal Thapa
Texas State University
San Marcos, TX
b_t220@txstate.edu

Micaela Rivas
University of Texas at San Antonio
San Antonio, TX
micaela.rivas@my.utsa.edu

Henry Griffith
San Antonio College
San Antonio, TX
hgriffith5@alamo.edu

Heena Rathore
Texas State University
San Marcos, TX
heena.rathore@txstate.edu

***Abstract*—Stress significantly impacts mental health, immune function, and overall well-being, reducing productivity and quality of life globally. While large language models (LLMs) have been explored in various health prediction domains, including mental health, activity monitoring, and sleep analysis, research on stress prediction remains limited, particularly regarding the identification of key biomarkers. In this study, we examine the capability of LLMs in predicting stress using wearable sensor data. By employing base prompting and systematically removing individual features, we assess the importance of various biomarkers on stress prediction accuracy. Our findings highlight essential biomarkers that improve model robustness, as well as features that contribute little to prediction accuracy, offering insights into more effective and resource-efficient stress monitoring solutions.**

***Index Terms*—Large Language Models, Mental Health, Stress**

I. Introduction

As mental health issues become increasingly prevalent, accurate stress prediction has emerged as a key area of focus [1]. Wearable sensors, including fitness trackers and smartwatches, provide continuous physiological data that, when combined with advanced machine learning models, hold potential for real-time stress assessment [2]. Recent advancements in large language models (LLMs) have opened new avenues for analyzing health data from wearables, enhancing the prediction of chronic diseases, mental health disorders, and activity patterns.

LLMs, trained on vast amounts of data, demonstrate a unique ability to handle complex, unstructured, and multimodal information [3]. This flexibility makes LLMs valuable tools in various healthcare tasks, ranging from disease prediction to interactive diagnostics [4]. Their considerable computational requirements at inference limit the applicability of LLMs for real-time applications versus traditional machine learning systems (MLS) [3]. Despite this limitation, their ability to process large multimodal inputs enhances their ability to produce context-sensitive predictions.

LLMs are also capable of providing explanations for the decisions they make, giving them a distinct edge over MLS. LLMs are currently being utilized for chronic disease prediction using wearable device data, healthcare documentation, and interactive health chatbots [5], which helps users understand complex machine-learning models through conversation. HeLM, a framework introduced by Belyaeva et al. [6], enables LLMs to use multimodal health data for the prediction of chronic disease risk, while HealAI [7] focuses on the generation of SOAP notes from doctor-patient conversations, demonstrating LLMs' ability to handle complex medical tasks reliably. These studies collectively highlight the promising potential of LLMs in medical analysis, health prediction, and aiding medical personnel in complex diagnostic scenarios.

The Health-LLM framework [8] assesses LLMs for their ability to predict various health conditions of interest using wearable sensor data, including stress levels. Although this work provides valuable initial benchmarks, further efforts are needed to better understand which captured biomarkers provide the most value in stress prediction. This study addresses this gap by evaluating which biomarkers, including heart rate, sleep patterns, and physical activity levels, are critical for accurate stress prediction. By identifying key indicators and eliminating less relevant features, we can improve the efficiency of wearable stress-monitoring devices, making them more resource-efficient and accurate for everyday use.

This paper makes the following contributions to the prediction of stress levels using wearable datasets:

- We conduct tests on multiple wearable stress-related datasets across four open-source LLM models, including BioMistralDare [9], Gemini, LLama, and various GPT variants, to evaluate how parameter count influences stress prediction.
- We investigate the impact of age on the accuracy of stress prediction. This is crucial for creating personalized and effective stress management interventions.
- We further examine whether gender influences the accuracy of stress prediction. This ensures that stress prediction models do not perpetuate biases and are effective for all individuals, regardless of gender.
- Finally, we identify the optimal set of biomarkers for accurate stress prediction. The selection of proper biomarkers can enhance model robustness by removing unnecessary features and allow for more efficient deployment on embedded systems

II. Datasets

We used two datasets for the evaluation of our work: PM-Data [11] and Wearable Stress and Affect Detection (WESAD) [12].

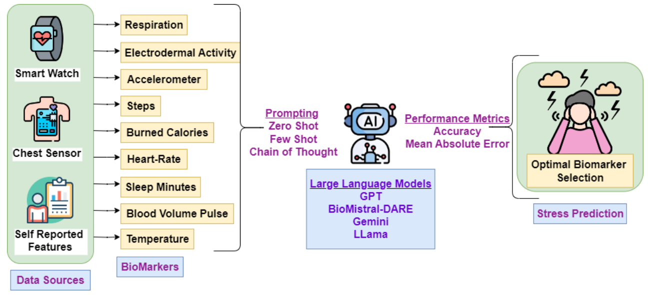

Fig. 1. LLM Framework for Optimal Biomarker Selection in Stress Prediction

A. *PMData*

The dataset combines the lifelogging data with sports-activity logs for 16 participants over 5 months using a ”Fitbit Versa 2” fitness smartwatch. It comprises a wide range of data features including heart rate, sleep patterns, calorie consumption, exercise data, subjective wellness scores, and food intake images. In PMData, the stress levels are self-reported on a scale from 1 to 5. The PMData biomarkers that we considered for our ablation study include the following biomarkers: 1) Age, 2) Gender, 3) Burned Calories, 4) Sleep Minutes 5) Steps 6) Resting Heart Rate 7) Stress Level (self-reported) 8) Mood (self-reported).

B. *WESAD*

This is a multimodal, publicly-available dataset generated using wearable devices for stress and affect detection. It provides high-resolution physiological and motion data values from the chest-worn “RespiBAN Professional” device and lower-resolution data from the wrist-worn “Empatica E4” device. Comprising 15 participants, the dataset includes a range of sensor modalities. The chest-worn devices capture keys including 1) three-axis accelerometer data (Ch_Accelerometer), 2) electrocardiogram (Ch_ECG), 3) electromyogram (Ch_EMG), 4) electrodermal activity (Ch_EDA), 5) body temperature (Ch_Temperature), and 6) respiration data (Ch_Respiration). The wrist-worn device captures the accelerometer (Wr_Accelerometer), blood volume pulse (Wr_BVP), electrodermal activity data (Wr_EDA), and temperature (Wr_Temp). The WESAD dataset uses multi-class labeling for different psychological states: Baseline, Amusement, Stress, Meditation, and Recovery.

TABLE I
EXAMPLE OF BASE PROMPT STRUCTURE FOR PMDATA

Base Prompt Example
The recent 14-day sensor readings show: [Steps]: [...] steps, [Burned Calories]: [...] calories, [Resting Heart Rate]: [...] beats/min, [SleepMinutes]: [...] minutes, [Mood]: [...] out of 5; What would be the predicted stress in a range from 1 to 5? The response should be a single discrete integer value. Use the following format to produce output: “The predicted stress level is [Your response here]”

III. METHODOLOGY

The proposed framework for optimal biomarker selection in stress prediction is illustrated in Fig. 1.

a) Baseline Setup: We adopted the prompting strategy described in Health-LLM [8] as the baseline for our experiments. This strategy served as the foundation for devising prompts across both datasets, incorporating relevant biomarkers. An example prompt is shown in Table I. Initially, we generated approximately 300 sample prompts for each mode during the preliminary experiments, increasing this number to 500 for further testing. Each prompt was divided into two key components:
-*Preamble*: Introduces the task and provides the history of the participant with a diverse set of biomarkers.
-*Example Format*: Provides a brief example(s) to demonstrate how to respond. The example format section includes a few examples to provide the model with more context for its response.

b) Prompting Strategies and Dataset-specific Implementation: We employed 1) zero-shot 2) few-shot and 3) Chain-Of-Thought (CoT) prompting strategy on both the datasets. For few-shot prompting, we used 3 example samples for PMData and 2 for WESAD. The task requirements for each dataset were as follows:

- PMData: We created a sliding window of 14 days and considered the daily stress rating as the ground label. This approach allowed the model to predict daily stress levels based on the previous 14 days of health and activity data. The model was asked to respond on a scale of 1 to 5.
- WESAD: We converted the original multi-class psychological states (Baseline, Amusement, Stress, Meditation, and Recovery) into a binary classification of stress and no stress. We selected specific biomarkers from both chest-worn and wrist-worn devices. The data was processed to ensure consistency, likely involving normalization across different sensors and participants. We generated 500 sample prompts by taking equal data samples from each participant. We created a fixed-size window for handling different sensor frequencies. We resampled all signals (chest and wrist sensors) from their original frequencies (ranging from 4 Hz to 700 Hz) to a common 20 Hz target frequency for consistency. For every window, we generated the label by taking into account the most common value within that particular window and labeled it as stress or no stress. We used this approach to synchronize the multi-modal sensor data to preserve the temporal information and created uniform inputs suitable for LLMs and machine learning (ML) models. The model was tasked with binary classification, determining whether the individual experienced stress or not.

c) Ablation Study for Biomarker Importance: To better understand the significance of each biomarker in stress prediction, we conducted an ablation study. This involved systematically removing individual biomarkers from the prompts and formulating new prompts to assess how the absence of

TABLE II
MAE AND ACCURACY FOR PMDATA USING TRADITIONAL MACHINE LEARNING MODELS

Mode	Decision Trees		Random Forest		SVM		Gradient Boosting		kNN		Naïve Bayes		Logistic Regression	
	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
Base (B)	0.4906	0.5887	0.3766	0.6666	0.3651	0.6695	0.4141	0.6450	0.3997	0.6624	0.4343	0.6320	0.4156	0.6248
B - Burned Calories	0.4892	0.5831	0.3766	0.6666	0.3766	0.6609	0.3983	0.6507	0.3853	0.6652	0.4040	0.6421	0.4127	0.6262
B - Resting Heart Rate	0.4719	0.5974	0.3781	0.6637	0.3766	0.6594	0.4069	0.6508	0.3983	0.6566	0.4113	0.6407	0.4141	0.6262
B - Sleep Minutes	0.4733	0.6032	0.3911	0.6479	0.3795	0.6551	0.4069	0.6421	0.4242	0.6392	0.4401	0.6248	0.4156	0.6320
B - Age	0.4921	0.5859	0.3896	0.6580	0.3867	0.6450	0.4127	0.6422	0.3983	0.6537	0.3997	0.6435	0.4141	0.6248
B - Age - Gender	0.5022	0.5771	0.3853	0.6551	0.3867	0.6450	0.4069	0.6465	0.3983	0.6537	0.3997	0.6435	0.4141	0.6248
B - Gender	0.5079	0.5743	0.3737	0.6651	0.3651	0.6695	0.4141	0.6450	0.3997	0.6624	0.4343	0.6320	0.4156	0.6248
B - Steps	0.5036	0.5916	0.3795	0.6623	0.3593	**0.6753**	0.4012	0.6493	0.4156	0.6523	0.3939	0.6507	0.4084	0.6306
B - Mood	0.5642	0.5469	0.4271	0.6305	0.4084	0.6204	0.4531	0.6061	0.4531	0.6147	0.4978	0.5902	0.4329	0.6017

TABLE III
MAE AND ACCURACY FOR PMDATA IN ZERO-SHOT AND FEW-SHOT PROMPTING

Mode	BioMistralDare		Gpt4		Gpt-3.5-Turbo		Gemini-Pro (zero-shot with CoT)		Gpt-3.5-Turbo (Few-shot)		Gemini-Pro (few-shots with CoT)	
	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
Base (B)	0.8031	0.3701	0.6041	0.5323	0.5000	0.5453	0.9100	0.3100	0.6800	0.4600	0.8133	0.3833
B - Burned Calories	0.8182	0.3535	0.6600	0.4100	**0.4468**	0.5957	0.9700	0.2967	0.7600	0.4100	0.8967	0.3733
B - Resting Heart Rate	0.8761	0.3540	0.6328	0.5022	0.4799	0.5738	0.9867	0.2967	0.7100	0.4000	**0.7667**	0.4100
B - Steps	0.8464	0.3292	0.5791	0.5269	0.4992	0.5501	0.9699	0.2843	**0.6700**	0.4600	0.8100	0.3833
B - Sleep Minutes	**0.6768**	0.4747	0.7600	0.3800	0.4600	0.5700	1.0000	0.2609	0.7400	0.4200	0.7833	0.4067
B + Age	0.8182	0.3333	0.5800	0.5000	0.5000	0.5521	0.9667	0.2867	0.8300	0.3500	0.9100	0.3667
B + Gender	0.8557	0.3505	0.7700	0.3700	0.4583	0.5729	0.9467	0.3133	0.7000	0.4300	0.8467	0.3733
B + Age + Gender	0.7928	0.3890	0.6725	0.4609	0.4940	0.5554	0.9493	0.2973	0.6800	0.5000	0.7867	0.4067
B - Mood	1.0443	0.2906	**0.4362**	0.6008	0.5253	0.5152	**0.8667**	0.3433	0.9500	0.2500	0.9667	0.2867

TABLE IV
MAE AND ACCURACY FOR LLAMA 2 AND GEMINI-PRO USING ZERO-SHOT AND FEW-SHOT LEARNING FOR PMDATA

Mode	Llama 2 (Zero-Shot)		Llama 2 (Few-Shot)		Gemini-Pro (Zero-Shot)		Gemini-Pro (Few-Shot)	
	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
Base (B)	0.8294 ± 0.0297	0.3144 ± 0.0286	**0.6148 ± 0.0063**	0.4891 ± 0.0049	0.8760 ± 0.0263	0.3233 ± 0.0213	**0.8353 ± 0.0154**	0.3580 ± 0.0085
B - Burned Calories	0.9553 ± 0.0084	0.2525 ± 0.0012	0.6379 ± 0.0374	0.4739 ± 0.0290	0.8913 ± 0.0098	0.3120 ± 0.0065	0.8947 ± 0.0195	0.3233 ± 0.0096
B - Resting Heart Rate	**0.7690 ± 0.0055**	0.3804 ± 0.0174	0.6648 ± 0.0098	0.4521 ± 0.0036	0.8400 ± 0.0402	0.3460 ± 0.0188	0.8507 ± 0.0125	0.3467 ± 0.0137
B - Sleep Minutes	0.8549 ± 0.0282	0.3274 ± 0.0190	0.6739 ± 0.0146	0.4518 ± 0.0174	0.9353 ± 0.0038	0.2853 ± 0.0077	0.8680 ± 0.0102	0.3267 ± 0.0109
B - Age	0.9650 ± 0.0049	0.2409 ± 0.0137	0.6159 ± 0.0278	0.4922 ± 0.0159	0.8446 ± 0.0180	0.3249 ± 0.0060	0.8493 ± 0.0105	0.3527 ± 0.0041
B - Age - Gender	0.8931 ± 0.0094	0.2892 ± 0.0083	0.6218 ± 0.0152	0.4770 ± 0.0030	0.8853 ± 0.0335	0.3273 ± 0.0223	0.8800 ± 0.0226	0.3400 ± 0.0071
B - Gender	0.8624 ± 0.0060	0.3309 ± 0.0117	0.6272 ± 0.0074	0.4825 ± 0.0061	0.8787 ± 0.0143	0.3213 ± 0.0115	0.8840 ± 0.0302	0.3327 ± 0.0260
B - Steps	0.9009 ± 0.0091	0.2924 ± 0.0037	0.6282 ± 0.0253	0.4853 ± 0.0201	0.9233 ± 0.0115	0.2853 ± 0.0172	0.9033 ± 0.0321	0.3193 ± 0.0120
B - Mood	0.8827 ± 0.0340	0.3075 ± 0.0195	0.7896 ± 0.0110	0.3642 ± 0.0185	**0.8207 ± 0.0159**	0.3540 ± 0.0166	0.9173 ± 0.0179	0.3133 ± 0.0084

specific features impacted model performance. These variations allowed us to evaluate the effectiveness of each feature for model predictions. Key performance metrics for evaluation were: 1) Accuracy 2) Mean Absolute Error (MAE).

d) Models: We conducted experiments using four state-of-the-art LLMs: GPT (4 and 3.5-Turbo), Llama2, BioMistral-DARE [9], and Gemini-Pro. This diverse selection of models enabled a comprehensive comparison between specialized biomedical models with fewer parameters and more general-purpose models with larger parameter sets [10].

e) Experimental Validation: We initially conducted preliminary experiments, as depicted in Tables III and VI. Following this, we conducted additional experiments using the Llama2 and Gemini-Pro models to validate the preliminary findings. To ensure consistency, we repeated the experiments on the same dataset three times to calculate the standard deviation. For a baseline comparison, we evaluated our work with different ML models namely Decision Trees, Random Forest, Support Vector Machine (SVM), Gradient Boosting, k-nearest Neighbors (kNN), Naive Bayes, and Logistic Regression.

IV. RESULTS

The results are presented in Table II, III, and IV for PMData and Table V, VI, and VII for WESAD.

A. PMData Results

1) Machine Learning Models: (Table II) The results from the machine learning models are summarized in Table II. The SVM model achieved the highest accuracy of 67.53% with the lowest MAE of 0.3593 when the ‘steps’ biomarker was excluded from the input data, indicating that this feature might not contribute significantly to stress prediction in this context. The accuracy was relatively similar for most of the machine learning models, with the lowest accuracy provided by Decision Trees when ‘Mood’ was removed from the input features.

2) Preliminary Results: (Table III) For our initial experiments, we used four LLM models: GPT-4, GPT-3.5-Turbo, BioMistral-DARE, and Gemini-Pro. We created the 'Base' prompt to include all the features excluding 'Age' and 'Gender'. In our ablation study, we added the age and gender separately to assess their impact on the performance of different models.

a) BioMistral-DARE: The BioMistral-DARE model achieved its lowest MAE when the 'sleep minutes' data feature was removed from the base prompt, indicating that including sleep data might lead to the inclusion of noise for stress prediction.

b) GPT: GPT-4 achieved its best MAE when the self-reported 'mood' was removed from the base prompt whereas GPT-3.5-Turbo achieved the best performance when 'burned calories' were excluded. In the few-shot prompting technique, GPT-3.5-Turbo exhibited less performance compared to the zero-shot setting. Among the various biomarkers, it achieved the best performance when 'steps' were removed from the base prompt.

c) Gemini-Pro: In the zero-shot with CoT, we observed inconsistent performance across different biomarkers, and the best performance was observed when 'mood' was removed from the prompt. There was a slight performance increment when we included few-shot (3-shot) prompting with CoT reasoning without explicitly explaining the type of reasoning required. The overall performance was better for the few-shot with CoT prompting techniques compared to zero-shot CoT and the best performance of 0.7667 MAE was achieved when we removed 'Resting Heart Rate' from the base prompt. This shows that including examples in the prompt helped contribute to an increase in performance.

d) Influence of Age and Gender: To assess the impact of demographic factors like age and gender, we calculated the average performance of the base prompt and compared it to the results when age and gender were included. Figure 2 highlights the average improvement across datasets for the zero-shot prompting technique. The results indicate that including age into the base, prompt led to a significant improvement of 8.49% compared to average performance. Similarly, adding gender also contributed to performance enhancement, though its impact was smaller, with only a 0.39% increase.

3) Ablation Study: (Table IV) To further our experiments, we incorporated the Llama2 and Gemini-Pro models, conducting each experiment three times to validate our results and calculate the standard deviation. We also modified the 'Base' prompt to include all features, such as Age and Gender. In our ablation study, we removed the age and gender biomarkers individually to evaluate their impact on the performance of both models. Furthermore, we also used traditional machine learning models to act as a baseline for comparison.

a) Gemini-Pro: For the zero-shot settings, the Gemini-Pro model achieved the best performance of 0.8207 MAE when the 'mood' biomarker was removed from the prompt. The Base model had the best performance with few-shot learning with 0.8353 MAE. While Gemini-Pro showed robustness in zero-shot settings, it still managed to improve performance from few-shot learning. In conclusion, adding examples to the prompt always helped increase the accuracy of Gemini-Pro, and the zero-shot technique always performed best when 'mood' was excluded from the prompt, suggesting that mood information isn't a significant biomarker in predicting stress.

b) Llama 2: For Llama 2, the few-shot learning approach consistently outperformed zero-shot across all biomarker configurations. For the zero-shot, the model achieved the best performance with 0.7690 MAE when the 'Resting heart rate' biomarker was removed from the prompt. For the few-shot, the model achieved the best performance during the base configuration when all of the biomarkers were present in the prompt. Overall, Llama 2 benefits significantly from few-shot learning, showing improved prediction accuracy when provided with examples, and performs best when all biomarkers are included.

B. WESAD Results

1) Machine Learning Models: (Table V) The results from the machine learning models for WESAD are summarized in the table V. The Random Forest model achieved the highest accuracy of 91.13% with the lowest MAE of 0.0887 when the 'gender' biomarker was excluded from the input data.

2) Preliminary Results: (Table VI)

a) BioMistral-DARE: For the WESAD dataset, BioMistral-DARE demonstrated its optimal performance when the 'chest_ECG' data was removed from the base prompt with 0.73 MAE. This suggests that ECG data might introduce noise for the WESAD dataset for stress state prediction. The model's performance deteriorated most when we removed the wrist accelerometer data from the base prompt with 0.86 MAE.

b) GPT: For the GPT models, for the zero-shot prompting techniques, GPT-4 and GPT-3.5-Turbo exhibited their best performance (MAE of 0.8067 and 0.8133 respectively) when the 'Ch_Accelerometer' data and 'Ch_ECG' data was excluded from the base prompt respectively. The GPT models didn't perform well with the dataset for zero-shot despite its tremendous parameter size and responded that the patient was 'stressed' for almost every sample. However, introducing examples drastically increased the accuracy of the GPT-3.5-Turbo. Comparing the results from zero-shot and few-shot for GPT-3.5-Turbo revealed the potential benefits of introducing examples to the GPT to improve stress prediction in WESAD data.

c) Gemini-Pro: The zero-shot prompting with CoT provided the best performance of 0.4825 MAE when we removed the 'Ch_Accelerometer' biomarker. The performance further improved when we employed few-shot prompting (2-shot) with CoT. The best performance was achieved when we removed the 'Wr_Respiration' from the base prompt.

d) Influence of age and gender: Keeping in mind that we removed the biomarkers in WESAD from the base prompt, we examined the influence of demographic biomarkers such as age and gender as depicted in Figure 2. Removing age had a negative impact on the performance, which shows that 'age'

TABLE V
MAE AND ACCURACY FOR WESAD USING TRADITIONAL MACHINE LEARNING MODELS

Mode	Decision Trees		Random Forest		SVM		Gradient Boosting		kNN		Naïve Bayes		Logistic Regression	
	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
Base (B)	0.1312	0.8688	0.1004	0.8996	0.2124	0.7876	0.1613	0.8387	0.2876	0.7124	0.4233	0.5767	0.3737	0.6263
B - Age	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Gender	0.1440	0.8560	0.0887	**0.9113**	0.2023	0.7977	0.1617	0.8383	0.2838	0.7162	0.4207	0.5793	0.3778	0.6222
B - Ch_Accelerometer	0.1759	0.8241	0.1425	0.8575	0.2534	0.7466	0.1733	0.8267	0.2891	0.7109	0.5425	0.4575	0.4658	0.5342
B - Ch_ECG	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Ch_EMG	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Ch_EDA	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Ch_Temperature	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Ch_Respiration	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Wr_Accelerometer	0.1538	0.8462	0.1372	0.8628	0.2695	0.7305	0.1989	0.8011	0.3365	0.6635	0.4989	0.5011	0.4282	0.5718
B - Wr_BVP	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Wr_EDA	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203
B - Wr_Temp	0.1489	0.8511	0.0929	0.9071	0.2128	0.7872	0.1602	0.8398	0.2590	0.7410	0.4094	0.5906	0.3797	0.6203

TABLE VI
MAE AND ACCURACY FOR WESAD IN ZERO-SHOT AND FEW-SHOT PROMPTING

Mode	BioMistralDare		Gpt4		Gpt-3.5-Turbo		Gemini-Pro (zero-shot with CoT)		Gpt-3.5-Turbo (Few-shot)		Gemini-Pro (few-shots with CoT)	
	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
Base (B)	0.7548	0.2452	0.8502	0.1498	0.8467	0.1533	0.4965	0.5035	0.6479	0.3521	0.2628	0.7372
B - Age	0.7357	0.2643	0.8467	0.1533	0.8333	0.1667	0.6312	0.3688	**0.5362**	0.4638	0.1986	0.8014
B - Gender	0.7410	0.2590	0.8467	0.1533	0.8400	0.1600	0.5775	0.4225	0.6642	0.3358	0.2446	0.7554
B - Ch_Accelerometer	0.7640	0.2710	0.8333	0.1667	0.8400	0.1600	**0.4825**	0.5175	0.5986	0.4014	0.2397	0.7603
B - Ch_ECG	**0.7321**	0.2679	0.8133	0.1867	**0.8133**	0.1867	0.6809	0.3191	0.5890	0.4110	0.2500	0.7500
B - Ch_EMG	0.7512	0.2488	0.8467	0.1533	0.8200	0.1800	0.6691	0.3309	0.6014	0.3986	0.2993	0.7007
B - Ch_EDA	0.7593	0.2407	0.8467	0.1533	0.8467	0.1533	0.5944	0.4056	0.6204	0.3796	0.2690	0.7310
B - Ch_Temperature	0.8376	0.1624	0.8467	0.1533	0.8467	0.1533	0.6043	0.3957	0.6875	0.3125	0.2345	0.7655
B - Ch_Respiration	0.8500	0.1500	0.8467	0.1533	0.8400	0.1600	0.5694	0.4306	0.5474	0.4526	**0.1655**	0.8345
B - Wr_Accelerometer	0.8557	0.1443	**0.8067**	0.1933	0.8333	0.1667	0.5603	0.4397	0.5850	0.4150	0.2639	0.7361
B - Wr_BVP	0.8462	0.1538	0.8133	0.1867	0.8523	0.1477	0.6042	0.3958	0.6547	0.3453	0.1888	0.8112
B - Wr_EDA	0.8434	0.1566	0.8267	0.1733	0.8267	0.1733	0.6619	0.3381	0.6408	0.3592	0.2937	0.7063
B - Wr_Temperature	0.8500	0.1500	0.8400	0.1600	0.8400	0.1600	0.6148	0.3852	0.5493	0.4507	0.2690	0.7310

TABLE VII
MAE AND ACCURACY FOR LLAMA 2 AND GEMINI-PRO USING ZERO-SHOT AND FEW-SHOT LEARNING FOR WESAD

Mode	llama 2 (Zero-shot)		llama 2 (Few-shot)		Gemini-Pro (Zero-shot)		Gemini-Pro (Few-shot)	
	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
Base (B)	**0.8724 ± 0.0009**	0.1276 ± 0.0009	0.8795 ± 0.0000	0.1205 ± 0.0000	0.5350 ± 0.0110	0.4650 ± 0.0110	0.1590 ± 0.0060	0.8410 ± 0.0060
B - Age	0.8766 ± 0.0007	0.1234 ± 0.0007	0.8803 ± 0.0010	0.1197 ± 0.0010	0.5470 ± 0.0090	0.4530 ± 0.0090	0.1596 ± 0.0010	0.8404 ± 0.0010
B - Gender	0.9033 ± 0.0005	0.1067 ± 0.0000	0.8794 ± 0.0003	0.1206 ± 0.0003	0.5190 ± 0.0150	0.4810 ± 0.0150	0.1693 ± 0.0070	0.8307 ± 0.0070
B - Ch_Accelerometer	0.9027 ± 0.0003	0.1073 ± 0.0003	0.8789 ± 0.0008	0.1211 ± 0.0008	0.5330 ± 0.0030	0.4670 ± 0.0030	0.1816 ± 0.0112	0.8184 ± 0.0112
B - Ch_ECG	0.8992 ± 0.0032	0.1008 ± 0.0032	0.8781 ± 0.0010	0.1219 ± 0.0010	0.5670 ± 0.0010	0.4330 ± 0.0010	0.1602 ± 0.0078	0.8398 ± 0.0078
B - Ch_EMG	0.8995 ± 0.0002	0.1005 ± 0.0002	0.8769 ± 0.0026	0.1231 ± 0.0026	0.5586 ± 0.0006	0.4414 ± 0.0006	**0.1504 ± 0.0096**	0.8496 ± 0.0096
B - Ch_EDA	0.9027 ± 0.0000	0.0973 ± 0.0000	0.8794 ± 0.0001	0.1206 ± 0.0001	0.5860 ± 0.0020	0.4140 ± 0.0020	0.1824 ± 0.0160	0.8176 ± 0.0160
B - Ch_Temperature	0.9000 ± 0.0000	0.1000 ± 0.0000	0.8778 ± 0.0015	0.1222 ± 0.0015	0.5830 ± 0.0290	0.4170 ± 0.0290	0.1713 ± 0.0047	0.8287 ± 0.0047
B - Ch_Respiration	0.9015 ± 0.0018	0.0985 ± 0.0018	0.8796 ± 0.0001	0.1204 ± 0.0001	0.5305 ± 0.0255	0.4695 ± 0.0255	0.1670 ± 0.0170	0.8330 ± 0.0170
B - Wr_Accelerometer	0.9028 ± 0.0002	0.0972 ± 0.0002	0.8768 ± 0.0020	0.1232 ± 0.0020	**0.4790 ± 0.0130**	0.5210 ± 0.0130	0.1918 ± 0.0094	0.8082 ± 0.0094
B - Wr_BVP	0.8938 ± 0.0018	0.1062 ± 0.0018	0.8772 ± 0.0032	0.1228 ± 0.0032	0.5536 ± 0.0004	0.4464 ± 0.0004	0.1662 ± 0.0078	0.8338 ± 0.0078
B - Wr_EDA	0.9028 ± 0.0040	0.0972 ± 0.0040	**0.8763 ± 0.0007**	0.1237 ± 0.0007	0.5230 ± 0.0050	0.4770 ± 0.0050	0.1754 ± 0.0030	0.8246 ± 0.0030
B - Wr_Temperature	0.9013 ± 0.0027	0.0987 ± 0.0027	0.8781 ± 0.0012	0.1219 ± 0.0012	0.5340 ± 0.0080	0.4660 ± 0.0080	0.1606 ± 0.0020	0.8394 ± 0.0020

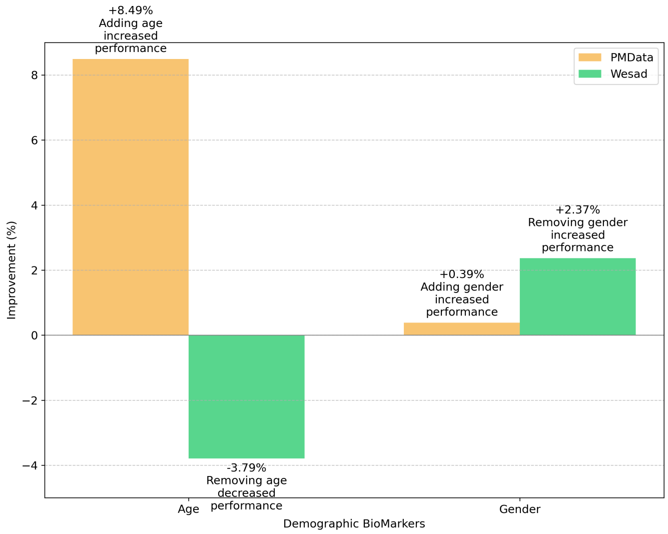

Fig. 2. Influence of Age and Gender on Performance

plays a vital role in the stress prediction for WESAD. We saw a performance decrease of 3.79% when we removed age from the prompt. However, we found that removing 'gender' has a positive impact on performance with zero-shot prompting as it led to an increase of 2.37% in performance.

3) Ablation Study: (Table VII)

a) Gemini-Pro: In the zero-shot configuration, Gemini-Pro achieved its best performance when we removed the 'Wr_Accelerometer' biomarkers, with MAE of 0.4790 and corresponding accuracy values of 52.10%. However, the few-shot configuration significantly improves Gemini-Proś performance across all the configurations, with the lowest MAE of 0.1504 when we excluded the 'Ch_EMG' biomarker and the highest accuracy of 84.96%. This demonstrates that Gemini-Pro performs better with few-shot learning, showing more accurate and consistent predictions compared to zero-shot.

b) Llama2: The Llama 2 model performed poorly on the WESAD dataset. For most samples, Llama2 predicted that the patient was in a 'stressed' state, but since the majority of our ground truth labels were 'no stress', its accuracy for the zero-shot prompting strategy was very low, with a best performance of 0.8274 MAE when all biomarkers were included in the base prompt.

V. Conclusion and Future Works

a) Model Performance: The study tested multiple wearable stress-related datasets on four open-source LLM models, including BioMistralDARE, Gemini-Pro, and GPT variants, to assess the impact of parameter count on stress prediction. Results showed that parameter size did not always correlate with prediction accuracy. Smaller models like GPT-3.5-Turbo performed comparably to larger ones like GPT-4 on different datasets (PMData and WESAD). Gemini-Pro even outperformed GPT-4 on WESAD, suggesting dataset characteristics and model architecture may play a more significant role than parameter count alone.

b) Impact of Age: Age was found to influence stress prediction accuracy, highlighting its importance in tailoring stress management interventions. Including age data in the models consistently improved performance, particularly in zero-shot settings for both PMData and WESAD datasets.

c) Impact of Gender: The study examined whether gender impacted prediction accuracy. Results indicated that gender did not significantly affect the models' performance as compared to age, suggesting that stress prediction models could be designed to avoid gender-based biases.

d) Optimal Biomarkers: Identifying key biomarkers was essential for improving model efficiency. PMData had model-specific optimal biomarkers, while WESAD showed better stress prediction accuracy when 'chest_ECG' was excluded. The analysis suggests that biomarker selection varies by dataset and model, but further research is needed to explore larger, more diverse datasets and biomarkers for more conclusive results.

e) Future Works: Future work will involve a comprehensive evaluation of additional stress datasets, such as CLAS [13] and MMASH [14]. Additionally, we plan to fine-tune models like FLAN-T5, BioMedGPT, and BioMistral to further assess the accuracy of stress predictions. Moreover, we intend to thoroughly evaluate the explainability of explanations provided by large language models.

References

[1] Y. Haque et al., "State-of-the-art of stress prediction from heart rate variability using artificial intelligence". *Cognitive Computation*, vol. 16, no. 2, pp.455-481, 2024.

[2] E. Lazarou, and T.P. Exarchos, "Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices". *AIMS Neuroscience*, vol. 11, no. 2, pp.76-102, 2024.

[3] X. Meng et al., "The application of large language models in medicine: A scoping review", *iScience*, vol. 27, no. 5, 2024.

[4] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review", *Informatics*, vol. 11, no. 3, p. 57, 2024.

[5] D. Slack et al., "Explaining machine learning models with interactive natural language conversations using TalkToModel", *Nature Machine Intelligence*, vol. 5, no. 8, pp. 873-883, 2023.

[6] A. Belyaeva et al., "Multimodal LLMs for health grounded in individual-specific data", *Workshop on Machine Learning for Multimodal Healthcare Data, Cham: Springer Nature Switzerland*, 2023.

[7] S. Goyal et al., "HealAI: A healthcare LLM for effective medical documentation", *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024.

[8] Y. Kim et al., "Health-LLM: Large language models for health prediction via wearable sensor data", *arXiv preprint arXiv:2401.06866*, 2024.

[9] Y. Labrak et al. "BioMistral: A collection of open-source pre-trained large language models for medical domains", *arXiv preprint arXiv:2402.10373*, 2024.

[10] H. Yang et al., "One LLM is not enough: Harnessing the power of ensemble learning for medical question answering", *medRxiv*, 2023.

[11] V. Thambawita et al., "PMData: A sports logging dataset", *Proceedings of the 11th ACM Multimedia Systems Conference (MMSys '20)*, pp. 231-236, 2020.

[12] P. Schmidt et. al., "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection", 2018.

[13] V. Markova et. al., "Clas: A database for cognitive load, affect and stress recognition", *Proceedings of the 2019 International Conference on Biomedical Innovations and Applications (BIA)*, pp. 1–4, 2019.

[14] A. Rossi et al., "Multilevel monitoring of activity and sleep in healthy people", *PhysioNet*, 2020.