Prenatal Cortisol Levels Estimation Using Heart Rate and Heart Rate Variability: A Weak Supervised Learning Based Approach

Rui Cao[1,*], Yong Huang[2], Amir M. Rahmani[2,3,4], and Karen Lindsay[5,6]

***Abstract*— Cortisol is a steroid hormone that regulates a wide range of vital signs throughout the body. However, current cortisol monitoring methods are inconvenient for everyday settings. Heart Rate (HR) and Heart Rate Variability (HRV) are easily collected biological parameters whose fluctuations highly correlate with cortisol, however, there does not exist a work attempting to estimate cortisol levels using these signals. In this paper, to the best of our knowledge, for the first time, we propose a machine learning-based salivary cortisol level estimation method using HR and HRV collected from pregnant women wearing an ECG chest strap. We first extract HR and HRV parameters from inter-beat-interval data derived from electrocardiogram signals. Then, we apply a feature selection algorithm to select the most contributing features and introduce a machine learning-based weak supervision method to address the unbalanced number of labels collected in real settings. Five machine learning algorithms are implemented to perform binary classification of baseline cortisol level (BL) versus two distinct cortisol levels (CL1 and CL2). One deep neural network is used to perform the classification across all three levels. As a pioneer study, we obtain prediction accuracy of up to 69% (BL VS. CL1), 71% (BL VS. CL2), and 60% (BL VS. CL1 VS. CL2).**

I. INTRODUCTION

Cortisol is a hormone that affects almost every organ and tissue across the human body. It plays a crucial role in helping individuals to respond to stress, fight infection, regulate blood sugar, maintain blood pressure, etc. [1]. During pregnancy, cortisol is a critical factor for fetal development and avoiding preterm birth [2]. For these reasons, cortisol level is considered an important indicator of both physical and mental health status. However, current methods for measuring cortisol are inconvenient as they require collecting samples from blood, urine, or saliva at a clinic.

Heart Rate and Heart Rate Variability are parameters extracted from electrocardiogram (ECG) or Photoplethysmography (PPG) signals [3]. These two commonly-used biosignals can be easily collected using a chest strap, smartwatch, or ring[4]. Both HR and HRV parameters have been shown to be significantly correlated with the fluctuation of cortisol levels [5]. However, there does not exist a work attempting to estimate cortisol levels using these signals collected from wearables.

In this paper, to the best of our knowledge, for the first time, we develop a cortisol level estimation method by only using HR and HRV parameters collected from pregnant mothers using a wearable ECG monitoring device. Many challenges are involved in this study such as spare cortisol measurements, missing data, presence of noises like motion artifacts, and unbalanced labels distribution since several factors such as the concentration, lasting time, distribution, and reaction to the psychological stimulation as well as the environment can not be controlled by researchers. Our method first derives inter-beat-interval data from ECG signals and extracts HR and HRV parameters. After that, a feature selection algorithm is applied to search for the most informative features to reduce the method's computational complexity. Then, we introduce a machine learning-based weak supervision method [6] to increase the number of cortisol level labels and balance labels distribution to improve the model performance. Five machine learning algorithms and one deep neural network were implemented for the estimation. As the first attempt, our method obtains acceptable cortisol levels estimation accuracy using only HR and HRV parameters recorded by a chest strap. The contribution of the study can be summarized into three folds:

- We propose a cortisol level estimation method using only HR and HRV parameters with an acceptable accuracy.
- We ran a human subject study on pregnant mothers and built predictive models to estimate their cortisol levels using HR and HRV extracted from ECG signals recorded by a chest strap. This shows the promises of our method to be used in everyday settings using wearable sensors.
- We introduce a novel solution to augment the sparsely labeled dataset based on weak supervision.

The subsequent sections are organized as follows. Section 2 describes the study design and data collection. Section 3 presents our proposed method and an entire pipeline of feature selection to the weak supervision algorithm. Section 4 presents the experimental results. Section 5 describes the analysis and discussion. Section 6 concludes the paper.

II. STUDY DESIGN & DATA COLLECTION

The salivary cortisol and HRV were collected from a crossover study that aimed to assess the effects of superimposed psychological stress on the postprandial metabolic response to a standardized breakfast meal during pregnancy. In this study, the Trier Social Stress Test[7], a standardized psychological stress challenge task was used to stimulate changes in the cortisol level.

[1]Dept. of Electrical Engineering and Computer Science, [2]Dept. of Computer Science, [3]School of Nursing, [4]Institute for Future Health (IFH), University of California, Irvine, [5]UCI Susan Samueli Integrative Health Institute, Susan Henry Samueli College of Health Sciences, University of California, Irvine, CA, [6]Department of Pediatrics, Division of Endocrinology, University of California, Irvine, School of Medicine, Orange, CA, (*correspondence e-mail: caor6@uci.edu)

The data collection process involved two visits to the clinical research facility, during which participants underwent either a control non-stress task or an acute psychosocial stress challenge task. Saliva samples were collected using Salimetrics oral swabs, which were placed under the tongue for 2 minutes at each collection time point. An initial saliva sample was collected before consuming a milkshake drink, which was the same at both visits. A second saliva sample was collected 15 minutes after the initial sample. At the same time as the second saliva sample, an Actiheart electrocardiograph monitor (CamNtech Ltd.) was placed on the participant's chest to measure heart rate and inter-beat interval continuously. All these were followed by the task period (15-minute duration), and post-task saliva sample collection immediately after the task (30 minutes post-baseline sample). Subsequent saliva sample collections occurred at 45, 60, 90, and 120 minutes post-baseline. Salivary cortisol values were measured from the saliva samples collected during the two visits. Due to the time of placing the Actiheart device, all the cortisol values were used in the following assessment study except the first two without the corresponding HR and IBI recordings.

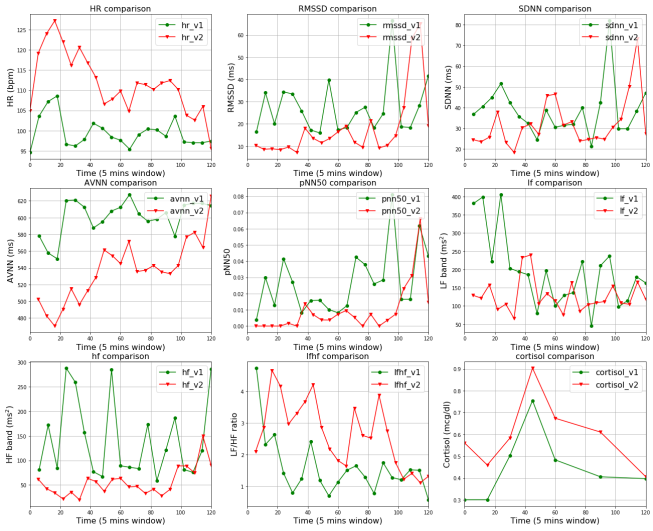

Fig. 1: HR, HRV and salivary cortisol values of one participant between two visits

TABLE I: Extracted HRV features and their descriptions

Feature	Unites	Description
IBI	ms	The time interval between two successive heartbeats.
RMSSD	ms	The root mean square of successive differences between adjacent normal NN intervals.
AVNN	ms	The average value of normal NN intervals.
SDNN	ms	The standard deviation of normal NN intervals.
pNN50	-	The proportion of the number of pairs of successive NN intervals.
LF	ms^2	Standard deviation of the signal.
HF	ms^2	The average value of the inhale peak intervals.
LF/HF	-	The standard deviation of the inhale peak intervals.

III. Method

A. Feature Extraction

Time-domain HRV parameters (i.e., RMSSD, AVNN, SDNN, and pNN50) and frequency-domain HRV parameters (i.e., LF, HF, and LF/HF ratio) were statistically extracted from 5-minute time windows using the Inter-beat Intervals provided by the Actiheart. This indicates that all the saliva cortisol values are matched to their nearest 5 minutes feature window. The extracted HRV parameters are briefly described in Table I. Motion artifacts are unavoidable during the data collection process. Therefore, we removed the abnormal IBI and HRV values affected by the motion artifacts before assessing cortisol levels. Figure 1 shows one participant's HR, HRV, and salivary cortisol in two colors corresponding the two visits.

B. Feature Selection

In order to provide generalizability and prevent overfitting in the cortisol levels assessment model, we introduced a feature selection to our algorithm. We implemented a filter-based feature selection method because of its benefits including less computational intensity and lower risk of overfitting [6]. This algorithm determines the relationship between various input features and target labels statistically. To compare the importance of each element in our assessment model, we applied Gini impurity gain in the filter-based feature selection method. We used a Random Forest Classifier based on a Decision-Tree structure to output the feature importance vectors. A condition on one of the features was assigned to each node inside the decision tree algorithm. We set the Gini impurity of the features chosen in every node as the splitting condition. The ultimate goal of these tree nodes is to categorize the data into two sets. Data with the same label should be put into one group in the ideal scenario. During the selection process, we calculate the contribution to decrease the impurity of each feature and utilize it as an informative measurement to rank features.

C. Features Labeling Method

We labeled the salivary cortisol values into three levels. Since all data collections were conducted during the morning (8 AM to 11 AM) from pregnant women, we took the average value of salivary cortisol measured in a previous study of pregnant women [8] as the threshold of baseline cortisol level (BL). As for the other two levels, the threshold was chosen carefully to separate the remaining cortisol values evenly so that their distribution is balanced for the following assessment model. We call cortisol level 1 (CL1) and level 2 (CL2), respectively. These labels distribution is shown in Figure 2. As can be seen from this table, unbalanced label distribution is an inherent challenge for further classification. The number of CL1 and CL2 is limited compared to the baseline. Besides, we only have 627 labeled windows compared to 1855 unlabeled windows (75%). These labels were collected from pregnant women using only stress tests to try to increase cortisol value during clinical visits. Such unbalanced distribution is unavoidable because of this study's

realistic nature. We cannot increase the frequency of high cortisol levels among the participants through other methods (besides short-term stress tasks) due to the health concern for both the mother and the fetus.

To address this problem, we exploited a weak supervision algorithm and tool called "Snorkel" into our labeling process to label the unlabeled feature windows. Snorkel is an end-to-end method that utilizes a weak supervision method to mark a training dataset when ground truth data is limited. According to our scenario, this method perfectly fits our case to solve the unbalanced nature of our labels. In this study, we considered all the 627 original labels collected directly from pregnant women as "strong" labels in the process of labeling function training. Each subject's strongly labeled windows were only used to mark their unlabeled windows. We call the remaining data points labeled by the Snorkel as "weak" data. These weakly labeled data were only used in our training process (not in the validation process) to ensure a fair evaluation of the accuracy of cortisol levels. In other words, our model's final performance is measured using only actual data collected from participants. The label's distribution after using the snorkel labeling function is summarized in Figure 2 as well. As a result, we have more labels for training.

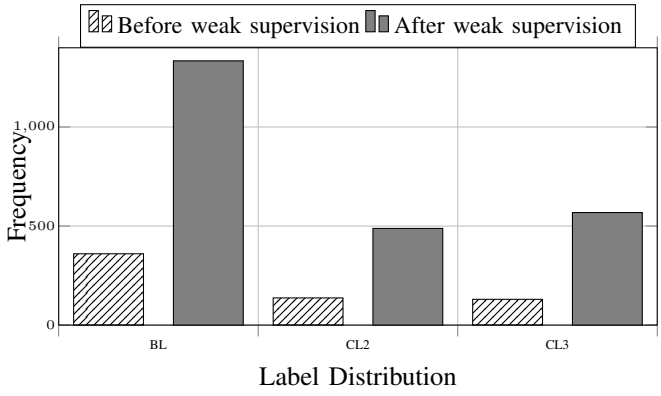

Fig. 2: Label distribution of 3 classes before and after using weak supervision labeling

D. Machine and Deep Learning based Predictive Models for Cortisol level Estimation

Machine learning algorithms were used to build the cortisol levels prediction models. We applied five classification approaches to build models, including AdaBoost, XGBoost, Random Forest, Support Vector Machine (SVM), and k-nearest neighbor (KNN). In addition to the machine learning classifiers, we also implemented a deep neural network (DNN) with two fully connected layers and ReLU activation in between to classify among three cortisol levels. The neural network's advantage is its capacity to learn non-linear representations to capture complex relations of features.

We used the Leave-one-subject-out cross-validation strategy to evaluate the performance of our classification models in terms of generalizability. We only included participants' strong labels into the test set during the cross-validation process. A combination of strong labels and weak labels from the rest of the participants was added to the training process.

IV. Results

Four features were selected for the classification between BL and CL1, including AVNN, LF, HF, and pNN50. Another combination of four features was chosen to classify BL and CL2, including AVNN, HR, LF/HF, and SDNN. The cortisol level estimation results using five classifiers and the Deep Neural Network are shown in Figure 3. All the values shown in the figure are the average accuracy across all participants. We summarize the final accuracy in Table II.

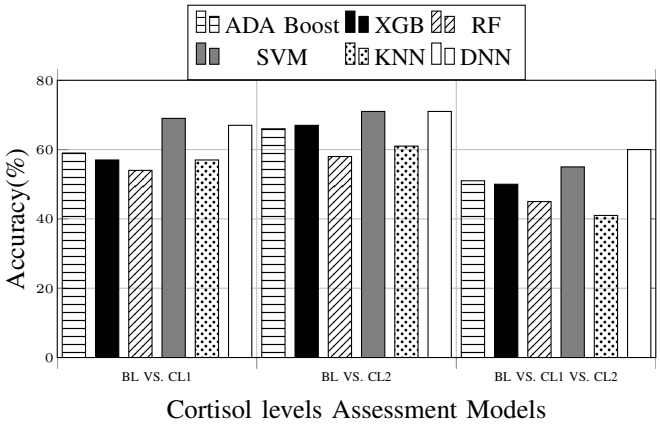

Fig. 3: Validation accuracy of all classifiers

TABLE II: Final accuracy of six models

Cortisol levels	ADA Boost	XGB	RF	SVM	KNN	DNN
BL VS. CL1	59	57	54	**69**	57	67
BL VS. CL2	66	67	58	**71**	61	**71**
BL VS. CL1 VS. CL2	51	50	45	55	41	**60**

The SVM classifier achieved the highest performance among the machine learning-based classification models, with an average accuracy of 69% between BL and CL1 and 71% between BL and CL2. The classification accuracy between BL and CL2 is always higher than BL VS. CL1 across all five machine learning-based models. The accuracy of the SVM classifier is acceptable given that only four HR and HRV features were used in the model, making it very lightweight. The deep neural network model had the same 71% accuracy in classifying BL and CL2 and performed the best among three-level classifications (i.e., BL VS. CL1 VS. CL2) with 60% precision.

V. Discussion

To the best of our knowledge, we are the first study to develop a salivary cortisol level estimation model using HR and HRV features among pregnant women. The only existing cortisol level assessment study in the literature uses Electroencephalogram (EEG) signals recorded by a helmet [9]. Although their binary classification accuracy between

low and high cortisol values is acceptable, their method is infeasible to be used for monitoring in everyday settings. Binary labeling of cortisol values to low and high is less accurate and practical as well. The range of cortisol values can change drastically in response to stressors, time of day, pregnancy state, etc [10]. Categorizing the values into two levels would not provide meaningful clinical interpretation of an individual's health state.

For these reasons, we increased the cortisol levels to three and used HR and HRV signals for the estimation. Unlike multi-channel EEG collected by a helmet, HR and HRV parameters can be easily monitored using a chest band, a smartwatch, or even a ring in everyday life, which is a feasible approach for most research studies and daily monitoring requirements. One would expect to produce a less acceptable accuracy because of increasing the number of cortisol level classes and using HR and HRV instead of EEG, which are sensitive to various noises and have less association with the nature of cortisol comparedal accuracy between different cortisol levels is promising despite being trained on a dataset collected in a harsher and more realistic setting (e.g., environmental noise, motion artifacts due to movements, unbalanced labels, etc.). The SVM classifier achieved the best performance with an accuracy of 69% for CL1 and 71% for CL2. Estimation of BL VS. CL2 has higher accuracy than BL VS. CL1 because more clear fluctuation among HR and HRV features happen for CL2. The considerable increase of cortisol values probably causes that change and helps our models to better differentiate the CL2. Our deep neural network has the best performance in three-level classification with 60% accuracy. The deep neural network works better than the five basic machine learning models when the labels are more diverse. One limitation of our study is that an average salivary cortisol value for pregnant women at the same pregnant stage and settings was not available in the literature. Thus, we considered the salivary value from pregnant women in similar settings as the threshold value for the baseline level, which may have resulted in an accuracy loss in our estimation model.

Identification and customization of suitable machine learning methods contribute to our promising results. The filter-based feature selection method presents the importance of HR and HRV parameters for different levels and reduces computational complexity. Our weak supervision labeling strategy using Snorkel mitigated the problem of having limited labels in our study which is one of the main issues in lab settings. We were not able to induce high cortisol levels by using intense stimulation due to the concern for maternal health, yet weak supervision addressed this problem significantly. In addition to providing more high-level labels for training, it also reduces the workload of saliva sample collection. We believe the proper use of these two methods enhanced the reliability of our cortisol level assessment results.

In summary, our novel method can assess salivary cortisol levels of pregnant women using HR and HRV parameters, which shows high promises to be used in everyday settings. Future work can focus on increasing assessment accuracy by incorporating more biosignals such as galvanic skin response (GSR) and applying more advanced machine learning methods.

VI. Conclusions

This paper proposed a novel salivary cortisol estimation method using HR and HRV parameters recorded by a chest strap on pregnant women. Our method produces acceptable accuracy despite using simple HR and HRV parameters. The estimation accuracy is up to 69% (BL VS. CL1), 71% (BL VS. CL2), and 60% (BL VS. CL1 VS. CL2), respectively, showing our method's promises in cortisol level estimation for a routine use. Moreover, our approach paves the way for future clinical research studies and wearable device manufacturers to estimate cortisol levels conveniently with an acceptable accuracy.

VII. ACKNOWLEDGEMENT

This research was supported in part by the U.S. National Science Foundation through the UNITE Project under Grant SCC CNS-1831918 and the National Institute of Health grant number K99/R00 HD-096109 (KL).

References

[1] L. Thau, J. Gandhi, and S. Sharma, "Physiology, cortisol," *StatPearls [Internet]*, 2021.

[2] K. Shaikh, S. Premji, K. Khowaja, S. Tough, A. Kazi, and S. Khowaj, "The relationship between prenatal stress, depression, cortisol and preterm birth: A review," *Open Journal of Depression*, vol. 2, no. 3, p. 24, 2013.

[3] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[4] R. Cao, I. Azimi, F. Sarhaddi, H. Niela-Vilen, A. Axelin, P. Liljeberg, A. M. Rahmani *et al.*, "Accuracy assessment of oura ring nocturnal heart rate and heart rate variability in comparison with electrocardiography in time and frequency domains: Comprehensive analysis," *Journal of Medical Internet Research*, vol. 24, no. 1, p. e27487, 2022.

[5] R. R. Looser, P. Metzenthin, S. Helfricht, B. M. Kudielka, A. Loerbroks, J. F. Thayer, and J. E. Fischer, "Cortisol is significantly correlated with cardiovascular responses during high levels of stress in critical care personnel," *Psychosomatic Medicine*, vol. 72, no. 3, pp. 281–289, 2010.

[6] Ratner, A. others, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, vol. 11, no. 3. NIH Public Access, 2017, p. 269.

[7] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[8] C. Obel, M. Hedegaard, T. Henriksen, N. Secher, J. Olsen, and S. Levine, "Stress and salivary cortisol during pregnancy," *Psychoneuroendocrinology*, vol. 30, no. 7, pp. 647–656, 2005.

[9] B. A. Savareh, A. Bashiri, M. M. Hatef, and B. Hatef, "Prediction of salivary cortisol level by electroencephalography features," *Biomedical Engineering/Biomedizinische Technik*, vol. 66, no. 3, pp. 275–284, 2021.

[10] K. Gustafsson, P. Lindfors, G. Aronsson, and U. Lundberg, "Relationships between self-rating of recovery from work and morning salivary cortisol," *Journal of occupational health*, vol. 50, no. 1, pp. 24–30, 2008.