

INTEGRATION OF TRANSCRIPTOMIC AND HIGH FREQUENCY TELEMETRY DATA VIA MACHINE LEARNING METHODS

by

TAO SHENG

(Under the Direction of Juan B. Gutierrez)

ABSTRACT

Machine learning has become one of the hottest topics in the last decade and is a driver of the current state of the art in various fields. In this study, we used machine learning techniques to integrate different frequency telemetry data and transcriptomic data for predicting and understanding malaria parasite infection. A python API was developed to query and download different data types stored in a cloud-based relational database. A deep neural network (DNN) model was built to predict pre-infection and the liver stage of malaria infection by using raw electrocardiogram (ECG) data. The ECG data were grouped into 24 hours of a day according to its recording time and then fed into the DNN model separately. The classification results varied largely with different ECG groups that recorded during different hours. The ECG recordings from H21 have the highest prediction F1 score of 0.831 and 0.781 on the validation and independent testing data set, respectively. While the classification accuracy dropped to 60% level with the ECG data recorded during daytime hours. Generally, the DNN model performed much better with the ECG data recorded during nighttime hours than daytime. We further proved that there were true ECG signal changes in early liver stage by integration of accelerometer data to the DNN model. We found that the differences of the model prediction accuracy with different ECG data groups were highly associated with the

corresponding accelerometer data of the hosts, which indicated that activity could impact the stability of ECG signals, thus affecting the DNN model prediction performance. Furthermore, the differential expression and gene set enrichment analysis of transcriptomic data showed the disruption of circadian rhythms in the liver stage. The disruption of circadian rhythms provided an intrinsic reason for cardiac involvement in malaria parasite infections.

INDEX WORDS: Machine Learning, Deep Neural Networks, Telemetry, Electrocardiogram, Accelerometers, Transcriptomics, Malaria

INTEGRATION OF TRANSCRIPTOMIC AND HIGH FREQUENCY TELEMETRY DATA VIA
MACHINE LEARNING METHODS

by

TAO SHENG

B.S. Nankai University, China, 2014

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

© 2019

Tao Sheng

All Rights Reserved

INTEGRATION OF TRANSCRIPTOMIC AND HIGH FREQUENCY TELEMETRY DATA VIA
MACHINE LEARNING METHODS

by

TAO SHENG

Approved:

Major Professor: Juan B. Gutierrez

Committee: Jessica C. Kissinger
Liang Liu
Jonathan Arnold

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

DEDICATION

This thesis is dedicated to my parents and my loved data driven science.

ACKNOWLEDGMENTS

I would like to thank Professor Gutierrez for his tremendous guidance and support throughout the doctoral process, particularly in the last year. Of course, I would like to thank my advisory committees: Dr. Jonathan Arnold, Dr. Liang Liu, and Dr. Jessica Kissinger for their mentorship and guidance in the past five years. Also, I would like to thank Dr. Ying Xu with whom I have collaborated and published several projects. In particular, I would like to thank Dr. Chi Zhang and Dr. Sha Cao (who are both assistant professors at Indiana University now). I learned various statistical techniques and machine learning methods from them. I would thank Dr. Elizabeth Trippe, Dr. Zerotti Woods, Jessica Brady, Karen Gonzalez and Dr. Diego Moncada for their invaluable supports for me in the past years. Also I would like to thank the MaPHIC-HAMMER Consortium for providing the data to analyze.

In addition, I want to express my deepest gratitude to the Institute of Bioinformatics at University of Georgia, which is a big and lovely family for providing me various opportunities. Particularly, Mrs. Sandra, Mrs. Julie and Mrs. April, they provided me with a very very strong support in life outside research. Finally, UGA is a lovely and memorable place. I feel so lucky to have the chance to be here and finish my Ph.D. here. No thank you is enough. I'm grateful to thank my family and girlfriend Genggeng for their love and spiritual supports in all the past years.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	x
CHAPTER	
1 BACKGROUND AND INTRODUCTION	1
1.1 MALARIA	1
1.2 TELEMETRY	4
1.3 ACCELEROMETER AND MALARIA	4
1.4 ELECTROCARDIOGRAM (ECG) AND MALARIA	5
1.5 MACHINE LEARNING AND ELECTROCARDIOGRAM (ECG)	7
1.6 CIRCADIAN RHYTHMS AND MALARIA	9
1.7 MOTIVATION OF THE STUDY	9
2 METHODS AND MATERIALS	11
2.1 EXPERIMENTAL DATA DESCRIPTION	11
2.2 TELEMETRY DATA	13
2.3 TELEMETRY QUERYING	18
2.4 DATA PREPROCESSING	21
2.5 IMBALANCE DATA PROCESSING	26
2.6 DEEP NEURAL NETWORK MODEL	27
2.7 ASSOCIATIONS BETWEEN ACCELEROMETER AND ECG	31

2.8	RNA-SEQ ANALYSIS	32
3	RESULTS	36
3.1	ACCELEROMETER CHANGES IN DIFFERENT HOUR GROUPS	36
3.2	DNN CLASSIFICATION RESULTS	37
3.3	ASSOCIATIONS BETWEEN ACCELEROMETER AND THE DNN MODEL	45
3.4	RNA-SEQ ANALYSIS RESULTS	48
4	CONCLUSION AND DISCUSSION	53
4.1	CONCLUSIONS	53
4.2	NOVELTIES	55
4.3	CAVEATS	56
4.4	APPLICATIONS	57
	BIBLIOGRAPHY	59
	APPENDIX	75
A	CONFUSION MATRICES	75
B	PYTHON MODULE FOR DATA ACCESS	87
C	DEEP NEURAL NETWORK ARCHITECTURES	91

LIST OF FIGURES

1.1	Malaria Prevalence	2
1.2	Malaria Parasite Life Cycle	3
1.3	ECG Morphology	6
2.1	E07B Experimental Time Line Summary	14
2.2	E06 Experimental Time Line Summary	15
2.3	E30 Experimental Time Line Summary	16
2.4	Examples of Telemetry Data	18
2.5	Framework of Python Class	19
2.6	Number of ECG Observations in Two Classes	23
2.7	Summary of DNN Architecture	28
2.8	Number of Epochs on H0 Training Dataset	29
3.1	Standard Deviation of Accelerometer During a Day	37
3.2	Comparison of Standard Deviations of Accelerometer Data Between Groups	38
3.3	Comparisons of Activity Levels Between Pre-infection And The Liver Stage	39
3.4	Classification Results of 13 Subjects	40
3.5	Classification Results of 24-Hour Groups	42
3.6	The Normalized Confusion Matrix of H21	43
3.7	ROC Comparisons of Different Hour Groups	44
3.8	Scatter Plots of Pre-infecetion Class	46
3.9	Scatter Plots of The Liver Stage	47
3.10	PCA of Normalized RNA-Seq Data	49

3.11 Differential Expressions of Circadian Genes	50
A.1 The Normalized Confusion Matrix of Hour 0	75
A.2 The Normalized Confusion Matrix of Hour 1	76
A.3 The Normalized Confusion Matrix of Hour 2	76
A.4 The Normalized Confusion Matrix of Hour 3	77
A.5 The Normalized Confusion Matrix of Hour 4	77
A.6 The Normalized Confusion Matrix of Hour 5	78
A.7 The Normalized Confusion Matrix of Hour 6	78
A.8 The Normalized Confusion Matrix of Hour 7	79
A.9 The Normalized Confusion Matrix of Hour 8	79
A.10 The Normalized Confusion Matrix of Hour 9	80
A.11 The Normalized Confusion Matrix of Hour 10	80
A.12 The Normalized Confusion Matrix of Hour 11	81
A.13 The Normalized Confusion Matrix of Hour 12	81
A.14 The Normalized Confusion Matrix of Hour 13	82
A.15 The Normalized Confusion Matrix of Hour 14	82
A.16 The Normalized Confusion Matrix of Hour 15	83
A.17 The Normalized Confusion Matrix of Hour 16	83
A.18 The Normalized Confusion Matrix of Hour 17	84
A.19 The Normalized Confusion Matrix of Hour 18	84
A.20 The Normalized Confusion Matrix of Hour 19	85
A.21 The Normalized Confusion Matrix of Hour 20	85
A.22 The Normalized Confusion Matrix of Hour 22	86
A.23 The Normalized Confusion Matrix of Hour 23	86

LIST OF TABLES

2.1	Summarization of Experiments	12
2.2	List of Experimental Subjects	12
2.3	Summary of Sampling Frequency	17
2.4	List of Experimental Subjects	17
2.5	List of Accelerometer Data Querying Table	24
2.6	Number of Accelerometer Observations After Segmentation	25
2.7	Number of RNA-Seq Replicates in Two Classes	33
2.8	Brief Summary of Circadian Gene Functions	34
3.1	Training And Testing Results of 24-Hour Groups	41
3.2	R-squared of Two Classes	45
3.3	Differential Expressions of Circadian Genes	51
3.4	Summary of DE Genes And Enriched Pathways	51
3.5	Enriched Circadian Pathways of Under-Expressed DE Genes	52

CHAPTER 1

BACKGROUND AND INTRODUCTION

1.1 MALARIA

Malaria is an infectious disease caused by the parasites of the *Plasmodium* genus. The World Health Organization (WHO) has estimated that 219 million malaria infection cases would occur and 435,000 deaths would be associated in 2017 [1]. Also, it is estimated that up to US\$ 3.1 billion in 2017 was funded for malaria control and elimination processes. The typical symptoms of malaria infection include high fever, muscle aches, severe anemia and fatigue, which is quite similar to other infectious diseases such as flu. The disease is most prominent in poor, tropical or subtropical areas with high abundances of anopheline mosquitoes, such as Africa, South-East Asia, Western Pacific, etc (Figure 1.1) [2, 3]. Mainly, the African region carries an extraordinarily high burden of malaria, with 92% of all cases reported and 93% of deaths in 2017 [1].

The parasite is carried by the infected mosquitoes. When the infected mosquitoes feed on human hosts, the mosquitoes inject the sporozoites, which are the motile stage of the *Plasmodium* parasite, along with their saliva [4]. The sporozoites then travel to the host liver and infect hepatocytes. A single sporozoite inside a hepatocyte can generate tens of thousands of new merozoites, which are later released into host's bloodstream. While the parasite is in the host's liver, the host usually does not have any clinical symptoms. This phase is known as the liver stage of the infection. When the parasites are released into the host's bloodstream, the clinical symptoms occur, and this phase is called the blood stage of the infection. Usually, the liver stage will last 7-10 days after the initial mosquito bite [5]. The life cycle is shown in Figure 1.2 below.

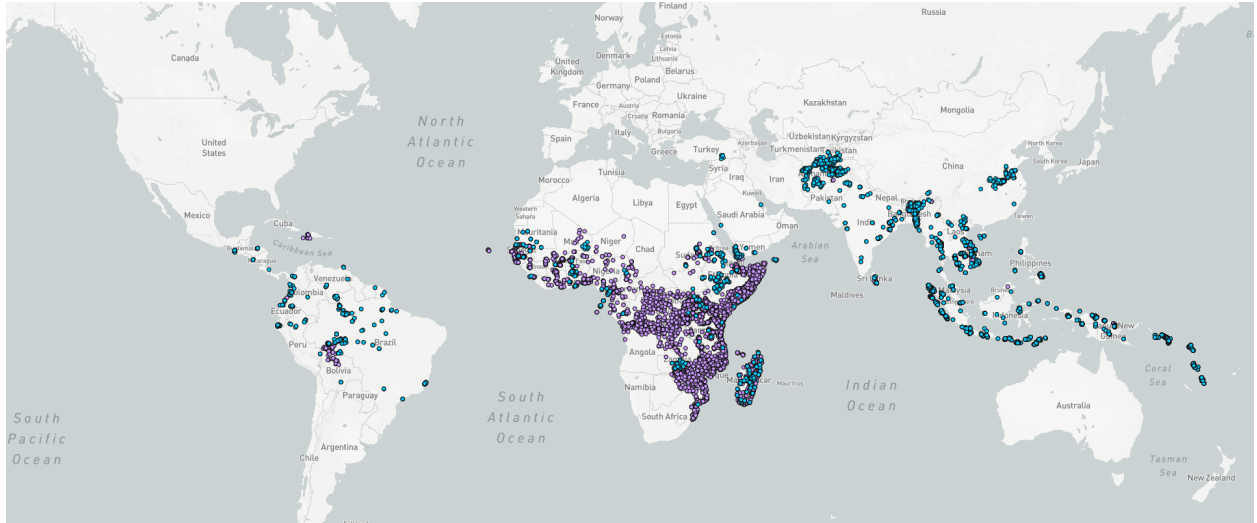


Figure 1.1: Malaria Prevalence

Each spot represents a village level clustering survey of malaria parasite prevalence with the blue dots denote *P. vivax* and the purple dots denote *P. falciparum*. The map was generated by the Malaria Atlas Project in 2018 [3]

To detect *Plasmodium* infection in the liver stage is quite tricky since the host does not show any symptoms of infection. When the parasites enter into the host's bloodstream, the symptoms can vary from clinically mild to severe. The symptoms of severe malaria infection usually involve the nervous system, respiration system, anemia, high fever, and in some cases, organ failure [6], which is very dangerous and can be fatal. Early detection of *Plasmodium* infection in its liver stage would be super beneficial to prevent complications. In Chapter 2, we are aiming to develop a machine learning method that can be used to detect the malaria infection in its liver stage before any of the symptoms occur by utilizing the telemetry data. Also, we are recognizing that the detection method should be convenient and easy to test since the host will not go to hospital for blood screen tests if there are no symptoms. Therefore, we are considering only to use the telemetry data such as the daily activity, the heartbeat rate, the periodic temperature, blood pressure and host's electrocardiogram

(ECG), which could be collected by wearable devices [7, 8], to build a machine learning model.

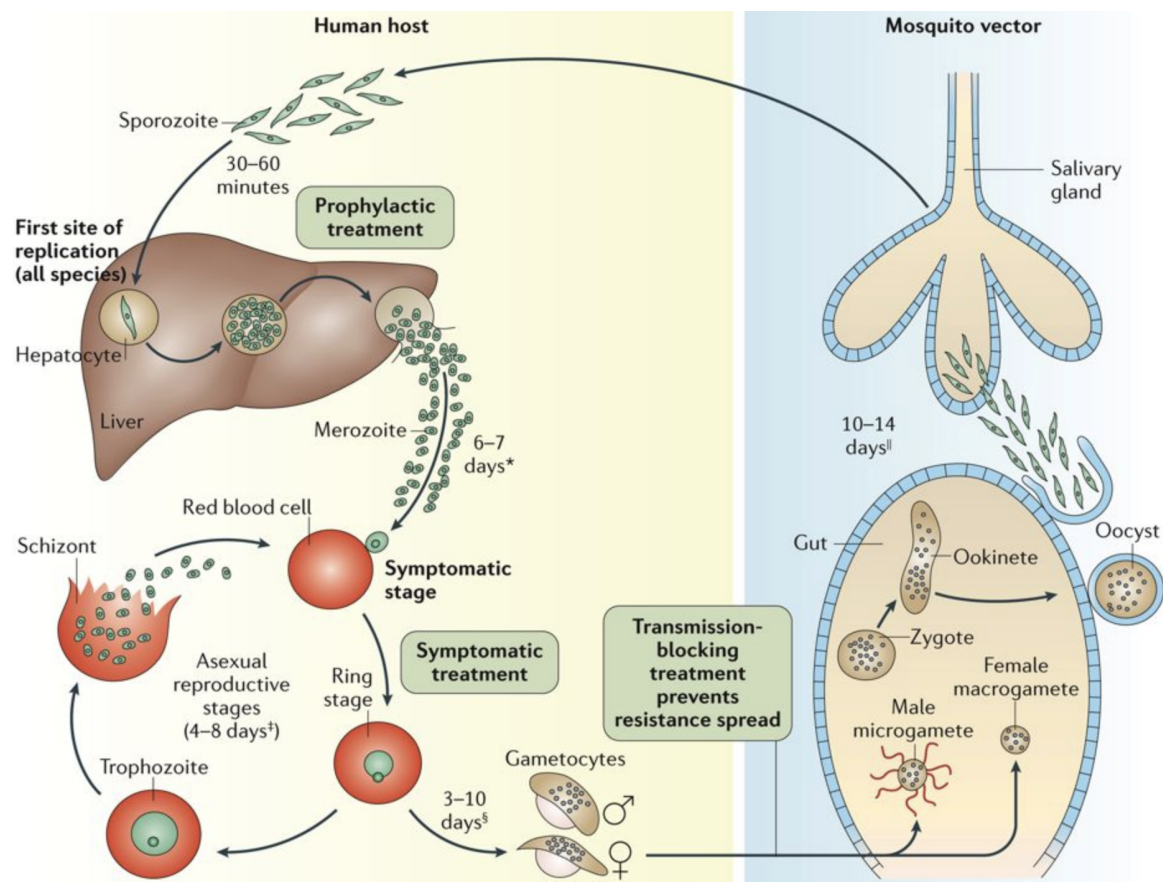


Figure 1.2: Malaria Parasite Life Cycle

The figure summarizes the standard life cycle of a malaria parasite infection. The infected mosquito takes a blood meal from a host, which injects the sporozoite stage of the parasite. The sporozoites then invade the host's liver cells, called hepatocytes. In the infected hepatocytes, the sporozoites replicate hundreds of thousands of merozoites, which are released into bloodstream [4].

1.2 TELEMETRY

With the popularization of wearable devices and smartphones like iPhone, Apple Watch and Fitbit, we can collect our daily activity, heart rates and Electrocardiogram (ECG) very easily [9]. The daily activity of one's own is not only an excellent way to monitor the rest/activity cycles, but also a good reflection of one's physical health. Hence, the collection and accumulation of those massive amounts of telemetry data give us an excellent opportunity to have a better monitor of our health conditions. For instance, a research group from Qatar [10] used wearable devices to monitor and predicting the sleep quality of 92 adolescents over one full week time period. In particular, they used physical activities data during awake time to predict the quality of sleep in the night time, and the predictive model is evaluated by the physical activity data during the sleep time. This example shows us that with a machine learning model and appropriate data, we can have a better monitor of our sleep during which we are usually unconscious.

1.3 ACCELEROMETER AND MALARIA

Activity data are usually collected by a small sensor that is attached to the host that is being monitored. The sensor can collect three axes of acceleration of the host in a given time period with specific recording frequency.

Accelerometer data reflects one's daily activity level, and it could be a great signal for predicting one's health condition through time series trajectories, such as the disruption in the activity levels that usually come with associated diseases. Sickness usually means lowered activity. One recent study used machine learning approaches to predict the risks of falls in a group of older women by using accelerometer data. The participants were grouped into two groups as high-fall risk and low-fall risk based on the accelerometer data and reported falls in the past year. Eleven models are combined in the study with a 10-fold cross-validation, and the ensemble model achieves a prediction accuracy of 73.7%, precision of 81.8% and AUC of

0.706 [11]. The study provided an excellent example of how the use of telemetry data from wearable devices could be informative when combined with machine learning methods. Also, several studies have used sensors to monitor daily activities of Parkinson's disease patients with the aim of 1. monitoring the progression of the disease [12]; 2. identifying the gait traits that are related to Parkinson's patients [13]; 3. Characterize the levels of tremor of Parkinson's patients [14].

Furthermore, in the field of bioinformatics, a genome-wide association study (GWAS) analyzed the genome of 91,105 UK Bio-bank participants and proved that 14 loci have significant relationships to physical activity and sleep duration in these participants. The study emphasizes the importance of physical activity and its contribution to the whole body health condition, especially in controlling blood pressure level [15].

1.4 ELECTROCARDIOGRAM (ECG) AND MALARIA

Electrocardiogram, ECG or EKG, is the process of recording the electrical activity of the heart over a period of time using electrodes placed on the skin. Commonly, those electrodes are placed on the patient's limbs and chest to detect the subtle electrical changes that are caused by the heart muscle's electrophysiological pattern of depolarizing and re-polarizing during each beat. During each heartbeat, a healthy heart has an orderly pattern of depolarization that makes the characteristic pattern of an ECG trace [16]. There are several main features of an ECG signal as shown in Figure 1.3 that are clinically important [17]:

- P-wave: A small voltage rise that represents the depolarization of the atria.
- PQ-interval: The time between the start of atrial depolarization and the start of ventricular.
- QRS-complex: The largest amplitude part of an ECG that represents the depolarization of the ventricles.
- QT-interval: The time between ventricular depolarization and repolarization.

Moreover, abnormal physiological heart rate variability (HRV) change is well known to associate with certain cardiac diseases or diseases that relate to heart dysfunction. The comprehensive measurements of ECG telemetry protocol have been studied in mice models [18].

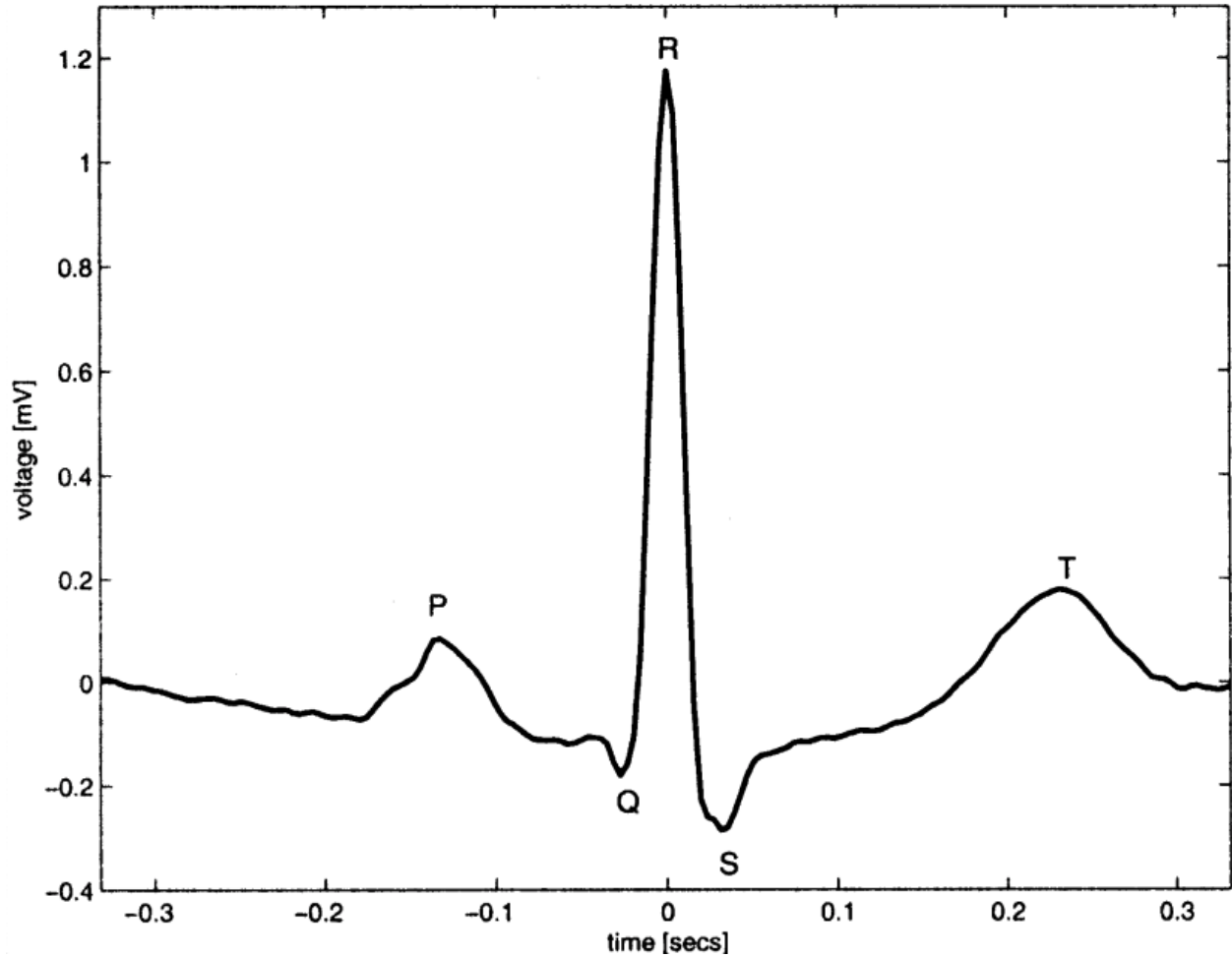


Figure 1.3: ECG Morphology

The electrocardiograph of a heart in normal sinus rhythm [17]

Malaria infection also has been implicated with heart failure [19]. For example, a 40-year-old man was reported to have been diagnosed with an acute coronary syndrome that was caused or associated with malaria infection or treatment [20]. Severe malaria infection mimics the symptoms of acute myocardial infarction that a traveler was reported to be diagnosed with acute myocardial infarction but later turned out to be malaria infection [21]. Although

vivax malaria is thought to be benign, one study has shown that this infection could also lead to multiple organ dysfunctions that include the cardiovascular system [22].

Besides malaria infection itself, the treatment of malaria could also lead to heart dysfunctions. Halofantrine is a medication used to treat malaria infection. However, researchers reported that electrocardiographic changes or specifically a prolonged QT interval were observed with treated patients [23]. Furthermore, quinine are used to be the most common type of medication for anti-malaria, but a study shows that before the drug treatment severe malaria-infected children show QT-shortage, while after therapy a prolonged QT was observed among the infected children [24]. The pre-treatment change of QT interval was proved to be independent of quinine treatment.

The electrocardiogram (ECG) signal contains a large amount of information, but it is challenging to interpret and requires lots of experience to detect the subtle ECG changes in patients manually, and the manual diagnosis could sometimes lead to misdiagnosis. A machine learning approach could help us to identify subtle ECG changes with consistency accurately. Hence, in this study, we will report a machine learning method that could detect early malaria infection when it is in the liver stage before the symptoms occur, with a relatively high and stable accuracy.

1.5 MACHINE LEARNING AND ELECTROCARDIOGRAM (ECG)

Machine learning and artificial intelligence offer a broad field of applications that include classification, clustering and regression problems. It uses computer algorithms and statistic models that “learn” certain hidden patterns from the given data, after which it will be applicable to perform a specific task or inference [25] effectively. Deep learning is one of the sub-fields under machine learning, and it has become one of the most widely used methods in recent ten years. Yann Lecun, Yoshua Bengio, and Geoffrey Hinton are the first three that introduced a backpropagation algorithm to the neural network, which let a neural network automatically change its internal parameters [26]. This backpropagation algorithm links more

neural network layers without gradient vanishing problem, thus taking the neural networks to a deep level. This method shows a much better performance in various prediction and classification settings [27, 28, 29].

The current state of the art in language processing [30], image recognition [29] and strategy games such as *alpha-go* [31] and *alpha-zero* [32] are dramatically improved by this subclass known as deep learning. Deep learning has also been successfully used in solving biomedical problems such as skin cancer diagnosis [33] and drug discovery [34]. The deep learning model is well suited for the “big data” era as the larger training set, the better [26].

In the field of electrocardiogram (ECG), machine learning methods have been deployed mainly from traditional methods to deep learning methods. For example, 1. A machine learning approach that combines principal component analysis (PCA) and least square support vector machine (LS-SVM) has been used to detect cardiac arrhythmias. Through analyzing three independent datasets of ECG signals, the method achieves accuracy higher than 96% [35]; 2. A robust, real-time machine learning model was built by AliveCor to detect silent atrial fibrillation by smartphone recordings of ECG signals [36]. These smartphone-based applications meet the current burst of personal health data from wearable devices quite well. As the data from the wearable device end is still exploding and accumulating, more rapid and robust machine learning based techniques should be developed to monitor individual health status better [36].

In the above traditional ECG data analysis methods, all the machine learning models are trying to build a prediction model based on the extracted ECG features, such as the length of QT interval, the QRS width and the T wave shape, etc [37]. Recently, a deep neural network (DNN) model from Andrew Y. Ng’s group was used to classify 12 cardiovascular rhythm classes by training on 91,232 ambulatory raw ECG recordings from more than 50,000 patients [38]. The model trains and predicts the raw ECG signals without any feature extractions, and it outperforms the average cardiologists’ classification with an impressive area under receiver operating characteristic (ROC) curve of 0.97, which shows that in the clinical setting, we

could use computers and machine learning approaches to have better diagnostic accuracy than an individual cardiologist. Meanwhile, another convolutional neural network was built to predict patients with asymptomatic left ventricular dysfunction (ALVD) by using 12-lead ECG data alone. The model is trained on 44,959 patients' ECG data from Mayo Clinic and tested on an independent ECG data set of 52,870 patients. The model achieves the accuracy of 85.7% and AUC of 0.93, which provides a powerful screening tool to ALVD disease [39]. All these deep learning methods inspired the possibility that a deep learning model could be built to predict the infectious malaria stage by directly using raw ECG data.

1.6 CIRCADIAN RHYTHMS AND MALARIA

Circadian rhythm or clock is an important biological mechanism for organisms to synchronize its internal biological processes with the outer environment. Studies have shown that the circadian rhythms also enable parasites like malaria *Plasmodium* to maximize its ability to live in the host [40, 41, 42]. The study demonstrates that the circadian rhythms of parasites do not depend on the light shift, but depend on the host's metabolic shifts [41], especially shift in blood glucose rhythms. Also, when the circadian rhythms between host and parasites were perturbed, it costs the parasites two-fold cost to back to stable state of replication and transmission [42]. Hence, it may be a beneficial way to prevent the host from parasites by slightly changing the circadian rhythms of glucose metabolites.

1.7 MOTIVATION OF THE STUDY

There are numerous studies that focus on using ECG data to predict a broad spectrum of heart diseases and abnormalities by machine learning methods [43, 44, 45, 46]. However, to the best of our knowledge, there is no existing method has been reported to predicting malaria infection by using ECG data alone. The combination of a large amount of computerized ECG data from experimental recordings and the development of machine learning methods could

provide us an extraordinary way to detect malaria infection early if the subject communities had wearable devices.

Deep neural networks can automatically detect the patterns and learn features from the raw input signals without too much requirement from data preprocessing and feature selections. And the DNN model tends to have a better performance with a larger amount of training data [26], which makes this approach quite suitable for our ECG data. Therefore, we are motivated to build a DNN model that takes the training data of the recorded ECG to aid in early detection of malaria infection.

With the predictive DNN model through training on ECG data alone, we are integrating accelerometer data to the machine learning model in the aim of improving the model performance and discovering the interactions of activity and ECG. The associations of accelerometer data and model performance are further analyzed in our study.

The telemetry information is the extrinsic performance of the host and it is driven by intrinsic complex biological mechanisms. Especially, the changes in ECG signals and accelerometer in malaria hosts are mainly associated with the intrinsic biological changes of the host. RNA-Seq data are also increased massively with the availability of the next-generation sequencing, and it is a perfect reflection of biological metabolisms of the model organisms. [47, 48]. Hence, the integration of RNA-Seq and telemetry data into the malaria infection prediction study could provide us not only the hint from extrinsic performance but also intrinsic mechanisms.

CHAPTER 2

METHODS AND MATERIALS

2.1 EXPERIMENTAL DATA DESCRIPTION

In this study, we are using telemetry and transcriptomic data provided by the *Malaria Host-Pathogen Interaction Center* and *Host Acute Models of Malaria* (MaHPIC-Hammer) project. MaHPIC was established with funding from the NIAID (September 2012 to September 2017) to characterize host-pathogen interactions during malaria infections of non-human primates and clinical studies via collaborations from various fields. The MaHPIC project used a “systems biology” strategy to understand the relationships between *Plasmodium* parasites and their non-human primate hosts. Specifically, the data used here are collected from infections in which *Plasmodium knowlesi* is the parasite strain that is injected into hosts to induce malaria infection, while two species of experimental host monkeys are included in the MaHPIC-Hammer project, one is *Macaca Mulatta* which is a non-resilient host to *P. knowlesi* and the other is *Macaca fascicularis* which is a *P. knowlesi* resilient host [49]. *P. knowlesi* is chosen because it is the parasite strain that is reported to be active both in human beings and Macaques [50, 51]. In total, MaHPIC result datasets were composed of thousands of files and several data types stored in PlasmoDB and other archived databases [52].

In the study, we have three experiments that are involved, namely Experiment06 (E06), Experiment07B (E07B, E07A failed and thus is not included) and Experiment30 (E30), which are summarized in the Table 2.1 below. For these three experiments, E30 and E06 are the malaria parasite non-resilient host group, while E07B represents malaria parasite resilient host group. There are two monkey subjects in E30, four monkey subjects in E06

and seven monkey subjects in E07B. All three experiments are infected by the *P. knowlesi* Malayan strain of the parasite.

Table 2.1: **Summarization of Experiments**

Summarization of the experimental IDs and the parasites, hosts and the number of replicates that are associated

Experiments ID	E30	E06	E07B
Parasite Strain	<i>P.knowlesi</i> (Malayan Strain)		
Host	<i>M. mulatta</i>		<i>M. fascicularis</i>
Number of Subjects	2	4	7

The experimental timeline for E07B (*M. mulatta*), the *P. knowlesi* resilient host group, is shown in the Figure 2.1 below [53]. Figure 2.1.A shows the overall experimental timeline for all the seven monkey subjects under E07B. The sensor collecting telemetry data was implanted into the experimental host 30 days before the experiments started, but it was activated 10 days before the injection of *P. knowlesi* parasite. The black line shows the predicted profile of infection. And each bar denotes the time to collect the host’s blood sample for downstream RNA-Seq analysis. Figure 2.1.B shows the actual parasite profile of infection in each. The bar in Figure 2.1.B corresponds to the timeline bar in Figure 2.1.A.

Table 2.2: **List of Experimental Subjects**

The subject name and id of each experiment is listed

Experiment	Subject Name	Subject ID
E06	RUf16	1918
E06	RIh16	1912
E06	RTe16	1915
E06	RCI15	1916
E07B	11C166	1409
E07B	12C53	1402
E07B	12C44	1407
E07B	11C131	1408
E07B	H12C8	1405
E07B	12C136	1401
E07B	H12C59	1404
E30	RKy15	1403
E30	REd16	1406

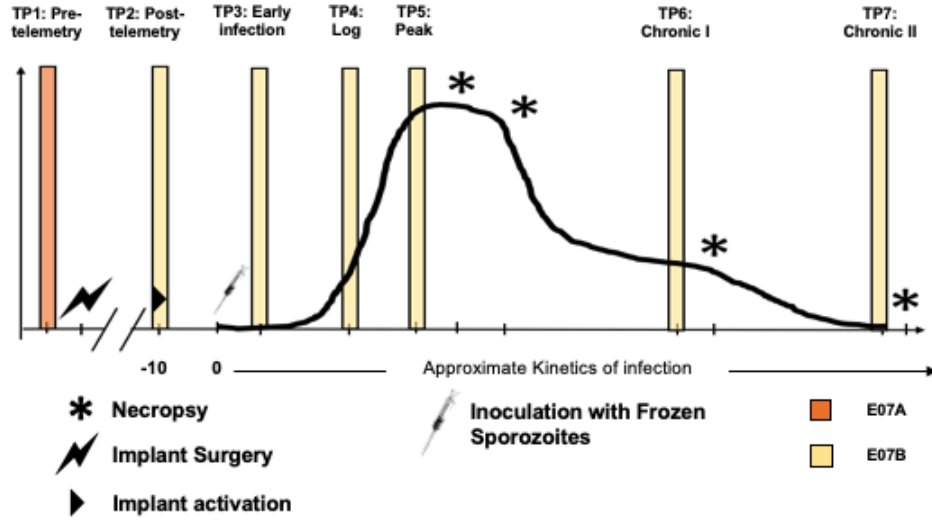
The similar experimental timeline summarization for Experiment 06 and Experiment 30 are shown in Figures 2.2 and 2.3 [53]. It is worth pointing out that both E06 and E30 are the *P. knowlesi* non-resilient host group, hence the euthanasia and necropsy occur early in all the monkey subjects in E06 and E30. Usually, euthanasia occurs at the peak of parasitemia. And another major difference is that the telemetry collection sensor was activated 13 days before the parasites injection, which offsets the missing days of telemetry collection due to early euthanasia and necropsy. E30 and E06 share a similar timeline.

In Table 2.1, we have listed that E30, E06, and E30 have 2, 4, and 7 monkey subjects respectively. The name of each subject and ID is shown in Table 2.2. The subject IDs are used as search keys for each subject in the cloud database, which is described in the following section 2.3.

2.2 TELEMETRY DATA

The continuous telemetry data of the host’s body temperature, three-dimensional activity and electrocardiogram (ECG) are collected. The ECG data are recorded by a standard 12-lead sensor. And the sampling frequency for the telemetry data are summarized in Table 2.3: 1 Hz for temperature, 10 Hz for the accelerometer and 1000 Hz for ECG respectively. The experimental recording timeline for each subject is approximate to one month, and for the continuous telemetry data, it will occupy a huge storage space in the local computers. For example, ECG data are sampled at a 1,000 Hz frequency, which equals 1,000 records per second, yielding 8.64×10^7 records per day. On average, one ECG recording from the experimental subject takes up to 300GB in storage space and the total 13 subjects’ ECG data equal around 4TB. Those large amounts of data are hard to store in local computers. Hence, all the telemetry data are stored in a server-based relational database, and the query method used to access them will be introduced in the next section 2.3.

A. IDEALIZED E07 INFECTION SCHEMATIC



B. ACTUAL E07 INFECTION SCHEMATIC

Time points below begin at TP2, as TP1 was part of a previous experiment E07A.

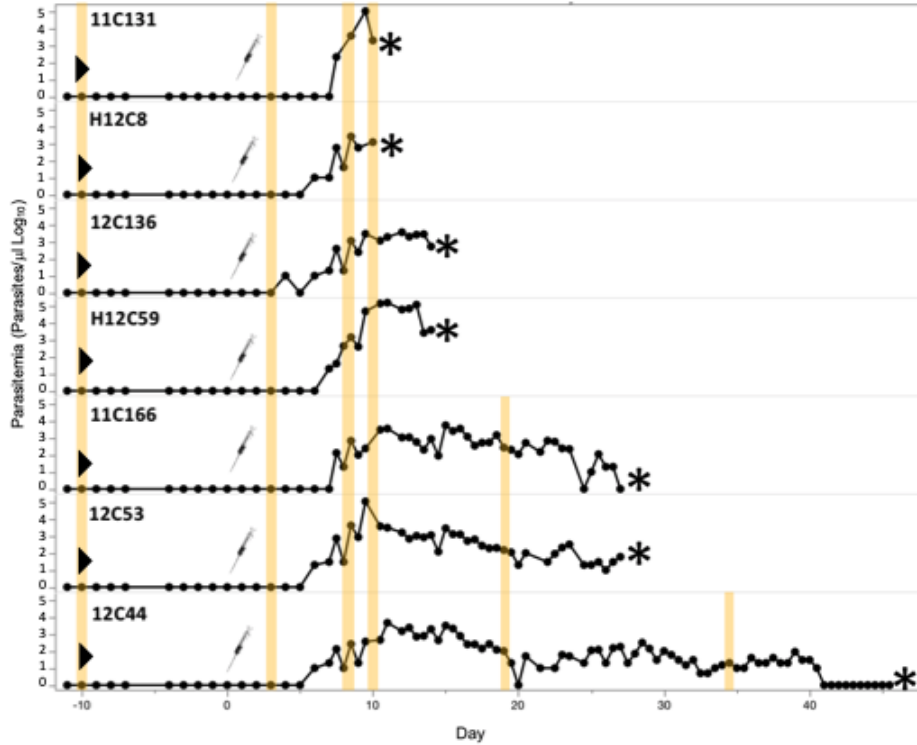
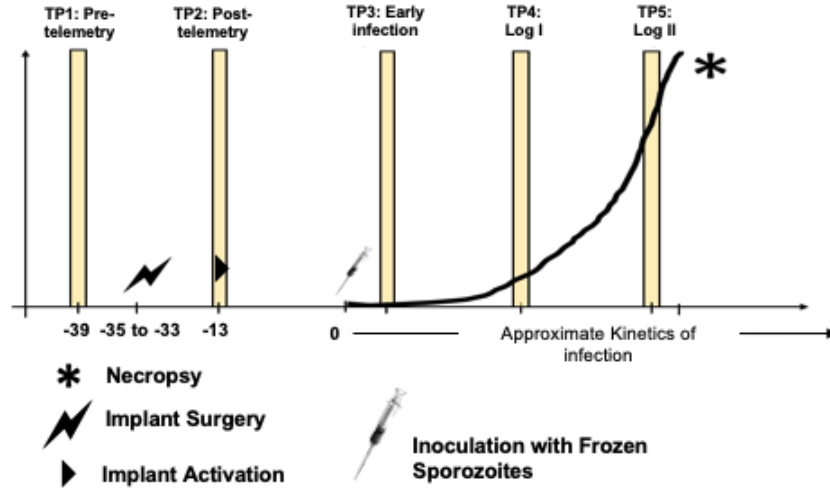


Figure 2.1: E07B Experimental Time Line Summary

The experimental timeline summary of the seven *M. mulatta* subjects that under Experiment 07B [53]. The black line shows the estimated blood parasite count in log₁₀/μl unit. And the vertical yellow bars show the time points that the blood samples from each individual host is collected. The sensor is activated at 10 days before injection of parasite and keeps collecting telemetry data until euthanasia and necropsy.

A. IDEALIZED E06 INFECTION SCHEMATIC



B. ACTUAL E06 INFECTION SCHEMATIC

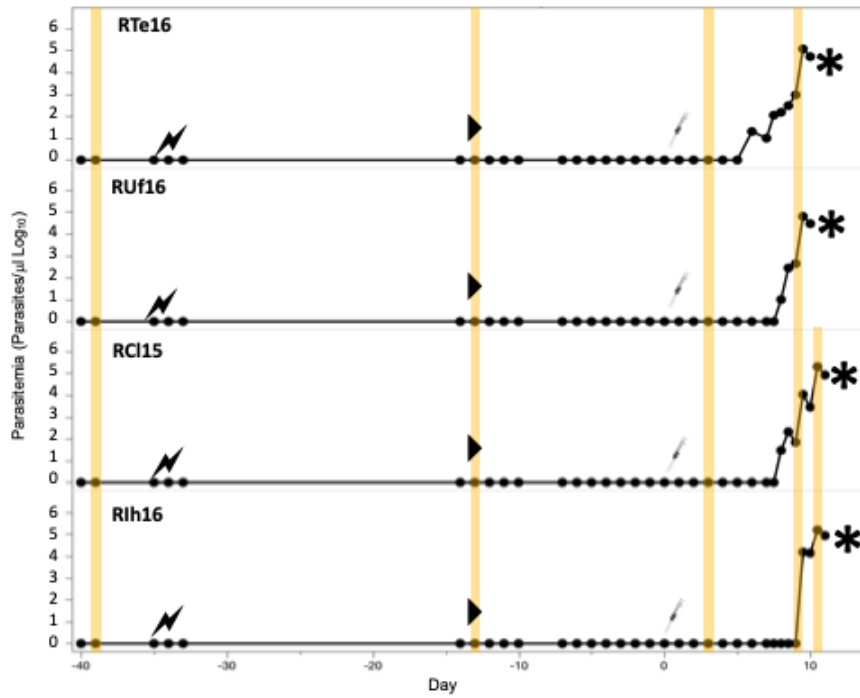
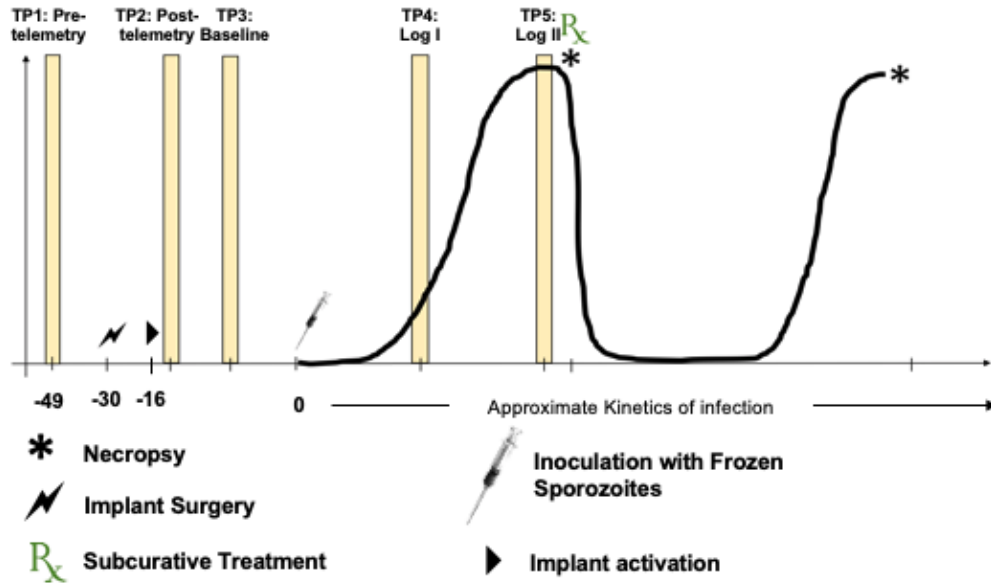


Figure 2.2: E06 Experimental Time Line Summary

The experimental timeline summary of the four *M.fascicularis* subjects that under Experiment 06 [53]. The black line shows the estimated blood parasite count in $\log_{10}/\mu\text{l}$ unit. And the vertical yellow bars show the time points that the blood samples from each individual host is collected. The sensor is activated at 13 days before injection of parasite and keeps collecting telemetry information until euthanasia and necropsy.

A. IDEALIZED E30 INFECTION SCHEMATIC



B. ACTUAL E30 INFECTION SCHEMATIC

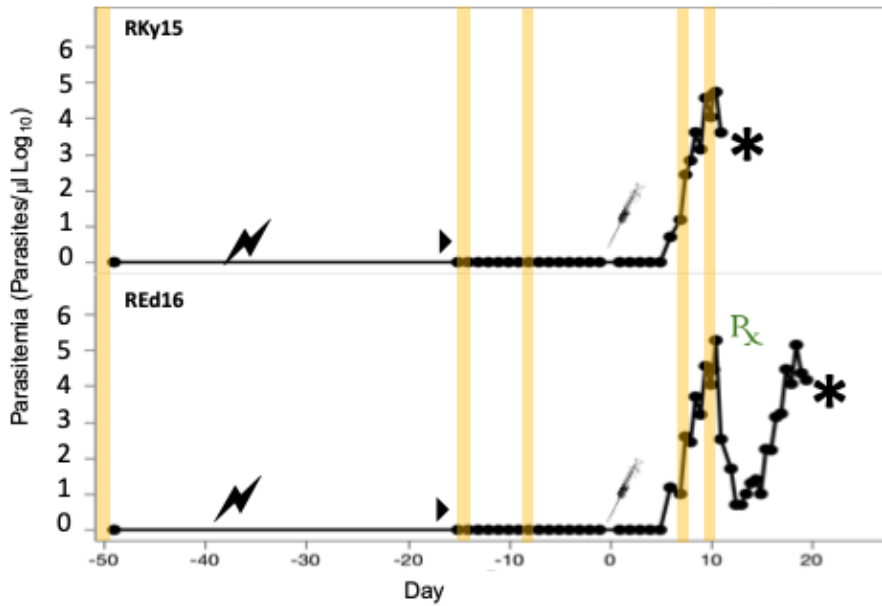


Figure 2.3: E30 Experimental Time Line Summary

The experimental timeline summary of the two *M. fascicularis* subjects that under Experiment 30 [53]. The black line shows the estimated blood parasite count in log₁₀/ μ l unit. And the vertical yellow bars show the time points that the blood samples from each individual host is collected. The sensor is activated at 13 days before injection of parasite and keeps collecting telemetry information until euthanasia and necropsy.

Table 2.3: **Summary of Sampling Frequency**

The sampling frequency for accelerometer, ECG and temperature is summarized.

	Temperature	Accelerometer	Electrocardiogram (ECG)
Sampling Frequency (Hz)	1	10	1000

Since the telemetry data are continuously recorded, we have telemetry information that spans from the pre-infection state to the host euthanasia. It is essential to separate the dataset into stages: pre-infection, liver infection and blood infection [4] as we perform classification. The stages of pre-infection and liver infection are determined by the experimental timeline (shown in Figure 2.1, Figure 2.2 and Figure 2.3), where the pre-infection stages were defined as the time period before the inoculation of parasites, while the liver stages were defined as the time period that after the parasite inoculation but before the occurrence of parasites in the blood stream. Based on that, the starting date and ending date of pre-infection and liver stages of each experiment group are listed in Table 2.4.

Table 2.4: **List of Experimental Subjects**

The subject name and id of each experiment is listed

Group	Start_Date	End_Date	Class
E06_preinfection	21-MAR-17	24-MAR-17	0
E06_Liver	12-APR-17	15-APR-17	1
E30_Preinfection	27-JUL-16	30-JUL-16	0
E06_Liver	19-AUG-16	22-AUG-16	1
E07B_Preinfection	10-JAN-17	13-JAN-17	0
E07B_Liver	27-JAN-17	30-JAN-17	1

Examples of telemetry data for temperature, accelerometer, and electrocardiogram (ECG) are shown in Figure2.4, which have the time ranges of 3 days, 10 seconds and 10 seconds, respectively. The example telemetry data are from subject of REd16 (1406) of experiment30 (E30). These three different data types of telemetry in Figure2.4 have very different features, such as the different number of periodic waves. For temperature in 3 days, there are only 2.5 periodic waves, while ECG has roughly 20 periodic waves in 10 seconds. To the best of our knowledge, there is no straight forward way to integrate those

desynchronized telemetry data with different frequency patterns [54]. In this study, we built the classification model by only using the raw ECG data alone, and the model performance is later proved to be linked to accelerometer data.



Figure 2.4: **Examples of Telemetry Data**

Visualization plot of sample telemetry data. A: The accelerometer data over a 10-second window; B: Temperature data over 3-day time range; C: Electrocardiogram (ECG) data over 10-second window range.

2.3 TELEMETRY QUERYING

An *object-oriented* python module [55] named **clsSKEDExeripment** (Appendix B) was written to query and download data from the server Oracle relational database named SKED.

This module is the python version of the data primitive class described in the SKED framework paper [56], which was originally coded in a MATLAB language. This python module follows the same functionality as the SKED framework paper see in Figure 2.5.

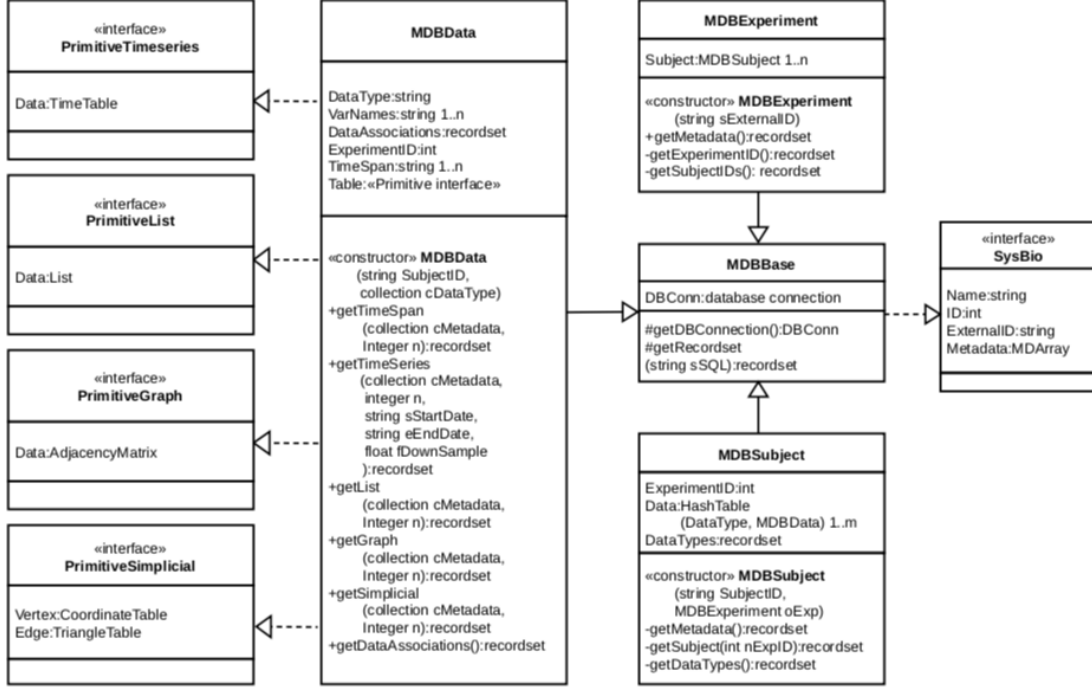


Figure 2.5: **Framework of Python Class**

The standard class diagram for querying and downloading data from the SKED relational database [56]. In the diagram, *c* stands for collection, *m* stands for matrix, *n* stands for integer, *s* stands for string and *o* stands for object. The boxes on the right are the class objects, and the left are the data types that are associated with each class object.

The following functions are the main functions that are associated with querying and downloading different types of telemetry and RNA-Seq data within different time range periods:

- `getMetadata(exp)`: which returns the meta-data information with the provided experiment number.
- `getSubjectIDs(exp)`: which returns all the subject IDs under the provided experiment number.

- `getExperimentID(exp)`: which returns all the experiment IDs under the provided experiment number.
- `getTimeSeriesRange(exp, sub_id)`: which returns the experimental time range of the provided experiment number of the specific subject.
- `getTSData(exp, sub_id, exp_id, time_start, time_end)`: which returns the telemetry time series data with the provided time range, experiment number, subject id, and experiment id.

```

1  from clsSKEDExeripment import *
2
3  exp = MahpicDB()
4  ### Generate the metadata of E06
5  E06_meta = exp.getMetadata('E06')
6  ### Get the subject ids of E06
7  E06_subject = exp.getSubjectIDs('E06')
8  ### Get the experiment ids of E06
9  E06_experiment = exp.getExperimentID('E06')
10 ### Get the time series range certain subject under E06
11 E06_TS_range = exp.getTimeSeriesRange('E06', '1918')
12 ### With the subject id 1918, experiment id 195 and time range
13 ### the continuous time series data will be downloaded
14 data = exp.getTSData('E06', '1918', '195', '22-MAR-17', '13-APR-17')

```

Listing 1: Data Query Example

More specifically, the subject ids refer to the subjects under each experiment as described in Table 2.2. For instance, under Experiment06 (E06) we have four monkey subjects; hence four subject ids will be returned if we call the function of `getSubjectIDs`. And the experiment ids refer to the different data types, such as accelerometer, electrocardiogram, RNA-Seq have its corresponding experiment id of each subject. The standard example query codes to query the continuous time series telemetry data from the SKED relational database will look like in the code in Listing 1. The advantages of this *object-oriented* python module are as follows:

- It is easy to query and download different data types with the given time range.

- Separating the data types and subjects will not overload the computing memory space.
- The query and downloading can be computed in parallel.
- It provides a way to integrate and store multiple data types.

2.4 DATA PREPROCESSING

2.4.1 ELECTROCARDIOGRAM (ECG)

The high frequency recorded Electrocardiogram (ECG) contains large signal information, while in the meantime, it also takes large computing resources to process and analyze. As we have shown in Section 2.2, a single record of ECG data is too large to load into computer memories. Our study aims to identify *Plasmodium* infection in liver stage, where there are no symptoms in the host.

The deep neural networks tend to perform better with a larger amount of training data, and it is ideal to sample the continuous ECG recordings into small pieces of segments and treat each segment as a single observation. The study from Andrew N.G group has reported that 30s segment of ECG data (sampled at 200Hz or 200 samples per second) can achieved prediction F1 score as high as 0.837 to classify 12 heart rhythm classes in a DNN architecture [38]. Based on that study, we proposed the following two approaches to sample the ECG data:

- Resample the 1000Hz ECG data into a 200Hz level and then take a time range of 15s, which will yield 30,000 points for each observation segment. The re-sampling method is to pick the first sample out of five consecutive samples from the continuous recordings.
- Keep the ECG sampling frequency at 1000Hz but to make 10s for each observation segment, which counts to 10,000 samples of each observation.

Of the two approaches, the second one is preferred and was picked for the downstream analysis. The latter approach will return more observations that can be used to train the DNN model without losing any information in the ECG signals. The only shortcoming of

the second approach is that the 10s (10,001 samples with 1000Hz) range of segment may not fully cover the ECG features of the class. However, the downstream analysis shows that the 10s segment is good enough to capture the features of ECG in malaria infection settings.

Physical activity levels have a strong impact to heart rates and cardiac polarization and depolarization process [57]. Therefore, the sampled ECG segments were further grouped into 24 groups to minimize the effects of activity within group. 24-hour groups represent all hours of the ECG recording segment in a day. The ECG signals also have circadian rhythm behaviors that vary a lot through a day [58, 59]. Thus, if the sampled ECG are not grouped into its collecting hour, the inter-group variance will dominate thus affecting the performance of the classification model. The number of observations of each hour in two classes after sampling are shown in Figure 2.6.

From Figure 2.6, we found that the number of observations varies a lot during some hours of the day. For instance, we have 5242 observations of pre-infection class and only 2,006 observations of the liver stage at hour 12. While half of the hours namely as hour 2, 6, 7, 8, 10, 16, 17 and 18, the two classes have a very close number of observations. This variation may be caused by that the sensor in collecting ECG is interrupted by the experiment operators somehow, which leads to unstable missing data in some time periods. However, this imbalanced number of observations will cause bias in the prediction model [60], which will be addressed in section 2.5.

2.4.2 ACCELEROMETER

Accelerometer data were queried and downloaded using the *object-oriented* python module described in section 2.3. The query subject ID and experiment ID are listed in the Table 2.5.

The acceleration of each subject was measured in three dimensions [7]. The root of the total sum of squares normalization method was applied to transform the three-dimensional

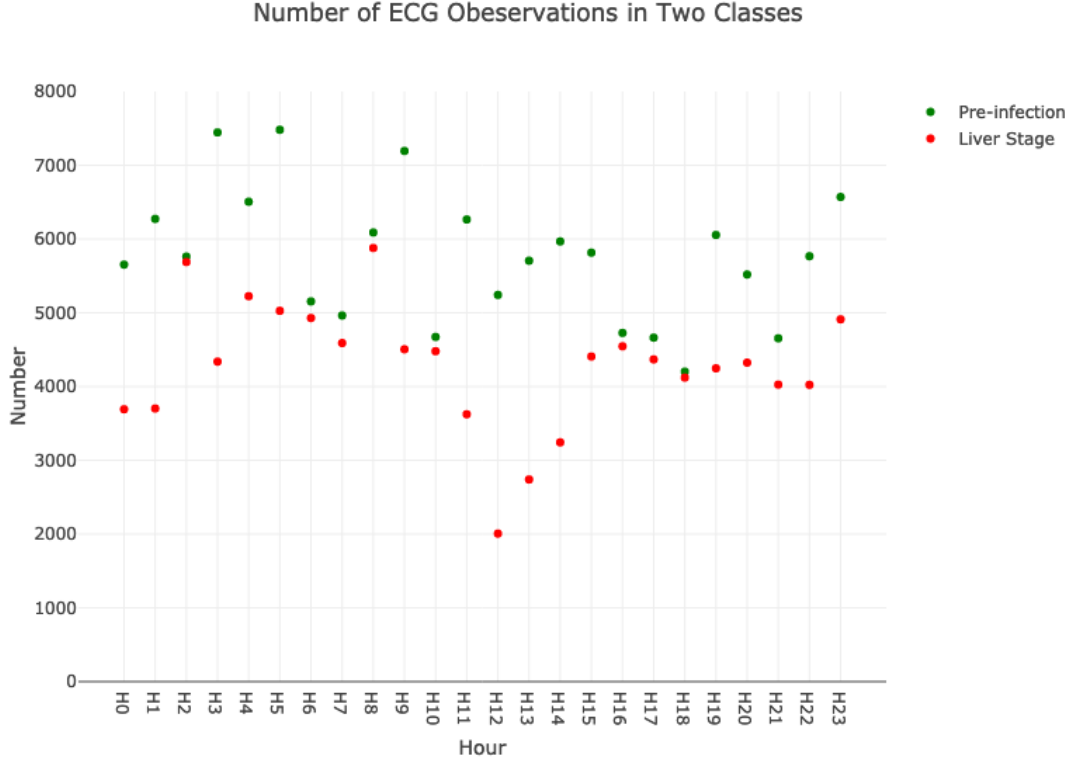


Figure 2.6: **Number of ECG Observations in Two Classes**

Y-axis denotes the number of observations after segmentation into 10s of each observation. X-axis denotes the hour of a day. The blue color shows the pre-infection class and orange shows the liver stage class.

accelerometer data into one dimension [61]. As shown in Equation 2.1 below, where x , y , and z are three dimensional measurements.

$$Accelerometer_{normalized} = \sqrt{x^2 + y^2 + z^2} \quad (2.1)$$

Acceleration itself does not convey too much information since each experimental subject is unique and has its characteristic of activity patterns, thus it may vary dramatically during the day or on certain days. The standard deviation of the accelerometer in a given time period of a day is more stable and contains more information about general activity changes

Table 2.5: **List of Accelerometer Data Querying Table**

Experiment	Species	Experiment ID	Subject ID	Data Type
E06	RIh16	1912	193	TEL_ACC
	RTe16	1915	197	
	RCI15	1916	202	
	RUf16	1918	207	
E07B	12C136	1401	216	
	12C53	1402	221	
	H12C59	1404	225	
	H12C8	1405	229	
	12C44	1407	233	
	11C131	1408	237	
	11C166	1409	241	
E30	RKy15	1403	147	
	REd16	1406	153	

in the time range [62]. Therefore, the standard deviation of normalized accelerometer data were calculated by the equation 2.2.

$$std_{ih} = \sqrt{\frac{\sum_{j=1}^{N_{ih}} (x_{ihj} - \bar{x}_{ih})^2}{N_{ih} - 1}} \quad (2.2)$$

where i denotes the subject i , h denotes the hour of the day, x_{ihj} is the normalized accelerometer of subject i at time h of observation j and N_{ih} is the total number of observations of subject i at time h .

The time ranges of pre-infection and liver stages of accelerometer data are listed in Table 2.4 and the data are queried and downloaded accordingly. Furthermore, we split the accelerometer data into 10-second segments as individual observations, which matches the ECG data processing standardization (described in section 2.4.1). After splitting the accelerometer data, we grouped the segments or observations into different hours of the day, which is recorded from 0 to 23 that represents 0:00 am to 23:00 pm. The hour 0 group covers all of the accelerometer data that are recorded during the time range of 0:00 am to 1:00 am, and the same standards apply to all hours. This grouping method provides us

additional information about experimental monkey subjects' activity changes through a day. The number of observations of each subject in the two infection stages after segmentation into 10s is listed in Table 2.6. As can be seen from Table 2.6, we have a different number of accelerometer observations between pre-infection and liver stages in each subject of each experiment. This imbalanced sample size will cause the problem of bias in the machine learning context [60], which will be addressed in section 2.5.

Table 2.6: Number of Accelerometer Observations After Segmentation

The table listed the number of accelerometer observations of each subject after segmented into 10s time range. 0 stands for the pre-infection stage and 1 stands for the liver stage.

Experiment	Subject	Stage	Number of Observations
E06	Rlh16	0	4151
		1	6011
	RTe16	0	4391
		1	1180
	RC115	0	3131
		1	3255
	RUf16	0	1300
		1	3458
E30	RKy15	0	14653
		1	16578
	REd16	0	12002
		1	15031
E07B	11C131	0	8390
		1	11546
	12C53	0	8913
		1	11999
	12C44	0	8907
		1	13462
	11C166	0	9344
		1	11910
	H12C8	0	9136
		1	13512
	12C136	0	8770
		1	11957
	H12C59	0	9468
		1	12111

2.5 IMBALANCE DATA PROCESSING

The continuous telemetry data are split into 10s of each and grouped into different hours of a day (described in section 2.4), and four days time range for pre-infection and liver stages, separately. The number of observations differ within the two classes. This imbalanced sample numbers will cause a skewed distribution of the dataset, which will cause bias in the classification setting [60]. The class has a larger number of observations is called majority class and the class with fewer observations is named minority class. There are mainly three existing methods that have been widely used in dealing with the imbalanced sample size problems [63]:

- A sampling-based method that involves the random sampling of the dataset by down-sampling the majority class and up-sampling the minority class or the combination of the two.
- An algorithm-based method that involves the modification of the existing algorithms by adding additional constraints, instead of changing the distribution.
- A feature selection based method that involves selecting different features of the class or putting weights on different features to balance the sample size.

In the study, the feature selection method is not applicable, since we are only using raw ECG data to train on the deep neural networks. The algorithm based method is applicable but hard to control the balance between the two class, as it is hard to measure how much additional constraints to put. Hence, we adopted the re-sampling method. However, DNN model requires a large amount of training data, randomly under-sampling the majority class is not suitable in this case. Based on the idea of the larger training dataset, the better performance [64], we choose to set 5,000 ECG observations as the standard number of for each class of each time hour. Specifically, we randomly over-sample the class of hours which has less than 5,000 ECG observations to 5,000 with replications, while under-sample the

of hours with more than 5,000 ECG observations to 5,000. This strategy was proved to be appropriate in later analysis.

2.6 DEEP NEURAL NETWORK MODEL

The segmented ECG recordings are stored in a tensor structure, which is powered by TensorFlow [65]. Each tensor object is a multi-dimensional array that has a data type and shape. Importantly, the tensor objects can reside in GPU memories, which can be later accelerated by the GPU computing resources. We use Keras [66] as the backend of the deep neural network model. Keras is a deep learning model API library that contains almost all the model operations from TensorFlow and provides useful abstractions to reduce the workload in coding. Particularly, all DNN architectures and hyperparameter functions such as neural layers, activation functions, cost functions, optimizers, and regularization terms are all standalone, thus can be changed or combined easily. Therefore, we chose TensorFlow with the backend of Keras to build and implement the deep neural network model.

With the even number of observations of 24 hours in two classes, we randomly picked out 10% of the observations to be used as an independent testing set, which is 1,000 10s ECG observations that involve 500 10s ECG observations from pre-infection and 500 10s ECG from the liver stage. And the rest 9,000 10s ECG observations are fed into the deep neural networks as the training and validation datasets.

2.6.1 ARCHITECTURE DEVELOPMENT

A deep convolutional neural network (CNN) model was built to classify the pre-infection and liver stages of malaria by only training on the raw ECG data without other information such as host species or ECG related features. Specifically, the convolutional neural network has 10 layers of neural networks that include one Conv1D with rectified linear unit activation layer, five fully connected dense layers, and four dropout [67] layers (summarized in Figure 2.7).

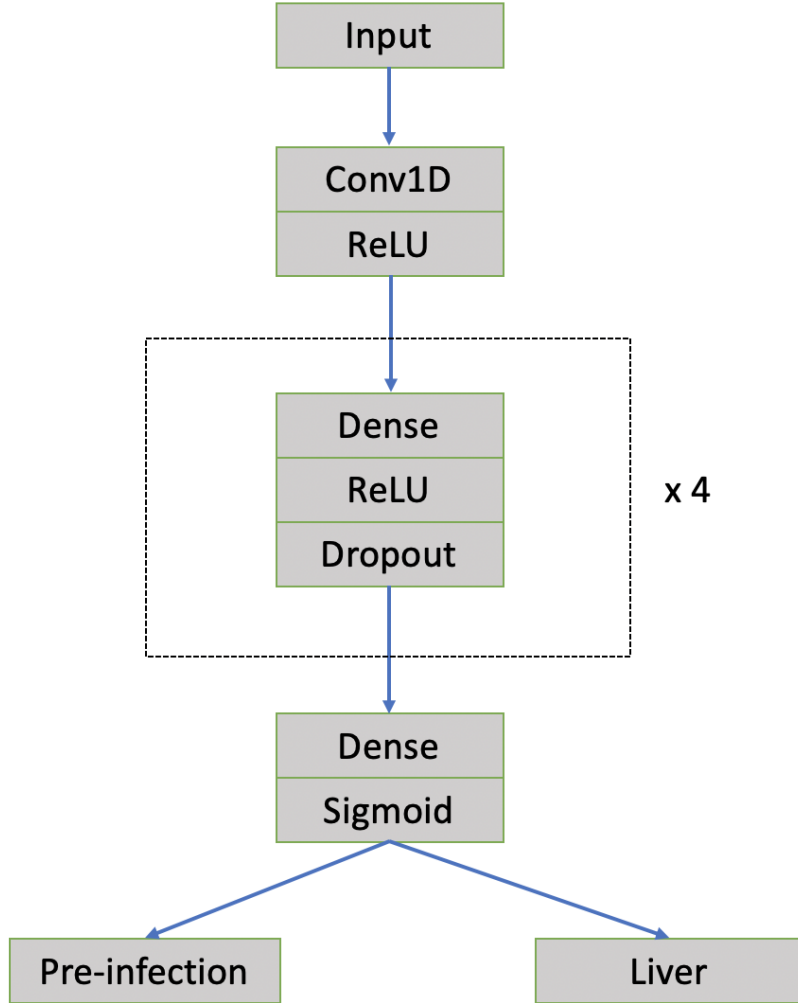


Figure 2.7: **Summary of DNN Architecture**

The deep convolutional neural network has a total of 10 neural layers that are listed as 1 Conv1D layer, 5 dense layers, and 4 dropout layers. The input is the batch of 10s incremental ECG records.

Overfitting is a widespread phenomena in deep neural networks that means the DNN model tends to perform much better in the training process than testing. Inclusion of the dropout layer could substantially reduce overfitting by randomly filtering out a subset of the input hidden units into the next neural network layer [67]. We applied a fraction of 0.5 to

the dropout layer. The weights of the neural networks were initialized as random uniform in the parameter space [68]. RMSprop optimizer was chosen for its ability in automatically adjusting the learning rate by the magnitudes of recent gradients [69].

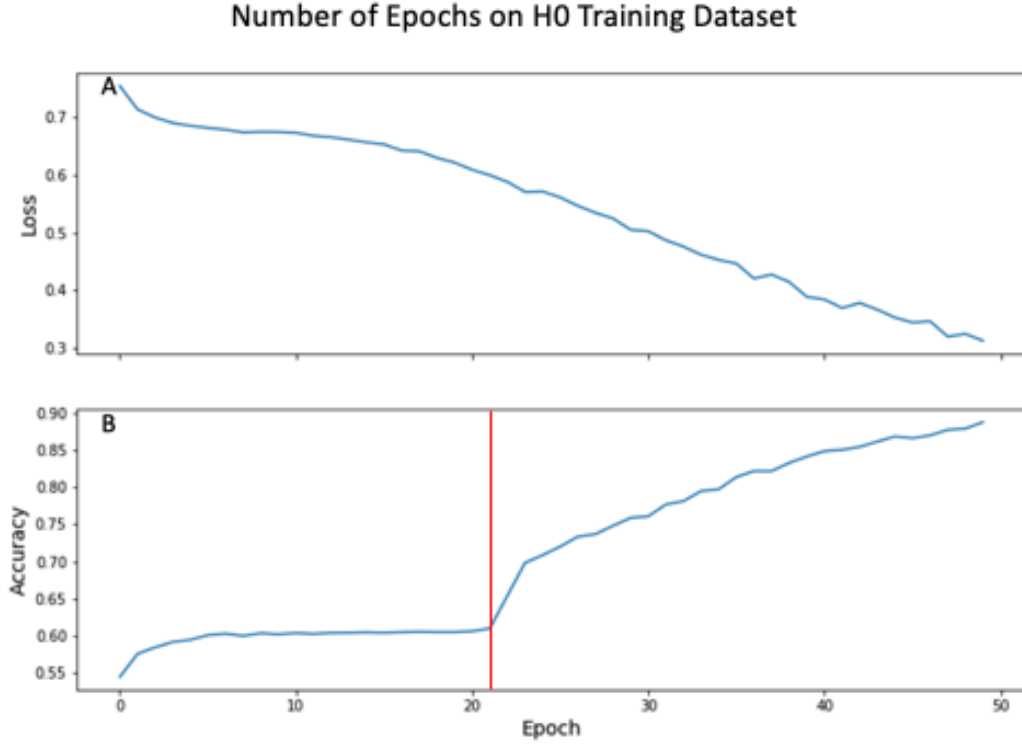


Figure 2.8: **Number of Epochs on H0 Training Dataset**

X-axis denotes the number of epochs, and Y-axis denotes the percentage value. Figure A. shows the decreasing loss value over epochs of H0 training dataset. Figure B. shows the increasing accuracy value over epochs of H0 training dataset. Red line is the cutoff at 22 epochs.

Overall, the hyper-parameters of the DNN architecture were chosen by running various combinations through grid searching. We mainly tuned the model by trying different combinations of neural network layers, different Dropout rates, and adding or dropping certain layers. The number of epochs was chosen by checking the accuracy and loss changes in the training dataset over a range of epochs (shown in Figure 2.8). In Figure 2.8.B, the prediction accuracy increased dramatically after 22 epochs in training dataset of H0, while the same accelerated decreasing of loss value is also observed in Figure 2.8.A. This indicates that

overfitting in the DNN model occurs after 22 epochs. In Figure 2.8.B, the training accuracy stays stable between epochs ranges of 10 to 22, while the loss value is still decreasing a little bit over each epoch. We set the epochs to be 20 to keep the performance of the model, and it was applied to all ECG hour groups of the day. To control the bias from the model, we kept the DNN model and the associated hyperparameters to be same for all 24 hour ECG groups.

2.6.2 MODEL EVALUATION

The independent testing dataset and a 10-fold cross-validation method were used to evaluate the model performance [70]. The training dataset were separated into 10 random subsets, and each subgroup is served as a validation dataset in the training process while the other 9 subgroups are fed into model training. The trained DNN model was further tested on the independent testing set which was picked out before model training. Classification results of 10-fold cross-validation and independent testing set of each hour are compared and shown in Section 3.2.

Receiver operating characteristic (ROC) and the area under ROC curve (AUC) were computed to further compare the model's ability in distinguishing pre-infection and liver stage of infection by different ECG hour groups [71]. The associated statistics: accuracy (Equation 2.3), sensitivity (recall) (Equation 2.4), specificity (Equation 2.5), PPV (precision) (Equation 2.6), and F1 score (Equation 2.7) are calculated [25]. F1 score is the harmonic mean of PPV (precision) and recall (sensitivity) that ranges from 0 to 1. The larger F1 score, the better performance of the model.

$$Accuracy = \frac{TP + TN}{N} \quad (2.3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2.4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2.5)$$

$$PPV = \frac{TP}{TP + FP} \quad (2.6)$$

$$F1 = 2 * \frac{PPV * Recall}{PPV + Recall} \quad (2.7)$$

Where N is the total number of samples, TP stands for the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

2.7 ASSOCIATIONS BETWEEN ACCELEROMETER AND ECG

We observed that the classification performance of the deep convolutional neural networks that we built in Section 2.6 varied dramatically with different ECG hour groups (Section 3.2). The model tends to perform better with the ECG data recorded during nighttime hours, while performs relatively worse with daytime ECGs, especially the ECG recorded during hours of H7 and H18. We suspected that the circadian changing of experimental host subjects' daily activities affect the stability of the ECG signal. The effects of activity on the heart rates and ECG signals have been well known [72, 73].

We tested the associations between accelerometer data and ECG signals through the regression analysis between the DNN model performance and the standard deviation of accelerometer data of each hour group. The rigid regression model was picked, as it can handle possible multicollinearity in our dataset by adding additional *L2-norm* penalty to the weights [74]. We used F1 scores of the model to represent the prediction performance of the DNN model [75].

- The regression model is formulated as Equation 2.8:

$$Y_h = \omega_i Xstd_{ih} + \alpha ||\omega_i||^2 + \epsilon_h \quad (2.8)$$

where Y_h denotes the F1 score of the DNN model at hour h , $Xstd_{ih}$ denotes the standard deviation of normalized accelerometer of subject i during hour h and ω_i denotes the weights associated to subject i .

- The corresponding objective function is then described as Equation 2.9:

$$\min_{\omega} ||Xstd_{ih}\omega_i - Y_h||^2 + \alpha ||\omega_i||^2 \quad (2.9)$$

We applied the regression model to the F1 scores of the DNN model of 24 hour ECG groups against the standard deviations of the corresponding normalized accelerometer data of pre-infection and liver stage, separately. The coefficient of determination (R^2) was calculated as the measurement of association level. In summary, we fixed the response variable of the regression model Y to be the F1 scores of the DNN model and the standard deviations of accelerometer data of two classes were applied individually.

2.8 RNA-SEQ ANALYSIS

The raw count RNA-Seq data of each experimental monkey subject was queried and downloaded from the relational SKED database by the python API module (Appendix B). In total we have 17 experimental RNA-Seq replicates of pre-infection from E06 and E30 non-resilient hosts, and 10 RNA-seq replicates of the corresponding liver stage as shown in Table 2.7. We did not include the resilient host group E07B in the RNA-Seq analysis, since there are no pre-infection data.

DESeq2 normalization method was applied to the RNA-Seq raw count data, which normalized the raw count of RNA-Seq by the median of ratios method [76]. The normalized RNA-Seq data of two classes were evaluated by principal component analysis (PCA), where

Table 2.7: **Number of RNA-Seq Replicates in Two Classes**

	Class	Experiment	Number
RNA-Seq	Pre-infection	E06	11
		E30	6
	Liver Stage	E06	6
		E30	4
		E07B	12

we decomposed each sample subject into two main components via eigenvalue decomposition [77].

2.8.1 DIFFERENTIAL EXPRESSION

Differences in the transcriptome between pre-infection and liver stage of malaria infection have the potential to reveal the intrinsic mechanisms of molecular changes underlying extrinsic telemetry shifts. Circadian or clock genes are the key regulatory components of circadian rhythms [78], whose expressions are extensively studied in mammals [79]. Circadian rhythms are also essential to control the maximum fitness of parasites to hosts [42], and govern the ECG signal stability.

We used DESeq2 [76] as the approach to find the differential expressed genes between pre-infection and liver stage group. Mathematically, the normalized read count values of gene i in sample j are modeled by a negative binomial distribution [80] as shown in Equation 2.10:

$$K_{ij} = NB(\mu_{ij}, \sigma_{ij}^2) \quad (2.10)$$

Mean parameter μ_{ij} of gene i in sample j was estimated by the product of condition dependent factor $q_{i,\rho(j)}$ (which was estimated from experimental conditions) and a size factor

s_j as Equation 2.11 shows. And the variance σ_{ij}^2 is modeled by adding a raw variance term $s_j^2 v_{i,\rho(j)}$ to the noise μ_{ij} (Equation 2.12).

$$\mu_{ij} = s_j * q_{i,\rho(j)} \quad (2.11)$$

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)} \quad (2.12)$$

The transcriptomic profile of the liver group was compared to pre-infection group. Since we have more than 5,000 gene features in the transcriptome, p-values of the differential expression tests are controlled by false discovery rate (FDR) with a p.fdr value of 0.05 [81]. In multiple hypothesis testings, type I error is prominent thus should be controlled.

We are particularly interested in the differential expressions of 7 circadian genes, which are the representatives of the circadian genes in the literature [82] (summarized in Table 2.8).

Table 2.8: **Brief Summary of Circadian Gene Functions**

Circadian Genes	Description [83]
PER1	A member of the period gene family that served as primary circadian pacemaker in the mammals.
CRY1	Cryptochrome circadian regulator 1: a key component of circadian core oscillator complex.
CRY2	Cryptochrome circadian regulator 2: a key component of circadian core oscillator complex.
ARNTL	Encodes a heterodimer with CLOCK, which activates transcriptions of PER1, PER2, PER3, CRY1 and CRY2.
NPAS2	Functions as a part of molecular clock regulator in mammals.
CREBBP	Ubiquitously expressed that act as a circadian transcriptional coactivator.
AANAT	Controls day/night rhythm in melatonin production in mammals.

2.8.2 GENE SET ENRICHMENT

Too many differentially expressed genes provide various pieces of information that can be noisy, thus misleading us from finding true underlying biological changes. Each of those significantly differentially expressed genes could serve in multiple biological functions and play different roles, which make the differential expression analysis alone hard to target the key metabolic shifts.

Geneset enrichment or pathway enrichment analysis was further done, which maps the differentially expressed (DE) genes to different biological pathways [84]. We used an in-house built script that based on a hypergeometric test to perform the enrichment analysis by referring to MSigDB pathways [85]. Mathematically, we used hypergeometric distribution to measure the statistical significance of enriched pathway p as in Equation 2.13.

$$Pr(p) = \frac{\binom{P}{p_k} \binom{N-P}{n-p_k}}{\binom{N}{n}} \quad (2.13)$$

Where p_k refers to the significantly DE genes in pathway p (out of n total number of significant DE genes) from a total number of N genes with a total of P number of genes in pathway p . FDR control was applied to all the P pathways of hypergeometric tests.

CHAPTER 3

RESULTS

3.1 ACCELEROMETER CHANGES IN DIFFERENT HOUR GROUPS

In Figure 3.1, we plotted the standard deviation of the normalized accelerometer in pre-infection stage of a day by taking six non-human primate subjects (two of each experiment) as examples. All the subjects have a higher standard deviation of activity in the daytime hours than the nighttime hours. This observation is consistent with our knowledge that during the nighttime hours the subjects are sleeping and their activities are limited. Interestingly, subject 12C136 and H12C8 from the resilient hosts in E07B (dashed lines in Figure 3.1) have a higher activity level when compared to the non-resilient subjects. We expected the resilient hosts to be more activity in the liver stage of malaria infection as they are more immune to the *Plasmodium* parasites than non-resilient hosts.

We further compared the standard deviation of normalized accelerometer data between pre-infection and liver stage at each hour of a day. Figure 3.2 shows that there are no significant changes of activity levels between pre-infection and liver stages of infection. Furthermore, the activity levels of each subject across the time were checked (three examples shown in Figure 3.3.A, 3.3.B, and 3.3.C). The three example subjects from three experimental batches are compared individually. There were no significant changes of the activity levels were found between pre-infection and liver stages, except for the hours of daily malaria parasite checking time in liver stage, namely 1:00 PM for E07B, 1:00 PM for E06, and 2:00 PM for E30. The observation is consistent with our expectation that during the liver stage of *Plasmodium* infection, the hosts do not have any prominent infection symptoms [5].

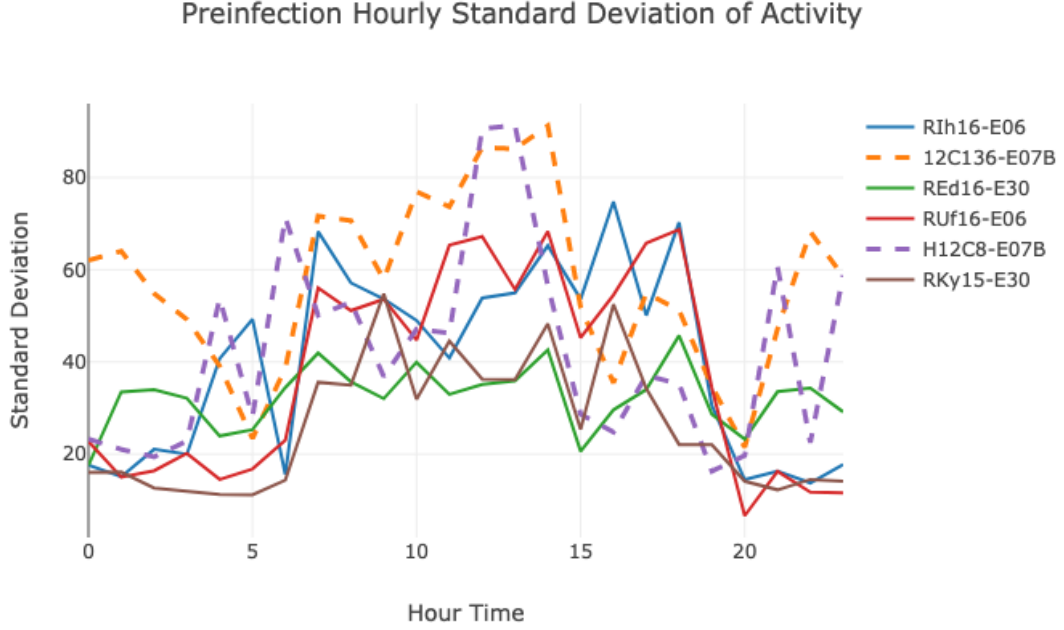


Figure 3.1: **Standard Deviation of Accelerometer During a Day**

The x-axis is the hour time of a day, where 0 stands for the time range from 0:00am to 1:00am and so on so forth. The y-axis is the standard deviation of accelerometer of each subject in the given time range. Six experimental subjects are plotted as examples. The dashed lines represent parasite resilient hosts, while solid lines represent non-resilient hosts.

3.2 DNN CLASSIFICATION RESULTS

3.2.1 PREDICTION RESULTS

The deep convolutional neural network model (described in section 2.6) was first applied to the 10s segments of raw ECG data from each subject separately. The prediction results of each subject were shown in Figure 3.4. We used F1 scores to measure the model performance, as it is the harmonic mean of precision and recall [86]. There are two subjects (RC115 and RUf16) from E06 parasite non-resilient hosts have slightly higher than average F1 scores, which could be caused by the more illness of the hosts, thus the ECG signals differ more between

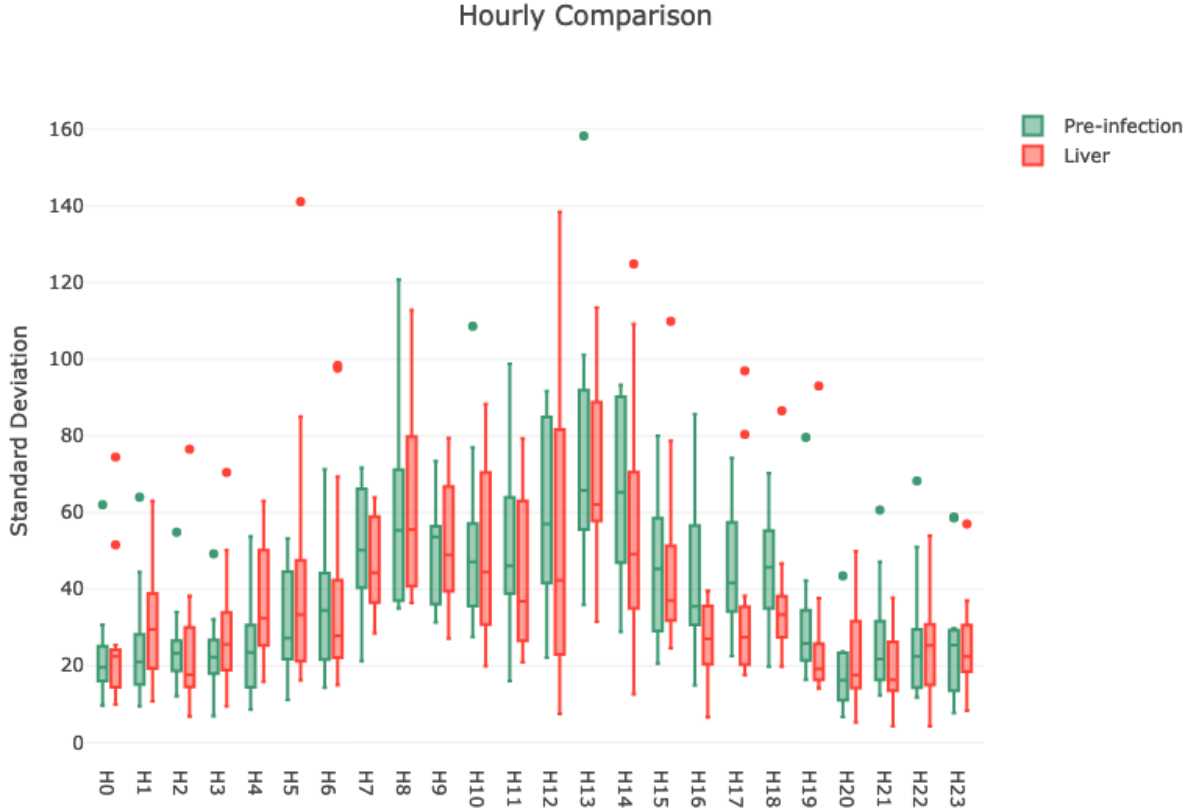


Figure 3.2: **Comparison of Standard Deviations of Accelerometer Data Between Groups**

The x-axis refers the hours of a day. The y-axis refers to the corresponding standard deviation of normalized accelerometer across all the subjects with both resilient and non-resilient hosts. Green color denotes the pre-infection stage, while red denotes the liver stage.

the two stages. While two subjects (12C136 and H12C8) from E07B have slightly lower F1 scores, which could be similarly caused by less illness and the ECG signals do not change at all between pre-infection and liver stage. The remaining 9 experimental subjects have the similar prediction performances with the average F1 score of 0.73. The slight differences of model prediction performance between each subject indicate that the stability of ECG

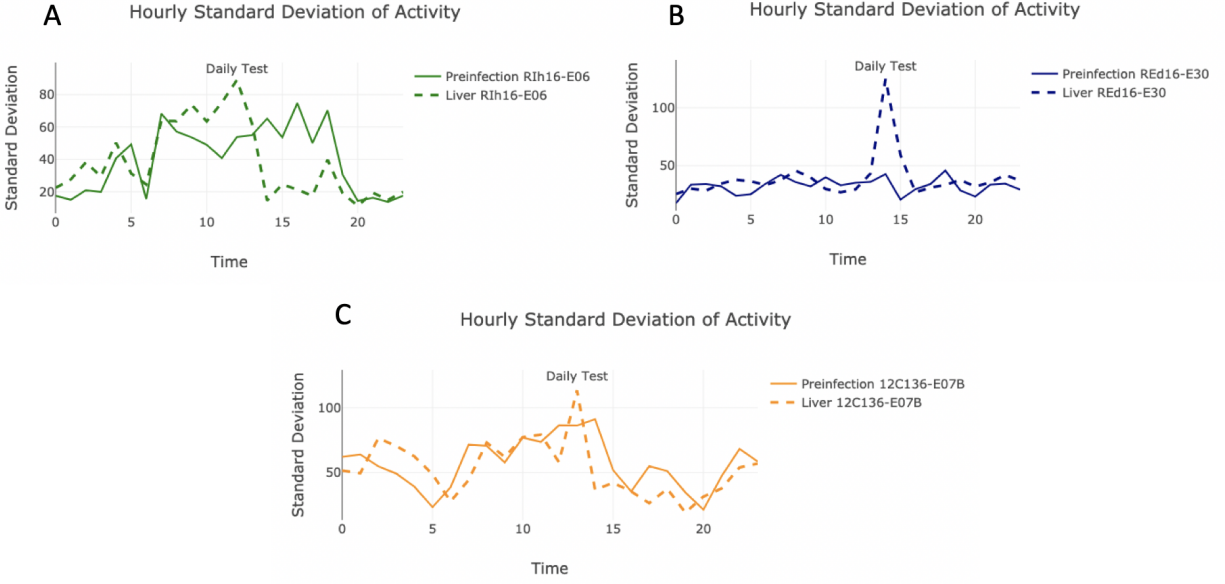


Figure 3.3: Comparisons of Activity Levels Between Pre-infection And The Liver Stage

All dashed lines represent the standard deviations of accelerometer in the liver stage, while the solid lines represent pre-infection stage. Figures A, B, and C show the sample subject RIh16-E06, REd16-E30, and 12C136-E07B, respectively. While E07B is *P.knowlesi* parasite resilient host, and E06 and E30 are non-resilient hosts. The time of daily experimental collections are denoted as shown in text.

signals have minor differences between subjects with the resilient hosts carry less differences between two classes and non-resilient hosts carry more.

The DNN model was further applied to the raw ECG segments from each hour group of all subjects with the same hyperparameter settings. We fed each ECG hour group into the DNN model individually, and the trained model was applied directly to the independent testing set. The input ECG data from different hours are the only differences in training and testing the DNN model of 24 groups. The prediction accuracy of the DNN model in distinguishing pre-infection and liver stage of *Plasmodium* infection were listed in Table 3.1 and shown in Figure 3.5.

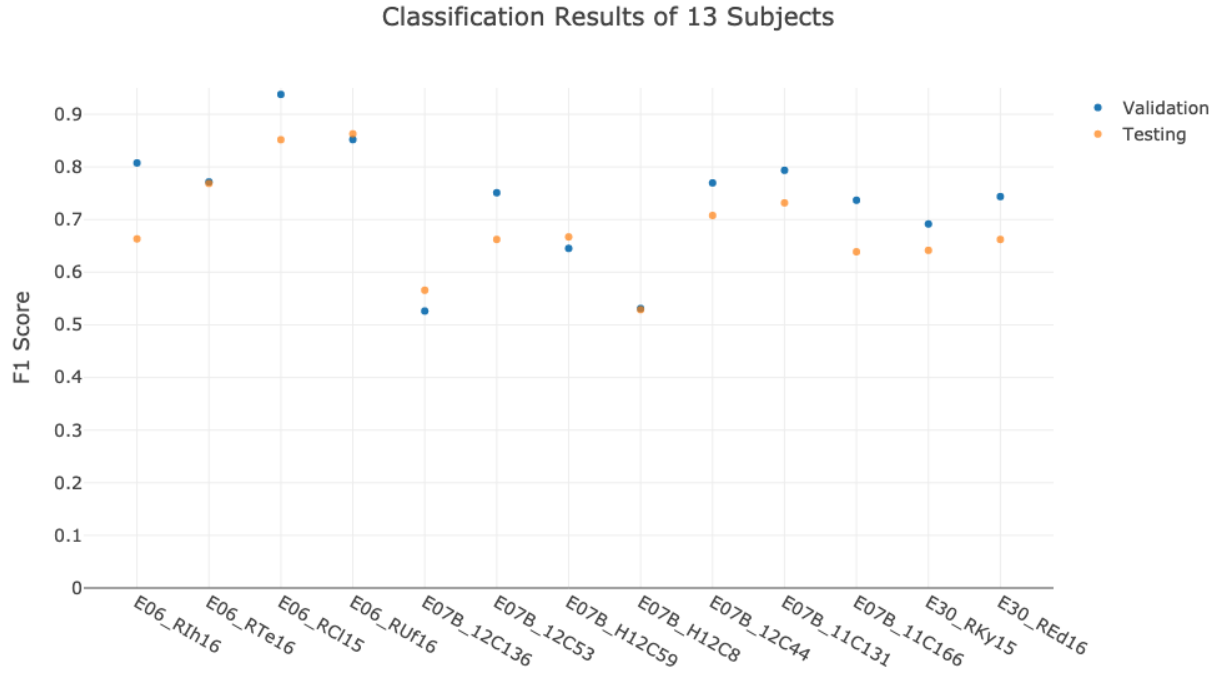


Figure 3.4: **Classification Results of 13 Subjects**

The x-axis refers to the subject and the corresponding experiment batch. The y-axis refers to prediction F1 scores (described in section 2.6.2. Blue dots are the F1 scores of 10-fold cross validation set, while orange dots are F1 scores of independent testing set.

Table 3.1: **Training And Testing Results of 24-Hour Groups**

The 10-fold cross validation results on training and validation dataset are listed along with the results of prediction accuracy on the independent testing dataset

Hour	10-fold CV Training Accuracy	10-fold CV Training F1 Score	Testing Accuracy	Testing F1 Score
0	0.653092	0.653215	0.619251	0.631339
1	0.671104	0.669698	0.661323	0.655559
2	0.573605	0.573375	0.559389	0.557685
3	0.715579	0.712945	0.650551	0.650551
4	0.785829	0.785955	0.721465	0.721465
5	0.726170	0.724754	0.687050	0.683222
6	0.699857	0.699438	0.636274	0.635201
7	0.547807	0.513195	0.557531	0.554147
8	0.604410	0.611169	0.555973	0.666674
9	0.622994	0.624431	0.620513	0.620513
10	0.685102	0.684519	0.617358	0.618392
11	0.670000	0.669629	0.642568	0.643405
12	0.727043	0.727427	0.709655	0.711453
13	0.677148	0.676845	0.662722	0.665067
14	0.695922	0.696103	0.650923	0.651107
15	0.639309	0.639399	0.611437	0.610928
16	0.672499	0.671495	0.577047	0.576847
17	0.623047	0.621170	0.555310	0.556319
18	0.552069	0.551991	0.504802	0.615472
19	0.692158	0.692442	0.640640	0.641867
20	0.760047	0.760401	0.668020	0.668356
21	0.841590	0.842496	0.759793	0.762112
22	0.788276	0.788982	0.697140	0.699375
23	0.838173	0.837655	0.737598	0.736952

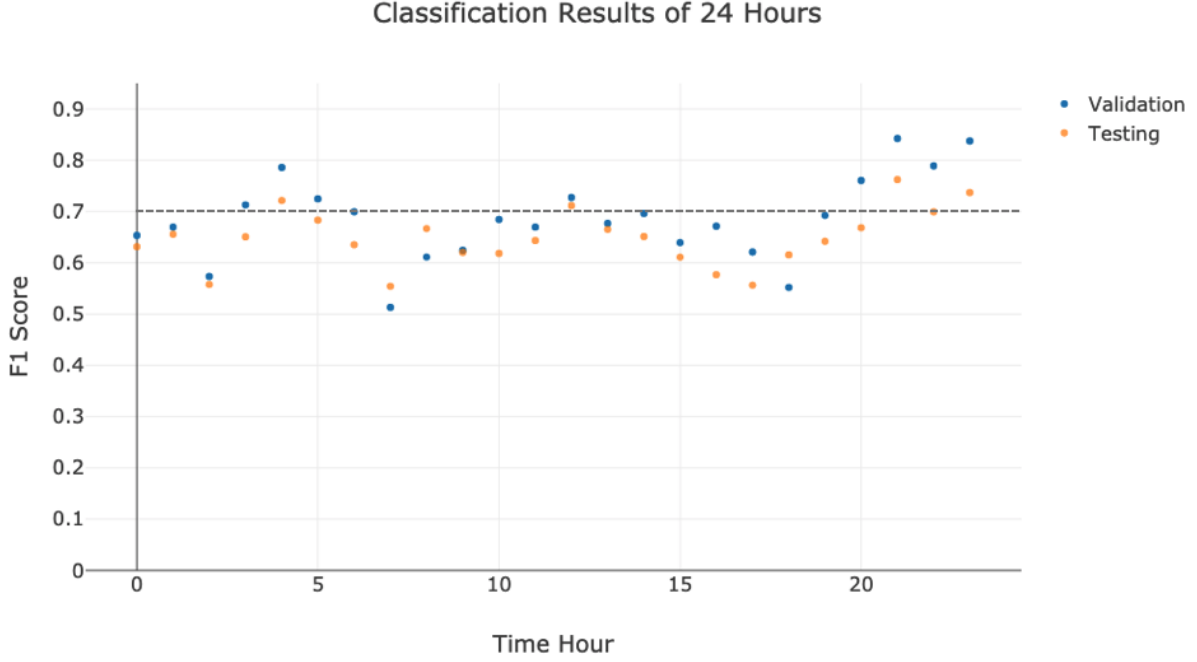


Figure 3.5: **Classification Results of 24-Hour Groups**

The x-axis refers to the hour group of a day that the ECG data was recorded from. The y-axis refers to the model prediction F1 score (described in Section 2.6.2). Blue dots are the F1 scores of 10-fold cross validation set, while orange dots are F1 scores of independent testing set. The dashed grey line is the horizontal line of 0.7.

The 10-fold cross-validation results and independent testing results in Figure 3.5 of each hour are quit close, which means the DNN model has relatively similar performance in validation and testing. Even the testing F1 scores of H7, H8, and H18 are slightly higher than training and validation set. Therefore, we think the DNN model did not have overfitting problems in the training process.

Interestingly, there were strong time-dependent patterns of the DNN model performance with the average F1 score of 0.65. The DNN model tends to have a higher F1 score with the inputs of ECG data recorded during nighttime hours than the daytime hours. Especially, the F1 scores of hour groups H21, H22, and H23 have an average increase of 0.1 from the average of all hour groups. The highest F1 score of the prediction model occurred at H21 with 0.84 in training and validation set and 0.76 in testing set, while the lowest F1 score occurred at

H7 with 0.51 in training and validation set and 0.55 in the testing set. The confusion matrix of H21 (Figure 3.6) also indicated that the model had a balanced prediction accuracy in distinguishing pre-infection and liver stage of *Plasmodium* infection (confusion matrices of other hour groups refer to Appendix A). We further evaluated the experimental recordings of ECG groups with bad prediction accuracy such as H2, H7, and H18. Interestingly, H7 and H18 are the daily light changing hours in the experimental cage, thus we expected these hours to have a worse prediction accuracy due to the outside disturbance.

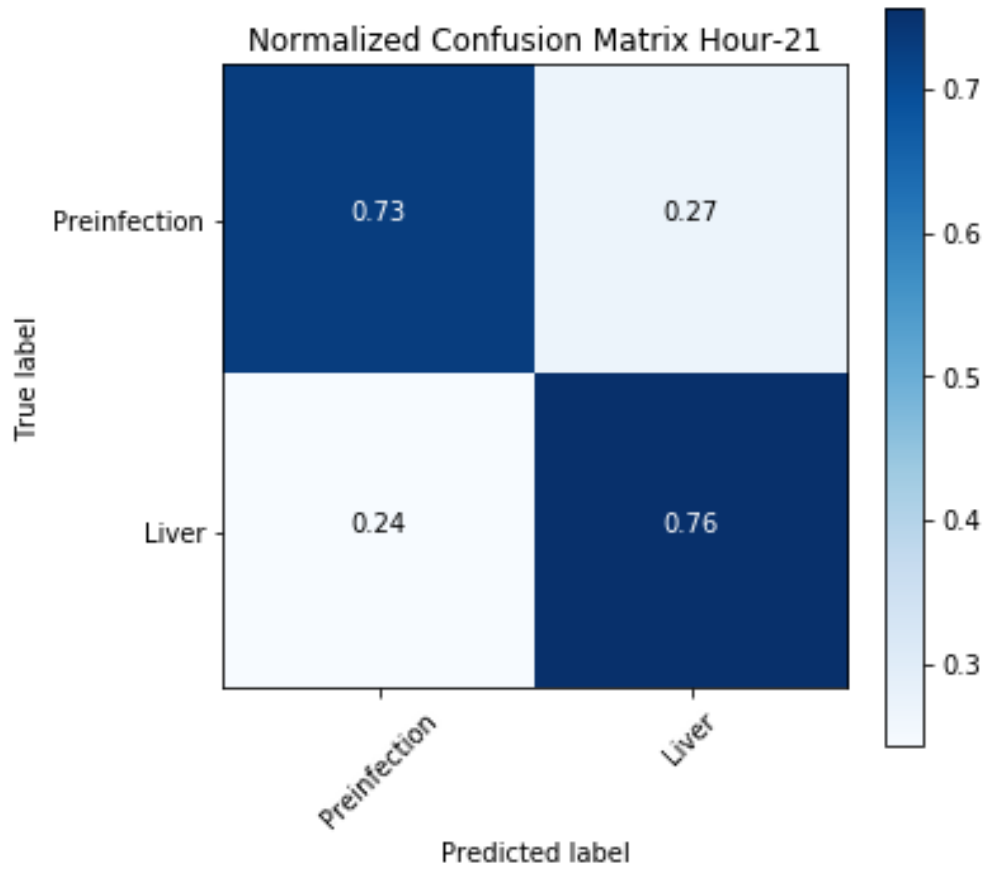


Figure 3.6: The Normalized Confusion Matrix of H21

The x-axis refers to the DNN model predicted ECG class label, while the y-axis is the true class label of the independent testing dataset. The color bar denotes the prediction accuracy range.

3.2.2 ROC COMPARISONS

We observed the large differences of model prediction performance in different hour groups of ECG data. The receiver operator characteristic (ROC) and area under curve AUC were further calculated to compare these hours (Figure 3.7). The ROC curve of H7 is very close to the diagonal line, which means that the ECGs recorded during H7 is randomly classified by the DNN model as shown in Figure 3.7. Also, the ROC curves of these hours were well separated, which confirmed that the DNN model had a very different prediction performance with the ECG data recorded in those hours.

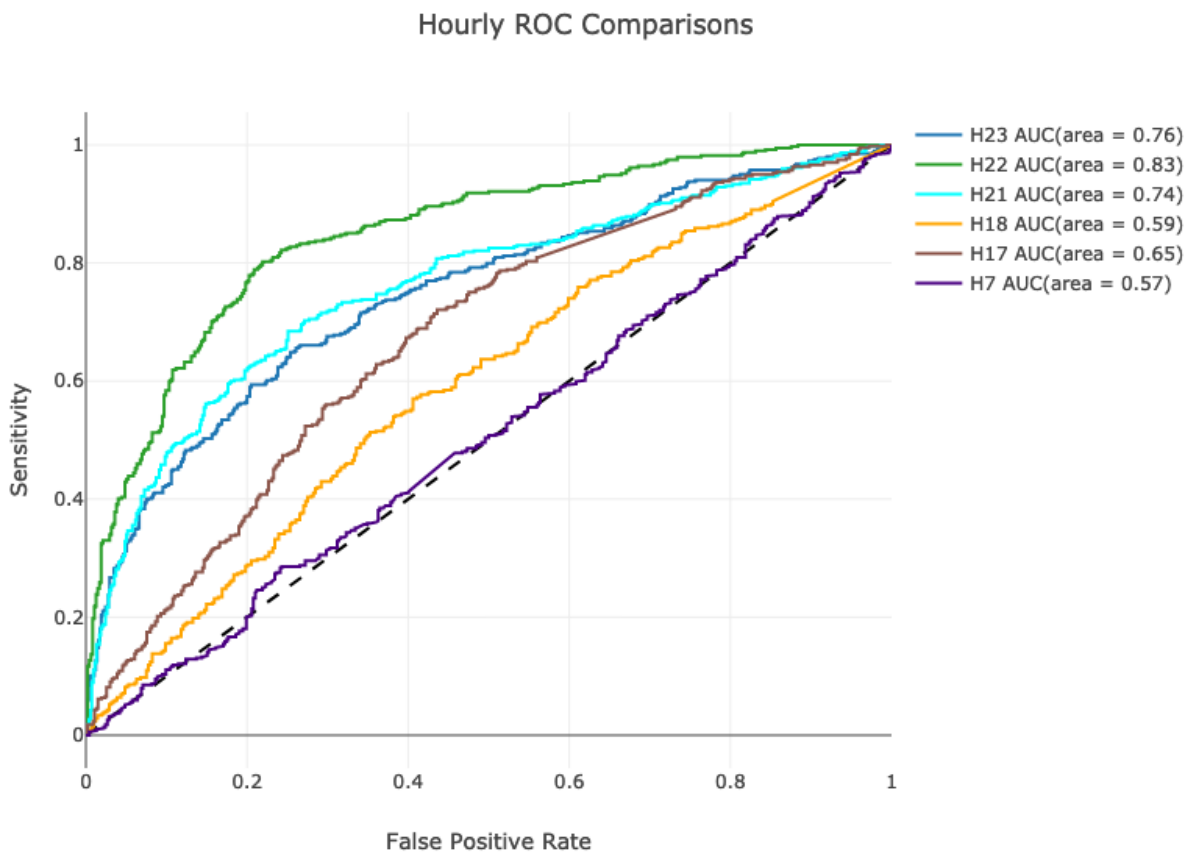


Figure 3.7: **ROC Comparisons of Different Hour Groups**

Receiver operating characteristics (ROC) of the three hour groups with low prediction accuracy and three hour groups of high prediction accuracy were shown with their AUC values. The ECG recorded during H22 had the highest AUC of 0.83, and H7 had the lowest AUC of 0.57.

We applied the DNN architecture with same hyperparameter settings to the raw 10s-segmented ECGs of different hour groups individually, hence, the input ECG data were the only difference between the groups that caused the large differences in DNN model prediction performance. We proposed two main reasons that caused the model prediction performance differences with the input ECG data recorded during different hours: 1. During the daytime hours, the experimental subjects have higher activity levels, which affect the stability of ECG signals; 2. The malaria parasites infection related ECG changes may be less prominent in daytime hours than nighttime hours. Nicole *et al.* [87] reported that *Plasmodium* parasites infected blood cells burst at midnight, which was consistent with our findings. It could also help to explain why our DNN model achieved a relatively high prediction accuracy of malaria parasite infection with the ECGs recorded during nighttime hours.

3.3 ASSOCIATIONS BETWEEN ACCELEROMETER AND THE DNN MODEL

The huge impact of physical activity changes to cardiac characteristics were well known [57]. We further did association analysis between the accelerometer data and the DNN model prediction performance to see the actual changes from pre-infection to the liver stage. We used F1 scores to measure the DNN model prediction performance as described in Section 2.6.2. The linear relationship between the F1 scores and the standard deviation of normalized accelerometer data in pre-infection and the liver stage were explored separately as shown in Figure 3.8 and Figure 3.9, respectively. In general, we found there were a stronger linear relationships between F1 scores and activity levels of each subject in pre-infection than the liver stage. For instance, the subjects of RC115 in E06, and 12C136, 12C44, 11C131 and 11C166 in E07B show stronger linearity in pre-infection than the liver stage.

Table 3.2: **R-squared of Two Classes**

The table lists the R-squared values of the ridge regression model in two classes separately

	Pre-infection	Liver
R^2	0.8176	0.5622

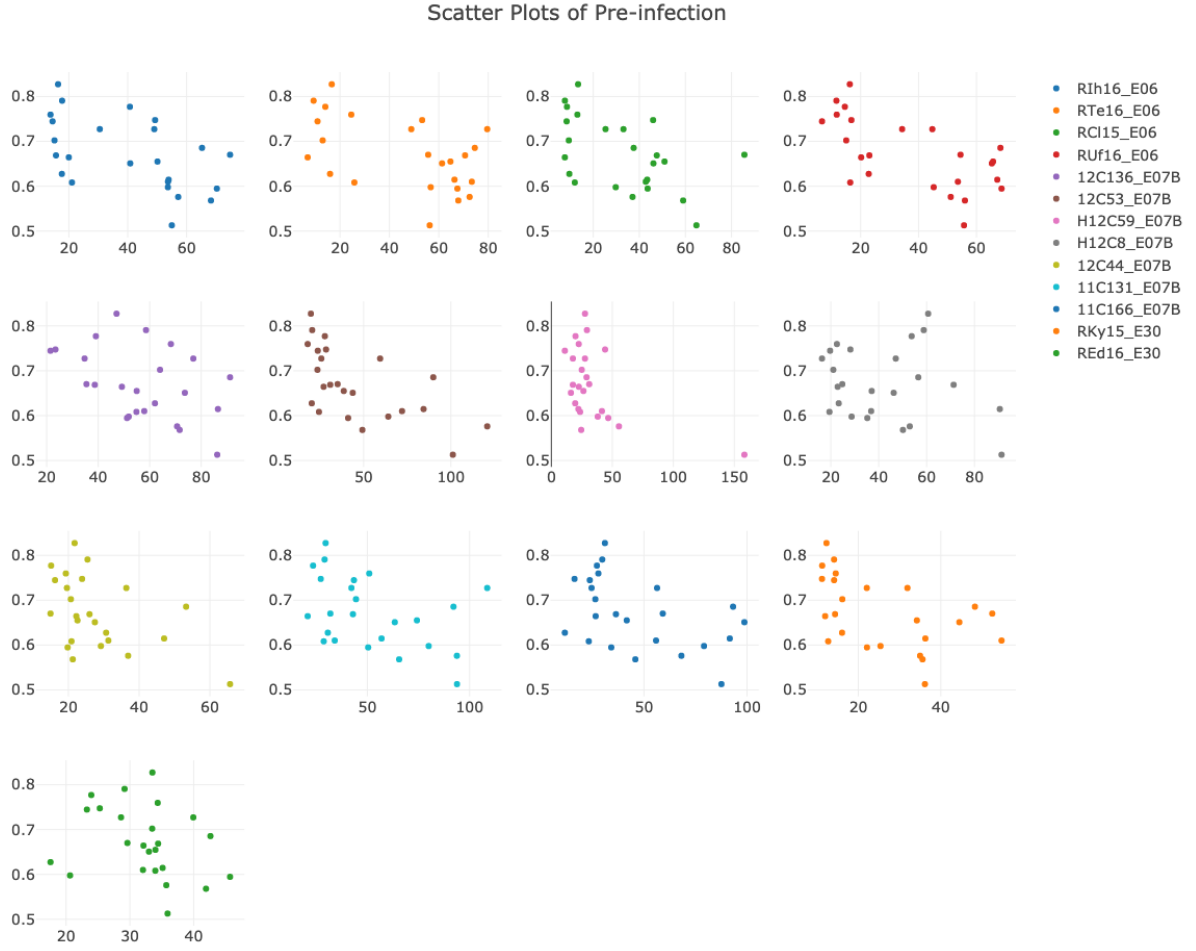


Figure 3.8: **Scatter Plots of Pre-infection Class**

The x-axis of each subgraph are the values of standard deviation of the normalized accelerometer. The y-axis are the F1 scores of the DNN model prediction.

The coefficients of determination (R^2 s) of the regression model of two classes were summarized in the Table 3.2. The R^2 dropped from 0.8176 in pre-infection to 0.5622 in the liver stage of infection. This big drop of R^2 indicated that the linear relationship between the the DNN model prediction performance and the accelerometer data was collapsed after parasite infection.

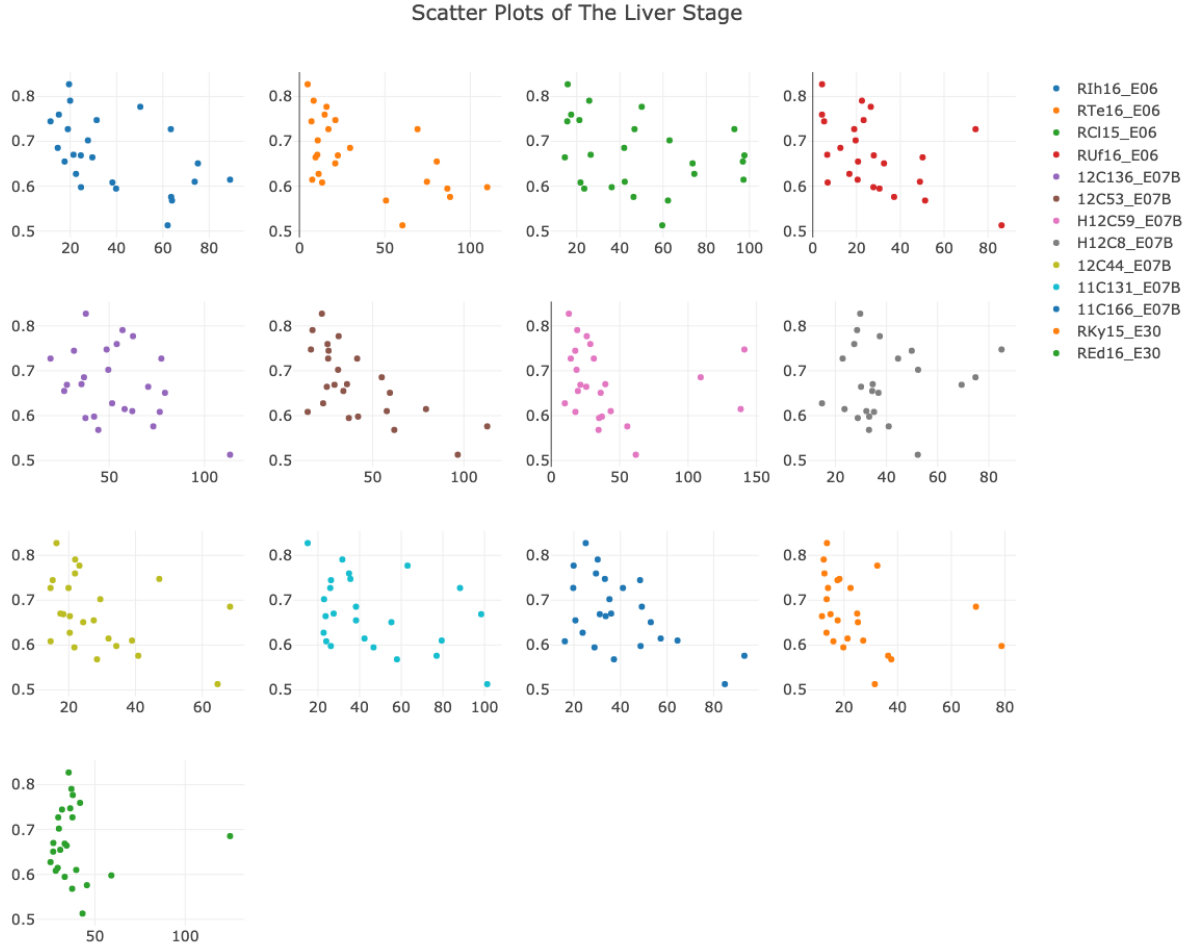


Figure 3.9: **Scatter Plots of The Liver Stage**

The x-axis of each subgraph are the values of standard deviation of the normalized accelerometer. The y-axis are the F1 scores of the DNN model prediction.

We have proved that our DNN model could effectively distinguish malaria pre-infection and the liver stage by using raw ECG data alone in Section 3.2. However, the model prediction performance was closely linked to the stability of ECG signals. Therefore, we could assume the F1 scores of the DNN model are good representatives of ECG signals. The disruption of associations between accelerometer data and model F1 scores from pre-infection to the liver

stage, thus equivalent to the disruption of associations between activity and ECG signals in the liver stage. As studies have shown that the physical activity levels could affect heart rates and the stability of ECG signals [88, 89, 90]. While we did not observe significant differences in daily activity levels of the hosts between pre-infection and liver stages (Figure 3.2). Hence, the disruption of association between activity and ECG in the liver stage proved that there are true differences of ECG signals between pre-infection and liver stage, which was not due to activity. The impact of parasite infection to the ECG signals become dominate in the liver stage.

3.4 RNA-SEQ ANALYSIS RESULTS

3.4.1 PRINCIPAL COMPONENT ANALYSIS

Principal component decomposition of the transcriptomic data of subjects from E06 and E30 (non-resilient hosts) was shown in Figure 3.10. The subjects of the two classes are well separated by transcriptomic data with only one sample subject from liver was clustered into pre-infection group by the top two principal components (PCs).

Transcriptomic profiles of sample subjects from pre-infection and the liver stages of *Plasmodium* non-resilient hosts can be well separated by the top two PCs. PCA results indicated that there are clear difference of transcriptomics between the two class.

3.4.2 DIFFERENTIALLY EXPRESSED GENES

A total of 1,164 genes are significantly differentially expressed (DE) between pre-infection and liver stage after false discovery rate (FDR) control ($q - value \leq 0.05$) [81]. Among these 1,164 significantly differentially expressed genes, 491 genes are underexpressed, and 673 genes are overexpressed in liver stage.

Interestingly, two of the major circadian regulatory genes CRY1 and CRY2 are significantly underexpressed in liver stage when compared to pre-infection (shown in Figure 3.11 and summarized in Table 3.3). CRY1 and CRY2 genes encode a key circadian core oscillator

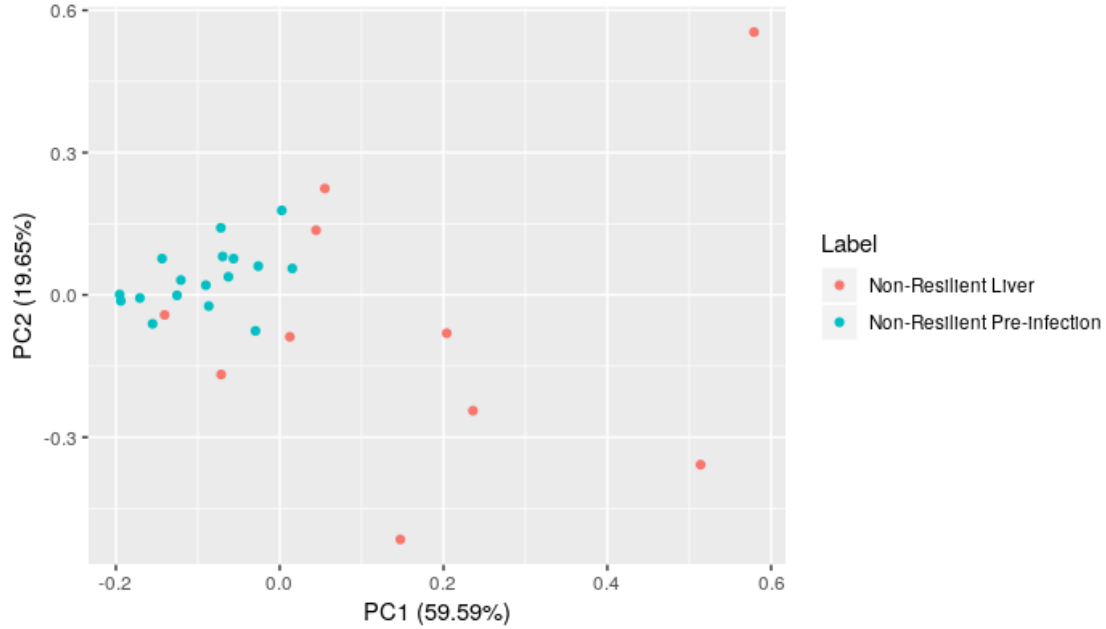


Figure 3.10: **PCA of Normalized RNA-Seq Data**

The x-axis denotes the first principal component of the transcriptomic profiles that accounts for 59.59% variance. The y-axis denotes the second major component that accounts for 19.65% variance. The blue dots represent sample subjects of pre-infection from E06 and E30(non-resilient hosts), and the red dots represent sample subjects of the liver stage also from E06 and E30.

complex that regulates circadian clock [82]. The underexpression of this core oscillator complex in liver stage indicates that *Plasmodium* parasites have already disturbed the circadian rhythms of the host at early liver stage. The circadian rhythms were reported to govern cardiac polarization, thus the disruption of circadian clock will potentially broke ECG signals [91]. Our findings that there are true ECG signal differences between pre-infection and liver stage are consistent with this disruption.

3.4.3 ENRICHED PATHWAYS

The significantly under- or over-expressed genes in liver stage are separately enriched to 2,145 biological pathways from MsigDB [85] that includes biological pathways from GO [92],

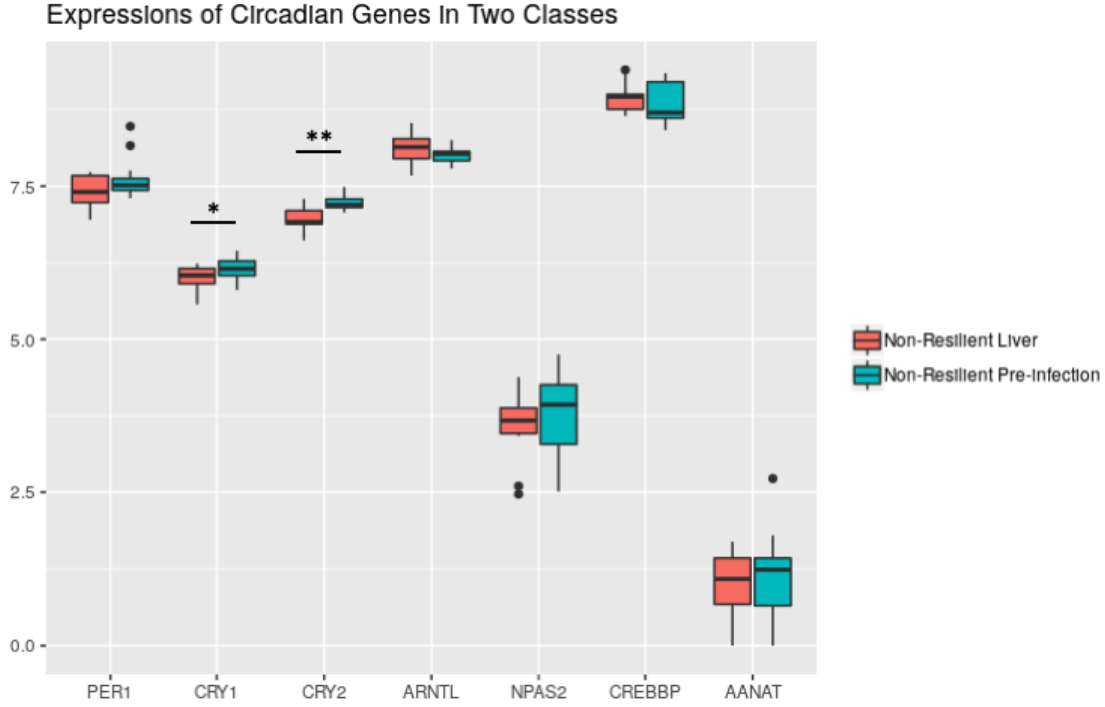


Figure 3.11: **Differential Expressions of Circadian Genes**

Transcriptomic expressions of 7 circadian genes of two classes from E06 and E30 non-resilient hosts are compared by boxplot. The * denotes the significant levels with ** group have a $q - value \leq 0.01$ and * groups have a $q - value \leq 0.05$.

Reactome [93], and KEGG databases [94]. There are 185 biological pathways are significantly enriched for underexpressed genes group ($q - value \leq 0.05$ after FDR control), while 420 pathways are significantly enriched for overexpressed genes group in liver stage (Table 3.4). It is worth noting that the significantly enriched biological pathways of underexpressed genes are the suppressed pathways in the liver stage, while the significantly enriched pathways of overexpressed genes are the over-activated pathways.

Two circadian rhythm pathways are significantly enriched for underexpressed DE genes group (Table 3.5). Both circadian clock and BMAL1 clock pathways are ranked within top 10% of the total reference pathways. There are no circadian related pathways enriched among

Table 3.3: **Differential Expressions of Circadian Genes**

Differential expression results of 7 circadian genes are listed. The p-values in the second column are after FDR control. The fold changes are the ratios of mean expression between liver stage and pre-infection group, smaller than 1 means the gene is underexpressed in liver stage and vice versa. Ranks are based on the significance level ($q\text{-value} \leq 0.05$), NA is not significantly differentially expressed.

Circadian Genes	q-value	fold.change	Rank
PER1	0.130622	0.977127	NA
CRY1	0.054979	0.973736	1198
CRY2	0.002174	0.964220	496
ARNTL	0.262859	1.013592	NA
NPAS2	0.248424	0.926175	NA
CREBBP	0.415372	1.010619	NA
AANAT	0.860955	0.957848	NA

Table 3.4: **Summary of DE Genes And Enriched Pathways**

Summary of significantly ($p \leq 0.05$ after FDR) over or under expressed genes and the corresponding number of significantly enriched pathways in liver stage compared to pre-infection.

	Genes	Pathways
Under Expressed	491	185
Over Expressed	673	420

overexpressed genes groups. The circadian clock pathway is the major clock in the hosts that receives input mainly from light and nutrient intake [93]. BMAL1 clock reacts as the major regulator for enhancing transcription of circadian clock genes [95].

Gene set enrichment results of underexpressed genes group have two major circadian pathways enriched, while no circadian pathways enriched for overexpressed genes group. This result indicates that some of circadian clocks are suppressed in liver stage, and the circadian

rhythms govern cardiac polarization, which may help to explain the intrinsic reasons why the ECG signals changed between pre-infection and the liver stage of infection.

Table 3.5: Enriched Circadian Pathways of Under-Expressed DE Genes

The rank of enriched pathways are based on the q-values. The smaller rank the more significant the pathway was enriched by underexpressed genes group in liver stage.

Pathway Names	q-value	Rank	Significant DE Genes
REACTOME CIRCADIAN CLOCK	0.042077	181	CRY1 CRY2 DBP MEF2C MEF2D AVP
REACTOME BMAL1 CLOCK NPAS2 ACTIVATES CIRCADIAN EXPRESSION	0.046680	184	AVP CRY1 CRY2 DBP

CHAPTER 4

CONCLUSION AND DISCUSSION

4.1 CONCLUSIONS

In this study, we used machine learning techniques to integrate a large amount of telemetry and transcriptomic data in the malaria parasite infection of non-human primates settings. A relational database was used for storing all the experimental measurements with different data types [56]. An *object-oriented* python module was therefore written to interact with the relational database including the functions of data querying and downloading.

We developed a deep convolutional neural network (CNN) with ten neural layers to classify *Plasmodium* parasite pre-infection and the early liver stage with the input of raw ECG data from 13 experimental *Macaca* monkey subjects. The raw ECG data were sampled into 10-seconds segment as an observation and further grouped into different hour groups of a day according to its recording time. Because the activity levels have a big impact to ECG signals, and thus the ECG signals vary significantly through the day [58, 59]. We have a minimized impact of activity within each hour group. If the ECG data were not grouped into different hour groups, the inter-group variances would dominate, thus affecting the overall prediction performance of our DNN model. The DNN model was trained on the 24 hours of grouped ECG data independently with the same architecture and same hyperparameter settings. And the DNN model was evaluated on an independent testing data set. The independent testing data sets were consisted of 1,000 10s-ECG observations for each hour (24,000 in total), which were randomly sampled before training.

We proved our DNN model could effectively distinguish malaria parasite pre-infection and the liver stage by feeding the raw ECG data alone. The DNN model achieved the highest

prediction accuracy of 0.760 on the independent testing ECG data recorded during H21, with the corresponding area under (AUC) the receiver operating characteristic curve (ROC) of 0.837. The DNN model had an average prediction accuracy of 0.638, an average F1 score of 0.647, and an average AUC of 0.671 on the independent testing ECG data of 24-hour groups. The prediction accuracy of the DNN model showed a strong time related pattern that the ECG data recorded during nighttime hours have significant higher F1 scores than daytime. ECG data recorded during hours of H7 and H18 had significantly worse prediction performances, which were proved to be light changing hours from experimental recordings.

There are true differences of ECG signals between pre-infection and the liver stage of malaria parasite infection. While we suspected the large differences in model prediction performance were caused by the impact of activity to the cardiac functions. In the nighttime hours, the ECG data were less impacted by the activity levels than the daytime hours. The explanation could help to explain the general patterns that the nighttime hours of ECG data tend to have better prediction accuracy than the day time ECG. With that, we think ECG changes are involved with malaria parasite infection especially dominate the ECG data during nighttime hours, which was consistent with the literature findings that malaria parasites infected cells burst at midnight [87].

We further proved the cardiac involvement in malaria parasite infection through the integration of accelerometer data, which was trying to further explain the large differences of model prediction performance with different ECG hour groups. Specifically, a ridge regression model was applied with the fixed response variable set to be the prediction F1 scores of 24 hours, and the accelerometer data of pre-infection or liver stage were set as explanatory variables separately. The coefficients of determination (R^2) of regression model dropped from 0.8176 of pre-infection to 0.5622 of the liver stage. This big drop of R^2 indicated the collapse of the associations between model prediction performance and the accelerometer data in the liver stage. As we assume the model prediction performance was highly linked to the stability of ECG signals, the collapse was equivalent to the disruption of associations

between ECG signals and the activity levels in the liver stage. Therefore, we concluded that the differences of ECG signals between pre-infection and the liver stage were not from the impact of activity to ECG, but from the effects of malaria parasites infection in the liver stage.

We further analyzed the RNA-Seq profiles of *Plasmodium Knowlesi* non-resilient hosts (E06 and E30) through differential expression analysis and gene set enrichment analysis. Interestingly, we found two primary circadian rhythm regulators (CRY1 and CRY2) were significantly underexpressed in liver stage when compared to pre-infection. Moreover, the corresponding circadian pathways were also significantly enriched for the underexpressed DE genes in the liver stage, which suggested the suppression of circadian rhythm pathways in the liver stage. As the cardiac polarization and repolarization were governed by circadian rhythms [91], the suppression of circadian rhythm pathways in the liver stage help to explain the intrinsic reasons of the ECG signal changes in the liver stage. We also found the mild cytokine-related pathways show significantly enriched with overexpressed DE genes in the liver stage. This early production of cytokine has been reported to be associated with protection of malaria parasites infection [96, 97].

In conclusion, through the integration of high-frequency telemetry data and transcriptomic profile, we found there are true ECG signal changes occur as early as in the liver stage of malaria parasite infection. We could utilize the ECG data alone to effectively early detect the malaria infection at liver stage by a DNN model. The ECG data with a lower activity impact have a better distinguishing ability. The RNA-Seq analysis suggested the changes of ECG signals in early liver stage may be caused by the suppression of circadian rhythms.

4.2 NOVELTIES

To systematically understand malaria parasites infection to non-human primates, we integrated different types of data from high-frequency telemetry data with different sampling frequencies such as 1,000 Hz ECG and 10 Hz accelerometer, to transcriptomic profile via

machine learning methods. The systematic approach not only provided us additional pieces of information, but also linked these pieces together that helped us better understand the interaction between malaria parasite infection and the hosts, from the disturbance of hosts' extrinsic behavior to the intrinsic biological pathway changes.

To the best of our knowledge, there are no existing studies to use ECG data alone to predict malaria parasite infection through machine learning methods. Though various machine learning methods have been applied to the ECG data in detection or classification of heart related diseases [38, 39, 98, 99], we are the first group to utilize raw ECG data in predicting malaria infection. We proved that there are cardiac involvement in malaria infection, and the ECG changes occur as early as in the liver stage. The predictive DNN model could successfully detect malaria parasite infection at early liver stage with a relatively high prediction accuracy by using the ECG data recorded during the host's inactive state.

4.3 CAVEATS

The study has several limitations. We only have experimental data from monkey subjects, which still have much differences to the human data, such as different sampling frequency of ECG. It remains to be experimentally validated and characterized that how ECG signals are disturbed during malaria parasites infections. We only have 13 non-human primate subjects involved in our analysis, which could bring bias to our findings. The 13 subjects are from two species, *Macaca Mulatta* (non-resilient hosts) and *Macaca Fascicularis*. If we could have more experimental subject replicates from the two species, the differences of telemetry and transcriptomic features could be further analyzed and grouped by species. Though we have shown both species have the similar patterns of telemetry data. Also, we do not have external data to test the generalizability of our DNN architecture that if a similar prediction performance pattern is still observed.

Though there is no golden standard for the minimum sample size in training deep neural network architectures, our model could be further improved by including additional sample

data sets. We only have 9,000 observations of each hour in training process and additional 1,000 observations as independent testing set. Also, the imbalanced sample size of each hour could also bring bias to our findings. Though the resampling method was applied to overcome the imbalanced sample size problem, we may have oversampled or undersampled certain hours that cause overfitting or underfitting problems to the classification model in some hour groups.

4.4 APPLICATIONS

The *object-oriented* python module could be further extended to be a universal API package that can be used to query and download data sets easily from relational databases. In the big data era, relational databases are widely used to store huge data sets with different data types [100, 101]. A user friendly python API will be super favorable in daily data querying and downloading processes.

The findings of cardiac involvement in malaria parasite infection in our study could foster the application of malaria infection detection through raw ECG signals in the future. In current clinics, malaria parasite infection is detected by blood smear that takes blood samples from patients and the parasites are detected under a microscopy [102]. This process is labor-intensive and requires an expert to check each blood smear manually. There is one study already proposed a machine learning method to automatically detect and diagnose malaria parasites on blood smear images [103]. However, it has the limitations that the patients should have the symptoms and seek for medical attention first. Our DNN model could detect the malaria parasite infection at early liver stage before occurrence of all the complications by using raw ECG data. The ECG signals are also quite easy to collect and measure with current technologies [7]. Therefore, the raw ECG data of human could also be included in malaria parasite infection detection in the future. An adjust DNN model is needed if trained on human ECG signals since the human ECG have different sampling frequency (200Hz) with monkey subject (1000Hz).

Other than the early detection of malaria parasite infection through raw ECG data, the detailed biological mechanisms behind cardiac changes could be further studied by integration of metabolics and proteomics data. A new drug target could be possibly developed to block the progression of malaria parasites infection by treating certain cardiac changes or related disruption of circadian rhythms.

BIBLIOGRAPHY

- [1] C. By-Nc-Sa, “World malaria report 2017,” p. 196.
- [2] R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay, “The global distribution of clinical episodes of *Plasmodium falciparum* malaria,” *Nature*, vol. 434, no. 7030, pp. 214–217, Mar. 2005. [Online]. Available: <https://www.nature.com/articles/nature03342>
- [3] S. I. Hay and R. W. Snow, “The Malaria Atlas Project: Developing Global Maps of Malaria Risk,” *PLOS Medicine*, vol. 3, no. 12, p. e473, Dec. 2006. [Online]. Available: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0030473>
- [4] M. A. Phillips, J. N. Burrows, C. Manyando, R. H. v. Huijsduijnen, W. C. V. Voorhis, and T. N. C. Wells, “Malaria,” *Nature Reviews Disease Primers*, vol. 3, p. 17050, Aug. 2017. [Online]. Available: <https://www.nature.com/articles/nrdp201750>
- [5] E. A. Ashley, A. P. Phyto, and C. J. Woodrow, “Malaria,” *The Lancet*, vol. 391, no. 10130, pp. 1608–1621, Apr. 2018. [Online]. Available: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(18\)30324-6/abstract](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)30324-6/abstract)
- [6] A. Trampuz, M. Jereb, I. Muzlovic, and R. M. Prabhu, “Clinical review: Severe malaria,” *Critical Care*, vol. 7, no. 4, pp. 315–323, 2003. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC270697/>
- [7] W. Chung, J. SJ, and Y. Lee, “A wireless sensor network compatible wearable u-healthcare monitoring system using integrated ECG, accelerometer and SpO2,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2008, pp. 1529–1532.

- [8] H. H. Asada, P. Shaltis, A. Reisner, and a. R. C. Hutchinson, "Mobile monitoring with wearable photoplethysmographic biosensors," *IEEE Engineering in Medicine and Biology Magazine*, vol. 22, no. 3, pp. 28–40, May 2003.
- [9] P. Krebs and D. T. Duncan, "Health App Use Among US Mobile Phone Owners: A National Survey," *JMIR mHealth and uHealth*, vol. 3, no. 4, Nov. 2015. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4704953/>
- [10] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, and S. Taheri, "Sleep Quality Prediction From Wearable Data Using Deep Learning," *JMIR mHealth and uHealth*, vol. 4, no. 4, Nov. 2016. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5116102/>
- [11] A. Hua, Z. Quicksall, C. Di, R. Motl, A. Z. LaCroix, B. Schatz, and D. M. Buchner, "Accelerometer-based predictive models of fall risk in older women: a pilot study," *npj Digital Medicine*, vol. 1, no. 1, p. 25, Jul. 2018. [Online]. Available: <https://www.nature.com/articles/s41746-018-0033-5>
- [12] "Toward Automated, At-Home Assessment of Mobility Among Patients With Parkinson Disease, Using a Body-Worn Accelerometer - Aner Weiss, Sarvi Sharifi, Meir Plotnik, Jeroen P. P. van Vugt, Nir Giladi, Jeffrey M. Hausdorff, 2011." [Online]. Available: <https://journals.sagepub.com/doi/full/10.1177/1545968311424869>
- [13] S. Del Din, A. Godfrey, and L. Rochester, "Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 3, pp. 838–847, May 2016. [Online]. Available: <http://ieeexplore.ieee.org/document/7078919/>
- [14] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian, and W. Grundfest, "Implementation of an iPhone for characterizing Parkinson's disease tremor through a wireless

- accelerometer application,” in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug. 2010, pp. 4954–4958.
- [15] “GWAS identifies 14 loci for device-measured physical activity and sleep duration | Nature Communications.” [Online]. Available: <https://www.nature.com/articles/s41467-018-07743-4>
- [16] H.-S. Yeo, J.-S. Hwang, W.-t. Han, K.-s. Shin, and Y.-H. Kim, “Electrode for measuring electrocardiogram and electrocardiogram device including the same,” US Patent US7672703B2, Mar., 2010. [Online]. Available: <https://patents.google.com/patent/US7672703B2/en>
- [17] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, “A dynamical model for generating synthetic electrocardiogram signals,” *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 289–294, Mar. 2003.
- [18] “Comprehensive multilevel in vivo and in vitro analysis of heart rate fluctuations in mice by ECG telemetry and electrophysiology | Nature Protocols.” [Online]. Available: <https://www.nature.com/articles/nprot.2015.139>
- [19] D. Franzen, J. M. Curtius, W. Heitz, H. W. Hpp, V. Diehl, and H. H. Hilger, “Cardiac involvement during and after malaria,” *The Clinical Investigator*, vol. 70, no. 8, pp. 670–673, Aug. 1992.
- [20] S. Bhat, J. Alva, K. Muralidhara, and S. Fahad, “Malaria and the heart,” *BMJ Case Reports*, vol. 2012, Nov. 2012. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4544263/>
- [21] H. Sulaiman, M. D. Ismail, M. Jalalonmuhali, N. Atiya, and S. Ponnampalavanar, “Severe Plasmodium falciparum infection mimicking acute myocardial infarction,” *Malaria Journal*, vol. 13, no. 1, p. 341, Aug. 2014. [Online]. Available: <https://doi.org/10.1186/1475-2875-13-341>

- [22] K. C. Nayak, S. L. Meena, B. K. Gupta, S. Kumar, and V. Pareek, “Cardiovascular involvement in severe vivax and falciparum malaria,” *Journal of Vector Borne Diseases*, vol. 50, no. 4, pp. 285–291, Dec. 2013.
- [23] J. E. Touze, J. Bernard, A. Keundjian, P. Imbert, A. Viguiier, H. Chaudet, and J. C. Doury, “Electrocardiographic Changes and Halofantrine Plasma Level During Acute Falciparum Malaria,” *The American Journal of Tropical Medicine and Hygiene*, vol. 54, no. 3, pp. 225–228, Mar. 1996. [Online]. Available: <https://www.ajtmh.org/content/journals/10.4269/ajtmh.1996.54.225>
- [24] L. Roggelin, D. Pelletier, J. N. Hill, T. Feldt, S. Hoffmann, D. Ansong, J. Sylverken, J. Burhenne, J. Fischer-Herr, P. Mehrfar, C. Thiel, G. D. Burchard, S. B. Nguah, and J. P. Cramer, “Disease-associated QT-shortage versus quinine associated QT-prolongation: age dependent ECG-effects in Ghanaian children with severe malaria,” *Malaria Journal*, vol. 13, no. 1, p. 219, Jun. 2014. [Online]. Available: <https://doi.org/10.1186/1475-2875-13-219>
- [25] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006.
- [26] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: <https://www.nature.com/articles/nature14539>
- [27] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond Short Snippets: Deep Networks for Video Classification,” 2015, pp. 4694–4702. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Ng_Beyond_Short_Snippets_2015_CVPR_paper.html
- [28] R. Collobert and J. Weston, “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” in *Proceedings of the*

- 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 160–167, event-place: Helsinki, Finland. [Online]. Available: <http://doi.acm.org/10.1145/1390156.1390177>
- [29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
- [30] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.
- [31] “Mastering the game of Go with deep neural networks and tree search | Nature.” [Online]. Available: <https://www.nature.com/articles/nature16961>
- [32] “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play | Science.” [Online]. Available: <http://science.sciencemag.org/content/362/6419/1140>
- [33] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: <https://www.nature.com/articles/nature21056>
- [34] E. Gawehn, J. A. Hiss, and G. Schneider, “Deep Learning in Drug Discovery,” *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201501008>
- [35] K. Polat and S. Gne, “Detection of ECG Arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine,” *Applied Mathematics and Computation*, vol. 186, no. 1, pp. 898–906,

- Mar. 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0096300306010125>
- [36] J. K. Lau, N. Lowres, L. Neubeck, D. B. Brieger, R. W. Sy, C. D. Galloway, D. E. Albert, and S. B. Freedman, “iPhone ECG application for community screening to detect silent atrial fibrillation: A novel technology to prevent stroke,” *International Journal of Cardiology*, vol. 165, no. 1, pp. 193–194, Apr. 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167527313002805>
- [37] A. Lyon, A. Bueno-Orovio, E. Zacur, R. Ariga, V. Grau, S. Neubauer, H. Watkins, B. Rodriguez, and A. Minchol, “Electrocardiogram phenotypes in hypertrophic cardiomyopathy caused by distinct mechanisms: apico-basal repolarization gradients vs. Purkinje-myocardial coupling abnormalities,” *EP Europace*, vol. 20, no. suppl.3, pp. iii102–iii112, Nov. 2018. [Online]. Available: <https://academic.oup.com/europace/article/20/suppl.3/iii102/5202143>
- [38] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Medicine*, vol. 25, no. 1, p. 65, Jan. 2019. [Online]. Available: <https://www.nature.com/articles/s41591-018-0268-3>
- [39] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger, S. J. Asirvatham, C. G. Scott, R. E. Carter, and P. A. Friedman, “Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram,” *Nature Medicine*, vol. 25, no. 1, p. 70, Jan. 2019. [Online]. Available: <https://www.nature.com/articles/s41591-018-0240-2>
- [40] M. Martinez-Bakker and B. Helm, “The influence of biological rhythms on host-parasite interactions,” *Trends in Ecology & Evolution*, vol. 30, no. 6, pp. 314–326,

- Jun. 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0169534715000713>
- [41] K. F. Prior, D. R. van der Veen, A. J. O'Donnell, K. Cumnock, D. Schneider, A. Pain, A. Subudhi, A. Ramaprasad, S. S. C. Rund, N. J. Savill, and S. E. Reece, "Timing of host feeding drives rhythms in parasite replication," *PLoS Pathogens*, vol. 14, no. 2, Feb. 2018. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5843352/>
- [42] A. J. O'Donnell, P. Schneider, H. G. McWatters, and S. E. Reece, "Fitness costs of disrupting circadian rhythms in malaria parasites," *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1717, pp. 2429–2436, Aug. 2011. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125626/>
- [43] J. Schlpfer and H. J. Wellens, "Computer-Interpreted Electrocardiograms: Benefits and Limitations," *Journal of the American College of Cardiology*, vol. 70, no. 9, pp. 1183–1192, Aug. 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0735109717387946>
- [44] A. P. Shah and S. A. Rubin, "Errors in the computerized electrocardiogram interpretation of cardiac rhythm," *Journal of Electrocardiology*, vol. 40, no. 5, pp. 385–390, Sep. 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0022073607000696>
- [45] M. E. Guglin and D. Thatai, "Common errors in computer electrocardiogram interpretation," *International Journal of Cardiology*, vol. 106, no. 2, pp. 232–237, Jan. 2006. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167527305004043>

- [46] K. Poon, P. M. Okin, and P. Kligfield, “Diagnostic performance of a computer-based ECG rhythm algorithm,” *Journal of Electrocardiology*, vol. 38, no. 3, pp. 235–238, Jul. 2005. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0022073605000488>
- [47] B. T. Wilhelm and J.-R. Landry, “RNA-Sequantitative measurement of expression through massively parallel RNA-sequencing,” *Methods*, vol. 48, no. 3, pp. 249–257, Jul. 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1046202309000632>
- [48] “Genome-wide genetic marker discovery and genotyping using next-generation sequencing | Nature Reviews Genetics.” [Online]. Available: <https://www.nature.com/articles/nrg3012>
- [49] V. I. Mwangi, “_USAntimalarial activity of methylene blue against selected drug resistant Plasmodium parasites,” Thesis, COHES - JKUAT, Jan. 2018. [Online]. Available: <http://localhost:8080/xmlui/handle/123456789/3535>
- [50] “Plasmodium knowlesi: Reservoir Hosts and Tracking the Emergence in Humans and Macaques.” [Online]. Available: <https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002015>
- [51] B. Singh, L. K. Sung, A. Matusop, A. Radhakrishnan, S. S. Shamsul, J. Cox-Singh, A. Thomas, and D. J. Conway, “A large focus of naturally acquired Plasmodium knowlesi infections in human beings,” *The Lancet*, vol. 363, no. 9414, pp. 1017–1024, Mar. 2004. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0140673604158364>
- [52] C. Aurrecochea, J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, and O. S. Harb, “PlasmoDB: a functional genomic database for malaria parasites,” *Nucleic acids research*, vol. 37, no. suppl_1, pp. D539–D543, 2008.

- [53] Y. Tang, C. J. Joyner, M. Cabrera-Mora, C. L. Saney, S. A. Lapp, M. V. Nural, S. B. Pakala, J. D. DeBarry, S. Soderberg, J. C. Kissinger, T. J. Lamb, M. R. Galinski, M. P. Styczynski, and the MaHPIC Consortium, “Correction to: Integrative analysis associates monocytes with insufficient erythropoiesis during acute *Plasmodium cynomolgi* malaria in rhesus macaques,” *Malaria Journal*, vol. 16, no. 1, p. 486, Dec. 2017. [Online]. Available: <https://doi.org/10.1186/s12936-017-2134-z>
- [54] K. Chen, Z. Yang, L. Hoang, J. Weiland, M. Humayun, and W. Liu, “An Integrated 256-Channel Epiretinal Prosthesis,” *IEEE Journal of Solid-State Circuits*, vol. 45, no. 9, pp. 1946–1956, Sep. 2010.
- [55] M. Lutz, *Learning Python: Powerful Object-Oriented Programming*. ”O’Reilly Media, Inc.”, Jun. 2013, google-Books-ID: ePyeNz2Eoy8C.
- [56] E. D. Trippe, J. B. Aguilar, Y. H. Yan, M. V. Nural, J. A. Brady, and J. B. Gutierrez, “Introducing Data Primitives: Data Formats for the SKED Framework,” *arXiv:1706.08131 [q-bio]*, Jun. 2017, arXiv: 1706.08131. [Online]. Available: <http://arxiv.org/abs/1706.08131>
- [57] G. L. Curtis, M. Chughtai, A. Khlopa, J. M. Newman, R. Khan, S. Shaffiy, A. Nadhim, A. Bhave, and M. A. Mont, “Impact of Physical Activity in Cardiovascular and Musculoskeletal Health: Can Motion Be Medicine?” *Journal of Clinical Medicine Research*, vol. 9, no. 5, pp. 375–381, May 2017. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5380169/>
- [58] M. Nakagawa, T. Iwao, S. Ishida, H. Yonemochi, T. Fujino, T. Saikawa, and M. Ito, “Circadian rhythm of the signal averaged electrocardiogram and its relation to heart rate variability in healthy subjects,” *Heart*, vol. 79, no. 5, pp. 493–496, May 1998. [Online]. Available: <https://heart.bmj.com/content/79/5/493>

- [59] “Circadian Variations in Atrial Fibrillatory Frequency in Persistent Human Atrial Fibrillation - BOLLMANN - 2000 - Pacing and Clinical Electrophysiology - Wiley Online Library.” [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-8159.2000.tb07040.x>
- [60] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, Jan. 2002. [Online]. Available: <https://content.iospress.com/articles/intelligent-data-analysis/ida00103>
- [61] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. Cardoso, “Preprocessing Techniques for Context Recognition from Accelerometer Data,” *Personal Ubiquitous Comput.*, vol. 14, no. 7, pp. 645–662, Oct. 2010. [Online]. Available: <http://dx.doi.org/10.1007/s00779-010-0293-9>
- [62] G. R. Norman, J. A. Sloan, and K. W. Wyrwich, “Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation,” *Medical Care*, vol. 41, no. 5, pp. 582–592, 2003. [Online]. Available: <https://www.jstor.org/stable/3768017>
- [63] R. Longadge and S. Dongre, “Class Imbalance Problem in Data Mining Review,” *arXiv:1305.1707 [cs]*, May 2013, arXiv: 1305.1707. [Online]. Available: <http://arxiv.org/abs/1305.1707>
- [64] “The Problem of Overfitting - Journal of Chemical Information and Modeling (ACS Publications).” [Online]. Available: <https://pubs.acs.org/doi/full/10.1021/ci0342472>
- [65] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: A system for large-scale machine learning,” in *12th USENIX Symposium on Operating Systems Design*

- and Implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>
- [66] F. Chollet and others, *Keras*, 2015. [Online]. Available: <https://keras.io>
- [67] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” p. 30.
- [68] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” *arXiv:1502.01852 [cs]*, Feb. 2015, arXiv: 1502.01852. [Online]. Available: <http://arxiv.org/abs/1502.01852>
- [69] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
- [70] A. Krogh and J. Vedelsby, “Neural network ensembles, cross validation, and active learning,” in *Advances in neural information processing systems*, 1995, pp. 231–238.
- [71] D. J. Hand and R. J. Till, “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems,” *Machine Learning*, vol. 45, no. 2, pp. 171–186, Nov. 2001. [Online]. Available: <https://doi.org/10.1023/A:1010920819831>
- [72] G. B. Moody, “ECG-based indices of physical activity,” in *Proceedings Computers in Cardiology*, Oct. 1992, pp. 403–406.
- [73] T. Penzel, J. W. Kantelhardt, R. P. Bartsch, M. Riedl, J. F. Kraemer, N. Wessel, C. Garcia, M. Glos, I. Fietze, and C. Schbel, “Modulations of Heart Rate, ECG, and Cardio-Respiratory Coupling Observed in Polysomnography,” *Frontiers in Physiology*, vol. 7, 2016. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fphys.2016.00460/full>

- [74] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970. [Online]. Available: <https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634>
- [75] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” *Journal of Biomedical Informatics*, vol. 35, no. 5, pp. 352–359, Oct. 2002. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1532046403000340>
- [76] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. 10, p. R106, Oct. 2010. [Online]. Available: <https://doi.org/10.1186/gb-2010-11-10-r106>
- [77] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
- [78] C. A. McClung, “Circadian genes, rhythms and the biology of mood disorders,” *Pharmacology & Therapeutics*, vol. 114, no. 2, pp. 222–232, May 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0163725807000368>
- [79] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, “A circadian gene expression atlas in mammals: implications for biology and medicine,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 16 219–16 224, 2014.
- [80] A. C. Cameron and P. K. Trivedi, “Regression Analysis of Count Data by A. Colin Cameron,” Sep. 1998. [Online]. Available: [/core/books/regression-analysis-of-count-data/4E25812F8271C66442FD9EE0FC6EEA57](#)
- [81] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

- [82] J. K. Kim and D. B. Forger, “A mechanism for robust circadian timekeeping via stoichiometric balance,” *Molecular Systems Biology*, vol. 8, p. 630, Dec. 2012. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3542529/>
- [83] N. A. O’Leary, M. W. Wright, J. R. Brister, S. Ciufu, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O’Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–745, Jan. 2016.
- [84] S. Hnzelmann, R. Castelo, and J. Guinney, “GSVA: gene set variation analysis for microarray and RNA-Seq data,” *BMC Bioinformatics*, vol. 14, no. 1, p. 7, Jan. 2013. [Online]. Available: <https://doi.org/10.1186/1471-2105-14-7>
- [85] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdttir, P. Tamayo, and J. P. Mesirov, “Molecular signatures database (MSigDB) 3.0,” *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.
- [86] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” 2011.
- [87] N. Mideo, S. E. Reece, A. L. Smith, and C. J. E. Metcalf, “The Cinderella syndrome: why do malaria-infected cells burst at midnight?” *Trends in Parasitology*, vol. 29, no. 1, pp. 10–16, Jan. 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1471492212001845>

- [88] K. Bertsch, D. Hagemann, E. Naumann, H. Schchinger, and A. Schulz, “Stability of heart rate variability indices reflecting parasympathetic activity,” *Psychophysiology*, vol. 49, no. 5, pp. 672–682, 2012. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.2011.01341.x>
- [89] L. Lger and M. Thivierge, “Heart Rate Monitors: Validity, Stability, and Functionality,” *The Physician and Sportsmedicine*, vol. 16, no. 5, pp. 143–151, May 1988. [Online]. Available: <https://doi.org/10.1080/00913847.1988.11709511>
- [90] K. F. Janz, “Validation of the CSA accelerometer for assessing children’s physical activity,” *Medicine & Science in Sports & Exercise*, vol. 26, no. 3, pp. 369–375, 1994.
- [91] D. Jeyaraj, S. M. Haldar, X. Wan, M. D. McCauley, J. A. Ripperger, K. Hu, Y. Lu, B. L. Eapen, N. Sharma, E. Ficker, M. J. Cutler, J. Gulick, A. Sanbe, J. Robbins, S. Demolombe, R. V. Kondratov, S. A. Shea, U. Albrecht, X. H. Wehrens, D. S. Rosenbaum, and M. K. Jain, “Circadian rhythms govern cardiac repolarization and arrhythmogenesis,” *Nature*, vol. 483, no. 7387, pp. 96–99, Feb. 2012. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3297978/>
- [92] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig, “Gene ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, p. 25, 2000.
- [93] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D’Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, and L. Matthews, “Reactome: a knowledgebase of biological pathways,” *Nucleic acids research*, vol. 33, no. suppl_1, pp. D428–D432, 2005.
- [94] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

- [95] N. Gekakis, D. Staknis, H. B. Nguyen, F. C. Davis, L. D. Wilsbacher, D. P. King, J. S. Takahashi, and C. J. Weitz, “Role of the CLOCK protein in the mammalian circadian mechanism,” *Science (New York, N.Y.)*, vol. 280, no. 5369, pp. 1564–1569, Jun. 1998.
- [96] A. J. Mitchell, A. M. Hansen, L. Hee, H. J. Ball, S. M. Potter, J. C. Walker, and N. H. Hunt, “Early cytokine production is associated with protection from murine cerebral malaria,” *Infection and immunity*, vol. 73, no. 9, pp. 5645–5653, 2005.
- [97] I. A. Clark and K. A. Rockett, “The cytokine theory of human cerebral malaria,” *Parasitology Today*, vol. 10, no. 10, pp. 410–412, 1994.
- [98] A. Lyon, A. Minchol, J. P. Martnez, P. Laguna, and B. Rodriguez, “Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances,” *Journal of the Royal Society, Interface*, vol. 15, no. 138, 2018.
- [99] J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, “A wavelet-based ECG delineator: evaluation on standard databases,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 570–581, Apr. 2004.
- [100] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Ieee, 2010, pp. 1–10.
- [101] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, and A. Ghodsi, “Spark sql: Relational data processing in spark,” in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. ACM, 2015, pp. 1383–1394.
- [102] “Techniques for the diagnosis of malaria. - PubMed - NCBI.” [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/3041624>

- [103] “Automated image processing method for the diagnosis and classification of malaria on thin blood smears | SpringerLink.” [Online]. Available: <https://link.springer.com/article/10.1007/s11517-006-0044-2>

APPENDIX

A CONFUSION MATRICES

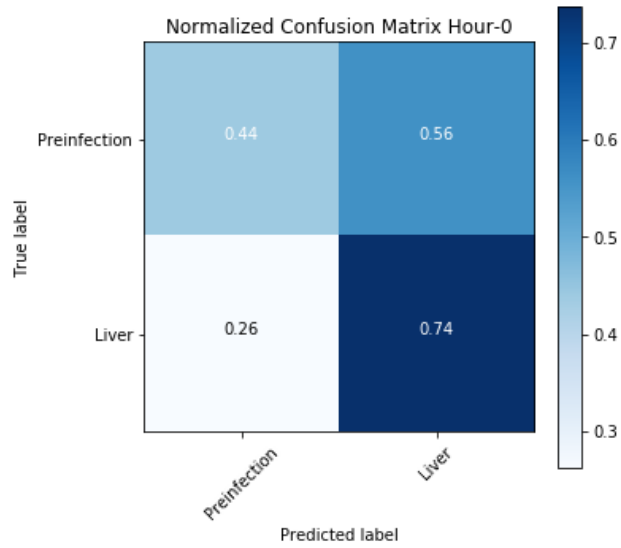


Figure A.1: **The Normalized Confusion Matrix of Hour 0**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

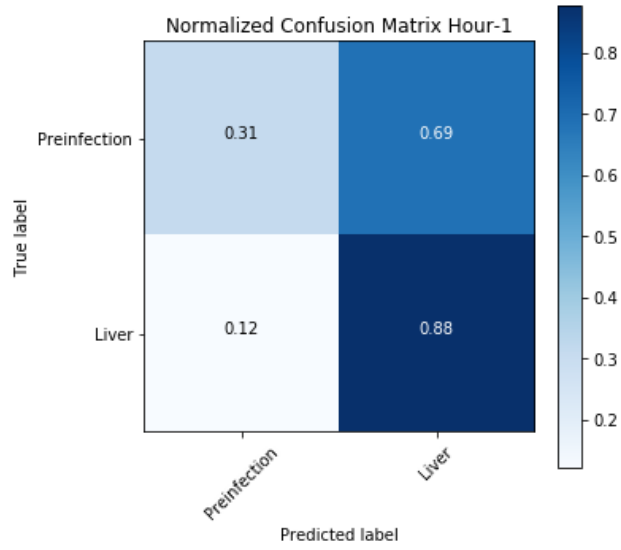


Figure A.2: **The Normalized Confusion Matrix of Hour 1**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

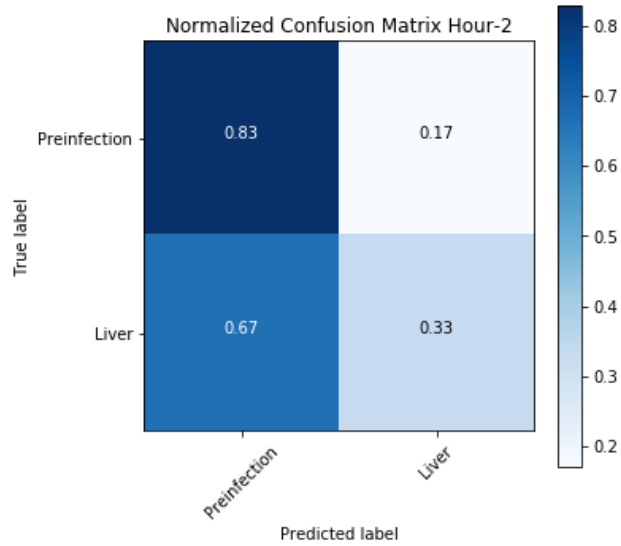


Figure A.3: **The Normalized Confusion Matrix of Hour 2**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

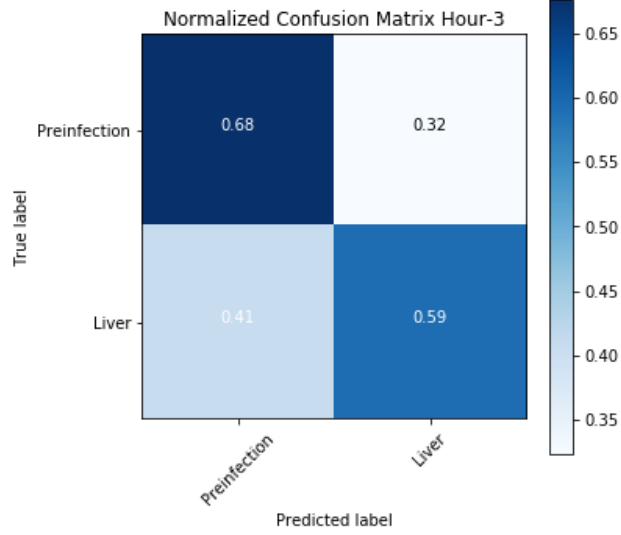


Figure A.4: **The Normalized Confusion Matrix of Hour 3**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

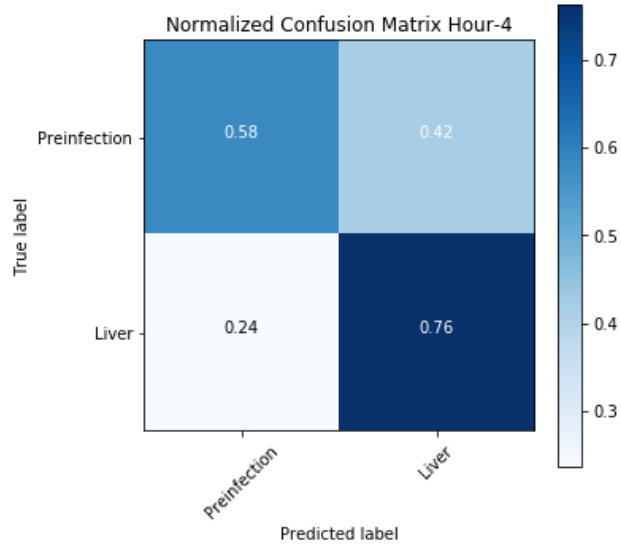


Figure A.5: **The Normalized Confusion Matrix of Hour 4**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

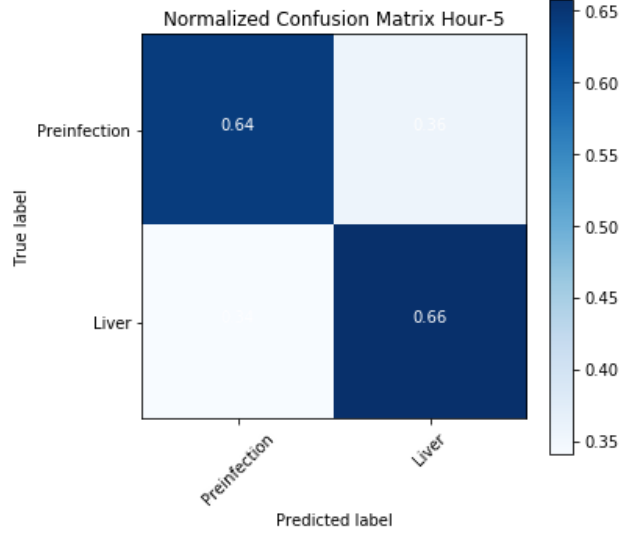


Figure A.6: **The Normalized Confusion Matrix of Hour 5**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

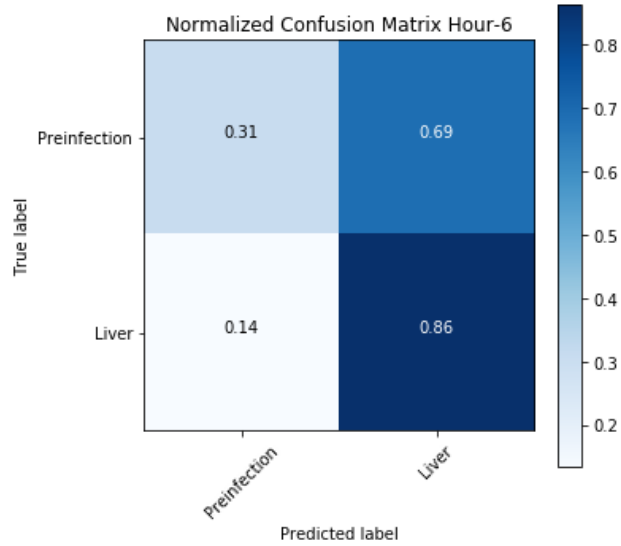


Figure A.7: **The Normalized Confusion Matrix of Hour 6**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

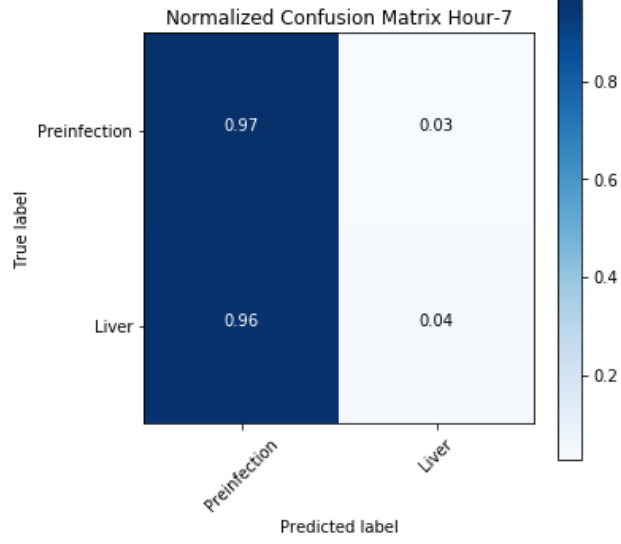


Figure A.8: **The Normalized Confusion Matrix of Hour 7**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

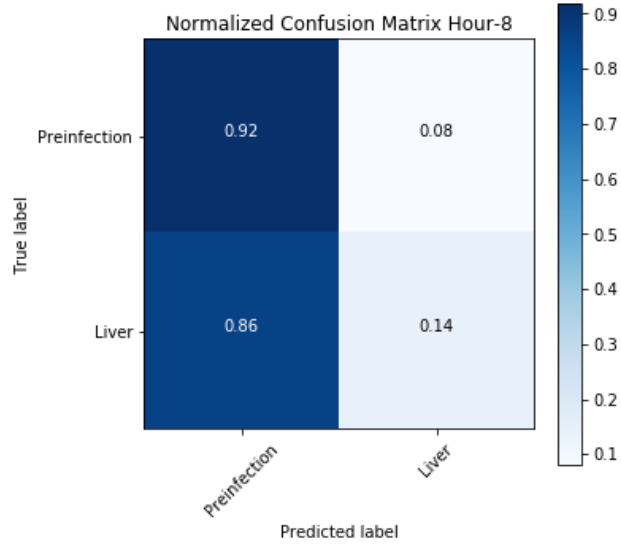


Figure A.9: **The Normalized Confusion Matrix of Hour 8**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

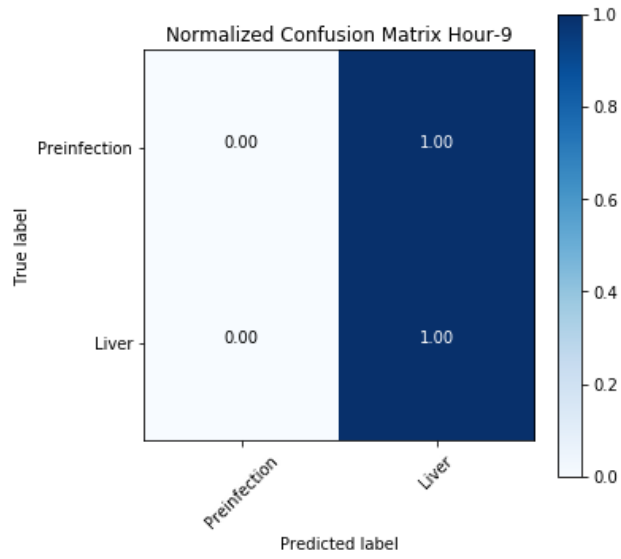


Figure A.10: **The Normalized Confusion Matrix of Hour 9**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

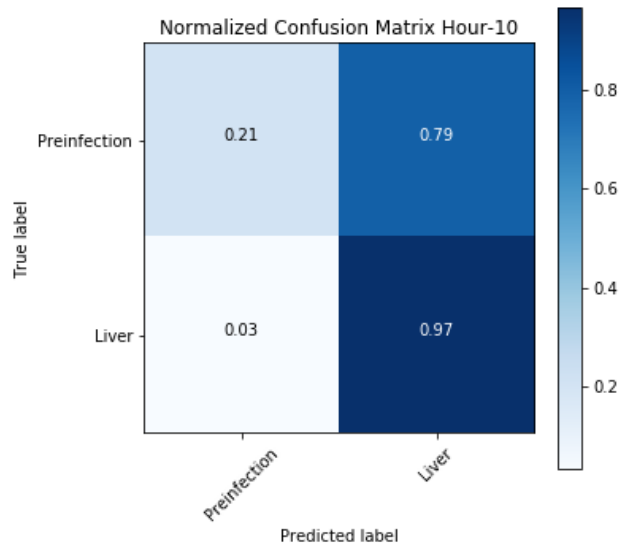


Figure A.11: **The Normalized Confusion Matrix of Hour 10**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

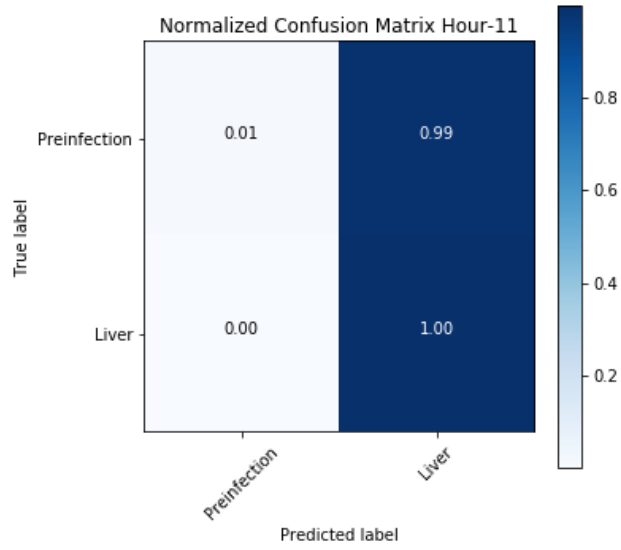


Figure A.12: **The Normalized Confusion Matrix of Hour 11**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

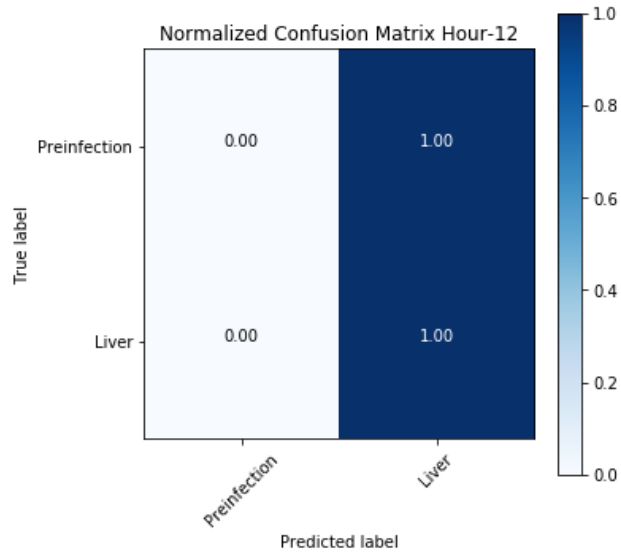


Figure A.13: **The Normalized Confusion Matrix of Hour 12**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

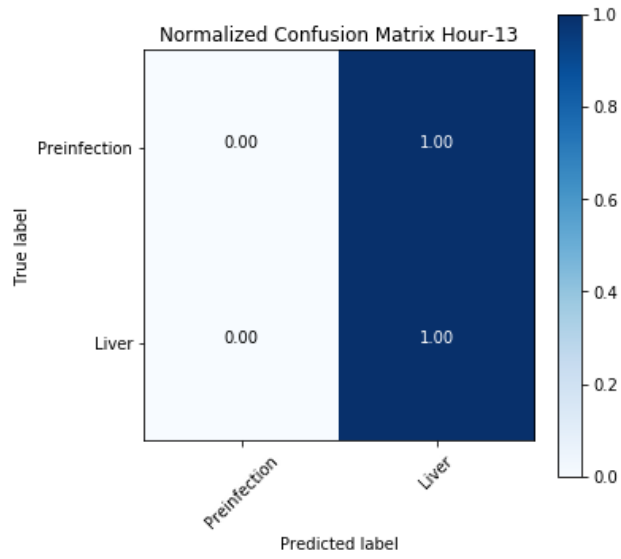


Figure A.14: **The Normalized Confusion Matrix of Hour 13**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

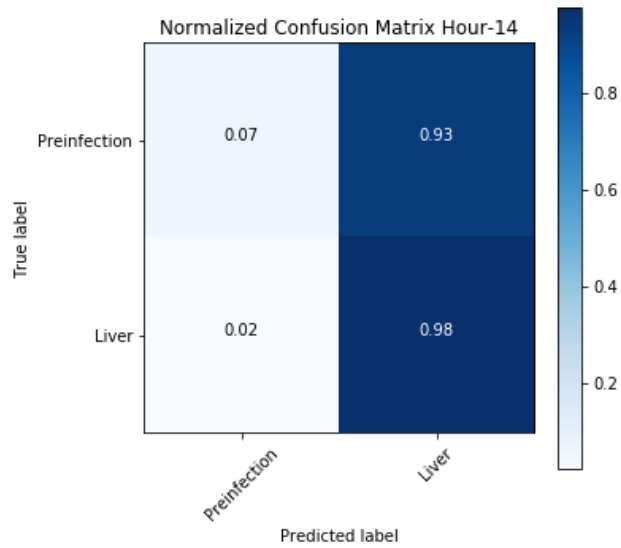


Figure A.15: **The Normalized Confusion Matrix of Hour 14**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

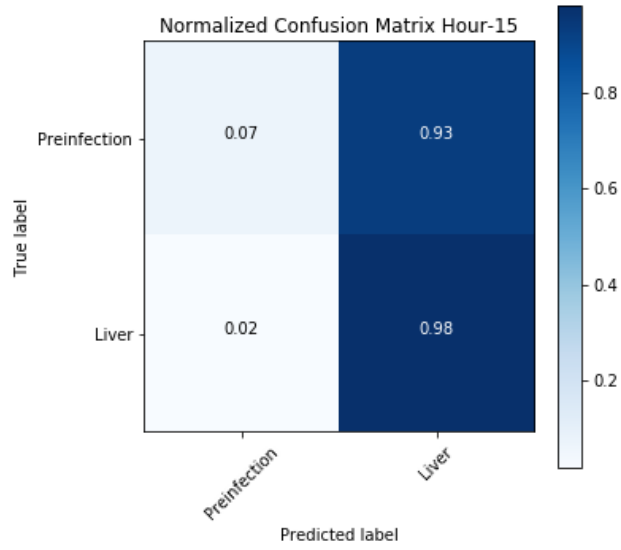


Figure A.16: **The Normalized Confusion Matrix of Hour 15**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

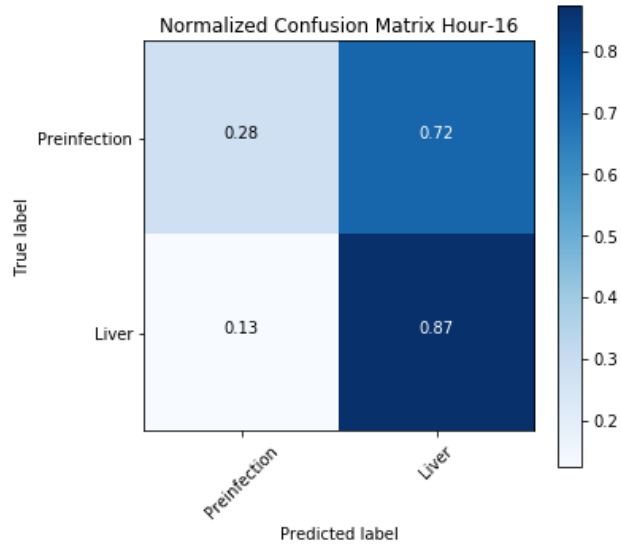


Figure A.17: **The Normalized Confusion Matrix of Hour 16**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

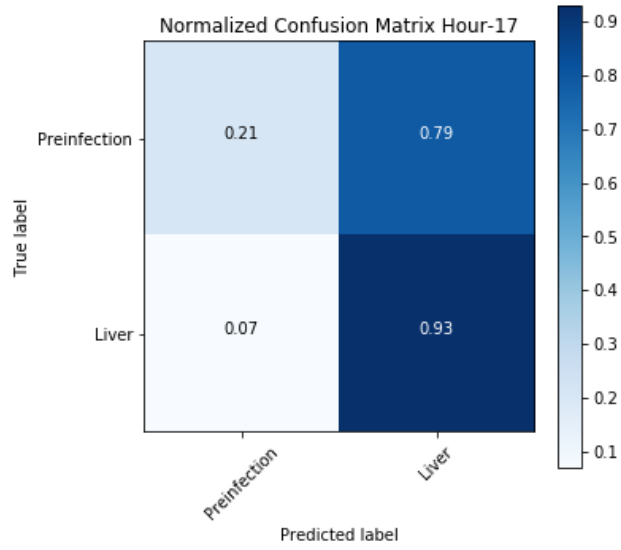


Figure A.18: **The Normalized Confusion Matrix of Hour 17**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

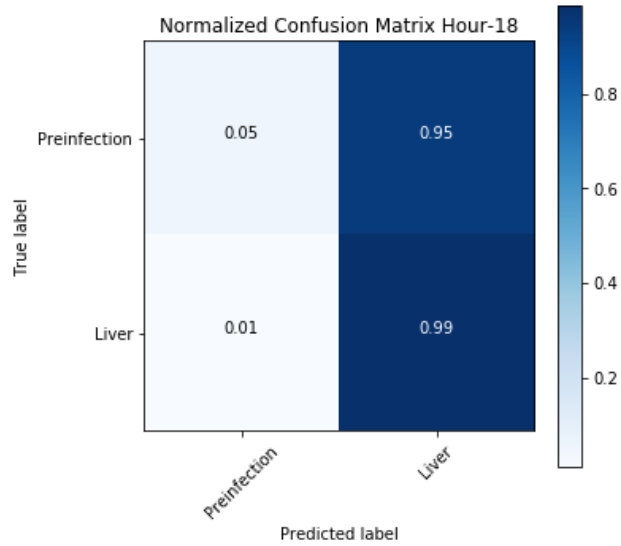


Figure A.19: **The Normalized Confusion Matrix of Hour 18**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

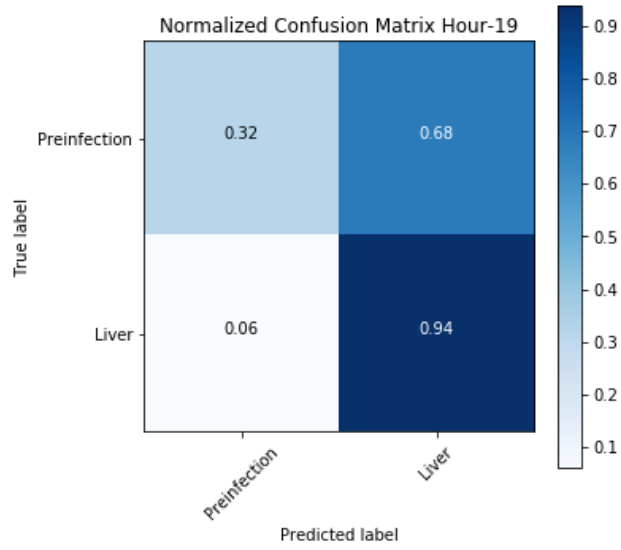


Figure A.20: **The Normalized Confusion Matrix of Hour 19**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

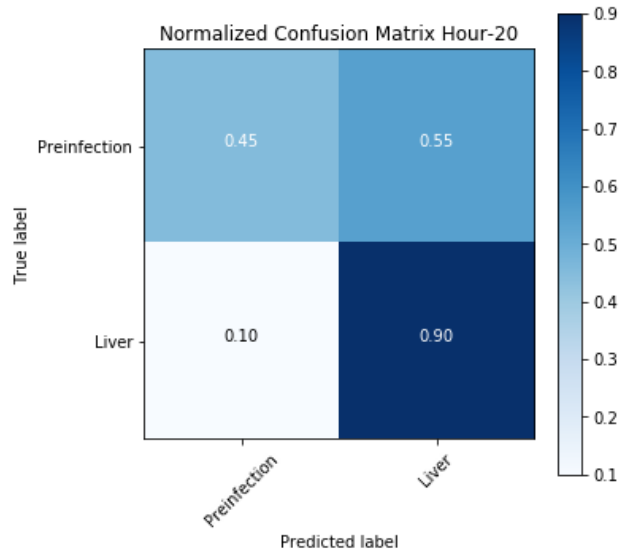


Figure A.21: **The Normalized Confusion Matrix of Hour 20**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

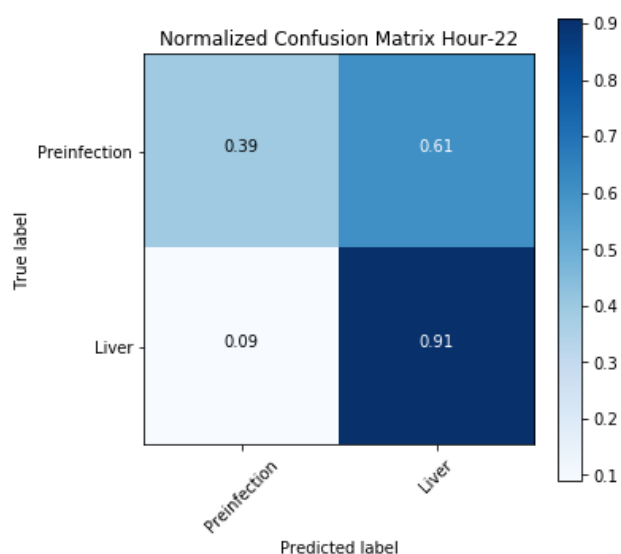


Figure A.22: **The Normalized Confusion Matrix of Hour 22**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

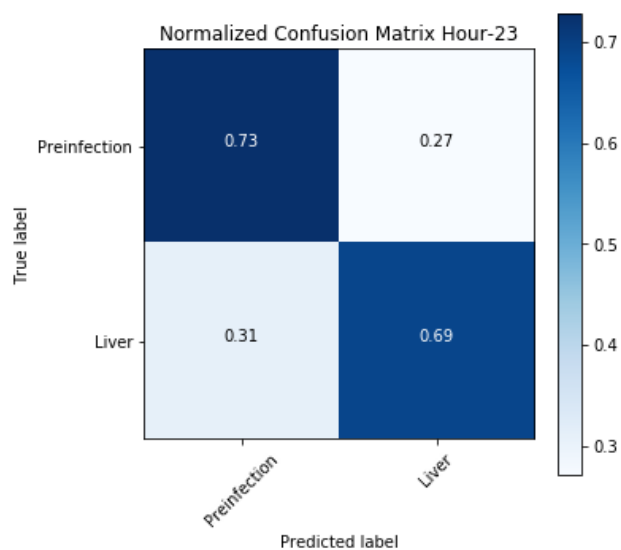


Figure A.23: **The Normalized Confusion Matrix of Hour 23**

X-axis is the DNN model predicted ECG label of testing dataset; Y-axis the true label of the testing dataset. The right color bar denotes the prediction accuracy range.

B PYTHON MODULE FOR DATA ACCESS

```
1 import cx_Oracle
2 import pandas as pd
3
4 ### Author: Tao sheng
5 ### Date: 2018-09-10
6
7 class MahpicDB:
8     """
9     class to initiate connection to Mahpic database
10    """
11    def __init__(self):
12        self.cnx = cx_Oracle.connect('myuser', 'mypasswd',
13        'jknfs1.gacrc.uga.edu:1521/mahpicdb1.mic.uga.edu')
14        self.cursor = self.cnx.cursor()
15
16    def getMetadata(self, exp):
17        """
18        Get Experimental metadata data by providing experimental ID
19        Such as: E07A, E07B, E06, E30 etc...
20        """
21        sql = "begin sked.get_protocols(experiment_id => '" +
22        str(exp) + "');end;"
23        self.cursor.execute(sql)
24        for ix, result in enumerate(self.cursor.getimplicitresults()):
25            res_pro = result.fetchall()
26            res_pro = pd.DataFrame(res_pro)
27        return res_pro
```

Listing B.1: Python module for query and downloading data from SKED database – Part A

```

1  def getExperimentID(self, exp):
2      """
3      Get Experimental Information by providing experiment ID
4      Such as: E07A, E07B, E06, E30 etc...
5      """
6      sql = "begin sked.get_experiments(experiment_id => '"
7      + str(exp) + "');end;"
8      self.cursor.execute(sql)
9      for ix, result in enumerate(self.cursor.getimplicitresults()):
10         res_exp = result.fetchall()
11         res_exp = pd.DataFrame(res_exp)
12         return res_exp
13
14  def getSubjectIDs(self, exp):
15      """
16      Get the subject IDs that are associated with the experiment ID
17      Such as: E07A, E07B, E06, E30 etc...
18      """
19      sql = "begin sked.get_subjects(experiment_id => '"
20      + str(exp) + "');end;"
21      self.cursor.execute(sql)
22      for ix, result in enumerate(self.cursor.getimplicitresults()):
23         res_sub = result.fetchall()
24         res_exp = pd.DataFrame(res_sub)
25         return res_sub
26
27
28
29  def getMetadatawithIDs(self, exp, id):
30      """
31      Get the subject IDs that are associated with the experiment ID
32      Such as: E07A, E07B, E06, E30 etc...
33      """
34      sql = "begin sked.get_subjects(subject_id =>'" +
35      str(id) + "', experiment_id => '" + str(exp) + "');end;"
36      self.cursor.execute(sql)
37      for ix, result in enumerate(self.cursor.getimplicitresults()):
38         res_met = result.fetchall()
39         if res_met == []:
40             return 'No Data'
41         else:
42             res_met = pd.DataFrame(res_met)
43             cols = pd.DataFrame(result.description)[0]
44             res_met.columns = cols
45             return res_met

```

Listing B.2: Python module for query and downloading data from SKED database – Part B

```

1  def getProtocolswithID(self, exp, id):
2      """
3      Get the protocol IDs that are associated with the experiment ID
4      Such as: E07A, E07B, E06, E30 etc...
5      """
6      sql = "begin sked.get_protocols(subject_id =>'"
7      + str(id) + "', experiment_id => '" + str(exp) + "');end;"
8      self.cursor.execute(sql)
9      for ix, result in enumerate(self.cursor.getimplicitresults()):
10         res_met = result.fetchall()
11         if res_met == []:
12             return 'No Data'
13         else:
14             res_met = pd.DataFrame(res_met)
15             cols = pd.DataFrame(result.description)[0]
16             res_met.columns = cols
17             return res_met
18
19
20  ### Time Series Range
21  def getTimeSeriesRange(self, exp, id):
22      """
23      Get the time series ranges and the protocol IDs
24      that are associated with the experiment ID
25      Such as: E07A, E07B, E06, E30 etc...
26      Usually try twice if failed at the first time.
27      """
28      sql = "begin sked.get_protocols(fields =>
29      'protocol_app_id', experiment_id => '" + str(exp)
30      ) + "', subject_id => '" + str(id) + "', data_type
31      => 'Telemetry');end;"
32      self.cursor.execute(sql)
33      for ix, result in enumerate(self.cursor.getimplicitresults()):
34         res_met = result.fetchall()
35         if res_met == []:
36             return 'No Data'
37         else:
38             sSQLs = "BEGIN sked.get_time_series_range(protocol_app_id => '"
39             + str(res_met[0][0]) + "', subject_id => '" + str(id) + "');end;"
40             self.cursor.execute(sSQLs)
41             for ix, result in enumerate(self.cursor.getimplicitresults()):
42                 time_range = result.fetchall()
43                 res_met.append(time_range)
44             return res_met

```

Listing B.3: Python module for query and downloading data from SKED database – Part C

```

1  ### Real Time Series
2  def getTSDData(self, exp, id, prot, minD, maxD):
3      """
4      Get the real time series data by providing the followings:
5      experimental id: E07A, E07B, E06, E30 etc...
6      subject id; protocol id; minimum date and maximum date with
7      the format of '22-NOV-16'
8      """
9      sql = "BEGIN sked.get_time_series(protocol_app_id => '"
10     + str(prot) + "', subject_id => '" + str(id
11     ) + "', min_date => '" + str(minD) + "', max_date => '"
12     + str(maxD) + "', transpose => '1');end;"
13     self.cursor.execute(sql)
14     for ix, result in enumerate(self.cursor.getimplicitresults()):
15         res_met = result.fetchall()
16     if res_met == []:
17         return 'No Data'
18     else:
19         res_met = pd.DataFrame(res_met)
20         cols = pd.DataFrame(result.description)[0]
21         res_met.columns = cols
22         return res_met

```

Listing B.4: Python module for query and downloading data from SKED database – Part D

```

1  ### Define F1 metric
2  def f1(y_true, y_pred):
3      def recall(y_true, y_pred):
4          """Recall metric.
5             Only computes a batch-wise average of recall.
6             Computes the recall, a metric for multi-label
7             classification of how many relevant items are selected.
8             """
9          true_positives = K.sum(K.round(K.clip(y_true * y_pred,
10             0, 1)))
11         possible_positives = K.sum(K.round(K.clip(y_true, 0, 1)))
12         recall = true_positives / (possible_positives + K.epsilon())
13         return recall
14
15     def precision(y_true, y_pred):
16         """Precision metric.
17            Only computes a batch-wise average of precision.
18            Computes the precision, a metric for multi-label
19            classification of how many selected items are relevant.
20            """
21         true_positives = K.sum(K.round(K.clip(y_true * y_pred, 0, 1)))
22         predicted_positives = K.sum(K.round(K.clip(y_pred, 0, 1)))
23         precision = true_positives / (predicted_positives + K.epsilon())
24         return precision
25     precision = precision(y_true, y_pred)
26     recall = recall(y_true, y_pred)
27     return 2*((precision*recall)/(precision+recall+K.epsilon()))

```

Listing C.5: Define F1 score in the DNN architectures

```

1  import keras
2  from keras.models import Sequential
3  from keras.layers import Dense, Dropout, Activation
4  from keras.optimizers import SGD
5  from keras import backend as K
6
7  ### Keras Sequential Models
8  model = Sequential()
9  model.add(Conv1D(128, 3, input_shape=(10001, 128), activation='relu'))
10 model.add(Dense(128, activation='relu'))
11 model.add(Dropout(0.5))
12 model.add(Dense(64, activation='relu'))
13 model.add(Dropout(0.5))
14 model.add(Dense(32, activation='relu'))
15 model.add(Dropout(0.5))
16 model.add(Dense(16, activation='relu'))
17 model.add(Dropout(0.5))
18 model.add(Dense(2, activation='sigmoid'))
19 ### Compilation
20 # For a binary classification problem
21 model.compile(optimizer='rmsprop',
22               loss='binary_crossentropy',
23               metrics=['accuracy', f1])
24
25 model.summary()

```

Listing C.6: Development of the DNN architectures

```

1  import keras
2  from keras.models import Sequential
3  from keras.layers import Dense, Dropout, Activation
4  from keras.optimizers import SGD
5  from keras import backend as K
6  import sklearn.metrics as skm
7  import tensorflow as tf
8  from sklearn.model_selection import train_test_split
9  from sklearn.preprocessing import normalize
10 from sklearn.metrics import confusion_matrix
11 import numpy as np
12 import pandas as pd
13 import matplotlib.pyplot as plt
14
15 ## Run the model
16 results = list()
17 sample_size = list()
18 fprs = list()
19 tprs = list()
20 aucs = list()
21 for k in range(0, 24):
22     fx_n = '/mnt/data0/tao/ECG/hdf5/ECG_H' + str(k) + '_Dataset.h5'
23     fy_n = '/mnt/data0/tao/ECG/hdf5/ECG_H' + str(k) + '_Label.h5'
24     x_n = 'ECG_H' + str(k)
25     with h5py.File(fx_n, 'r') as hf:
26         X_H = hf["x_n"][:]
27     with h5py.File(fy_n, 'r') as hf:
28         Y_H = hf["x_n"][:]
29     x_train, x_test, y_train, y_test = train_test_split(X_H, Y_H, test_size=0.1)
30     y_train = keras.utils.to_categorical(y_train, num_classes=2)
31     y_test = keras.utils.to_categorical(y_test, num_classes=2)
32     history = model.fit(x_train, y_train, epochs=20, batch_size=128,
33                         validation_data=(x_test, y_test))
34     preds = model.predict(x_test, verbose=1)
35     gt = np.argmax(y_test, axis=1)
36     binary_gt = gt == 0
37     binary_probs = preds[..., 0]
38     fpr, tpr, _ = skm.roc_curve(binary_gt.ravel(), binary_probs.ravel())
39     auc = skm.auc(fpr, tpr)
40     results.append([k, max(history.history['acc']), max(history.history['f1']),
41                    max(history.history['val_acc']), max(history.history['val_f1'])])
42     sample_size.append([len(x_train), len(x_test)])
43     fprs.append(fpr)
44     tprs.append(tpr)
45     aucs.append(auc)

```

Listing C.7: Feeding the DNN model of ECG in 24-hours independently