博 士 論 文

Establishment of a diagnostic method for premenstrual dysphoric disorder using machine learning algorithms and hippocampal factors in rats

（機械学習を用いたラット海馬での発現遺伝子を指標とする

月経前不快気分障害の診断法の確立）

ちょう　ゆうぇい
張　育薇
CHANG　YU-WEI

Contents

General Introduction

Introduction to premenstrual dysphoric disorder

Premenstrual dysphoric disorder (PMDD) is a disease with emotional and physical symptoms during the luteal phase in women of childbearing age. It is a newly defined disease which had not been put into *Diagnostic and Statistical Manual of Mental Disorder 5th Edition* (DSM-5) until 2013 (American Psychiatric Association, 2013). Before this update, the terms of premenstrual tension (PMT) and premenstrual syndrome (PMS) were included in the WHO International Classification of Diseases (ICD) but with no specific diagnostic definition. Currently, while PMDD is the term used for the patients with severe symptoms meeting the diagnostic criteria in DSM-5, PMS is for the ones with rather mild symptoms (Yonkers, O'Brien, & Eriksson, 2008).

Epidemiology

It is suggested that there are up to 80% of women having one or more symptoms of PMDD, and the prevalence of PMDD is around 1.8-5.8% (American Psychiatric Association, 2013; Biggs and Demuth, 2011). Due to the high number of people affected and the property of recurring monthly, PMDD is a disease influencing not only the patients themselves but also the world economic system. In previous studies, the severity of the symptoms was demonstrated to have a relationship with productivity impairment and absenteeism. The patients had intention to reduce work hours, and were prone to be late or leave early (Hardy & Hunter, 2021; Heinemann, Minh, Filonenko, & Uhl-Hochgräber, 2010). Although the exact figure of economic loss brought by PMDD is not yet clear, it is a disease deserving further attention.

Pathophysiology

The exact pathological mechanism of PMDD remains controversial. However, there are three main suggested hypotheses. The first and the most general one is that the chronic exposure followed by rapid withdrawal of the progesterone (P4) metabolites, allopregnanolone (ALLO), in the luteal phase can lead to insensitive gamma-aminobutyric acid A receptors (GABAR), and induce anxiogenic effect. ALLO is a positive modulator of GABAR. It induces sedative effect but can also have paradoxical adverse effect to cause symptoms of PMDD (Bixo, Johansson, Timby, Michalski, & Bäckström, 2018). In more detail, previous studies showed that the α4, α6, δ subunits of GABAR are sensitive to ALLO, and GABAR with the α5 subunits play important roles in hippocampus mediating recognition, learning, and memory. The change of these subunits under stress situation may be the reason for PMDD. However, whether the stress situation is able to cause direct change of the subunits, or whether these changes are brought out by ALLO remains equivocal due to no direct evidence (Locci & Pinna, 2017). Moreover, there are still patients with PMDD symptoms onsetting from the early luteal phase or even from the ovulation period, which is the period with increasing ALLO concentration but not withdrawal. Thus, this theory cannot explain all the situations in the patients, and still needs further modification (Hantsoo & Epperson, 2020).

The second hypothesis is that PMDD is related to the fluctuation of inflammatory proteins during the menstrual cycle. Chronic inflammation has been suggested in the etiologies of depression and other disorders with symptoms similar to PMDD, and thus it was also considered to be a reason for symptoms of PMDD (Brennan *et al.,* 2009; Miller, Maletic, and Raison, 2010). This theory is supported by some studies. In a

previous work, PMS patients showed elevated high-sensitivity C-reactive protein (hs-CRP) compared to control groups (Gold, Wells, & Rasor, 2016). In another study, interleukin-4 (IL-4), IL-10, IL-12, and interferon-γ (IFN-γ) were significantly higher in the PMDD patients. However, due to the cross-sectional property of the study, the temporal relationship between inflammatory factors and the symptoms of PMDD remains unclear (Ronnenberg *et al.*, 2014).

The third hypothesis is simply the combination of hypothesis one and two (Ziomkiewicz-wichary, 2017). There is a possibility that PMDD is a disease caused by the coaction of dysfunctional signaling pathway between ALLO and GABAR, and the fluctuation of immune system in the menstrual cycle simultaneously.

In conclusion, until now, there is no definitive answer to the exact pathological mechanism of PMDD. The theories above may investigate some parts of the cause but not the whole picture of it. Besides, there were inconsistent experimental methods and results in the previous studies which also made the important pathological factors of PMDD yet unclear. Thus, PMDD is still a disease with no objective diagnostic method but ambiguous diagnostic criteria.

Diagnosis

According to DSM-5, there are seven criteria to diagnose PMDD. Here only gives a summary. First, there have to be at least five symptoms in the week before the onset of menses. One or more of them must be mood lability, irritability, depression or anxiety. The list of symptoms also has to contain one or more symptoms of decreased interest in usual activities, difficulty in concentration, lethargy, change in appetite, hypersomnia or insomnia, a sense of being overwhelmed, and physical symptoms such as breast

tenderness or muscle pain. Second, the symptoms have to start to improve within a few days after the onset of menses, and become minimal or absent in the week after menses. The situation has to be in most of the menstrual cycles in the preceding year. Third, the severity of symptoms during the menstrual cycle should be confirmed by daily ratings for at least two symptomatic cycles.

In the narration above, there are vague expressions in the existing PMDD diagnosis criteria. The diagnostic process has to depend on subjective mood records from patients and subjective judgements from doctors. This not only makes the diagnostic process complicated and time-consuming, but also makes it difficult to find the proper treatment for each patient.

Treatment

The mainstream therapies of PMDD are selective serotonin reuptake inhibitors (SSRIs) and GABA-modulating drugs. SSRIs are the gold standard treatments for PMDD. They can provide rapid effect at low doses. In a study giving patients a 2-day 20 mg fluoxetine treatment, nearly two-thirds of the patients had significant improvement in irritability, sadness, anxiety, and mood swing (Steinberg, Cardoso, Martinez, Rubinow, & Schmidt, 2012). This unique property suggested the therapeutic effect of SSRIs might not be entirely serotonergic. Some of the studies suggested that SSRIs could promote the conversion of P4 to ALLO or could somehow enhance ALLO synthesis (Devall *et al.*, 2015; Pinna, Costa, & Guidotti, 2009). However, another study showed that SSRIs could increase peripheral ALLO level in patients with low baseline of ALLO, but had an opposite effect in the patients with high baseline ALLO level

(Gracia *et al.*, 2009). There are more clinical observation and research needed for the effect of SSRIs on PMDD patients.

GABA-modulating drugs are another option. Brexanolone is the first treatment approved by the Food and Drug Administration (FDA) for postpartum depression. It is a synthetic compound of ALLO, having the ability to upregulate GABAR pathway. Due to its ability of rapid symptom relief, brexanolone is also expected to be used for PMDD treatment (Meltzer-Brody *et al.*, 2018). Another medicine is sepranolone. Sepranolone (UC1010) is a new medicine for PMDD, which was under Phase IIa clinical trials in 2020. It is an isoallopregnanolone which can inhibit the effect of ALLO on GABAR. In the trials, it provided significant reduction of 75% in mood symptoms (Bixo *et al.*, 2017). However, in the Phase IIb trials, the placebo response was 33% higher than that in the Phase IIa study. It turned out that there was no statistical difference between the treatment groups, and the study results of UC1010 remained inconclusive.

The gold standard treatment, SSRIs, can only treat two-thirds of the patients, and has a risk of acting conversely in different individuals. There also had a new medicine, sepranolone, developed for PMDD treatment in the past few years, but it failed to pass the clinical trials. The treatments for PMDD still need to be improved and evaluated to be able to cope with the symptoms and individual patients precisely.

Problem statement and aim

As described above, PMDD is a disease with unknown details in mechanism, having ineffective diagnosis and treatment. To break through this obstacle, the key is to find the important factors of PMDD symptoms and use them to build an objective diagnostic method. Moreover, because there are multiple symptoms and multiple

reported biological factors in PMDD pathology, new tools which can deal with more than one feature at the same time were used to achieve the goals.

The tool chosen to investigate these critical factors was artificial intelligence (AI). AI has been widely used for data classification and prediction. More specifically, machine learning and deep learning methods in AI play important roles in achieving these objectives (Miotto, Wang, Wang, Jiang, & Dudley, 2017). There have already been applications in various biological fields in the last decade, including clinical medical research and animal behavior research. In the clinical field, AI has already been used for pathological diagnoses, surgical assistance, monitoring devices, and so on (Yu, Beam, & Kohane, 2018). In animal behavior research, AI has been used for applications such as separating different chirping patterns and detecting the appearance of specific species in wild grassland (Valletta, Torney, Kings, Thornton, & Madden, 2017). However, when it comes to experimental animals, the use of AI has only recently begun, and has been relatively restricted to behavior and pose estimation (Mathis & Mathis, 2020). Since PMDD shows contradictory experimental results in human patients and experimental animals, this study aimed to establish an objective diagnostic method utilizing hippocampal factors in rats by investigating the diagnostic factors of PMDD using AI tools.

In this study, pseudopregnant rats were used to mimic the hormone environment during the luteal phase in humans, and then the data of them were clustered by their severity of PMDD-like symptoms using unsupervised machine learning implementing the k-means algorithm. Next, the factors, which were most related to symptom severity, were extracted through an original two-step feature selection by the use of RNA sequencing (RNA-seq) results. These selected factors were considered as a diagnostic

set for PMDD symptoms and were used to build predictive machine learning models as diagnostic systems. Another group of rats that underwent the same experimental procedure was then used to evaluate the feasibility of the diagnostic systems.

There are three chapters in this thesis. Chapter 1 includes the introduction of the five methods and tools used for data analyses in the present study, the reasons why they were chosen, and the challenges occurred. Chapter 2 is the pilot study of the main experiment. It is referred to an attempt of using a P4 injection schedule to intimate the hormone environment of PMDD in rats and the behavioral tests performed. Chapter 3 is the main experiment. It includes the use of pseudopregnancy rats, the behavioral tests and their analyses, and how the target experimental group was determined by the use of an unsupervised machine learning algorithm. It also contains the process of choosing the important factors by analyzing RNA-seq results, the way to use the important factors to build the diagnostic systems by supervised machine learning algorithms, and the evaluation and prediction of the systems.

Chapter 1 Methodology

Introduction

In the present study, the latest techniques were attempted to be used in sections of data analyses. These tools were expected to help to provide efficient experimental operations and accurate experimental results. There were five tools. First, DeepLabCut, a deep learning package, was used to analyze the videos of behavioral tests. Second, k-means, an unsupervised machine learning algorithm, was used to cluster data of behavioral test results. Third, Usegalaxy.org, an online computational biological platform, was used to processed RNA-seq results. Fourth, an original two-step feature selection method, composed of manual selection and cross validation algorithm, was used to pick up the key genes of PMDD symptoms in the RNA-seq results. Finally, six supervised machine learning models were built to classify data from another group of rats undergoing the same experiment procedure by the use of the key gene set. Whether the clustering results from the behavioral tests and that from the mRNA expression levels of the key genes were identical or not would be verified.

Most of the tools are in the field of artificial intelligence (AI). The word “AI” refers to any mechanical system processing and acting rationally by mimicking human intelligence. The main functions include learning and problem solving (Norvig, 1995). It has several subfields, such as vision, robotics, natural language processing, machine learning, and decision theory (Pamela, 2004). AI is not a fancy concept which has already been raised for more than fifty years. However, it has not been widely used before the advancement of hardware in the last decade.

Machine learning (ML) is a subfield of AI which most of the tools in the present study belong with. ML is the study of computer algorithms which separates data by mathematical formula and has ability to improve performance by learning from experience (Mitchell, 2017). There are three main approaches in ML, which are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a way of using input vectors of data to find their corresponding target vectors. The cases which supervised learning can handle include classification problems with output of discrete categories, and regression problems with output of continuous variables. Unsupervised learning is a way to discover groups of data with only input vectors but no target vectors. It can be used for clustering and density estimation. The last one, reinforcement learning, is a way to find the best action to take by giving rewards or punishments. It is used for decision making of consecutive motion, such as game playing and autonomous driving (Christopher, 2006). Beside these three approaches, there is still a popular field included in the broad family of ML, which is deep learning. It will be explained below in the section of analysis of behavioral test video.

Analysis of behavioral test videos

Analysis of behavioral test videos and the difficulties

For analysis of behavioral test videos, human manual scoring is the unchangeable gold standard (Sturman *et al.*, 2020). Most research workers are able to point out specific postures or actions of animals correctly after training. The analysis results will not have much difference among annotators when the movements of animals are simple, short and clear. However, human manual scoring has limits which reduces reliability.

First, humans make mistakes. Humans can unconsciously take down wrong records when being tired or absent-minded. Second, humans can have different judgments when encountering complex or continuous movements. For example, one of the frequently-used behavioral parameters, rearing, is an action of an animal raising itself by hind legs. Although it can be described by words clearly, rearing can be an action with different patterns such as against a wall or not, rising high or low, by a single foreleg or both, or in order to explore or just to smell. Therefore, it is hard to have a common answer among annotators to determine whether a movement is rearing or not. At last, the scoring procedure by humans can be time-consuming. Since people can only analyze one video at a time to avoid missing details, it is easy to realize human manual scoring is not an efficient way if there were multiple tests or a large number of test animals.

Behavior analysis using EthoVision XT

To solve these issues, automated analysis methods with assistance of computers were brought out in the past two decades. EthoVision software series is probably the most popular one with more than 500 citations of its original paper. Here is how EthoVision works. First, it distinguishes the target object from the background by gray scale (contour-based tracking). Then, it finds the mathematical center and area of the object. At last, it calculates the specific measurement of behavior assigned by the users (Noldus, Spink, & Tegelenbosch, 2001). It has already been used in different kinds of experimental animals such as mice, rats, zebra fish, and in different kinds of tests such as open field test, elevated plus-maze test, Morris Water Maze test, and so on (Desland, Afzal, Warraich, & Mocco, 2014).

However, there are also deficiencies in EthoVision (version: XT, with no add-on modules). First, the principle of capturing objects by gray scale puts limitations on the colors of the animal and the background. It is better for the users to use animals with black hair in white background or *vice versa*. However, when the rats with white hair were used in the black background, the software could not separate the rats from reflected light only by the difference in their gray scale. The tracked points in the center of the rats even sometimes disappeared when the rats were resting in the corner of the apparatus, making the analysis results doubtful. Second, the tracked points cannot be customized freely. Basically, there can be only one tracked point per one object. The users cannot track different parts of the body of the animal for precise observation. Last but not least, the analytic results cannot be adjusted. There is no function to revise the points by the unit of frame. The users have a chance of finding their analytic works in vain in the very last moment. There are barely any studies focusing on the comparison of EthoVision XT and other tracking tools. However, a previous study comparing EthoVision 2.3 and an authors' original tracking tool raised similar problems mentioned above: (1) there can be only one individual under one detection; (2) the luminosity conditions can be strict (Delcourt *et al.*, 2006).

Due to the reasons above, the following AI-based tool was chosen to resolve the difficulties.

Behavior analysis using DeepLabCut

DeepLabCut (DLC) is an analysis tool based on AI deep learning structure, developed by the research teams of Harvard University and the University of Tubingen (Nath *et al.*, 2019). Before DLC, here is the short introduction to deep learning. Deep

learning is one of the ML methods. It was built with human brain structure composed of layers full of “neurons”. These neurons can tune the calculating parameters by themselves, and adjust the output results by the input from the last layer (Waldrop, 2019) (figure 1-1). The properties make deep learning a method with high accuracy compared to other shallow ML algorithms, and suitable for solving complex problems such as high dimensional data processing or image recognition (Goodfellow, Bengio, & Courville, 2016). In addition, deep learning technology has also been brought into the latest released EthoVision in 2021 (version: XT 16) to improve the tracking system according to the official page of EthoVision series.

DLC gathers the advantages of deep learning, and is now a powerful tool used for multiple situations in behavioral research. Since the publication in Nature Protocols in 2019, DLC has already had citations more than 700, and has been used in humans, wild animals and experimental animals for pose estimation, investigation of neural science, and even cellular image analysis. There are the benefits of DLC as follows, corresponding to the drawbacks of EthoVision XT mentioned before. First, it recognizes animals in video directly, but not by gray scale, reducing the chance of mixing up the target animals and the background or reflected light. Second, the users are able to decide the tracked points of their interest anywhere and at any amount. For example, for research of hand movement, the tracked points can be all the knuckles in the forelegs of a mouse; for research of facial expression, the tracked points can be as many as the muscles in interest. Last, the analytic results of DLC can be adjusted by label refinements or by editing the wrong frames. Label refinements can bring out better analytic results by improving the training model; while the manual correction of the wrong frames can directly improve the results. There is a study investigating the

analysis abilities of DLC and other two commercial solutions, EthoVision (version: XT 14) and TSE Multi Conditioning System (Sturman *et al.*, 2020). The authors demonstrated that among the three solutions, DLC is the one bringing out results with the most similar accuracy to human manual scoring, especially in the results of fine movements such as head dips in the open arms in elevated plus-maze test.

On the contrary, there are still some weak points in DLC. First, the technical requirement of using DLC is high. DLC has graphical user interface (GUI) for the users who are not familiar with programming languages. However, the operation is not intuitive. Users have to install and open the GUI by typing commands in the terminal application, and deal with version conflict of the built-in programming packages or other early-stage problems. If the users knew little about programming, a huge obstacle had already shown up in the first step. Second, the requirement of hardware is also high. It is better to have a graphics processing unit (GPU) than a central processing unit (CPU) to have enough processing capacity and efficiency. Last, DLC is a package designed to analyze recorded video, therefore it cannot track points which are invisible. Tracked points can disappear in some frames because of motion blur or other reasons. However, since DLC is a frequently updating package, the problems can be expected to be improved with progressive functions in the near future.

Analysis of behavioral test results

Analysis of behavioral test results and the difficulties

Generally, the behavioral test results of the experimental group and the control group are compared one by one with traditional statistical methods such as student t test or *post hoc* tests followed by analysis of variance (ANOVA) test. These methods focus

on the relationship among observations. If the null hypothesis (H_0) is rejected, then the alternative hypothesis (H_1) is established and there exists a difference between the means of the groups. This is how traditional statistical methods work, being widely acceptable due to the rigorous inference procedure. However, they have a limitation. Traditional statistical methods cannot deal with plenty of features at the same time. By the results of traditional statistical methods, the difference among groups in every single test can be discovered, but the relationship of all the test results in every individual cannot be compared and understood. For example, a famous behavioral test, open field test, has several parameters such as total traveled distance, ratio of time in center zone, number of rearing, number of grooming, number of passing through center and peripheral zone, and so on. It is difficult for traditional statistical methods to know if an individual showed anxiety-like behavior in all the parameters related to anxiety by only comparing the differences of the means of the parameters among groups. Therefore, for a disease with multiple symptoms such as PMDD, instead of traditional statistical methods, unsupervised learning in AI was chosen in the present study to separate healthy and sick individuals.

Unsupervised learning

As mentioned in the introduction of this chapter, unsupervised learning is a way to discover the subgroups of data with only input vectors but no target vectors. It contains algorithms built by different formulas, and thus it is a notion hard to be precisely defined (Estivill-Castro, 2002). There are two main kinds of tasks which unsupervised learning can handle, clustering and density estimation. Algorithms for clustering can put similar data points into a specific group; while algorithms for density estimation use

observed data to estimate the distribution or other parameters for the density function. In the present study, neither the individuals which had severe PMDD symptoms, nor the exact border criteria to separate healthy and sick individuals by the behavioral test results was unknown. Thus, the algorithms for clustering were needed to help to put the individuals with similar severity of symptoms together by considering multiple parameters of the behavioral tests at the same time.

Candidate algorithms in unsupervised learning

The next step is to choose a suitable clustering algorithm. In the field of AI, there are countless algorithms for different tasks. Only the most popular ones, k-means and density-based spatial clustering of applications with noise (DBSCAN), were considered. K-means is a centroid-based clustering. The number of centroids is decided in the beginning, then each point is assigned to the cluster of the nearest centroid. DBSCAN belongs to density-based clustering. It connects points which meet the specific density threshold (Oded & Lior, 2005). The schematics of these two clustering methods and the comparison are shown in table 1-1 (Fabian, Gael, Alexandre, Vincent, & Bertrand, 2011).

The principals of k-means

Although there are both advantages and disadvantages in k-means and DBSCAN, k-means was chosen in the present study. There are several reasons. First, k-means is sensitive to noise. It can serve as an outlier trimmer and remove the needless points at the beginning of the analysis. Second, the number of centroids can be customized. This makes the results can be adjusted to a certain degree after data observation. At last, k-

means provides linear results. The goal in this section is to separate healthy and sick individuals by the severity of PMDD symptoms, which is corresponding to the degree of behavioral change in the behavioral tests. By using k-means, the clusters can be obtained with linear relationship by the order of degree of behavioral change, and thus can bring out results with higher interpretability and reliability.

Processing of RNA sequencing raw data

Processing of RNA sequencing raw data using Usegalaxy.org

Processing of RNA sequencing (RNA-seq) data is included in the field of bioinformatics. It usually requires certificate programs or professional training beforehand, otherwise it is hard to get started by self-study. Besides, to process RNA-seq raw data also needs the skill of programming since most of the tools are R or python-based packages. Researchers who are not familiar with bioinformatics often choose to outsource the tasks to other laboratories or private enterprises. However, outsourcing means there is going to be more cost but less operating flexibility. Due to the reasons above, there is need to be an alternative solution.

Galaxy is an open-sourcing python-based software developed by a team gathering Pennsylvania State University, Johns Hopkins University, Oregon Health & Science University, and the Galaxy Community (Afgan *et al.*, 2016). It is a platform collecting analytic methods for genomics, proteomics, and transcriptomics. The tools are turned into GUI form so the users without programming skills are also able to build their own tasks. The Galaxy team also provides tutorials on the site of Galaxy Training!: https://training.galaxyproject.org/. The articles include not only the step-by-step instructions but also the explanation of related principles, and thus it is easy for the

users without basic knowledge of bioinformatics to perform learning by doing. In the present study, the tools in Usegalaxy.org were used for reference-based RNA-seq data analysis to obtain the whole mRNA expression levels in hippocampus, trying to discover the genomic difference between healthy and sick individuals. The procedures include quality check by FastQC, trimming by Cutadapt, mapping by HISAT2, and annotation by featureCounts.

Feature selection method for the important genes

Conventional choices and the difficulties

After processing the RNA-seq raw data, the mRNA expression levels of the genes in the samples were obtained. The next step is to only keep the important ones. There have already been tools to find the genes with significant statistical differences between the experimental group and the control group in Galaxy, such as DESeq and edgeR. However, from these tools, only the group of genes with significant differences but not their ranking of importance can be acquired. Thus, the tools in ML were chosen to find the genes whose expression levels have the closest relationship to the severity change of PMDD symptoms.

Feature selection

The method is called feature selection, a process of finding the subsets in need of the features. There are three types of the feature selection methods: filter, wrapper, and embedded (Guyon & Elisseeff, 2003). Filter method scores each feature by their divergence or correlation, and sets specific threshold or number to choose features. Wrapper method picks up or excludes some of the features in each test, and chooses

features by the score of predictive performance. Embedded method works like filter method, but chooses features in the process of training. The comparison of the three feature selection methods is in figure 1-2 (Mohd Rosely, Salleh, & Zain, 2019).

After comparing the three methods, filter method was chosen for the following reasons. First, it is the fastest method. When there is neither computer with high calculating ability, nor time to wait for time-consuming calculations, the efficiency is always the first consideration. Second, it is the only method that is independent of specific ML algorithm. The results brought out by filter method are not deeply affected by one specific classifier, and thus different algorithms can be tried to find the most suitable one for the present experiment. Finally, it is the method that is less prone to overfitting. Overfitting is the situation when a ML model performs well in the training data but poorly in the test data. If a model is overfitted, it only remembers the correct answer by rote, but not learning the pattern from the training data and applying it to the test data. In the present study, since there only had a small number of samples and the data was easily overfitted, filter method was chosen to guarantee the accuracy of the results.

Filter feature selection method

At first, a filter feature selection method built by SelectKBest and Linear Support Vector Machine (Linear SVM) in a repeated k-fold grid search cross-validation (Grid Search CV) was used. SelectKBest is an algorithm for feature selection. It selects the best features by univariate statistical tests, and only keeps the *k* highest scoring features (Blesson, 2017). Linear SVM is one of the popular algorithms often used for data classification. It can find a line or hyperplane with the maximal margin to separate two

groups of data (Chang & Lin, 2008). Grid Search CV is a cross-validation method for parameter tuning. It splits data into *k* groups and performs multiple tests to check the performance of ML model. In other words, Grid Search CV can verify models under different sets of features to find the set of features with the best performance (Ranjan, Kumar Verma, & Radhika, 2019).

Here is how this filter feature selection method worked. First, SelectKBest is used to make Grid Search CV able to obtain all the combinations of all the features. Then, Linear SVM is used to perform the classification under each combination of the features. At the end, Grid Search CV find the combination with the best mean model accuracy. In the present experiment, the mRNA expression levels of genes from the RNA-seq results were the features. The genes in the combination with the best mean model accuracy were the most important factors for PMDD symptoms.

However, the dataset of the RNA-seq results was a small dataset with high dimensions. There were only 12 samples but 32,545 gene expression levels detected in all. After several attempts, the best accuracy remained low and the selected gene set remained suspicious.

The original two-step feature selection method

For this reason, a pre-processing step was added to reduce the dimensions of the dataset before executing the filter feature selection method. During the testing phase, the genes with small numbers of their expression levels were found more likely to be picked up, and had a relatively larger influence on the selection model. Besides, there were a certain number of unknown genes whose functions had not been studied yet. Thus, the genes which failed to meet the following three standards were manually

filtered out: (1) genes with mean expressions increasing/decreasing by the order of severities of symptoms; (2) genes without zero in their expression levels; and (3) genes that were not in the unknown list in Database for Annotation, Visualization and Integrated Discovery (DAVID) functional annotation tool.

By adding the pre-processing step, an original two-step feature selection method was established. In the first step, the genes which might disturb the later analysis were filtered out, and the dimensions of the dataset were reduced. The number of the genes went from 32,545 to 5,047. In the second step, the remained genes from the last step were used to execute the filter feature selection method described above. After the two steps, the set of the most important 17 genes was found.

Predictive machine learning models

The last step in the present experiment is using the set of the most important genes to build supervised learning models to classify the individuals, compare the classifying results with the clustering results of the behavioral test results, and evaluate the performance of the models. Since the cluster to which each individual belongs assigned by k-means has already been determined, the supervised learning algorithms were then used to find the "rules" between the classification results and the expression levels of the set of the most important genes. The six supervised learning algorithms, which were the most popular ones with different principles were chosen. They were Linear SVM, Random Forest (RF), Bag SVM, Neural Network (NN), Light Gradient Boosting Machine (Light GBM), and Extreme Gradient Boosting (XGBoost). The illustrations of the operating principles of the supervised learning models are in figure 1-3.

Linear SVM

The introduction of Linear SVM is also in the paragraph of feature selection. It is a method which can separate two groups of data by a line or hyperplane with the maximal margin. More specifically, the points in the margin of each group are detected, and the hyperplane is calculated by the use of the coordinates of the margin points (Joachims, 1998).

- Advantages: usually performs well in most of the problems compared to other algorithms, good to use in samples with number of dimensions greater than the number of samples, hard to have bias when there are outliers.
- Disadvantages: need more time to process, the principle is not well known, cannot do well if the boundary points are noise.

Random Forest (RF)

RF is an ensemble learning method grouped by multiple decision tree (DT) predictors. DT is one of the earliest algorithms and it separates data by decision logic. The process starts from the starting point called “root node”. By the process of decision making, the branches and levels of the tree increase until the “leaf nodes” are reached (Quinlan, 1986). However, it has a risk of overfitting when a tree is extremely complicated with excessive levels.

RF is a solution to the overfitting situation in DT. It gathers multiple DTs to train the different parts of the dataset and decide the final answer of classification by averaging or voting by all the DT trees (Breiman, 2001). In this way, the result will not depend on only one single tree, and will not be affected easily if there is a small change in the input data.

- Advantages: efficient on large dataset with even thousands of features, able to deal with missing data, able to deal with irrelevant features.
- Disadvantages: can still be overfitted if the dataset is noisy, have to choose the number of DTs manually, difficult to interpret.

Bag SVM

Bag SVM is an ensemble learning method consisting of several SVM models in a bagging classifier. Similar to RF, it fits base models on random subsets of data and decides the final prediction by aggregation or voting from the predictions of base models (Breiman, 1996). Bagging classifiers are often compared with boosting classifiers. The comparison of these two is in table 1-2.

Neural Network (NN)

NN is a model with structure imitating the human brain. In the biological brain, there are neurons connected to each other. When new information comes in, the neuroplasticity occurs to adapt to the new situation. In NN, there are input and output neurons connected by hidden layers. There are also weighted synapses being able to adjust signal strength. It gives out results by learning from former layers, and has the ability to tune itself (Rumelhart & Hintont, 1982). Deep learning mentioned before is a classic method based on NN.

- Advantages: able to deal with unknown situations, suitable for non-mathematical models, easy to build with less formal statistical knowledge.
- Disadvantages: takes a long time to process large network, needs a large amount of input data, prone to be overfitted due to the complexity of model structure.

Light Gradient Boosting Machine (Light GBM)

Light GBM is a gradient boosting machine based on the "leaf-wise" DT algorithm brought out by the Microsoft team. Compared to the traditional "level-wise" method, the leaf-wise algorithm can reduce more loss, give out results with better accuracy, and perform faster (Ke *et al.*, 2017).

Extreme Gradient Boosting (XGBoost)

XGBoost is a model similar to Light GBM but based on the traditional "level-wise" algorithm. It is an ensemble method of adaptive boosting algorithm. Adaptive boosting algorithm works by making *n* number of DTs, and brings out results by passing on the incorrect parts or the weaknesses from the last tree to the next tree to adjust and strengthen the whole model (Chen & Guestrin, 2016). Although XGBoost usually performs well and is good for optimization, the training process is rather time-consuming.

To sum up, there were five tools for the sections of data analysis in the present study. For the behavioral tests, DLC was used for analyzing the videos; while k-means was for clustering the result data. For the RNA-seq results, Usegalaxy.org was used for processing the raw data; while the original two-step feature selection method containing manual filtration, SelectKBest, Linear SVM, and Grid Search CV pipeline was for the extraction of the most important genes. Lastly, for testing whether the expression levels of the selected genes could bring out the same classifying result as the behavioral tests did, the six classifiers of Linear SVM, RF, Bag SVM, NN, Light GBM, and XGBoost were built and evaluated.

Figures and tables

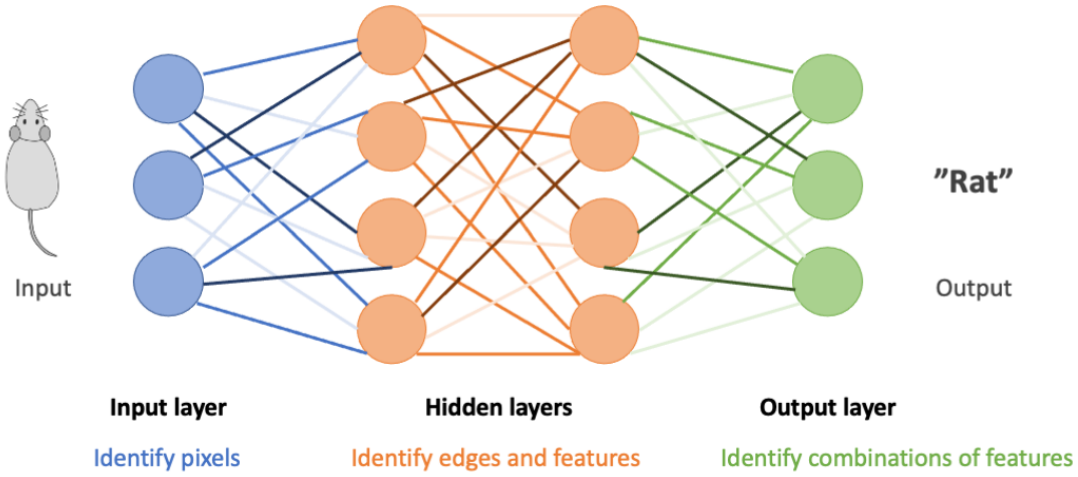


Figure 1-1. The concept of the deep learning network.

The network recognizes the input picture or video frame by identifying it from the pixel level to the combination of features level. By investigating input precisely, the multiple-layer process brings out output answer with higher accuracy.

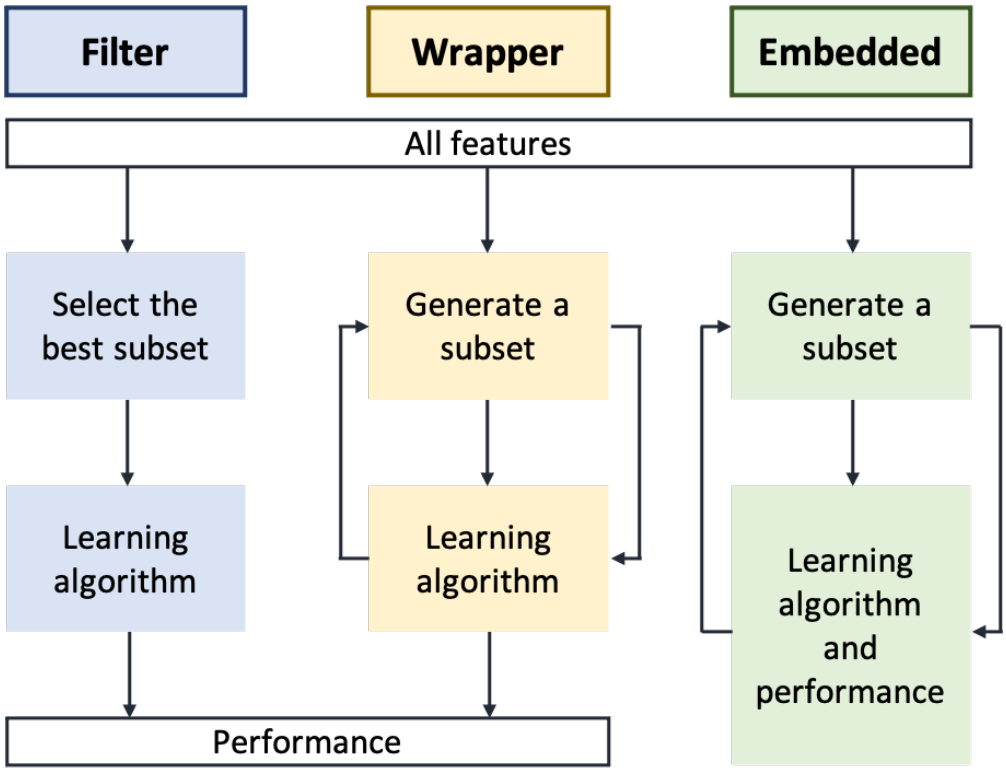


Figure 1-2. Work flow of the feature selection methods.

The three main feature selection methods are filter, wrapper, and embedded. All of them go through the steps of generation of subsets of features, and evaluation of the performance of algorithm. The major difference among them is the degree of independence of the steps inside. For example, the steps in filter method are independent to one another, making filter method easy to be customized.

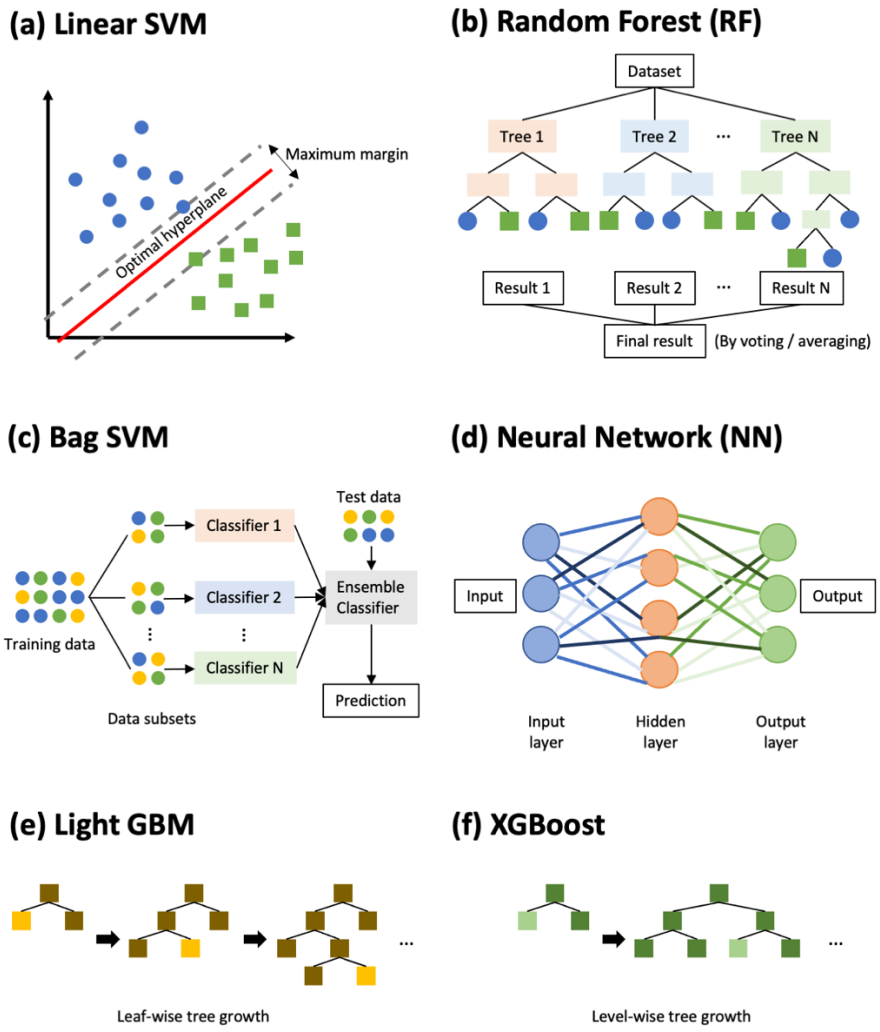


Figure 1-3. Principles of the six classifiers.

The six classifiers used in the present study are the most classical and popular ones for AI tasks nowadays. (a) Linear SVM separates data by drawing a margin line or hyperplane. (b) RF decides the classes of data by concluding the results from the decision trees inside. (c) Bag SVM wraps up several SVM classifiers to improve the classification result. (d) NN mimics the structure of the human brain. (e, f) Light GBM and XGBoost are boosting algorithms, improving the results by loss reduction.

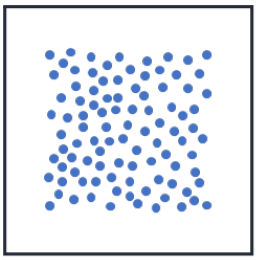
	k-means	DBSCAN
Schematic	 	 
Number of clusters	Pre-determined	Data-driven
Sensitivity to noise	Sensitive	Insensitive
Missing value handling	No	No
Advantages	Be able to process fast and multiple times, efficient in large dataset	Be able to find clusters in any shape, insensitive to the order of data
Disadvantages	Results can be inconsistant, the generation of initial centroids is random	Border points can be in more than one cluster, finding the density threshold can be difficult

Table 1-1. Comparison of k-means and DBSCAN.

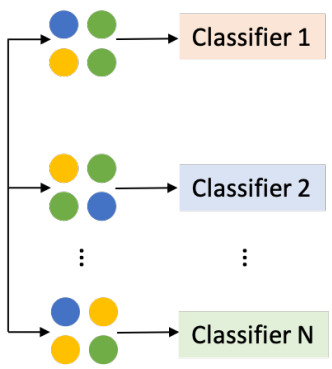
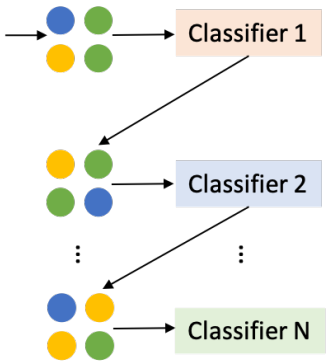
	Bagging	Boosting
Schematic	 Parallel	 Sequential
Similarities	Both use voting system and combine models of same types.	
Differences	Individual models are built separately.	Each model is influenced by the previous one.
Advantages	Be able to against noise	Good to be optimized
Disadvantages	Needs a lot of comparable classifiers	Cannot against noise

Table 1-2. Comparison of bagging and boosting algorithms.

Chapter 2 Analysis of a PMDD Rat Model with Exogenous Hormone Evocating Method: A Pilot Experiment

Introduction

Due to the unknown in the detail pathological mechanism, using animals with deficiencies of possible genes or reproducing hormone environment similar to the human luteal phase have been the two broad categories used in the previous PMDD studies. For example, there were knockout mice deficient in brain-derived neurotrophic factor gene or bromodomain containing 1 gene used to evaluate the degree of behavioral dysfunction under the specific gene deficiency (Marrocco *et al.*, 2020; Rajkumar *et al.*, 2020). There were also animals produced by giving exogenous or endogenous P4 evocation through different procedures and dosages, including capsule implants, daily injection, and pseudopregnancy induction (Li *et al.*, 2012; Moran and Smith, 1998; Smith, Freeman, and Neill, 1975). In the present study, since the object was to investigate the relationship between PMDD symptoms and their related endogenous factors in a comprehensive way, construction of luteal phase-like environment was chosen instead of transgenic techniques.

A simple review of P4 evocation methods used in previous studies is in table 2-1. In earlier years, ovariectomy was more prevalent to be used for sex hormones withdrawal in the beginning or the end of experiments. It was considered to be an effective measure to eliminate the effect from primary reproductive system (Ho, Olsson, Westberg, Melke, & Eriksson, 2001; Smith *et al.*, 1998). On the other hand, after the 2000s, exogenous sex hormone replacement methods without ovariectomy and pseudopregnancy were more prone to be performed. The former was performed by P4

(and/or estradiol) withdrawal with the use of injection schedule or subcutaneous capsule; while the latter was operated by mounting with vasectomized male, injection with mare serum gonadotropin, or vaginal mechanical stimuli (Löfgren, 2009). There was also a special case that providing P4 by subcutaneous injection 4 hours before experiment (Schneider & Popik, 2009). However, most of the studies followed the principles of a long-time exposure and rapid withdrawal of sex hormone. In 2012, Li *et al.* compared four P4 21-day withdrawal methods, and found the multiple P4 withdrawal schedule (m-PWD method in table 2-1) was the most sufficient way to induce behavioral change of premenstrual dysphoria (Li *et al.*, 2012). Therefore, in the present study, the m-PWD method was chosen to imitate the hormone change which was considered to induce PMDD symptoms.

Furthermore, due to lack of standard experimental process of PMDD study, the behavioral tests were also selected by literature review. According to the studies by Kaplan's group, it is better that a disease model should have same structure with human pathology or behavior. It should fit predictive validity (pharmacological correlation), face validity (behavioral isomorphism), or construct validity (similarity of underlying mechanism) (Kaplan, 1973). However, species differences also make the relevance of the three criteria questioned. The pharmacokinetics, the patterns of behavioral responses, and the underlying factors can all have a chance to be different between animal and human, especially in psychological diseases (Belzung & Griebel, 2001). Despite the practical difficulties mentioned above, the three validities are still the gold standard to follow. Therefore, for the disease with neither authorized treatment nor detail pathogenesis such as PMDD, reaching face validity as far as possible became the priority when choosing the suitable behavioral tests.

Although there are a large number of symptoms of PMDD described in DSM-5, anxiety, depression, and difficulties with social interaction were the three main focuses of the behavioral tests in the previous studies using rats. For anxiety, open field test (OFT), elevated plus maze test (EPM), and marble burying behavior test were the most popular ones (Islas-preciado, Lo´pez-rubalcava, Gonza´ lez-olvera, Gallardo-tenorio, & Estrada-camarena, 2016; Löfgren, Johansson, Meyerson, Turkmen, & Bäckström, 2009). For depression / anhedonia, forced swim test (FST) and saccharin / sucrose preference test were the choices (Schneider & Popik, 2007). And for difficulties with social interaction, tube test, social preference test, food competition test, and so on were used (Ho *et al.*, 2001; Li *et al.*, 2012; Löfgren, 2009). Considering of the scale of the present study, three behavioral tests of OFT, EPM, and FST were selected to evaluate the primary locomotor function, and anxiety-like and depression-like behaviors due to their properties of simplicity and time-saving.

In this chapter, whether the m-PMD method could provide high concentration of P4, and whether it could bring out behavioral changes of PMDD symptoms by OFT, EPM, and FST was investigated.

Materials and methods

Animals

Eight- to 10-week-old female Wistar-Imamichi rats were obtained from the Imamichi Institute for Animal Reproduction (Tsuchiura, Japan) and acclimatized for 1 week before the experimental schedule. Animals were maintained under a 12/12 h light/dark cycle (lights on at 07:00), 4 rats per cage, and given *ad libitum* access to food and water. The estrus cycle stage was accessed daily by observing vaginal smear for

two cycles, and only the rats with regular cycles were included in the following experiments. The estrous status included proestrous, estrous, metestrous, and diestrous (proestrous: nucleated epithelial cells, estrous: non-nucleated cornified cells, metestrous: cornified cells with leukocytes, diestrous: large number of leukocytes). All experiments were conducted in accordance with the Guidelines for the Care and Use of Laboratory Animals, Graduate School of Agriculture and Life Sciences, The University of Tokyo.

P4 injection schedule

P4 injection was performed under one, two or three injection cycles. Each cycle included seven days. Each rat in the experimental groups was injected intraperitoneally with 6 mg/rat of P4 dissolved in 0.2 ml sesame oil or the vehicle saline per day in the former five days followed by withdrawal in the latter two days (n=4-5 / group). The 3-week experimental schedule of the P4 injection group is in figure 2-1. On the last day of each cycle, the following behavioral tests were performed.

Behavioral tests

The behavioral tests were performed during the light period in the following order and time: OFT at 10:00, EPM at 12:00, and FST at 14:00. Each test was followed by a 1-h interval. The sizes and shapes of the apparatuses are in figure 2-2. The analysis of the behavioral test results was performed by EthoVision (version: XT) followed by the official manual.

- OFT

The OFT apparatus consisted of four white wooden walls and a black polypropylene floor without a ceiling (75 × 75 × 45 cm) and was placed in a room with brightness of 60–80 lx. Each rat was placed in the center of the floor at first, and allowed to freely explore the field for 10 min. The open field was divided into a center zone (45 × 45 cm) and a peripheral zone. Avoidance of the center zone is considered as anxiety-like behavior. Other parameters included total distance traveled and number of rearing.

- EPM

The EPM apparatus was elevated to a height of 50 cm from the floor and had four arms (10 × 50 cm); two of the arms had 40-cm high walls (closed arm; east and west arms), while the other two had only 5-mm high ridges (open arms; north and south arms). The arms were merged at a central square (10 × 10 cm) to form a plus shape. Each rat was placed in the central square facing an open arm at first, and allowed to explore the maze freely for 10 min. The rats that fell off the apparatus were excluded from the later analyses. A higher percentage of time spent in closed arms was considered as anxiety tendency. Other parameters included total distance traveled and number of head dips in open arms.

- FST

The FST apparatus was a cylindrical box with a diameter of 38 cm filled with water to a height of 30 cm. Each rat was placed in the center of the water at first, and

left in the box for 6 min. A higher percentage of immobile time is considered as having depression tendency. Total distance traveled was also recorded.

Statistics

Two-way ANOVA test and pairwise Tukey's *post hoc* test was performed by Python using statsmodels package. Only the results with p-value < 0.05 in the two-way ANOVA test were performed Tukey's *post hoc* test.

Results

Serum P4 concentration on the test day of each group is in figure 2-3. Our target group is the experimental group in the 3rd week (E3 group). The E3 group had a significantly higher P4 concentration with a mean over 160 ng/ml (168 ± 32 ng/ml) compared to other groups. The vehicle group in the 3rd week tended to increase but with no statistical difference.

The behavioral test results are in figure 2-4. There are the results of OFT (figure 2-4 a-c) and EPM (figure 2-4 d-f). The results of FST were excluded due to the confusion between target animal and reflected light. The E3 group had a rather higher time ratio in closed arms in EPM but also a longer error bar for larger individual difference (figure 2-4 e). It also had a smaller number of head dips in open arms (figure 2-4 f). However, there were no statistically significant differences among groups in all the results.

Discussion

In this chapter, m-PWD method was chosen to imitate the hormone environment evoking PMDD symptoms, and the anxiety- and depression-like behaviors were also

evaluated by three behavioral tests and EthoVision (version: XT). As a result, the E3 group had a significant higher P4 concentration; however, there were no statistical differences in the behavioral test results.

For the results of serum P4 concentration, the highest number in the P4 injection group was in the E3 group (168 ± 32 ng/ml). The concentration of 168 ng/ml was higher than the highest progesterone level in a natural pregnant period, which has been indicated to be about 130 ng/ml (Pepe & Rothchild, 1974). This suggested that daily injection might be a method inducing excessive progesterone which could have a problem to reflect the physiologically normal situation.

For the behavioral test results of the P4 injection group, there were no significant differences among groups in the present study. The reason could be that EthoVision (version: XT) was not able to provide correct analytic results due to its principle of capturing objects. It could only distinguish the target animals and the background by their difference in gray scale. Moreover, when there was reflected light or immobile target animals, the software could not obtain the accurate location of the animal and brought out improper results. A previous study in 2006 indicated the similar disadvantages of early version of EthoVision. There were strict setting conditions and much background noise when detection (Delcourt *et al.*, 2006). Thus, instead of EthoVision (version: XT), there was better to have a new tool to be used in the future analysis of behavioral tests in the present study.

In conclusion, there were two concerns from the present study. One was that m-PWD was a good method to evoke high P4 concentration but might not present the true situation under the normal physiology. The other was that EthoVision (version: XT) could bring out inaccurate results and might not be suitable for the present behavioral

test analysis. These two issues should be solved in the main experiment. Thus, rats with endogenous P4 evocating method and an AI-based tool for the analysis of the behavioral tests were used in the next chapter.

Figures and tables

D1	D2	D3	D4	D5	D6	D7 TEST
1.P4 2.P4 3.P4	1.P4 2.P4 3.P4	1.P4 2.P4 3.P4	1.P4 2.P4 3.P4	1.P4 2.P4 3.P4	1.x 2.x 3.x	Open Field EPM Forced Swim
D8	**D9**	**D10**	**D11**	**D12**	**D13**	**D14 TEST**
1.- 2.P4 3.P4	1.- 2.P4 3.P4	1.- 2.P4 3.P4	1.- 2.P4 3.P4	1.- 2.P4 3.P4	1.- 2.x 3.x	Open Field EPM Forced Swim
D15	**D16**	**D17**	**D18**	**D19**	**D20**	**D21 TEST**
1.- 2.- 3.P4	1.- 2.- 3.P4	1.- 2.- 3.P4	1.- 2.- 3.P4	1.- 2.- 3.P4	1.- 2.- 3.x	Open Field EPM Forced Swim

Figure 2-1. Experimental schedule.

All the rats were divided into three groups, and operated 5-day P4 intraperitoneal injection followed by 2-day withdrawal cycle in 1, 2, or 3 weeks. On the last day of each cycle, three behavioral tests were performed to evaluate the degrees of anxiety- and depressive-like behaviors.

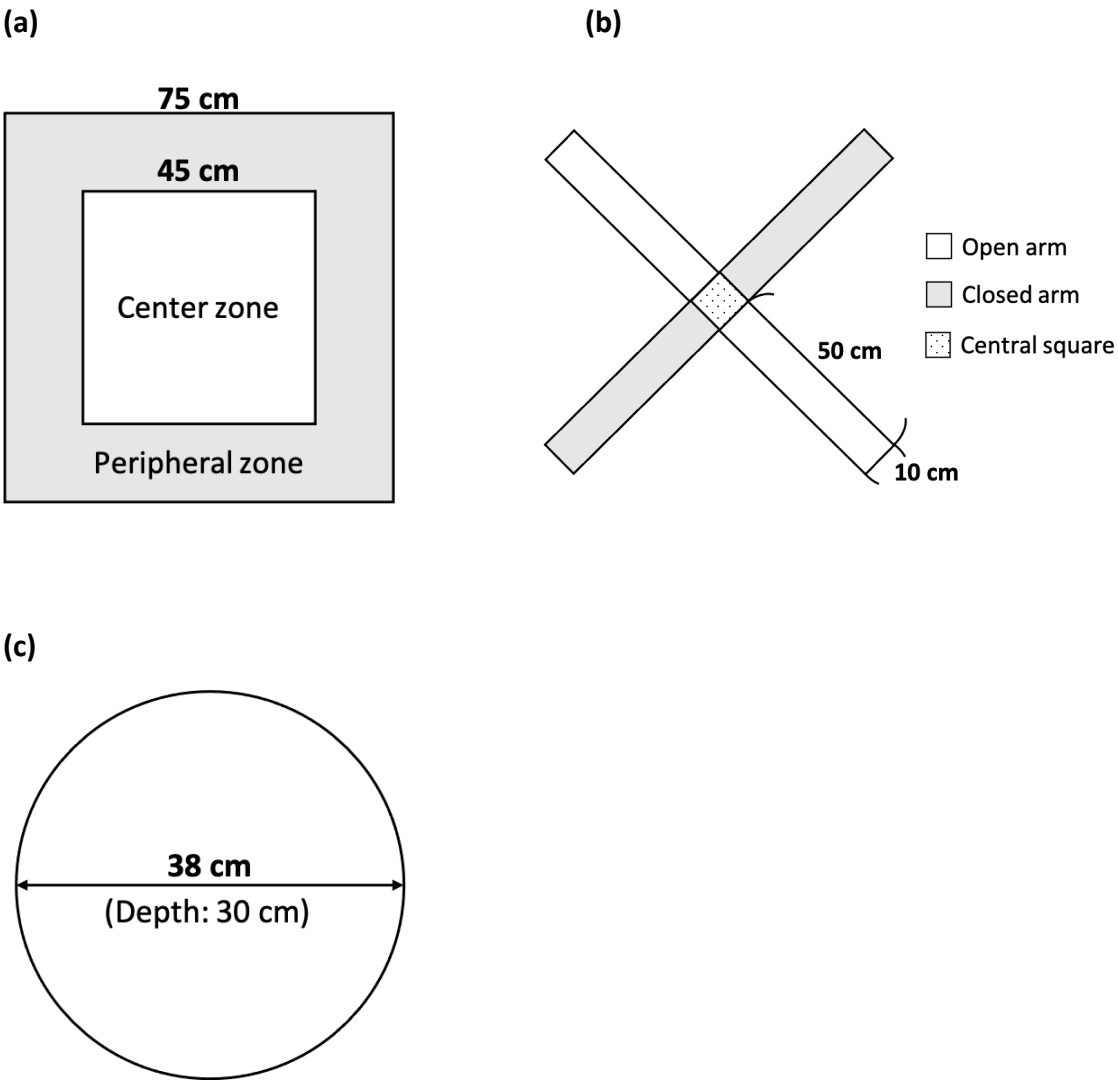


Figure 2-2. The apparatuses of the three behavioral tests.

(a) Open field test (OFT). (b) Elevated plus maze (EPM). (c) Forced swim test (FST).

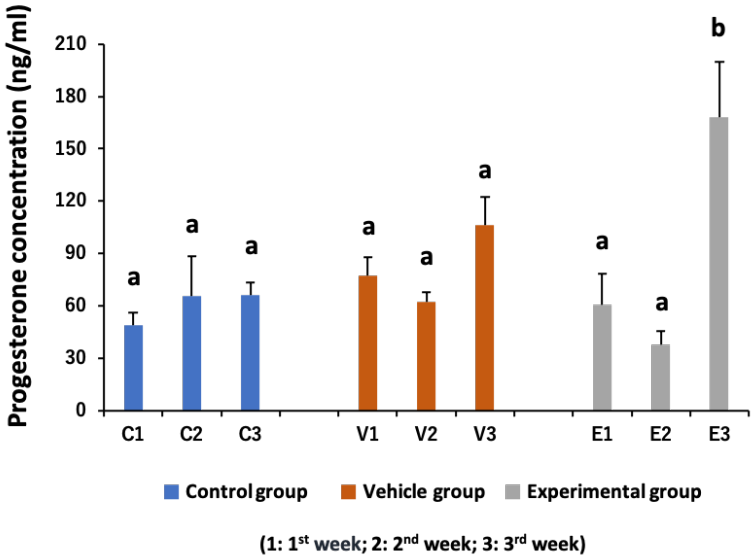


Figure 2-3. Serum P4 concentration.

Serum progesterone concentration on the last day of each group (n=4-5). Different letters indicate statistical significance. The error bars represent the standard deviations.

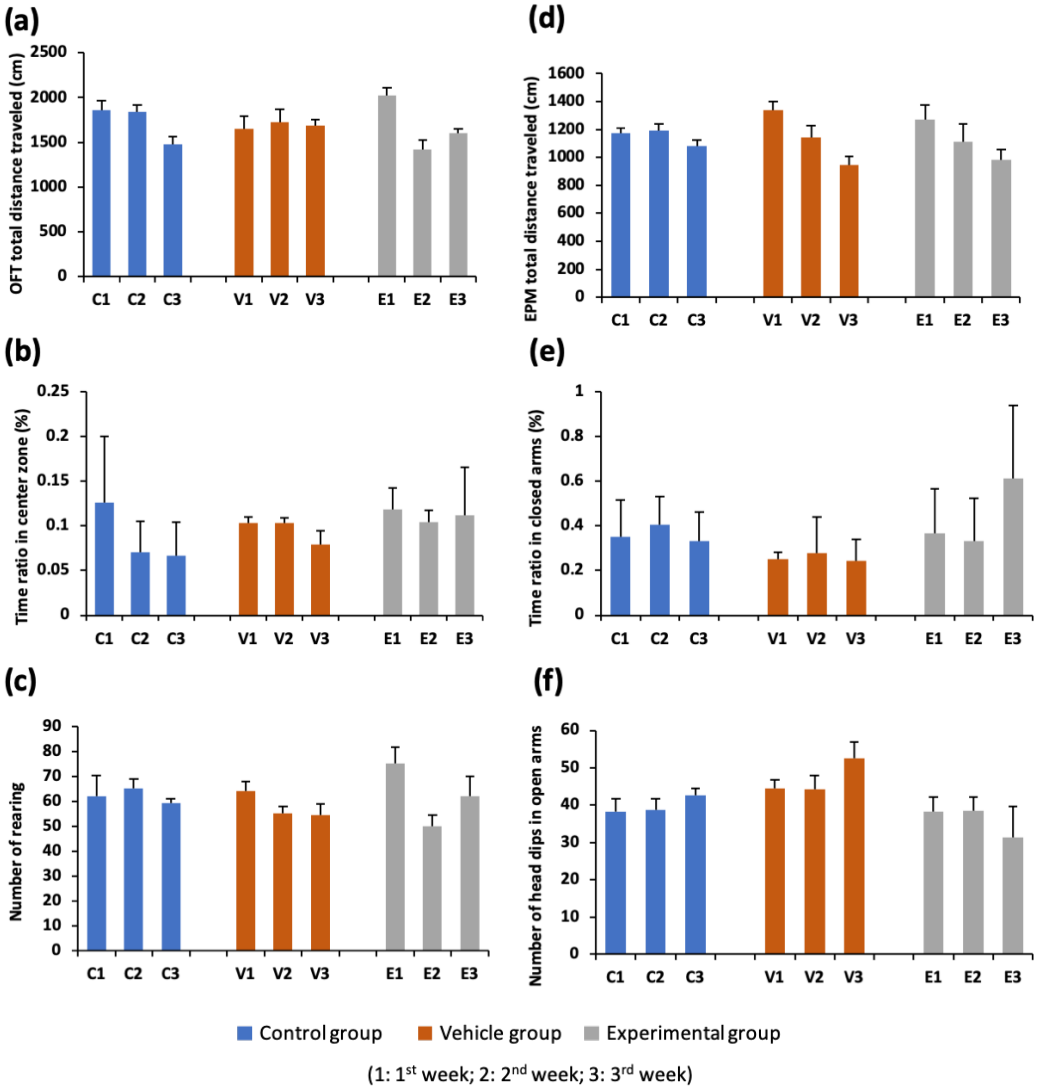


Figure 2-4. Results of the behavioral tests.

(a-c) Results of open field test. (a) Total distance traveled in the OFT. (b) Time ratio in the center zone in the OFT. (c) Rearing number in the OFT. (d-f) Results of elevated plus-maze test. (d) Total distance traveled in the EPM. (e) Time ratio in closed arms in the EPM. (f) Number of head dips in open arms in the EPM.

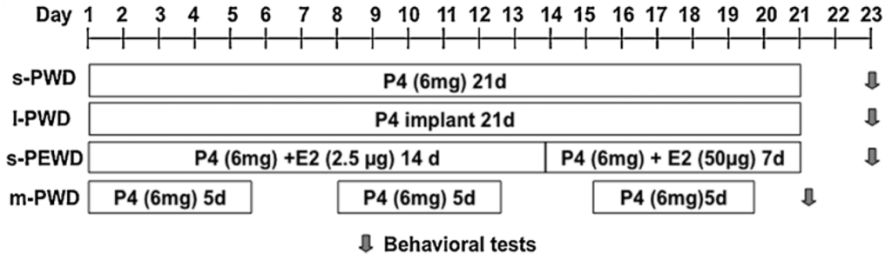
Research	P4 evocation method
Smith *et al.,* 1998	1. Pseudopregnancy: Pregnant mare serum gonadotropin dissolved in saline was injected intraperitoneally (20 IU/0.2 ml) on postnatal day 27. Human chorionic gonadotropin was injected (10 IU/0.2 ml) on postnatal day 29, which was also considered day 0 of pseudopregnancy. 2. Progesterone withdrawal: On day 11 of pseudopregnancy, the rats were ovariectomized.
Ho *et al.,* 2001	Ovariectomized rats were primed with estradiol benzoate (estradiol-3-benzoate in oleate 5mg/ ml; diluted in sesame oil, 0.5 mg/rat in 0.1 ml, s.c.; 0 h), estradiol (1, 3, 5 [10]-estratiene-3, 17b-diol; 0.5 mg/rat in 0.1 ml sesame oil, s.c.; 32 h), and progesterone (4-pregnene-3, 10-dione; 0.5 mg/rat in 0.1 ml sesame oil s.c.; 44 h).
Hsu and Smith, 2003	Progesterone-filled capsules of silicone tubing (10 mm per 100 grams of body weight) were implanted subcutaneously in the lower aspect of the back.
Löfgren, 2009	Progesterone 5 mg/kg + estradiol 10 µl/kg once a day (i.p. for 6 and s.c. for 10 days), followed by a 24-hour withdrawal.
Schneider and Popik, 2009	Progesterone 1 mg/kg, s.c. in corn oil in a volume of 0.4 ml/rat, 4 hr every time before testing
Li *et al.,* 2012	

Table 2-1. Progesterone evocation methods in previous studies.

Chapter 3 Main Experiment

Introduction

In this chapter, developing new option of experimental animal and analytic tool for behavioral test were the two main priorities. Exogenous P4 evocation methods including m-PWD elevated serum P4 concentration high but exceeded. To avoid exogenous effects, pseudopregnancy was chosen. Pseudopregnancy is a method developing sign of pregnancy without implanted embryo, and can maintain high-level progesterone during a long period (Chan, 2013). In fact, both external and internal ways are able to induce pseudopregnancy. The external way uses subcutaneous injection of pregnant mare serum gonadotropin followed by human chorionic gonadotropin (Bar-Ami *et al*., 2006); while the internal ways is performed by vaginal mechanical stimuli using glass rod or vibrating probe (De Feo, 1963, 1966; Frye & Bayon, 1998). After all, the internal way of mechanical vaginal-cervical stimuli using glass rod was chosen.

Another issue in the previous study was the tool of analysis for behavioral test. Although there have always been novel analytic tools of behavioral tests presented by different research teams, DLC is one of the most widely spread tools in the recent years. DLC is a deep learning-based package for pose estimation, especially used in experimental animals (Nath *et al.*, 2019). When compared to EthoVision (version: XT14) and another commercial tool of TSE Multi Conditioning System, DLC showed results with similar accuracy to human manual scoring in OFT, EPM, and FST, especially in the results of fine movements (Sturman *et al.*, 2020). Besides, the popularity of DLC can be demonstrated by the citation number of the original journal

paper, which has come to 917 at the time of writing. Thus, DLC was chosen to be used in the present study instantly as it might be the next main current of behavioral test analysis.

Moreover, besides DLC, other AI tools and algorithm were also used in the present study for different purposes. They were (1) unsupervised machine learning algorithm, k-means, for data clustering; (2) original two-step feature selection method built by manual selection and Grid Search CV pipeline for filtering out the important genes; and (3) six supervised machine learning algorithms for building diagnostic system. The backgrounds and operating details of these methods are in chapter 1.

In this chapter, the behavioral test results of pseudopregnant rats were clustered by k-means to obtain the three sub-groups of pseudopregnant rats with mild to severe PMDD symptoms. Then, the RNA-seq results of the three sub-groups were processed by the two-step feature selection method to obtain the important genes related to PMDD symptoms. Eventually, six supervised machine learning diagnostic systems built by the behavioral test results and the important genes were evaluated by a new group of rats undergoing the same experimental schedule as the original pseudopregnant rats. The feasibilities of the six systems were compared by their model and prediction accuracies.

Materials and methods

Animals

Eight- to 10-week-old female Wistar-Imamichi rats were obtained from the Imamichi Institute for Animal Reproduction (Tsuchiura, Japan) and acclimatized for 1 week before the experimental schedule. Animals were maintained under a 12/12 h light/dark cycle (lights on at 07:00), 3-4 rats per cage, and given *ad libitum* access to

food and water. The estrus cycle stage was accessed daily by vaginal smear for two cycles, and only the rats with regular cycles were included in the following experiments. The estrous status included proestrous, estrous, metestrous, and diestrous (proestrous: nucleated epithelial cells, estrous: non-nucleated cornified cells, metestrous: cornified cells with leukocytes, diestrous: large number of leukocytes). All experiments were conducted in accordance with the Guidelines for the Care and Use of Laboratory Animals, Graduate School of Agriculture and Life Sciences, The University of Tokyo.

Pseudopregnancy

Pseudopregnancy was induced by mechanical vaginal-cervical stimuli with a glass rod at 18:00 of the proestrous stage. The day of stimulation was designated as day 0 and pseudopregnant status was assessed daily using vaginal smear. After day 0 as proestrous and day 1 as estrous status, the field of microscope became similar to diestrous status which was filled with leukocytes until day 16-18. Only the rats with pseudopregnant period until day 16 were included in the following experiments.

Experimental schedule on the test days

On days 0, 8, and 16 of the pseudopregnant period, three behavioral tests of OFT, EPM and FST, hippocampus sampling, and serum sampling were performed (n=6–7/day). Day 0 was chosen since some of the patients had PMDD onset in the ovulation period. The treatment with ovulation inhibitors could alleviate the symptoms (Ryu & Kim, 2015). The animals undergoing experimental schedule on day 0 were not induced pseudopregnancy. The operation of the three behavioral tests was identical to that in the previous study. The time schedule of the test days shows in figure 3-1.

Behavioral test result: analysis by DLC

The analysis of behavioral test results was performed by DLC followed by customized codes written in Python. The thorough install and usage instructions of DLC is in Nath *et al.*, 2019 and the website of Mathis Laboratory: http://www.mackenziemathislab.org/deeplabcut.

- DLC

DLC 2.1.9 was used to track the points of interest (figure 3-2). All training was operated on Google Colaboratory, and label refinements were processed through GUI. All training videos were randomly selected. The OFT network was trained using 40 frames from 12 videos for 395,000 iterations, 0.00103 loss, and a 0.02 learning rate (lr). After label refinement, the second training ended with 650,000 iterations, 0.00088 loss, and a 0.002 lr. The EPM network was trained using 30 frames from 12 videos for 220,000 iterations, 0.00121 loss, and a 0.02 lr. After label refinements, the third training was ended with 430,000 iterations, 0.00148 loss, and a 0.02 lr. The FST network was trained using 40 frames from 20 videos for 785,000 iterations, 0.00077 loss, and a 0.001 lr. After label refinement, the second training ended with 700,000 iterations, 0.00111 loss, and a 0.002 lr. The x, y coordinates of each frame of each rat generated by DLC were then calculated into parameters in later analyses. The training python script was an official script for Google Colaboratory Notebooks of DeepLabCut Toolbox: https://colab.research.google.com/github/AlexEMG/DeepLabCut/blob/master/examples/COLAB_YOURDATA_TrainNetwork_VideoAnalysis.ipynb. Outliers were checked and fixed by trajectory plots generated by DLC and customized python scripts.

- DLC coordinates analysis

Customized python scripts were used to convert x, y coordinates into the required parameters. For the OFT, points of interest included the head center, body center, and tail base. The total length of the head center to body center and body center to tail base shorter than the body length when the rat rested was used for the calculation of rearing. For the EPM, points of interest included the head center and body center. Only the frames with the head center outside of the open arms while the body center was inside the open arms for over 30 frames (1 s) were counted as head dip behavior. For the FST, points of interest included the head center, body center, tail base, left side of the body center (lbc), and right side of the body center (rbc). The percentage of immobile time was determined by the rate of change of the polygon area formed by head center, lbc, tail base, and rbc. When the rate of change was below 60 px^2/frame, the animal was considered immobile. Total distance traveled and percentage of time spent in specific areas were calculated using coordinates of the body center in all tests. The customized python scripts for the three behavioral tests were uploaded to Github: https://github.com/RoyKudo/PMDD/tree/main/Behavioral%20test%20analysis.

Behavioral test result: clustering approach

To separate the animals with different syndrome severities of anxiety- and depression-like behaviors, a clustering of the behavioral test results was performed using unsupervised machine learning implementing the k-means algorithm. The number of centroids was determined using the evaluation metrices of elbow method and silhouette score method. In elbow method, the number (k) which flattens the line in the chart is considered as the optimal number of centroids. In silhouette score method, the

number (k) which obtains the highest score is considered as the optimal number. Features for clustering were the most commonly used parameters of the three behavioral tests, including time spent in the center zone in the OFT, time spent in the closed arms in the EPM, and immobile time in the FST. Python script: https://github.com/RoyKudo/PMDD/blob/main/Machine%20learning/k-means%20clustering.py.

The statistical significance of clusters in k-means was already tested by evaluation metrices, and thus the results generated were not calculated p-value or other additional measurements of statistically significant difference. Moreover, machine learning is hypothesized that it operates by data in population but not by a small number of samples. Therefore, the hypothetical tests after machine learning algorithms are not usually necessary (Mehryar, Afshin, & Ameet, 2018).

Hippocampus sampling

Each rat was decapitated and right or left hippocampal sample was taken. Hippocampal tissues were separated from the brain immediately after cervical dislocation, and the RNAs were extracted using the following procedure. Hippocampal tissue was homogenized in 1 ml Trizol buffer with stainless beads in cell-destructive equipment (Shake Master) at 75 xg, 5 min, and centrifuged (13,860 xg, 3 min, 4 °C). Next, the supernatant was added 0.2 ml chloroform, vortexed, and centrifuged (13,860 xg, 15 min, 4 °C). The new supernatant (0.4 ml) was added in the equivalent of isopropyl alcohol, vortexed, and centrifuged (13,860 xg, 15 min, 4 °C). The fluid was removed, 0.5 ml 70% EtOH was added, and centrifugation (13,860 xg, 10 min, 4 °C) was performed to wash the RNA pellets. After drying, 15 µl sterilized diethyl

pyrocarbonate (DEPC)-treated water was added, and the concentration and quality of RNA were measured by UV-Vis spectrophotometry in Thermo Scientific NanoDrop One instrument software. RNAs in all the samples were checked with the ratio of absorbance at 260 nm and 280nm around 1.9. The extracted RNAs were used for RNA-seq.

cDNAs were then synthesized as follows: 2 μg of RNA was added 1 μl Oligo (dT) primer (Invitrogen Item No. N8080128), followed by the use of SuperScript II Reverse Transcriptase (Invitrogen Item No. 18064071) according to the manufacturers' instructions. cDNA was used to perform quantitative polymerase chain reaction (qPCR) to measure the mRNA expression levels of the selected factors. The cDNA from one sample was diluted to 10, 100, and 1000 folds to serve as standards. The cDNA from all the samples were diluted to 100 folds and added in 0.5 mM of forward and reverse primers by the following recipe: 5 μl 100-fold cDNA, 4 μl sterilized ultrapure water, 0.5 μl forward primer, 0.5 μl reverse primer, 10 μl SYBER qPCR Mix. The total volume was 20 μl. The qPCR reaction condition was 95℃ 10 min. followed by 45 cycles of 95℃ 15 sec., 60℃ 15 sec., and 72℃ 1 min. All the expression levels of the samples were the relative levels compared to the house keeping gene of hypoxanthine-guanine phosphoribosyl transferase (HPRT).

- Processing of RNA-seq raw data

Total RNA with more than 2 μg of 12 samples was used, followed by Poly(A) selection for rRNA removal and sequence analysis by the DNBSEQ-G400 sequencing platform. Raw reads were trimmed, mapped, and annotated to the rat genome (rn6). All procedures were performed according to an official tutorial provided by Galaxy

(Reference-based RNA-Seq data analysis: https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html). Here is the brief summary: FastQC was used to perform the quality check. After trimming by Cutadapt, HISAT2 was used for mapping, and featureCounts was used for annotation. KEGG pathway analysis and GO term enrichment analysis were performed using DAVID.

Serum sampling

Each rat was decapitated and the blood were taken from jugular vessels close to the rupture surface. Serum was separated from the whole blood by centrifuge (672 xg, 15 min, 4 °C) after 1-hour still standing at the room temperature.

- Enzyme-linked Immunosorbent Assay (ELISA)

ELISA of serum P4, estradiol (E2), and ALLO were performed. E2 was also included due to its relationship with anxiogenic effect of P4 and ALLO, and chronic immune response during the estrous cycle (Bekhbat & Neigh, 2018; Costa, Spence, Smith, & Ffrench-Mullen, 1995; Laconi, Casteller, Gargiulo, Bregonzio, & Cabrera, 2001). Serum was diluted to 800X and the P4 concentration was measured by Progesterone ELISA Kit (Item No.582601, Cayman), according to the manufacturers' instructions. E2 was measured by 5X diluted serum and serum 17β-estradiol enzyme immunoassay kit (Catalog Number KB30-H1, Arbor Assays). ALLO was measured after extracted and by the use of allopregnanolone enzyme immunoassay kit (Catalog Number K061-H1, Arbor Assays). The extraction steps of serum ALLO are as follows. 100 µl of serum sample was put into glass tube with 900 µl of sterilized ultrapure water and 2 ml of diethyl ether. The tube was shortly vortexed 4-5 times, and put into -80 ℃

freezer for 20 – 30 min. to stratify. Then, the liquid supernatant with steroids and diethyl ether was put into another glass tube and the diethyl ether was evaporated by dry heat cabinet of 55 ℃. The steps from adding in 2 ml of diethyl ether were repeated 2 times to improve the extraction efficiency. After the evaporation of diethyl ether, 500 µl of ELISA assay buffer were added to dissolve the steroids, and this solution was used as the sample fluid for the following ELISA.

Feature selection

Feature selection was performed by an original two-step feature selection method. First, the annotated genes from the results of RNA-seq were identified by the following three criteria: (1) genes with mean expressions increasing/decreasing by the order of severities of symptoms; (2) genes without zero in their expression; and (3) genes that were not in the unknown list in DAVID. Then, a filter feature selection method composed of Linear SVM and SelectKBest in GridsearchCV cross-validation were used. Python script of the filter feature selection method: https://github.com/RoyKudo/PMDD/blob/main/Machine%20learning/Feature%20selection.py.

Supervised machine learning models

Six classifiers of different supervised machine learning algorithms were built, including Linear SVM, RF, Bag SVM, NN, Light GBM, and XGBoost. (1) Linear SVM: a linear model that creates a line or a hyperplane to separate data. (2) RF: an ensemble learning method grouped by multiple decision tree predictors that uses a tree-like structure to separate the data. (3) Bag SVM: another ensemble learning method

consisting of SVM models in a bagging classifier, fitting base models on random subsets of data and making final predictions by aggregating or voting on the predictions of the base models. (4) NN: a model with structure imitating the human brain, having input and output neurons connected by hidden layers and weighted synapses and giving results by learning from former layers and the ability to tune itself. (5) Light GBM: a gradient boosting machine based on a leaf-wise decision tree algorithm that has the ability to reduce more loss and perform fast. (6) XGBoost: a model similar to Light GBM but based on a level-wise algorithm, which is good for optimization but rather slow.

All classifiers were fed with the qPCR results of the identified genes from the feature selection step. With the ability to deal with missing values, Light BGM and XGBoost used raw data with 10.76% missing values. Other classifiers used the imputed data processed by the Multivariate Imputation by Chained Equation (MICE) imputation method. The model accuracies were evaluated by Leave-One-Out cross-validation, which runs validations across the number of samples to provide reliable and unbiased estimations for small datasets.

Code availability

All code files for analysis are available at: https://github.com/RoyKudo/PMDD.

Results

Serum P4 concentration of the pseudopregnancy rats.

The serum P4 concentration during the pseudopregnant period of the is in figure 3-3. The highest P4 concentration showed on day 8 with the number of 106 ± 20 ng/ml,

which was a result more comparable to the highest level in natural pregnant period of 130 ng/ml than that of the m-PWD rats in the previous study.

Clustering according to the results of the behavioral tests.

Due to the individual variety of PMDD onset and progression, all the data from days 0, 8, and 16 were gathered together and subjected to clustering using the k-means algorithm. The number of the clustering were determined “five” since it was the elbow of the curve in the graph of elbow method (figure 3-4 a) and obtained the highest score in silhouette score method except the number of two (figure 3-4 b). The clustering result by k-means with five clusters is in figure 3-4 c (n=55). Cluster 1, 2 and 3 (cluster green, orange, and red) were the three groups showed both anxiety- and depression-like behaviors progressing from mild to severe. Cluster 4 and 5 (cluster blue and purple) were the two groups showed only anxiety-like behavior but no depression-like behavior. To investigate the symptoms of anxiety and depression at the same time, only cluster 1, 2 and 3 were used for later analysis (n=32) (figure 3-5 a). The results of the main parameters of the three behavioral tests indicated that both anxiety- and depression-like behavior progressed from mild to severe in the green, orange, and red clusters (figure 3-5 g-i). These three clusters were referred to as clusters 1, 2, and 3. Figure 3-5 b-f show the other ancillary parameters in the behavioral tests. Cluster 3 was relatively low in both total distance traveled in the FST and number of head dips in open arms in the EPM, and relatively high in total distance traveled in the EPM at a degree similar to that of cluster 2. Regarding the results of total distance traveled and number of rearing in the OFT, the three clusters were at about the same level. The distribution of cluster 1, 2, and 3 on day 0, 8, and 16 is in table 3-1. For the total number of individuals in the three

clusters in all the three test days, cluster 3 had the fewest number of four, and cluster 2 had the highest number of 17. The individuals of cluster 3 were showed on day 0 (n=2) and 8 (n=2) but not on day 16.

Serum P4, E2, and ALLO concentration on the test days.

Serum P4, E2, and ALLO concentration on day 0, 8, and 16 are in figure 3-6. In figure 3-6 a, although mean P4 concentration was slightly increased from cluster 1 to 3, there were still some individuals with higher P4 concentration in cluster 2 than those in cluster 3. In figure 3-6 b, serum E2 concentration showed variously especially in cluster 2; while cluster 3 still had the highest mean P4 concentration. In figure 3-6 c, mean serum ALLO concentration was significantly increased from cluster 1 to 3 with the difference between cluster 2 and 3 was larger than that of cluster 1 and 2.

Determination of PMDD diagnostic factors using the original two-step feature selection method.

After separating all the individuals into three clusters, RNA-seq was performed to investigate the gene expression level in the hippocampus of each cluster, and 32,545 genes were obtained by annotation. Next, to identify the genes most related to the change in PMDD symptoms, a feature selection model consisting of SelectKBest, Linear SVM, and Grid Search CV was performed. However, the result of RNA-seq was a small dataset with few samples (n=12) but high dimensions (n=32,545). As a result, no matter how the model was tuned, the accuracy remained low and the number of selected genes remained high. Therefore, the original two-step feature selection method was performed.

The first step in the method efficiently reduced the number of genes from 32,545 to 5047. The KEGG PATHWAY database and the results of the GO term enrichment analysis of the 5047 genes are shown in figure 3-7 and 3-8. Then, in the second step, the 5047 genes were inputted and iterations were run until only one gene remained (table 3-2). Still, the result of the multi-class classification of clusters 1 to 3 was unsatisfying. Although the number of genes remaining dropped sharply (from 5043 to 6 in the second iteration), the low accuracy persisted (69.40% in the fourth iteration). Thus, feature selection for a binary classification of putting cluster 1 with cluster 2 versus cluster 3 was also performed. This time, the number of genes remaining gradually declined, and the accuracy reached 100.00% in the range of 48 to 16 remaining genes. After that, the only gene remaining after four iterations in the multi-class classification of clusters 1 to 3, and the 16 genes that maintained the accuracy at 100.00% with the smallest number of genes remaining in the binary classification of clusters 1+2 and cluster 3 were then identified. The list of the identified 17 genes is also shown in table 3-2.

To confirm the RNA-seq results, qPCR was performed to verify the gene expression levels of the 17 identified genes (figure 3-9). Figure 3-9 a shows a scatterplot of the RNA-seq results for two of the identified genes, Marcks and Hhex. From this figure, the distribution of cluster 3 was far from those of clusters 1 and 2 could be verified. However, clusters 1 and 2 both had one data point for each group separated from the other data points, making the separation of clusters 1 and 2 difficult. A similar distribution can also be seen in figure 3-9 b, which shows a scatterplot of the qPCR results for Marcks and Hhex. Based on the scale difference of the axes, this was not identical to figure 3-9 a; however, it showed similar characteristics in clusters. For this

reason, the qPCR results of the 17 identified genes were used to perform the second screening of the RNA-seq results and subsequent tests.

Verification and comparison of the supervised machine learning models.

Six classification models were built using the 17 identified genes. The model accuracies of both multi-class (three-class) classifications of cluster 1 to 3 and the binary (two-class) classification of cluster 1+2 versus 3 are listed in table 3-2. As shown in the table, NN had the best model accuracies for both classifications (80.00% and 100.00%), followed by Bag SVM in the three-class classification (72.73%) and as good as RF and Bag SVM in the two-class classification (both 100.00%). On the other hand, Linear SVM and XGBoost had the lowest accuracies in the three-class classification (both 54.55%), and considerably lower accuracy in the two-class classification (72.73%) compared with the others.

Prediction performance of the models.

After constructing the models, the previous pseudopregnancy process and behavioral tests were performed again with a new group of rats (n = 26). The behavioral test results were standardized and used to predict the cluster of each rat by the same k-means model. The qPCR results of the 17 identified genes in this new group of rats were obtained to examine the prediction ability of the classifiers. The prediction accuracies are also shown in table 3-3. RF had the best accuracies (96.00% and 100.00%) in the three- and two-class classifications, respectively, while NN (78.00%) had the lowest in the three-class and Light GBM (91.00%) in the two-class classification. Figure 3-10 shows the confusion matrices of RF and NN for the three-

class predictive classification. RF only misclassified one cluster 2 data point to cluster 1. However, NN seemed to be unable to distinguish the clusters by putting all the data points into cluster 2. The confusion matrices of the other classifiers are shown in figure 3-11, and the outline of the methods and tools used in this study is shown in figure 3-12.

Discussion

In this chapter, pseudopregnant rats were chosen to imitate the hormone environment in the human luteal phase, and k-means were used to cluster the behavioral data to form the three groups of rats with mild to severe anxiety- and depression-like behaviors. Besides, DLC was successfully applied for the analysis of behavioral test videos. Then, the original two-step feature selection method was used to identify 17 important genes from the hippocampus, and these 17 genes and the foregoing clustering result were used to build the six supervised learning classifiers. Among the classifiers, RF was the best classifier for the present task with the best prediction accuracy in both two- and three-class classifications.

The method of pseudopregnancy followed by k-means is possible to differentiate PMDD and other premenstrual illnesses.

In the clustering result, there were five clusters with two of them only having anxiety-like behavior but no depression-like behavior. It is an interesting result as it shows that pseudopregnancy is a method providing endogenous environment inducing different physiological changes in different individuals, and k-means has the potential to be used for the separation of different diseases. Individual difference is a huge difficulty for studying PMDD and other mood disorders (Eisenlohr-Moul *et al.*, 2020). By use of

pseudopregnant animals and k-means, animals with different properties in syndromes were successfully generated and differentiated. Especially that the property of unsupervised learning, the number of clusters being determined by evaluation metrics or data-driven training, is important for investigating new facts or rules which have not been noticed before. From the result in the present study, it is possible to find the key factors which can separate the patients with anxiety-like behavior and depression-like behavior, and the patients only with anxiety-like behavior. K-means makes it possible to differentiate PMDD and Anxiety disorder.

It is also possible to add in more parameters, such as other behavioral test results or environmental factors, to investigate more about the symptoms and their related factors. Since k-means is such a powerful tool, future studies and databases which can gather more information and experimental results of mood disorders sharing similar pathological pathways are expected. In this way, more details in these disorders can be recognized, and more precise diagnoses and treatments can be provided to the patients.

The three sex hormones of P4, E2, and ALLO cannot be the diagnostic criteria of PMDD symptoms.

In figure 3-6, the relationship among sex hormones, clusters, and the pseudopregnant period was revealed. Serum P4 tended to affected by the progress of the pseudopregnancy period but not the clusters. Serum E2 seemed to be higher in the beginning of the pseudopregnant period and had larger individual difference. Serum ALLO had a relatively clear increasing tendency along with the clusters than the other two. These results agree with the conclusions that it is ALLO but not P4 or E2 has the

closest relationship with PMDD (Timby *et al.*, 2016). However, the great variation makes none of them suitable for being the diagnostic criterion of PMDD symptoms.

The hippocampus was chosen due to its relationship with PMDD mechanism.

Amygdala, prefrontal cortex, and hippocampus are the most studied brain areas studied the most in PMDD research. Amygdala and prefrontal cortex have been more studied by the use of resting-state functional magnetic resonance imaging (fMRI) scan in PMS/PMDD patients. One of the studies indicated that amygdala activity is related to anxiety proneness of PMDD (Gingnell, Morell, Bannbers, Wikström, & Poromaa, 2012). Another study demonstrated that the volume of amygdala is greater, and the functional connectivity (FC) between amygdala and other brain areas are altered in the PMS patients. For instance, the FC increases between amygdala and prefrontal cortex, but decreases between amygdala and hippocampus (Deng *et al.*, 2018). On the other hand, hippocampus has been more studied in its relationship with PMDD mechanism. The receptors of P4, E2, and GABAR in the hippocampus are related to the functional changes in PMDD patients. These receptors are sensitive to the fluctuation of the corresponding hormones, affect the structure and composition of hippocampus, and finally bring out PMDD-related symptoms (Wiklund, 2017). A previous study even indicated that the proliferation of hippocampal neuron progenitor cells can be promoted by ALLO, which may be one of the reasons of emotional dysfunction in PMDD (J. M. Wang, Johnston, Ball, & Brinton, 2005). Due to the essential role in PMDD symptoms, such as emotional, cognitive, learning, and memory function changes, hippocampus was chosen to be the research target in the present study.

As for the effect of the mainstream treatment of PMDD, SSRIs, in the hippocampus, it is still a controversial issue. There is not yet a study mainly focusing on the direct relationship between SSRIs and hippocampus. There were studies about SSRIs normalizing hippocampal output to treat mood symptoms in major depression disorder (Dale *et al.*, 2016). There was also a study about low E2 state being able to increase hippocampal serotonin transporter activity (Bertrand *et al.*, 2005). However, there was also another study showing that progesterone withdrawal (one of the hypothesized PMDD etiology) did not alter hippocampal serotonin level (Li *et al.*, 2012). Thus, further research for the effect of SSRIs in the hippocampus under PMDD situation is still needed.

Most of the hippocampal 17 key genes are associated with PMDD symptoms.

For the 17 selected genes, they are not closely related and do not gather in the same pathway. They also have a wide range of mRNA expression levels from tens to thousands of digits. However, some of them or their families have already been suggested to be involved in the development of PMDD symptoms such as anxiety and depression, or to be able to be used to differentiate mood disorders. For example, the vasopressin pathway is known to play a role in anxiety and depression (Morales-medina, Witchey, & Caldwell, 2016). The nuclear factor-κB (NF-κB) family has been proven to be a mediator of depressive behavior (Koo, Russo, Ferguson, Nestler, & Duman, 2010). Family with sequence similarity 65, member C (Fam65c) has been indicated to be involved in distinguishing major depressive and anxiety disorders (Rao, Yin, Xiang, & So, 2020). These findings suggest that our method can identify factors that have actual relationships with and close to PMDD. The list of the factors and their

previous studies related to mood disorders is in table 3-4. Furthermore, some of the 17 genes has also been studied their relationship with the characters in the PMDD hypothesized pathological pathway. The activation of chemokines and cytokines, such as NF-κB, can be prevented by ALLO (Balan, Beattie, O'Buckley, Aurelian, & Morrow, 2019). Cysteine-rich angiogenic inducer 61 (Cyr61) has been suggested as a progesterone down-regulated gene (Hanekamp *et al.*, 2003). G protein-coupled receptor 101 (Gpr101) was found expressed by a subgroup of GABAR in ventral tegmental area (VTA) and substania nigra pars compacta (SNC) of the limbic system (Paul, Tossell, & Ungless, 2019). In addition, due to the wide range of expression levels of the 17 key genes, the genes with low expression levels could be hard to be measured by the use of qPCR. Thus, a triplicate experiment is better to be operated and the primer should be selected carefully. The primers of the 17 genes used in qPCR are listed in table 3-5.

RF is the classifier with the best performance in the present task.

Regarding the results of the six classifiers, the model accuracies determined by Leave-One-Out cross-validation did not always have absolute positive relationship with the predictive accuracies. For instance, RF had a rather low model accuracy of 63.64%, but the highest prediction accuracy of 96.00% in the three-class classification. XGBoost had a model accuracy of 72.73%, but achieved a prediction accuracy of 96.00% in the two-class classification. Conversely, NN had higher model accuracies in the three-class classification (80.00% and 100.00%), but rather low ones (78.00% and 96.00%) in the two-class classification. It seems that NN could not learn patterns from the models correctly when there were only a small number of samples. By considering all the performances, RF is the best classifier with the highest prediction accuracies in both

classifications, and the opposite, NN, might have a problem with overfitting. Because of the operating complexity, NN is a remarkable classifier but considered to be prone to overfitting, especially when using small size dataset (Lawrence, Giles, & Tsoi, 1997). If there is more data inputted into the present NN model in the future, it is possible that NN will have better predictive performance.

The operating principle of classifier is more influential to the classification result than the missing data imputation method.

The way to handle missing value is an issue for classification tasks. In the present study, MICE was used to impute the missing values in the qPCR results. MICE is an imputation algorithm which imputes the blanks by referring to the values in other columns of features (van Buuren & Groothuis-Oudshoorn, 2011). By the use of random forest algorithm, it provides relatively correct result than other methods, such as imputation using mean, median, or constant values. Due to the ability of missing value tolerance, data imputed by MICE was not used in Light GBM and XGBoost in this study. However, these two classifiers had lower model accuracies, indicating that missing values actually had influence on the classification results even the rate of missing value was only 10.76%. At the same time, the performance of prediction of Light GBM and XGBoost were not that poor (91.00% and 96.00% in two-class classification), showing that the patterns of data still could be learned when there were about 10% missing values. It seems that the prediction accuracy is more about the operating principles of the classifiers than the missing value processing.

The AI diagnostic method in the present study is objective, improvable and promising.

Diagnostic machine learning techniques have been studied and developed in the past ten years. Heart disease, diabetes, hepatic and renal disease are the main targets. By using different algorithms and attributes, these AI tools have already been able to provide diagnostic accuracies around 95% and the accuracies are still increasing (Fatima & Pasha, 2017; Somnay *et al.*, 2017). On the contrast, the use of AI methods in psychological diseases is just started in the recent three years. The methods has been used for diagnosis of autism, monitor of mental health physiological indicators, or supplementary psychotherapies (Abbas, Garberson, Liu-Mayo, Glover, & Wall, 2020; Ćosić, Popović, Šarlija, Kesedžić, & Jovanovic, 2020). There is still only few of formal reports of AI application using biological factors in the diagnosis of psychological disease. However, the researchers interviewed by Harvard Business Review and The Verge believe that this field is a promising and exciting help for the suffering patients. Bringing in machines can be a new solution (Garg & Glick, 2018; Zarley, 2019). This is also the vision which the present study is striving for. When comparing to the existing diagnostic method in DSM-5, the new method in the present study is superior in the following two aspects. One is that it is an objective method. It provides the diagnostic standard by using the expression levels of measurable genetic factors but not subjective judgements. Another is that, it is a method which is improvable. By gathering data continuously, the current diagnostic system can be more useful and accurate.

In conclusion, in the end of the present study, the expression changes of the 17 genes in hippocampus were found being able to separate the individuals with different severity of PMDD symptoms, and RF can be the most suitable classifier for the current

task. Moreover, the original two-step feature selection method consisting of a manual selection followed by a filter method was presented and proved to have the ability to find the true factors of disease. It is expected to provide a new solution for dimension reduction problems in genetic studies.

Figures and tables

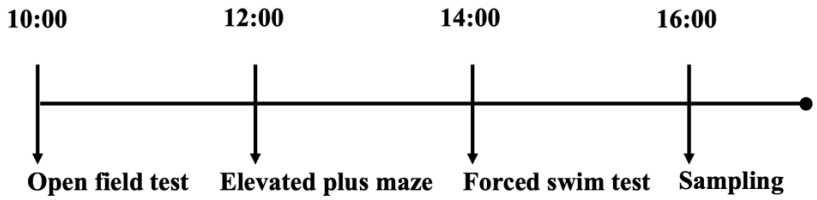


Figure 3-1. Experimental schedule of day 0, 8 ,16 in the pseudopregnant period.

The experimental schedule started from 10:00. Between each step, there were approximately 1-h interval. After the last behavioral test of FST, all the rats were decapitalized, and the hippocampus and serum were sampled immediately.

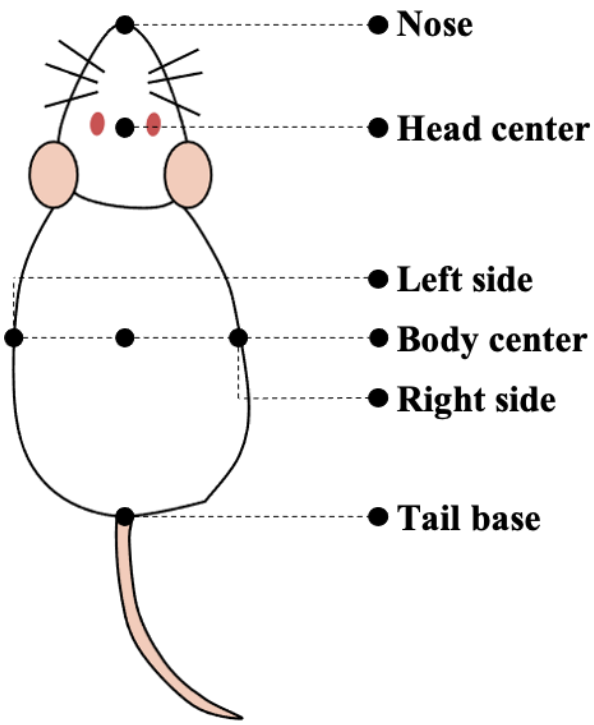


Figure 3-2. The points of interest used for the animal tracking in DLC.

In DLC, the points of interest have to be taught to the algorithm beforehand, and the algorithm will know what to capture in the video. OFT: head center, body center, tail base. EPM: head center, body center. FST: head center, body center, tail base, left and right side of the body center.

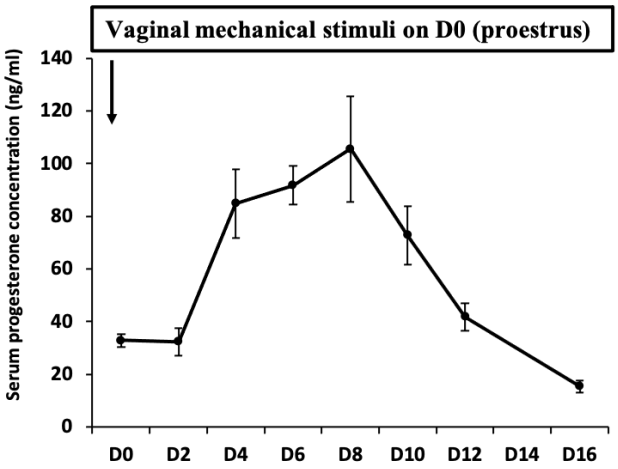
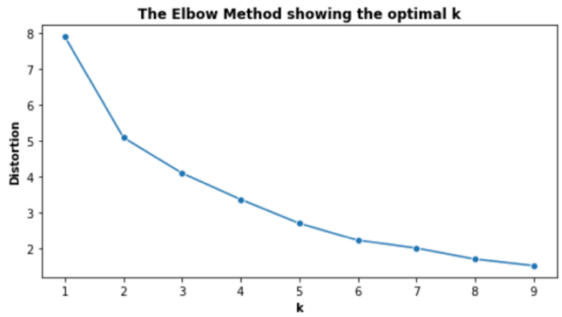


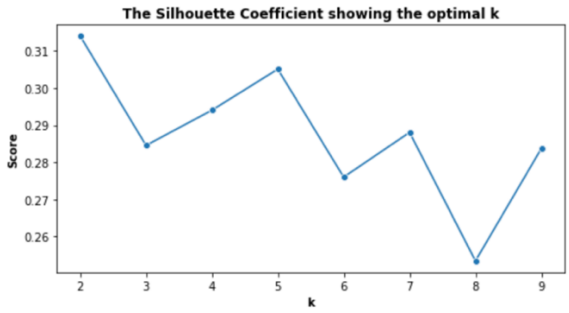
Figure 3-3. Serum progesterone concentration during the pseudopregnant period (n=4).

The pseudopregnant period was around 16 days. There was a peak on day 8 with the concentration of 106 ± 20 ng/ml.

(a)



(b)



(c)

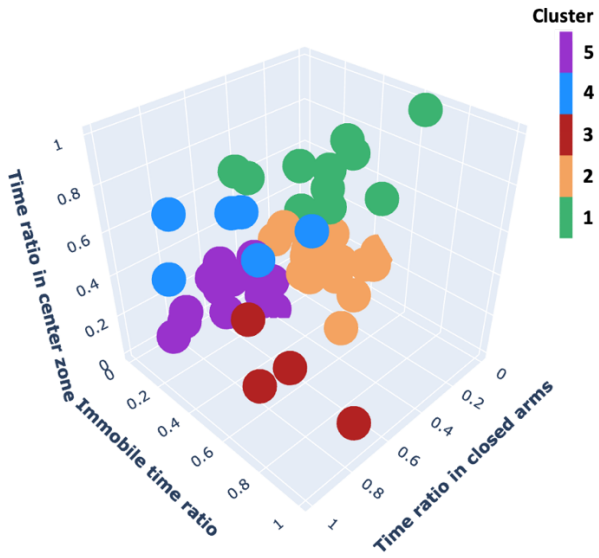


Figure 3-4. The plot of elbow method and silhouette score method, and the clustering result with five clusters.

(a) The evaluation result of elbow method. (b) The evaluation result of silhouette score method. (c) The clustering result with five clusters.

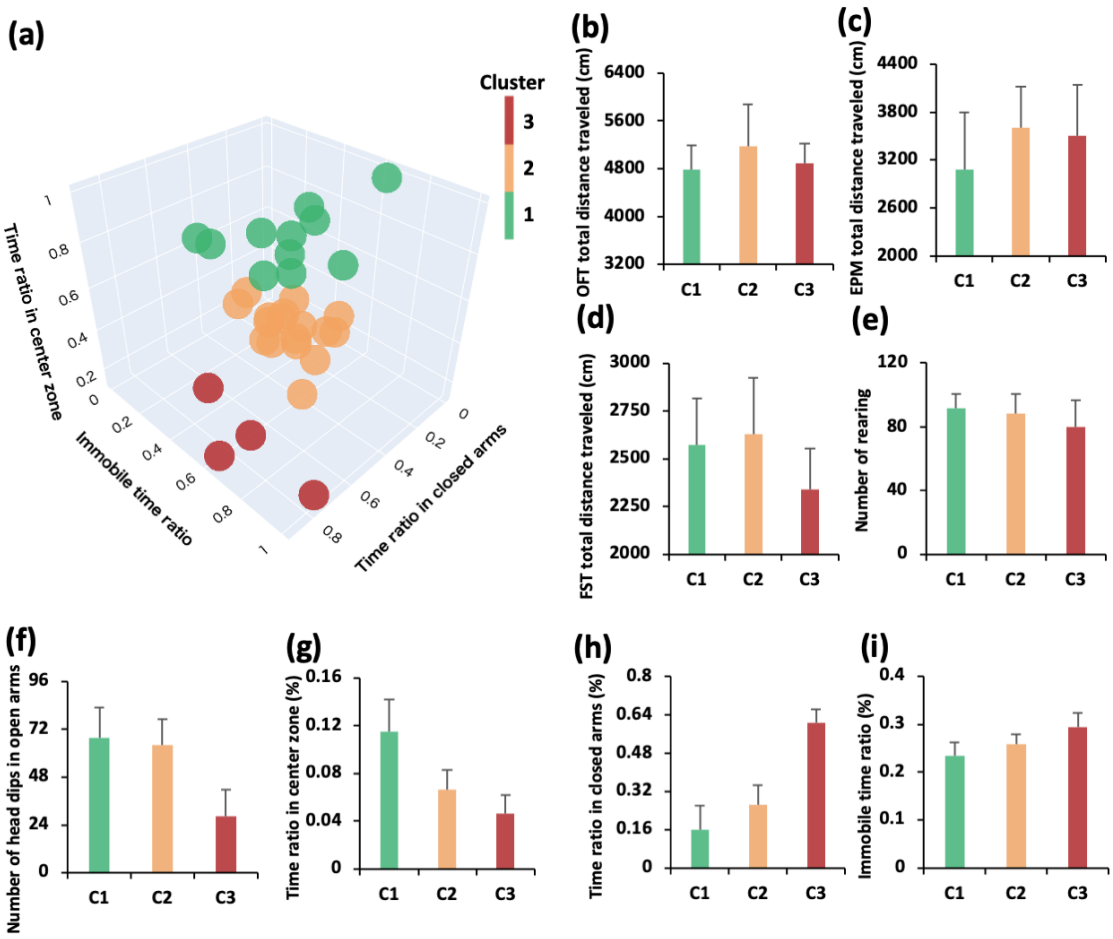


Figure 3-5. Clustering and behavioral test results of pseudopregnancy groups.

(a) k-means clustering result. (b-i) Behavioral test results of clusters 1–3. (b) Total distance traveled in the OFT. (c) Total distance traveled in the EPM. (d) Total distance traveled in the FST. (e) Rearing number in the OFT. (f) Number of head dips in open arms in the EPM. (g) Time ratio in the center zone in the OFT. (h) Time ratio in closed arms in the EPM. (i) Immobile time ratio in the FST.

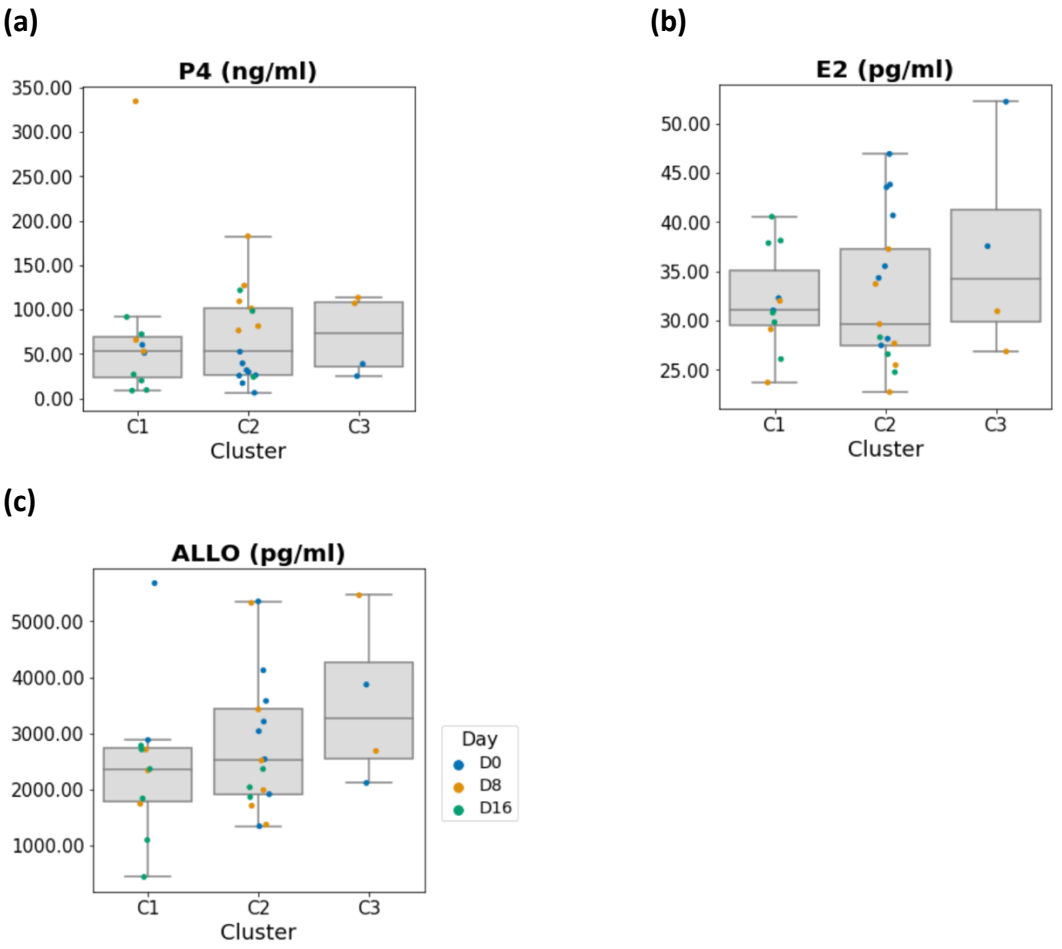


Figure 3-6. Serum P4 (a), E2 (b), and ALLO (c) concentration of cluster 1, 2, and 3 on pseudopregnant day 0, 8, and 16.

All the three hormones tended to increase in the order of cluster 1 to 3. However, the individual differences made great variations. None of them could be the diagnostic criterion of PMDD symptoms.

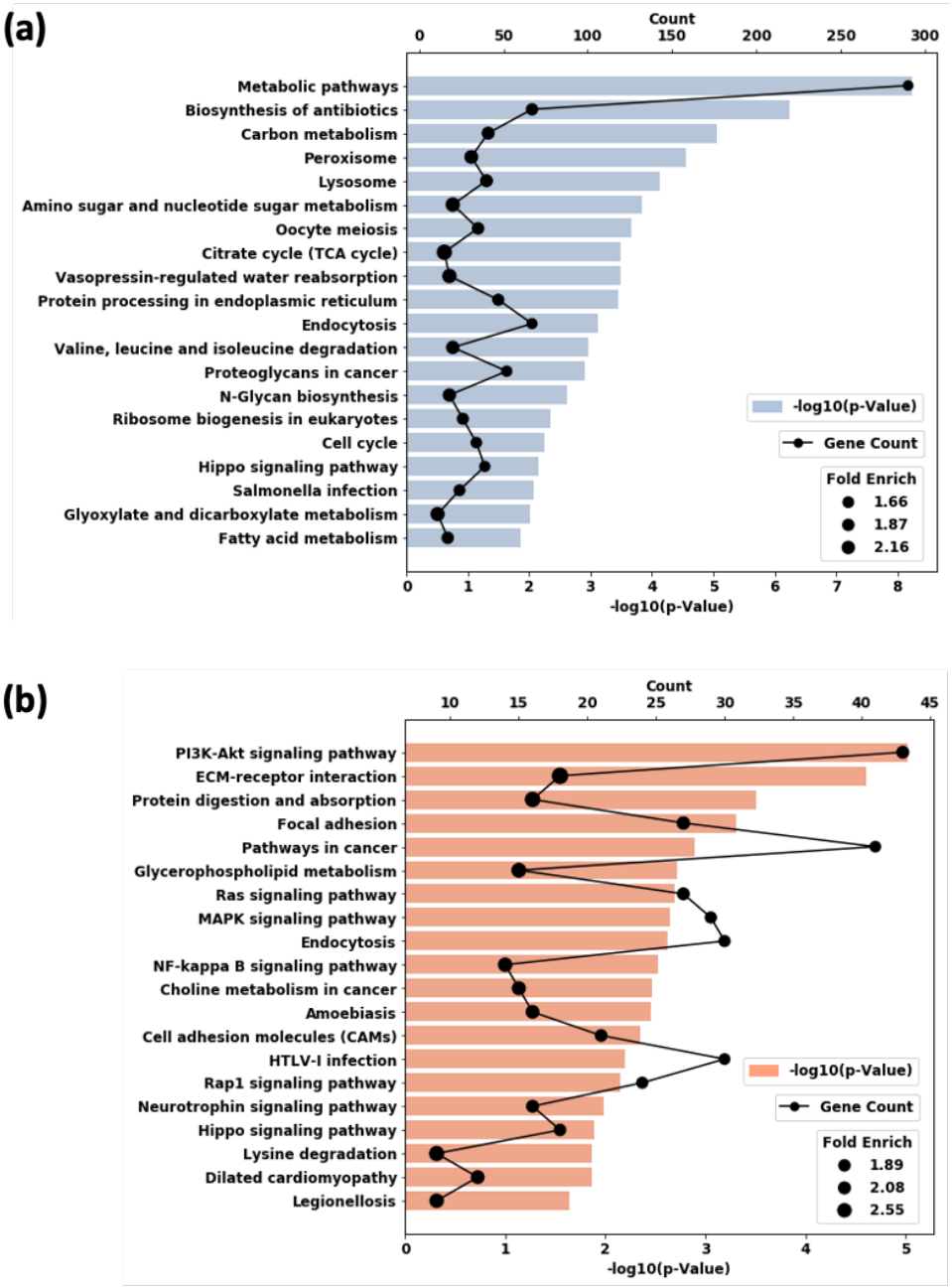


Figure 3-7. KEGG pathway analysis result for the selected 5047 genes.

Top 20 KEGG pathway results of (a) results of upregulated genes (3517 genes). (b) results of downregulated genes (1530 genes) of cluster 1 to 3.

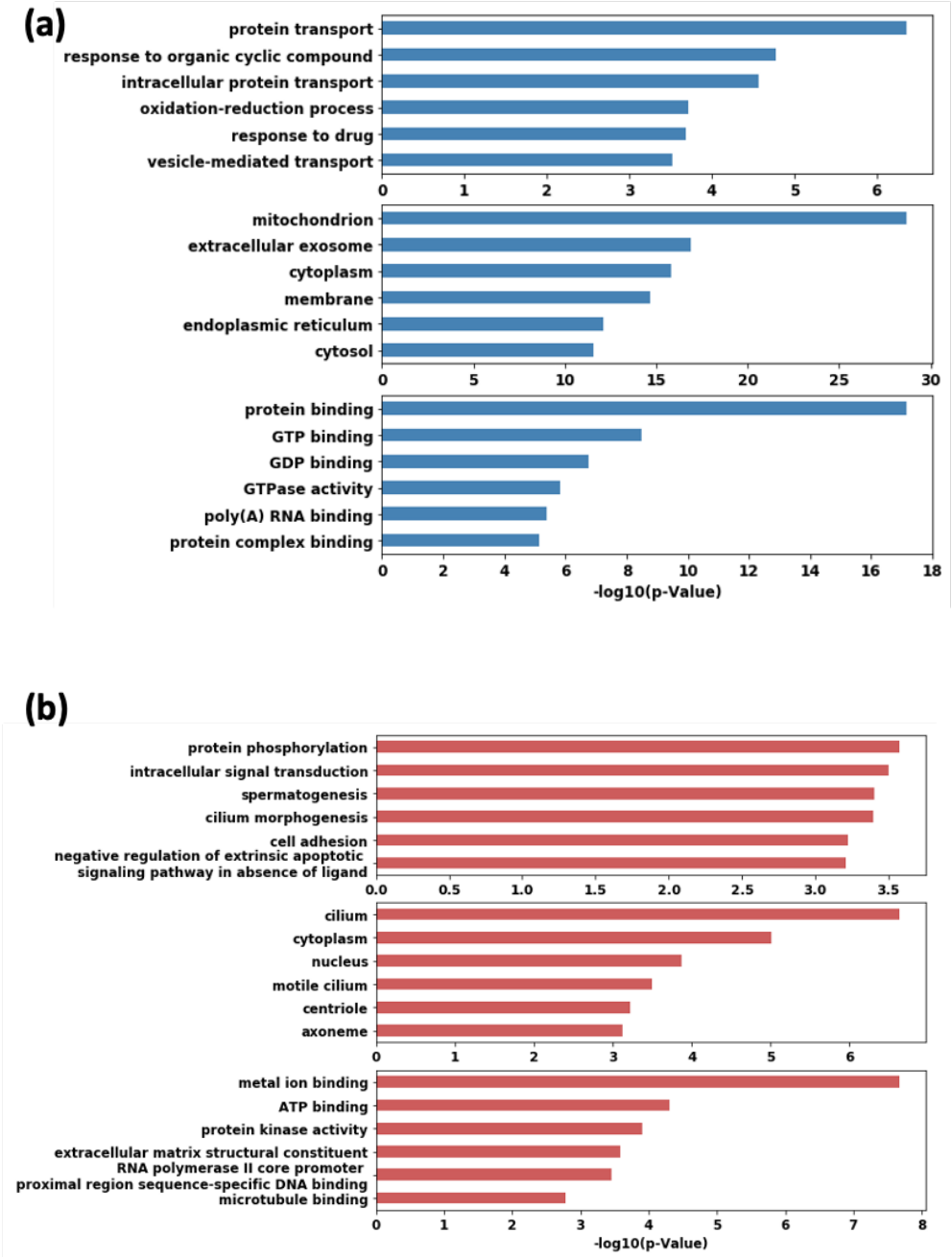
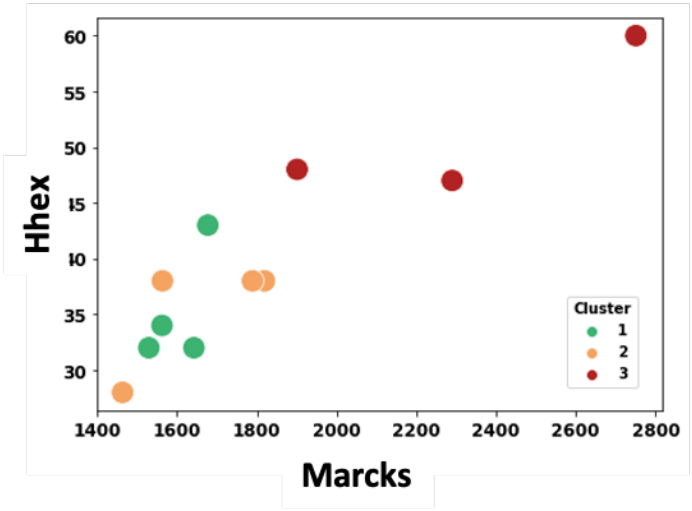


Figure 3-8. GO term enrichment analysis result for the selected 5047 genes.

Top six GO analyses of the biological process, cellular component, and molecular function of (a) results of upregulated genes (3517 genes). (b) results of downregulated genes (1530 genes) of cluster 1 to 3.

(a)



(b)

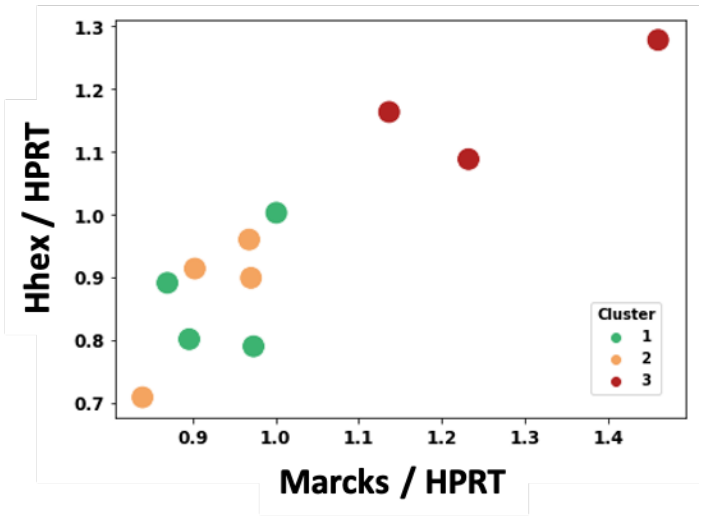
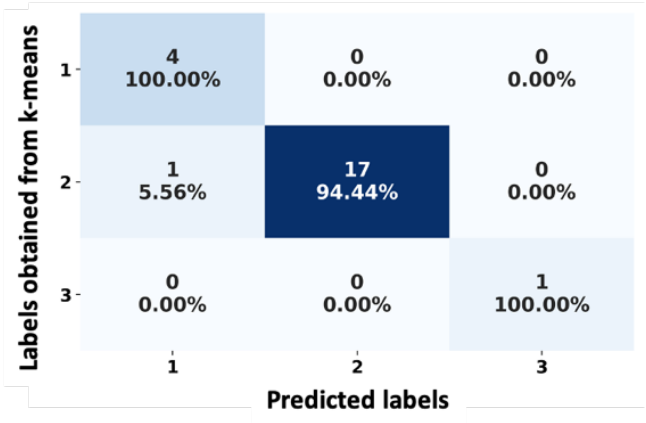


Figure 3-9. Scatterplots of Marcks and Hhex.

(a) Plot by results of RNA-sequencing. (b) Plot by results of qPCR. The scatterplots were used to compare whether the result of RNA-seq and qPCR were identical or not, and to observe the distribution of data. In the plots, cluster 3 could be separated easily. On the contrast, the separation of cluster 1 and 2 could be difficult.

(a)



(b)

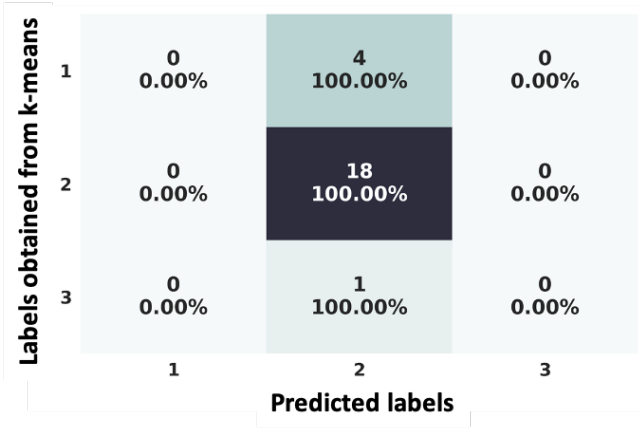


Figure 3-10. Confusion matrices of predicted data in C1 vs. C2 vs. C3 multi-class classifications.

(a) Results of RF. (b) Results of NN. RF only misclassified one cluster 2 data point to cluster 1. However, probably duo to the issue of overfitting, NN was unable to distinguish the clusters and put all the data into cluster 2.

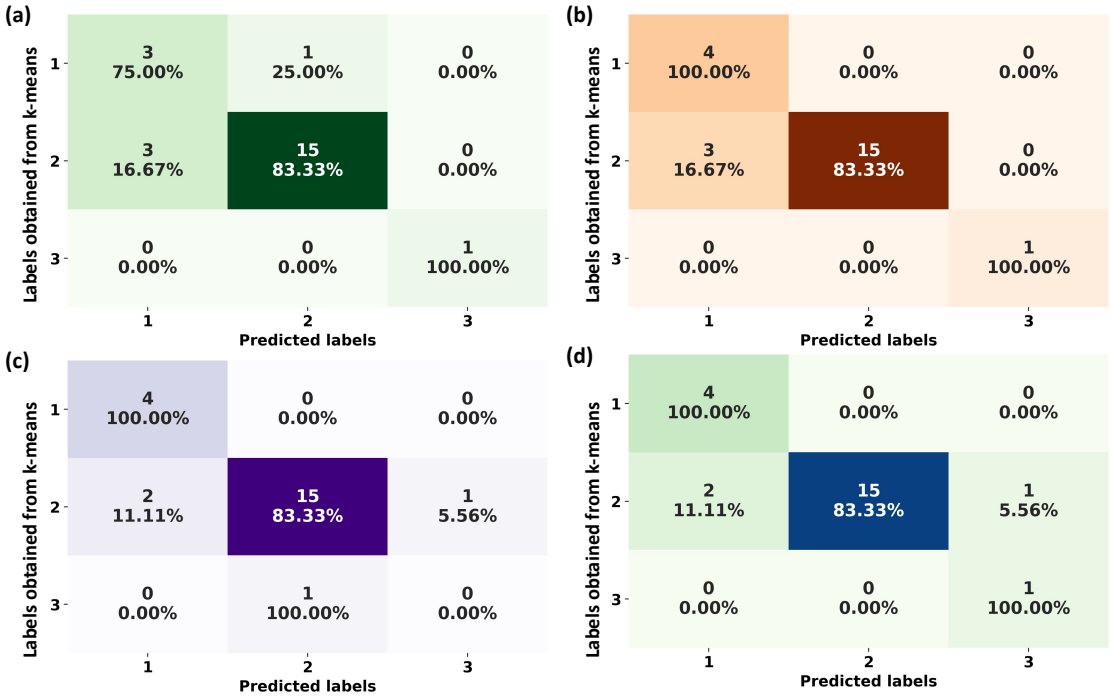


Figure 3-11. Confusion matrices for the predicted data in C1 vs. C2 vs. C3 multi-class classifications.

(a) Results of Linear SVM. (b) Results of Bag SVM. (c) Results of Light GBM. (d) Results of XGBoost. The algorithms other than RF and NN also struggled with the differentiation between cluster 1 and 2 mostly.

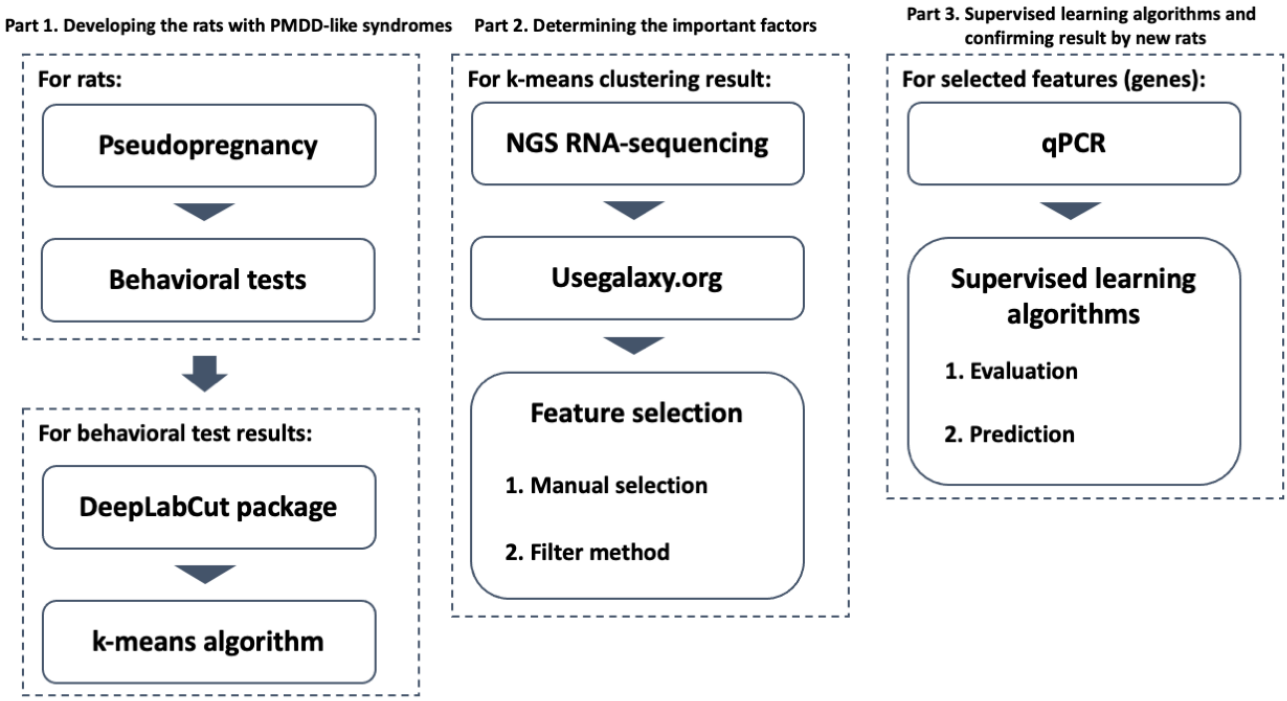


Figure 3-12. Outline of the methods and tools used in the present study.

	C1	C2	C3	3 clusters
D0	2	8	2	12
D8	3	6	2	11
D16	6	3	0	9
Total	11	17	4	32

Table 3-1. The distribution of cluster 1, 2, and 3 on pseudopregnancy day 0, 8, and 16.

There were more individuals of cluster 2 and 3 in day 0 and 8 than in day 16. Due to the fact of alleviation of PMDD symptoms before menstrual period, the results in the present study could be thought comparable to the real situation in the human patients.

Type	Iteration	Accuracy	Features remaining
C1 vs. C2 vs. C3	1	0.472	5043
	2	0.472	6
	3	0.667	2
	4	0.694	1
C1+C2 vs. C3	1	0.889	248
	2	0.889	179
	3	0.944	133
	4	0.917	60
	5	1	48
	6	1	40
	7	1	35
	8	1	31
	9	1	28
	10	1	25
	11	1	22
	12	1	20
	13	1	19
	14	1	18
	15	1	17
	16	1	16
	17	0.917	15
	18	0.889	13
	19	0.889	7
	20	0.861	1

Avpr2, Cd22, Cfap44, Cyr61, Efcc1, Fam65c, Gpr101, Gemin7l1, Hhex, Klhl41, Lrrc74b, Marcks, Nfkbil1, Nlrc4, Ppp1r42, Spc24, LOC108352928 (total: 17 genes)

Table 3-2. Feature selection processes and results of C1 vs. C2 vs. C3 multi-class classifications and C1+C2 vs. C3 binary classifications.

Due to the low accuracy in C1 vs. C2 vs. C3 multi-class classification, only one of the last gene of Gemin7l1 was selected. In total, 17 genes were identified as the most important gene set for PMDD according to cross-validation model accuracy.

Classifier	C1 vs. C2 vs. C3 Model Accuracy	C1 vs. C2 vs. C3 Prediction Accuracy	C1 + C2 vs. C3 Model Accuracy	C1 + C2 vs. C3 Prediction Accuracy
Linear SVM	54.55%	83.00%	90.91%	100.00%
Random Forest	63.64%	96.00%	100.00%	100.00%
Bag SVM	72.73%	87.00%	100.00%	100.00%
Neural Network	80.00%	78.00%	100.00%	96.00%
Light GBM	63.64%	83.00%	90.91%	91.00%
XGBoost	54.55%	87.00%	72.73%	96.00%

Table 3-3. Cross-validation model accuracy and prediction accuracy of C1 vs. C2 vs. C3 multi-class and C1+C2 vs. C3 binary classifications by 6 classifiers.

Linear SVM, Random Forest, Bag SVM, Neural Network used data after imputation. Light GBM and XGBoost used data with missing values (missing value percentage: 10.76%). Due to the difficulty of separating cluster 1 and 2, the model accuracies were relatively low in all the classifiers. However, all the prediction accuracies were higher than 75%. The prediction accuracies of Linear SVM, RF, and Bag SVM even reached 100% in the binary classification. The results demonstrated that it is possible to use supervised machine learning algorithms to build diagnostic systems for difficult diseases such as PMDD.

Gene name	Functions related to mood disorders	Research
Avpr2	The vasopressin pathway potentially contributes to stress-related disorders, including anxiety and depression.	Morales-medina, Witchey, and Caldwell, 2016
Cyr61	Related to acute immobilization stress (anxiety and depression)	Kurumaji *et al.,* 2008
Klhl41	Related to chronic unpredictable mild stress (CUMS) induced depression	Sun *et al.,* 2018
Nfkbil1	Related to stress induced anhedonia (depression)	Koo *et al.,* 2010
Fam65c	Can be used to separate Major Depression Disorder and Anxiety Disorder	Rao *et al.,* 2020
Marcks	Related to depressed suicide	Pandey *et al.,* 2003
Cfap44	Related to cognitive impairment in hippocampus	Song *et al.,* 2019
Hhex	Related to chronic restraint stress (anxiety and depression)	Gray *et al.,* 2014
Cd22	Related to Major Depression Disorder	Mellon *et al.,* 2016
Spc24	Related to disturbances in circadian rhythm, and indirectly related to multiple mood disorders including Major Depression Disorder and Anxiety Disorder	Polo *et al.,* 2017
Efcc1	Related to sexual differences in brain regions of Alzheimer's disease	Sun *et al.,* 2019
Lrrc74b	One of the schizophrenia-associated copy number variants	Campbell and Granato, 2020
Nlrc4	One of the inflammasome subtype of Nlrp3, involved in stress, depression, and comorbid illnesses	Iwata, Ota, and Duman, 2013
Gpr101	Related to higher immobile time in FST (depression)	Mantas *et al.,* 2021
LOC108352928	Uncharacterized gene	-
Ppp1r42	Little-known gene	-
Gemin7l1 (Gemin7)	Related to sexual-biased psychopathology	Phillips *et al.,* 2019

Table 3-4. The list of the 17 key factors and their previous studies related to mood disorders.

Gene name	Primer	Up or down regulated
Avpr2	F: ACCCTTCTTCCTCGTGCAG	Down
NM_019136.1	R: AGCAGCATGAGCAACACAAA	
Cyr61	F: GGATCTGTGAAGTGCGTCCT	Up
NM_031327.2	R: CTGCATTTCTTGCCCTTTTT	
Klhl41	F: CACTGAAGTCAATGACATATGGAAG	Down
NM_057191.1	R: CCCGAAGCATATCGAATCTC	
Nfkbil1	F: TCAAGGAGAAGGAACTGTGTGA	Up
NM_212509.2	R: GGTCCCCTTGAGCCTCTT	
Fam65c	F: ATCACAGAGTTGGGCACCAT	Down
ENSRNOT00000035463.5	R: TCAGAGTCCAAGGGGTTCC	
Marcks	F: GTCGCCTTCCAAAGCAAAT	Up
NM_001271090.2	R: AAAGTTGGCGTGCAGCTC	
Cfap44	F: GAGAGGAGCAAGACATGCAA	Down
ENSRNOT00000059680.4	R: CCGACTGGAAGCCAGAGTAG	
Hhex	F: AATCGCAGAGCTAAATGGAGA	Up
NM_024385.1	R: TCCAAACTGTCCAACTCATCC	
Cd22	F: TGTGGCCGTGGAGATAGATA	Down
NM_001107503.1	R: ACACATGATGGCTTGTCTGG	
Spc24	F: GACTGCTGGAGATGCAGGAC	Up
ENSRNOT00000045384.5	R: TCCTTCAGTTCAAGAAGGCTCT	
Efcc1	F: GAAGTTGGTGGATGTGCTACAA	Down
NM_001163921.1	R: CTTCCCCAGTGCTCTTTTTG	
Lrrc74b	F: CAGGGTCAACAATGTGTTAGAAGA	Down
ENSRNOT00000002557.7	R: GAAGACCCAACCCCAACTTC	
Nlrc4	F: CTTGAAAACTGGAACTATCCTGTG	Down
NM_001309432.1	R: TCTTCCAAGTTCTGATGAAAAAGA	
Gpr101	F: GAAAGGTGGCCAGACAGC	Down
NM_001108258.1	R: CAGCTCAATGTCTGCCCTAGTA	
LOC108352928	F: GTACCAGTCGGTGAGCAAAGT	Down
ENSRNOT00000082495.1	R: CTTCACAGCGAAGGGAAACT	
Ppp1r42	F: TCTTAAACCCAGAAAAGAAGAAACC	Up
NM_001127569.1	R: GTTTTTACAAAGAGAGAGGTCATCAA	
Gemin7l1 (Gemin7)	F: GGGCGTAAGAGCAGAGCTT	Up
XM_003749742.3	R: CACCGGCACAGGAATGAT	

Table 3-5. The 17 key genes, their primers used in qPCR, and the direction of regulation.

General discussion

In the present study, a diagnostic method for PMDD symptoms using AI tools and hippocampal factors in rats was successfully established. DLC was used for analyzing the behavioral test videos, and k-means for clustering the individuals by their severity of symptoms. Also, Usegalaxy.org was used to process RNA-seq raw data, and an original two-step feature selection method was established to find the set of the most important factor genes. The clustering result and the expression of the selected genes were then used to build supervised learning algorithms to form the new diagnostic system for PMDD symptoms (figure 4-1). In this chapter, issues of PMS/PMDD and the benefits/limitations of the diagnostic method in the present study would be discussed.

The possible evolutionary advantages of PMDD syndromes.

The emotional change in PMDD is now generally considered pathological and detrimental since it has been added into DSM-5 in 2013. On the contrast, there were also researchers approving that the occurrence of PMDD symptoms has evolutionary advantages. They conjectured that the anxiety or depression status in the luteal phase in women might be able to break down the infertile pair bonds of couples unsuitable for each other and be beneficial to reproduction (Morriss and Keverne, 1974; Wilson, 1992). There was also another assumption that the lower sexual desire and irritable mood might protect women from the increased vulnerability to infection during the luteal phase (Kinney & Tanaka, 2009). However, these hypotheses were criticized for either their unreplicable methodological flow or lack of direct evidence, making them failing to be widely accepted (Gillings, 2014; Ziomkiewicz-wichary, 2017).

The association among PMS/PMDD, postpartum depression (PD), and menopausal symptoms (MS) is still unclear.

PMS/PMDD, PD, and MS are the three female-specific mood disorders. Due to the difficulties in investigating the detail mechanism, PD and MS are also the diseases with limited understanding and treatment on a similar note with PMS/PMDD (Freeman, 2002). The comparative studies are still rare, making the only clear experimental conclusion at this time is that the hormone-related etiology is shared in the three diseases (Lee *et al.*, 2015; Parry & Newton, 2001). As for the question that whether there is direct relationship between the onset of PMS/PMDD and MS, the answer is still controversial (Freeman, Sammel, Rinaudo, & Sheng, 2004; Richards, Rubinow, Daly, & Schmidt, 2006).

The clustering method in the present study can be used to separate not only the healthy/sick but also PMS/PMDD individuals.

The method for separating healthy and sick animals by using unsupervised machine learning algorithm was established in the present study. In the clustering results, there were three clusters with different symptom severities corresponding to healthy, premenstrual syndrome (PMS), and PMDD. In the clinical setting, the differentiation of PMS and PMDD is also subjective and ineffective (Biggs & Demuth, 2011). The present clustering result with three clusters showed that when diagnosing PMDD, it is also possible to separate PMS from PMDD at the same time. Collection of more data from experimental animals and human patients can enable the specific borders regarding the behaviors and key factors between PMS and PMDD to be identified, allowing these two terms to be distinguished more easily in the future.

The original two-step feature selection method can be a new option for RNA-seq data analysis.

The original two-step feature selection method for RNA-seq small dataset analysis, which consists of a step of manual extraction and a subsequent step of machine learning feature selection, was employed in the present study. Due to the complexity and high dimensionality, RNA-seq data have been considered to be hard to handle and the proper tools have been lacking (Jabeen, Ahmad, & Raza, 2018). Other studies have already tried to analyze RNA-seq data using different types of classifiers or normalization methods (Evans, Hardin, & Stoebel, 2018; L. Wang, Xi, Sung, & Qiao, 2018). In the present study, there was only a small dataset containing few samples. Some of the features, such as the ones with several zeros and the ones with relatively small numbers, could easily affect the feature selection model and the selected list. Therefore, the data was preprocessed by reducing the feature dimensions of 84.49% to make it simpler and clearer, and this step also made the result more interpretable. Moreover, by this feature selection procedure, a set of diagnostic genes for PMDD symptoms instead of finding only one key gene was provided. By taking this procedure, the factors whose expression changes were the most relevant to the definition of the severity order of clusters 1 to 3 could be discovered, and therefore separating the degrees of disease severity could be performed in a more secure and dependable way. The concept of this method is similar to Gene Set Enrichment Analysis (GSEA) presented by Subramanian *et al.* in 2005. Since the genes with low expression levels can also have downstream effects, focusing on a set of gene instead of one single gene can be expected to make the result more reproducible and more interpretable (Subramanian *et al.*, 2005).

There are also limitations in the present study.

The present study still had some limitations. First, PMDD symptoms could not be induced in all rats, and thus it might not reflect the true level of difference among clusters. Second, the used factors are in the hippocampus, and they cannot be practiced directly in clinical settings. A further study on PMDD diagnostic sets in serum or cerebrospinal fluid is therefore required. Finally, the use of rat as test subject may involve some divergence from humans on the species level. In fact, there were also very few research teams using experimental primates to study PMS/PMDD and possible treatments (Qiao, Zhao, Wei, Zhang, & Wang, 2013; Rapkin, Pollack, Raleigh, Stone, & McGuire, 1995). However, there were similar problems of inconsistent experimental animal species and methods bringing out the results with less conviction. The establishment of official experimental process and the unification of disease models in different species are the urgent challenges in PMDD research.

In conclusion, a methodology using machine learning to construct efficient diagnostic procedures for pluricausal diseases, including PMDD, was established in the present study. This is the first study applying machine learning to syndrome classification in experimental animal. The procedures and tools developed in the present study are expected to be applied to other mood disorders and complex diseases in the future. They can be used to perform differential diagnoses between two or more different but similar diseases and contribute to tailored treatment. The data from the future studies can be used to build disease database that includes data from both experimental animals and human patients, which is expected to make the comparison of pathological change among species more diligently.

Figures and tables

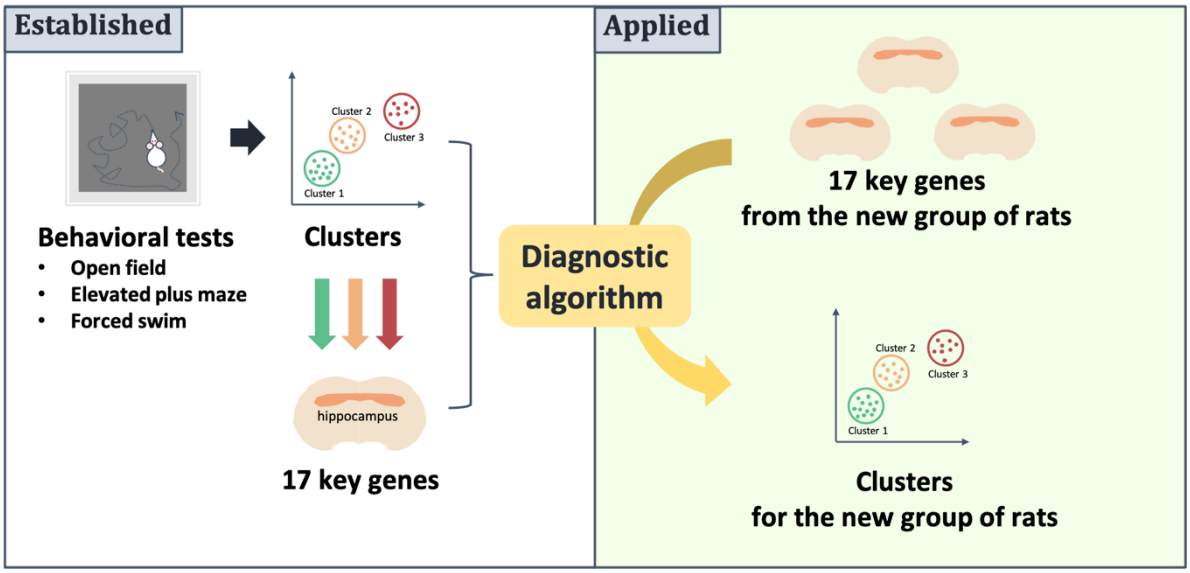


Figure 4-1. Overview of the present study.

In the establishing stage of the diagnostic algorithm, the clusters with different severity of PMDD syndromes were obtained from the three behavioral tests, open field test, elevated plus maze test, and forced swim test, and then were used to identify the most related 17 key genes. In the applying stage, the expression levels of the 17 key genes from the new group of rats were used as the input to the diagnostic algorithm, and successfully produced the output of the clusters for this group of rats.

References

Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E., & Wall, D. P. (2020). Multi-modular AI Approach to Streamline Autism Diagnosis in Young Children. *Scientific Reports*, *10*(1), 1–8. https://doi.org/10.1038/s41598-020-61213-w

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, *44*(W1), W3–W10. https://doi.org/10.1093/nar/gkw343

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: fifth edition*.

Balan, I., Beattie, M. C., O'Buckley, T. K., Aurelian, L., & Morrow, A. L. (2019). Endogenous Neurosteroid (3α,5α)3-Hydroxypregnan-20-one Inhibits Toll-like-4 Receptor Activation and Pro-inflammatory Signaling in Macrophages and Brain. *Scientific Reports*, *9*(1), 1–14. https://doi.org/10.1038/s41598-018-37409-6

Bar-Ami, S., Bendel, N., Leschiner, S., Levin, E., Veenman, L., & Gavish, M. (2006). The effects of prostaglandin F2α treatment on peripheral-type benzodiazepine receptors in the ovary and uterus during pseudopregnancy of rats. *Biochemical Pharmacology*, *71*(4), 472–478. https://doi.org/10.1016/j.bcp.2005.10.046

Bekhbat, M., & Neigh, G. N. (2018). Sex differences in the neuro-immune consequences of stress: Focus on depression and anxiety. *Brain, Behavior, and Immunity*, *67*, 1–12. https://doi.org/10.1016/j.bbi.2017.02.006

Belzung, C., & Griebel, G. (2001). Measuring normal and pathological anxiety-like behaviour in mice: A review. *Behavioural Brain Research*, *125*(1–2), 141–149. https://doi.org/10.1016/S0166-4328(01)00291-1

Bertrand, P. P., Paranavitane, U. T., Chavez, C., Gogos, A., Jones, M., & van den Buuse, M. (2005). The effect of low estrogen state on serotonin transporter function in mouse hippocampus: A behavioral and electrochemical study. *Brain Research*, *1064*(1–2), 10–20. https://doi.org/10.1016/j.brainres.2005.10.018

Biggs, W. S., & Demuth, R. H. (2011). Premenstrual syndrome and premenstrual dysphoric disorder. *American Family Physician*, *84*(8), 918–924. https://doi.org/10.4135/9781452232607.n30

Bixo, M., Ekberg, K., Poromaa, I. S., Hirschberg, A. L., Jonasson, A. F., Andréen, L., Timby, E., Wulff, M., Ehrenborg, A., Bäckström, T. (2017). Treatment of premenstrual dysphoric disorder with the GABAA receptor modulating steroid antagonist Sepranolone (UC1010)—A randomized controlled trial. *Psychoneuroendocrinology*, *80*, 46–55. https://doi.org/10.1016/j.psyneuen.2017.02.031

Bixo, M., Johansson, M., Timby, E., Michalski, L., & Bäckström, T. (2018). Effects of GABA active steroids in the female brain with a focus on the premenstrual dysphoric disorder. *Journal of Neuroendocrinology*, *30*(2), 1–9. https://doi.org/10.1111/jne.12553

Blesson, G. (2017). A study of the effect of random projection and other dimensionality reduction techniques on different classification methods. *Journal of Interdisciplinary Studies and Research*, *XVIII*(01).

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brennan, A. M., Fargnoli, J. L., Williams, C. J., Li, T., Willett, W., Kawachi, I., Qi, L., Hu, F. B., Mantzoros, C. S. (2009). Phobic anxiety is associated with higher serum concentrations of adipokines and cytokines in women with diabetes. *Diabetes Care*, *32*(5), 926–931. https://doi.org/10.2337/dc08-1979

Campbell, P. D., & Granato, M. (2020). Zebrafish as a tool to study schizophrenia-associated copy number variants. *DMM Disease Models and Mechanisms*, *13*(4). https://doi.org/10.1242/dmm.043877

Chan, C. (2013). Pseudopregnancy. *Clinical Veterinary Advisor*, 411–412. https://doi.org/10.1177/003172170909100225

Chang, Y., & Lin, C. (2008). Feature Ranking Using Linear SVM. *JMLR: Workshop and Conference Proceedings*, 53–64.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Christopher, M. B. (2006). *Pattern Recognition and Machine Learning*. https://doi.org/10.1007/978-3-030-57077-4_11

Ćosić, K., Popović, S., Šarlija, M., Kesedžić, I., & Jovanovic, T. (2020). Artificial intelligence in prediction of mental health disorders induced by the COVID-19 pandemic among health care workers. *Croatian Medical Journal*, *61*(3), 279–288. https://doi.org/10.3325/cmj.2020.61.279

Costa, A. M. N., Spence, K. T., Smith, S. S., & Ffrench-Mullen, J. M. H. (1995). Withdrawal from the endogenous steroid progesterone results in GABA(A)

currents insensitive to benzodiazepine modulation in rat CA1 hippocampus. *Journal of Neurophysiology*, *74*(1), 464–469. https://doi.org/10.1152/jn.1995.74.1.464

Dale, E., Pehrson, A. L., Jeyarajah, T., Li, Y., Leiser, S. C., Smagin, G., Olsen, C. K., Sanchez, C. (2016). Effects of serotonin in the hippocampus: How SSRIs and multimodal antidepressants might regulate pyramidal cell function. *CNS Spectrums*, *21*(2), 143–161. https://doi.org/10.1017/S1092852915000425

De Feo, V. J. (1963). Temporal aspect of uterine sensitivity in the pseudopregnant or pregnant rat. *Endocrinology*, (November), 305–316. Retrieved from http://press.endocrine.org/doi/abs/10.1210/endo-72-2-305

De Feo, V. J. (1966). Vaginal-Cervical Vibration: A Simple and Effective Method for the Induction of Pseudopregnancy in the Rat. *Endocrinology*, *79*(2), 440–442. https://doi.org/10.1210/endo-79-2-440

Delcourt, J., Beco, C., Ylieff, M. Y., Caps, H., Vandewalle, N., & Poncin, P. (2006). Comparing the EthoVision 2.3 system and a new computerized multitracking prototype system to measure the swimming behavior in fry fish. *Behavior Research Methods*, *38*(4), 704–710. https://doi.org/10.3758/BF03193904

Deng, D., Pang, Y., Duan, G., Liu, H., Liao, H., Liu, P., Liu, Y., Li, S., Chen, W., Wen, D., Xuan, C., Li, M. (2018). Larger volume and different functional connectivity of the amygdala in women with premenstrual syndrome. *European Radiology*, *28*(5), 1900–1908. https://doi.org/10.1007/s00330-017-5206-0

Desland, F. A., Afzal, A., Warraich, Z., & Mocco, J. (2014). Manual versus Automated Rodent Behavioral Assessment: Comparing Efficacy and Ease of Bederson and Garcia Neurological Deficit Scores to an Open Field Video-Tracking System.

Journal of Central Nervous System Disease, *6*, JCNSD.S13194. https://doi.org/10.4137/jcnsd.s13194

Devall, A. J., Santos, J. M., Fry, J. P., Honour, J. W., Brandão, M. L., & Lovick, T. A. (2015). Elevation of brain allopregnanolone rather than 5-HT release by short term, low dose fluoxetine treatment prevents the estrous cycle-linked increase in stress sensitivity in female rats. *European Neuropsychopharmacology*, *25*(1), 113–123. https://doi.org/10.1016/j.euroneuro.2014.11.017

Eisenlohr-Moul, T. A., Kaiser, G., Weise, C., Schmalenberger, K. M., Kiesner, J., Ditzen, B., & Kleinstäuber, M. (2020). Are there temporal subtypes of premenstrual dysphoric disorder?: Using group-based trajectory modeling to identify individual differences in symptom change. *Psychological Medicine*, *50*(6), 964–972. https://doi.org/10.1017/S0033291719000849

Estivill-Castro, V. (2002). Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65–75. https://doi.org/10.1145/568574.568575

Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, *19*(5), 776–792. https://doi.org/10.1093/bib/bbx008

Fabian, P., Gael, V., Alexandre, G., Vincent, M., & Bertrand, T. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1289/EHP4713

Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, *09*(01), 1–16. https://doi.org/10.4236/jilsa.2017.91001

Freeman, E. W. (2002). Treatment of depression associated with the menstrual cycle: Premenstrual dysphoria, postpartum depression, and the perimenopause. *Dialogues in Clinical Neuroscience*, *4*(2), 177–191. https://doi.org/10.31887/dcns.2002.4.2/efreeman

Freeman, E. W., Sammel, M. D., Rinaudo, P. J., & Sheng, L. (2004). Premenstrual syndrome as a predictor of menopausal symptoms. *Obstetrics and Gynecology*, *103*(5 I), 960–966. https://doi.org/10.1097/01.AOG.0000124804.81095.7f

Frye, C. A., & Bayon, L. E. (1998). Seizure activity is increased in endocrine states characterized by decline in endogenous levels of the neurosteroid 3 alpha,5 alpha-THP [published erratum appears in Neuroendocrinology 1998 Dec;68(6):436]. *Neuroendocrinology*, *68*(4), 272–280.

Garg, P., & Glick, S. (2018). AI's potential to diagnose and treat mental illness. *Harvard Business Review*. Retrieved from https://hbr.org/2018/10/ais-potential-to-diagnose-and-treat-mental-illness

Gillings, M. R. (2014). Were there evolutionary advantages to premenstrual syndrome? *Evolutionary Applications*, *7*(8), 897–904. https://doi.org/10.1111/eva.12190

Gingnell, M., Morell, A., Bannbers, E., Wikström, J., & Poromaa, I. S. (2012). Menstrual cycle effects on amygdala reactivity to emotional stimulation in premenstrual dysphoric disorder. *Hormones and Behavior*, *62*(4), 400–406. https://doi.org/10.1016/j.yhbeh.2012.07.005

Gold, E. B., Wells, C., & Rasor, M. O. N. (2016). The Association of Inflammation with Premenstrual Symptoms. *Journal of Women's Health*, *25*(9), 865–874. https://doi.org/10.1089/jwh.2015.5529

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press, Cambridge, MA, USA.

Gracia, C. R., Freeman, E. W., Sammel, M. D., Lin, H., Sheng, L., & Frye, C. (2009). Allopregnanolone levels before and after selective serotonin reuptake inhibitor treatment of premenstrual symptoms. *Journal of Clinical Psychopharmacology*, *29*(4), 403–405. https://doi.org/10.1097/JCP.0b013e3181ad2064

Gray, J. D., Rubin, T. G., Hunter, R. G., & McEwen, B. S. (2014). Hippocampal gene expression changes underlying stress sensitization and recovery. *Mol Psychiatry*, *19*(11), 1171–1178. https://doi.org/10.1038/mp.2013.175

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3*, 1157–1182. https://doi.org/https://doi.org/10.1162/153244303322753616

Hanekamp, E. E., Gielen, S. C. J. P., Smid-Koopman, E., Kühne, L. C. M., De Ruiter, P. E., Chadha-Ajwani, S., Brinkmann, A. O., Grootegoed, J. A., Burger, C. W., Huikeshoven, F. J., Blok, L. J. (2003). Consequences of loss of progesterone receptor expression in development of invasive endometrial cancer. *Clinical Cancer Research*, *9*(11), 4190–4199.

Hantsoo, L., & Epperson, C. N. (2020). Allopregnanolone in premenstrual dysphoric disorder (PMDD): Evidence for dysregulated sensitivity to GABA-A receptor modulating neuroactive steroids across the menstrual cycle. *Neurobiology of Stress*, *12*(December 2019), 100213. https://doi.org/10.1016/j.ynstr.2020.100213

Hardy, C., & Hunter, M. S. (2021). Premenstrual symptoms and work: Exploring female staff experiences and recommendations for workplaces. *International*

Journal of Environmental Research and Public Health, *18*(7). https://doi.org/10.3390/ijerph18073647

Heinemann, L. A. J., Minh, T. Do, Filonenko, A., & Uhl-Hochgräber, K. (2010). Explorative Evaluation of the Impact of Severe Premenstrual Disorders on Work Absenteeism and Productivity. *Women's Health Issues*, *20*(1), 58–65. https://doi.org/10.1016/j.whi.2009.09.005

Ho, H. P., Olsson, M., Westberg, L., Melke, J., & Eriksson, E. (2001). The serotonin reuptake inhibitor fluoxetine reduces sex steroid-related aggression in female rats: An animal model of premenstrual irritability? *Neuropsychopharmacology*, *24*(5), 502–510. https://doi.org/10.1016/S0893-133X(00)00219-0

Hsu, F. C., & Smith, S. S. (2003). Progesterone withdrawal reduces paired-pulse inhibition in rat hippocampus: Dependence on GABAAreceptor α4 subunit upregulation. *Journal of Neurophysiology*, *89*(1), 186–198. https://doi.org/10.1152/jn.00195.2002

Islas-preciado, D., Lo´pez-rubalcava, C., Gonza´ lez-olvera, J., Gallardo-tenorio, A., & Estrada-camarena, E. (2016). Environmental enrichment prevents anxiety-like behavior induced by progesterone withdrawal in two strains of rats. *Neuroscience*, *336*, 123–132. https://doi.org/10.1016/j.neuroscience.2016.08.050

Iwata, M., Ota, K. T., & Duman, R. S. (2013). The inflammasome: Pathways linking psychological stress, depression, and systemic illnesses. *Brain Behav Immun.*, *31*, 105–114. https://doi.org/10.1016/j.bbi.2012.12.008

Jabeen, A., Ahmad, N., & Raza, K. (2018). *Machine Learning-Based State-of-the-Art Methods for the Classi fi cation of RNA-Seq Data*.

Joachims, T. (1998). Making large-Scale SVM Learning Practical. *Technical Report*.

Kaplan, A. (1973). *The Conduct of Inquiry: Methodology for Behavioural Science*. Aylesbury, Buckinghamshire: International Textbook Company.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 3147–3155.

Kinney, D. K., & Tanaka, M. (2009). An evolutionary hypothesis of depression and its symptoms, adaptive value, and risk factors. *Journal of Nervous and Mental Disease*, *197*(8), 561–567. https://doi.org/10.1097/NMD.0b013e3181b05fa8

Koo, J. W., Russo, S. J., Ferguson, D., Nestler, E. J., & Duman, R. S. (2010). Nuclear factor-κB is a critical mediator of stress-impaired neurogenesis and depressive behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2669–2674. https://doi.org/10.1073/pnas.0910658107

Kurumaji, A., Ito, T., Ishii, S., & Nishikawa, T. (2008). Effects of FG7142 and immobilization stress on the gene expression in the neocortex of mice. *Neuroscience Research*, *62*(3), 155–159. https://doi.org/10.1016/j.neures.2008.08.001

Laconi, M. R., Casteller, G., Gargiulo, P. A., Bregonzio, C., & Cabrera, R. J. (2001). The anxiolytic effect of allopregnanolone is associated with gonadal hormonal status in female rats. *European Journal of Pharmacology*, *417*(1–2), 111–116. https://doi.org/10.1016/S0014-2999(01)00865-2

Lawrence, S., Giles, C. L., & Tsoi, A. C. (1997). Lessons in neural network training: overfitting may be harder than expected. *Proceedings of the National Conference on Artificial Intelligence*, 540–545.

Lee, Y. J., Yi, S. W., Ju, D. H., Lee, S. S., Sohn, W. S., & Kim, I. J. (2015). Correlation between postpartum depression and premenstrual dysphoric disorder: Single center study. *Obstetrics & Gynecology Science*, *58*(5), 353. https://doi.org/10.5468/ogs.2015.58.5.353

Li, Y., Pehrson, A. L., Budac, D. P., Sánchez, C., & Gulinello, M. (2012). A rodent model of premenstrual dysphoria: Progesterone withdrawal induces depression-like behavior that is differentially sensitive to classes of antidepressants. *Behavioural Brain Research*, *234*(2), 238–247. https://doi.org/10.1016/j.bbr.2012.06.034

Locci, A., & Pinna, G. (2017). Neurosteroid biosynthesis down-regulation and changes in GABAA receptor subunit composition: a biomarker axis in stress-induced cognitive and emotional impairment. *British Journal of Pharmacology*, *174*(19), 3226–3241. https://doi.org/10.1111/bph.13843

Löfgren, M. (2009). *Behavioral effects of female sex steroid hormones – models of PMS and PMDD in Wistar rats*.

Löfgren, M., Johansson, I., Meyerson, B., Turkmen, S., & Bäckström, T. (2009). Physiology & Behavior Withdrawal effects from progesterone and estradiol relate to individual risk-taking and explorative behavior in female rats. *Physiology & Behavior*, *96*(1), 91–97. https://doi.org/10.1016/j.physbeh.2008.08.015

Mantas, I., Saarinen, M., Xu, Z. Q. D., & Svenningsson, P. (2021). Update on GPCR-based targets for the development of novel antidepressants. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-021-01040-1

Marrocco, J., Einhorn, N. R., Petty, G. H., Li, H., Dubey, N., Hoffman, J., Berman, K. F., Goldman, D., Lee, F. S., Schmidt, P. J., McEwen, B. S. (2020). Epigenetic intersection of BDNF Val66Met genotype with premenstrual dysphoric disorder

transcriptome in a cross-species model of estradiol add-back. *Molecular Psychiatry*, *25*(3), 572–583. https://doi.org/10.1038/s41380-018-0274-3

Mathis, M. W., & Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, *60*, 1–11. https://doi.org/10.1016/j.conb.2019.10.008

Mehryar, M., Afshin, R., & Ameet, T. (2018). *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, USA.

Mellon, S. H., Wolkowitz, O. M., Schonemann, M. D., Epel, E. S., Rosser, R., Burke, H. B., Mahan, L., Reus, V. I., Stamatiou, D., Liew, C. C., Cole, S. W. (2016). Alterations in leukocyte transcriptional control pathway activity associated with major depressive disorder and antidepressant treatment. *Translational Psychiatry*, *6*(March), e821. https://doi.org/10.1038/tp.2016.79

Meltzer-Brody, S., Colquhoun, H., Riesenberg, R., Epperson, C. N., Deligiannidis, K. M., Rubinow, D. R., Li, H., Sankoh, A. J., Clemson, C., Schacterle, A., Jonas, J., Kanes, S. (2018). Brexanolone injection in post-partum depression: two multicentre, double-blind, randomised, placebo-controlled, phase 3 trials. *The Lancet*, *392*(10152), 1058–1070. https://doi.org/10.1016/S0140-6736(18)31551-4

Miller, A. H., Maletic, V., & Raison, C. L. (2010). Inflammation and its discontents: The role of cytokines in the pathophysiology of major depression. *Psiquiatria Biologica*, *17*(2), 71–80. https://doi.org/10.1016/j.psiq.2010.04.001

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mitchell, T. M. (2017). *Machine Learning*.

Mohd Rosely, N. F. L., Salleh, R., & Zain, A. M. (2019). Overview Feature Selection using Fish Swarm Algorithm. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012068

Morales-medina, J. C., Witchey, S. K., & Caldwell, H. K. (2016). The Role of Vasopressin in Anxiety and Depression. *Melatonin, Neuroprotective Agents and Antidepressant Therapy*. https://doi.org/10.1007/978-81-322-2803-5

Moran, M. H., & Smith, S. S. (1998). Progesterone withdrawal I: Pro-convulsant effects. *Brain Research*, *807*(1–2), 84–90. https://doi.org/10.1016/S0006-8993(98)00782-3

Morriss, G. M., & Keverne, E. B. (1974). Premenstrual tension. *Lancet*, *2*, 1317. https://doi.org/10.1016/0002-9378(50)90475-3

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, *14*(7), 2152–2176. https://doi.org/10.1038/s41596-019-0176-0

Noldus, L. P. J. J., Spink, A. J., & Tegelenbosch, R. A. J. (2001). EthoVision: A versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, & Computers*, *33*(3), 398–414.

Norvig, S. J. R. P. (1995). *Artificial intelligence: a modern approach*. Alan Apt.

Oded, M., & Lior, R. (2005). *The data mining and knowledge discovery handbook*.

Pamela, M. (2004). *Machines Who Think*. Natick, Massachusetts: A K Peters, Ltd. https://doi.org/10.1126/science.254.5036.1291-a

Pandey, G. N., Dwivedi, Y., Ren, X., Rizavi, H. S., Roberts, R. C., Conley, R. R., & Tamminga, C. (2003). Altered expression and phosphorylation of myristoylated

alanine-rich C kinase substrate (MARCKS) in postmortem brain of suicide victims with or without depression. *Journal of Psychiatric Research*, *37*(5), 421–432. https://doi.org/10.1016/S0022-3956(03)00047-5

Parry, B. L., & Newton, R. P. (2001). Chronobiological basis of female-specific mood disorders. *Neuropsychopharmacology*, *25*(5), S102–S108. https://doi.org/10.1016/S0893-133X(01)00340-2

Paul, E. J., Tossell, K., & Ungless, M. A. (2019). Transcriptional profiling aligned with in situ expression image analysis reveals mosaically expressed molecular markers for GABA neuron sub-groups in the ventral tegmental area. *European Journal of Neuroscience*, *50*(11), 3732–3749. https://doi.org/10.1111/ejn.14534

Pepe, G. J., & Rothchild, I. (1974). A comparative study of serum progesterone levels in pregnancy and in various types of pseudopregnancy in the rat. *Endocrinology*, *95*(1), 275–279. https://doi.org/10.1210/endo-95-1-275

Phillips, O. R., Onopa, A. K., Hsu, V., Ollila, H. M., Hillary, R. P., Hallmayer, J., Gotlib, I. H., Tayor, J., Mackey, L., Singh, M. K. (2019). Beyond a Binary Classification of Sex: An Examination of Brain Sex Differentiation, Psychopathology, and Genotype. *J Am Acad Child Adolesc Psychiatry*, *58*(8), 787–798. https://doi.org/10.1016/j.jaac.2018.09.425

Pinna, G., Costa, E., & Guidotti, A. (2009). SSRIs act as selective brain steroidogenic stimulants (SBSSs) at low doses that are inactive on 5-HT reuptake. *Current Opinion in Pharmacology*, *9*(1), 24–30. https://doi.org/10.1016/j.coph.2008.12.006

Polo, A., Singh, S., Crispo, A., Russo, M., Giudice, A., Montella, M., Colonna, G., Costantini, S. (2017). Evaluating the associations between human circadian

rhythms and dysregulated genes in liver cancer cells. *Oncology Letters*, *14*(6), 7353–7359. https://doi.org/10.3892/ol.2017.7109

Qiao, M., Zhao, Q., Wei, S., Zhang, H., & Wang, H. (2013). Premenstrual dysphoria and luteal stress in dominant-social-status female macaques. *Evidence-Based Complementary and Alternative Medicine*, *2013*. https://doi.org/10.1155/2013/393862

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Rajkumar, A. P., Qvist, P., Donskov, J. G., Lazarus, R., Pallesen, J., Nava, N., Winther, G., Liebenberg, N., Cour, S. H., Paternoster, V., Fryland, T., Palmfeldt, J., Fejgin, K., Mørk, A., Nyegaard, M., Pakkenberg, B., Didriksen, M., Nyengaard, J. R., Wegener, G., Mors, O., Christensen, J. H., Børglum, A. D. (2020). Reduced Brd1 expression leads to reversible depression-like behaviors and gene-expression changes in female mice. *Translational Psychiatry*, *10*(1). https://doi.org/10.1038/s41398-020-00914-2

Ranjan, G. S. K., Kumar Verma, A., & Radhika, S. (2019). K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*, 9–13. https://doi.org/10.1109/I2CT45611.2019.9033691

Rao, S., Yin, L., Xiang, Y., & So, H. C. (2020). Analysis of genetic differences between psychiatric disorders: Exploring pathways and cell-types/tissues involved and the ability to distinguish the disorders by polygenic scores. Retrieved from https://arxiv.org/abs/2003.07105

Rapkin, A. J., Pollack, D. B., Raleigh, M. J., Stone, B., & McGuire, M. T. (1995). Menstrual cycle and social behavior in vervet monkeys. *Psychoneuroendocrinology*, *20*(3), 289–297. https://doi.org/10.1016/0306-4530(94)00060-N

Richards, M., Rubinow, D. R., Daly, R. C., & Schmidt, P. J. (2006). Premenstrual symptoms and perimenopausal depression. *American Journal of Psychiatry*, *163*(1), 133–137. https://doi.org/10.1176/appi.ajp.163.1.133

Ronnenberg, A. G., Houghton, S. C., Nobles, C., Zagarins, S. E., Faraj, J. L., & Whitcomb, B. W. (2014). Association of inflammation markers with menstrual symptom severity and premenstrual syndrome in young women. *Human Reproduction*, *29*(9), 1987–1994. https://doi.org/10.1093/humrep/deu170

Rumelhart, D. E., & Hintont, G. E. (1982). Learning Representations by Back-Propagating Errors. *Nature*, *323*(9), 533–536. https://doi.org/10.7551/mitpress/1888.003.0013

Ryu, A., & Kim, T. H. (2015). Premenstrual syndrome: A mini review. *Maturitas*, *82*(4), 436–440. https://doi.org/10.1016/j.maturitas.2015.08.010

Schneider, T., & Popik, P. (2007). Increased depressive-like traits in an animal model of premenstrual irritability. *Hormones and Behavior*, *51*(1), 142–148. https://doi.org/10.1016/j.yhbeh.2006.09.006

Schneider, T., & Popik, P. (2009). An animal model of premenstrual dysphoric disorder sensitive to antidepressants. *Current Protocols in Neuroscience*, (SUPPL. 46), 1–10. https://doi.org/10.1002/0471142301.ns0931s46

Smith, M. S., Freeman, M. E., & Neill, J. D. (1975). The control of progesterone secretion during the estrous cycle and early pseudopregnancy in the rat: Prolactin,

gonadotropin and steroid levels associated with rescue of the corpus luteum of pseudopregnancy. *Endocrinology*, *96*(1), 219–226. https://doi.org/10.1210/endo-96-1-219

Smith, S. S., Gong, Q. H., Li, X., Moran, M. H., Bitran, D., Frye, C. A., & Hsu, F. C. (1998). Withdrawal from 3α-OH-5α-pregnan-20-one using a pseudopregnancy model alters the kinetics of hippocampal GABA(A)-gated current and increases the GABA(A) receptor α4 subunit in association with increased anxiety. *Journal of Neuroscience*, *18*(14), 5275–5284. https://doi.org/10.1523/jneurosci.18-14-05275.1998

Somnay, Y. R., Craven, M., McCoy, K. L., Carty, S. E., Wang, T. S., Greenberg, C. C., & Schneider, D. F. (2017). Improving diagnostic recognition of primary hyperparathyroidism with machine learning. *Surgery (United States)*, *161*(4), 1113–1121. https://doi.org/10.1016/j.surg.2016.09.044

Song, S. Y., Meng, X. W., Xia, Z. Y., Liu, H., Zhang, J., Chen, Q. C., Liu, H. Y., Ji, F. H., Peng, K. (2019). Cognitive impairment and transcriptomic profile in hippocampus of young mice after multiple neonatal exposures to sevoflurane. *Aging* (Vol. 11). https://doi.org/10.18632/aging.102326

Steinberg, E. M., Cardoso, G. M. P., Martinez, P. E., Rubinow, D. R., & Schmidt, P. J. (2012). Rapid response to fluoxetine in women with premenstrual dysphoric disorder. *Depression and Anxiety*, *29*(6), 531–540. https://doi.org/10.1002/da.21959

Sturman, O., Von Ziegler, L., Schläppi, C., Akyol, F., Privitera, M., Slominski, D., Grimm, C., Thieren, L., Zerbi, V., Grewe, B., Bohacek, J. (2020). Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming

commercial solutions. *Neuropsychopharmacology*, *45*(11), 1942–1952. https://doi.org/10.1038/s41386-020-0776-y

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun, L. L., Yang, S. L., Sun, H., Li, W. Da, & Duan, S. R. (2019). Molecular differences in Alzheimer's disease between male and female patients determined by integrative network analysis. *Journal of Cellular and Molecular Medicine*, *23*(1), 47–58. https://doi.org/10.1111/jcmm.13852

Sun, X., Song, Z., Si, Y., & Wang, J. H. (2018). microRNA and mRNA profiles in ventral tegmental area relevant to stress-induced depression and resilience. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *86*(May), 150–165. https://doi.org/10.1016/j.pnpbp.2018.05.023

Timby, E., Bäckström, T., Nyberg, S., Stenlund, H., Wihlbäck, A. C. N., & Bixo, M. (2016). Women with premenstrual dysphoric disorder have altered sensitivity to allopregnanolone over the menstrual cycle compared to controls—a pilot study. *Psychopharmacology*, *233*(11), 2109–2117. https://doi.org/10.1007/s00213-016-4258-1

Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour*, *124*, 203–220. https://doi.org/10.1016/j.anbehav.2016.12.005

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Waldrop, M. (2019). What are the limits of deep learning? *Proceedings of the National Academy of Sciences of the United States of America*, *116*(4), 1074–1077. https://doi.org/10.1073/pnas.1821594116

Wang, J. M., Johnston, P. B., Ball, B. G., & Brinton, R. D. (2005). The neurosteroid allopregnanolone promotes proliferation of rodent and human neural progenitor cells and regulates cell-cycle gene and protein expression. *Journal of Neuroscience*, *25*(19), 4706–4718. https://doi.org/10.1523/JNEUROSCI.4520-04.2005

Wang, L., Xi, Y., Sung, S., & Qiao, H. (2018). RNA-seq assistant: Machine learning based methods to identify more transcriptional regulated genes. *BMC Genomics*, *19*(1), 1–13. https://doi.org/10.1186/s12864-018-4932-2

Wiklund, L. (2017). Premenstrual Dysphoric Disorder - A Review of Neural and Cognitive Changes in Women with PMDD. *Bachelor Degree Project in Cognitive Neuroscience*. https://doi.org/10.1016/j.mcna.2019.02.007

Wilson, H. C. (1992). A critical review of menstrual synchrony research. *Psychoneuroendocrinology*, *17*(6), 565–591. https://doi.org/10.1016/0306-4530(92)90016-Z

Yonkers, K. A., O'Brien, P. S., & Eriksson, E. (2008). Premenstrual syndrome. *The Lancet*, *371*(9619), 1200–1210. https://doi.org/10.1016/S0140-6736(08)60527-9

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zarley, B. D. (2019). Meet the scientists who are training AI to diagnse mental illness. *The Verge*. Retrieved from https://www.theverge.com/2019/1/28/18197253/ai-mental-illness-artificial-intelligence-science-neuroimaging-mri

Ziomkiewicz-wichary, A. (2017). Premenstrual syndrome : A short review. *Zdrowie Publiczne i Zarządzanie*, *15*(2), 156–160. https://doi.org/10.4467/20842627OZ.17.016.6786

Acknowledgements

Throughout the writing of this thesis, I have received a great deal of support and assistance.

I would first like to thank Professor Masugi Nishihara. Your incisive suggestions and supports encouraged me to have faith to accomplish this study. I would like to thank Associate Professor Keitaro Yamanouchi. Your insightful feedback helped to make up for the shortcomings in this study. I would like to thank my supervisor, Associate Professor Takashi Matsuwaki. It is impossible to accomplish this study without your knowledge and guidance, and your generous and patient support.

I would like to thank Professor Chia-Wei Sun and Doctor Chun-Jung Huang. You taught me the basics of data analysis and your precious advice give the directions of the works. I would like to thank Wei-Yuan Chang, Li-Yu Chen, Kuo-Yen Luo, Saya, and Rei. Thank you for your help when I got lost in the ocean of programming logic, methods and errors. I would like to thank the members of the Laboratory of Veterinary Physiology. Thank you for your precise advice to my experiment and operation methods.

I would like to thank my family, friends, and pets. Thank you for supporting both my research and private life.

I would also like to thank all the experimental animals involved in this study. May the pure souls rest in peace.