Computational Psychiatry: New Perspectives on Depression Research and Clinical Application

Authors: Qu Jianxin, Wu Yin, Liu Jinting, Li Hong, Li Hong

Date: 2019-08-26T00:00:00+00:00

Abstract

Depression is a complex and heterogeneous psychiatric disorder that imposes a substantial global disease burden. Although symptom-based diagnostic approaches have been widely adopted across various domains, such methods are not conducive to elucidating pathological mechanisms. Moreover, these diagnostic approaches suffer from low predictive validity, hindering accurate evaluation and comparison of therapeutic efficacy across different treatment modalities. Computational psychiatry methods can address these challenges through two complementary approaches: theory-driven and data-driven methods, thereby enhancing the understanding, prevention, and treatment of depression. Theory-driven approaches, grounded in empirical knowledge or hypotheses, employ computational modeling techniques to conduct multi-level analyses of data; data-driven approaches, leveraging machine learning algorithms to analyze high-dimensional data, improve the accuracy of depression diagnosis and prognosis, consequently enhancing treatment precision. The development and integration of theory-driven and data-driven methods, coupled with the consolidation of expertise and resources, will more effectively advance the prevention and treatment of depression.

Full Text

Computational Psychiatry: A New Perspective on Depression Research and Clinical Application

OU Jianxin[1]; WU Yin[1]; LIU Jinting[1]; LI Hong[1],[2],[3]
(1 School of Psychology, Shenzhen University; 2 Shenzhen Key Laboratory of Affective and Social Cognitive Science; 3 Shenzhen Institute of Neuroscience, Shenzhen 518060, China)

Abstract: Depression is a complex and heterogeneous mental disorder that imposes a heavy burden of disease globally. Although symptom-based diagnostic

methods have been widely applied across various domains, this approach is not conducive to exploring pathological mechanisms. Additionally, its low predictive validity makes it difficult to accurately assess and compare the efficacy of different treatment options. Computational psychiatry addresses these issues through two complementary approaches—theory-driven and data-driven methods—thereby enhancing our understanding, prevention, and treatment of depression. Theory-driven approaches employ computational modeling for multilevel analysis based on empirical knowledge or hypotheses, while data-driven approaches utilize machine learning algorithms to analyze high-dimensional data, improving diagnostic and predictive accuracy for depression and consequently enhancing treatment precision. The development and integration of these two approaches, combined with the consolidation of talent and resources, will more effectively advance depression prevention and treatment.

Keywords: depression; computational psychiatry; computational models; machine learning; diagnosis; treatment

Depression is a common yet often overlooked mental illness that has become a leading cause of "mental disability" worldwide, particularly among young and middle-aged adults [**?**, **?**]. In March 2018, the World Health Organization (WHO) reported that approximately 300 million people globally suffer from depression, with about 800,000 deaths by suicide annually [**?**, **?**]. Depression significantly contributes to the overall global disease burden, causing immense suffering for patients and their families [**?**, **?**, **?**]. Therefore, systematic and in-depth research on depression to develop effective prevention and treatment strategies is urgently needed. However, symptom-based diagnostic methods for depression have proven ineffective in clinical applications, with low predictive validity [**?**, **?**]. Subsequent developments in biological psychiatry have also failed to adequately explain the psychopathological mechanisms of depression. With advances in computer science and biotechnology, the emerging field of computational psychiatry seeks to unlock the "black box" of depression by introducing computational and statistical methods. From both theory-driven and data-driven perspectives, it aims to reveal the brain' s information processing mechanisms and the pathogenesis of mental illness, utilizing high-dimensional complex data to identify psychiatric disorders [**?**, **?**]. Computational psychiatry represents a cutting-edge intersection of computational neuroscience and psychiatry that will substantially advance our understanding of the pathogenesis, diagnosis, treatment, and prevention of mental illness.

1.1 Definition of Depression

While depression typically refers to temporary low mood and sadness, major depressive disorder is a mental illness involving multiple factors including genetics, the nervous system, and cognition. Depressive disorder is a mood disorder characterized by low mood, anhedonia, strong feelings of guilt and fatigue, low self-worth, and accompanying symptoms such as sleep disturbances, appetite changes, and difficulty concentrating [**?**, **?**]. Research has found that depressed

individuals typically exhibit negative cognitive biases and poor emotion regulation [?, ?], and show impairments in approach-motivated behavior and reward learning processes [?, ?, ?, ?]. Notably, depressive disorder exhibits high heterogeneity. Depression can be divided into multiple subtypes with strong comorbidity—nearly two-thirds of patients with major depressive disorder also suffer from anxiety disorder [?, ?]. Additionally, the duration and number of depressive episodes vary across patients [?, ?]. However, these perspectives only describe the symptomatic manifestations of depression and fail to explain or control the occurrence, development, maintenance, and remission of depressive symptoms.

1.2 Diagnosis of Depression

Currently, depression diagnosis primarily relies on the International Classification of Diseases (ICD) and the Diagnostic and Statistical Manual of Mental Disorders (DSM) published by the American Psychiatric Association (APA). ICD and DSM define and diagnose mental disorders based on phenomenology and symptom clusters. For major depressive disorder, DSM-V diagnostic criteria require five or more of nine symptoms—including persistent depressed mood, anhedonia, insomnia or hypersomnia, significant weight change, and fatigue—to be present for most of the day, nearly every day, for two weeks [?, ?].

Despite its widespread use in basic research and clinical practice, this diagnostic approach has several significant problems. First, the method relies on patient self-reports, which are influenced by personal beliefs and self-awareness abilities, while subsequent clinical interviews depend on patients' verbal expression skills and clinicians' diagnostic experience and competence. This vague symptom description hinders objective clinical diagnosis. Second, this disease classification system struggles to explain comorbidity phenomena [?, ?]. Depression can be divided into multiple subtypes and shares symptomatic similarities with other mental illnesses such as bipolar affective disorder (BD), post-traumatic stress disorder (PTSD), and anxiety disorders. The presence of similar symptoms across different diseases may indicate shared underlying mechanisms (e.g., endophenotypes; [?, ?]). Current diagnostic methods simply categorize these as distinct disorders, potentially overlooking commonalities and hindering in-depth investigation of comorbidity mechanisms. Third, ICD and DSM provide qualitative rather than quantitative symptom descriptions, crudely dichotomizing the presence or absence of symptoms regardless of severity, which prevents individualized interventions based on symptom intensity. Finally, psychiatric diagnosis commonly uses dichotomous classification to separate patient and healthy populations, ignoring at-risk individuals who do not yet meet diagnostic criteria but have high depression risk.

The subjectivity of self-reports and clinical interviews, comorbidity between depression and other mental illnesses, and dichotomous classification of symptoms and disease all constrain diagnostic effectiveness and utility. In response, biological psychiatry aims to explore the functions and mechanisms of the nervous

system underlying mental disorders [?, ?, ?, ?] and attempts to use biomarkers as diagnostic criteria for depression. Although biological psychiatry examines multi-level biological pathological mechanisms and emphasizes changes in the internal environment, partially compensating for ICD and DSM limitations, scholars have noted that causal relationships between biological changes and symptoms are nearly impossible to establish one-to-one [?, ?], because mental health is linked not only to complex brain functions but also to individuals' environments and experiences [?, ?, ?, ?]. Due to the multi-system, multi-level influences and strong heterogeneity of mental illness [?, ?], biological psychiatry still faces a major gap: the lack of an appropriate intermediate description between molecular and clinical symptom levels [?, ?, ?, ?, ?]. Computational psychiatry introduces computational modeling to bridge micro-level (molecules, cells, etc.) and macro-level (behavior, environment, etc.) processes, attempting to reveal influence pathways between different levels and systems—for instance, how changes in a particular factor level affect entire system changes and subsequently trigger behavioral modifications [?, ?].

2.1 Definition and Goals

Computational psychiatry integrates methods from psychiatry, psychology, neuroscience, behavioral economics, and machine learning to establish computational models of brain function based on neural and cognitive phenomena associated with mental illness, such as specific models of neuronal networks and abstract models of high-level cognitive abilities [?, ?]. It aims to predict the degree of psychological dysfunction and evaluate treatment efficacy through detailed, multi-dimensional computational models [?, ?].

2.2.1 Theory-Driven Approaches

Theory-driven approaches originate from computational neuroscience [?, ?, ?, ?]. Computational models have been used to study individuals'information representation and processing, seeking explanations rather than statistical features, and concretizing the cognitive and neurophysiological process from "unobservable brain states" to behavioral measurements as mathematical models [?, ?, ?, ?, ?]. Most models assume that cognitive style (information processing mode) is closely linked to information acquisition, interactively influencing learning and the formation of beliefs or cognitive schemas, which explicitly or implicitly affect individual behavior or decision-making. Accumulated abnormal learning experiences form the core component of maladaptive cognitive styles, which may be important factors inducing mental illness.

Theory-driven computational modeling is a unique method for explaining the computational properties of the human brain. By representing model parameters of specific psychological processes, this method objectively and quantitatively explains the weight and influence pathways of components in cognitive and emotional functional abnormalities in depressed patients. It overcomes problems such as subjectivity and low precision in symptom-based diagnosis,

providing objective quantitative indicators for severity determination and treatment efficacy evaluation. If certain cognitive tasks show high sensitivity and specificity for depression identification, they can be used to distinguish healthy and depressed groups, improving objectivity and efficiency of identification, and providing precise cognitive-behavioral intervention plans based on patients' specific cognitive function impairments.

2.2.2 Data-Driven Approaches

Researchers have obtained rich molecular biological and neuroimaging data through biotechnology and brain imaging techniques. Combined with symptom and self-report inventory results, these form complex multi-dimensional datasets that reflect the complexity of mental illness etiology, where factors at each level cannot be arbitrarily ignored. However, traditional methods struggle to integrate such multi-dimensional data. Machine learning offers an alternative approach that helps analyze large and complex datasets by accumulating experience and updating algorithms through training [**?**, **?**]. Currently, machine learning methods have been widely applied to emotion state classification [**?**, **?**, **?**, **?**], mood fluctuation monitoring [**?**, **?**, **?**, **?**, **?**], and diagnosis of mental illnesses such as depression [**?**, **?**, **?**, **?**, **?**] and schizophrenia [**?**, **?**, **?**, **?**, **?**, **?**, **?**]. Therefore, data-driven approaches objectively reveal effective clinical manifestation patterns within datasets by integrating patient data from different levels and systems, thereby distinguishing healthy and depressed groups, evaluating drug and intervention efficacy, and predicting depression risk.

3 Theory-Driven Computational Psychiatry Research

Models based on theory mainly include four categories: biophysically based models focusing on activities at cellular, synaptic, and cortical levels; connectionist and neural models reflecting neural and behavioral levels; algorithmic models such as reinforcement learning (RL) and drift diffusion models (DDM) that abstractly simulate brain computation and behavior execution; and normative models based on Bayesian theory that examine whether behavior and neural activity are consistent with normative theory [**?**, **?**, **?**, **?**]. Currently, reinforcement learning models, drift diffusion models, and generative models based on the Bayesian brain hypothesis are most widely used in mental illness research. These models can quantify finer psychophysiological processes (e.g., attention, decision-making) from behavior to neural levels. The following sections primarily introduce results from these three models in depression research.

3.1 Reinforcement Learning Models

Reinforcement learning refers to the process by which a learning agent learns to associate situations with actions—that is, to execute goal-directed behavior through a series of trial-and-error explorations based on reward maximization principles [**?**, **?**]. The agent can perceive environmental states and take actions

to change them. Regarding behavioral outcomes, actions not only affect reward acquisition but also trigger subsequent situation-based reward expectations [?, ?].

Reinforcement learning theory suggests that habitual behavior values are formed through trial-and-error experiences without explicit decision model encoding, executing rapidly and considered model-free reinforcement learning. In contrast, goal-directed behavior requires deliberative planning and is considered model-based reinforcement learning because real-time value calculation requires internal model participation [?, ?, ?, ?]. Studies by Daw et al. (2011) and Wunderlich et al. (2012), as well as genetic research by Doll et al. (2016), have demonstrated the existence of these two learning types.

In model-free learning, the simplest and most widely used model is the Rescorla-Wagner model [?, ?]. Based on this model, the value (V_t) computed by the brain in the current trial t can be expressed as:

$$V_t = V_{t-1} + \alpha \cdot PE_{t-1}$$

where represents the learning rate and PE is the prediction error. In the previous trial (t-1), PE equals the difference between the obtained reward (R) and the expected value (V):

$$PE_{t-1} = R_{t-1} - V_{t-1}$$

The learning rate is the core variable in model-free reinforcement learning, reflecting the efficiency with which the agent uses prediction error information to update the value function. Value functions can be flexibly updated based on the agent' s motivational state and cognition of non-rewarding environments using algorithms [?, ?, ?, ?]. Building on the Rescorla-Wagner model, researchers developed the more advanced temporal difference (TD) model, which incorporates future rewards (see [?, ?, ?] for details). For more reinforcement learning model algorithms, see [?, ?].

The classic reinforcement learning paradigm presents participants with a pair of meaningless stimuli to choose from, with different choices producing different outcomes—one stimulus is associated with high-probability reward or punishment (70%-80%), while the other is associated with low-probability reward or punishment (20%-30%). As trial numbers increase, participants discover the associations between rewards/punishments and specific stimuli, tending to approach rewards and avoid punishments. This paradigm is most commonly used to study reward and punishment processing in depressed populations. The following sections introduce how reinforcement learning models explain anhedonia mechanisms in depression.

Depression-related anhedonia is strongly linked to abnormal reward processing. Rothkirch et al. (2017) used this paradigm to study reward and punishment processing mechanisms in unmedicated major depressive disorder patients. Results

showed that patients could adjust their behavior based on monetary reinforcement, and neural activity in the ventral striatum and anterior insula processing reward and punishment information remained intact. However, activity in the medial orbitofrontal cortex was reduced, and reward prediction error signals influenced by this region were negatively correlated with anhedonia. Therefore, depressed patients may have impairments in experiencing pleasure or reward, a phenomenon related to medial orbitofrontal cortex processing of prediction error signals. Additionally, numerous studies indicate that reward prediction error signals are associated with dopamine in the ventral tegmental area/substantia nigra and their projection regions such as the striatum, prefrontal cortex, and hippocampus, with reward prediction error signals typically showing a decreasing trend in depression [?, ?, ?, ?, ?, ?]. However, some studies using similar paradigms found that dopaminergic reward prediction error processing functions normally in depressed populations [?, ?]. These inconsistent results may relate to patient heterogeneity, task difficulty, and antidepressant medication status. Different depression types and severity levels, as well as varying drug dosages and efficacy, increase within-group variance and may mask experimental condition effects. Additionally, overly complex tasks may produce floor effects in depressed patients' performance, potentially caused by learning deficits but also by attention maintenance and executive function impairments. Therefore, examining learning processes in depressed patients requires rigorous experimental designs.

Unlike prediction error reflecting model-free learning, model-based learning requires deliberative planning. Research shows that depressed patients with impaired cognitive functions use model-based learning less frequently, with related neural activity showing dysfunction in lateral and medial prefrontal cortex, orbitofrontal cortex, and hippocampus [?, ?, ?, ?, ?, ?]. A possible reason is that past learning has formed beliefs that situations are uncontrollable, affecting subsequent motivation to explore and use model-based learning strategies to update existing beliefs [?, ?], similar to learned helplessness phenomena. Beyond cognition and memory, contextual factors such as stress also affect model-based reinforcement learning. Stress damages prefrontal cortex structure and function [?, ?, ?], shifting behavior from goal-directed to habitual control [?, ?, ?]. Individuals with high chronic stress levels show less model-based learning under acute stress and weaker executive control [?, ?]. Moreover, compared to less depressed groups, more severely depressed groups are more affected by stress in model-based learning, showing more habitual behavior, possibly due to impaired executive control function [?, ?, ?, ?, ?]. Therefore, stress may exacerbate depressive symptoms by affecting planning, execution, and control of goal-directed behavior.

In summary, reinforcement learning model-based depression research has primarily found that depressed populations show impairments in processing reward prediction errors and expected values, exhibiting lower reward sensitivity [?, ?]. Additionally, numerous studies consistently indicate that depressed populations show decreased model-based learning performance, possibly related to impair-

ments in working memory, executive control function, and value processing.

3.2 Drift Diffusion Model

The drift diffusion model (DDM) is a sequential sampling model primarily used to describe rapid binary decision processes. These simple binary decisions are brief, with reaction times typically less than 1000-1500 ms, involving only a single judgment process without multi-stage decision-making or reasoning. DDM assumes that decisions result from noisy evidence accumulation over time from a starting point to a decision threshold or boundary, generating a response [?, ?, ?, ?, ?, ?]. The model is expressed as:

$$dy(t) = v(\Delta u) \cdot dt + \sigma \cdot dW$$

where y(t) represents the amount of accumulated information at time point t, Δu refers to the utility difference between two choices, v is the drift rate (the amount of information accumulated per unit time, i.e., information processing speed), and is the Gaussian noise parameter in the dW Wiener process (a continuous-time stochastic process). These parameters vary across trials, and changes in any parameter affect the reaction time distributions for correct and incorrect responses [?, ?, ?, ?, ?, ?, ?]. For example, trials with larger drift rates show higher accuracy and shorter reaction times, while those with smaller drift rates show the opposite pattern. Research indicates that DDM can effectively provide precise quantitative explanations for binary perceptual decision outcomes [?, ?, ?, ?, ?, ?, ?, ?, ?], prompting researchers to analyze individual differences in various information processing stages to distinguish different populations and even identify abnormal information processing stages and their causes in different patient groups.

Recently, many researchers have used DDM to analyze different information processing stages [?, ?], including vision [?, ?], attention [?, ?], and emotion [?, ?, ?, ?, ?, ?, ?, ?, ?]. Mental illness research has also begun adopting DDM methods [?, ?, ?, ?, ?]. For example, researchers using DDM have identified cognitive, behavioral, and physiological markers of depression.

Pe et al. (2013) used an emotional flanker task to explore the relationship between rumination tendency, depressive tendency, and attentional bias. Results showed that rumination tendency rather than depressive tendency explained the depressed group' s attentional bias toward negative information. Dillon et al. (2015) used a traditional flanker task to examine dominant response bias (i.e., tendency to respond to distractors), response inhibition, and executive control functions in major depressive disorder individuals. They found that major depressive disorder participants had lower drift rates than healthy controls, showing slower responses but higher accuracy. This result is consistent with previous research, but the difference was not caused by simple speed-accuracy trade-offs because the decision threshold in the major depressive disorder group remained unchanged [?, ?]. Additionally, anhedonia scores were negatively correlated with drift rate. These results suggest that major depressive disorder

individuals have reduced dominant response bias and weakened executive control function, possibly related to decreased striatal dopamine secretion rather than simple speed-accuracy trade-offs. Overall, DDM can extract different components of decision processes from reaction times and accuracy rates—something traditional behavioral data analysis cannot achieve. Furthermore, DDMs fitted to observed data can estimate the values of different components driving behavior [**?**, **?**].

Compared to traditional group mean difference tests, DDM analysis improves result specificity and sensitivity. Regarding specificity, DDM can identify decision components that explain behavioral performance. Regarding sensitivity, DDM can detect subtle differences in reaction times and accuracy rates, even non-significant differences [**?**, **?**, **?**]. Recently, researchers have developed neural drift diffusion models (NDDM) that combine behavioral and brain imaging data [**?**, **?**, **?**, **?**]. Hierarchical drift diffusion models (HDDM) use hierarchical Bayesian methods to simultaneously estimate subject-level and group-level parameters, enabling estimation of parameter uncertainty in posterior distributions [**?**, **?**, **?**, **?**]. Additionally, Krajbich and Rangel (2011) extended DDM to three-choice scenarios. For more DDM principles, algorithms, and toolkits, see [**?**, **?**, **?**, **?**, **?**].

3.3 Generative Models

The Bayesian brain hypothesis posits that the brain constructs a generative model of incoming sensory information and continuously updates it. This continuous updating relies on energy minimization [**?**, **?**], known as the free-energy principle. The free-energy principle states that biological systems or individuals can maintain their states when facing changing environments [**?**, **?**], striving to avoid "surprise" that causes fluctuations in free energy and ensuring psychological states remain within physiological bounds (free energy) [**?**, **?**]. Specifically, human learning is driven by prediction errors arising from mismatches between existing beliefs and new input information, while mental illness stems from reduced precision in prior sensory information encoding, leading to maladaptive reasoning based on imprecise sensory information [**?**, **?**].

Generative models based on Bayesian theory can precisely measure unobservable neural and cognitive processes [**?**, **?**]. Currently, the most commonly used generative model is dynamic causal modeling (DCM), which focuses on using Bayesian model reduction—that is, performing Bayesian model inversion and comparison—to elegantly handle multiple models for a single dataset or multiple datasets for a single model [**?**, **?**]. Based on generative models and brain functional connectivity hypotheses, researchers have applied DCM to functional magnetic resonance imaging (fMRI) and magneto-/electroencephalography (M/EEG) data.

When DCM was first introduced for fMRI data analysis, Friston et al. (2003) proposed a dynamic model of neuronal activity:

$$\dot{x} = (A + \sum B^{(j)} u_j)x + Cu$$

This bilinear model describes neural state x dynamics as the effect of synaptic connections between neural nodes or brain regions (endogenous connectivity A) and experimentally controlled manipulations u on the system. Experimental manipulations may directly affect neuronal states (direct input C) or modulate effective connectivity between different nodes (modulatory input B) [?, ?]. The optimal DCM is determined by specific hypotheses about pathological processes. For example, nonlinear DCM is often used for abnormal synaptic plasticity problems caused by region-specific abnormal modulation [?, ?], while stochastic DCM and spectral DCM are commonly used to detect resting-state brain functional network abnormalities [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?], with spectral DCM showing higher prediction accuracy and greater sensitivity to group differences in effective connectivity [?, ?].

Many researchers have used DCM to analyze task-based fMRI, resting-state fMRI, MEG, and EEG data, finding abnormalities in neural networks related to visual, attentional, and emotional processing in major depressive disorder patients. In emotional processing, Lu et al. (2012) recorded MEG signals in major depressive disorder patients during a facial emotion task. Results showed significantly weakened top-down effective connectivity from dorsolateral prefrontal cortex to amygdala, while bottom-up connectivity from amygdala to anterior cingulate cortex and connectivity from anterior cingulate cortex to dorsolateral prefrontal cortex were significantly enhanced. Impaired frontal cortex effective connectivity to lower brain regions may cause emotional information processing dysfunction in major depressive disorder patients. In resting state, compared to controls, major depressive disorder patients showed enhanced causal effective connectivity from right insula, right putamen, and right caudate to rostral anterior cingulate cortex, but reduced causal effective connectivity from bilateral dorsolateral prefrontal cortex and left orbitofrontal cortex to rostral anterior cingulate cortex, again suggesting cognitive control impairment [?, ?]. In self-evaluation, major depressive disorder patients showed hyperactivation of effective connectivity from medial prefrontal cortex to posterior cingulate cortex, suggesting excessive "hyperregulation" of self-processing on emotion, which may be an important cause of self-reflection or rumination in major depressive disorder [?, ?, ?, ?, ?].

Overall, using DCM to fit data and using these fitted parameters (e.g., causal connection strength between two brain regions) as sufficient statistics can concisely describe data characteristics, potentially providing concise and effective indicators for depression diagnosis. Models generated from observed data allow model inversion—predicting brain activity from known psychological phenomena and inferring behavioral changes from brain activity. Generative models quantify model complexity through Bayesian model comparison, increasing model robustness, reproducibility, and generalizability [?, ?]. For more DCM principles, fMRI-based and MEG/EEG-based DCM, and its applications in mental illness, see [?, ?, ?, ?].

Integrating these three models, computational simulations of human perception,

attention, executive control, and decision-making processes can more finely link discovered neurophysiological bases with existing behaviors. Model parameters (specific processes) also provide concise and effective measurement indicators for subsequent research and intervention. Different models have unique advantages: at the behavioral level, DDM focuses on perception, attention, and simple decision-making processes during specific tasks, while reinforcement learning models focus more on the dynamic influence of reward/punishment processing and value processing on learning. At the neural level, neuroimaging-based DCM can simulate causal relationships between brain region activities under resting-state or task-state conditions. Therefore, systematic investigation of cognitive functions in depressed populations requires complementary advantages from different models. Although depression mechanism research continues, with accumulating data and evolving methods, its pathogenesis remains controversial. Thus, when many unknown influences exist, using data-driven methods (e.g., machine learning) to study depression may be a feasible approach.

4 Data-Driven Computational Psychiatry Research

The most commonly used method in data-driven computational psychiatry research is machine learning. The most frequently used categories in machine learning are supervised learning and unsupervised learning, primarily including classification, regression, and clustering methods. This article focuses on supervised learning methods, briefly introducing the principles of these three method types and applications of machine learning in depression diagnosis, prediction, and treatment evaluation.

In machine learning algorithms, each instance in a dataset is represented by the same set of features, which may be continuous, categorical, or binary. If instances have labels (known categories), the learning process is called supervised learning; otherwise, it is unsupervised learning. Supervised learning includes classification and regression methods. Classification methods use labeled data samples to build classifiers for predicting new samples' categories, while regression methods fit data to continuous functions to mark data with continuous variables. Thus, supervised learning often builds classifiers to distinguish patient and healthy groups or establishes prediction models to forecast depression occurrence.

The main steps of supervised learning are feature extraction (and selection), model training and testing, and model evaluation. First, researchers obtain behavioral, genetic, brain imaging, and other data from subjects (e.g., patients vs. healthy groups), forming a feature X dataset of N subjects $\times$ P variables, with target variable Y as known labels or categories, and perform feature extraction [**?**, **?**]. Researchers can screen features of interest using feature selection methods such as t-tests for classification or Pearson correlation for regression, or retain all features. Then, model training and testing are conducted using selected algorithms, such as the commonly used support vector machine (SVM) method (see below). Finally, model evaluation uses in-sample estimates from

training data to predict out-of-sample estimates for new data. Common model evaluation methods include cross-validation (CV): the training set is divided into K subsamples, with each subsample used for model validation while the remaining K-1 samples are used for training. Each subsample is validated once, and results from K iterations are averaged or transformed to produce a final single estimate model. This method is commonly called K-fold cross-validation, with 5-fold or 10-fold cross-validation being most widely used [?, ?]. Classification validation results can be evaluated using accuracy, sensitivity, specificity, precision, and area under the receiver operating characteristic curve (AUC); regression validation results require correlation coefficients, mean absolute error, mean squared error, etc. [?, ?].

Numerous specific machine learning algorithms exist, with selection depending on research questions. Current machine learning methods can be divided into classification algorithms, regression algorithms, and clustering algorithms. For other classification methods and their statistical principles, see [?, ?] and [?, ?]. In classification algorithms, k-nearest neighbors, decision trees, support vector machines, and AdaBoost can classify both nominal and numerical data, with support vector machines being widely used. A support vector machine is a hyperplane (n-dimensional plane) that effectively distinguishes different category samples in the training set while maximizing the margin between the hyperplane and the nearest training samples [?, ?, ?, ?]. This algorithm has high classification accuracy and good generalization ability, avoiding data overfitting [?, ?, ?, ?].

In regression algorithms, linear regression is commonly used, including least absolute shrinkage and selection operator (LASSO), ridge regression, relevance vector regression (RVR), and elastic net regression. Cui and Gong (2018) evaluated computational time and prediction accuracy of different regression algorithms, finding ridge regression had high efficiency and accuracy. Ridge regression minimizes prediction error and the sum of squared regression coefficients, removing unimportant parameters through shrinkage and better handling multicollinearity and overfitting [?, ?].

Classification and regression methods are commonly used in supervised learning, while clustering is often used in unsupervised learning, primarily aiming to find intrinsic distribution structures in datasets. The k-means algorithm is one of the most common clustering algorithms, requiring user-specified k random centroids. Each sample point is assigned to its nearest cluster centroid, causing centroid changes. This process repeats until centroids no longer change [?, ?, ?, ?, ?]. More efficient clustering algorithms such as bisecting k-means and hierarchical clustering are not detailed here.

Currently, supervised and unsupervised learning methods based on these algorithms have been widely applied to depression diagnosis, prediction, and treatment evaluation. The following sections primarily introduce clinical applications of supervised learning methods.

4.1 Disease Diagnosis and Prediction

Supervised learning applied to mental illness diagnosis and prediction includes three main stages. First, researchers write algorithms they believe can learn successfully. Second, they use these algorithms to train on (behavioral, physiological, brain imaging, etc.) data, calculating differences or distances between learning results and actual results (illness presence), adjusting internal parameters (weights) to minimize these differences and produce optimal learning results, marking data features as specific mental illnesses. Finally, they use these models to predict new data (potentially from patients or non-patients) [**?**, **?**, **?**, **?**]. If learning is successful, most or all labels will be correctly identified. Beyond these main steps, researchers often perform pattern localization on results, such as interpreting specific or important vectors or visualizing graphics showing feature weights for label prediction—weight maps.

In recent years, many studies have diagnosed and predicted major depressive disorder using machine learning methods. For example, Patel et al. (2015) used supervised learning with subjects' age, mini-mental state examination scores, and structural MRI data, applying alternating decision tree algorithms to predict late-life depression, achieving 88.89% sensitivity and 85.71% specificity—high prediction accuracy. Some studies have used social media text expressions to predict depression. Compared to naive Bayes and decision tree algorithms, support vector machine algorithms showed higher prediction sensitivity (83.3%) and specificity (82.6%) [**?**, **?**, **?**, **?**, **?**]. Another study used multi-view bi-clustering algorithms with multi-perspective sensory data as input to build models distinguishing depressed and healthy groups, validated with support vector machine classifiers, achieving 87% accuracy [**?**, **?**]. Additionally, some researchers have used unsupervised learning methods for depression diagnosis and prediction. Drysdale et al. (2017) performed resting-state fMRI on 1,188 depression patients and conducted hierarchical clustering analysis on abnormal functional connectivity data from limbic system and frontal-striatal networks, producing four neurophysiological subtypes (diagnostic classifiers). Through multisite validation and out-of-sample validation, clustering showed high sensitivity and specificity (82-93%).

However, these studies are cross-sectional. Better methods for testing prediction accuracy should use longitudinal studies, applying generated machine learning models to predict individuals' future depression risk based on past data and comparing predictions with actual outcomes to validate model accuracy. Additionally, increasing research has focused on how individual patients' brain imaging data can predict their future mental illness probability, involving Bayesian model selection and generative embedding processes—see [**?**, **?**] for details.

4.2 Treatment Effect Prediction and Scheme Selection

Researchers have extended machine learning methods beyond distinguishing depressed patients from healthy individuals to treatment effect determination and

treatment scheme selection. Previous studies show different antidepressants have varying efficacy across patients, but this individual difference is important for personalized treatment selection. Current research has begun using univariate or multivariate markers to predict treatment effects. For example, DeRubeis et al. (2014) examined treatment efficacy across different patient groups, finding that married, employed individuals with rich experience but antidepressant tolerance could alleviate depressive symptoms through cognitive behavioral therapy (CBT), while comorbid personality disorder and depressive disorder patients responded better to antidepressant medication. Williams et al. (2015) used a single marker—amygdala response to supraliminal and subliminal emotional face stimuli—to predict specific drug efficacy for different depression patients. Results showed that patients with weakened amygdala activation when processing subliminal rewarding and threatening faces indicated general antidepressant sensitivity (Cohen' s d = 0.63-0.77), while patients with hyperactivated amygdala when processing subliminal sad faces indicated resistance to serotonin-norepinephrine reuptake inhibitors (Cohen' s d = 1.5). Thus, subliminal face processing as a cognitive marker can effectively predict patients' sensitivity to specific antidepressants.

Beyond using single markers to test and predict drug treatment effects, many studies have attempted to use more dimensional markers to improve prediction accuracy. Rush et al. (2006) conducted a longitudinal study comparing depression treatment efficacy, noting that among approximately 4,000 major depressive disorder patients, over 50% were sensitive to citalopram, showing higher remission rates and lower relapse rates in subsequent stages, while insensitive patients still showed low remission rates and high relapse rates even after switching medications. Additionally, researchers analyzed symptoms in 4,041 depression patients, using machine learning models to determine which individuals would benefit most from specific antidepressants. The final model considered 164 latent variables involving somatic symptoms, insomnia, exposure time to past traumatic events, and other aspects, achieving good prediction accuracy [**?**, **?**, **?**, **?**, **?**].

Based on treatment effect prediction, machine learning methods can establish optimal treatment plans for patients. DeBattista et al. (2011) used referenced-electroencephalogram (rEEG) to predict drug efficacy for patients. Compared to Rush et al.'s (2006) medication-switching treatment plan based on drug efficacy, using rEEG data (over 1,800 patients, total 405 days, over 17,000 drug trials, 74 available biomarkers) through automatic treatment-selection algorithms significantly improved treatment outcomes. If these results receive further validation, rEEG-based schemes will greatly enhance treatment efficiency and become a simple, inexpensive, objective, and predictive antidepressant selection procedure.

Overall, exponentially growing artificial intelligence-based mental illness research has shown certain effectiveness in depression diagnosis, prediction, and treatment [**?**, **?**]. Currently, depression diagnosis and treatment typically use neuroimaging data, wearable device sensor data, social media data, and survey

data, employing support vector machine methods and regression algorithms to build classifiers or prediction models [?, ?, ?, ?], with these models achieving relatively high prediction accuracy. Additionally, in depression treatment evaluation, many researchers use neuroimaging data, clinical treatment data, survey data, and common drug and biomarker data, detecting prediction model accuracy from different algorithms through longitudinal tracking. Currently, such methods are increasingly used clinically, though further promotion requires more mature algorithms and deeper research on depression pathological mechanisms.

5.1.1 Focus on Individuals

Previous studies typically compared depressed and healthy groups to identify abnormal physiological and psychological functions in depressed populations, yielding certain research findings (e.g., \cite{孙也婷等，印刷中}; \cite{文宏伟，陆菁菁，何晖光，2018}). However, with deepening research and new technologies, this group-comparison approach is inefficient and yields unsatisfactory results for studying complex mental illnesses. Depression arises from interactions among genetic, cognitive, emotional, and environmental factors, with each factor' s effects varying by individual circumstances (e.g., age, personal experience, family environment, even cultural background). Therefore, treatment efficacy may also vary by individual, meaning treatment plans effective for early-adult depression patients may not apply to elderly or adolescent patients [?, ?, ?, ?]. Compared to traditional psychiatric research, computational psychiatry methods focus more on individuals, can identify individuals at depression risk, and can locate core features that may cause depression based on their prediction models. Thus, this "personalized treatment" can more efficiently help individuals achieve physiological and behavioral improvements [?, ?, ?, ?].

5.1.2 Full Utilization of Research Data

Robinson and Chase (2017) argue that using computational modeling methods requires quantitatively described research hypotheses, which facilitates using parameters to simulate single or multiple separate processes. This approach more fully utilizes obtained data and more accurately predicts human behavior. First, comparisons among multiple prediction models can determine model applicability and estimate model fitting strength [?, ?]. Second, for mental illness patients, research on cognitive and learning processes is particularly important as these processes are closely related to the emergence and maintenance of many mental illnesses, including depression [?, ?, ?, ?, ?], anxiety [?, ?, ?, ?], and schizophrenia [?, ?, ?, ?, ?]. Learning processes are constantly changing, requiring monitoring of trial-by-trial variance in choices, which computational modeling can achieve. Traditional psychological analysis methods generally reflect data characteristics through means and standard deviations of reaction times and accuracy rates, ignoring some subtle or dynamic changes. Third, computational models can link behavioral observation indicators with physio-

logical and neural-level changes, providing valuable information for explaining internal mechanisms underlying behavior. Finally, computational models provide parameters for complex cognitive and decision-making processes, such as meta-parameters commonly used in pathology or variance of model parameters.

In summary, computational modeling methods process multi-level, multi-system data rather than relying solely on single self-reports and clinical interviews, promoting researchers' understanding of depression pathological mechanisms and improving diagnostic and treatment objectivity. Second, computational modeling methods can provide detailed descriptions of individual depression severity and symptom characteristics (e.g., comorbidity), partially overcoming diagnostic and treatment difficulties caused by comorbidity and heterogeneity. Finally, computational model parameters can provide quantitative descriptions for specific psychological processes or symptom manifestations (e.g., learning rates in reward processing) rather than dichotomous results, allowing clinicians to judge individual depression severity based on these values and propose more precise treatment plans.

5.2.1 Theory-Driven Research

Although projects studying depression and other mental illnesses using computational modeling methods are gradually increasing, and these methods can explain how different levels and their interactions from biological molecules to behavioral environments affect depression's emergence, development, maintenance, and recovery, their effectiveness still faces skepticism—are computational models mature enough to adequately determine and predict mental illness [**?**, **?**, **?**, **?**]?

First, while models can reveal patient response biases, preferences, or value updating processes through different parameter assignments [**?**, **?**, **?**, **?**], these parameters cannot cover all aspects of a psychological process. Second, the biological significance of some model parameters is ambiguous. Different researchers understand the same parameter from different perspectives, but can these parameters truly match concepts in psychiatry, psychology, or neurobiology [**?**, **?**]? For example, in reinforcement learning models, Husain and Roiser (2018) propose that the value in the Softmax function based on reinforcement learning models can be interpreted as inconsistency or exploration, but what psychological process does actually represent? Which neurophysiological indicators does it correspond to? Third, are findings from theory-driven research reproducible [**?**, **?**, **?**, **?**]? As mentioned above, is reward prediction error processing function normal or abnormal in depressed populations during model-free learning? Solving these problems depends both on mathematical model development itself—such as whether model parameters clearly reflect psychophysiological processes and whether examined data dimensions are comprehensive—and on interdisciplinary researcher collaboration to update dynamic pathological mechanism models of depression from molecular (genes, cells), neural and endocrine system, and behavioral/environmental levels. Finally, previous studies have described characteristics of cognition and decision-making in depression patients

using physiological, brain imaging, and behavioral indicators, but these studies often only reveal correlations [?, ?]. In clinical neuroscience, using TMS or transcranial direct-current stimulation (tDCS), pharmacological manipulations or treatments combined with imaging techniques can reveal causal relationships across multiple levels and systems including biological molecules, neural systems, and behavior—for example, rTMS and tDCS stimulation of left dorsolateral prefrontal cortex can also alleviate symptoms of depression or bipolar disorder [?, ?, ?].

5.2.2 Data-Driven Research

Depression research using traditional methods has progressed slowly, possibly because depression and similar diseases have substantial complexity and heterogeneity, while machine learning methods can extract main features from complex phenomena, potentially overcoming difficulties caused by depression' s complexity and heterogeneity. Currently, machine learning-based diagnosis and treatment evaluation research for depression and other mental illnesses has shown initial success but still has limitations. First, in supervised learning methods, sample labels are crucial for machine learning modeling. For example, researchers typically use clinical scales to determine whether individuals have depression, such as the Hamilton Depression Scale [?, ?, ?, ?]. However, as mentioned above, symptom-based disease diagnosis is ineffective, and using this result to determine sample labels already carries defects of symptom-based methods—the labels themselves may be inaccurate. Using inaccurate labels to hope for accurate prediction results may be unfeasible. Moreover, single categorical or continuous labels may lose much additional information (e.g., comorbidity conditions), cannot comprehensively reflect data characteristics, and may easily cause bias in machine learning processes, leading to gaps between model predictions and real results. In response, some researchers propose using unsupervised learning methods to improve this situation. Unsupervised learning, based on existing datasets without label input, uses clustering and association analysis for dimensionality reduction, reducing bias caused by human definition or classification. Additionally, when sample sizes are large and data dimensions are numerous, deep learning methods (e.g., recurrent neural networks, convolutional neural networks) learn sample features without manual coding, and the hierarchical network structures formed through learning can improve classification and prediction accuracy to some extent. Therefore, for identifying complex and heterogeneous mental illnesses like depression, accurate labels, sufficient information, and appropriate methods can better explain various subtypes and comorbidity phenomena, thereby developing corresponding diagnostic and treatment plans and preventive measures—this may also be a future development trend.

Second, machine learning requires large samples; small sample sizes easily cause large measurement errors and model variability [?, ?], while research data sharing can better solve sample size problems. Additionally, overly homogeneous

samples cause model identification results to correlate with samples rather than the disease itself [**?**, **?**]. This raises new questions: can original model prediction effectiveness be maintained when using different imaging instruments or applying to new samples? Since machine learning cross-validation uses two homologous subsamples, whether models generated in new samples from different sources (e.g., different genders, ages, education levels, cultural backgrounds) have generalizability and reliability still requires validation with completely new samples. Furthermore, researchers may selectively report only high-prediction models for publication, creating a risk that machine learning model prediction accuracy may be overestimated—a concern requiring researcher attention.

5.2.3 Combining Theory-Driven and Data-Driven Methods

Both data-driven and theory-driven methods are processes of appropriate statistical inference based on datasets from measurements to draw conclusions. Theory-driven methods only perform data dimensionality reduction when selecting measurement indicators based on researchers' experience and knowledge, while data-driven methods typically measure all possible variables and use machine learning for dimensionality reduction or global analysis, resulting in relatively higher explanatory or predictive rates. Currently, researchers have limited theoretical understanding of depression and other mental illnesses, and excessive reliance on past experience often leads to biased results and low explanatory rates. However, dimensions or factors obtained through machine learning are often difficult to interpret, and measurement processes require substantial work —important reasons why these methods cannot be well applied in research and clinical fields. Huys et al. (2016) argue that when theory-driven model parameters are sufficient statistics, they can capture fundamental patterns driving complex observation processes, maximizing data-driven effectiveness. Therefore, combining theory-driven and data-driven methods can improve research and application efficiency and reliability [**?**, **?**]. Researchers can integrate existing knowledge and experience based on dimensionality reduction results from machine learning algorithms, analyze causes of dimension formation and contribution rates of each dimension, deeply understand depression pathological mechanisms and core features, and provide theoretical guidance for depression prevention and treatment. Researchers can also use more mature and appropriate algorithms, such as deep learning, (broad sense) reinforcement learning, and even transfer learning, to better fit or construct depression pathological mechanisms, providing precise prevention and treatment plans.

5.2.4 Integration of Talent and Resources

Computational psychiatry research requires interdisciplinary and multi-method integration, but few people master multiple disciplines. Therefore, such research urgently needs interdisciplinary talent cooperation and exchange, including joint laboratory construction, instrument and technology sharing, and complementary research methods \cite{谢小华，冯建峰，2019}.

Depression computational psychiatry research aims to generate new diagnostic schemas and propose precise treatment plans, but achieving this goal still requires time [?, ?, ?, ?]. It must be emphasized that computational psychiatry development cannot be separated from drug development, the development cycle of biomarkers, drugs, and medical devices, and the development and complementarity of preclinical research, clinical laboratory development, longitudinal and cross-sectional studies, prospective studies, and randomized controlled trials [?, ?, ?, ?, ?, ?, ?, ?]. Computational psychiatry researchers need to cultivate interdisciplinary awareness, understand methods and latest developments in different disciplines, leverage their own disciplinary strengths, cooperate with researchers from other disciplines, and deeply understand depression pathological mechanisms from multiple levels.

Additionally, data and code resource sharing is crucial. Data-driven research datasets have large dimensions, and statistical demands for sample size increase accordingly. Individual researchers have limited resources, and obtained data volumes often fail to meet requirements. To obtain accurate prediction models using machine learning, researchers need to increase sample size while collecting multi-level information. Currently, collecting patient data from multiple levels including genes, neurotransmitters, neural systems, brain, behavior, and environment requires substantial human and material resources, urgently requiring data resource sharing. This requires researchers from multiple fields to collect physiological, brain imaging, behavioral indicators, and other data from patients according to the same standards and parameters, attach data analysis code, and share with other researchers. This approach not only improves data resource utilization but also promotes optimization of various algorithms and facilitates reproducibility testing of computational psychiatry research results.

In summary, computational psychiatry research has promoted in-depth understanding of depression pathogenesis and patients' cognitive processing. Computational psychiatry methods focus on individuals, fully utilize individuals' multi-level data, making diagnosis objective, precise, and individually specific, but also face challenges such as difficult model parameter interpretation, insufficient algorithm optimization, and inconvenient data collection. However, as long as researchers from different disciplines leverage their expertise, cooperate and exchange, share resources, innovate methods, and complement each other' s advantages, depression computational psychiatry research is expected to reveal depression pathology from multi-level, multi-system interactions from micro to macro levels. Its research results are expected to translate into precise prevention and treatment for depression patients, ultimately reducing depression incidence and improving people' s mental health levels.

孙也婷, 陈桃林, 何度, 董再全, 程勃超, 王淞, . . . 龚启勇. (印刷中) 基于精神影像和人工智能的抑郁症客观生物标志物研究进展. 生物化学与生物物理进展, 1-45.

文宏伟, 陆菁菁, 何晖光. (2018). 机器学习在神经精神疾病诊断及预测中的应用. 协和医学杂志, 9(1), 19-24.

谢小华, 冯建峰. (2019). 上海市脑与类脑智能基础转化应用研究的现状及展望. 心理学通讯, 2(02), 84-87.

Note: Figure translations are in progress. See original paper for figures.