ARPNet: Antidepressant Response Prediction Network for Major Depressive Disorder

Buru Chang [1,†], Yonghwa Choi [1,†], Minji Jeon [1], Junhyun Lee [1], Kyu-man Han [2], Aram Kim [3], Byung-joo Ham [2,4,*] and Jaewoo Kang [1,5*]

[1] Department of Computer Science and Engineering, Korea University, Seoul, Korea; buru_chang@korea.ac.kr (B.C.); yonghwachoi@korea.ac.kr (Y.C.i); mjjeon@korea.ac.kr (M.J.);ljhyun33@korea.ac.kr (J.L.);

[2] Department of Psychiatry, Korea University Anam Hospital, Korea University College of Medicine, Seoul, Korea; han272@korea.ac.kr (K.-M. H.)

[3] Department of Biomedical Sciences, Korea University College of Medicine, Seoul, Korea; chloe_ark@korea.ac.kr (A.K.)

[4] Brain Convergence Research Center, Korea University Anam Hospital, Seoul, Korea

[5] Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul, Korea

[*] Correspondence: hambj@korea.ac.kr (Byung-joo Ham); kangj@korea.ac.kr (Jaewoo Kang)

[†] These authors contributed equally to this work.

Abstract: Treating patients with major depressive disorder is challenging because it takes 6-8 weeks for antidepressants prescribed for the patients to take effect. This limitation may result in increased risks and treatment costs. To address this limitation, an accurate antidepressant response prediction model is needed. Recently, several studies have proposed models that extract useful features such as neuroimaging biomarkers and genetic variants from patient data, and use them as predictors for predicting the antidepressant responses of patients. However, it is impossible to utilize all the different types of predictors when making a clinical decision on what drugs to prescribe for a patient. Although a machine learning-based antidepressant response prediction model has been proposed to overcome this problem, the model cannot find the most effective antidepressant for a patient. Based on a neural network, we propose an Antidepressant Response Prediction Network (ARPNet) model capturing high-dimensional patterns from useful features. Based on a literature survey and data-driven feature selection, we extract useful features from patient data, and use the features as predictors. In ARPNet, the patient representation layer captures patient features and the antidepressant prescription representation layer captures antidepressant features. Utilizing the patient and antidepressant prescription representation vectors, ARPNet predicts the degree of antidepressant response. The experimental evaluation results demonstrate that our proposed ARPNet model outperforms machine learning-based models in predicting antidepressant response. Moreover, we demonstrate the applicability of ARPNet in downstream applications in use case scenarios. Our source code that implements ARPNet is available at http://github.com/dmis-lab/arpnet.

Keywords: major depressive disorder; antidepressant response prediction; patient representation; neural network

1. Introduction

Treatment of major depressive disorder (MDD), which is one of the most common mental illnesses, is challenging. The main reason is that it takes a long time (6-8 weeks) for the effects of an antidepressant to be seen in a patient. It also takes several months for psychiatrists to find that some drugs are ineffective. As a result, only 11–30% of patients reach clinical remission from initial

treatment [1–4]. This limitation of MDD treatment may result in increased risks and treatment costs. To address this limitation, an accurate prediction model that can predict antidepressant responses is needed.

Recently, several studies that extract useful predictors for predicting the antidepressant responses from patient data such as genetic and brain imaging information [5,6] have been proposed. For example, neuroimaging biomarkers [7], genetic variants [8–12], and DNA methylation [13] have been proven to be effective in improving prediction performance. However, it is impossible for even an experienced psychiatrist to utilize all predictors to make a clinical decision on what drugs to prescribe for a patient.

To overcome this limitation, a machine learning-based antidepressant response prediction model [14] has been proposed. Machine-learning based approach can be used to detect high-dimensional patterns of all potential predictors extracted from various patient data. This model extracts useful predictors from the data of the multicentre clinical trial on MDD (STAR*D) [15–18] and predicts the antidepressant "citalopram" response in MDD patients. Using the Elastic network model [19], this model first selects useful predictors such as sociodemographic features and depression severity checklists from patient data for predicting antidepressant responses. The model then predicts whether a patient will reach clinical remission based on the 25 selected predictors.

However, the above antidepressant response prediction model has some limitations. The prediction model is designed for predicting the response of only one antidepressant, "citalopram" (**Limitation 1**). In addition, the model only predicts whether the patient is responding to the antidepressant. The model cannot measure the degree of antidepressant response (**Limitation 2**). This model is limited as it cannot be used for finding the most effective drugs for MDD patients among many antidepressant candidates (**Limitation 3**).

To address these limitations, we propose an **A**ntidepressant **R**esponse **P**rediction **Net**work (ARPNet) model based on a neural network architecture. Our proposed ARPNet model is designed for real-world antidepressant prescription scenarios where the prescription consists of a combination of one or more antidepressants (related to **Limitation 1**). When an antidepressant combination and patient data including the demographic, genetic, and brain MRI information of a patient are given, ARPNet can predict the patient's Hamilton Depression Rating Scale (HAM-D) score at the next visit (related to **Limitation 2**). Our ARPNet model is capable of prescribing the most effective antidepressant combination by predicting not only whether the patient will respond to the prescribed antidepressant combination but also the degree of response (HAM-D score) (related to **Limitation 3**). Moreover, ARPNet trains patient representation vectors that capture patient characteristics, and antidepressant representation vectors that capture the features of antidepressants. The trained patient and antidepressant representation vectors could be used in downstream applications such as drug discovery for new antidepressant development and patient clustering in clinical decision support systems. The vectors can also be used to find similar patients who are likely to respond to the same drugs.

The contributions of our study are summarized as follows.

- We propose an antidepressant response prediction network model called ARPNet based on a neural network architecture. Unlike the previous models, our proposed model predicts whether the patient will reach clinical remission from depression, and the degree of antidepressant response. By predicting the degree of antidepressant response, our proposed model can prescribe the most effective antidepressants.
- ARPNet can be used in a real-world antidepressant prescription scenario where psychiatrists prescribe one or more antidepressants to patients with MDD. ARPNet predicts the antidepressant response to a given combination of antidepressants, which could not be done by previously proposed models.
- Through the literature-based and data-driven feature selection process, ARPNet extracts useful features from patient data including the demographic, genetic, and brain MRI information of patients.

- The trained patient and antidepressant representation vectors of ARPNet could be utilized in various downstream applications such as drug discovery and clinical decision making.

The remainder of this paper is organized as follows. In Section 2, we introduce the data, including the prescription records of patients with MDD, which is used for antidepressant response prediction. We describe the feature selection process which involves extracting useful predictors from patient data. We then describe our proposed ARPNet model which uses selected predictors and the prescription records of patients. In Section 3, we conduct experiments to assess ARPNet, and analyze the experimental results. In Section 4, we conclude the paper.

2. Materials and Method

In this section, we first describe the MDD patient data used in our study. Next, we introduce our feature selection process which involves selecting useful features as predictors for improving performance in antidepressant response prediction. Then, we state the antidepressant response prediction task. Last, we introduce our proposed model ARPNet.

2.1. Data Description

In this study, we use the data of 121 patients with MDD collected from Korea University Anam Hospital and Kyung Hee University Hospital, Seoul, Korea. All the patients who are ethnically Korean were examined by trained psychiatrists using a structured clinical questionnaire. The depression severity of the patients was measured using the 17-item Hamilton Depression Rating (HAM-D 17) scale at 0 (baseline), 4, and 8 weeks. In this paper, we employ the HAM-D 17 scale for the predictions on the degree of the antidepressant response. The collected MDD patient data includes the demographic, genetic, and brain MRI information of only patients who have consented to the use of their data for this study. Some patients did not consent to the use of certain data, so the data available for each patient may vary. If one patient's i-th feature is missing, we randomly choose one of the i-th features of all the remaining patients, and use it as the i-th feature of the patient. Specifically, 67, 71, 96, and 91 patients consented to the use of their demographic, brain MRI , and genetic information, respectively.

We confirm that the Institutional Review Board of Korea University Anam Hospital approved this study (ED16162). The statistics are summarized in Table 1.

Table 1. Selected features.

Neuroimaging Biomarkers		Genetic Variants		DNA Methylation		
Feature Name	Coefficient	Feature Name	Coefficient	Feature Name	Gene Name	Coefficient
lh_G_front_inf_Orbital_thickness	−5.0191	HTR1A:p.Ala155Gly	−7.2980	cg03829016	SLC6A4	−89.1863
lh_G_cingul_Post_dorsal_thickness	−3.9096	PAPLN:p.Gly838Glu	−7.0918	cg17075252	BDNF	−64.8215
lh_G_front_middle_thickness	−3.2466	TPH2:p.Arg225Gln	−6.3323	cg06373684	SLC6A4	−62.5500
lh_G_cingul_Post_ventral_thickness	−3.1023	ABCB1:p.Ala599Thr	−5.8213	cg26741280	SLC6A4	−58.5697
rh_G_insular_short_thickness	−2.5389	TNFSF14:p.Ala77Val	−4.1921	cg15462887	BDNF	−56.0164
rh_G_and_S_cingul_Mid_Ant_thickness	−2.2651	SCN5A:p.Pro1090Leu	−3.0044	cg18354203	BDNF	−50.8337
lh_G_and_S_cingul_Mid_Post_thickness	−1.8986	OPRM1:p.Gln402His	−2.8859	cg10241426	SLC6A4	−50.3451
lh_G_oc_temp_lat_fusifor_thickness	−1.8417	CYP2D6:p.Gly169Arg	−2.4712	cg06260077	BDNF	−47.2429
rh_G_front_middle_thickness	−1.5880	MTHFR:p.Ile248Val	−2.4050	cg07919246	BDNF	−46.7545
lh_G_insular_short_thickness	−1.5748	BDNF:p.Arg109Gln	−2.4010	cg15014679	BDNF	−38.1958
rh_G_cingul_Post_ventral_thickness	1.2208	BMP7:p.Arg154Gln	3.3401	cg10558494	BDNF	43.3355
rh_G_parietal_sup_thickness	1.4562	FKBP5:p.Val437Phe	3.3898	cg16737991	IL11	45.6990
lh_G_and_S_cingul_Mid_Ant_thickness	1.9680	IL11:p.Arg98Pro	3.4588	cg04672351	BDNF	48.9424
rh_G_and_S_cingul_Mid_Post_thickness	2.1854	OPN1SW:p.Ile302Val	3.5439	cg06961290	SLC6A4	50.3463
lh_G_oc_temp_med_Lingual_thickness	2.7142	TPH2:p.Ser41Tyr	3.7021	cg17882499	BDNF	59.1599
rh_G_front_inf_Opercular_thickness	3.3577	DRD4:p.Ala84Thr	3.7587	cg07238832	BDNF	70.9359
rh_G_cingul_Post_dorsal_thickness	4.3347	CDH17:p.Tyr79Cys	4.4870	cg11241206	BDNF	87.2378
rh_G_front_inf_Triangul_thickness	5.1329	TPH1:p.Arg248*	4.5572	cg07159484	BDNF	96.8888
rh_G_front_inf_Orbital_thickness	6.6912	RORA:p.Pro15Leu	7.2040	cg05016953	SLC6A4	111.0701
lh_G_parietal_sup_thickness	9.0480	PML:p.Arg755His	8.3033	cg01636003	BDNF	181.9174

2.2. Feature Selection

We use a two-step feature selection process to extract the most useful features from patient data. The first step involves selecting features identified by a literature survey. The second step involves further refining the features identified in the first step using data-driven methods. We do not perform feature selection on demographic information because it already contains features considered useful by psychiatric experts.

2.2.1. Literature-Based Feature Selection

From the brain MRI and genetic information, we select the following three literature-based features.

Neuroimaging biomakers Neuroimaging biomarker is consist of brain cortical thickness, which was automatically estimated using FreeSurfer (software version 5.0, http://surfer.nmr.mgh.harvard.edu). Using brain MRI information, we conducted a literature survey to identify neuroimaging biomarkers of treatment response in patients with MDD [7]. Among the reported biomarkers, we found 94 features in our MRI information. We selected 62 MRI features excluding 32 features which are not found in more than 30% of patients.

Genetic variants Several studies [8–12] have been conducted to identify the effects of genetic variants on antidepressant response. The studies report that variation of 27 significant genes such as 5-HTTLPR, BDNF, and SLC6A4 is associated with antidepressant response.

We found additional genes using biomedical literature-based tools such as BEST [20] and VarDrugPub [21]. BEST, which is a biomedical entity search tool, returns biomedical entities frequently mentioned with query terms in abstracts of published articles. We searched for all the antidepressant prescriptions on BEST, and obtained the top 23 ranked genes. We also found 12 genes related to antidepressants using VarDrugPub, which is a database containing relationships among genomic variants, drugs, and diseases extracted from the literature.

DNA methylation Genetic variants and epigenetic modification known as DNA methylation [13] are known to affect the treatment for depression. The authors reported that SLC6A4, BDNF, IL11, and MAOA genes and their CpG sites are associated with various antidepressants. We selected 136 CpG sites as DNA methylation features from the four genes.

2.2.2. Data-Driven Feature Selection

We identified and extracted useful features from the literature and filtered unnecessary features using the available APIs. Using Elastic net [19] which is one of the popular regression analysis methods, we select the more informative features among the extracted features for antidepressant response prediction. The estimates from the elastic net method are defined as follows.

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|Y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1), \tag{1}$$

where X is an input vector, Y is a true output value and β is the weight of the input vector. We used Elastic net to select the more informative features for treatment of patients with MDD from each of the three feature groups. We concatenate each feature with a one-hot encoded 14-dimensional antidepressant vector as input X where the number of antidepressants in our patient data is 14. The difference between HAM-D scores at initial and last visits of patients are used as output Y.

For each feature group, we split the data into 5 subsets and used one subset as the test set and the remaining subsets as the training set to reduce the risk of overfitting. Then we conducted 10-fold cross validation on the training set to find the hyper parameters of each Elastic net-based feature selection model. After selecting hyper parameters for each feature selection model, we validated the models on the test set and coefficients of the feature. By averaging the coefficients of the features from the

5 subsets, we select 20 features with the 10 highest coefficients and the 10 lowest coefficients for each of the feature groups. The coefficients of the selected features are summarized in Table 2, respectively.

Table 2. Statistics of feature selection data.

Feature	# of Features at First-Step	# of Features at Second-Step
Demographic information	127	127
Neuroimaging biomarkers	62	20
Genetic variants	156	20
DNA methylation	136	20

2.3. Problem Statement

When a patient sees a psychiatrist, the psychiatrist may prescribe an antidepressant to the patient based on the data of the patient. To prescribe the most effective antidepressant, the psychiatrist predicts the antidepressant response of the patient. Antidepressant response prediction is defined with this real-world antidepressant prescription scenario in mind.

Based on data P including the demographic, brain MRI, and genetic information of a patient, the current HAM-D score H of the patient, and the time interval ΔT until the next visit, antidepressant response prediction models predict the HAM-D score of the patient at the next visit for the given antidepressants A. The demographic information includes height, weight, and sext information of patients. Real-valued information such as height and weight is represented with log-scale. And discriminated information such as sex and occupation is encoded with a one-hot vector. The predicted HAM-D score $\hat{Y}$ is computed as follows:

$$\hat{Y} = f(P, H, \Delta T, A | \theta), \tag{2}$$

where θ is a set of parameters of the prediction model. The parameters are trained in the training phase to minimize the loss that is defined in the next subsection.

2.4. ARPNet

Figure 1 shows the architecture of ARPNet. Our ARPNet model consists of the following three components: a patient representation layer, an antidepressant prescription representation layer, and a prediction layer. The notations used in this paper are summarized in Table 3.

Figure 1. Architecture of our proposed ARPNet model.

Table 3. Notations.

Notation	Description
	Patient Representation Layer
$P, \mathcal{P}$	Patient and set of patients
X_P	Input vector of the representation layer
H_P	Current HAM-D score of the patient
V_{demo}	Demographic information of the patient
V_{bio}	Neuroimaging biomarkers of the patient
V_{gene}	Genetic variants of the patient
V_{methyl}	DNA methylation of the patient
ΔT_P	Visit interval between two consecutive visits of the patient
W_P, b_P	Weight matrix and bias term of the patient representation layer
V_P	Patient representation vector
	Antidepressant Prescription Representation Layer
$A, \mathcal{A}$	Antidepressant and set of antidepressants
E_A	Antidepressant representation matrix
X_A	Input of the antidepressant prescription representation layer
$\hat{X}_A$	Lookuped antidepressant prescription representation
V_A	Antidepressant representation vector
	Prediction Layer
X	Input vector of the prediction layer
$Y, \hat{Y}$	True and predicted HAM-D score
W, b	Weight matrix and bias term of the prediction layer

2.4.1. Patient Representation Layer

The patient representation layer is designed to capture an abstractive patient feature from patient data. The input $X_P \in \mathbb{R}^{127+20+20+20+1+1=189}$ of this layer is constructed as the concatenation of the current HAM-D score $H \in \mathbb{R}$ and time interval $\Delta T \in \mathbb{R}$, which are the useful features selected in Section 2.2, as follows.

$$X_P = [V_{demo}; V_{bio}; V_{gene}; V_{methyl}; H; \Delta T], \tag{3}$$

where $V_{demo} \in \mathbb{R}^{127}$, $V_{bio} \in \mathbb{R}^{20}$, $V_{gene} \in \mathbb{R}^{20}$, and $V_{methyl} \in \mathbb{R}^{20}$ are the representation of demographic information, neuroimaging biomarkers, genetic variants, and DNA methylation, respectively.

To extract the abstractive patient feature from interactions between selected features, the concatenated input X_p is forwarded to a non-linear layer with a non-linear function such as ReLU, tanh, and sigmoid. In this paper, we use ReLU as the non-linear function.

$$V_P = \sigma(X_P \cdot W_P + b_P), \tag{4}$$

where $V_P \in \mathbb{R}^d$ is the d-dimensional patient representation vector, and $W_P \in \mathbb{R}^{189 \times d}$ and $b_P \in \mathbb{R}^d$ are the weight matrix and bias term of the patient representation layer, respectively. $\sigma(\cdot)$ is the ReLU function.

The two major contributions of the computed patient representation vector are summarized as follows. The patient feature representation is automatically trained from patient data to improve the performance in antidepressant response prediction. Moreover, the similarities between patients could be computed by representing patients as real-valued vectors. The similarities between patients are useful in downstream applications such as a clinical decision support system. For example, physicians can find patients who are similar to a given patient using patient representation vectors. The patient representation vectors are trained for antidepressant response prediction. Therefore, patients who have similar vector representations are predicted to respond to the same drugs.

2.4.2. Antidepressant Prescription Representation Layer

A prescription for a patient consists of one or more antidepressants. The prescription representation layer is designed to capture antidepressant features from one antidepressant or the combination of antidepressants. We first associate each antidepressant $A \in \mathcal{A}$ with a d'-dimensional real-valued vector. $\mathcal{A}$ is the set of antidepressants in our patient dataset. We denote an antidepressant representation matrix that consists of the d'-dimensional antidepressant vectors as $E_A \in \mathbb{R}^{|\mathcal{A}| \times d'}$. The prescription is represented as a one-hot vector as follows.

$$X_A = [x_{A_1}, x_{A_2}, ..., x_{A_{|\mathcal{A}|}}], \tag{5}$$

$$x_{A_i} = \begin{cases} 1, & \text{if the } i\text{-th antidepressant is in the prescription,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Finally, the prescription representation is computed as follows.

$$\hat{X}_A = E_A(X_A), \tag{7}$$

$$V_A = Sum(\hat{X}_A), \tag{8}$$

where E_A is the antidepressant representation look-up function, and $\hat{X}_A$ denotes the representations of the drugs in the prescription X_A. $V_A \in \mathbb{R}^{d'}$ is the sum of the antidepressant representations in $\hat{X}_A$ and is the final antidepressant prescription representation that is utilized in the prediction layer described in the next subsection.

2.4.3. Prediction Layer

The prediction layer is a linear regression approach based on the patient representation and the prescription representation. An input of this layer is constructed by concatenating the patient representation and the prescription representation as follows.

$$X = [V_P; V_A], \tag{9}$$

where $X \in \mathbb{R}^{d+d'}$ is the input of the prediction layer. Then, the input X is forwarded to the regression layer as follows.

$$\hat{Y} = X \cdot W + b, \tag{10}$$

where $W \in \mathbb{R}^{(d+d') \times 1}$ and $b \in \mathbb{R}^1$ are the weight matrix and the bias term of the regression layer, respectively.

To optimize the parameters θ of our proposed model, we define the loss function of our model using Mean Square Error (MSE) as follows.

$$MSE = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (Y_i - \hat{Y}_i)^2, \tag{11}$$

where Y_i is a true HAM-D score of the i-th prescription record data. The model parameters are randomly initialized with a Gaussian distribution and trained by minimizing the MSE using the Adam optimizer [22] with a learning rate 0.001 and a batch size 16. To address the overfitting problem, we apply dropout [23] and L2 regularization. Based on cross-validation which will be described in the next section, we find the optimal hyperparameter setting of our proposed ARPNet with grid search.

We implemented our proposed model using TensorFlow v1.6.0 (https://www.tensorflow.org) and Python v2.7.12 (https://www.python.org). Our experiments were conducted on Ubuntu 16.04.5 with two TITAN X (Pascal) GPUs.

3. Experiments and Results

In this section, we evaluate our proposed ARPNet model by comparing it with several machine learning-based baseline models. We conduct experiments to answer the following research questions:

RQ1. Does ARPNet outperform the baseline models in predicting the degree of antidepressant response?

RQ2. Can ARPNet predict whether a patient reaches clinical remission, which is a task proposed in previous work.

RQ3. Is ARPNet useful for prescribing the most effective antidepressant combination to patients?

RQ4. How can we use the trained patient and antidepressant representation vectors in downstream applications?

We describe the experimental setting and report the evaluation results on the prediction of the degree of antidepressant response (Section 3.1) and the prediction of whether a patient reaches clinical remission with the given antidepressant (Section 3.2). We introduce the use case scenarios of our ARPNet model. (Sections 3.3 and 3.4).

3.1. Task 1: Prediction of the Degree of Antidepressant Response

3.1.1. Dataset Preparation

The prescription records of patients contain the HAM-D scores of the patients measured at the initial visit and 1, 4, and 8 weeks (or 6 months) after the initial visit. Some prescription records are missing the prescription record for the week in which patients did not visit. The data $d = (P, A, H, \Delta T)$ used in this experiment is generated from all pairs of two consecutive visit records (r_i and r_{i+1}) of a patient when the prescription records $R_P = \{r_1, r_2, ..., r_{|R_P|}\}$ of the patient are given. ΔT is the interval between the i-th and $i+1$-th visits. Our dataset $\mathcal{D}$ consists of 273 data generated from 395 records of 121 patients. The randomly sampled 10% of data is used as the test set, and the remaining data are used as the training set. The statistics of the constructed dataset D are summarized in Table 4.

Table 4. Statistics of datasets used in experimental evaluation.

Statistics	Dataset $\mathcal{D}$	Dataset $\mathcal{D}'$
# of data	273	121
# of training data	243	108
# of test data	30	13

3.1.2. Baselines

For the qualitative evaluation, we compare the performance of ARPNet with that of the following six machine learning baseline models: Support Vector Machine Regressor with a linear kernel (Linear SVR), Ridge Regressor, Gradient Boosting, Multi-Layer Perceptron (MLP) Regressor, K-Nearest Neighbors Regressor, and Random Forest Regressor. We concatenate the predictor features selected in Section 2.2, one-hot encoded prescribed antidepressants, current HAM-D score of a patient, and the interval between visits as an input of the machine learning baseline models.

To find the optimal hyper-parameters for our proposed ARPNet model and the baseline models, we perform 5-fold cross validation on the training set.

3.1.3. Metric

Antidepressant response prediction is a regression task. To assess the performance of the models, we employ Root Mean Square Error (**RMSE**) and **R-squared** which are the most common evaluation metrics used for regression models. RMSE is computed as follows.

$$RMSE = \sqrt{\frac{1}{|\mathcal{D}_{test}|} \sum_{i=1}^{|\mathcal{D}_{test}|} (y_i - \hat{y}_i)^2}. \tag{12}$$

R-squared is computed as follows.

$$R - squared = 1 - \frac{\sum_{i=1}^{|\mathcal{D}_{test}|} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{|\mathcal{D}_{test}|} (y_i - \bar{y})^2}, \tag{13}$$

where $\bar{y}$ denotes the average value of actual HAM-D scores in our dataset $\mathcal{D}_{test}$. A lower RMSE indicates that the model achieves better prediction performance. A higher R-squared also indicates better performance.

3.1.4. Results

To reduce the randomness of the machine learning-based models that randomly initialize their parameters, we perform the training and evaluation five times using the optimal hyper-parameters obtained from cross-validation. We then report the mean values of the performance scores.

The left side of Table 5 shows the evaluation results on the prediction of the degree of antidepressant response (Task 1). Our model ARPNet outperforms all the baseline models in terms of RMSE and R-squared. Specifically, ARPNet obtains 27.9%, 12.2%, 25.2%, 14.9%, and 12.8% performance improvements in terms of RMSE over LinearSVR, Ridge, Gradient Boosting, K-Nearst Neighbors, and Random Forest, respectively. ARPNet also obtains 372.9%, 31.5%, 174.0%, 44.4%, and 34.1% performance improvements in terms of R-squared over Support Vector Machine Regressor with a linear kernel (Linear SVR), Ridge Regressor, Gradient Boosting, Multi-Layer Perceptron (MLP) Regressor, K-Nearest Neighbors Regressor, and Random Forest Regressor, respectively.

Table 5. Experimental evaluation results. The performance improvement of ARPNet over the baseline with the best performance is reported in the *Improv.* row.

| Model | Task 1 | | Task 2 | | | | |
	RMSE	R-Squared	Sensitivity	Specificity	Precision	F1-Score	Accuracy
Linear SVR	4.5774	0.1168	0.3200	0.4250	0.1662	0.2167	0.3846
Ridge Regressor	3.7609	0.4199	0.7417	0.2833	0.4300	0.4509	0.4400
Gradient Boosting Regressor	4.4122	0.2016	0.4000	0.6250	0.4000	0.4000	0.5385
K-Nearst Neighbors Regressor	3.8805	0.3824	0.2000	0.5000	0.2000	0.2000	0.3846
Random Forest Regressor	3.7867	0.4118	0.2000	0.4750	0.1717	0.1835	0.3692
ARPNet	**3.3022**	**0.5523**	**0.8000**	**0.8750**	**0.8000**	**0.8000**	**0.8462**
Improv.	12.2%	31.5%	7.9%	40.0%	74.4%	86.0%	57.1%

These observations demonstrate that ARPNet is more accurate than the baseline models in predicting the degree of antidepressant response (Answer to RQ1).

3.2. Task 2: Prediction of the Clinical Remission of Patients

Prediction of the clinical remission of patients is a binary classification task. We assume that a patient responded to the prescribed antidepressant if the HAM-D score of the patient measured at

their last visit is less than half the HAM-D score measured at the initial visit. A patient's response to an antidepressant is calculated as follows.

$$Y'_P = \begin{cases} 1, & \text{if } \frac{H_{|R_P|}}{H_1} \leq 0.5 \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

where Y'_P is the label of the patient P and H_i is the HAM-D score of the patient at the i-th visit. Prediction models predict the clinical remission of patients using Algorithm 1.

Algorithm 1 Prediction of the clinical remissions of patients

---

Input: P: Patient, R_P: Prescription records $\{r_1, r_2, \cdots, r_{|R_P|}\}$ of P

Output: $\hat{Y}'_P$: Predicted clinical remission of patient P.

1: **procedure**
2:      Initialize current HAM-D score $H \leftarrow H_1 \in r_1$
3:      **for** i **in** $[1, |R_P|)$ **do**
4:          i-th record's HAM-D score $H_i \leftarrow H$
5:          Predict $(i+1)$-th record's HAM-D score $\hat{H}_{i+1}$ from r_i
6:          Current HAM-D score $H \leftarrow \hat{H}_{i+1}$
7:      **end for**
8:      $|R_P|$-th record's HAM-D score $H_{|R_P|} \leftarrow H$
9:      **if** $\frac{H_{|R_P|}}{H_1} \leq 0.5$ **then**
10:          **Return** $\hat{Y}'_P = 1$
11:      **else**
12:          **Return** $\hat{Y}'_P = 0$
13:      **end if**
14: **end procedure**

3.2.1. Dataset Preparation

Using Equation (14), we constructed the dataset $\mathcal{D}'$ which contains 121 data from the prescription records of 121 patients. Like Task 1, 10% of the data is randomly sampled and used as the test set, and the remaining data are used as the training set. To find the optimal hyper-parameters for the prediction models, we perform 5-fold cross validation on the training set. The statistics of the constructed dataset $\mathcal{D}'$ are summarized in Table 4.

3.2.2. Metric

To evaluate performance on the prediction of clinical remission, we employ the following five evaluation metrics that are widely used in binary classification tasks: **Sensitivity**, **Specificity**, **Precision**, **F1-score**, and **Accuracy**. Sensitivity and Specificity are computed as follows,

$$Sensitivity = \frac{TP}{TP + FN}, \; Specificity = \frac{TN}{TN + FP}, \tag{15}$$

where TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative respectively. Precision and F1-score are computed as follows,

$$Precision = \frac{TP}{TP + FP}, \; F1 - score = \frac{2TP}{2TP + FN + FP}. \tag{16}$$

Accuracy is computed as follows,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{17}$$

For all the evaluation metrics, the higher the score, the better the performance of a model on prediction.

3.2.3. Results

In Section 3.1, we perform the training and evaluation five times using the optimal hyper-parameters to reduce the randomness of the machine learning-based prediction models. We report the average of the five evaluation results for each model. The right side of Table 5 shows the evaluation results on the prediction of the clinical remission of patients (Task 2). Our ARPNet model outperforms all the baseline models on all the evaluation metrics (Answer to RQ2). ARPNet achieved performance improvements of at least 40.0%, 74.4%, 86.0%, and 57.1% in Specificity, Precision, F1-score, and Accuracy, respectively, over the best baseline model. In terms of Sensitivity, ARPNet achieves performance improvements of 100% over all the baselines except for K-Nearst Neighbors Regressor. Like in the evaluation result of Ridge Regressor, sensitivity generally shows a negative correlation between specificity and precision. Nevertheless, ARPNet obtains performance improvements in all the evaluation metrics.

We observe that ARPNet achieves a greater performance improvement in Task 2 than in Task 1. Figure 2 shows two representative cases: (1) the patient improves over time; (2) the condition of the patient worsens after four weeks of treatment. In both cases, ARPNet accurately predicts the outcome of the patient as the predicted HAM-D score of the patient is almost the same as their actual HAM-D score. On the other hand, K-Nearest Neighbors and Gradient Boosting were unable to accurately predict the HAM-D scores of patients. Their prediction performance decreased as the interval between visits increased. As it takes a long time (more than 8 weeks) to treat MDD, inaccurate antidepressant response predictions can result in ineffective prescriptions and increased risks.
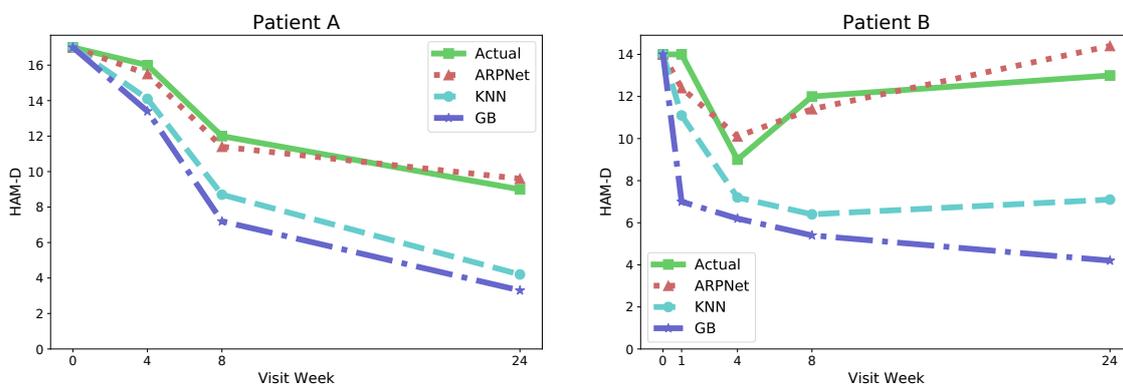

Figure 2. Examples of predicting the clinical remission of patients. KNN and GB denote the K-Nearst Neighbors and Gradient Boosting Regressors, respectively.

3.3. Use Case Scenario: Prescribing the Most Effective Antidepressant

As ARPNet can predict the degree of the antidepressant response, ARPNet can be used to prescribe the most effective antidepressants. We conduct an additional qualitative evaluation to show the applicability of ARPNet. We compute the predicted HAM-D scores of all the possible antidepressant candidates. The possible antidepressant candidates include 14 single antidepressants and 91 combinations of any two antidepressants. Figure 3 shows the result of a sample patient as an example. In the figure, the tables list the predicted HAM-D score of each antidepressant candidate for Patient A at their visit. The green line indicates the change in the actual HAM-D score of Patient A. The red line indicates the change in the HAM-D score predicted by ARPNet for the prescribed antidepressant MIRTAZAPINE. The blue line denotes the change in the HAM-D score when Patient A at the initial visit is prescribed the antidepressant combination of MIRTAZAPINE and MILNACIPRAN which is predicted to be the most effective. In this phase, it is predicted that MIRTAZAPINE with MILNACIPRAN (Rank 1) is more effective than MIRTAZAPINE alone (Rank 3). At the visit in Week 8,

the antidepressant combination of ESCTALOPRAM and AMITRIPTYLINE which is predicted to be the most effective antidepressant is prescribed. The cyan line denotes the change in the HAM-D score when the psychiatrist changes the prescription to the most effective antidepressants predicted at the visit in Week 8. As shown in this example, the HAM-D score of the patient decreases faster than the actual HAM-D score by prescribing the most effective antidepressants at every visit.
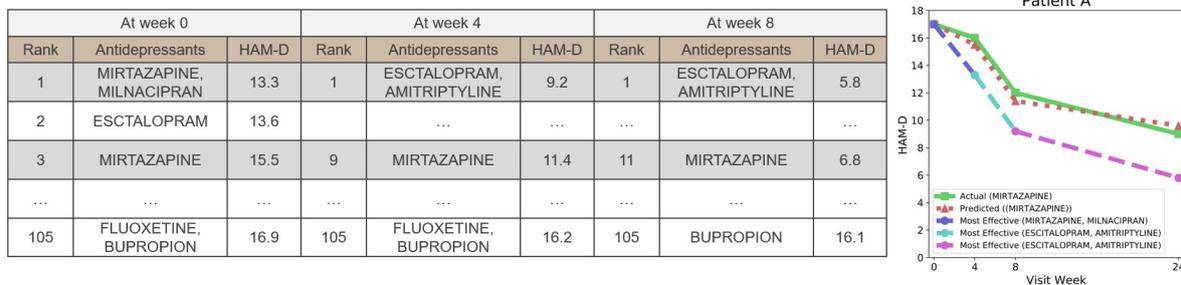
	At week 0			At week 4			At week 8	
Rank	Antidepressants	HAM-D	Rank	Antidepressants	HAM-D	Rank	Antidepressants	HAM-D
1	MIRTAZAPINE, MILNACIPRAN	13.3	1	ESCTALOPRAM, AMITRIPTYLINE	9.2	1	ESCTALOPRAM, AMITRIPTYLINE	5.8
2	ESCTALOPRAM	13.6		…	…	…		…
3	MIRTAZAPINE	15.5	9	MIRTAZAPINE	11.4	11	MIRTAZAPINE	6.8
…	…	…	…	…	…	…	…	…
105	FLUOXETINE, BUPROPION	16.9	105	FLUOXETINE, BUPROPION	16.2	105	BUPROPION	16.1

Figure 3. The use case where the most effective antidepressant is prescribed. The first tables show the ranked predicted HAM-D scores of Patient A at their first visit. A green line indicates an actual change in the HAM-D score of Patient A. A red line indicates a change in the HAM-D score predicted by ARPNet for the prescribed antidepressant MIRTAZAPINE. The blue line indicates the changes in the predicted HAM-D score when Patient A is prescribed MIRTAZAPINE and MILNACIPRAN which are predicted to be the most effective at first visit. The cyan and pink lines indicate the predicted HAM-D score changes of Patient A when the antidepressants are changed to ESCITALOPRAM and AMITRIPTYLINE. ESCITALOPRAM and AMITRIPTYLINE are the most effective antidepressants predicted by ARPNet in Weeks 4 and 8.

These observations demonstrate that ARPNet can compute the predicted HAM-D scores of all the possible antidepressant candidates to prescribe the most effective antidepressant, which can lead to more effective treatment and reduced costs and risks (Answer to RQ3).

3.4. Use Case Scenario: Using Pre-Trained Representation Vectors

Through the patient representation layer, ARPNet trains the patient representation vectors to capture high-dimensional patterns for improving performance in antidepressant response prediction. By representing the patients as real-valued vectors, we can compute the antidepressant response-aware similarity between patients. This antidepressant response-aware similarity can be used for precision prescription in evidence-based medicine [24,25]. Figure 4 shows an example of using patient representation vectors in clinical decision support systems. Patients in the prescription record database are represented as patient representation vectors, using patient data and prescription records. Based on their representation vectors, the patients are clustered using a clustering method that computes the antidepressant response-aware similarity between the patients. When a new patient's data is given, the patient's representation vector is computed using the patient representation layer of ARPNet. Clinical decision support systems compute the similarities between the patient and patients in the cluster in which the patient is likely to be included, and provide psychiatrists with the prescription records of patients very similar to those of the patient for evidence-based medicine. Psychiatrists prescribe the most effective antidepressant to their patient by referring to the antidepressant response records of similar patients.

This use case demonstrates that pretrained patient representation vectors can be utilized as clinical evidence in downstream applications for precision medicine (Answer to RQ4).
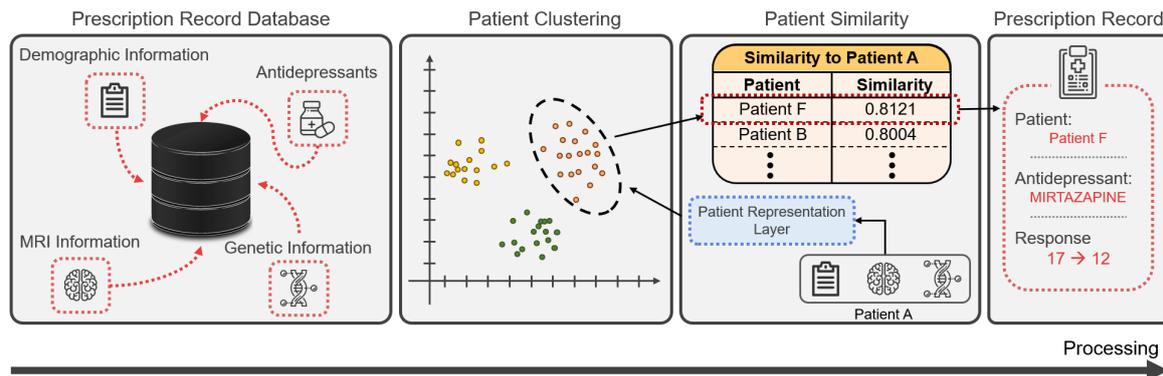
Figure 4. A use case of clinical decision support systems using the similarity between patients.

4. Conclusions

In this paper, we proposed ARPNet which is a new antidepressant response prediction network model for major depressive disorder. ARPNet utilizes several useful features from patient data including the demographic, brain MRI , and genetic information of patients. Based on a literature survey, features such as neuroimaging biomarkers were extracted from brain MRI information. Also, genetic variants and DNA methylation were extracted from genetic information using biomedical literature-based tools. To make the extracted features more predictive, we filtered them by a data-driven feature selection process. Based on a neural network architecture, ARPNet predicts not only whether patients will respond to an antidepressant, but also the degree of the antidepressant response. ARPNet considers a real-world antidepressant prescription scenario where psychiatrists prescribe one or more antidepressants to patients with MDD. In the experiments, ARPNet outperformed the machine learning-based baseline models in both the prediction of the degree of antidepressant response and the prediction of clinical remission of patients. Finally, we used real-world use cases to show the applicability of ARPNet in downstream applications. In future work, we plan to further evaluate ARPNet on real patients.

Author Contributions:

Conflicts of Interest:

References

1. Rush, A.J.; Wisniewski, S.R.; Warden, D.; Luther, J.F.; Davis, L.L.; Fava, M.; Nierenberg, A.A.; Trivedi, M.H. Selecting among second-step antidepressant medication monotherapies: predictive value of clinical, demographic, or first-step treatment features. *Arch. Gen. Psychiatry* **2008**, *65*, 870–880.
2. Rost, K.; Nutting, P.; Smith, J.L.; Elliott, C.E.; Dickinson, M. Managing depression as a chronic disease: A randomised trial of ongoing treatment in primary care. *Bmj* **2002**, *325*, 934.
3. Cipriani, A.; Brambilla, P.; Furukawa, T.A.; Geddes, J.; Gregis, M.; Hotopf, M.; Malvini, L.; Barbui C. Fluoxetine versus other types of pharmacotherapy for depression. *Cochrane Database Syst. Rev.* **2013**, CD004185.
4. Cipriani, A.; Furukawa, T.A.; Salanti, G.; Geddes, J.R.; Higgins, J.P.; Churchill, R.; Watanabe, N.; Nakagawa, A.; Omori, I.M.; McGuire, H.; et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: A multiple-treatments meta-analysis. *Lancet* **2009**, *373*, 746–758.
5. Phillips, M.L.; Chase, H.W.; Sheline, Y.I.; Etkin, A.; Almeida, J.R.; Deckersbach, T.; Trivedi, M.H. Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: Neuroimaging approaches. *Am. J. Psychiatry* **2015**, *172*, 124–138.
6. Investigators, G.; Investigators, M.; Investigators, S.D. Common genetic variation and antidepressant efficacy in major depressive disorder: A meta-analysis of three genome-wide pharmacogenetic studies. *Am. J. Psychiatry* **2013**, *170*, 207–217.

7. Fonseka, T.M.; MacQueen, G.M.; Kennedy, S.H. Neuroimaging biomarkers as predictors of treatment outcome in Major Depressive Disorder. *J. Affect. Disord.* **2018**, *233*, 21–35. Are there Biomarkers for Mood Disorders? doi:10.1016/j.jad.2017.10.049.

8. Breitenstein, B.; Scheuer, S.; Holsboer, F. Are there meaningful biomarkers of treatment response for depression? *Drug Discov. Today* **2014**, *19*, 539–561. doi:10.1016/j.drudis.2014.02.002.

9. Lin, E.; Lane, H.Y. Genome-wide association studies in pharmacogenomics of antidepressants. *Pharmacogenomics* **2015**, *16*, 555–566, PMID: 25916525, doi:10.2217/pgs.15.5.

10. Opmeer, E.M.; Kortekaas, R.; Aleman, A. Depression and the role of genes involved in dopamine metabolism and signalling. *Prog. Neurobiol.* **2010**, *92*, 112–133. doi:10.1016/j.pneurobio.2010.06.003.

11. Lin, E.; Chen, P.S.; Chang, H.H.; Gean, P.W.; Tsai, H.C.; Yang, Y.K.; Lu, R.B. Interaction of serotonin-related genes affects short-term antidepressant response in major depressive disorder. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2009**, *33*, 1167–1172. doi:10.1016/j.pnpbp.2009.06.015.

12. Fabbri, C.; Di Girolamo, G.; Serretti, A. Pharmacogenetics of antidepressant drugs: An update after almost 20 years of research. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2013**, *162*, 487–520, doi:10.1002/ajmg.b.32184.

13. Lisoway, A.J.; Zai, C.C.; Tiwari, A.K.; Kennedy, J.L. DNA methylation and clinical response to antidepressant medication in major depressive disorder: A review and recommendations. *Neurosci. Lett.* **2018**, *669*, 14–23. The path toward personalized medicine in psychiatry: Promises and pitfalls, doi:10.1016/j.neulet.2016.12.071.

14. Chekroud, A.M.; Zotti, R.J.; Shehzad, Z.; Gueorguieva, R.; Johnson, M.K.; Trivedi, M.H.; Cannon, T.D.; Krystal, J.H.; Corlett, P.R. Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry* **2016**, *3*, 243–250.

15. Rush, A.J.; Fava, M.; Wisniewski, S.R.; Lavori, P.W.; Trivedi, M.H.; Sackeim, H.A.; Thase, M.E.; Nierenberg, A.A.; Quitkin, F.M.; Kashner, T.M.; et al. Sequenced treatment alternatives to relieve depression (STAR* D): rationale and design. *Control. Clin. Trials* **2004**, *25*, 119–142.

16. Trivedi, M.H.; Rush, A.J.; Wisniewski, S.R.; Nierenberg, A.A.; Warden, D.; Ritz, L.; Norquist, G.; Howland, R.H.; Lebowitz, B.; McGrath, P.J.; et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR* D: Implications for clinical practice. *Am. J. Psychiatry* **2006**, *163*, 28–40.

17. Warden, D.; Rush, A.J.; Trivedi, M.H.; Fava, M.; Wisniewski, S.R. The STAR* D Project results: A comprehensive review of findings. *Curr. Psychiatry Rep.* **2007**, *9*, 449–459.

18. Sinyor, M.; Schaffer, A.; Levitt, A. The sequenced treatment alternatives to relieve depression (STAR* D) trial: A review. *Can. J. Psychiatry* **2010**, *55*, 126–135.

19. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320.

20. Lee, S.; Kim, D.; Lee, K.; Choi, J.; Kim, S.; Jeon, M.; Lim, S.; Choi, D.; Kim, S.; Tan, A.C.; et al. BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. *PLoS ONE* **2016**, *11*, e0164680, doi: 10.1371/journal.pone.0164680.

21. Lee, K.; Kim, B.; Choi, Y.; Kim, S.; Shin, W.; Lee, S.; Park, S.; Kim, S.; Tan, A.C.; Kang, J. Deep learning of mutation-gene-drug relations from the literature. *BMC Bioinform.* **2018**, *19*, 21. doi:10.1186/s12859-018-2029-1.

22. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

23. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.

24. Sackett, D.L. *Evidence-Based Medicine How to Practice and Teach EBM*; WB Saunders Company: Philadelphia, PA, USA, 1997.

25. Parimbelli, E.; Marini, S.; Sacchi, L.; Bellazzi, R. Patient similarity for precision medicine: A systematic review. *J. Biomed. Inform.* **2018**, *83*, 87–96.