

# **Bioinformatics and machine learning-driven key genes screening for vortioxetine**

Sabire Kılıçarslan<sup>1</sup> . Meliha Merve Hız-Çicekliyurt<sup>2,\*</sup>

<sup>1</sup>Department of Medical System Biology, School of Graduate Studies, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye <sup>2</sup>Department of Medical Biology, Faculty of Medicine, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye

**Abstract** − Vortioxetine is a pharmacological agent that acts as a serotonin modulator and stimulant, with safety and tolerability being important health issues. This study aimed to use bioinformatic and machine learning methods to find differentially expressed genes (DEG) between rats exposed to vortioxetine and matched controls. The GSE236207 dataset (Rattus norvegicus) was obtained from the National Center for Biotechnology Information (NCBI) and analyzed with R, followed by genetic ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) enrichment analyses, and String's protein-protein interaction network was established to identify important genes. The original datasets were preprocessed in the second step by detecting and correcting missing and noisy data and then merged. After feature selection for the cleaned dataset, machine learning algorithms such as the Knearest neighbors' algorithm, Naive Bayes, and Support Vector Machine (SVM) were used. In addition, an accuracy of 0.90 was observed with SVM. Leveraging these techniques, the study linked IGFBP7, KLRA22, PROB1, SHQ1, NTNG1, and LOC102546359 to vortioxetine exposure. The bioinformatic analysis revealed 18 upregulated genes and 27 downregulated genes, with all approaches identifying only one common locus, LOC102546359, responsible for noncoding ribonucleic acid (ncRNA) synthesis. The crucial point is that this locus bears no connection to any disease or trigger mechanism, thereby bolstering the safety of vortioxetine.

**Keywords:** *DEG, depression, LOC102546359, machine learning, systems biology*

# **1. Introduction**

Vortioxetine has brought a new approach in terms of both treatment effectiveness and side effect profile. Vortioxetine, which has a unique mechanism of action, is a new-generation antidepressant approved for the treatment of major depressive disorder (MDD) in different countries. Unlike other therapeutic agents used in the treatment of MDD, vortioxetine works as a 5-HT1A receptor agonist, 5-HT3, 5-HT7, and 5-HT1D receptor antagonist, 5-HT1B receptor partial agonist, and inhibitor of the 5-HT transporter (SERT). With this feature, it is described in the literature as an antidepressant with multimodal activity. This multimodal activity is used in the first-line treatment of the disease or subsequent therapy of the failure of other antidepressants [1].

Vortioxetine's unique pharmacodynamic profile makes the molecule safer and better to tolerate. Studies have shown that the risk of treatment discontinuation due to side effects is lower and that it is one of the most tolerable drugs in the antidepressant group. While nausea and vomiting are the most common side effects associated with vortioxetine, studies have shown that these effects are mild or moderate and typically temporary within 2 weeks. Reports indicate that the effects are permanent at a rate of only 2% and increase dose-dependently [2].



Artificial intelligence (AI) methods in healthcare offer revolutionary advancements in areas such as medical diagnosis, treatment planning, disease prediction, and prevention. Deep learning algorithms enable rapid and accurate detection of abnormalities in medical imaging, facilitating early diagnosis. Machine learning analyzes patient data to propose personalized treatment strategies, while big data analytics predict disease risks, allowing for preventive measures. Additionally, AI-supported systems enhance hospital management and operational efficiency, making healthcare services more effective and accessible. Consequently, AI technologies significantly benefit healthcare professionals and patients, advancing healthcare services to new heights [3-6].

Applying data mining methods to analyze microarray gene data poses substantial difficulties. The considerable number of features in the data makes the categorization process more complex. Therefore, to address these problems successfully, performing the necessary feature selection in the microarray dataset for gene analysis is crucial. Artificial intelligence methods have proven useful in streamlining feature selection processes [6].

The study employed a bioinformatics and artificial intelligence queue, effectively identifying significant differentially expressed genes using vortioxetine. Detecting complex patterns in large datasets and identifying key genetic markers are both possible with the help of machine learning algorithms. Bioinformatics tools, on the other hand, make it easier to analyze and comprehend significant amounts of genomic data. These methods enhance comprehension of the molecular mechanism of vortioxetine, thereby promoting the progress of our knowledge regarding the safety and tolerability of the drug. The primary aim of this research is to assess the impact of vortioxetine on gene expression in the absence of depressive symptoms or cognitive impairment.

This paper primarily contributes to the following areas:

Bioinformatic and machine learning-based screening of vortioxetine response genes. Understand the mechanism of action of vortioxetine without a background in depression. Evaluate the safety and tolerability of drugs in non-depressive conditions under machine learning models – a novel idea based on machine learning to predict the vortioxetine response. A machine learning-based comparative study on how feature selection with AI and GEO2R can improve drug response from sequence data.

## **2. Material and Methods**

## **2.1. Dataset**

The GSE236207 database was employed for conducting the studies, and the dataset was carefully curated by removing irrelevant and unrelated data. For this purpose, we generated a subset of data that consisted of six rats exposed to vortioxetine, and the six rats were fed a standard diet without any manipulation. A total of 4570 features were extracted from a dataset containing 45738 features. The study used comparative bioinformatics and machine learning methodologies, including Naive Bayes (NB), k-nearest neighbor (kNN), and support vector machines (SVM) algorithms, to identify the genes associated with using vortioxetine. The study did not use data from humans or animals; all the data utilized can be freely accessed from the Gene Expression Omnibus database. The datasets used in this study are open files in the GEO databases and do not contain any material pertaining to humans or animals; thus, no ethics board approval is necessary.

## **2.2. Determination of the Differently Expressed Genes (DEG)**

The R software-limma package [7] analyzed GSE236207 transcript counts to identify differentially expressed genes. To compare gene expression in vortioxetine-exposed and non-exposed samples, we used a moderated t-test with the eBayes method in the lymphatic package in R [8]. This step improves analysis reliability by using empirical Bayes moderation. The error detection rate (FDR) approach [9] was used to adjust the computed p values, with the Benjamini-Hochberg (BH) procedure [9] used to control errors. The selection

thresholds for DEGs were p-values  $\leq 0.05$ , Log2FC  $\leq$  -1, and Log2FC  $\geq$  1. The visual representation of the differential expression analysis's outcomes was provided by the R package's "umap" function.

# **2.3. Functional Analysis of DEGs**

Webgestalt [10] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [11] are web-based platforms that have been used to conduct gene enrichment analysis. These platforms use the Genetic Ontology (GO) and KEGG databases. Genes were visually represented using the Venn diagram available at [12]. The Venn diagram depicted differentially expressed genes' up and down-regulation outcomes (DEGs).

# **2.4. Protein-Protein Interaction (PPI)**

A network diagram representing differentially expressed genes was constructed to examine the interactions between genes via protein-protein interactions via the STRING website [13] with a confidence threshold exceeding 0.4.

# **3. Artificial Intelligence and Predictive Model of Vortioxetine Action**

# **3.1. Artificial Intelligence Experimental Design**

Firstly, it is crucial to thoroughly analyze the dataset to identify and remove any missing or conflicting data. As a consequence of this analysis, it was determined that there were no instances of missing or incompatible data in the dataset. After this step, we selected the properties using the nsFilter function from the gene filter library, setting a threshold of 0.90. Three different machine learning algorithms (kNN, NB, and SVM) evaluated the collected dataset.

kNN is an algorithm used for regression and classification problems in controlled machine learning models where predictions based on the similarity of observations are made.

A statistical learning system that applies structural risk minimization concepts, Vapnik's SVM is utilized for classification and curve modification tasks [14]. SVM utilizes structural risk minimization, a principle in statistical learning theory, as opposed to the experimental risk minimization approach, where the classification function is created by reducing the mean square error in the dataset [15].

Based on Bayes' theory, the NB classifier is a simple but powerful algorithm for categorization and prediction, especially in microarray data and image processing. This algorithm assumes that certain properties are independent of other properties. Vapnik's SVM is a statistical learning method used for classification and curve adjustment problems based on the principles of structural risk minimization [16].

We utilized the varImp feature of the Caret software to determine the top 20 genes most strongly associated with the consumption of vortioxetine. This work used the svmRadial approach to train the SVM model, while the caret R package was employed to train the kNN model and NB model using their default values. The varImp function in the R package supported all models. The flow chart showing the operation of the study is presented in Figure 1.



**Figure 1**. Flowchart of the proposed method

## **4. Results**

#### **4.1. Bioinformatics-Based Evaluation**

The GSE236207 datasets were partitioned, and any unrelated data was eliminated from the dataset. We exclusively analyzed the data from the dorsal hippocampus (dHipp) tissue of sham controls and control-dietfed animals exposed to vortioxetine to assess the entire genome microarray. Our objective was to investigate the potential overlap and shared regulation of biological networks in response to vortioxetine. We analyzed the data using a bioinformatic approach and three distinct artificial intelligence algorithms to achieve this.

Table 1 presents a list of genes resulting from the bioinformatic-based evaluations. The results showed an upregulation of 18 genes and a downregulation of 26.

Our study identified crucial pathways enriched in vortioxetine treatment, including cellular responses to endogenous stimuli such as hormones or hypoxia in biological processes. The statistical significance of the results is indicated by the GO enrichment analysis (false discovery rate (FDR) equals 1) for the DEGs in biological processes and cellular components, but they are susceptible to type 1 error.

SLC22A6 FMOD LGFBP2 CPEB1 STEAP1 KCNJ13 AQP1 KL





**Table 1**. Representation of up and down-regulated genes according to DEGs results

**Gene Symbol**<br>GNMT

**Gene Regulation- Direction DOWN**



RN45S GNMT RN5-8S KRT24 RN45S CDKN1C RN18S MDFIC LOC103692073 OLR1482 LOC100361143 SPSB2 LIAS SFRP1 KIF12 SLFN14 LFI47 TAL2 TPH1 MX1 GNGT1 OGN CCL22 GLYCAM1 SNX22 VGLL3 OLR852 CBLL1 METTL21C LGF2 OLR1407 PRLR TTYH3 C1QTNF3 CSAP TNNT2

**Gene Regulation-Direction UP**

**Gene Symbol**<br>RN45S



**Top 10 Biological Process Enriched** 

**Figure 2**. Bubble chart of commonly upregulated genes enriched biological processes

Figure 2 shows the significantly enriched GO:BP terms shared by all the upregulated genes for the vortioxetine response. The size of the dots is the normalized enrichment score (NES) and the color is the statistical significance as -log10 (p-adjust).





Figure 3 shows the significantly enriched GO:CC terms shared by all the upregulated genes for the vortioxetine response. The size of the dots is the NES and the color is the statistical significance as -log10 (p-adjust).



**Top 10 Biological Process Enriched** 

**Figure 4.** Bubble chart of commonly down-regulated genes enriched biological processes

Figure 4 shows the significantly enriched GO:BP terms shared by all the down-regulated genes for the vortioxetine response. The size of the dots is the NES and the color is the statistical significance as -log10 (padjust).



**Figure 5.** Bubble chart of commonly down-regulated genes enriched cellular component

Figure 5 shows the significantly enriched GO:CC terms shared by all the down-regulated genes for the vortioxetine response. The size of the dots is the NES, and the color is the statistical significance as -log10(padjust)

We used the STRING database to look at the PPI network. We selected the organism Rattus norvegicus by navigating to the multiple protein tab and inputting the names of all associated genes. After that, we analyzed the gene-gene interaction network. The connections' shape and color indicate the relationship's type and intensity. The STRING results have 43 nodes and five edges, with an average grade of 0.233 and a UFE enrichment p-value of 0.271. Figure 6 shows that CDKN1C, LGF2, PRLR, and LGFBP2 are key factors related to vortioxetine treatment.



**Figure 6***.* KEGG pathway related to DEG

The KEGG pathway analysis showed that the DEGs were significantly linked to pathways involved in the serotonergic synapse with key components of upregulated tryptophan hydroxylase 1 (TPH1) gene (Figure 6).

## **4.2. Machine Learning-Based Evaluation**

Experimental evaluations have been conducted to determine the most influential genes that react to vortioxetine treatment utilizing machine learning techniques. SVM with Radial Basis (svmRadial), NB, and kNN have been used as machine learning algorithms in experimental assessments. The findings from the experiments have allowed us to identify the top 20 genes for each model. Table 3 displays the most significant genes associated with vortioxetine response using machine learning approaches.

**Table 3.** Five common genes and one chromosomal location that identified in 3 different machine-learning approaches

<b>Gene Symbol</b>	<b>Gene Name</b>
IGFBP7	insulin-like growth factor binding protein 7
KLRA22	killer cell lectin-like receptor subfamily A, member 22
PROB <sub>1</sub>	proline-rich basic protein 1
SHO <sub>1</sub>	SHQ1, H/ACA ribonucleoprotein assembly factor
NTNG1	netrin G1
LOC103691895	protamine-like
LOC103691173	60S ribosomal protein L39-like
LOC102546359	-

By employing three machine learning algorithms (SVMRadial, NB, and kNN), we successfully identified five genes, and one chromosomal location consistently identified by all three methods. The genes IGFBP7, KLRA22, PROB1, SHQ1, and NTNG1 exhibit various biological functions.

IGFBP7 modulates growth factors response to retinoic acid and cortisol [17], while KLRA22 appears to be a major component affecting immune responses [18-19]. PROB1 encodes proline-rich basic protein 1, although its function is still unidentified [20]. NTNG1 is associated with nervous system development and synaptic plasticity and is active in glutamatergic synapses [21]. Additional notable chromosomal sites include LOC103691895, which encodes a protamine-like protein, and LOC103691173, which encodes a ribosomal protein. The gene LOC102546359 is currently uncharacterized. The genes and their corresponding chromosomal positions are frequently linked to significant biological processes within the specific environment under investigation. This association may offer novel opportunities for comprehending disease mechanisms or developmental biology.

Furthermore, IGFBP7, KLRA22, PROB1, SHQ1, and NTNG1 are the most noteworthy genes consistently identified by three distinct artificial intelligence techniques. These genes have the potential to respond to vortioxetine.

## **5. Discussion**

Upon bioinformatic analysis of the upregulated genes, we observed that the process of ribosome biogenesis (Rn45s, Rn18s, and Rn5-8s) was activated by a given dose of vortioxetine. Liu et al. [22] have shown that aripiprazole, clozapine, and lithium down-regulate ribosomal biogenesis, decreasing protein synthesis. Fusco et al. demonstrated the significance of ribosome localization in neurons and the specialized translation machinery [23]. In that scheme, regulating neuronal ribosomal biogenesis and its subsequent location is a crucial concern for health. Powell et al. provided evidence of escitalopram's ability to stimulate the growth of new neurons in the hippocampus, a brain region. This was accompanied by changes in gene expression involved in the formation of axons and microtubules and the production of ribosomes [24]. Our informatics analysis has demonstrated that vortioxetine promotes the generation of new ribosomes and improves protein synthesis, which may result in the restructuring of neurons, which is crucial for neuropsychiatric disorders.

Another notable discovery is that vortioxetine enhances the activity of the mitochondrial lipoic acid synthetase enzyme, which has been extensively researched in the literature due to its connection with neurodegeneration and its potential antidepressant effects [25-29]. The literature shows that the number and turnover rate of olfactory receptors are reduced in depression [30,31]. Our analysis also indicates that vortioxetine increases olfactory receptor (Olr1407, Olr852) expression during treatment. This aligns with the existing literature in a similar manner. Vortioxetine enhances the expression of tryptophan hydroxylase, a crucial enzyme involved in the production of serotonin, a monoamine neurotransmitter.

Upon analysis of the DEG-based enriched biological processes, it was shown that the synthesis of aromatic amino acids and pathways associated with serotonin and tryptophan metabolic processes were upregulated. Conversely, the synthesis of aspartate and serine amino acids and purine nucleotide synthesis pathways were downregulated. In line with previous research, our data demonstrate that vortioxetine stimulates tryptophan synthesis and its conversion to serotonin [32].

The artificial intelligence methods were able to detect 18 common elements. However, only five genes and one chromosomal site were ultimately identified, with four undefined and seven uncharacterized. In artificial intelligence analyses, three algorithms demonstrated common DEG: IGFBP7, KLRA22, PROB1, SHQ1, and NTNG1. These genes were uncommon with bioinformatic analyses; only the one chromosomal location responsible for ncRNA synthesis, LOC102546359 was commonly found in all approaches.

#### **6. Conclusion**

The main goal has been to determine gene expression variations in non-depressed rats exposed to vortioxetine. We conducted differential gene expression analysis focusing on gene sets and pathway analysis to investigate genes and molecular-signature pathways that may be specific to vortioxetine response. IGFBP7, KLRA22, PROB1, SHQ1, and NTNG1 are the top upregulated genes upon machine learning analysis. These genes were uncommon with bioinformatic analyses; only the one chromosomal location responsible for ncRNA synthesis, LOC102546359 was commonly found in all approaches. None of these investigations identified any pathways that are unrelated to the use of vortioxetine and could potentially induce other diseases, particularly cancer. This is done to assess safety and gain insight into the response mechanism while avoiding the influence of disease heterogeneity associated with vortioxetine usage. Future research endeavors should focus on developing precise and reliable clinical assays that distinguish between individuals who exhibit a positive therapeutic response to vortioxetine and those who may have an adverse or sensitized reaction to the drug. This differentiation is critical to minimizing the potential risks of misdiagnosis and inappropriate treatment.

#### **Author Contributions**

All the authors equally contributed to this work. They all read and approved the final version of the paper.

#### **Conflicts of Interest**

All the authors declare no conflict of interest.

#### **Ethical Review and Approval**

No approval from the Board of Ethics is required.

#### **Acknowledgments**

This work was supported by the Office of Scientific Research Projects Coordination at Çanakkale Onsekiz Mart University, Grant number: TYL-2018-2584.

#### **References**

- [1] G. Chen, A.-M. Højer, J. Areberg, G. Nomikos, *Vortioxetine: Clinical pharmacokinetics and drug interactions*, Clinical Pharmacokinetics 57 (6) (2018) 673–686.
- [2] A. J. Krupa, K. Wojtasik-Bakalarz, M. Siwek, *Vortioxetine - pharmacological properties and use in mood disorders. The current state of knowledge*, Psychiatria Polska 57 (6) (2023) 1109–1126.
- [3] S. Kiliçarslan, E. Dönmez, *Improved multi-layer hybrid adaptive particle swarm optimization based artificial bee colony for optimizing feature selection and classification of microarray data,* Multimedia Tools and Applications 83 (26) (2024) 67259-67281.
- [4] S. Kiliçarslan, *A novel nonlinear hybrid HardSReLUE activation function in transfer learning architectures for hemorrhage classification*, Multimedia Tools and Applications 82 (4) (2023) 6345-6365.
- [5] I. Pacal, *Deep learning approaches for classification of breast cancer in ultrasound (US) images*, Journal of the Institute of Science and Technology 12 (4) (2023) 1917–1927.
- [6] I. Pacal, *MaxCerVixT: A novel lightweight vision transformer-based approach for precise cervical cancer detection*, Knowledge-Based Systems 289 (2024) 111482.
- [7] Version 4.2.2,<https://www.r-project.org/> (Accessed 30 May 2024).
- [8] G. K. Smyth, *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments*, Statistical Applications in Genetics and Molecular Biology 3 (3) (2004).
- [9] G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, *ClusterProfiler: An R package for comparing biological themes among gene clusters*, Omics: A Journal of Integrative Biology 16 (5) (2012) 284–287.
- [10] Web-based gene set analysis toolkit, <https://www.webgestalt.org/option.php> (Accessed 30 May 2024).
- [11]Kyoto encyclopedia of genes and genomes, <https://www.genome.jp/kegg/> (Accessed 30 May 2024).

[12] J. C. Oliveros, (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. <https://bioinfogp.cnb.csic.es/tools/venny/> (Accessed 30 May 2024).

[13] Protein-protein interaction networks functional enrichment analysis, https://string-db.org/ (Accessed 30 May 2024).

- [14]V. N. Vapnik, *The vicinal risk minimization principle and the SVMs*, in The Nature of Statistical Learning Theory, V. N. Vapnik (Ed.), New York, NY: Springer, 2000, pp. 267–290.
- [15]Ö. Akay, M. Tunçeli, *Use of the support vector regression in medical data analysis,* Experimental and Applied Medical Science 2 (4) (2021) 242–256.
- [16]N. Cristianini, E. Ricci, *Support vector machines*, in Encyclopedia of Algorithms, M.-Y. Kao (Ed.), Boston, MA: Springer, 2008, pp. 928–932.
- [17]J. L. Januzzi et al*.*, *IGFBP7 (Insulin-like growth factor–binding protein-7) and neprilysin inhibition in patients with heart failure*, Circulation: Heart Failure 11 (10) (2018) e005133.
- [18]F. Gays, S. Taha, C. G. Brooks, *The distal upstream promoter in Ly49 genes, Pro1, is active in mature NK cells and T cells, does not require TATA boxes, and displays enhancer activity*, The Journal of Immunology 194 (12) (2015) 6068–6081.
- [19]Ø. Nylenna *et al.*, *The genes and gene organization of the Ly49 region of the rat natural killer cell gene complex*, European Journal of Immunology 35 (1) (2005) 261–272.
- [20]J. A. Karolak *et al.*, *Variants in SKP1, PROB1, and IL17B genes at keratoconus 5q31.1-q35.3 susceptibility locus identified by whole-exome sequencing*, European Journal of Human Genetics 25 (1) (2027) 73–78.
- [21]Archer, Hayley L., et al. *NTNG1 mutations are a rare cause of Rett syndrome*, American Journal of Medical Genetics Part A 140 (7) (2006) 691-694.
- [22]Z. S. J. Liu *et al.*, *Effects of psychotropic drugs on ribosomal genes and protein synthesis*, International Journal of Molecular Sciences 23 (13) (2022) 7180.
- [23]C. M. Fusco *et al.*, *Neuronal ribosomes exhibit dynamic and context-dependent exchange of ribosomal proteins*, Nature Communications, 12 (1) (2021) 6127.
- [24]T. R. Powell *et al.*, *The genome-wide expression effects of escitalopram and its relationship to neurogenesis, hippocampal volume, and antidepressant response*, American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 174 (4) (2017) 427–434.
- [25]L. Akotkar *et al.*, *Antidepressant effect of alpha lipoic acid in rats exposed to chronic unpredictable mild stress: Putative role of neurotransmitters and 5ht3 receptor*, Future Pharmacology 3 (2) (2023) 407–425.
- [26]B. P. Brennan et al*.*, *A placebo-controlled trial of acetyl-l-carnitine and α-lipoic acid in the treatment of bipolar depression*, Journal of Clinical Psychopharmacology 33 (5) (2013) 627–635.
- [27]J. Kleinkauf-Rocha, L. D. Bobermin, P. de M. Machado, C.-A. Gonçalves, C. Gottfried, A. Quincozes-Santos, *Lipoic acid increases glutamate uptake, glutamine synthetase activity and glutathione content in C6 astrocyte cell line*, International Journal of Developmental Neuroscience 31 (3) (2013) 165–170.
- [28]M. R. Salazar, *Alpha lipoic acid: a novel treatment for depression*, Medical Hypotheses 55 (6) (2000) 510–512.
- [29]M. C. C. Silva et al*.*, *Evidence for protective effect of lipoic acid and desvenlafaxine on oxidative stress in a model depression in mice*, Progress in Neuro-Psychopharmacology and Biological Psychiatry 64 (4) (2016) 142–148.
- [30]I. Croy, T. Hummel, *Olfaction as a marker for depression*, Journal of Neurology 264 (4) (2017) 631–638.
- [31]Q. Li et al*.*, *Reduced amount of olfactory receptor neurons in the rat model of depression*, Neuroscience Letters, 603 (2015) 48–54.
- [32]A. Eskelund et al., *Drugs with antidepressant properties affect tryptophan metabolites differently in rodent models with depression-like behavior*, Journal of Neurochemistry 142(1) (2017) 118–131.