



**«Digital Applications and Innovation»
Postgraduate Program (MSc)
2020-2021**

Capstone Project:

“Models for diagnosing schizophrenia, schizoaffective disorder, bipolar disorder and depression using genetic data. The aim of the capstone project is the analysis of a set of genetic data from patients with schizophrenia and depression with machine learning techniques, in order develop and evaluate a predictive model.”

**Lazarou Evgenia
Student ID: ΠΜΚ201905
Supervising Professor: Themis Exarchos**

Table of contents.

Abstract	4
Introduction - Problem Description.....	7
Schizophrenia.	7
Differences in schizophrenia and schizoaffective disorder.	19
Depression.....	20
Bipolar Disorder.	31
Description of Tool / Method / Technical Solution.	38
Programming Language: Python.....	38
GenBank.....	38
Dataset: Gene expression of schizophrenia and bipolar.....	38
Methods.	38
Application of Tool / Method / Technical Solution.....	43
Presentation of the results.	60
Dataset.	60
Our estimated results.....	60
Training.....	60
Feature Selection and Cross-Validation.	61
The Model - Random Forest Classifier.....	61
Conclusions.....	62
Bibliography.....	63
Annexes.	65

Abstract.

A disease is a specific irregular condition that negatively affects and influences the structure or function of all or part of an organic entity (for example a human organism), and that is not because of any immediate external injury¹. Diseases are regularly known to be medical conditions that are associated with specific symptoms, manifestations, and signs. A disease may be caused by outer factors such as pathogens and microbes or by internal dysfunctions. Internal dysfunctions of the immune system can produce a wide range of different diseases, including different types of immunodeficiency, hypersensitivity, allergies, and autoimmune disorders.

In humans, disease is regularly utilized more extensively to refer to any condition that causes pain, dysfunction, distress, social issues, or death to the individual afflicted, or even comparable problems for those in contact with the person. In this more extensive sense, diseases sometimes include injuries, disorders, disabilities, syndromes, infections, isolated symptoms, deviant behaviors, and atypical variations of structure and capacity, while in different contexts and for other purposes these may be considered distinguishable and recognizable categories. Diseases can affect individuals not only physically, but also mentally, as contracting and living with a disease can alter the affected individual's perspective on himself or life.

There are four principal types of disease:

1. Infectious diseases,
2. Deficiency diseases,
3. Hereditary diseases (including both genetic diseases and non-genetic hereditary diseases), and
4. Physiological diseases.

In many cases, terms such as disease, disorder, morbidity, sickness and illness are used interchangeably; however, there are situations when specific terms are considered preferable^[1].

Diseases can likewise be classified and characterized in other ways, such as communicable versus non-communicable diseases.

“Communicable diseases are the diseases that are caused by infectious agents and can be transmitted from an infected individual to other people, animals, or other sources in the environment. In contrast, non-communicable diseases are the diseases that are not transferred from an infected person to another via any means and are mostly caused by factors like improper lifestyle and eating habits”².

¹ <https://en.wikipedia.org/wiki/Disease>

² <https://microbenotes.com/communicable-vs-non-communicable-diseases/>

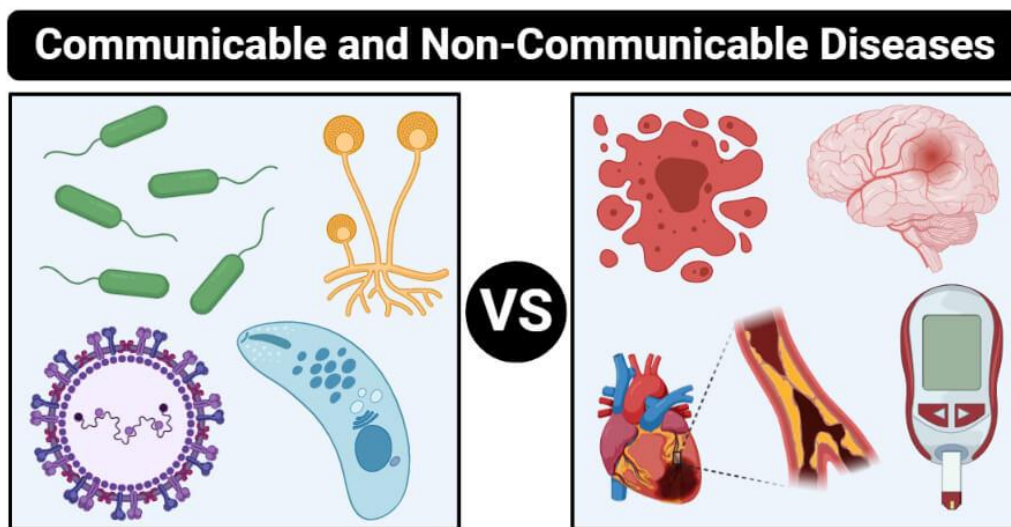


Image 1 - Communicable and Non-Communicable Diseases³

The top worldwide causes of death in 2019⁴, in order of total number of lives lost, were associated with three wide themes: cardiovascular (ischaemic heart disease, stroke), respiratory (chronic obstructive pulmonary disease, lower respiratory infections) and neonatal conditions – which include birth asphyxia and birth trauma, neonatal sepsis and infections, and preterm birth complications. Despite the fact that the world's biggest killer is ischaemic heart disease, responsible for 16% of the world's total deaths, in developed countries, the diseases that cause the most serious health conditions overall tend to be the neuropsychiatric conditions, such as schizophrenia, bipolar disorder, depression, and anxiety. In the future, given the aging population and worsening social problems, the number of people with mental and behavioural disorders will increase considerably.

Mental disorder⁵ is an expansive, generic name for a category of illnesses that may include affective or emotional instability, behavioral and mood dysregulation, cognitive dysfunction or impairment. Specific illnesses known as mental illnesses include major depression, generalized anxiety disorders, bipolar disorder, schizophrenia, and attention deficit hyperactivity disorder, etc. Mental illness can have biological roots (e.g., anatomical, chemical, or genetic) or psychological (e.g., trauma or conflict) origins. It can impair the affected person's ability to communicate, work or study.

According to the above study conducted in 2019, common mental illnesses around the globe include depression, which affects about 264 million, bipolar disorder, which affects about 45 million, dementia, which affects about 50 million, and schizophrenia and different psychoses, which affects about 20 million people. Stress, genetics, nutrition, perinatal infections and exposure to environmental hazards are likewise contributing variables to mental issues.

Despite the fact that these diseases affect an increasing percentage of the world's population, the causes that cause them remain under investigation.

³ "Communicable vs non-communicable diseases", <https://microbenotes.com/communicable-vs-non-communicable-diseases/>

⁴ <https://www.who.int/>

⁵ <https://en.wikipedia.org/wiki/Disease>

In this work, a genetic analysis in three serious mental conditions - schizophrenia, schizoaffective disorder bipolar disorder and depression will be conducted and models for diagnosing schizophrenia and bipolar disease using genetic data will be developed and presented. The aim of this work is to analyze a set of genetic data from patients with schizophrenia, bipolar disorder and schizoaffective disorder, as well as healthy patients, with machine learning techniques, in order to develop and evaluate a predictive model.

Objective: Schizophrenia, bipolar disorder and schizoaffective disorder are chronic and debilitating neuropsychiatric disorders. It has been suggested that and genetics underlie the pathophysiology of these psychiatric disorders.

Methods: In this work were investigated: 84 bipolar patients, 58 schizophrenia patients, 13 schizoaffective patients, and 391 healthy controls using genetic analysis and machine learning methods. A number of genetic and other data were mined from a predefined dataset. These data were compared between different predictive models and used for machine learning to classify patients with mental disorders and healthy controls.

Results: In classifying schizophrenia patients and healthy controls via the predefined dataset, the selected machine learning algorithms and the predictive models showed an encouraging level of performance. The overall accuracy reached 87%.

Conclusion: In conclusion, the possibility firmly exists to classify schizophrenia patients and healthy controls using predictive models as well as that there is an underlying pattern of genetics among patients with schizophrenia.

Keywords: schizophrenia; bipolar disorder; depression; schizoaffective disorder; machine learning; genetic data;

Introduction - Problem Description.

In many countries, schizophrenia, schizoaffective, bipolar and other mental disorders have been consistently overdiagnosed and affective disorders underdiagnosed, particularly among vulnerable and lower socioeconomic groups. The general causes of such misdiagnoses include overreliance on the classic thought disorder symptoms as pathognomonic of schizophrenia or other psychoses and, for affective disorders, lack of clearly characterized boundaries between normal and abnormal behaviors and failure to realize that patients with affective illness can manifest cognitive symptoms. In addition to the above factors, misdiagnosis among some specific social groups results from such factors as cultural differences in language and mannerisms, difficulties in relating between patients and therapists, and other myths related to mental and affective disorders.

In the modern era, due to the recent findings and correlations, scientists have significantly advanced the understanding of the genetic architecture of schizophrenia. In this work, a genetic analysis in three mental conditions (schizophrenia, schizoaffective disorder, bipolar disorder) will be conducted while the developed models for diagnosing schizophrenia, schizoaffective disorder, and bipolar disorder using genetic data setting will set the example on how this knowledge may be translated into advances that will improve diagnosis and patient health control.

Schizophrenia.

Schizophrenia⁶ is a severe mental disorder, affecting more than 20 million individuals worldwide. Psychoses, including schizophrenia, are characterized by distortions in thinking, perception, emotions, language, sense of self, self-appreciation, and behaviour. Regular psychotic experiences include hallucinations (hearing, seeing or feeling things that are not there) and delusions (fixed false beliefs or suspicions that are firmly held even when there is evidence to the contrary). The problem can make it hard for individuals influenced to work or concentrate ordinarily.

Particularly incriminating acts the Stigma and the social discrimination that is also leading to an absence of access to health and social services. Furthermore, people with psychosis are at high risk of exposure to human rights violations, such as long-term confinement in institutions.

Schizophrenia typically begins in late adolescence or early adulthood. Treatment with medicines and psychosocial support is usually effective. With appropriate treatment and social support, affected individuals can lead a productive life and be integrated in society. Facilitation of assisted living, supported housing and supported employment and work environments can act as a foundation from which individuals with severe mental disorders, including schizophrenia and bipolar diseases, can accomplish various recovery goals as they frequently face obstacles in acquiring or retaining a house to live and ordinary employment.

⁶ <https://en.wikipedia.org/wiki/Schizophrenia>

Schizophrenia is a complex disorder, caused by both genetic and environmental factors and their associations. Research on pathogenesis has customarily centered around neurotransmitter systems in the brain, particularly those involving dopamine. Schizophrenia has been considered a separate disease for over a century, but in the absence of clear biological markers, diagnosis has historically been based on signs, symptoms, and manifestations.

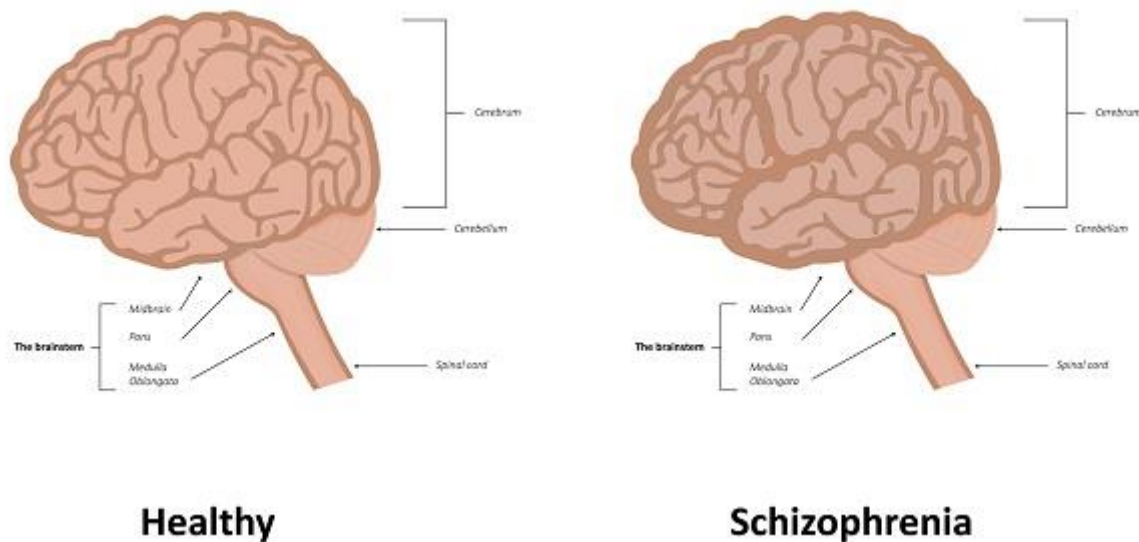


Image 2 - Schizophrenia Brain vs Normal Brain⁷

Types of Schizophrenia.

There are actually several distinct types of schizophrenia depending on the individual's symptoms, however, for the most part; the fundamental types of schizophrenia include paranoid schizophrenia, catatonic schizophrenia, disorganized or hebephrenic schizophrenia, residual schizophrenia, and undifferentiated schizophrenia.

Paranoid schizophrenia.

This type of schizophrenia is the most common expression of the disease. There are possibilities to be developed later in life than in other forms. Symptoms include hallucinations and/or delusions, but the abilities to speak and feel emotions may not be affected.

Catatonic schizophrenia.

This is the rarest schizophrenia diagnosis, characterised by unusual, restricted, and unexpected movements. The patient may regularly switch between being dynamic or

⁷ 5. "Mutated Genes in Schizophrenia Map to Brain Networks" - Schizophrenia networks in the prefrontal cortex area of the brain. Dr. Mary-Claire King, University of Washington, <https://www.nih.gov/news-events/nih-research-matters/mutated-genes-schizophrenia-map-brain-networks>

extremely still. In other cases, the patient may not talk much, and/or impersonate others' speech and movement.

Hebephrenic schizophrenia.

Otherwise called 'disorganised schizophrenia', this type of schizophrenia typically develops when the individual is 15-25 years old. Symptoms include disorganised and confused behaviours and thoughts, alongside short-enduring delusions and hallucinations. The individual may have disorganised speech patterns and difficulties to be understood.

Individuals living with disorganised schizophrenia frequently show practically little or no emotions in their facial expressions, voice tone, or mannerisms.

Residual schizophrenia.

The individual is possible to be diagnosed with residual schizophrenia if he/she has a history of psychosis, however only experience the negative symptoms (such as slow movement, poor memory, lack of concentration and poor hygiene).

Undifferentiated schizophrenia.

The diagnosis of the patient may have some indications of paranoid, hebephrenic, or catatonic schizophrenia, but these signs are not obviously fit into one of the previous types alone.

Genetics of Schizophrenia.

Genetics constitute a crucial risk factor to schizophrenia. In the last decade, molecular genetic research and examination has produced novel discoveries, infusing optimism about discovering the biological roots and foundation of schizophrenia. The genetic architecture of schizophrenia has demonstrated to be highly complex and perplexing, heterogeneous and polygenic. The disease risk is established by different regular genetic variants of only very small individual effect and by rare, exceptionally penetrant genetic variants of larger effects.

Numerous studies have reported that genetic variants play a significant and critical role in the pathogenesis of schizophrenia. In the past few years, the genetics knowledge base of schizophrenia has advanced extensively. These advances generally experienced **Genome-Wide Association Studies**⁸ (GWAS) and **Copy Number Variations**⁹ (CNVs) studies. Moreover, exome sequencing by utilizing the next-generation DNA sequencers has been introduced. In the following years, the high-throughput sequencing is expected to play an important role and assume a significant part in the diagnosis of the condition. These discoveries have produced new hypotheses and speculations about etiology of schizophrenia and other psychotic disorders and they can demonstrate future research strategies and examination techniques. Therefore, it is expected that further studies and investigations will yield further insights.

⁸ https://en.wikipedia.org/wiki/Genome-wide_association_study

⁹ https://en.wikipedia.org/wiki/Copy_number_variation

Genome-wide association study (GWAS) of schizophrenia (SCZ)

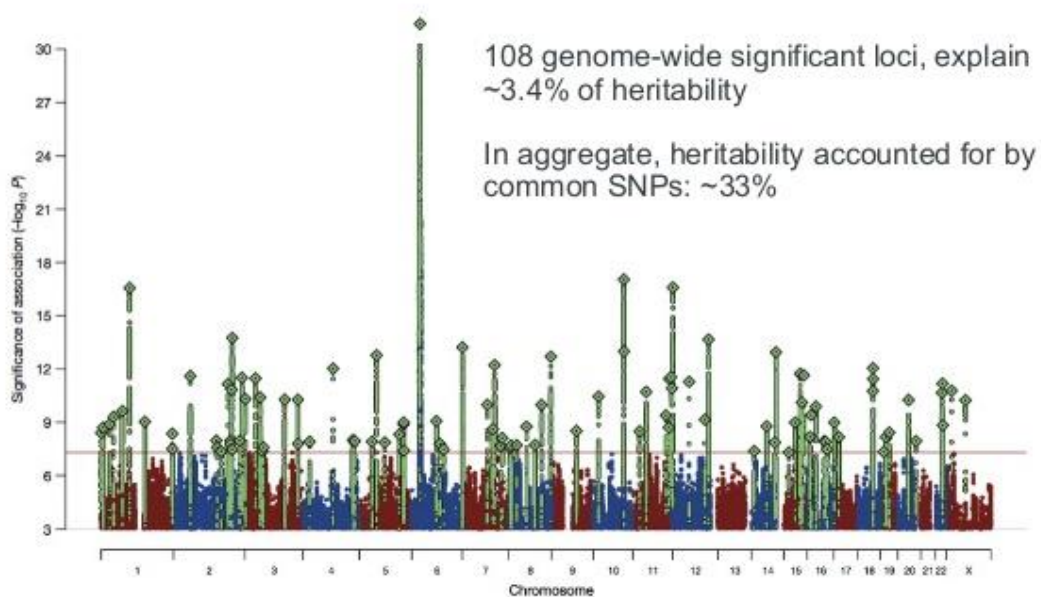


Image 3 - Genome-wide association study of Schizophrenia¹⁰

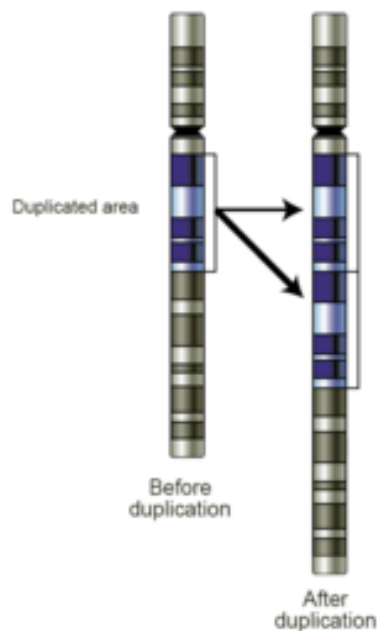


Image 4 - This gene duplication has created a copy number variation. The chromosome now has two copies of this section of DNA, rather than one¹¹.

¹⁰ <https://www.slideshare.net/yasirmhm/analysis-of-genomewide-association-studies-uncovers-genetic-loci-shared-between-schizophrenia-and-cognitive-traits>

¹¹ https://en.wikipedia.org/wiki/Gene_duplication

A fundamental message emerging from genome-wide association studies of copy number variations (CNVs) associated with the disease is that its genetic basis does not necessarily conform to classical nosological disease boundaries. Certain CNVs confer not only high relative risk of schizophrenia but also of other psychiatric disorders. The structural variations associated with schizophrenia can involve several genes and the phenotypic syndromes, or the 'genomic disorders', have not yet been characterized.

Since **genes are considered to be the strongest risk factor for schizophrenia**, some grasp of this complex research domain is relevant in many clinical contexts.

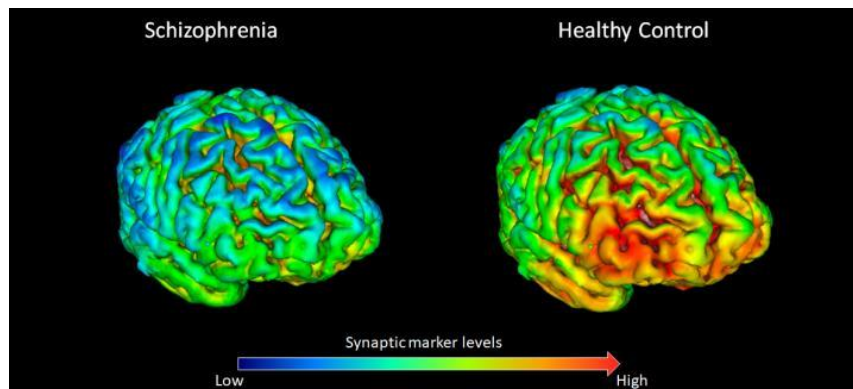


Image 5 - New evidence revealed that patients with Schizophrenia have reduced brain connections (the synaptic protein SV2A)¹²

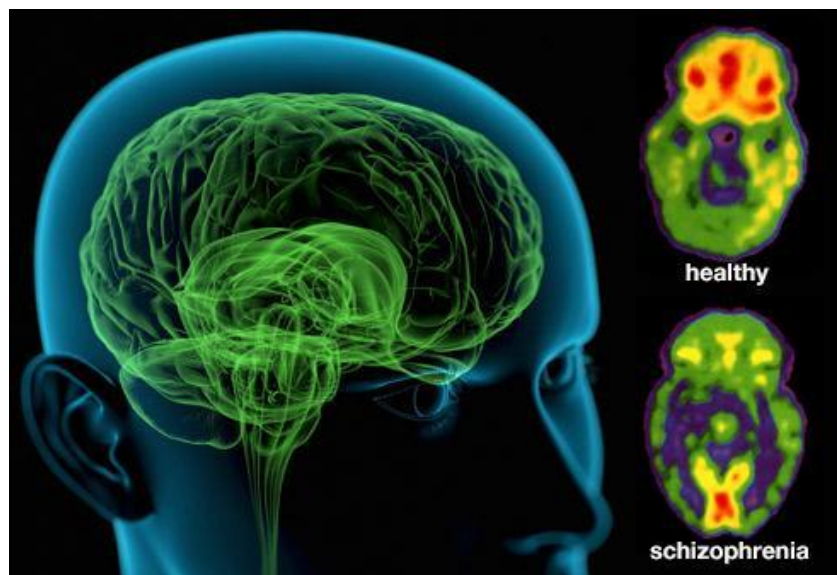


Image 5 - A Visual Guide to Schizophrenia¹³

¹² "New Study Finds Evidence For Reduced Brain Connections In Schizophrenia" - Credit: E. Onwordi at MRC London Institute of Medical Sciences (LMS),

<https://scienmag.com/new-study-finds-evidence-for-reduced-brain-connections-in-schizophrenia/>

¹³ "A Visual Guide to Schizophrenia", <https://www.webmd.com/schizophrenia/ss/slideshow-schizophrenia-overview>

Historical Data.

In the first half of the 20th century, family studies demonstrated that the rate of schizophrenia was higher in relatives of patients with schizophrenia than in the general population (Rüdin, 1916; Kahn, 1923; Schulz, 1932; Kallmann, 1938). Twin studies documented that the concordance rate (i.e., both twins suffering from schizophrenia) was elevated in monozygotic (MZ) twins compared to dizygotic (DZ) twins (Luxenburger, 1928; Kallmann, 1946; Slater, 1953). These early twin studies were later criticized for various methodological reasons (Rosenthal, 1959, 1962; *vide infra*). From the 1960s, improved twin (Kringlen, 1967; Fischer, 1973) and adoption studies (Heston, 1966; Rosenthal et al., 1968; Kety et al., 1975; Tienari et al., 1985) became crucial in determining the familial clustering and concordance rates for schizophrenia. By indicating a strong genetic component in the etiology of the illness, the studies contributed to undermine the psychoanalytical hypothesis of schizophrenic causation, claiming that schizophrenogenic rearing was either a necessary or sufficient cause for developing schizophrenia. The basic intuition behind the twin studies is the following: given that MZ twins (sharing 100% of their genes) and DZ twins (sharing 50% of their genes) share the environment they are raised in, higher concordance rates in MZ over DZ twins most likely result from genetic similarity. Estimates of concordance rates for schizophrenia, based on European twin studies from 1963 to 1987, show higher rates for MZ (48%) than for DZ twins (17%; Gottesman, 1991), and similar concordance rates were reported in European and Japanese twin studies from 1992 to 1999—41%–65% for MZ vs. 0%–28% for DZ twins (Cardno and Gottesman, 2000). A meta-analysis (Sullivan et al., 2003) of twin studies estimates the genetic liability to schizophrenia at 81% (95% CI, 73%–90%), whereas shared environmental influences were estimated to be 11% (95% CI, 3%–19%). Finally, a few studies of children of discordant MZ twins found a similar risk of schizophrenia spectrum disorders in the children of the affected and unaffected MZ twin (Gottesman and Bertelsen, 1989; Kringlen and Cramer, 1989), presumably indicating that unaffected MZ twins carry silent (non-expressed) susceptibility genes for schizophrenia. By contrast, for children of discordant DZ twins, the risk was higher in the children of the affected DZ twin compared to the children of the unaffected DZ twin (Gottesman and Bertelsen, 1989)¹⁴.

Adoption studies have documented that schizophrenia spectrum disorders are more frequent in adopted-away children of mothers with schizophrenia than in their control adoptees (Heston, 1966; Rosenthal et al., 1968; Kety et al., 1975, 1994). A cross-fostering study (Wender et al., 1974) found that children of healthy parents, adopted by a family where one of the parents later developed schizophrenia, did not have an increased risk of developing schizophrenia. Other studies (Heston, 1966; Higgins, 1976) found that children of mothers with schizophrenia had the same risk of developing the disorder independent of whether they were raised by their biological mothers or by adopting parents with no history of mental illness. A Finish adoption study (Tienari et al., 1985, 2004) found that markedly dysfunctional rearing environments (the adoptive families were initially assessed and classified on a scale ranging from “1. healthy” to “5. severely disturbed”) predicted schizophrenia spectrum disorders in adopted-away children of mothers with schizophrenia but not in their genetically undisposed controls. Interestingly, similar results were reported in the Danish High-Risk study (Mednick et al., 1987), which found

¹⁴ Genetics of Schizophrenia: Overview of Methods, Findings and Limitations
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5480258/>

increased risk of schizophrenia in children of mothers with schizophrenia, who were exposed to unstable parenting or raised in public childcare institutions (Parnas et al., 1985).

The Human Genome Project.

The Human Genome Project (HGP) was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. All our genes together are known as our "genome."¹⁵ After the Human Genome Project was completed in 2001, research on the genetic underpinnings of schizophrenia has expanded and opened the door to accelerating our understanding of molecular mechanisms of human diseases¹².

Linkage analysis approach.

The first DNA-based method was "linkage analysis", which aimed at discovering genomic regions in samples of affected extended or nuclear families and sibling pairs without implicating a specific allelic variant¹³.

"Linkage analysis is based on the observation that genetic markers, which are located physically close on the same chromosome, tend to be inherited together, i.e., they remain "linked" during meiosis".

Study: Genome-wide Association Analysis Identifies 14 New Risk Loci for Schizophrenia¹⁶:

Schizophrenia is a heritable disorder with substantial public health impact. We conducted a multi-stage genome-wide association study (GWAS) for schizophrenia beginning with a Swedish national sample (5,001 cases, 6,243 controls) followed by meta-analysis with prior schizophrenia GWAS (8,832 cases, 12,067 controls) and finally by replication of SNPs in 168 genomic regions in independent samples (7,413 cases, 19,762 controls, and 581 trios). In total, 22 regions met genome-wide significance (14 novel and one previously implicated in bipolar disorder). The results strongly implicate calcium signaling in the etiology of schizophrenia, and include genome-wide significant results for CACNA1C and CACNB2 whose protein products interact. We estimate that ~8,300 independent and predominantly common SNPs contribute to risk for schizophrenia and that these collectively account for most of its heritability. Common genetic variation plays an important role in the etiology of schizophrenia, and larger studies will allow more detailed understanding of this devastating disorder¹⁷.

Many chromosomal regions may contain schizophrenia susceptibility loci. Notably, these loci do not themselves confer risk but they may harbor variants that do. These results also made it clear that the power of the linkage design was too weak to address genomic loci with small effects; the sample size requirement necessary to detect linkage was simply practically

¹⁵ <https://www.genome.gov/human-genome-project/What>

¹⁶ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3827979>

¹⁷ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3892949/>

unachievable. Hence, other DNA-based methods were required to key in on the genes potentially involved in the etiology of schizophrenia.

Gene	Name	Location	Function(s)
<i>CACNA1C</i>	Calcium channel, voltage-dependent, L type, alpha 1C subunit	12p13.3	Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization
<i>COMT</i>	Catechol-O-methyltransferase	22q11.21	Key enzyme in degradation of dopamine and norepinephrine
<i>CSMD1</i>	CUB and Sushi multiple domains 1	8p23.2	One of the proteins that modulate the classical complement pathway, part of the immune system
<i>CYP2D6</i>	Cytochrome P450 2D6	22q13.1	Key enzyme in drug metabolism
<i>C10orf26</i>	Chromosome 10 open reading frame 26	10q24.32	Unknown
<i>DISC1</i>	Disrupted in schizophrenia 1	1q42	Neurite outgrowth, cortical development, synaptic function
<i>DRD1</i>	Dopamine receptor D1	5q35.1	D1 receptors regulate neuronal growth and development, mediate behavioral responses, and modulate D2 receptor-mediated events
<i>DRD2</i>	Dopamine receptor D2	11q23	D2 receptors regulate motor activities and information processing in the brain
<i>DTNBP1</i>	Dystrobrevin binding protein 1	6p22	Neurodevelopment and synaptic transmission
<i>HLA-DQB1</i>	major histocompatibility complex, class II, DQ beta 1	6p21.3	Plays a central role in the immune system by presenting peptides derived from extracellular proteins
<i>HTR2C</i>	Serotonin receptor 2C	Xq24	Modulate mood, food intake behavior, and feeling of satiety
<i>MCR4</i>	Melanocortin 4 receptor	18q22	Modulate food intake behavior and feeling of satiety
MHC region	Major histocompatibility complex	6p21-22	Immune function; neurodevelopment, synaptic plasticity
<i>MIR137</i>	MicroRNA 137	1p23.3	Post-transcriptional regulation of messenger RNAs; neuron maturation, adult neurogenesis
<i>MTHFR</i>	Methylenetetrahydrofolate reductase	1p36.3	Key enzyme in folate metabolism
<i>TCF4</i>	Transcription factor 4	18q21.2	Neuronal transcriptional factor, neurogenesis
<i>TPH1</i>	Tryptophan hydroxylase 1	11p15.3	Key enzyme in biosynthesis of serotonin
<i>ZNRF804A</i>	Zinc finger protein 804A	2q32.1	Transcription factor, neuronal connectivity in the dorsolateral prefrontal cortex

Image 6 - Recent findings in genetics of schizophrenia - Genes and their functions¹⁸

The genome-wide association studies (GWAS) employ a case-control study design, interrogate the genome purely empirically for associations between common genomic variants or loci and the disorder. The reasoning behind the GWAS approach is that if specific allele variants are found more frequently in patients than in their controls, then the allele variants may be indicative of a genetic association. Associations were found at dopamine receptor D2, in several genes involved in glutamatergic neurotransmission and synaptic plasticity, and in tissues with central immune functions. A proportion of the variance in genetic liability is apparently not accounted for by common genetic variants. Associations between schizophrenia and the >100 susceptibility loci, the identified CNVs and SNVs. Common alleles of only a very small effect, which do not individually achieve significance but which collectively form a substantial polygenic component of schizophrenia risk,

Associations between common (SNPs) or uncommon (CNVs, SNVs) genetic variants and schizophrenia, though statistical facts, are not necessarily indexes of causal pathways, Associations that are non-specific to schizophrenia but indicative of a genetic vulnerability to several mental disorders.

Gene expression profiles:

DTNBP1 - dystrobrevin binding protein 1¹⁹.

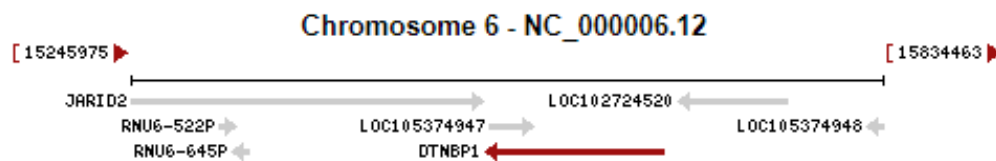
This gene encodes a protein that may play a role in organelle biogenesis associated with melanosomes, platelet dense granules, and lysosomes. A similar protein in mouse is a component of a protein complex termed biogenesis of lysosome-related organelles complex

¹⁸ "Recent findings in genetics of schizophrenia - Genes and their functions",

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3892949/>

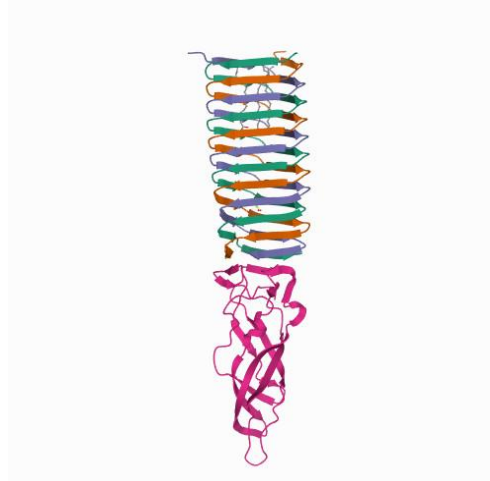
¹⁹ <https://www.ncbi.nlm.nih.gov/gene/84062>

1 (BLOC-1), and binds to alpha- and beta-dystrobrevins, which are components of the dystrophin-associated protein complex (DPC). Mutations in this gene are associated with Hermansky-Pudlak syndrome type 7. This gene may also be associated with schizophrenia.



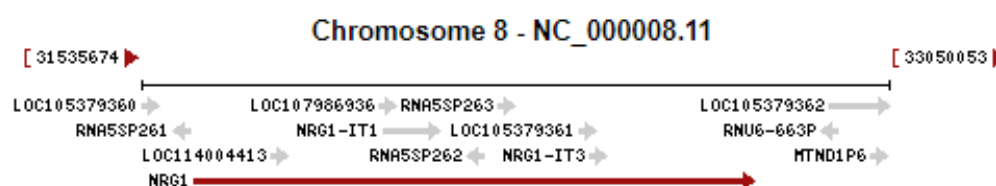
6P22²⁰.

Photorhabdus Virulence Cassette (PVC) PAAR repeat protein Pvc10 in complex with a T4 gp5 beta-helix fragment modified to mimic Pvc8, the central spike protein of PVC.



NRG1- neuregulin 1²¹.

The protein encoded by this gene is a membrane glycoprotein that mediates cell-cell signaling and plays a critical role in the growth and development of multiple organ systems. An extraordinary variety of different isoforms are produced from this gene through alternative promoter usage and splicing. These isoforms are expressed in a tissue-specific manner and differ significantly in their structure, and are classified as types I, II, III, IV, V and VI. Dysregulation of this gene has been linked to diseases such as cancer, schizophrenia, and bipolar disorder (BPD).



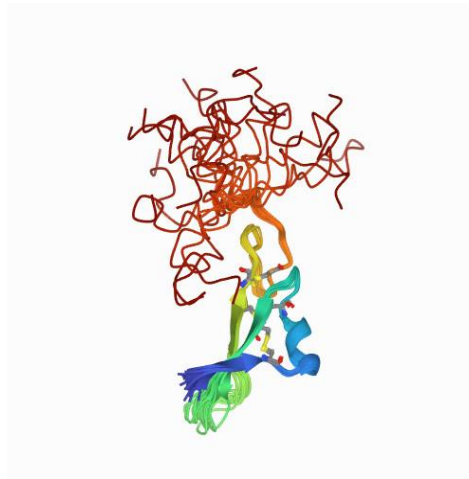
1HAE²².

²⁰ <https://www.rcsb.org/structure/6P22>

²¹ <https://www.ncbi.nlm.nih.gov/gene/3084>

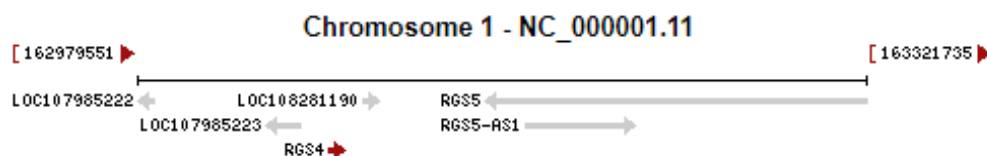
²² <https://www.rcsb.org/structure/1HAE>.

Heregulin-alpha epidermal growth factor-like domain, NMR, 20 structures.



RGS4 - regulator of G protein signaling 4²³.

Regulator of G protein signaling (RGS) family members are regulatory molecules that act as GTPase activating proteins (GAPs) for G alpha subunits of heterotrimeric G proteins. RGS proteins are able to deactivate G protein subunits of the Gi alpha, Go alpha and Gq alpha subtypes. They drive G proteins into their inactive GDP-bound forms. Regulator of G protein signaling 4 belongs to this family. All RGS proteins share a conserved 120-amino acid sequence termed the RGS domain. Regulator of G protein signaling 4 protein is 37% identical to RGS1 and 97% identical to rat Rgs4. This protein negatively regulate signaling upstream or at the level of the heterotrimeric G protein and is localized in the cytoplasm. Alternatively spliced transcript variants have been found for this gene.



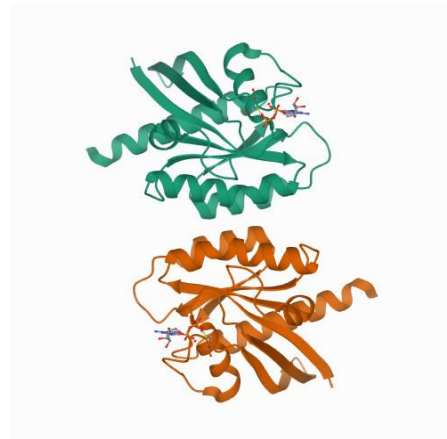
1Q21²⁴.

Crystal Structures at 2.2 angstroms resolution of the catalytic domains of normal RAS protein and an oncogenic mutant complexed with GSP.

Neuregulin 1 (NRG1) was originally identified as a 44-kD glycoprotein that interacts with the NEU/ERBB2 receptor tyrosine kinase to increase its phosphorylation on tyrosine residues. It is known that an extraordinary variety of different isoforms are produced from the NRG1 gene by alternative splicing. These isoforms include heregulins (HRGs), glial growth factors (GGFs) and sensory and motor neuron-derived factor (SMDF). They are tissue-specifically expressed and differ significantly in their structure.

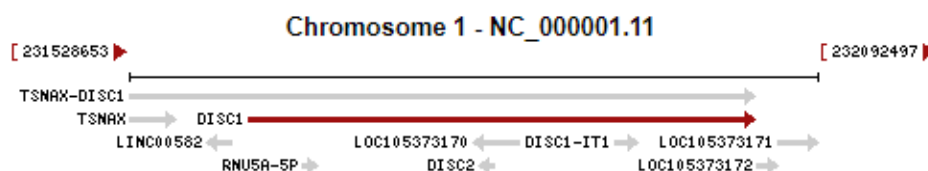
²³ <https://www.ncbi.nlm.nih.gov/gene/5999>

²⁴ <https://www.rcsb.org/structure/1Q21>



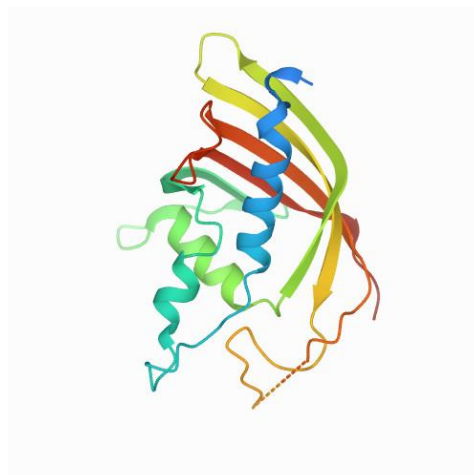
DISC1 - scaffold protein²⁵.

This gene encodes a protein with multiple coiled coil motifs which is located in the nucleus, cytoplasm and mitochondria. The protein is involved in neurite outgrowth and cortical development through its interaction with other proteins. This gene is disrupted in a t(1;11)(q42.1;q14.3) translocation which segregates with schizophrenia and related psychiatric disorders in a large Scottish family. Alternate transcriptional splice variants, encoding different isoforms, have been characterized.



1Q42²⁶.

Crystal structure analysis of the Candida albicans Mtr2.



²⁵ <https://www.ncbi.nlm.nih.gov/gene/27185>

²⁶ <https://www.rcsb.org/structure/1Q42>

Gene	Name	Location	Function	Associations
DTNBP1	dystrobrevin binding protein 1	6p22.3	This gene encodes a protein that may play a role in organelle biogenesis associated with melanosomes, platelet dense granules, and lysosomes.	Schizophrenia.
NRG1	neuregulin 1	8p12	The protein encoded by this gene is a membrane glycoprotein that mediates cell-cell signaling and plays a critical role in the growth and development of multiple organ systems.	Cancer, schizophrenia, and bipolar disorder (BPD).
RGS4	regulator of G protein signaling 4	1q23.3	Regulator of G protein signaling (RGS) family members are regulatory molecules that act as GTPase activating proteins (GAPs) for G alpha subunits of heterotrimeric G proteins.	Schizophrenia.
DISC1	scaffold protein	1q42.2	This gene encodes a protein with multiple coiled coil motifs which is located in the nucleus, cytoplasm and mitochondria.	Schizophrenia and related psychiatric disorders.
DAOA	D-amino acid oxidase	12q24.11	This gene encodes the peroxisomal enzyme D-amino acid oxidase.	Schizophrenia, bipolar disorder (BPD).
COMT	catechol-O-methyltransferase	22q11.21	Catechol-O-methyltransferase catalyzes the transfer of a methyl group from S-adenosylmethionine to catecholamines, including the neurotransmitters dopamine, epinephrine, and norepinephrine.	Schizophrenia.
PRODH	proline dehydrogenase 1	22q11.21	This gene encodes a mitochondrial protein that catalyzes the first step in proline degradation.	Schizophrenia.
CACNA1C	calcium voltage-gated channel subunit alpha1 C	12p13.33	This gene encodes an alpha-1 subunit of a voltage-dependent calcium channel.	Schizophrenia.
MC4R	melanocortin 4 receptor	18q21.32	The protein encoded by this gene is a membrane-bound receptor and member of the melanocortin receptor family.	Schizophrenia.
MIR137	microRNA 137	1p21.3	microRNAs (miRNAs) are short non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs.	Schizophrenia.

Table 1 – Genes associated with schizophrenia and their functions

Differences in schizophrenia and schizoaffective disorder.

Study: The continuing story of schizophrenia and schizoaffective disorder: One condition or two²⁷?

It has been long debated whether schizophrenia and schizoaffective disorder represent a single disorder or two distinct conditions (Cheniaux et al., 2008; Malhi et al., 2008). These syndromes share psychotic symptoms, yet schizoaffective disorder involves a concurrent mood disorder and does not require evidence of a decline in role functioning.

More specifically, schizoaffective disorder is a chronic and ongoing mental health condition portrayed principally by symptoms of schizophrenia, for example, hallucinations or delusions, and symptoms of a mood disorder, such as mania and depression.

Numerous individuals with schizoaffective disorder frequently inaccurately diagnosed at first with bipolar disorder or schizophrenia. Because schizoaffective disorder is less well-examined than the other two conditions, many interventions are borrowed from their treatment approaches.

Schizoaffective is moderately rare, with a lifetime prevalence of only 0.3%. Males and females experience schizoaffective disorder at the same rate, but males often develop the illness at an earlier age. Schizoaffective disorder can be managed effectively with medication and treatment. Co-occurring substance use disorders are a serious risk and require integrated treatment.

In spite of the fact that schizophrenia and schizoaffective disorder remain separable in diagnostic frameworks, the validity of the distinction and differentiation is unsure.

Causes of schizoaffective disorder.

The specific cause of schizoaffective disorder is obscure. A combination of causes may contribute to the development of schizoaffective disorder.

Genetics.

Genetics also constitute a crucial risk factor to schizoaffective disorder. This does not necessarily mean that if a relative has the disorder, the individual will absolutely occur with the disorder as well. In any case, it implies that there is a more prominent possibility for the individual to develop the illness.

Brain chemistry and structure.

As in Schizophrenia, the brain function and structure may be a risk factor to schizoaffective disorder as well. Cerebrum examines are assisting to advance research in the brain's area.

²⁷ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6370594/>

Stress.

Stressful events and upsetting occasions such as a relative's death, end of a marriage or loss of a job can trigger symptoms or an onset of schizoaffective disorder.

Drug use.

Psychoactive drugs such as LSD have been connected to the development of schizoaffective disorder.

Diagnosis.

Schizoaffective disorder can be challenging to diagnose because it has symptoms of both schizophrenia and either depression or bipolar disorder.

There are two significant types of schizoaffective disorder: **bipolar type** and **depressive type**.

To be diagnosed with schizoaffective disorder an individual must have the following symptoms.

- Delusions or hallucinations for two or more weeks in the absence of a major mood episode.
- A period during which there is a significant mood disorder, either depression or mania, that occurs at the same time that symptoms of schizophrenia are present
- Symptoms that meet criteria for a major mood episode are present for the majority of the total duration of the illness.
- The abuse of drugs or a medication are not responsible for the symptoms of the disorder.

Methods and Findings until now.

Limitations and challenges that confront this field of research are the intrinsic difficulty of the object of inquiry, to methodological or technological inadequacies and insufficiencies, or in many cases to a mistaken definition of the examination issue. In the last decade, genetic research in schizophrenia and schizoaffective disorder has experienced a new dawn infused by a regained optimism due to newly developed, far more advanced molecular, technological and statistical methods.

Depression.

Depression (major depressive disorder or clinical depression)²⁸ is a typical but serious mood disorder. It causes severe symptoms that affect how the individual who suffers

²⁸ https://en.wikipedia.org/wiki/Major_depressive_disorder

feels, thinks, and handles daily activities, such as sleeping, eating, or working. To be diagnosed with depression, the symptoms must be present for at least fourteen days. Major depressive disorder is a disabling syndrome characterized by persistent low mood and reduced enjoyment along with additional signs and symptoms including reduced concentration, energy, and self-esteem and altered appetite and sleep quality.

Types of Depression.

Some types of depression are slightly different, or they may develop under special or unique circumstances, such as:

Persistent depressive disorder.

Persistent depressive disorder (also called dysthymia) is a depressed mood that lasts for at least two years. An individual diagnosed with persistent depressive disorder may have episodes of significant depression along with periods of less severe symptoms, but symptoms must last for two years to be considered persistent depressive disorder.

Postpartum depression.

Postpartum depression is substantially more serious than the “baby blues” (relatively mild depressive and anxiety symptoms that typically clear within fourteen days after delivery) that numerous women experience after giving birth. Women with postpartum depression experience full-blown major depression during pregnancy or after delivery (postpartum depression). The feelings of extreme sadness, anxiety, and exhaustion that accompany postpartum depression may make it difficult for these new mothers to complete daily care activities for themselves and/or for their children.

Psychotic depression.

Psychotic depression occurs when an individual has extreme depression in addition to some type of psychosis, such as having disturbing false fixed beliefs (delusions) or hearing or seeing disturbing things that others cannot hear or see (hallucinations). The psychotic symptoms typically have a depressive “theme,” such as delusions of guilt, poverty, or illness.

Seasonal affective disorder.

Seasonal affective disorder is described by the onset of depression during the cold weather months, when there is less natural daylight. This depression generally lifts during spring and summer. Winter depression, typically accompanied by social withdrawal, increased sleep, and weight acquire, predictably returns every year in seasonal affective disorder.

Bipolar disorder.

Bipolar disorder is different from depression, but it is included in this list because a person with bipolar disorder experiences episodes of extremely low moods that meet the criteria for major

depression (called “bipolar depression”). An individual with bipolar disorder additionally experiences extraordinary high – euphoric or irritable – moods called “mania” or a less severe form called “hypomania.”

Instances of different kinds of depressive disorders recently added to the diagnostic classification of DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) incorporate problematic mood dysregulation disorder (diagnosed in children and adolescents) and premenstrual dysphoric disorder (PMDD).

Signs and Symptoms of Depression.

If an individual have been encountering some of the following signs and symptoms most of the day, nearly every day, for at least two weeks, he/she may be suffering from depression:

- Feelings of hopelessness, or pessimism
- Persistent sad, anxious, or “empty” mood
- Feelings of guilt, worthlessness, or helplessness
- Moving or talking more slowly
- Irritability
- Loss of interest or pleasure in hobbies and activities
- Decreased energy or fatigue
- Feeling restless or having trouble sitting still
- Difficulty concentrating, remembering, or making decisions
- Appetite and/or weight changes
- Difficulty sleeping, early-morning awakening, or oversleeping
- Thoughts of death or suicide, or suicide attempts
- Aches or pains, headaches, cramps, or digestive problems without a clear physical cause and/or that do not ease even with treatment

At this point we should mention that not everyone who has depression experiences all the above symptoms. Some individuals experience only a few symptoms while others may experience many of them.

Recent research suggests that depression is caused by a combination of genetic, biological, environmental, and psychological factors. Regarding the genetic complexity, in which multiple genes act in concert with non-genetic factors to produce a risk of major depressive disorder (MDD) could explain the pathogenesis of depression (or schizophrenia as mentioned earlier). The complex genetics of risk should reinvigorate research on the epidemiology and classification of mental disorders and explain the complex patterns of disease transmission within families.

Major depressive disorder (MDD) is a highly disabling mental illness involving an imbalance in brain chemicals, and it majorly contributes to the global burden of disease. According to the World Health Organization, an estimated 350 million people of all ages suffer from depression disorder globally. In a systematic review, the summary estimate of the prevalence of depression or depressive symptoms among medical students was 27.2%, and that of suicidal

ideation was 11.1%. A psychiatric disorder is not a sign of specific characteristics, but it reveals an opposite result with a highly prevalent heritability that accounts for major psychological, physical, and social impairments.

Historical Data.

Evidence for a genetic component to mood disorders has been documented consistently using family, twin, and adoption studies. The first genetic studies of mood disorders were conducted more than 70 years ago and included assessment of concordance rates for monozygotic and dizygotic twins with mood disorders. These early studies did not distinguish between bipolar depression and MDD-recurrent unipolar (MDD-RU). A recent review of twin studies in MDD-RU estimated heritability at 37%, with a substantial component of unique individual environmental risk but little shared environmental risk.

Family studies of MDD-RU have shown that first-degree relatives of MDD-RU probands are at increased risk of MDD-RU disorders compared with first-degree relatives of control probands. There was a twofold to fourfold increased risk of MDD-RU among the first-degree relatives of MDD-RU probands. Characteristics of MDD-RU disorders that yield a more heritable phenotype include early onset (ie, before age 30 years) and a high degree of recurrence. A third characteristic that may identify a separate group of disorders is the presence of psychosis. Additional genetic subtypes of MDD-RU may be identified through examination of comorbidities with panic disorder, other anxiety disorders, and alcoholism.

Genetics of Depression.

In the past 40 years, many twin and family members' studies of Major Depression have established that liability to Major Depression has a non-deterministic genetic component to its etiology as substantiated by a twin heritability of 31%–42% (*Sullivan et al., 2000*). This degree of twin heritability implies the need for large samples for reliable and reproducible gene identification (*Levinson et al., 2014, Wray et al., 2012*) yet complex diseases with similar twin heritabilities have had considerable success (e.g., type 2 diabetes). Twin heritability provides support for the logic of genetic searches but does not elucidate the most critical feature, the genetic architecture (the number of loci underlying liability to a complex disorder along the effect sizes and frequencies of the loci) (*Sullivan et al., 2012*).

The genetic study designs applied to Major Depression have mirrored those used in other common, complex disorders. The three major genetic approaches in depression, besides GWASs, are linkage analysis, candidate gene studies, and re-sequencing studies.

There have been several recent studies addressing the genetic architecture of depression. This review serves to assess what is known now about the genetics of depression, how it has expanded our knowledge and understanding of its mechanisms and systems, and how the information and knowledge can be leveraged to improve the consideration of individuals affected. We identify four priorities for how the field of Major Depression genetics research may move forward in future years, namely by increasing the sample sizes available for

genome-wide association studies (GWASs)²⁹, greater inclusion of diverse ancestries and low-income nations, the closer integration of psychiatric genetics with electronic medical records, and the development of the neuroscience toolkit for polygenic disorders.

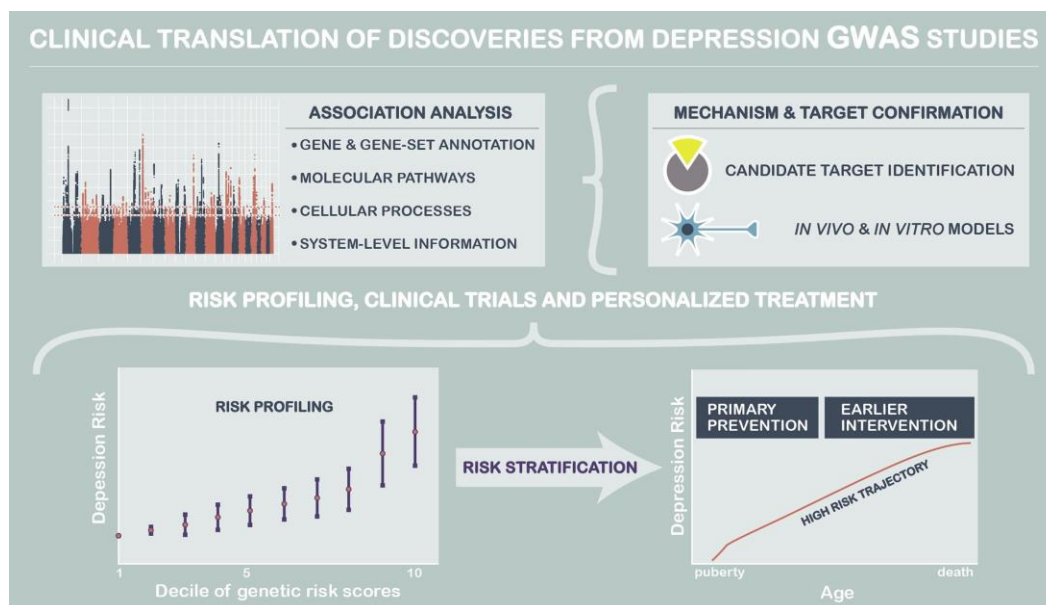


Image 7 - Clinical Translation of Genetic Associations from GWASs³⁰

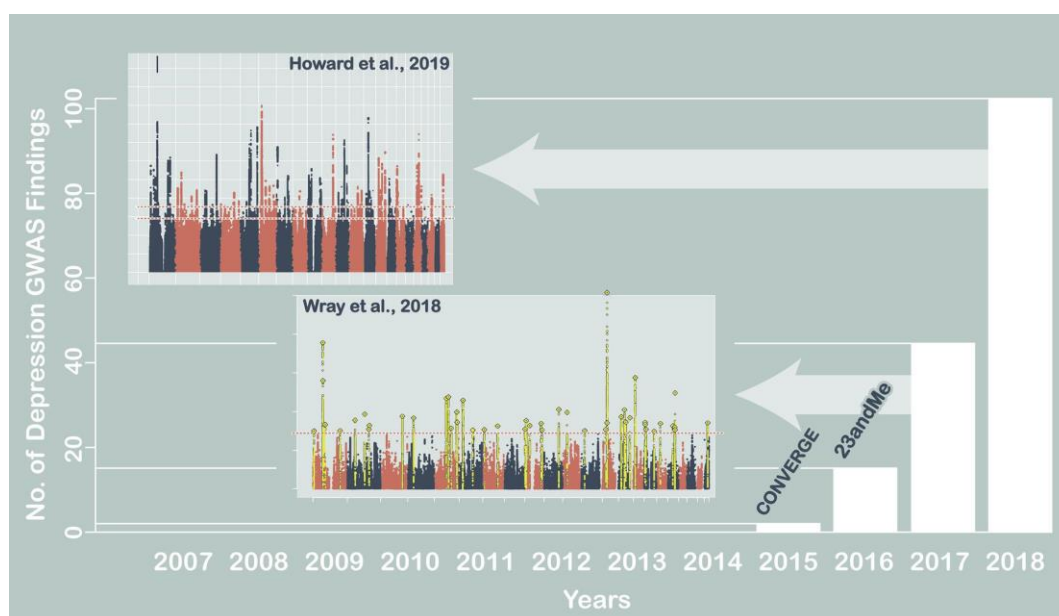


Image 8 - Recent Progress in Genome-wide Loci Discovery from GWASs³¹

²⁹ A genome-wide association study (GWAS) is a new approach that involves rapidly scanning several hundred thousand (up to 5 millions) markers across the complete sets of DNA of many people to find genetic variations associated with a particular trait.

³⁰ <https://www.sciencedirect.com/science/article/pii/S0896627319302855>

The clinical translation of genetic associations from GWASs starts with a MD-variant association set of test statistics and downstream gene, gene set, pathway, and other enrichment analyses

³¹ The image shows how the number of genome-wide loci discovered to date has increased from 44 to 102 in 12 months through the inclusion of additional participants, most of which have been categorized as MD cases or controls using criteria that would fall short of DSM or ICD diagnostic criteria.

Study: Genetics Factors in Major Depression Disease³²

In a meta-analysis, Gatt et al. attempted to identify genes associated with DD and common genes shared by the five severe psychiatric disorders: MDD, anxiety (including panic disorders), schizophrenia, bipolar disorder, the attention deficit–hyperactivity syndrome. Image 9 displays the data for the genes whose products are involved in monoaminergic neurotransmission.

Associations shown in meta-analyses between DDs and polymorphic variants of the genes linked to the exchange of monoamine neuromediators^{*}.

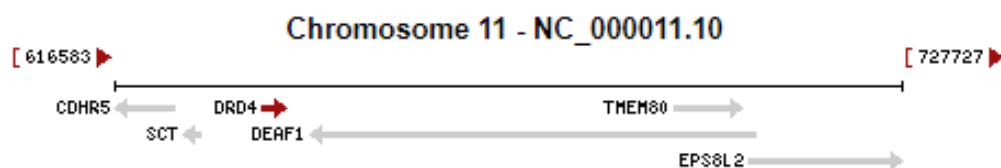
	Gene	Polymorphism	Total number of studies included in meta analyses	OR	Risk allele
<i>DRD4</i>	Dopamine receptor D4	48-bp VNTR	5	1.73	2-repeat
<i>HTR1A</i>	5-hydroxytryptamine receptor 1A	rs6295 (C1019G)	9	0.82	G- allele
<i>MAOA</i>	Monoamine oxidase A	VNTR polymorphism in promotor region	9	1.23	L- allele
<i>PCLO</i>	Piccolo presynaptic cytomatrix protein	rs2522833	3	NR	C- allele
<i>SLC6A3</i> or <i>DAT1</i>	Solute carrier family 6 member 3	40-bp VNTR	3	2.06	9/10 - genotype
<i>SLC6A4</i> or <i>5-HTT</i>	Solute carrier family 6 member 4	44-bp Ins/Del (5-HTTLPR)	24	1.11- 1.23 [*]	S- allele
<i>TPH2</i>	Tryptophan hydroxylase 2	rs4570625	6	0.77- 0.83	G- allele
		rs11178997	4	0.75	T- allele
		rs17110747	5	0.79 - 0.84	G- allele

Image 9 - Associations shown in meta-analyses between DDs and polymorphic variants of the genes linked to the exchange of monoamine neuromediators

Gene expression profiles:

DRD4 - dopamine receptor D4³³.

This gene encodes the D4 subtype of the dopamine receptor. The D4 subtype is a G-protein coupled receptor which inhibits adenylyl cyclase. It is a target for drugs which treat schizophrenia and Parkinson disease. Mutations in this gene have been associated with various behavioral phenotypes, including autonomic nervous system dysfunction, attention deficit/hyperactivity disorder, and the personality trait of novelty seeking. This gene contains a polymorphic number (2-10 copies) of tandem 48 nt repeats; the sequence shown contains four repeats.

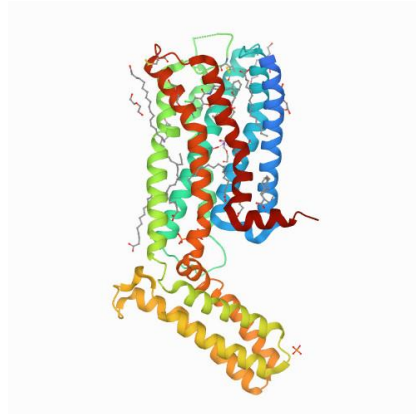


³² <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6065213/>

³³ <https://www.ncbi.nlm.nih.gov/gene/1815>

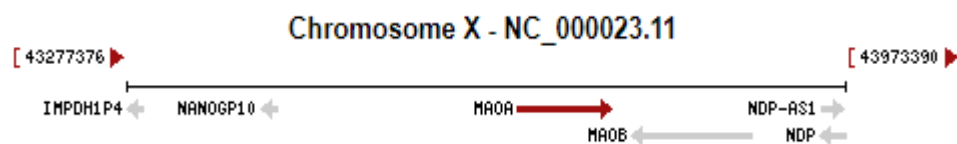
5WIV³⁴.

Structure of the sodium-bound human D4 Dopamine receptor in complex with Nemonapride.



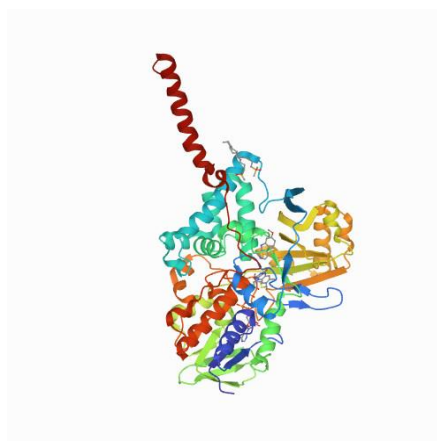
MAOA - monoamine oxidase A³⁵.

This gene is one of two neighboring gene family members that encode mitochondrial enzymes which catalyze the oxidative deamination of amines, such as dopamine, norepinephrine, and serotonin. Mutation of this gene results in Brunner syndrome. This gene has also been associated with a variety of other psychiatric disorders, including antisocial behavior. Alternatively spliced transcript variants encoding multiple isoforms have been observed.



2Z5Y³⁶.

Crystal Structure of Human Monoamine Oxidase A (G110A) with Harmine.



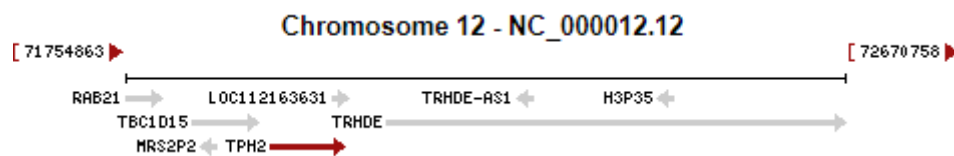
³⁴ <https://www.rcsb.org/structure/5WIV>

³⁵ <https://www.ncbi.nlm.nih.gov/gene/4128>

³⁶ <https://www.rcsb.org/structure/2Z5Y>

TPH - tryptophan hydroxylase 2³⁷.

This gene encodes a member of the pterin-dependent aromatic acid hydroxylase family. The encoded protein catalyzes the first and rate limiting step in the biosynthesis of serotonin, an important hormone and neurotransmitter. Mutations in this gene may be associated with psychiatric diseases such as bipolar affective disorder and major depression.



4V06³⁸.

Crystal structure of human tryptophan hydroxylase 2 (TPH2), catalytic domain.



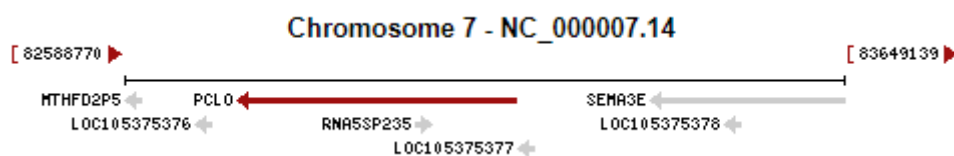
PCLO - piccolo presynaptic cytomatrix protein³⁹.

The protein encoded by this gene is part of the presynaptic cytoskeletal matrix, which is involved in establishing active synaptic zones and in synaptic vesicle trafficking. Variations in this gene have been associated with bipolar disorder and major depressive disorder. Two transcript variants encoding different isoforms have been found for this gene.

³⁷ <https://www.ncbi.nlm.nih.gov/gene/121278>

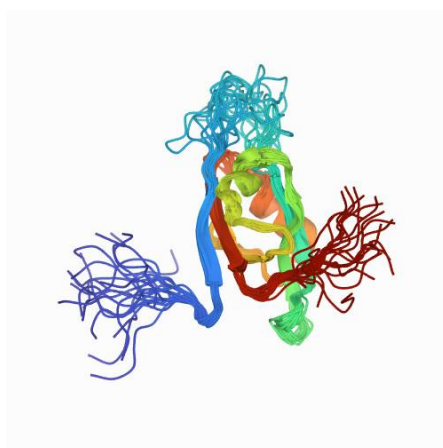
³⁸ <https://www.rcsb.org/structure/4V06>

³⁹ <https://www.ncbi.nlm.nih.gov/gene/27445>



1UJD⁴⁰.

Solution Structure of RSGI RUH-003, a PDZ domain of hypothetical KIAA0559 protein from human cDNA.



Gene	Name	Location	Function	Associations
DRD4	dopamine receptor D4	11p15.5	This gene encodes the D4 subtype of the dopamine receptor.	Depression, autonomic nervous system dysfunction, attention deficit/hyperactivity disorder.
MAOA	monoamine oxidase A	Xp11.3	This gene is one of two neighboring gene family members that encode mitochondrial enzymes which catalyze the oxidative deamination of amines, such as dopamine, norepinephrine, and serotonin.	Depression, psychiatric disorders, antisocial behavior.
TPH2	tryptophan hydroxylase 2	12q21.1	This gene encodes a member of the pterin-dependent aromatic acid hydroxylase family.	Bipolar affective disorder and major depression
PCLO	piccolo presynaptic cytomatrix protein	7q21.11	The protein encoded by this gene is part of the presynaptic cytoskeletal	Bipolar disorder and major

⁴⁰ <https://www.rcsb.org/structure/1UJD>

			matrix, which is involved in establishing active synaptic zones and in synaptic vesicle trafficking.	depressive disorder.
HTR1A	5-hydroxytryptamine receptor 1A	5q12.3	This gene encodes a G protein-coupled receptor for 5-hydroxytryptamine (serotonin), and belongs to the 5-hydroxytryptamine receptor subfamily.	Depression.
SLC6A3	solute carrier family 6 member 3	5p15.33	This gene encodes a dopamine transporter which is a member of the sodium- and chloride-dependent neurotransmitter transporter family.	idiopathic epilepsy, attention-deficit hyperactivity disorder, dependence on alcohol and cocaine, parkinson disease
SLC6A4	solute carrier family 6 member 4	17q11.2	This gene encodes an integral membrane protein that transports the neurotransmitter serotonin from synaptic spaces into presynaptic neurons.	Depression.

Table 2 – Genes associated with depression and their functions

Mitochondrial dysfunctions in major depression.

Study: Molecular correlates of mitochondrial dysfunctions in major depression: Evidence from clinical and rodent studies⁴¹.

Major depressive disorder (MDD) is one of the most prevalent stress-related mental disorders worldwide. Several biological mechanisms underlying the pathophysiology of MDD have been proposed, including endocrine disturbances, neurotransmitter deficits, impaired neuronal plasticity, and more recently, mitochondrial dysfunctions. In this review, we provide an overview of relevant molecular correlates of mitochondrial dysfunction in MDD, based on findings from clinical studies and stress-induced rodent models. We also compare differences and similarities between the phenotypes of MDD patients and animal models.

Several biological mechanisms underlying the pathophysiology of Major Depressive Disorder have been proposed, including:

- Endocrine disturbances,
- Neurotransmitter deficits,
- Impaired neuronal plasticity, and more recently,
- Mitochondrial dysfunctions.

⁴¹ <https://pubmed.ncbi.nlm.nih.gov/32979495/>

Diagnosis of Depression.

The diagnostic criteria for major depressive disorder (Image 9) are provided in the American Psychiatric Association's Diagnostic and Statistical Manual (DSM)⁴² (American Psychiatric Association, 2013)⁴³. These criteria are generally preferred over the very similar criteria for moderate depressive disorder adopted by the 10th edition of the World Health Organization's International Classification of Diseases (ICD)⁴⁴ (World Health Organization, 1993)⁴⁵. These two diagnostic systems are highly overlapping, and both require that the symptoms must be present most of the time for a 2-week period and are not better accounted for by another condition (schizophrenia, bipolar and schizoaffective disorder).

DSM-5 Major Depressive Disorder	ICD-10 Moderate Depressive Episode
Five or more symptoms, at least one of which must come from the "A" criteria:	Six or more symptoms, including two from the following:
"A" criteria	1. Depressed mood
1. Depressed mood	2. Loss of interest and enjoyment
2. Markedly diminished interest or pleasure in almost all activities	3. Reduced energy leading to increased fatigability and diminished activity
"B" criteria	Three or more typical symptoms from the following:
1. Significant weight loss/gain or decrease/increase in appetite	1. Reduced concentration and attention
2. Insomnia or excessive sleep	2. Reduced self-esteem and self-confidence
3. Psychomotor agitation or retardation	3. Ideas of guilt and unworthiness (even in mild type of episode)
4. Fatigue or loss of energy	4. Bleak and pessimistic views of the future
5. Feelings of worthlessness or excessive/inappropriate guilt	5. Ideas or acts of self-harm or suicide
6. Diminished concentration or indecisiveness	6. Disturbed sleep
7. Recurrent thoughts of death, suicidal ideation, plans or an attempt	7. Diminished appetite

Image 9 - Diagnostic Criteria for Major Depressive Disorder and Depressive Episode⁴⁶

Methods and Findings until now.

At present, the criteria for Major Depressive Disorder (MDD) diagnosis and treatment depend on different various signs and symptoms that do not always fit into specific diagnostic categories, such as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) as mentioned earlier. The entirety of the potential causes and symptoms for a set of past experiences have to be examined, including individual private medical information and

⁴² <https://www.psychiatry.org/psychiatrists/practice/dsm>

⁴³ <https://www.psychiatry.org/>

⁴⁴ https://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf

⁴⁵ <https://www.who.int/>

⁴⁶ <https://www.sciencedirect.com/science/article/pii/S0896627319302855>

confidential material, which increases stigma and makes diagnosis even more difficult (as in the case of schizophrenia).

Association of schizophrenia with depression.

Study: Depression and Schizophrenia: Cause, Consequence, or Trans-diagnostic Issue?⁴⁷

Historically, there have been some ideas and thoughts that the presence of mood symptoms in schizophrenia may be a good prognostic indicator, with patients who have high levels of affective symptoms appearing more on the “bipolar” rather than deficit/autistic end of a psychosis continuum model. However on the contrary, evidence suggests that depression is linked to poorer outcomes in schizophrenia. As example, depression is the most significant factor in completed suicide in schizophrenia, more so than acting on command hallucinations. Depression also has long term consequences for functional recovery and quality of life. Conley et al report that those with schizophrenia and depression were significantly more likely to relapse, to be a safety concern (violent, arrested, victimized, suicidal), have greater substance-related problems and report poorer life satisfaction, mental functioning, family relationships, and medication adherence. In UHR samples, reporting of psychotic-like experiences in the presence of moderate depression also raises risk for suicidal ideation. Kelleher et al show in a large community sample that attenuated psychotic experiences were relatively common among young people who had a diagnosis of moderate depressive disorder, and that the combination of experiences in this sample was significantly associated with suicidal behavior: ie, patients did not need to present with severe depression or have formal psychotic symptoms to be at heightened risk. In addition, depression has been linked to increased risk of transition from UHR to first-episode psychosis (FEP), suggesting that in this group depression also indicates a poorer outcome. However, the relationships are not straightforward nor is there sufficient evidence to suggest direct causality. In longitudinal studies depression and positive symptoms may co-occur but not necessarily predict each other over time.

Bipolar Disorder.

Bipolar disorder⁴⁸ (formerly called manic-depressive illness or manic depression) is a mental disorder that causes unordinary variations in mood, energy, activity levels, concentration, and the ability to carry out everyday tasks.

Types of Bipolar Disorders.

There are three main types of bipolar disorder. All these types of bipolar disorders include clear changes in mood, energy, and activity levels.

⁴⁷ <https://academic.oup.com/schizophreniabulletin/article/43/2/240/2503507>

⁴⁸ https://en.wikipedia.org/wiki/Bipolar_disorder#Diagnosis
<https://www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml>

Bipolar Disorder Type I.

Bipolar Disorder Type I is defined by manic episodes that last at least 7 days, or by manic symptoms that are so severe that the individual needs immediate hospitalization. Normally, depressive episodes occur as well, ordinarily lasting at least fourteen days. Episodes of depression with blended features (having depressive symptoms and manic symptoms at the same time) are also conceivable.

Bipolar Disorder Type II.

Bipolar Disorder Type II is defined by a pattern of depressive episodes and hypomanic episodes, but not the all out manic episodes that are ordinary of Bipolar I Disorder.

Cyclothymic Disorder.

Cyclothymic Disorder (likewise called Cyclothymia) is characterized by periods of hypomanic symptoms as well as periods of depressive symptoms lasting for at least 2 years (or 1 year in children and adolescents). However, the symptoms do not meet the diagnostic requirements for a hypomanic episode and a depressive episode.

Newly released genetic results have demonstrated that the two major, traditionally distinct types of psychosis – schizophrenia and bipolar disorder – may share causative factors in common - as well as and symptomatology. Understanding the genetic factors contributing to the shared and disorder-specific symptoms will be crucial for improving on the one hand diagnosis and on the other hand treatment.

Diagnosis of Bipolar Disorder.

Bipolar disorder is commonly diagnosed during adolescence or early adulthood, but onset can occur at any point of an individual's life. Its diagnosis is based on the self-reported experiences, events and considerations of the individual, abnormal behavior reported by family environment, friends or colleagues, observable signs of the disease as assessed by a clinician, and ideally a medical diagnosis. The most widely used criteria for diagnosing bipolar disorder are from the American Psychiatric Association's (APA) Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) and the World Health Organization's (WHO) International Statistical Classification of Diseases and Related Health Problems, 10th Edition (ICD-10).

It should be mentioned that bipolar disorder is difficult to diagnose and is often mistaken for other types of psychosis. Unfortunately, this misdiagnosis creates a great amount of negative outcomes. Patients will probably receive inadequate or inappropriate treatment that will not diminish the symptoms or impairment of the disorder and may even further destabilize their mood and activity levels. These individuals are under these circumstances at risk for experiencing numerous social and professional impairments, alcohol or drugs abuse, and suicidal behavior. An accurate diagnosis and appropriate treatment of bipolar disorder are necessary to prevent this chain of potentially disastrous events. For this reason, it is particularly

important to analyze the genes that may be associated with the disorder and their correlation with the disease.

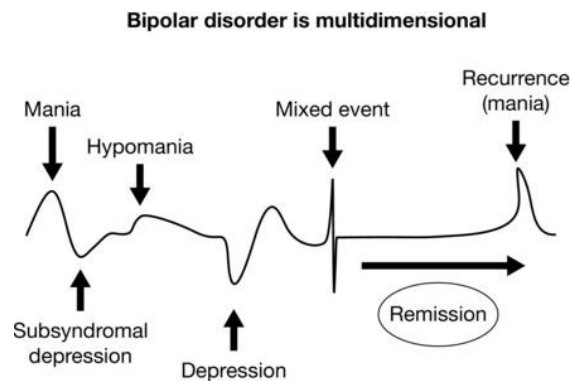
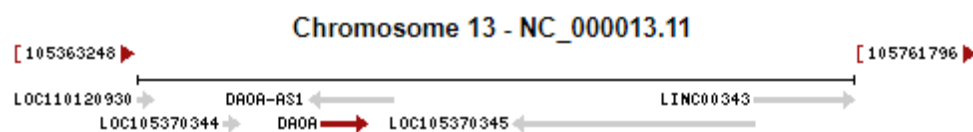


Image 10 - Diagnosis and treatment of patients with bipolar disorder⁴⁹

Gene expression profiles:

DAOA - D-amino acid oxidase activator⁵⁰.

This gene encodes a protein that may function as an activator of D-amino acid oxidase, which degrades the gliotransmitter D-serine, a potent activator of N-methyl-D-aspartate (NMDA) type glutamate receptors. Studies also suggest that one encoded isoform may play a role in mitochondrial function and dendritic arborization. Polymorphisms in this gene have been implicated in susceptibility to schizophrenia and bipolar affective disorder. Alternatively spliced transcript variants encoding different isoforms have been identified.



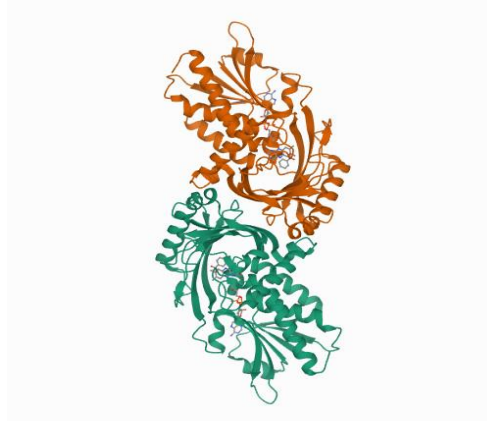
2DU8⁵¹.

Crystal structure of human D-amino acid oxidase.

⁴⁹ <https://pubmed.ncbi.nlm.nih.gov/26172568/>

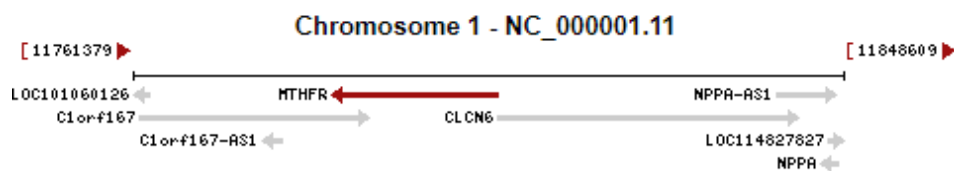
⁵⁰ <https://www.ncbi.nlm.nih.gov/gene/267012>

⁵¹ <https://www.rcsb.org/structure/2DU8>



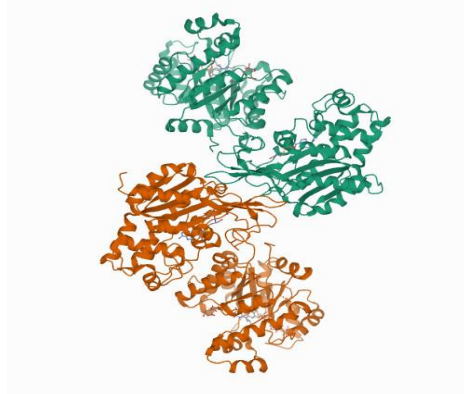
MTHFR - methylenetetrahydrofolate reductase⁵².

The protein encoded by this gene catalyzes the conversion of 5,10- methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine remethylation to methionine. Genetic variation in this gene influences susceptibility to occlusive vascular disease, neural tube defects, colon cancer and acute leukemia, and mutations in this gene are associated with methylenetetrahydrofolate reductase deficiency.



6FCX⁵³.

Structure of human 5,10-methylenetetrahydrofolate reductase (MTHFR).



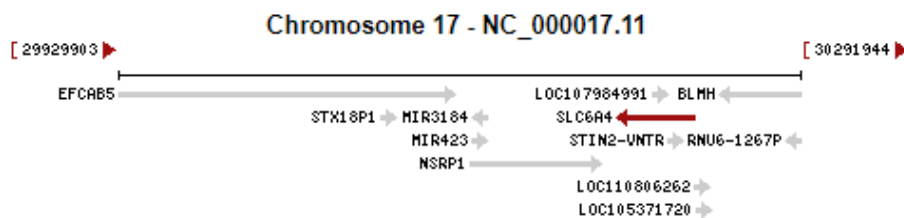
SLC6A4 - solute carrier family 6 member 4⁵⁴.

⁵² <https://www.ncbi.nlm.nih.gov/gene/4524>

⁵³ <https://www.rcsb.org/structure/6FCX>

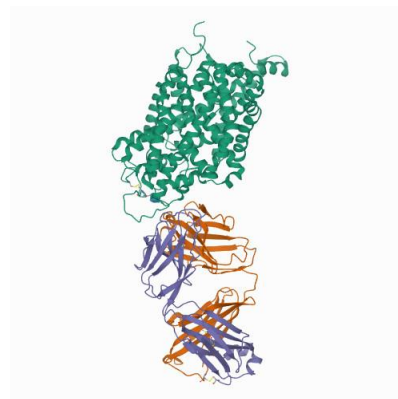
⁵⁴ <https://www.ncbi.nlm.nih.gov/gene/6532>

This gene encodes an integral membrane protein that transports the neurotransmitter serotonin from synaptic spaces into presynaptic neurons. The encoded protein terminates the action of serotonin and recycles it in a sodium-dependent manner. This protein is a target of psychomotor stimulants, such as amphetamines and cocaine, and is a member of the sodium:neurotransmitter symporter family. A repeat length polymorphism in the promoter of this gene has been shown to affect the rate of serotonin uptake. There have been conflicting results in the literature about the possible effect, if any, that this polymorphism may play in behavior and depression.



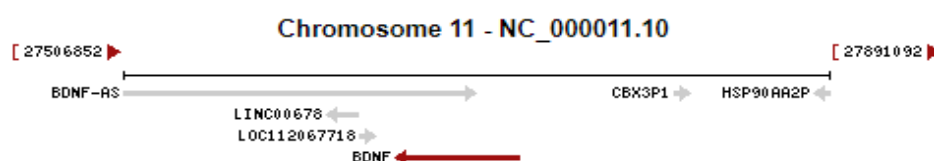
5I6Z⁵⁵.

X-ray structure of the ts2 human serotonin transporter.



BDNF - brain derived neurotrophic factor⁵⁶.

This gene encodes a member of the nerve growth factor family of proteins. Alternative splicing results in multiple transcript variants, at least one of which encodes a preproprotein that is proteolytically processed to generate the mature protein. Binding of this protein to its cognate receptor promotes neuronal survival in the adult brain. Expression of this gene is reduced in Alzheimer's, Parkinson's, and Huntington's disease patients. This gene may play a role in the regulation of the stress response and in the biology of mood disorders.

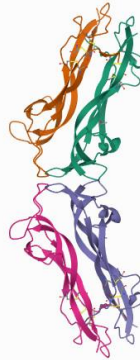


⁵⁵ <https://www.rcsb.org/structure/5I6Z>

⁵⁶ <https://www.ncbi.nlm.nih.gov/gene/627>

1BND⁵⁷.

Structure of the brain-derived neurotrophic factor/neurotrophin 3 heterodimer.



Association of schizophrenia with bipolar disorder.

Bipolar disorder (BD) and schizophrenia (SCZ) are severe psychiatric disorders and among the leading causes of disability worldwide (Whiteford et al., 2013). Both disorders have significant genetic components with heritability estimates ranging from 60% to 80% (Nöthen et al., 2010). Recent genetic and epidemiological studies have demonstrated substantial overlap between these two disorders, with a genetic correlation from common variation near 0.6–0.7 (Lee et al., 2013) and high relative risks (RR) among relatives of both BD and SCZ patients (RRs for parent/offspring: BD/BD: 6.4, BD/SCZ: 2.4; SCZ/BD: 5.2, SCZ/SCZ: 9.9) (Lichtenstein et al., 2009). Despite shared genetics and symptomatology, the current diagnostic systems (DSM Library, World Health Organization) adhere to historical distinctions from the late 19th century and represent BD and SCZ as independent categorical entities differentiated on the basis of their clinical presentation, with BD characterized by predominant mood symptoms, mood-congruent delusions, and an episodic disease course and SCZ considered a prototypical psychotic disorder. Identifying genetic components contributing to both disorders will provide insight into the biology underlying the shared symptoms of the disorders.

While the shared genetic component is substantial, studies to date have also implicated genetic architecture differences between these two disorders (Curtis et al., 2011, Ruderfer et al., 2014). A polygenic risk score (PRS) created from a case-only SCZ versus BD genome-wide association study (GWAS) significantly correlated with SCZ or BD diagnosis in an independent sample (Ruderfer et al., 2014), providing the first evidence that differences between the disorders also have a genetic basis. An enrichment of rare, moderate to highly penetrant copy-number variants (CNVs) and de novo CNVs are seen in SCZ patients (Gulsuner and McClellan, 2015, Kirov et al., 2012, Marshall et al., 2017, Stone et al., 2008, Szatkiewicz et al., 2014), while the involvement of CNVs in BD is less clear (Green et al., 2016). Although the role of de novo single nucleotide variants in BD and SCZ has been investigated in only a handful of studies, enrichment in pathways associated with the

⁵⁷ <https://www.rcsb.org/structure/1BND>

postsynaptic density has been reported for SCZ but not BD (Fromer et al., 2014, Kataoka et al., 2016). Identifying disorder-specific variants and quantifying the contribution of genetic variation to specific symptom dimensions remain important open questions. Characterizing these genetic differences will facilitate an understanding of the dimensions of the disorders instead of the dichotomous diagnosis. For example, we have shown that SCZ patients with greater manic symptoms have higher polygenic risk for BD (Ruderfer et al., 2014). These findings demonstrate shared genetic underpinnings for symptoms across disorders and may enable us to characterize patients by genetic liability to symptom dimensions, thereby informing disease course and treatment.

Gene	Name	Location	Function	Associations
MTHFR	methylenetetrahydrofolate reductase	1p36.22	The protein encoded by this gene catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine remethylation to methionine.	Bipolar disorder (BPD).
NRG1	neuregulin 1	8p12	The protein encoded by this gene is a membrane glycoprotein that mediates cell-cell signaling.	Cancer, schizophrenia, and bipolar disorder (BPD).
DAOA	D-amino acid oxidase	12q24.11	This gene encodes the peroxisomal enzyme D-amino acid oxidase.	Schizophrenia, bipolar disorder (BPD).
COMT	catechol-O-methyltransferase	22q11.21	Catechol-O-methyltransferase catalyzes the transfer of a methyl group from S-adenosylmethionine to catecholamines, including the neurotransmitters dopamine, epinephrine, and norepinephrine.	Schizophrenia and bipolar disorder (BPD).
SLC6A4	solute carrier family 6 member 4	17q11.2	This gene encodes an integral membrane protein that transports the neurotransmitter serotonin from synaptic spaces into presynaptic neurons.	Bipolar disorder (BPD), behavior and depression.
BDNF	brain derived neurotrophic factor	11p14.1	This gene encodes a member of the nerve growth factor family of proteins.	Bipolar disorder (BPD), mood disorders.

Table 3 – Genes associated with bipolar disorder and their functions

Description of Tool / Method / Technical Solution.

For the purposes of this work, the Python Programming Language and GenBank - the NIH genetic sequence database were used.

Programming Language: Python

Programming Language Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented methodology intend to assist programmers and software engineers write clear, logical code for small and large-scale projects. Python is commonly used in machine learning projects and it was used for the analysis of our dataset.

GenBank.

GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis. The GenBank database is designed to provide and encourage access to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no limitations on the utilization, dissemination or distribution of the GenBank data. For the purposes of the presented work, I conducted research for the genes associated with schizophrenia, schizoaffective disorder, bipolar disorder & depression.

Dataset: Gene expression of schizophrenia and bipolar.

The dataset contained more than 546 features (variables), including the characteristics organism, individual, sex, age, organism part, cell type, disease and genetic data. All these characteristics are important features for the diagnosis of schizophrenia and bipolar disorder. The classes of the gene expression of schizophrenia and bipolar dataset:

Class 0 (bipolar disorder): 84 observations
Class 1 (normal): 391 observations
Class 2 (schizoaffective disorder): 13 observations
Class 3 (schizophrenia): 58 observations

Methods.

In this work machine learning algorithms⁵⁸ were used to build models based on sample data, known as "training data", in order to make predictions or decisions without being explicitly

⁵⁸ https://en.wikipedia.org/wiki/Machine_learning

programmed to do so. Machine learning algorithms are generally used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Machine learning approaches are traditionally divided into three main categories, depending on the nature of the "signal" or "feedback" available to the learning system:

- Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
- Reinforcement learning: A program interacts with a dynamic environment in which it must perform a certain goal.

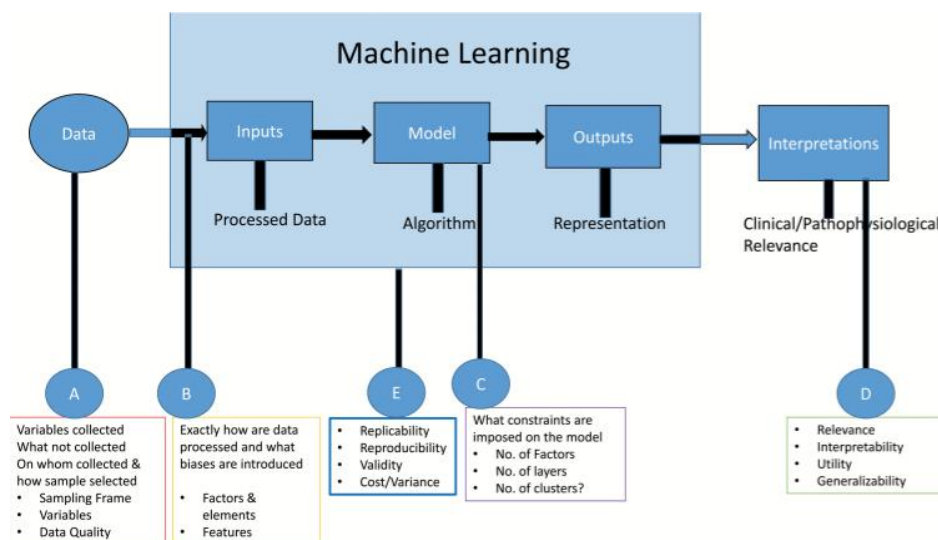


Image 11 - Machine Learning⁵⁹

The Classifiers used to develop and evaluate a predictive model were:

- The Naïve Bayes (GaussianNB) Classifier;
- The Decision Trees Classifier;
- The Random Forest Classifier;

Naïve Bayes.

Naïve Bayes Algorithm⁶⁰ is an effective algorithm for the classification task. Even if there is a need for working on a data set with a large amount of data, it is suggested to try Naïve Bayes

⁵⁹ Will Machine Learning Enable Us to Finally Cut the Gordian Knot of Schizophrenia, <https://pubmed.ncbi.nlm.nih.gov/29986110/>

⁶⁰ https://en.wikipedia.org/wiki/Naive_Bayes_classifier

approach. To understand the Naïve Bayes Algorithm we need to understand the Bayes theorem. Bayes theorem works on conditional probability. Conditional probability describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Using the conditional probability, we have the ability to calculate the probability of an event using its prior knowledge.

Below is the formula for calculating the conditional probability, where:

$$P(H | E) = \frac{P(E | H) * P(H)}{P(E)}$$

- $P(H)$ is the probability of hypothesis H being true. This is known as the prior probability.
- $P(E)$ is the probability of the evidence (regardless of the hypothesis).
- $P(E|H)$ is the probability of the evidence given that hypothesis is true.
- $P(H|E)$ is the probability of the hypothesis given that the evidence is there.

Naïve Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).

The MAP for a hypothesis is:

$$\begin{aligned} \text{MAP}(H) &= \max(P(H|E)) \\ &= \max((P(E|H)*P(H))/P(E)) \\ &= \max(P(E|H)*P(H)) \end{aligned}$$

$P(E)$ is evidence probability, and it is used to normalize the result.

Naïve Bayes Algorithm assumes that all the features are unrelated to each other. Presence or absence of a feature does not influence the presence or absence of any other feature.

Decision Tree.

Decision Tree⁶¹ algorithm belongs to the group of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

Important Terminology related to Decision Trees.

⁶¹ https://en.wikipedia.org/wiki/Decision_tree_learning

- **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
- **Leaf / Terminal Node:** Nodes that do not split are called Leaf or Terminal nodes.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning.
- **Branch / Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.
- **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

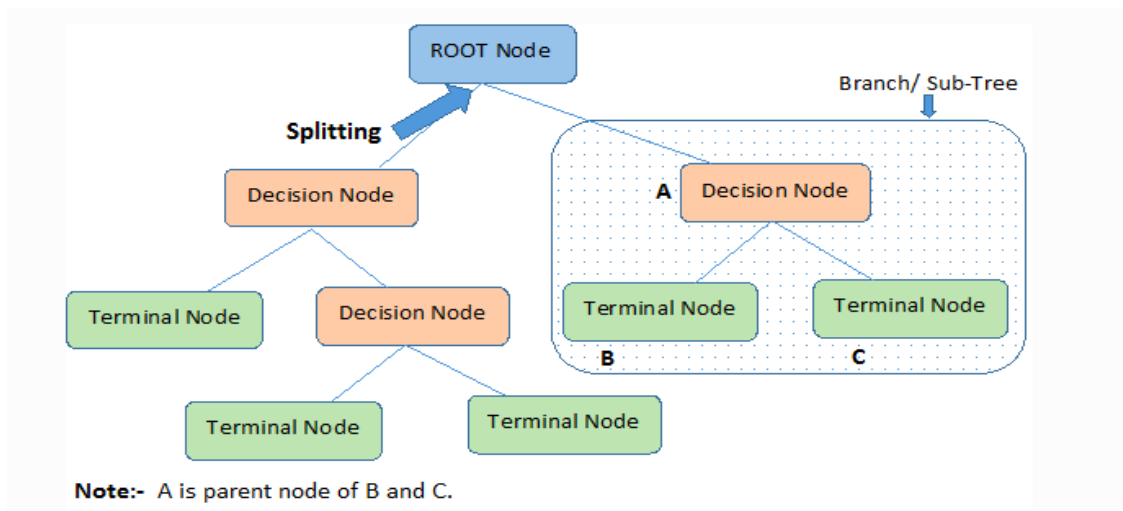


Image 12 - Decision Tree (Source Wikipedia)

The algorithm selection is also based on the type of target variables. Some algorithms used in Decision Trees:

ID3 → (extension of D3)

C4.5 → (successor of ID3)

CART → (Classification And Regression Tree)

CHAID → (Chi-square automatic interaction detection Performs multi-level splits when computing classification trees)

MARS → (multivariate adaptive regression splines).

Random Forest.

Random Forest or Random Decision Forests⁶² is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The

⁶² https://en.wikipedia.org/wiki/Random_forest

general idea of the bagging method is that a combination of learning models increases the overall performance.

One advantage of the Random Forest algorithm is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. In our case we need to examine a Random Forest in classification.

Below is presented a Random Forest with two trees:

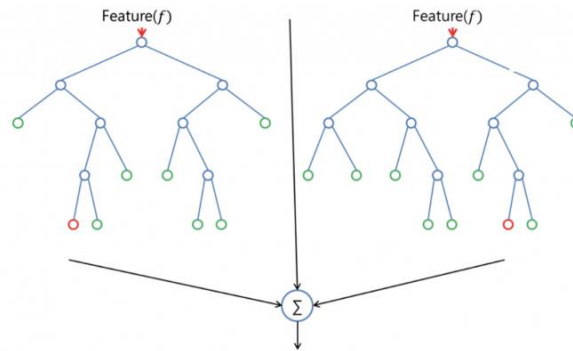


Image 13 - Random Trees (Source Wikipedia)

The training algorithm for Random Forest applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \dots, x_n$ with responses $Y = y_1, \dots, y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, \dots, B$:

- Sample, with replacement, n training examples from X, Y ; call these X_b, Y_b .
- Train a classification or regression tree f_b on X_b, Y_b .

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x' :

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

or by taking the majority vote in the case of classification trees.

Application of Tool / Method / Technical Solution.

For data analysis (using Python) in Schizophrenia, bipolar disorder, and schizoaffective disorder - as already mentioned, machine learning techniques were used while the applying classifiers were: the Naïve Bayes (GaussianNB) Classifier, the Decision Trees Classifier, and the Random Forest Classifier.

In the first step, NumPy (an open-source numerical Python library) and Pandas (Python-based data analysis toolkit) were imported I used the Pandas in order to provide us with some sets of powerful tools like DataFrame and Series to analyze our data, and NumPy module to provide us objects for multi-dimensional arrays. In the same context I imported the sklearn library objects, and I selected repeated k-fold cross-validation to provide us an improved estimated performance of our machine learning models.

```
import numpy as np
import pandas as pd
from numpy import mean
from numpy import std

from sklearn.impute import SimpleImputer
from sklearn.feature_selection import SelectKBest, mutual_info_regression
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, recall_score
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler

pd.options.mode.chained_assignment = None
```

Next, when I tried to load the data, I noticed an issue in the line 1066 (some weird codes instead of values - especially after the column 547), So I kept only the necessary columns. Furthermore, the gene data where stored as a gene per line which was not very useful so I transformed the matrix resulted in a table where each person was stored per line. The people data were easier to handle. So I joined (thus the index int transformation) the gene data to the corresponding person.

```
# Load the gene data
genes = pd.read_table("Adjusted_expression_values.txt", index_col=0, usecols=list(range(0,547))).transpose()
genes.index = genes.index.astype('int64')

# Load the people data
people = pd.read_table("E-MTAB-8018.sdrf.txt", index_col=0)
people.index = people.index.astype('int64')

# Join both datasets into one
data = people.join(genes, how='inner')
```

Then, during the transformation process, I transformed the disease column to a Categorical type since it's our target output. I investigated if any of the person's features can be used on our models. I extracted the Age and the Sex characteristics which seemed reasonable to be added. After, I check the data that had some missing values. Because the available data were

not that large, I could not just delete the observations that had missing values. So, I decided to fill the missing values with the corresponding mean value.

In the next step, I transformed the data into numpy arrays which is the desired form of the sklearn models.

```
# Transform our output to categorical
data['Characteristics[disease]'] = pd.Categorical(data['Characteristics[disease]'])

# Extract, patch and rename the age and sex columns of each person to the data
col_in_model = [2, 3] + list(range(28,47311))
X = data.iloc[:, col_in_model]
X['BinarySex'] = X['Characteristics[sex]'].apply(lambda x: 0 if x == 'female' else 1)
X['Age'] = X['Characteristics[age]']
X.drop(['Characteristics[sex]', 'Characteristics[age]'], axis=1, inplace=True)

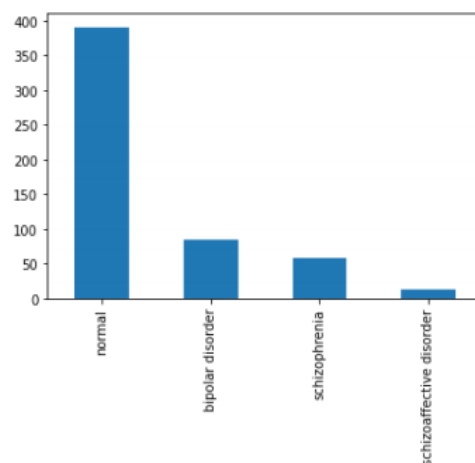
# Fill the missing values with the corresponding mean value
impute_mean = SimpleImputer(missing_values=np.nan, strategy='mean')
impute_mean.fit(X)
X = pd.DataFrame(impute_mean.transform(X))
y = np.array(data['Characteristics[disease]'].cat.codes)
```

Regarding the Features, I observed that a large amount of them created an unwanted "noise" in our models and that had as a result a low general performance.

So, on this step, I tried to rank all the features. I used the (sklearn.feature_selection) because our features are continuous and can be negative as well. I Kept the top 100.

```
ranked_features_100 = SelectKBest(mutual_info_regression, k=100).fit_transform(X, y)
```

After investigating the dataset, it is obvious that there is a significant imbalance between classes.



In order to solve this issue, I used some standard techniques shown below (Over and Under Sampling).

For the minority classes I used Oversampling technique with K-NN (3) and brought them up to 250 observations per class and I used Undersampling for the majority class where I brought it down to 250, in order to have a better balanced dataset.

It was necessary to write the validation myself (due to the sampling) in order to avoid any overfitting of the model.

Class 0 (bipolar disorder): 84 observations

Class 1 (normal): 391 observations

Class 2 (schizoaffective disorder): 13 observations

Class 3 (schizophrenia): 58 observations

```
data['Characteristics[disease]'].value_counts().plot.bar()
```

```
def evaluate_model(model, X_train, y_train, sampling=False, random_state=5):
    k = 10
    over_strategy = {0:250, 2:250, 3:250}
    oversample = SMOTE(sampling_strategy=over_strategy, k_neighbors=2, random_state=random_state)
    under_strategy = {1:250}
    undersample = RandomUnderSampler(sampling_strategy=under_strategy, random_state=random_state)
    cv = RepeatedStratifiedKFold(n_splits=k, n_repeats=1, random_state=random_state)
    rec_sum = 0

    for train_index, test_index in cv.split(X_train, y_train):
        X_train_fold, y_train_fold = X_train[train_index], y_train[train_index]
        X_test_fold, y_test_fold = X_train[test_index], y_train[test_index]

        if sampling:
            X_over, y_over = oversample.fit_resample(X_train_fold, y_train_fold)
            X_sampled, y_sampled = undersample.fit_resample(X_over, y_over)
            model.fit(X_sampled, y_sampled)
        else:
            model.fit(X_train_fold, y_train_fold)

        rec_sum += recall_score(y_test_fold, model.predict(X_test_fold), average='micro')
    #     print(confusion_matrix(y_test_fold, model.predict(X_test_fold)))
    #     print(f"Recall: {recall_score(y_test_fold, model.predict(X_test_fold), average='micro')}")

    return rec_sum/k
```

Then I had to establish a baseline model in order to keep improving the data. Firstly, I decided to use the simplest classifier for multiclass problems - so I chose the Naïve Bayes (GaussianNB). Then I experimented with both Decision Trees and Random Forest.

Multi-Class Classification.

In the case of our dataset, we have a multi-class classification problem (4-class classification problem). More specifically, we have a dataset that has four class labels, Class 0 (bipolar), Class 1 (normal), Class 2 (schizoaffective disorder), and Class 3 (schizophrenia). Therefore, we had to find **TP**, **TN**, **FP** and **FN** for each individual class.

1. **True positive**: correctly classified or detected.
2. **False positive**: incorrectly classified or detected. It represents the type I error.
3. **False negative**: incorrectly rejected. It represents the type II error.
4. **True negative**: correctly rejected.

Since we have all the necessary metrics for every class from the confusion matrix, we need to calculate the performance measures for each class.

Confusion Matrix for each model:

Naive Bayes (No Sampling)

Run 0:

True Class / Predicted Class

[Class 0 [8 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]

Precision: 0.775569131419

Recall: 0.8575569545481

Run 1:

True Class / Predicted Class

[Class 0 [7 1 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]]

Precision: 0.6792414258225

Recall: 0.83745569545489

Run 2:

True Class / Predicted Class

[Class 0 [8 1 0 0]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]]

Precision: 0.6841569121345

Recall: 0.84215519546413

Run 3:

True Class / Predicted Class

[Class 0 [8 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.68013569141465

Recall: 0.84101512195315

Run 4:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 40 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]
Precision: 0.67951571985413
Recall: 0.8397569859531

Run 5:
True Class / Predicted Class
[Class 0 [4 0 0 4]
Class 1 [0 39 0 0]
Class 2 [1 0 0 1]
Class 3 [6 0 0 0]]
Precision: 0.62114569545489
Recall: 0.82452569545489

Run 6:
True Class / Predicted Class
[Class 0 [4 0 0 4]
Class 1 [0 39 0 0]
Class 2 [2 0 0 0]
Class 4 [3 0 0 2]]
Precision: 0.6240256911138103
Recall: 0.82995569121929

Run 7:
True Class / Predicted Class
[Class 0 [5 1 0 2]
Class 1 [0 39 0 0]
Class 2 [2 0 0 0]
Class 3 [5 0 0 0]]
Precision: 0.6235756923235622
Recall: 0.831311569545489

Run 8:
True Class / Predicted Class
[Class 0 [8 0 0 0]
Class 1 [0 39 0 0]
Class 2 [1 0 0 0]
Class 3 [6 0 0 0]]
Precision: 0.7841456954522476
Recall: 0.84813569161623

Run 9:
True Class / Predicted Class
[Class 0 [7 0 0 1]
Class 1 [0 39 0 0]
Class 2 [1 0 0 0]
Class 3 [3 1 0 2]]
Precision: 0.7831456924245956
Recall: 0.50813569545413

Total Precision: 0.7365956954548923
Total Recall: 0.8498316498316498

Naive Bayes (With Sampling)

Run 0:

True Class / Predicted Class

[Class 0 [6 3 0 0]:

Class 1 [0 39 0 0]

Class 2 [0 1 0 0]

Class 3 [1 5 0 0]]

Precision: 0.6035956954548945

Recall: 0.7721548821548822

Run 1:

True Class / Predicted Class

[Class 0 [3 6 0 0]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [3 3 0 0]]

Precision: 0.5935956954548912

Recall: 0.7621548821242421

Run 2:

True Class / Predicted Class

[Class 0 [3 6 0 0]

Class 1 [0 39 0 0]

Class 2 [0 1 0 0]

Class 3 [1 5 0 0]]

Precision: 0.6135956912161975

Recall: 0.7567548821548813

Run 3:

True Class / Predicted Class

[Class 0 [3 6 0 0]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [3 3 0 0]]

Precision: 0.6135956954548921

Recall: 0.7459548821445419

Run 4:

True Class / Predicted Class

[Class 0 [3 5 0 0]

Class 1 [0 40 0 0]

Class 2 [0 1 0 0]

Class 3 [0 6 0 0]]
Precision: 0.60956956910201345
Recall: 0.7391548120544052

Run 5:

True Class / Predicted Class

[Class 0 [2 6 0 0]

Class 1 [0 39 0 0]

Class 2 [0 2 0 0]

Class 3 [3 3 0 0]]

Precision: 0.5998136954548931

Recall: 0.7215548821548513

Run 6:

True Class / Predicted Class

[Class 0 [2 6 0 0]

Class 1 [0 39 0 0]

Class 2 [1 1 0 0]

Class 3 [1 3 0 1]]

Precision: 0.5915956954891374

Recall: 0.7201542221548112

Run 7:

True Class / Predicted Class

[Class 0 [4 3 0 1]

Class 1 [0 39 0 0]

Class 2 [0 2 0 0]

Class 3 [1 4 0 0]]

Precision: 0.6035456954721332

Recall: 0.72911548821140742

Run 8:

True Class / Predicted Class

[Class 0 [5 3 0 0]

Class 1 [0 39 0 0]

Class 2 [0 1 0 0]

Class 3 [1 5 0 0]]

Precision: 0.6135923154540112

Recall: 0.7210212221544513

Run 9:

True Class / Predicted Class

[Class 0 [4 4 0 0]

Class 1 [0 39 0 0]

Class 2 [0 1 0 0]

Class 3 [2 4 0 0]]

Precision: 0.6099856913142240

Recall: 0.7321548821548822

Total Precision: 0.6135956954548930
Total Recall: 0.7821548821548822

Decision Tree (No Sampling)

Run 0:

True Class / Predicted Class

[Class 0 [7 0 1 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [3 0 0 3]]

Precision: 0.6416256954548950

Recall: 0.8692336026936027

Run 1:

True Class / Predicted Class

[Class 0 [7 0 1 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [2 0 0 4]]

Precision: 0.6397293312125402

Recall: 0.8715013026659521

Run 2:

True Class / Predicted Class

[Class 0 [3 0 1 5]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 1 0]]

Precision: 0.6458971952658112

Recall: 0.8451336026936027

Run 3:

True Class / Predicted Class

[Class 0 [5 0 0 4]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [3 0 0 3]]

Precision: 0.641356920208960

Recall: 0.8601336026902011

Run 4:

True Class / Predicted Class

[Class 0 [5 0 1 2]

Class 1 [0 40 0 0]

Class 2 [0 0 0 1]

Class 3 [2 0 2 2]]
Precision: 0.6525246915156430
Recall: 0.8701016026931415

Run 5:

True Class / Predicted Class
[Class 0 [4 0 0 4]
Class 1 [0 39 0 0]
Class 2 [0 0 0 2]
Class 3 [1 0 0 5]]
Precision: 0.6697256954548921
Recall: 0.8526936026936027

Run 6:

True Class / Predicted Class
[Class 0 [7 0 1 0]
Class 1 [0 39 0 0]
Class 2 [1 0 0 1]
Class 3 [5 0 0 0]]
Precision: 0.6597259245523318
Recall: 0.8599936052523540

Run 7:

True Class / Predicted Class
[Class 0 [6 0 0 2]
Class 1 [0 39 0 0]
Class 2 [2 0 0 0]
Class 3 [2 0 1 2]]
Precision: 0.6697256954548914
Recall: 0.8601315026106012

Run 8:

True Class / Predicted Class
[Class 0 [6 0 0 2]
Class 1 [0 39 0 0]
Class 2 [1 0 0 0]
Class 3 [4 0 1 1]]
Precision: 0.6697256954548922
Recall: 0.8626936026936027

Run 9:

True Class / Predicted Class
[Class 0 [6 0 0 2]
Class 1 [0 39 0 0]
Class 2 [1 0 0 0]
Class 3 [2 0 0 4]]
Precision: 0.6690468954548901
Recall: 0.8701036024096013

Total Precision: 0.6697256954548944

Total Recall: 0.8626936026936027

Decision Tree (With Sampling)

Run 0:

True Class / Predicted Class

[Class 0 [3 0 1 5]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]]

Precision: 0.5701156940132226

Recall: 0.8102241851851411

Run 1:

True Class / Predicted Class

[Class 0 [6 0 1 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [2 0 0 4]]

Precision: 0.6035956940138914

Recall: 0.8320204851121361

Run 2:

True Class / Predicted Class

[Class 0 [5 0 1 3]

Class 1 [0 39 0 0]

Class 2 [0 0 0 1]

Class 3 [5 0 0 1]]

Precision: 0.6012350540697812

Recall: 0.8301135851851464

Run 3:

True Class / Predicted Class

[Class 0 [6 0 1 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [3 0 2 1]]

Precision: 0.610191294345616

Recall: 0.8421141851855223

Run 4:

True Class / Predicted Class

[Class 0 [5 0 0 3]

Class 1 [0 40 0 0]

Class 2 [0 0 0 1]

Class 3 [3 0 1 2]]

Precision: 0.6004350120693063
Recall: 0.8302025851834852

Run 5:

True Class / Predicted Class

[Class 0 [4 0 0 4]

Class 1 [0 39 0 0]

Class 2 [0 0 0 2]

Class 3 [4 0 0 2]]

Precision: 0.603595694013894

Recall: 0.8351851851851851

Run 6:

True Class / Predicted Class

[Class 0 [0 0 0 8]

Class 1 [0 39 0 0]

Class 2 [0 0 1 1]

Class 3 [1 0 0 4]]

Precision: 0.6002526912122914

Recall: 0.82959518514481352

Run 7:

True Class / Predicted Class

[Class 0 [3 0 0 5]

Class 1 [0 39 0 0]

Class 2 [2 0 0 0]

Class 3 [5 0 0 0]]

Precision: 0.5925956940654201

Recall: 0.8301149651851851

Run 8:

True Class / Predicted Class

[Class 0 [5 0 1 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [2 0 2 2]]

Precision: 0.6004956914608912

Recall: 0.8292923851851851

Run 9:

True Class / Predicted Class

[Class 0 [7 0 1 0]

Class 1 [0 39 0 0]

Class 2 [0 0 0 1]

Class 3 [2 0 1 3]]

Precision: 0.6035956940153098

Recall: 0.8401251346521878

Total Precision: 0.6035956940138992

Total Recall: 0.8351851851851851

Random Forest (No Sampling)

Run 0:

True Class / Predicted Class

[Class 0 [8 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.5913956954541078

Recall: 0.8453020153125092

Run 1:

True Class / Predicted Class

[Class 0 [7 0 0 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.5980956954321232

Recall: 0.8401024562235099

Run 2:

True Class / Predicted Class

[Class 0 [8 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.6001236912341352

Recall: 0.855378953535352

Run 3:

True Class / Predicted Class

[Class 0 [8 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.6001236912341352

Recall: 0.8553562353535352

Run 4:

True Class / Predicted Class

[Class 0 [8 0 0 0]

Class 1 [0 40 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Total Precision: 0.5991956954541078

Total Recall: 0.84093635312530166

Run 5:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 39 0 0]

Class 2 [2 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.60012369123452

Recall: 0.840101351234652

Run 6:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 39 0 0]

Class 2 [2 0 0 0]

Class 3 [5 0 0 0]]

Precision: 0.60012369123454

Recall: 0.840101351212852

Run 7:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 39 0 0]

Class 2 [2 0 0 0]

Class 3 [5 0 0 0]]

Precision: 0.600123691341352

Recall: 0.8412125145635352

Run 8:

True Class / Predicted Class

[Class 0 [8 0 0 0]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]]

Total Precision: 0.600414565454107

Total Recall: 0.8501015356465302

Run 9:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Total Precision: 0.599589745453328

Total Recall: 0.8501301354545352

Total Precision: 0.6013956954541078

Total Recall: 0.8553535353535352

Random Forest (With Sampling)

Run 0:

True Class / Predicted Class

[Class 0 [7 0 0 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]]

Precision: 0.702045692563444

Recall: 0.8704146936024893

Run 1:

True Class / Predicted Class

[Class 0 [8 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [4 0 0 2]]

Precision: 0.719645395454413

Recall: 0.8764056936033912

Run 2:

True Class / Predicted Class

[Class 0 [9 0 0 0]

Class 1 [0 39 0 0]

Class 2 [0 0 0 1]

Class 3 [4 0 0 2]

Precision: 0.719995693212013

Recall: 0.8801034936023422

Run 3:

True Class / Predicted Class

[Class 0 [7 0 0 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [5 0 0 1]]

Precision: 0.719545695454896

Recall: 0.8736026936026935

Run 4:

True Class / Predicted Class

[Class 0 [8 0 0 0]

Class 1 [0 40 0 0]

Class 2 [1 0 0 0]

Class 3 [4 0 0 2]]

Precision: 0.710145130134847
Recall: 0.8701326936096912

Run 5:

True Class / Predicted Class

[Class 0 [6 0 0 2]

Class 1 [0 39 0 0]

Class 2 [1 0 0 1]

Class 3 [5 0 0 1]]

Precision: 0.704413695024832

Recall: 0.8696526931456975

Run 6:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 39 0 0]

Class 2 [2 0 0 0]

Class 3 [4 0 0 1]]

Precision: 0.719545695454896

Recall: 0.8613269315693526

Run 7:

True Class / Predicted Class

[Class 0 [8 0 0 0]

Class 1 [0 39 0 0]

Class 2 [2 0 0 0]

Class 3 [5 0 0 0]]

Precision: 0.699545660664812

Recall: 0.8636026931212901

Run 8:

True Class / Predicted Class

[Class 0 [8 0 0 0]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [6 0 0 0]]

Precision: 0.709545692456289

Recall: 0.8696926932535021

Run 9:

True Class / Predicted Class

[Class 0 [7 0 0 1]

Class 1 [0 39 0 0]

Class 2 [1 0 0 0]

Class 3 [4 0 0 2]]

Precision: 0.710250293564832

Recall: 0.8626228542323454

Total Precision: 0.719545695454896

Total Recall: 0.8736026936026935

Respectively in the code:

```
X_train = np.copy(X)
y_train = np.copy(y)

rs = 224
model = GaussianNB()
res1 = evaluate_model(model, X_train, y_train, sampling=False, random_state=rs)
print("Naive Bayes (No Sampling)")
print(f"Recall: {evaluate_model(model, X_train, y_train, sampling=False, random_state=rs)}")
print("-----")

model = GaussianNB()
res2 = evaluate_model(model, X_train, y_train, sampling=True, random_state=rs)
print("Naive Bayes (With Sampling)")
print(f"Recall: {evaluate_model(model, X_train, y_train, sampling=True, random_state=rs)}")
print("-----")

model = DecisionTreeClassifier(random_state=rs)
res3 = evaluate_model(model, X_train, y_train, sampling=False, random_state=rs)
print("Decision Tree (No Sampling)")
print(f"Recall: {evaluate_model(model, X_train, y_train, sampling=False, random_state=rs)}")
print("-----")

model = DecisionTreeClassifier(random_state=rs)
res4 = evaluate_model(model, X_train, y_train, sampling=True, random_state=rs)
print("Decision Tree (With Sampling)")
print(f"Recall: {evaluate_model(model, X_train, y_train, sampling=True, random_state=rs)}")
print("-----")

model = RandomForestClassifier(random_state=rs)
print("Random Forest (No Sampling)")

print(f"Recall: {evaluate_model(model, X_train, y_train, sampling=False, random_state=rs)}")
print("-----")

model = RandomForestClassifier(random_state=rs)
print("Random Forest (With Sampling)")
print(f"Recall: {evaluate_model(model, X_train, y_train, sampling=True, random_state=rs)}")
print("-----")
```

```
Naive Bayes (No Sampling)
Recall: 0.836936026936027
-----
Naive Bayes (With Sampling)
Recall: 0.8406397306397306
-----
Decision Tree (No Sampling)
Recall: 0.8405387205387205
-----
Decision Tree (With Sampling)
Recall: 0.8591245791245792
-----
Random Forest (No Sampling)
Recall: 0.8533670033670034
-----
Random Forest (With Sampling)
Recall: 0.8680808080808081
```

Total Recalls of our Models.

It is easy to observe from the results of the Confusion Matrix that the Random Forest Classifier had the best Precision and Recall results of all three. So, I decided to use the Random Forest as the selected Classifier.

Hyperparameter optimization or Tuning⁶³.

After choosing the Random Forest as our model, I chose a set of optimal hyperparameters (Parameter Tuning)^[4] to control the learning process and get the maximum our model had to offer. For the Parameter Tuning, I used the number of estimators, the max depth of the tree and minimum samples needed for a node to split. After checking all the possible combinations, the parameters below reached the best **Recall Score (87.35)**.

Max Depth: None

Number of Estimators: 50

Minimum Sample for Node Split: 2

```
import itertools

n_estimators = [10, 50, 100, 200]
max_depth = [4, 6, 10, 12, None]
min_samples_split = [2, 5, 10]

max_recall = -1
depth_estimators_list = [list(zip(x, n_estimators)) for x in itertools.permutations(max_depth, len(n_estimators))]

for depth in max_depth:
    for estimators in n_estimators:
        for split in min_samples_split:
            model = RandomForestClassifier(random_state=rs, max_depth=depth, n_estimators=estimators, min_samples_split=split)
            recall = evaluate_model(model, X, y, sampling=True, random_state=rs)
            if recall > max_recall:
                max_recall = recall
                print(depth, estimators, split)
                print(recall)
```

4	10	2	0.8076430976430977
4	10	5	0.8076767676767677
4	50	2	0.8204713804713805
4	200	2	0.8371043771043769
4	200	10	0.8389225589225588
6	50	2	0.8461952861952862
6	50	10	0.8516498316498318

⁶³ https://en.wikipedia.org/wiki/Hyperparameter_optimization

6 200 5
0.8534343434343434
10 10 5
0.8551178451178452
10 100 5
0.857104377104377
12 50 2
0.8645117845117843
None 50 2
0.8735690235690237

Presentation of the results.

Dataset.

These dataset⁶⁴ contained more than 540 examples, including individual's characteristics (such as the age and the sex) and genetic data (such as cell type). These basic characteristics were important features for a first attempt to create a model for the diagnosis of schizophrenia, bipolar disorder and schizoaffective disorder. With an approach to analyze data by dividing them in two main categories - mentally ill and healthy individuals (binary classification), I did not have the expected results. Then, with a second approach to analyze all features, also given the large number of examples of our dataset, and after the application of the necessary techniques, an effort was made to give the best possible models Recall. Of course, due to the unequal distribution in the classes of the dataset, we did not have good results (Precision, Recall) in the classes with a limited number examples.

Our estimated results.

Development of a model for the diagnosis of schizophrenia, bipolar disorder, and schizoaffective disorder - contribution to the diagnosis of psychotic diseases using basic characteristics and genetic data (from the datasets available to us). The main objective was the maximum exploitation of our data, in order to achieve better training of our diagnosis models.

Training.

Preprocessing of the dataset was performed; this included the correction of issues in specific lines (codes instead of values). I kept only the necessary columns. Transformation of the disease column to a categorical type since it's our target output. I investigated the individual's features that was related to the diseases we study, and could be used on our models. I

⁶⁴ <https://www.kaggle.com/peterwu19881230/gene-expression-of-schizophrenia-and-bipolar>

extracted the characteristics age and the sex (as we had already mentioned). After checking the data, I found some missing values in our dataset and I filled the missing values with the corresponding mean value. Due to the small number of examples, I did not proceed with any removal. To correctly train the predictive model, I needed to separate the data into two groups: training data and test data (10-fold cross-validation).

Feature Selection and Cross-Validation.

Feature selection was performed. Feature selection was applied on the training set of each category (repeated 10-fold cross-validation to provide us an improved estimated Recall of our machine learning models). In our case, we employed feature selection methods that consider the overall set of features collectively, compared to feature ranking methods that assess each feature independently.

The Model - Random Forest Classifier.

Random Forest is one of the most common data mining techniques, employed in several different domains, including clinical applications. A key element of the Random Forest that makes their usage appealing in the medical domain is that they can be transformed to rules and provide transparency and interpretation in the decisions made (in contrast to other mining techniques). In our case, Random Forest Classifier was the algorithm with the best performance. It is easy to observe from the results of the Confusion Matrix, as well, that the Random Forest Classifier had the best Precision and Recall results of all three. So, I decided to use the Random Forest as the selected Classifier.

Conclusions.

First attempts.

In a first attempt to train our models I chose to approach the problem as a binary problem (mentally ill or not). This problem with the Binary Classification approach (and the given dataset) was quite easy to solve. The Multi-Class Classification for all conditions (healthy, schizophrenia, bipolar disorder and schizoaffective disorder) proved to be quite challenging compared to the Binary Classification approach. I managed to improve and reach an overall great accuracy of the model over 87%. However, more data, - especially on the minority classes, could make this problem to have greater medical interest, but also to definitely improve the overall performance and results.

Methodology.

In this pattern, a methodology based on data mining techniques (feature selection, using algorithms, predictive models) could be employed in the future for the development of a recommendation tool and a diagnostic model for Schizophrenia and other psychotic diseases. In order to assist professionals, firstly in requesting the necessary information from the patients to reach a potential diagnosis and secondly to support the diagnosis of the diseases as a whole, the predictive models can be a significant important tool. Following the extensive analysis on the association of these psychiatric diseases with the genetic factor, the creation of advanced diagnostic models is considered imperative.

The results.

The results of this work are indicative and based on limited data. It would be helpful having access to more genetic data from patients with psychotic diseases, and the models used for each diagnosis to be accurately in line with clinical knowledge and guidelines. An increase in the overall accuracy and recall could be presented, which could be attributed to the additional and more sophisticated features used by the professionals. Further machine learning application Schizophrenia and other psychotic diseases in real clinical settings would assist the future potential of the presented approach.

Bibliography.

1. "[The top 10 causes of death](#)", World Health Organization, retrieved October 2020
2. "[Mental Disorders](#)", World Health Organization, retrieved October 2020.
3. "[Uncovering the Genetic Architecture of Major Depression](#)", Science Direct, retrieved October 2020.
4. "[Genetics of Schizophrenia: Overview of Methods, Findings and Limitations](#)", PubMed Central® (PMC), retrieved October 2020
5. "[Identification of Key Genes and the Pathophysiology Associated With Major Depressive Disorder Patients Based on Integrated Bioinformatics Analysis](#)", National Center for Biotechnology Information, retrieved October 2020
6. "[Identification of diagnostic markers for major depressive disorder by cross-validation of data from whole blood samples](#)", PubMed Central® (PMC), retrieved November 2020
7. "[Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network](#)", Springer Nature Switzerland AG, retrieved November 2020
8. "[Molecular Genetics of Schizophrenia](#)", National Center for Biotechnology Information, retrieved November 2020
9. "[European-ancestry and African American families with schizophrenia](#)", PubMed Central® (PMC), retrieved November 2020
10. "[Identification of loci associated with schizophrenia by genome-wide association and follow-up](#)", PubMed Central® (PMC), retrieved November 2020
11. "[Common polygenic variation contributes to risk of schizophrenia and bipolar disorder](#)", PubMed Central® (PMC), retrieved November 2020
12. "[Common variants conferring risk of schizophrenia](#)", PubMed Central® (PMC), retrieved November 2020
13. "[Psychosis Genetics: Modeling the Relationship Between Schizophrenia, Bipolar Disorder, and Mixed \(or “Schizoaffective”\) Psychoses](#)", National Center for Biotechnology Information, retrieved December 2020
14. "[Computational Psychiatry and the Challenge of Schizophrenia](#)", PubMed Central® (PMC), retrieved November 2020
15. "[Machine learning for genetic prediction of psychiatric disorders](#)", PubMed Central® (PMC), retrieved November 2020

16. "[An Investigation of Psychosis Subgroups With Prognostic Validation and Exploration of Genetic Underpinnings](#)", JAMA Network, retrieved December 2020
17. "[Exome arrays capture polygenic rare variant contributions to schizophrenia](#)", National Center for Biotechnology Information, retrieved December 2020
18. "[The NHGRI-EBI Catalog of human genome-wide association studies](#)", National Human Genome Research Institute, retrieved December 2020
19. "[Synopsis of Genetic Association Studies](#)", Schizophrenia Research Forum, retrieved December 2020
20. "[Cross-sectional similarities and differences between schizophrenia, schizoaffective disorder and mania or mixed mania with mood-incongruent psychotic features](#)", PubMed Central® (PMC), retrieved December 2020
21. "[Phenomenology of Schizophrenia and the Representativeness of Modern Diagnostic Criteria](#)", PubMed Central® (PMC), retrieved January 2021
22. "[Blood-based biomarkers predicting response to antidepressants](#)", PubMed Central® (PMC), retrieved January 2021
23. "[Prevalence of Depression, Depressive Symptoms, and Suicidal Ideation Among Medical Students](#)", National Human Genome Research Institute, retrieved January 2021
24. "[Multimorbidity and depression: A systematic review and meta-analysis](#)", PubMed Central® (PMC), retrieved January 2021
25. "[Weighted Gene Coexpression Network Analysis Identifies Specific Modules and Hub Genes Related to Major Depression](#)", PubMed Central® (PMC), retrieved January 2021
26. "[Identification of Key Genes and the Pathophysiology Associated With Major Depressive Disorder Patients Based on Integrated Bioinformatics Analysis](#)", PubMed Central® (PMC), retrieved January 2021
27. "[Fronto-parietal and temporal brain dysfunction in depression: A fMRI investigation of auditory mismatch processing](#)", PubMed Central® (PMC), retrieved January 2021
28. "[Depression in Schizophrenia: Associations With Cognition, Functional Capacity, Everyday Functioning, and Self-Assessment](#)", PubMed Central® (PMC), retrieved January 2021
29. "[Depression during first episode psychosis and subsequent suicide risk: A systematic review and meta-analysis of longitudinal studies](#)", Science Direct, retrieved January 2020

Annexes.

[1]"[Regents Prep: Living Environment: Homeostasis](#)". Oswego City School District Regents Exam Prep Center. Archived from the original on 25 October 2012. Retrieved November 2020.

[2]"[What is the Human Genome Project?](#)" Retrieved November 2021

[3]"[Genetic linkage](#)", Retrieved November 2020

[4]"[Hyperparameter optimization](#)", Retrieved November 2020

