IDENTIFICATION OF CRITICAL PROTEINS ASSOCIATED WITH LEARNING PROCESS FOR DOWN SYNDROME

HANDAN KULAN

PH.D. THESIS

Submitted to the School of Graduate Studies of
Kadir Has University in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Engineering

İSTANBUL, April, 2020

DECLARATION OF RESEARCH ETHICS /
METHODS OF DISSEMINATION

I, HANDAN KULAN, hereby declare that;

- this Ph.D. thesis is my own original work and that due references have been appropriately provided on all supporting literature and resources;
- this Ph.D. thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- I have followed *Kadir Has University Academic Ethics Principles prepared in accordance with The Council of Higher Education's Ethical Conduct Principles.*

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

Furthermore, both printed and electronic copies of my work will be kept in Kadir Has Information Center under the following condition as indicated below (SELECT ONLY ONE, DELETE THE OTHER TWO):

☐ The full content of my thesis will be accessible from everywhere by all means.
☐ The full content of my thesis will be accessible only within the campus of Kadir Has University.
☐ The full content of my thesis will not be accessible for ____ years. If no extension is required by the end of this period, the full content of my thesis will be automatically accessible from everywhere by all means.

HANDAN KULAN

.....

KADİR HAS UNIVERSITY

SCHOOL OF GRADUATE STUDIES

ACCEPTANCE AND APPROVAL

This work entitled IDENTIFICATION OF CRITICAL PROTEINS ASSOCIATED WITH LEARNING PROCESS FOR DOWN SYNDROME prepared by HANDAN KULAN has been judged to be successful at the defense exam on ..... and accepted by our jury as Ph.D. thesis.

APPROVED BY:

Assoc. Prof. Dr. Tamer Dağ (Advisor) .....
Kadir Has University

Prof. Dr. Feza Kerestecioğlu .....
Kadir Has University

Prof. Dr. Nevcihan Duru .....
Kocaeli University

Assoc. Prof. Dr. Tansal Güçlüoğlu .....
Yıldız Teknik University

Asst. Prof. Dr. Hatice Bahar Şahin .....
Kadir Has University

I certify that the above signatures belong to the faculty members named above.

.....
Prof. Dr. Sinem Akgül Açıkmeşe
Dean of School of Graduate Studies
DATE OF APPROVAL: .....

TABLE OF CONTENTS

IDENTIFICATION OF CRITICAL PROTEINS ASSOCIATED WITH LEARNING PROCESS FOR DOWN SYNDROME

ABSTRACT

The protein profiles of people with DS are observed by applying biochemical techniques in laboratory. However, the list of analyzed proteins is long and not all proteins in list are not related to DS. Thus, for the analysis and the treatment of DS, protein expression levels have been analyzed by applying statistical procedures and machine learning techniques. In this thesis, compared to previous works, different preprocessing steps, feature selection and classification techniques are applied to define the subsets of proteins for datasets. These subsets differentiate mice more accurately. When these subsets which affect the critical pathways of specific DS aspects are analyzed, it is monitored that selected proteins have vital roles in the processes, such as apoptosis, learning and memory, signaling pathways, immune system and Alzheimers disease (AD). The subsets of proteins selected in this thesis can be applied to interpret the causes of different symptoms in DS and can be utilized to foster effective drugs for the cure of DS.

Keywords: Down syndrome, protein expression, feature selection, memory, learning, signal pathway, immune system

DOWN SENDROMUNDA OGRENME SURECI ILE ILISKILI KRITIK PROTEINLERIN BELIRLENMESI

ÖZET

DS protein profilleri laboratuvarda biyokimyasal teknikler uygulayarak gözlemlenmek tedir. Fakat, elde edilen protein listesi uzundur ve listedeki her protein DS ile alakalı değildir. Bu yüzden, DS analizi ve tedavisinde, protein ifade miktarları istatiksel metodlar ve makine öğrenmesi teknikleri uygulayarak analiz edilmektedir. Bu tezde, önceki çalışmalara kıyasla, farklı ön değerlendirme adımları, özellik seçimi ve sınıflandırma teknikleri, farklı veri setleri için protein altkümeleri belirlenmesi için uygulanmıştır. Bu protein altkümeleri fareleri daha doğru şekilde ayrıştırır. Spesifik DS özelliklerinin kritik yolaklara etki eden bu altkümelerdeki proteinler tek tek analiz edildiğinde, seçilmiş proteinlerin öğrenme ve hafıza, sinyal yolakları, Alzheimer hastalığı, bağışıklık sistemi ve hücre ölümü gibi önemli süreçlerde rol aldığı gözlemlenmiştir. Bu tezde seçilen protein alt kümelerinden DS un farklı semptomlarını anlamak için yararlanılabilinir ve DS tedavisinde etkili ilaçlar geliştirmek için kullanılabilinir.

Anahtar Sözcükler: Down sendromu, protein ifadesi, özellik seçimi, hafıza, öğrenme, sinyal yolakları, bağışıklık sistemi

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Assoc. Prof. Dr. Tamer Dag for his support and all the things he has done for me.

I would like to thank to Prof. Dr. Feza Kerestecioglu, Prof. Dr. Nevcihan Duru, Assist. Prof. Dr. Hatice Bahar Sahin and Assoc. Prof. Dr. Tansal Gucluoglu for reviewing and commenting on the manuscript of this thesis.

I wish to thank 2211A- TUBITAK BDEB National Scholarship Program for PhD Students foundation for supporting me.

My parents, I know this thesis and PhD study will not be complete without your support and your faith in me.

to my charismatic dad,

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

A	Adenosine
AA	Amino Acid
AD	Alzheimers Disease
AdaBoost	Applied Adaptive Boosting
ADARB1	Adenosine Deaminase RNA Specific B1
AKT	Protein kinase B (PKB)
AMPA	α -Amino-3-Hydroxy-5-Methylisoxazole-4-Propionic Acid
ANN	Artificial Neural Network
APP	Amyloid Precursor Protein
ARC	Activity Regulated Cytoskeleton
BAX	BCL2 Associated X
BCL2	B-Cell Lymphoma 2
BDNF	Brain Derived Neurotrophic Factor
BH3	BCL-2 Homology Domain 3
BRAF	v-Raf Murine Sarcoma Viral Oncogene Homolog B
BRK	Breast Tumor Kinase
CAMKII	Ca^{2+} /Calmodulin-Dependent Protein Kinase II
CaNA	Carbonic Anhydrase
CB	Cerebellum
CDK5	Cyclin-Dependent Kinase 5
cDNA	Complementary DNA
CFC	Context Fear Conditioning
CFOS	FBJ Murine Osteosarcoma Viral Oncogene Homolog
CHAF1B	Chromatin Assembly Factor 1 Subunit B
CHD	Congenital Heart Defect
CLEC7A	C-Type Lectin Domain Family 7 Member A
CN	Calcineurin
CR	Cortex
CREB	cAMP Response Element Binding Protein

CS	Context Shock
CS-m	Context Shock-memantine
CS-s	Context Shock-saline
CTTNB1	Catenin Beta-1
c-CS	control - Context Shock
c-SC	control - Shock Context
c-CS-m	control - Context Shock-memantine
c-CS-s	control - Context Shock-saline
c-SC-m	control - Shock Context-memantine
c-SC-s	control - Shock Context-saline
dbEST	Database of Expressed Sequence Tags
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
DS	Down Syndrome
DSCR1	Down Syndrome Critical Region 1
DYRKA1	Dual Specificity Tyrosine Phosphorylation-Regulated Kinase 1A
EGR1	Early Growth Response Protein 1
ERBB2	v-erb-b2 Avian Erythroblastic Leukemia Viral Oncogene Homolog 2
ERBB4	v-erb-a Erithroblastic Leukemia Viral Oncogene Homology 4
ERK	Extracellular Signal-Regulated Kinase
ESR	Eritrosit Sedimation Rate
FDR	False Discovery Rate
FFNN	Feed Forward Neural Network
GABAA	γ aminobutyric acid A
GAD2	Glutamate Decarboxylase 2
GFAP	Glial Fibrillary Acidic Protein
GJA1	Gap Junction Alpha-1 Protein
GluR3	Glutamate Receptor 3

GSK3B	Glycogen Synthase Kinase 3 Beta
GTP	Guanosine Triphosphate
HCG	Human Chorionic Gonadotropin
HP	Hippocampus
Hsa21	Human Chromosome21
H3MeK4	Methylated Lysine 4 on Histone H3
I	Inosine
ID	Intellectual Disability
IDEA	Individuals with Disabilities Education Act
IEG	Immediate Early Gene
IL1B	Interleukin 1 Beta
ITSN1	Intersectin 1
JNK	c-Jun N-Terminal Kinases
KNIME	Knostanz Information Miner
KNN	K Nearest Neighbor
LOOCV	Leave-One-Out Cross Validation
LTP	Long Term Potentiation
L/M	Learning and Memory
MAPK	Mitogen Activated Protein Kinase
MeCP2	Methyl-CpG-Binding Protein 2
MEK	Mitogen Activated Protein (MAP) Kinase
mir155	MicroRNA 155
MLP	Multi Layer Perceptron
Mmu 10	Mouse Chromosome 10
Mmu 16	Mouse Chromosome 16
Mmu 17	Mouse Chromosome 17
mRNA	Messenger RNA
MTOR	Mechanistic Target Of Rapamycin Kinase
Mx1	Interferon-induced GTP-binding protein Mx1
NAM	Negative Allosteric Modulator
NL	Non Learning

NMDA	N-methyl-D-Aspartate
NMDAR	N-methyl-D-Aspartate Receptor
NN	Neural Network
nNOS	Neuronal Nitric Oxide Synthase
NR1	N-Methyl-D-Aspartate Receptor Subunit
NR2A	N-Methyl D-Aspartate 2A
NR2B	N-Methyl D-Aspartate Receptor Subtype 2B
NTRK2	Neurotrophic Receptor Tyrosine Kinase 2
ORF	Open Reading Frame
PAPP-A	Pregnancy-Associated Plasma Protein-A
PCA	Principal Component Analysis
PKA	Protein Kinase A
PKC	Protein Kinase C
PKCA	Protein Kinase C Alpha
PKCAB	Protein Kinase C Alpha/Beta
PKCG	Protein Kinase C Gamma
PRMT2	Protein Arginine N-Methyltransferase 2
PSD95	Postsynaptic Density Protein 95
PTK6	Protein Tyrosine Kinase 6
P70S6	Ribosomal Protein S6 Kinase Beta-1 (S6K1)
p75NTR	p75 Neurotrophin Receptor NGFR
RAPTOR	Regulatory-Associated Protein of mTOR
RBF	Radial Basis Function
RCAN1	Regulator of Calcineurin 1
RefSeqP	Protein Coding Genes Annotated in the Reference Sequence Database
RMS	Root Mean Square
RNA	Ribonucleic Acid
RO4938581	3- bromo-10- (difluoromethyl)- 9H-benzo [f]imidazo [1,5-a] [1,2,4] triazolo [1,5-d] [1,4] diazepine
RPPA	Reverse Phase Protein Arrays

RPS6	Ribosomal Protein S6
SC	Shock Context
SC-m	Shock Context-memantine
SC-s	Shock Context-saline
SHH	Sonic Hedgehog
SOD1	Superoxide Dismutase 1
SOM	Self Organizing Map
SRC	Sarcoma
SVM	Support Vector Machines
SVR	Support Vector Regression
SYP	Synaptophysin
TH	Tyrosine Hydroxylase
TMPRSS2	Transmembrane Protease Serine 2
TRKA	Tropomyosin Receptor Kinase A
TRKB	BDNF Activated NTRK2
t-SC	trisomic - Shock Context
t-CS-s	trisomic - Context Shock-saline
t-SC-s	trisomic - Shock Context-saline
3LME	Three Level Mixed Effects

1. INTRODUCTION

DS is the most accepted genetic basis of intellectual disability (ID) and individuals with DS demonstrate latency in motor skill progress. Most individuals with DS can accomplish primary skills on their own developmental time and demonstrate communicative intent, in spite of limitations in their verbal ability (Parker et al., 2010).

The DS can be caused by errors of nondisjunction or a Robertsonian translocation between chromosome 21 and another chromosome. The additional copy of chromosome 21 is accountable for almost 160 protein–coding genes and five microRNAs (Sturgeon and Gardiner, 2011). Overexpression of these proteins such as protein modifiers, transcription factors, RNA (Ribonucleic Acid) splicing factors, adhesion molecules and many biochemical pathway components lead to learning and memory (L/M) deficits. Furthermore, people with DS show abnormalities in count of neurons and cellular texture in brain regions, such as the cerebellum, cortex and hippocampus (Chapman and Hesketh, 2000; Nadel, 2003; Silverman, 2007). Hippocampus has critical roles in the consolidation of information from short–term memory to long–term memory. Cerebellum takes a role in motor control and some cognitive functions such as attention and language. Cortex is the highly developed region of brain and responsible for perceiving, producing, thinking and understanding language. DS is also affiliated with comparably great incidences of autism and an AD like dementia (Head et al., 2015).

Mouse models are utilized in the examination of many human abnormalities. Due to the austerity and the huge incidence rate of DS, researchers have employed mice for the progress of cure for DS. However, it is hard to represent DS in mice as orthologs

of the Hsa21 genes project to mouse chromosomes 10, 16 and 17 (Yu et al., 2010; O'Doherty, 2005). The trisomic mice, Ts65Dn mice which contains 5 microRNA genes and 88 orthologs of Hsa21 protein coding genes and Tc1 mice that is trisomic for 120 HSA21 protein–coding genes have been used as DS mice models (Davisson, 1993; Rueda, 2012). These trisomic mice exhibit similar characteristics to the DS, including anomalies in learning and synaptic plasticity. Using these trisomic mice, protein expression profiles of DS patients have been assessed by computational learning methods in order to comprehend mechanism of DS. Thanks to computational learning methods, datasets that show protein expression profiles of mice are processed. After preprocessing step, the important proteins are obtained by applying feature selection methods and healthy and unhealthy mice are discriminated based on these proteins.

In the recent years, many drugs have been observed to rescue one or more abnormalities in trisomic mice. Untreated trisomic mice were unsuccessful to learn unless they are injected with drug, they can grasp successfully. These successes have promoted noticeable passion for clinical tests to treat DS. For the cure of the DS, numerous attempts have been tried for developing drugs. More than 20 drugs that have different effects, such as aminobutyric acid A (GABAA) receptor antagonists, acetylcholinesterase suprassants, N–methyl–D–aspartate receptor (NMDAR) antogonist and the green tea component have been shown to be impressive for recovering performance in L/M efforts (Gardiner, 2014; Braudeau et al., 2011; Block, 2018; Costa, Scott-McKean and Stasko, 2007; Chang and Gold, 2008; Corrales, 2013; Das et al., 2013; Busciglio, 2013; Gardiner, 2010, Chen and Lipton, 2005; Lipton, 2007; Kamat, 2013; Olivares, 2012).

The protein profiles of different datasets are observed by applying biochemical techniques in laboratory. However, the list of obtained proteins is long and not all proteins in list are not related to DS. Thus, it is necessary to determine exact protein subset which is critical in DS. In previous works, protein expression profiles based on learning outcome, age, sex, brain regions and subcellular fraction of brain regions were evaluated by statistical methods (Higuera, Gardiner and Cios, 2015; Ahmed et

al., 2014, 2015; Eicher and Sinha, 2017; Feng et al., 2017). Feature selection from protein profiles based on learning outcome (succesful, rescued learning with drugs, failed) provides the critical proteins in learning process. The result of age related abnormalities over different brain regions displays the importance of aging in DS. In addition, analysis based on subcellular parts of brain shows the importance of brain regions in DS. However, statistical techniques only show the change such as decrease or increase in protein profiles and do not exactly determine the critical proteins in DS. In addition to the statistical techniques, the machine learning algorithms are also practiced to determine critical proteins. However, the type and parameter of applied techniques were not appropriate and preprocessing steps are not very efficient in previous works.

In this thesis, the preprocessing step includes filling of missing values and normalization is performed in a different way when compared to other works. Missing values are filled with the mean value of related sample's protein expression level in the same class. Z score normalization (Abdi and Lynne, 2010) is done to inhibit the huge impact of proteins with greater effects on classification. After the preprocessing step, the forward feature selection technique is applied for determining the protein subsets for different datasets. These datasets show importance of learning outcome, age, brain regions, subcellular parts of brain regions and different type of mouse models. Naive Bayes learner is applied for the learning process in forward feature selection. It is efficient in multiclass classification is applied (Tsoumakas and Katakis, 2007; Aly, 2005). After selecting features, DNN (Deep Neural Network), gradient boosting tree, random forest and SVM classification techniques are used to discriminate control and trisomic Ts65Dn mice. The accuracy result of this work turned out to be higher than Feng et al. (2017) for all classification methods. The detailed analysis to determine critical proteins in successful learning, failed learning and rescued learning are also performed. The accuracy results of this work are higher than Higuera et al. (2015) for all classification methods. In addition, the pathway analyses are done in order to figure out the molecular mechanism of DS and foster powerful drugs for the treatment of DS.

The contribution of this work is the application of different steps to protein expression datasets. The preprocessing steps, feature selection and classification techniques are applied in a different way in order to differentiate healthy and unhealthy mice more accurately. The obtained higher classification accuracies for all classification methods substantiate the efficieny of different processing steps applied in this work.

The selected proteins are very important in order to understand the cause and cure of the DS. The biological processes can be understood by analyzing the pathways on which the selected proteins affect one by one or aggregately. The selected proteins can be effective in specific DS aspects such as ID and affects motor, cognitive, linguistic, personal or social skills. Thus, the evaluation of proteins is important in order to understand the causes of different aspects for DS. After understanding the cause of the DS, the treatments can be possible by developing the effective drugs.

The rest of the thesis is designed as follows: Chapter 2 explains the background information and related works. Datasets used for this work and data preprocessing steps are explained in Chapter 3. The feature selection algorithm to identify the critical proteins are described in Chapter 4. Chapter 5 introduces the classification methods and Chapter 6 illustrates the pathway analysis of the selected protein subsets. The results part in Chapter 7 shows the feature selection and classification results. Finally, in conclusions, the general evaluation of this thesis is presented and possible extensions to the work are introduced.

2. BACKGROUND INFORMATION AND RELATED WORKS

2.1 Down Syndrome

This chapter presents general information on DS such as its diagnosis, treatment, therapies and drugs. To understand DS, many researchers are working on different mice models. These mice models and current research efforts to observe the important proteins associated with the DS are also explained in this chapter.

DS is the most accepted genetic reason of ID and affects almost one in 700 live births worldwide (Parker et al., 2010). DS is characterized with anomalies at the cellular, electrophysiological, molecular and behavioral level. People with DS have a characteristic facial display and weak muscles in childhood. All affected individuals struggle cognitive delays, but ID is frequently mild to moderate. With DS, the count of neurons and cellular texture becomes unusual in brain regions, such as the cortex, hippocampus and cerebellum (Chapman and Hesketh, 2000; Nadel, 2003; Silverman, 2007). Furthermore, people with DS are at danger for specific types of blood disorders, like autoimmune disorders, leukemia and an AD like dementia (Head et al., 2015).

DS was first described by Jean-Étienne Dominique Esquirol in 1838 and later by Édouard Séguin in 1844 (Neri and Opitz, 2009). However, DS was first characterized in 1862 by English physician John Langdon Down and named after him. Down published a report in 1866 and recognized DS as a specific type of mental disorder (Hickey and Summar, 2012; Down, 1867). By the 20^{th} century, DS became the most noticeable type of mental disorder. In the ancient times, many newborns with disorders were either abandoned or killed. Various historical items of art are

considered to illustrate DS (Bernal and Briceno, 2006). In the 20^{th} century, numerous individuals with DS were stereotyped. Some of the related medical troubles were handled and most people with DS died in childhood. With the upsurge of the eugenics movement, a lot of countries started programs of restricted sterilization of individuals with DS (Prost and Nasreen, 2013).

With the invention of karyotype methods in the 1950s, it became possible to pinpoint anomalies of chromosomal number. In 1959, Jerome Lejeune determined that DS was associated with an additional, third copy of human chromosome 21 (Lejeune, Turpin and Gautier, 1959). The extra copy of 21^{st} chromosome is accountable for nearly 160 protein–coding genes and five microRNAs. The overproduction of genes coded by the additional copy of Hsa21 in DS is considered to be adequate to disrupt numerous distinct biological actions and pathways, such as influencing development of brain and cause learning and memory defects.

Figure 2.1 shows the karyotype depiction of chromosomes (Ghani, 2019). People with DS have one extra 21 chromosome. Karyotype analysis of both parents is advised for analysis of DS.

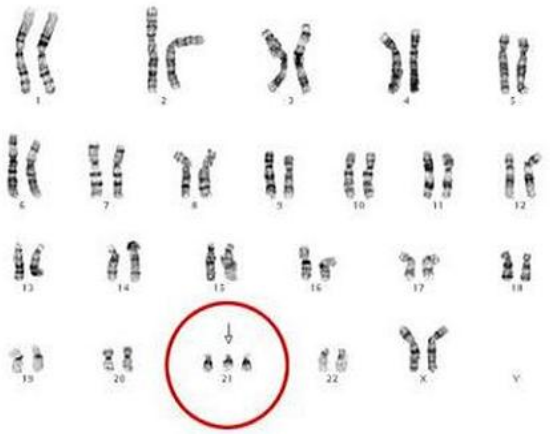

Figure 2.1: Karyotype representation of chromosomes (Ghani, 2019)

Figure 2.2 depicts nondisjunction of chromosomes during meiosis stage (Sleigh, 2019). When one cell divides in two, pairs of chromosomes are divided and one of the pairs goes to one daughter cell, the other pair goes to the other daughter

cell. In nondisjunction, both chromosomes from one pair go into one cell and no chromosomes for that pair go into the other cell. This process causes an extra copy of 21^{st} chromosome.

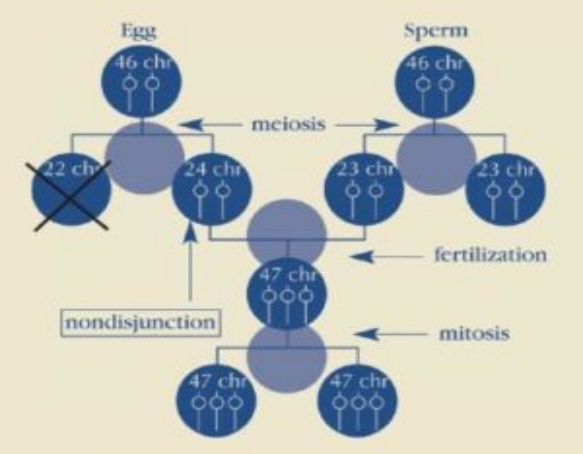

Figure 2.2: Nondisjunction of chromosomes during meiosis (Sleigh, 2019)

Another reason of DS is a Robertsonian translocation (Cooper, 2019) between chromosome 21 and another chromosome as shown in Figure 2.3. Robertsonian translocation is occurred when the whole long arms of two acrocentric chromosomes in which the centromere is located quite near one end of the chromosome are merged.

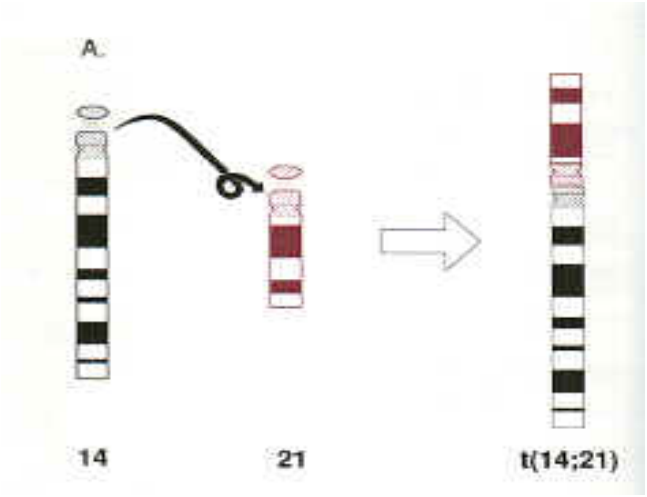

Figure 2.3: Robertsonian translocation (Cooper, 2019)

The major signs of DS are physical development, intellectual impairments and other health dilemmas. Figure 2.4 shows deficiencies that can be seen with a person diagnosed by DS (Pameer, 2019). DS is associated with several distinct physical

characteristics, such as short and broad neck, flattened facial facets, smaller than average head and ears. Even though babies with DS are typically of average size at birth, they have smaller growth rate when compared to children without the disorder. As a result of poor muscle tone, a person with DS has delays during walking.

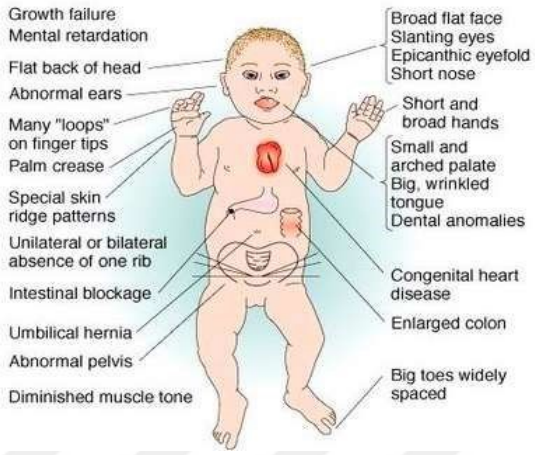

Figure 2.4: Abnormalities of person with DS (Pameer, 2019)

People with DS have intellectual disabilities. Delayed improvement of language and speech, poor attentiveness and impulse regulation and learning problems are common characteristics. In addition, they are usually at much higher risk for developing other health problems, such as vision problems, digestion problems and heart disease. Understanding the molecular anomalies that form the basis of the DS can lead to the development of cures to prevent or alleviate at least some facets of the phenotype. Thus, functions of specific Hsa21 genes linked to anomalies detected in DS have been examined. This examination provided the discoveries of efficient treatments for pharmacotherapies that can stop or proper the abnormalities and cognitive deficits.

By 1990, 13 Hsa21 genes with biological associations had been identified (Gardiner et al., 1990). Before this time, gene mapping was a slow and laborious process. The pace of Hsa21 gene identification raised in the 1990s thanks to the development of the techniques of exon trapping and cDNA (Complementary DNA). These techniques identified transcribed sequences from defined genomic regions (Chen et al., 1996; Dahmane et al., 1998; Tassone et al., 1995). Together with the augmentation of

the number of dbEST (database of expressed sequence tags) entries from human tissues, there has been a rapid rise in complete cDNA open reading frame (ORF) sequences (Gardiner and Yaspo, 1998). In 2000, the genomic sequence of Hsa21q was disclosed (Hattori et al., 2000). Table 2.1 lists the classes and numbers of HSa21q genes present in that initial annotation.

Table 2.1: Initial annotation of Hsa21 (Hattori et al., 2000)

	# protein coding genes
Known protein coding	127
Protein similarity	13
Domain similarity	17
Novel	68
Total	225

Initial annotation had few gaps and identified 225 genes and gene models. The dbEST extended in both human and model organism entries and the mouse genome was sequenced. Thus, the Hsa21 gene catalog was reviewed and supplemented with experimental validation and information on conservation in mouse (Gardiner et al., 2002, 2003; Reymond et al., 2001, 2002). Data from the most recent annotation of Hsa21 (Sturgeon and Gardiner, 2011) are presented in Table 2.2.

Table 2.2: Annotation of Hsa21 (Sturgeon and Gardiner, 2011)

	# protein coding genes
RefSeqP	161
MicroRNA	5
Novel; ORF > 50 AA	146
Novel; ORF < 50 AA, repetitive or incomplete	250
Total	562

The most impressive discrepancy in comparison with the 2000 gene list is the more than two-fold raise in the number of genes. This is the result of the mRNA (messenger RNA) and dbEST databases.

2.2 Diagnosis of DS

People at risk of having a baby with DS can take screening and diagnostic tests. While the screening tests can predict the probability of DS, diagnostic tests can

decisively inform whether a baby will have DS. Thus, only diagnostic tests can clearly identify whether the baby will have DS. These tests are explained below.

2.2.1 Screening Tests

The chance of having a DS baby increases with parents' age. Women aged 30–35 years or above can take genetic screening during pregnancy. Screening is an affordable way to substantiate whether more detailed diagnostic tests are required or not. Screening tests include ultrasound, integrated screen, nuchal translucency testing, triple screen, quadruple screen, cell-free DNA (Deoxyribonucleic Acid).

Thanks to sound waves in ultrasound imaging, the inside of the body is shown and the pregnancy status is determined. Nuchal translucency testing is done at 11 –14 weeks. In this test, an ultrasound can quantify the open space in tissue back of the fetus neck. When abnormalities are present in a baby with DS, more fluid tends to collect in this neck tissue. Triple screen or quadruple screen procedure is done at 15–18 weeks of pregnant women. This test quantifies the volume of several substances in the mother's blood. Integrated screen links results from first-trimester blood test with second-trimester quadruple screening results. It measures blood level of some substances like alpha fetoprotein, estriol, HCG (Human Chorionic Gonadotropin), plasma protein–A (PAPP–A) and the pregnancy hormone known as human chorionic gonadotropin (HCG). It checks certain protein levels in the mother's blood. Abnormalities in protein levels are indicators of the baby with DS. Cell –free DNA test is a blood test that evaluates fetal DNA present in the mother's blood.

2.2.2 Diagnostic Tests

Diagnostic tests are more definite for the diagnosis of DS. Diagnostic tests are amniocentesis, chorionic villus sampling and percutaneous umbilical blood sampling. Chorionic villus sampling is done at 8 –12 weeks. In the procedure, a small sample of placenta is taken for the analysis of fetal chromosomes. Amniocentesis is done at 15

–20 weeks and a tiny amount of amniotic fluid is taken for analysis. The cells from the fluid are then cultured. Then, karyotype analysis is performed for detecting DS. Percutaneous umbilical blood sampling is done after 20 weeks and a little sample of blood from the umbilical cord is analyzed for chromosome abnormalities.

2.3 Treatments and Therapies for the DS

While there is no effective cure for DS, there are therapies to enhance the life standard of a person with DS. Medications for DS change by individual. The exact stage of cure relies on the individual, considering the person's health, age, strengths and limitations.

Treatments are primary care to monitor growth, medical specialists and vaccinations based on the needs of the patient, speech therapy to enhance the ability to communicate, occupational therapy to support refine motor skills and make daily tasks painless, physical therapy to aid strengthen muscles and boost motor skills, behavioral therapy to help manage the emotional challenges (Winders, Wolter-Warmerdam and Hickey, 2018; Kumin, 1996; Costa, 2011; Kishnani et al.,2010; Fidler, Hepburn and Rogers, 2006; Phelps, 2010).

Speech-language therapy enhances the child's communication ability and it is given at toddler stage. It deals with communication and language skills, cognitive skills and strengthening the oral muscles. Hearing loss is frequently seen with people diagnosed by DS. Because of the anatomical distinctions in children with DS, they are susceptible to fluid detention behind the eardrum which in turn results in hearing loss. This causes life-long problems for speech and understanding (Phelps, 2010).

Physical therapy concentrates on improvement of how the person moves. People with DS generally have weak muscles and shorter hands. Thus, physical treatment can alleviate any obstacles caused by these features. A physical treatment is likely to consist bolstering and toning muscles, recovering balance and correcting posture (Winders, Wolter-Warmerdam and Hickey, 2018). Occupational therapy is carried out

to develop the daily abilities which are essential for living a healthy life. Contrary to physical therapists, occupational therapists operate on boosting fine motor skills and the achievement of daily tasks like brushing teeth, getting dressed and eating. As the child grows, the target of the therapy concentrates on getting skills like using a computer (Daunhauer and Fidler, 2011).

Assistive technology involves in equipments that aid a person with an impairment to function better. These equipments consist of hearing aids, large-button mobile phones, seat cushions, walking aids and large-letter keyboards. Tablets and computers are helpful for children with DS who have conflict in performing motor movements (Al-Moghyrah, 2017).

In USA, thanks to the U.S. Individuals with Disabilities Education Act (IDEA), children with DS take exclusive education until they either finish high school or reach the age of 21.

2.4 Prevalence of DS

Throughout the world based on 2010 data, DS occurs in about 1 per 1000 births (Weijerman and de Winter, 2010) and causes about 17,000 deaths per year (Lozano et al., 2012). Mostly, children are born with DS in places where abortion is not permitted and pregnancy usually appears at an older age. About 1.4 per 1000 live births in the US (Parker et al., 2010) and 1.1 per 1000 live births in Norway are influenced (Malt et al., 2013). The amount of pregnancies with DS is more than twice higher with many naturally aborting countries (Kliegma, 2011). It is the reason of 8% of all congenital disabilities. Also, thanks to prenatal screening and abortions, the prevalence of DS decreases.

Maternal age increases the probability of having a pregnancy with DS. At age 20, the probability is one in 1441; at age 50 it is one in 44. The father's older age is also a hazard component in women older than 35. Also, DS liability increases with women age.

2.5 Life Expectancy for DS People

The lifespan of people with DS increased enormously between 1960 and 2007. In 1960, on the average, persons with DS lived around age 10. In 2007, persons with DS lived around age 47. Numerous factors can impact how long a person with DS lives.

If babies with DS are born weighing less than 1,500 grams, they are 24 times more probable to die in the first 28 days of life in comparison to baby with normal weight (between 2,500 grams and 4,000 grams). Black or African-American babies with DS have little chance of living beyond the first year of life compared with white babies with DS. Further investigation is required to understand the cause of this. Infants with DS who have a congenital heart defect (CHD) are five times more probable to die in the first year of life. Between 1983 and 2003, about 93% of babies born with DS lived to one year of age. In the same course of time, about 88% of babies born with DS endured to 20 years of age. The number of babies with DS who die before one year has reduced over time. By comparison, the percentage of death during the first year of life reduced from 1.5% during 1979-1983 to 0.9% during 1999-2003.

2.6 Alzheimer's Disease Risk

DS increases the risk of AD. Approximately all adults with DS display the neuropathological modifications of AD by the age of 40 years. This linkage promotes an understanding in the advancement of AD and offers special understandings for AD in the overall community. Amyloid-β builds up in the brain through the lifetime of people with DS, which contributes a special change to grasp the temporal advancement of AD and the dementia initiation.

Studying the function of APP in DS might cause comprehending its function in both sporadic AD and familial AD. The age reliance in the progress of AD in DS can stimulate research into the appearance of AD. Juxtaposition of the biomarker profiles, genetic profiles and risk profiles of adults with DS with those of individuals

with AD can aid to figure out specific pathways for dementia. Analysis of brain scans and other tests aid to diagnose AD.

2.7 Mice Models Used for the Analysis of DS

Investigation of molecular actions at the level of protein modification contributes a direct assessment of functional feedbacks. Thus, protein expression data have been assessed by computational learning methods using mice models in order to diagnose and treat DS. Nonetheless, DS is hard to model in mice as orthologs of Hsa21 genes project to three mouse chromosomes, indicated as Mmu10, Mmu16, and Mmu17. Figure 2.5 displays mapping of three different mouse chromosomes to HSA21 (Antonarakis et al., 2014).

Figure 2.6 shows the mapping of Hsa21 genes on mice chromosomes Mmu10, Mmu16, and Mmu17 and different type of trisomic mice models (Sturgeon et al., 2012). Using these trisomic mice models, DS can be analyzed. In Figure 2.6, genes within three regions of HSA21 shown at the left project to segments of mouse chromosomes 10, 17, and 16 as indicated. Two mouse models, Ts65Dn and Tc1, are shown at right. The Ts65Dn is trisomic for the telomeric segment of MMU16. The Tc1 mouse model carries normal complement of mouse chromosomes and Hsa21 that has internal deletions (indicated by grey circles). The number of HSA21 Reference Sequence database protein-coding genes (RefSeqP) mapping to HSA21 regions. The regions in mouse models are indicated at bottom part of figure.

Ts65Dn and Tc1 mouse models are used in this thesis. Mouse models have become very helpful for scientists to analyze human ailments. Humans have a lot of similar genes with mice. Since, mice are bred quickly, relatively cheap and perform experiments not applicable to humans, they are frequently used. The ideal DS mouse model is that all genes placed on Hsa21 are triplicated. This is difficult as the genes placed on Hsa21 are expanding over three mouse chromosomes (chromosomes 10, 16 and 17).

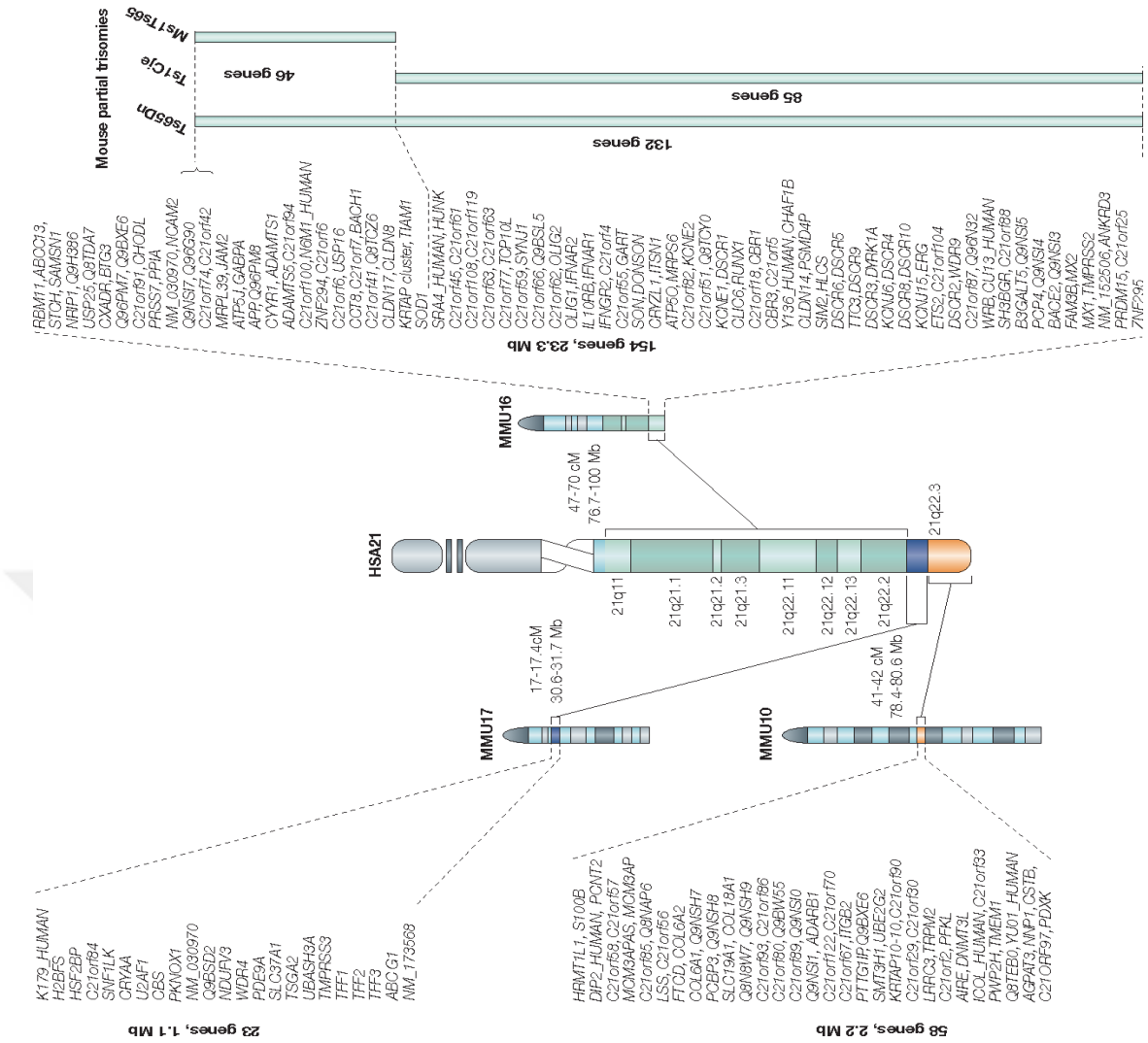

Figure 2.5: Mapping of three different mouse chromosomes to HSA21 (Antonarakis et al., 2014)

The first designed mouse model is the Ts16 and it was designed in the 1970s. This model has an additional copy of mouse chromosome 16 that incorporates significant part of the genes placed on Hsa21. Unfortunately, these mice do not endure past birth and thus cannot be utilized to inspect the progress and function of the aging progresses or the nervous system. These mice have additional genes from mouse chromosome 16 that are not triplicated in DS (Heyn, 2005).

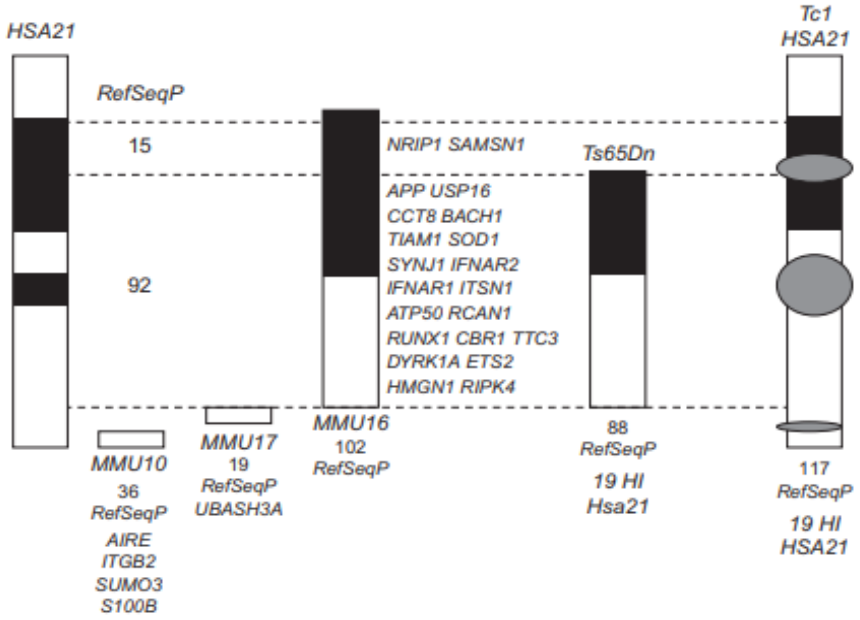

Figure 2.6: Hsa21 and the regions of trisomy in the Ts65Dn and Tc1 mouse models of DS (Sturgeon et al., 2012)

2.7.1 Ts65Dn Mouse Model

The Ts65Dn mouse is trisomic for the Mmu16 region from mir155 to Zfp295, a region that displays excellent conserved linkage with Hsa21 (Davisson et al., 1990; Gardiner et al., 2003). The Ts65Dn is partly trisomic. This means that it encompasses an additional copy of many, but not all mouse genes identical to Hsa21. It expresses some characteristics of DS such as impaired learning, behavior deficits, developmental delay and weight problems.

These mice are at dosage asymmetry for the region from APP through Tmprss2 that contains nearly half of the genes on Hsa21. Ts65Dn mice are not at dosage asymmetry for the most distal segment of Hsa21, which demonstrates conserved synteny with Mmu10 and Mmu17. On the other hand, the region triplicated in Ts65Dn incorporates the Hsa21 segment from D21S55 to MX1 that has been recorded to consist of genes chargeable for numerous DS features. Ts65Dn mice displays unsatisfactory efficiency in the Morris water maze (Braudeau et al., 2011a; Escorihuela et al., 1995, 1998; Olson et al., 2007; Rueda et al., 2010; Stasko and Costa, 2004), contextual

fear conditioning (Faizi et al., 2011; Kleschevnikov et al., 2012; Hyde et al., 2001), object recognition (Braudeau et al., 2011a; Kleschevnikov et al., 2012; Lockrow et al., 2011; Fernandez et al., 2007), other memory tests (Demas et al., 1996, 1998; Hunter et al., 2003), as well as deficient long term potentiation (LTP) (Siarey et al., 1997; Costa and Grybko, 2005; Garcia-Cerro et al., 2014; Lysenko et al., 2014; Fernandez et al., 2007; Kleschevnikov et al., 2004; Filippo et al., 2010), as well as age related worsening of neuronal cultures spontaneously impacted by DS and AD (Cooper et al., 2001; Salehi et al., 2009).

The Ts65Dn strain is aneuploid, which means that it has an additional chromosome conveying a portion of Mmu16 orthologous to Hsa21 and so utilized widely for the investigation of DS (Reeves et al., 1995). Nonetheless, this extra chromosome also encompasses 10 Mb of Mmu17 including 60 mouse genes that are not orthologous to Hsa21, thus reducing the benefit of this model (Duchon et al., 2011).

2.7.2 Tc1 Mouse Model

The Ts65Dn mice do not represent a complete trisomy. The model contains extra genes on Hsa21 that may cause anatomical, behavioral or physical differences that are irrelevant to DS. Thus, scientists look for improving mouse models to make them more similar to human habituates.

Tc1 mice model carries roughly the entire replica of Hsa21 (nearly 92% of all genes). This model was constructed by utilizing a technique called transfer of irradiation microcell mediated chromosome. Over a series of transfer, Hsa21 was removed from a donor cell and independently inserted into small cells. These microcells were combined to recipient mouse embryonic stem cells. The cell encompassing the chromosome 21 largest fragment (90%) was selected for insertion into mouse embryos at initial phase of improvement. Afterwards, the embryos were reimplanted into the mother. The derived mice were fused to normal mice to create the mouse model known as Tc1. More than 40% of the mice acquires the Hsa21 fragment.

The Tc1 mouse type is practically trisomic for nearly 120 Hsa21 protein coding genes. They exhibit features related to the DS phenotype, containing anomalies in learning (O'Doherty et al., 2012; Morice, 2008; Galante et al., 2009). The Tc1 mice not only demonstrates some of the DS features present in other mouse models. It also displays heart deficiencies that are similar to those that make trisomy 21 the main reason of congenital heart disease. None of the other mouse models mimic DS heart defects so well. The considerable harm of Tc1 mice is that they are mosaic, thus not every cell in the Tc1 mouse has a copy of Hsa21 and every mouse is distinct. This becomes an intricate matter as it is more lengthy and more costly to figure out whether or not the tested cells were trisomic after the experiment. A larger number of mice may be required in order for the statistics to be worthwhile. The Tc1 mouse contains various genes not present in Ts65Dn mice. Tc1 is thus a more integrated model of DS. However, Tc1 has absence for a few genes that are present in Ts65Dn mice.

2.8 Drugs for DS

To cure the DS, numerous attempts are done for developing drugs. Over the last few years, some drugs have been observed to recover one or more anomalies in the trisomic mice. Untreated trisomic mice have been unsuccessful to learn but if they are first instilled with drug, they can learn correctly, thus, learning is recovered.

Comparison of the trisomic mice protein profiles when they are unsuccessful and when their learning is recovered with drugs shows numerically important variations in protein levels related with rescued learning. These successes have promoted to noticeable passion for clinical trials to improve drugs. Greater than 20 drugs with varied qualities, such as γ aminobutyric acid A (GABAA) receptor antagonists, the green tea component and acetylcholinesterase suppressants have been shown to be useful for rescuing performance in L/M (Gardiner, 2014; Braudeau et al., 2011; Block, 2018; Costa, Scott-McKean and Stasko, 2007; Chang and Gold, 2008; Corrales, 2013; Das et al., 2013; Busciglio, 2013; Gardiner, 2010; Chen and Lipton, 2005; Lipton,

2007; Kamat, 2013; Olivares, 2012).

The memantine drug is currently in operation for remedy of learning impairments in DS. The structure of memantine is illustrated in Figure 2.7 (DrugCentral, 2019).

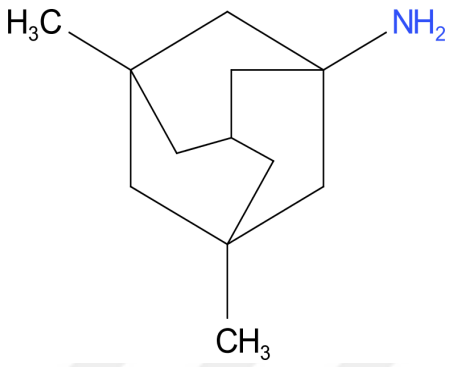

Figure 2.7: Memantine (DrugCentral, 2019)

Even though memantine is notable to adjust excitatory neurotransmission via antagonizing activity of NMDA (N-methyl-D-Aspartate) receptor as shown in Figure 2.8 (DrugsDetails, 2018), limited knowledge is available about its responses on protein expression, either solo or with learning situations. Memantine prevents Ca^{2+} cytotoxicty by supressing excitatory neurotransmission.

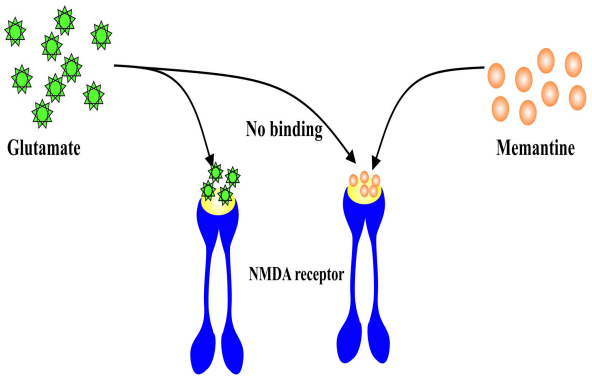

Figure 2.8: NMDA receptor and memantine (DrugsDetails, 2018)

GABAA-mediated inhibition is known to be a vital system for supplying the L/M modifications found in trisomic mice. Figure 2.9 (Ezza and Khadrawyb, 2014) shows glutamate effect.

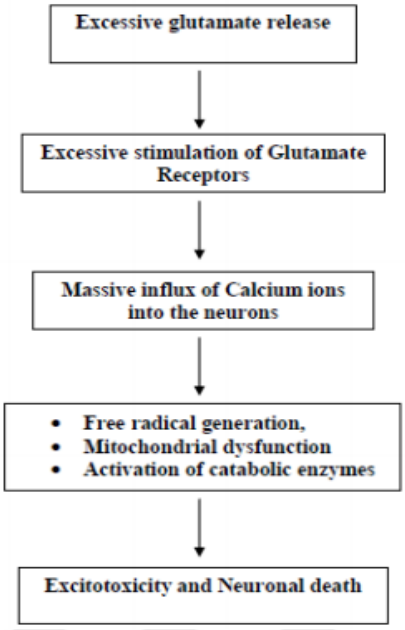

Figure 2.9: Glutamate effect on NMDA receptor (Ezza and Khadrawyb, 2014)

When glutamate binds to NMDA receptor, it increases calcium influx. This causes free radicals and these free radicals damage the cells. However, drug memantine prevents binding of glutamate to NMDA receptor. Thus, it prevents cell deaths.

Another drug, RO4938581 (3- bromo-10-(difluoromethyl)- 9H-benzo[f]imidazo[1,5-a][1,2,4] triazolo [1,5-d] [1,4] diazepine) is shown in Figure 2.10 (Davies, 2019) is GABAA receptor negative allosteric modulator (NAM).

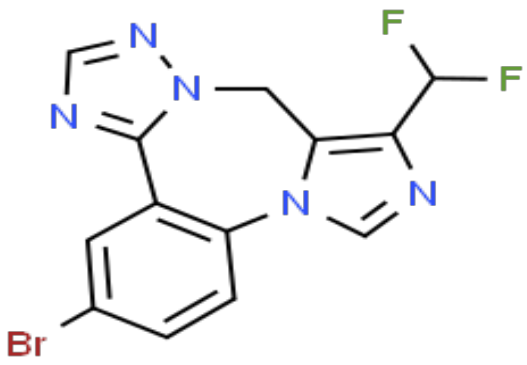

Figure 2.10: RO4938581 (3- bromo-10-(difluoromethyl)- 9H-benzo[f]imidazo[1,5-a][1,2,4] triazolo [1,5-d] [1,4] diazepine) (Davies, 2019)

RO4938581 has a binding and functional affinity specific for GABAA receptors which has a5 subunit. Just like memantine, it is also used for refining protein abnormalities in the Ts65Dn. It has been demonstrated that RO4938581 enhances L/M and recover neurogenesis, without the side effects, such as anxiety and convulsions.

2.9 Analysis of Protein Profiles

The datasets of protein profiles are divided into classes of mice after experimenting in CFC (Context Fear Conditioning) with and without treatment of memantine.

2.9.1 Context Fear Conditioning

In CFC protocols (Fanselow, 1990), CS group is inserted into a cage, waiting a few minutes to analyze the context. Subsequently, an electric shock is given. It is awaited from control mice to connect the condition with an electric shock and freeze after re-expose to the identical cage. The SC group is placed in a cage for checking the reaction of the shock only. After the placement in the cage, the electric shock is applied instantly. It is forecasted that wild type mice do not understand to connect the cage with shock and do not freeze after re-expose the identical cage. Thus, CS mice have learning capacity. However, SC mice are not capable of learning. Protein responses after CFC have been recorded. The trisomic (such as Ts6Dn, Tc1) CS group of mice cannot learn and not freeze. However, if the trisomic CS group of mice is instilled with drug, learning can be recovered.

2.9.2 Reverse Phase Protein Arrays

The protein expression amounts of each mice are quantified with RPPA (Tibes et al., 2006) that is a high-output method. Protein samples from each mice are robot-assisted placed onto nitrocellulose-coated microscope slides. RPPA detects protein expression through antibody and antigen interaction and provides a quantitative assesment of the differential expression of proteins.

2.9.3 Classes of Mice

Table 2.3 shows the class of mice and the learning outcome. As seen in Table 2.3, the dataset is divided into eight classes of mice according to the expression profiles of proteins after training in CFC with and without injection of memantine. The control mice and trisomic mice are shown as c and t, respectively. Control mice are healthy mice and learn successfully. These mice are divided into groups based on context shock or drug mematine is applied or not. CS stands for mice that are exposed to context shock and SC represents the shock context only. The letter m stands for drug memantine and s shows saline. Saline is a salt solution and used as a control of drug.

Table 2.3: Classes and learning outcome of datasets.

Class	Learning Outcome
c−SC−s	No Learning
c−SC−m	No Learning
c−CS−s	Successful Learning
c−CS−m	Successful Learning
t−SC−s	No Learning
t−SC−m	No Learning
t−CS−s	Failed Learning
t−CS−m	Rescued Learning

2.10 Related Works

The protein profiles of different datasets are observed by applying biochemical techniques in laboratory. However, the list of obtained proteins is long and not all proteins in list are not related to DS. Thus, it is necessary to determine exact protein subset which is critical in DS. In the literature, protein anomalies in DS have been observed by using a variety of techniques to select proteins. Table 2.4 shows a summary of the techniques that can be applied in the literature.

Firstly, Ahmed et al. (2014) examined a three level mixed effects (3LME) statistical analysis method of the Ts65Dn and normal mice protein profiles with and without exposure to Context Fear Conditioning (CFC). Then, Higuera et al. (2015) inspected

Table 2.4: Applied techniques in the literature.

	Techniques
Ahmed et al. (2013, 2014, 2015, 2017)	3LME statistical model
Higuera, Gardiner and Cios (2015)	SOM
Eicher and Sinha (2017)	Linear SVM
Block et al. (2018)	3LME statistical model
Feng et al. (2017)	Decision Tree+Random Forest+SVM

the profiles by applying unsupervised learning method, Self Organizing Map (SOM). It pinpoints the critical proteins for three cases; successful learning, rescued learning with memantine and failed learning. However, Eicher and Sinha (2017) stated that the problem was more relevant to a classification problem rather than a clustering problem. They used the linear SVM (Support Vector Machines) to figure out proteins that are distinct among two classes or groups of classes.

Block et al. (2018) applied 3LME and used another drug RO4938581 for recovering protein abnormalities. Feng et al. (2017) applied adaptive boosting (AdaBoost) for feature selection. Then, they applied random forest, decision tree and SVM classification techniques for discriminating normal and trisomic mice. Using Tc1 mice model, Ahmed et al. (2015) also analyzed protein profiles across brain regions and compared protein expression abnormalities of Tc1 and Ts65Dn mice. In addition, Ahmed et al. (2017) analyzed age process in DS by comparing the protein expression profiles of old and young mice. The detailed information about these studies can be found in the subsections below.

2.10.1 3LME Statistical Method

3LME statistical method is a multilevel model in which parameters vary at more than one level. The units of level 1 are nested within groups of level 2. Then, the groups are themselves nested within supergroups of level 3.

Ahmed et al. (2015) and Block et al. (2018) applied 3LME statistical methods in their works. Ahmed et al. (2015) measured levels of 85 protein expression in

the hippocampus and the cortex of Ts65Dn mice. They observed that more than 40 responses detected in control mice for successful learning was seen in Ts65Dn for failed learning. In addition, they showed that the cure with memantine did not normalize the starting protein levels. However, it induced responses in nearly half of proteins and resulted in normalization of protein endpoint levels. This work provides an initial insight of the complications related with pharmacological learning rescue in the Ts65Dn. They assessed more than 80 protein levels in subcellular fractions from cortex and hippocampus of mice trained in CFC. More than half of the protein levels adjusted in one or more portions. 37 protein levels altered in the nuclear fraction of hippocampus alone. Anomalies in thirteen protein levels were recorded in brains of patients with AD. Furthermore, Ahmed et al. (2014) investigated the subcellular fraction of hippocampus and cortex protein profiles of mice treated with memantine. Out of 84 proteins displayed in one or more fractions, expression levels of 72 and 65 were noticeable in cytosolic and nuclear fractions, and 28 in membrane fractions. In all three subcellular fractions of hippocampus, half of the proteins replied to one or more stimuli (successful learning, memantine treatment, or successful learning with memantine) increased. In hippocampus, memantine caused many alterations similar to those seen after CFC and the levels of proteins associated with AD anomalies. In cortex, proteins in the nuclear fraction were less clear and specific. Cortex also demonstrated a great number of proteins decreasing, especially in the membrane and nuclear fractions.

In order to explain how the complexities of the detected protein alterations might be incorporated into a response to L/M, they contemplated functional relationships between the proteins measured by reverse phase protein arrays (RPPA). They chose LTP (long term potential) as it is an important cellular mechanism underlie L/M. The LTP pathway consisted of 70 proteins. It involved signaling through NMDAR to MAPK (Mitogen Activated Protein Kinase), plus contributions from PKC (Protein Kinase C), PKA (Protein Kinase A), and calcineurin complexes, and elements of the mTOR (Mechanistic Target Of Rapamycin Kinase) pathway. Fifteen RPPA proteins are units of the LTP pathway. They extended the pathway by adding

the 30 RPPA proteins that instantaneously connected with components of the LTP pathway. A total of 35 RPPA proteins answered in NL (Non Learning). Memantine treatment corrected 22 of these responses. Successful learning enhanced modulation of the LTP pathway while memantine significantly changed these normal responses. Comparison of these patterns is beneficial in comprehending the vital features of molecular responses to L/M.

Ahmed et al. (2013) also used Tc1 mice which is functionally trisomic for nearly 120 Hsa21 genes to analyze the expression of 93 protein levels in hippocampus, 88 proteins in cortex and 64 proteins in cerebellum. They showed that 26 proteins changed in expression levels for at least one brain region. They compared protein anomalies in Tc1 mice with the expression level of Ts65Dn mice. While there were similarities, there were anomalies unique to the Tc1 mice. Moreover, Ahmed et al. (2017) applied 3LME statistical technique to understand age process in DS. They analyzed protein expression changes in different brain regions based on age. They showed that the number of protein abnormalities are higher at 12 months than at 6 months. The number of anomalies in cerebellum decreased while the number in cortex increased significantly with age.

Block et al. (2018) used 3LME statistical technique to observe recovering performance of Ts65Dn mice. Numerous drugs and short molecules were observed to rescue the L/M defaults. Thus, Block et al. (2018) used GABAAa5- selective modulator, RO4938581, for rescuing protein abnormalities of Ts65Dn mice. Oral intake of RO4938581 to Ts65Dn mice recovered LTP and neurogenesis. It enhanced L/M without the side effects, such as concern and spasm, observed with nonselective GABAA receptor modulators (Marti'nez-Cue' et al., 2013).

RO4938581 connects directly to the a5 subunit of the GABAA receptor. No transcription factors and other proteins are known to be markers. Nonetheless, RO4938581 cure may cause differences in protein post-translational or transcription modifications as RO4938581 connection reduce inhibition. The correlated rise in excitatory

neurotransmission causes activity dependent plasticity which stimulates many genes transcription (Flavell and Greenberg, 2008). Furthermore, since GABAA receptors are centered at dendrites of cells in juxtaposition to NMDA receptors, inhibition of GABAA receptor activity may provide to activation of NMDA receptors with responses in signaling pathways.

Block et al. (2018) measured 91 protein levels pertinent to brain tasks by applying the 3LME. 44 of the 52 abnormalities in trisomic Ts65Dn mice were amended by RO4938581. They also compared drug memantine and drug RO4938581 responses. They identified similar and different outcomes of the two drugs.

2.10.2 SOM Method

SOM is an artificial neural network (ANN) that is trained using unsupervised learning to map high-dimension inputs to a low dimensional discretised representation. It conserves the underlying structure of its input space. Higuera, Gardiner and Cios (2015) applied SOM technique to their work.

They stated that the statistical study done by Ahmed et al. (2014, 2015) was not adequate to pinpoint all variations in protein profiles. They claimed that machine learning methods realize these demands. They used SOM to group protein profiles by using 77 protein levels gathered from control mice and Ts65Dn mice both with and without treatment of the memantine.

The SOM method recognized fewer subgroups of proteins forecasted to make the vital supports to successful learning, failed learning and rescued learning. They specified a set of class-specific groups which was created from nearby nodes containing samples from a single class or a node with at least 80% of its samples gathered from one mouse. Then, Gardiner and Cios (2015) carried out the Wilcoxon rank-sum method. They figured out that protein amounts were remarkably distinct between each pair of groups and defined those proteins as distinct between two classes.

For the experiments, they evaluated SOMs created only control group mice to assess molecular processes in successful learning. The classes in control groups are explained in Table 2.3. In this SOM, c-CS (control - Context Shock) mice were apparently disconnected from c-SC (control - Shock Context) mice. Also, there were no nodes consisting both CS and SC evaluations which demonstrate that disparities in the amounts of these proteins distinguish successful learning from the absence of stimulation to learn. Next, they contemplated the four comparisons related to successful learning, c-CS-s/m (control - Context Shock- saline/memantine) vs. c-SC-s/m (control - Shock Context- saline/memantine). They used the Wilcoxon-rank sum method to pinpoint the protein sets differentiated markedly in each of the four comparisons. All comparisons differed in 11 proteins. The chance that the proteins are vital to successful learning was checked by yielding a SOM employing only these 11 proteins. The disconnection of CS (Context Shock) from SC (Shock Context) mice was preserved, i.e., these proteins were enough for discerning successful learning from lack of stimulation to learn. Furthermore, the quantity of mixed CS-s (Context Shock-saline) plus CS-m (Context Shock-memantine) nodes twofolded like the number of mixed SC-s (Shock Context- saline) and SC-m (Shock Context- memantine) nodes. To sum up, they propose that the 11 proteins are essential for successful grouping or segregation. Together these results robustly help the relevant biological significance to learning of the 11 proteins.

Then, Gardiner and Cios (2015) analyzed SOMs with the equivalent groups of trisomic mice, to analyse failed learning and its recovery by memantine. A similar implementation of SOM to the trisomic mice dataset generated very distinct outcomes. Trisomic t-CS-s (trisomic- Context Shock-saline) mice are unsuccessful to learn. When the four sets of trisomic mice were grouped, the SOM exhibited a t-CS-s cluster of nodes directly nearby to t-SC (trisomic- Shock Context) nodes and nodes that mixed CS and SC calculations. In the trisomic SOM, only 15% of CS nodes consisted both CS-s and CS-m computations. In control mice, 30% of CS nodes were mixed c-CS-s (control - Context Shock-saline) and c-CS-m (control - Context Shock-memantine). These SOM properties pointed out that protein responses when

trisomic mice were unsuccessful to learn in CFC similar responses in mice that were not promoted to learn. Five proteins from ten proteins varied in failed learning were common to the subset of 11 important proteins in successful learning.

In third case, Gardiner and Cios (2015) created SOMs with mix of control and trisomic sets to analyze dissimilarities in learning. Same set of ten proteins were differentiating between t-CS-s and both c-CS-s and c-CS-m. They demonstrated the disparities in the levels of proteins differentiate successful learning from the absence of stimulation to learn. They stated that protein reactions when trisomic mice are unsuccessful to learn in CFC look like reactions in mice that not boosted to learn. The outcomes show that SOM can aid to diagnose protein anomalies in DS mice.

2.10.3 Linear SVM Method

Given labeled training data, SVM algorithm try to find a hyperplane that classifies the data points. In Linear SVM, the hyperplane divides a plane in two parts where in each class lay in either side.

Eicher and Sinha (2017) thought that the discrimination of healty and unhealthy mice based on their protein profiles was naturally related to classification difficulty instead of clustering dilemma. They stated that the decision of proteins which can discriminate two classes was needed. Furthermore, they claimed that classification techniques could give greater accuracy than clustering techniques as relabeling clusters might decrease the accuracy. In addition, accuracy quantified more effectively by applying quantitative procedures like cross validation, training and testing expectation rather than a visual way. Thus, they applied linear SVM for discriminating proteins. They analyzed the expression amount of 77 protein responses gathered from the nuclear cortex of normal and Ts65Dn mice, with and without memantine therapy and with and without CFC. They selected features to choose proteins that take a important role in each model. For each classification, feature selection was carried out in two stages. The result was discriminatory protein subsets for the two

classes. The first phase of feature selection utilized weight values of hyperplane. In the second step, the Wilcoxon rank sum test is applied. The efficiency of the classifiers was higher than classifiers using previous works. Also, the large part of distinctly identified features was also statistically important according to the Wilcoxon rank sum test which is a nonparametric statistical test and calculates the difference between two paired group.

The classification efficiency of the linear SVM algorithm was higher than the techniques applied in previous works. However, for resolving critical proteins for more than two classes as an input to Higuera et al. (2015), Eicher and Sinha (2017) combined classes to create new positive and negative classes. The outputs of combined class were not compared with the Higuera et al. (2015) efficiently. In order linear SVM was not to effective for multi-class of proteins, multiclass classification techniques were required.

2.10.4 Decision Tree, Random Forest, SVM Methods

Feng et al. (2017) decreased feature subset from 77 to 30 features by using AdaBoost method. AdaBoost converts a set of weak classifiers into a strong one in order to increase overall accuracy. In addition, they used Decision Tree, Random Forest and SVM classification techniques to differentiate normal and trisomic Ts65Dn mice. Decision Tree is a flowchart-like structure in which leaves represent class labels and branches show features to those class labels. The paths from root to leaf indicate classification rules. Random forest contains a large number of individual decision trees that create an ensemble. Each individual tree in the random forest splits out a class prediction. The class with the most votes becomes model prediction. SVM defines decision boundaries and separate different class memberships according to decision plane.

Feng et al. (2017) observed that the selected protein sets gave better classification outcomes. They also generated a combined Adaboost and Decision Tree feature

selection methods to determine the critical proteins . These proteins were relevant to the DS phenotype and biological pathways that could differentiated classes of DS mice and non-DS mice successfully. All proteins determined in their study were linked with known functions in the pathways of DS. Unsupervised PCA (Principal Component Analysis) analysis and hierarchical or agglomerative clustering analysis also proved that these proteins were important and associated with several symptoms of DS.

Nonetheless, they did not take into account control and Ts65Dn mice, with and without memantine therapy and with and without CFC stimulation subsets. They could only discriminate the control group from the trisomic group. Therefore, their work did not describe systematic procedure that was applied with Higuera et al. (2015) by examining the subgroups. Also, AdaBoost was a very proficient technique for solution of the two-class classification problem. Nonetheless, in going from two-class to multiclass classification, naive AdaBoost has restrainted the multiclass classification problem to multiple two-class problems.

2.11 Proposed Method

In our work, the different preprocessing steps, feature selection and classification techniques are applied in order to differentiate classes of mice more accurately. The flowchart of applied steps to protein expression datasets can be seen in Figure 2.11.

The datasets are preprocessed in filling missing values and normalization steps. 15 tissue samples that are three replicates of a five-point dilution series were obtained per mouse. Compared to previous works, the effect of dilution ratio is considered in this work and missing values are replaced with the average expression value of equivalent sample in same class. Rather than min-max normalization, Z score normalization which preserves range is done to inhibit the huge impact of proteins on classification. After the preprocessing step, the forward feature selection technique is applied to determine specific proteins for the DS. The multiclass classification

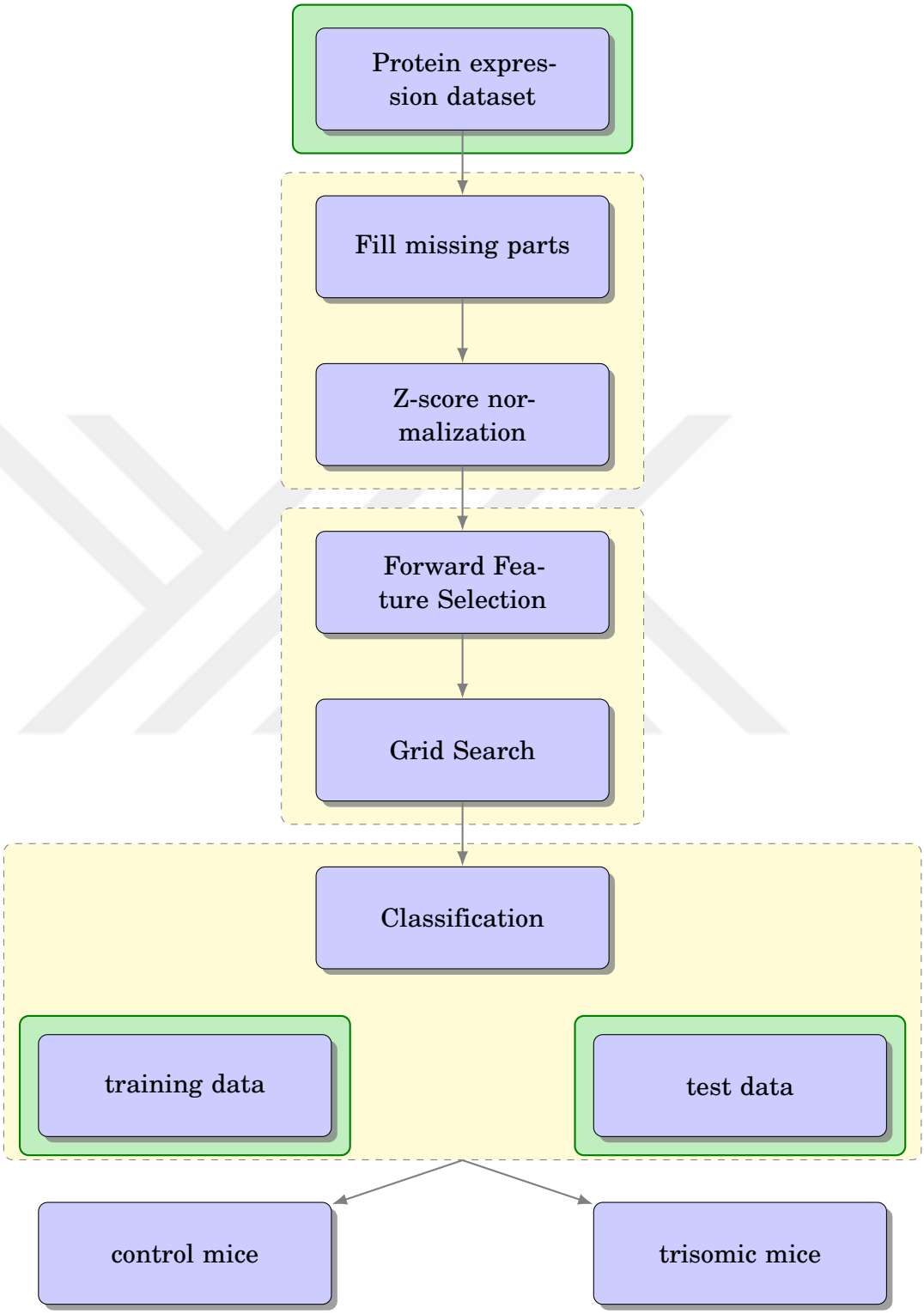

Figure 2.11: Flowchart of applied steps to protein expression datasets

algorithms are required to figure out which proteins are distinct when there are more than two classes. Thus, naive Bayes learner for multiclass classification is used in forward feature selection. After selecting features, grid search is done to determine appropriate parameters for classification techniques, DNN, gradient boosting tree, random forest and SVM. Thanks to these methods, control and trisomic mice are differentiated. The accuracy result of this work turned out to be higher than Feng et al. (2017) for all classification methods. The detailed analysis to determine critical proteins in successful learning, failed learning and rescued learning is also performed. DNN, gradient boosting tree, random forest and SVM classification methods are done after applying forward feature selection. The accuracy results of this work are higher than Higuera et al. (2015) for all classification methods.

3. DATASETS AND DATA PREPROCESSING

In this chapter, the datasets used for this thesis are described. In this thesis, five different datasets are used to differentiate mice based on:

- Learning Outcome (succesful, rescued with drug, failed),
- Drug (memantine, RO4938581),
- Age (young, old),
- Mice type (Ts65Dn, Tc1),
- Fractions of brain region (nuclear, cytosolic)

First dataset is obtained from University of California Irvine Machine Learning Repository (Dua and Graff, 2017). The other datasets are obtained from Prof. Gardiner at Colorado University (Globaldownsyndrome, 2019). The information on these datasets are explained in the following sections. The datasets are also presented along with the CD attached to this thesis. Later, the preprocessing steps for these datasets are explained.

3.1 Datasets

3.1.1 Dataset to Differentiate Mice for Learning Outcome

The data contains the expression profiles of 77 proteins obtained from the nuclear fraction of cortex. These 77 proteins have functions for brain structure and development. In this dataset, there are 38 control mice and 34 trisomic mice which are shown as c and t, respectively. Control mice are healthy mice and learn successfully. These mice are divided into groups based on context shock or drug mematine is ap-

plied or not. CS stands for mice that are exposed to context shock and SC represents the shock context only. In CS group, mice are exposured to shock and waited some minutes to learn from this shock. In SC group mice, it is not held so mice do not context shock to the environment. The letter m stands for drug memantine and s shows saline. Table 3.1 describes format of the dataset.

Table 3.1: Description of protein expression dataset.

Mice	**P1**	**P2**	**..**	**..**	**..**	**P77**	**Class**
mouse 1	0.504	0.747				1.676	c-CS-m
mouse 2	0.515	0.689				1.744	c-CS-m
mouse 3	0.509	0.730				1.926	c-CS-m
..							
mouse 72	0.303	0.461				1.371	t-SC-s

The rows show the individual mice. The columns show the expression amounts of 77 proteins that generated noticeable signals. The last column of each row shows the class of each mice. The value of protein expression levels in Table 3.1 corresponds the intensity of antigen-antibody binding in RPPA. It is an arbitrary unit that is a unit of measurement to show the ratio of substance to a predetermined reference measurement. High value in protein profiles indicates more protein expression in inspected tissue.

15 samples are extracted from each mouse, resulting in 1080 samples. Table A.1 in Appendix A shows the mouse ID and the first eleven protein expression levels and classes of two mice.

Table 3.2 shows the class of mice, the number of mice in the class and the learning outcome. As seen in Table 3.2, the dataset is divided into eight classes of mice based on the profiles of 77 proteins after training in CFC with and without injection of memantine.

Table 3.2: Classes, number and learning outcome of mice.

Class	Number of Mice	Learning Outcome
c−SC−s	9	No Learning
c−SC−m	10	No Learning
c−CS−s	9	Successful Learning
c−CS−m	10	Successful Learning
t−SC−s	9	No Learning
t−SC−m	9	No Learning
t−CS−s	7	Failed Learning
t−CS−m	9	Rescued Learning

3.1.2 Datasets to Differentiate Mice for Drugs

These datasets are based on two different drugs- memantine and RO4938581. The memantine drug dataset contains measurement from 9 mice (t−CS−m) and 15 measurements are taken from each mouse as explained previous subsection.

The RO4938581 drug dataset shows the normalized data from whole lysate hippocampus of Spanish Ts65Dn ten months old mice. Data treated with drug RO4938581 or saline as control group. Table 3.3 describes format of the RO4938581 dataset. Each sample is associated with three replicate spots of a five point dilution series. The dataset shows 91 protein profiles of 43 mice.

Table 3.3: Description of RO4938581 dataset.

Mice	P1	P2	P3	..	..	P91	Class
mouse 1	0.414	-	0.624			0.114	c-8581
mouse 2	0.450	-	0.658			0.116	c-8581
mouse 3	0.452	-	0.649			0.117	c-8581
..							
mouse *43*	0.967	0.391	0.855			0.862	t-s

Table 3.4 shows the class of mice, the number of mice in the class and the learning outcome. The dataset is divided into four classes of mice which are control group and Ts65Dn group mice treated with saline or RO4938581. There are 22 control mice and 21 trisomic mice shown as c and t, respectively. The 8581 stands for drug RO4938581 and s shows saline.

Table 3.4: Classes in RO4938581 dataset.

Class	Number of Mice	Learning Outcome
c−s	12	Successful Learning
c−8581	10	Successful Learning
t−s	11	No Learning
t−8581	10	Rescued Learning

3.1.3 Datasets to Differentiate Mice for Age

Table 3.5 shows the classes of young mice and old mice datasets, the number of mice in the corresponding classes.

Table 3.5: Classes in the young mice and old mice datasets.

Class	Number of Young Mice	Number of Old Mice
c−CB	6	5
c−CR	6	5
c−HP	6	5
t−CB	5	5
t−CR	5	5
t−HP	5	5

The dataset consists of young and old mice protein expression levels generated from cortex, cerebellum and hippocampus of the mice. The dataset shows protein expression profiles of control mice and trisomic mice that are shown as c and t, respectively. There are five trisomic and six control young mice. Old mice consist of five trisomic and five control mice. The letters CB, CR and HP stand for cerebellum, cortex and hippocampus, respectively.

The young mice dataset describes the information of 33 samples from five trisomic and six control mice at three different brain regions of 63 proteins. Table 3.6 describes format of the young mice dataset.

There are 20 measurements taken from each protein per sample: four replicates of a five-point dilution series that are stepwise dilution of protein. Therefore, for control mice, there are 18x20= 360 calculations, and for trisomic mice, there are 15x20 = 300 measurements.

Table 3.6: Description of young mice dataset.

Mice	P1	P2	P3	..	..	P63	Class
mouse 1	0.489	0.208	0.398			0.639	c-HP
mouse 2	0.456	0.197	0.387			0.632	c-HP
mouse 3	0.427	0.185	0.366			0.605	c-HP
..							
mouse *33*	0.461	0.246	0.448			0.590	t-HP

The old mice dataset contains the information of 89 protein expressions which are taken from five trisomic and five control mice across three brain regions. 15 measurements are taken from each protein per sample: three replicates of a five-point dilution series. There are 15x15=225 measurements for control and trisomic groups. Table 3.7 describes the format of old mice dataset.

Table 3.7: Description of old mice dataset.

Mice	P1	P2	P3	..	..	P89	Class
mouse 1	5.821	2.346	5.678			0.384	c-CR
mouse 2	6.011	2.256	5.423			0.437	c-CR
mouse 3	5.876	2.193	5.591			0.412	c-CR
..							
mouse *30*	-	-	-			0.443	c-CB

3.1.4 Datasets to Differentiate Mice for Mice Type

These datasets show the protein expression data from cortex, cerebellum and hippocampus. Data were taken from Tc1 mice, Ts65Dn mice and their controls.

Ts65Dn mice dataset that explained as young mice dataset in previous subsection. It includes the expression levels of 63 proteins obtained from cortex, cerebellum and hippocampus. There are six control mice and five trisomic Ts65Dn mice. There are 20 measurements obtained from each protein per sample: four replicates of a five-point dilution series.

Table 3.8 describes format of the Tc1 mice dataset. The data from Tc1 mice contains the expression levels of 90 proteins obtained from cortex, cerebellum and hippocampus of 8 moths old mice. In the dataset, there are seven control mice and seven

trisomic Tc1 mice. 20 samples are extracted from each mouse.

Table 3.8: Description of Tc1 mice dataset.

Mice	P1	P2	P3	..	..	P90	Class
mouse 1	0.279	0.477	0.259			-	t-CR
mouse 2	0.288	0.490	0.255			-	t-CR
mouse 3	0.290	0.472	0.251			-	t-CR
..							
mouse *42*	0.155	0.404	0.476			0.802	t-CB

Table 3.9 shows the classes of mice and the number of mice in the classes. The datasets are divided into six classes of mice which represent expression data across brain regions of control and Tc1 group mice.

Table 3.9: Classes in the Tc1 mice dataset.

Class	Number of Tc1 Mice
c–CB	7
c–CR	7
c–HP	7
t–CB	7
t–CR	7
t–HP	7

3.1.5 Datasets Used to Differentiate Mice for Fractions of Brain Region

This dataset shows the protein expression data from nuclear and cytosolic fraction of cortex. This dataset was obtained from Ts65Dn mice and control mice which were treated with memantine or saline. Table 3.10 describes format of the protein expression data from nuclear and cytosolic fraction of cortex.

Table 3.10: Description of protein expression dataset from fractions of cortex.

Mice	P1	P2	P3	..	..	P79	Class
mouse 1	1.0518	1.405	0.412			2.866	t-cyto-s
mouse 2	0.892	1.205	0.348			2.845	t-cyto-s
mouse 3	0.903	1.197	0.366			2.747	t-cyto-s
..							
mouse *72*	0.238	0.339	0.332			0.807	c-nuc-m

15 samples are gathered from each mouse. The dataset is partitioned into eight classes of mice which represent expression data at cytosolic and nuclear fraction

of cortex. There are control group and Ts65Dn group mice treated with saline or memantine. The dataset contains the expression levels of 79 proteins obtained from nuclear and cytosolic fraction of cortex.

Table 3.11 shows the class of mice and the number of mice in the class.

Table 3.11: Classes in the dataset from fractions of cortex.

Class	Number of Mice
c−cyto−s	9
c−cyto−m	10
c−nuc−s	9
c−nuc−m	10
t−cyto−s	7
t−cyto−m	10
t−nuc−s	7
t−nuc−m	10

3.2 Data Preprocessing

Data preprocessing is a critical issue for data mining (Alasadi and Bhaya, 2017). It prepares raw data for further processing and resolves real-world data problems that include incomplete data. These data likely contain many errors. Data preprocessing consist of cleaning, feature extraction and selection, instance selection, normalization, transformation steps. The output is the final training set.

Data processing step is important since there can be unrelated, unnecessary or unreliable data. Also, if raw data is used, knowledge exploration will be more challenging and require reasonable processing time. In this thesis, data preprocessing step consists of handling missing value, normalization of datasets and feature selection.

3.2.1 Handling Missing Value

For the datasets that are used in this thesis, a number of protein levels have missing values. Using the data in this from can result to misleading predictions for the unknowns. The missing points are substituted by the mean expression levels of the

equivalent sample of mice in the same class. For example, if a mouse is missing the first sample expression level, the missing value is substituted by the average expression value of first sample in the same class of mice.

This subsitution method used in this thesis is different from previous studies. In the previous studies, missing values were substituted with mean value of all protein expression in the same class of mice. 15 tissue samples that are three replicates of a five-point dilution series were obtained for each mouse. We considered the effect of dilution ratio and applied different calculations to handle missing values.

Table 3.12 presents example of the sample data that have missing values. Columns represent Sample ID, dilution ratio, replicate, nine protein expression levels and class of mice. The sample data in Table 3.12 belong to the same class; namely c-cs-m. P1 column shows the protein expression levels of protein one. Three expression levels in third dilution ratio of Sample 2 are missing.

Table 3.12: Missing value example of sample data.

Sample ID	Dilution	Replicate	P1	P2	P3	P4	P5	P6	P7	P8	P9	Class
2	1	1	0.2929	0.2455	0.7521	0.1686	0.2256	0.4729	0.9345	0.3337	1.7124	c-cs-m
2	1	2	0.2101	0.1824	0.5525	0.1026	0.1113	0.2928	0.5676	0.2621	1.4518	c-cs-m
2	1	3	0.2267	0.1691	0.5689	0.1195	0.1254		0.5367	0.2549	1.4360	c-cs-m
2	2	1	0.2297	0.1805	0.5454	0.1067	0.1180	0.3207	0.5341	0.2513	1.4653	c-cs-m
2	2	2	0.1582	0.1528	0.0920	0.0935	0.3504	0.5577	1.6326			c-cs-m
2	2	3	0.1636	0.1614	0.1027	0.1060	0.3225	0.5713	1.6571			c-cs-m
2	**3**	1			0.1004	0.1070	0.3852	0.5693	1.7325			c-cs-m
2	**3**	2			0.1038	0.0965	0.3693	0.6235	1.7903			c-cs-m
2	**3**	3			0.1016	0.0971	0.3663	0.6491	1.7366			c-cs-m
2	4	1	0.1573	0.1543	0.0966	0.0976	0.3591	0.6540	1.7204			c-cs-m
2	4	2	0.1619	0.1395	0.3882	0.0934	0.0856	0.3413	0.3641	0.1605	1.5170	c-cs-m
2	4	3	0.1812	0.1535	0.4158	0.0941	0.1030	0.3409	0.3980	0.1717	1.6041	c-cs-m
2	5	1	0.5418	0.6688	0.2448	0.3133	0.3048	0.1961	0.6129	1.0875	1.1994	c-cs-m
2	5	2	0.4974	0.6623	0.2244	0.3036	0.3115	0.1937	0.6034	1.1106	1.2441	c-cs-m
2	5	3	0.5797	0.7521	0.2571	0.3233	0.3233	0.2193	0.6616	1.2934	1.4180	c-cs-m
3	1	1	0.7747	0.9413	0.4267	0.2844	0.2640	0.2029	1.0663	1.5607	1.8201	c-cs-m
3	1	2	0.8614	0.9675	0.4441	0.3068	0.2728	0.2269	1.1725	1.6870	2.0388	c-cs-m
3	1	3	0.8646	0.9929	0.4409	0.3292	0.2714	0.2232	1.1505	1.7492	1.9095	c-cs-m
3	2	1	0.7926	0.9193	0.3471		0.2432	0.2030	0.9279	1.5316	1.8244	c-cs-m
3	2	2	0.8313	0.9490	0.3445		0.2528	0.2054	0.9774	1.5965	1.8408	c-cs-m
3	2	3	0.8607	1.0137	0.3715	0.3457	0.2604	0.2111	1.0260	1.8023	1.8809	c-cs-m
3	**3**	1	0.7191	0.8612	0.2793	0.2899	0.2353	0.1913	0.9034	1.4744	1.7711	c-cs-m
3	**3**	2	0.7337	0.8381	0.2787	0.2898	0.2303	0.1995	0.9204	1.5225	1.6598	c-cs-m
3	**3**	3	0.7977	0.9378	0.3147	0.3346	0.2403	0.2096	0.9759	1.6961	1.7955	c-cs-m
3	4	1	0.7014	0.8735	0.2847	0.3040	0.2083	0.1604	0.7043	1.5029	1.7138	c-cs-m
3	4	2	0.7584	0.9005	0.2854	0.3308	0.2190	0.1689	0.7782	1.6034	1.7766	c-cs-m
3	4	3	0.8095	0.9091	0.3037	0.3424	0.2155		0.7424	1.7950	1.7136	c-cs-m
3	5	1	0.7012	0.8056	0.2953	0.3685	0.2395	0.1935	0.8503	1.4730	1.5208	c-cs-m
3	5	2	0.7097		0.2998	0.3808	0.2329	0.2015	0.8930	1.6801	1.5318	c-cs-m
3	5	3	0.7326		0.3089	0.3939	0.2430	0.1934	0.8787	1.7945	1.5427	c-cs-m

Sample 2, Dilution ratio 3

1. Replicate - missing value
2. Replicate - missing value
3. Replicate - missing value

Sample 3, Dilution ratio 3

1. Replicate - 0.719190141
2. Replicate - 0.733772061
3. Replicate - 0.79775641

When these missing values are handled, the dilution rate is considered. Sample 2 and 3 belong to the same class: c-cs-m. So, the missing values will be the average of third dilution levels of Sample 2 and 3. Thus, the missing values of sample 2 will be the average of 0.719190141, 0.733772061 and 0.79775641. The value is 0.7502395373.

The previous works did not consider the dilution ratio and took consideration of expression levels in the same class. The result is 0.5573819498. There is a big difference between missing value result of this work and previous works.

Table 3.13 shows the completed representation of Table 3.12. All missing values in Table 3.12 are replaced by the mean expression levels of equivalent sample in the same dilution ratio and class of mice.

Table 3.13: Complete representation of example sample data.

Sample ID	Dilution	Replicate	P1	P2	P3	P4	P5	P6	P7	P8	P9	Class
2	1	1	0.2929	0.2455	0.7521	0.1686	0.2256	0.4729	0.9345	0.3337	1.7124	c-cs-m
2	1	2	0.2101	0.1824	0.5525	0.1026	0.1113	0.2928	0.5676	0.2621	1.4518	c-cs-m
2	1	3	0.2267	0.1691	0.5689	0.1195	0.1254	0.2837	0.5367	0.2549	1.4360	c-cs-m
2	2	1	0.2297	0.1805	0.5454	0.1067	0.1180	0.3207	0.5341	0.2513	1.4653	c-cs-m
2	2	2	0.1582	0.1528	0.0920	0.0935	0.3504	0.5577	1.6326	0.1295	0.1753	c-cs-m
2	2	3	0.1636	0.1614	0.1027	0.1060	0.3225	0.5713	1.6571	0.1295	0.1753	c-cs-m
2	**3**	1	0.7502	0.8790	0.1004	0.1070	0.3852	0.5693	1.7325	0.1564	0.1742	c-cs-m
2	**3**	2	0.7502	0.8790	0.1038	0.0965	0.3693	0.6235	1.7903	0.1564	0.1742	c-cs-m
2	**3**	3	0.7502	0.8790	0.1016	0.0971	0.3663	0.6491	1.7366	0.1564	0.1742	c-cs-m
2	4	1	0.1573	0.1543	0.0966	0.0976	0.3591	0.6540	1.7204	0.1113	0.1665	c-cs-m
2	4	2	0.1619	0.1395	0.3882	0.0934	0.0856	0.3413	0.3641	0.1605	1.5170	c-cs-m
2	4	3	0.1812	0.1535	0.4158	0.0941	0.1030	0.3409	0.3980	0.1717	1.6041	c-cs-m
2	5	1	0.5418	0.6688	0.2448	0.3133	0.3048	0.1961	0.6129	1.0875	1.1994	c-cs-m
2	5	2	0.4974	0.6623	0.2244	0.3036	0.3115	0.1937	0.6034	1.1106	1.2441	c-cs-m
2	5	3	0.5797	0.7521	0.2571	0.3233	0.3233	0.2193	0.6616	1.2934	1.4180	c-cs-m
3	1	1	0.7747	0.9413	0.4267	0.2844	0.2640	0.2029	1.0663	1.5607	1.8201	c-cs-m
3	1	2	0.8614	0.9675	0.4441	0.3068	0.2728	0.2269	1.1725	1.6870	2.0388	c-cs-m
3	1	3	0.8646	0.9929	0.4409	0.3292	0.2714	0.2232	1.1505	1.7492	1.9095	c-cs-m
3	2	1	0.7926	0.9193	0.3471	0.1630	0.2432	0.2030	0.9279	1.5316	1.8244	c-cs-m
3	2	2	0.8313	0.9490	0.3445	0.1630	0.2528	0.2054	0.9774	1.5965	1.8408	c-cs-m
3	2	3	0.8607	1.0137	0.3715	0.3457	0.2604	0.2111	1.0260	1.8023	1.8809	c-cs-m
3	**3**	1	0.7191	0.8612	0.2793	0.2899	0.2353	0.1913	0.9034	1.4744	1.7711	c-cs-m
3	**3**	2	0.7337	0.8381	0.2787	0.2898	0.2303	0.1995	0.9204	1.5225	1.6598	c-cs-m
3	**3**	3	0.7977	0.9378	0.3147	0.3346	0.2403	0.2096	0.9759	1.6961	1.7955	c-cs-m
3	4	1	0.7014	0.8735	0.2847	0.3040	0.2083	0.1604	0.7043	1.5029	1.7138	c-cs-m
3	4	2	0.7584	0.9005	0.2854	0.3308	0.2190	0.1689	0.7782	1.6034	1.7766	c-cs-m
3	4	3	0.8095	0.9091	0.3037	0.3424	0.2155	0.3331	0.7424	1.7950	1.7136	c-cs-m
3	5	1	0.7012	0.8056	0.2953	0.3685	0.2395	0.1935	0.8503	1.4730	1.5208	c-cs-m
3	5	2	0.7097	0.7222	0.2998	0.3808	0.2329	0.2015	0.8930	1.6801	1.5318	c-cs-m
3	5	3	0.7326	0.7222	0.3089	0.3939	0.2430	0.1934	0.8787	1.7945	1.5427	c-cs-m

3.2.2 Normalization

Normalization is the process of discarding methodological biases from the data. Bias may happen due to many factors, including adjustments in temperature over the time of experiment, changes in conditions of sample processing, instrument calibrations.

In this thesis, all evaluations are normalized with Z-score to ward off proteins with higher amounts affect on the classification outcome incorrectly. Since Z-score normalization protects range (maximum and minimum), Z-score normalization is used in this thesis rather than min-max normalization which was applied in Higuera et al. (2015). Min-max normalization transforms linearly x to $y = (x - min)/(max - min)$, where min and max are the minimum and maximum values in X, where X is the set of x values that are observed. The minimum value in X is correspond to 0 and the maximum value in X is correspond to 1. Therefore, the entire range of X values from min to max are mapped to the range 0 to 1.

With Z-score normalization as stated in (3.1), mean of the scores is subtracted from each score and then divided into the standard deviation (Abdi and Lynne, 2010). Z-score is applied to avoid higher impact of proteins on the classification outcome.

$$Z = \frac{x - \mu}{\sigma}, \tag{3.1}$$

Z-score normalization preserves range (maximum and minimum) and introduces the dispersion of the series (standard deviation / variance). If data follows a gaussian distribution, the comparison between series will be easier. Z-scores allow for a very straight elimination of systematic errors.

4. FEATURE SELECTION

In this chapter, before developing a classification model, dimensionality reduction which is an important step for comprehending the knowledge about the class is applied. Dimensionality reduction decreases the amount of features for diagnosis of the vital variables. It has the effect of declining the computational cost. For dimensionality reduction, feature selection and feature extraction methods have been used. Feature selection chooses a subset of features, while feature extraction produces a new feature set from original features (Khalid, Khalil and Nasreen, 2014; Kaushik, 2016).

For feature selection, the forward feature selection method is used in this thesis. The reason for using feature selection method rather than feature extraction method is the determination of exact protein subset. In feature extraction method like PCA, the dataset is compressed onto a lower-dimensional feature subspace in order to obtain most of the relevant information. By this way, the patterns in data can be identified based on the correlation between features. However, the determination of protein subset which differentiates mice more accurately is needed in this work, thus feature selection is used.

Three classes of feature selection algorithms are filter methods, wrapper methods and embedded methods. These feature selection methods improve accuracy and efficiency of classifier methods. Generally, a feature subset selection algorithm involves a feature evaluation criterion and a search algorithm. The evaluation criterion determines the feature subset capacity to discriminate one class from another. The search algorithm explores the potential solution space. With filter selection method, the feature selection is independent of the classifier. Also, a

statistical measure is used to assign a score to each feature. The features are ranked by the score. On the other hand, wrapper method selects features using classifier and obtains better performances. Thus, one of the wrapper feature selection method, forward feature selection, is used in this work.

Feature selection is the method of choosing a subset of related features in model structure (Rawale, 2018). It is one of the main concepts in machine learning. It greatly affects the performance of model as the features have a great effect on the function. Having unrelated features diminishes the accuracy of the models and makes model learn based on unrelated features. Some benefits of applying feature selection method before modeling data are accuracy improvement, reduction of overfitting and training time. The accuracy of a model is enhanced if the exact subset is selected. Less relevant data results in less chance to make decisions.

The forward feature selection is the heuristic procedure. It discovers the ideal feature subset by repetitively choosing features depending on the classifier efficiency. It starts with an empty feature subset and appends one feature for each spin. This one feature is removed from the set of all features which are not in the feature subset. Then, it is added to the set and set gives highest classifier performance. This procedure is reiterated until the desired number of features are appended. Forward feature selection does not inspect all possible subsets and does not ensure to obtain the optimal subset. However, it decreases the search time when compared to exhaustive feature selection (ScienceDirect, 2014).

The target of feature selection is to choose a subset of the whole input features set. The subset can forecast the output Y with accuracy relative to the complete input set X, and with great decrease in the computational cost. The forward feature selection method starts by evaluating all feature subsets that contain one input. In another words, it begins by quantifying the Leave-One-Out Cross Validation (LOOCV) error of the one element subsets, $X1$, $X2$, ..., XM, where M is the input dimensionality; so that we can obtain the best individual feature, $X1$. The algorithm of forward feature

selection is shown in Figure 4.1

1. Gather training data set from the explicit scope.
2. Mix up the data set.
3. Split it into N partitions.
4. For each partition ($i = 0, 1, ..., N - 1$)
   - Let Outer Trainset(i) = all partitions except i.
   - Let Outer Testset(i) = the i'th partition
   - Let Inner Train(i) = randomly chosen 70% of the Outer Trainset(i).
   - Let Inner Test(i) = the remaining 30% of the Outer Trainset(i).
   - For $j = 0, 1, ..., m$
     Investigate for the best feature set with j components, fsij.using leave-one-out on Inner Train(i)
     Let Inner TestScore(ij) = RMS score of fsij on Inner Test(i).
     End loop of (j).
   - Select the fsij with the best inner test score.
   - Let Outer Score(i) = RMS score of the selected feature set on Outer Testset(i)
5. Return the mean Outer Score.

Figure 4.1: Forward feature selection algorithm

In our study, the features are proteins and the class is the class of mice such as c-CS-m, t-SC-s (trisomic- Shock Context- saline).

The forward feature selection is applied using the Knostanz Information Miner (KNIME) as shown in Figure 4.2 (Berthold et al., 2009). It is the open source software for developing data science. KNIME Analytics Platform provides to comprehend data and shape data science workflows.

Inside the search loop, the dataset is partitioned into a learning set (70%) and a validation set (30%). Learning set is used for the development of the model in the selection of the variables. Validation set measures an error rate approximation. Naive Bayes learner that is efficient in multi classification problem is used for the learning process. Bayes theorem as stated in (4.1) offers a solution of measuring the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier undertakes that the impact of the predictor (x) value on a given class (c) is independent of the values of other predictors. This presumption is known as class conditional independence (Chen et al., 2009).

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \tag{4.1}$$

where $P(c|x)$ is the target class posterior probability, $P(c)$ is the class prior probability, $P(x|c)$ is the possibility that is the chance of class predictor, $P(x)$ is the predictor prior probability.

Despite the underlying assumption of conditional independence, naive Bayes executes well with more-than-two-classes problem. In previous works, the implemented algorithms suffered from an effective multiclass classification technique. In this thesis, this deficiency is eliminated with naive Bayes algorithm in forward feature selection method.

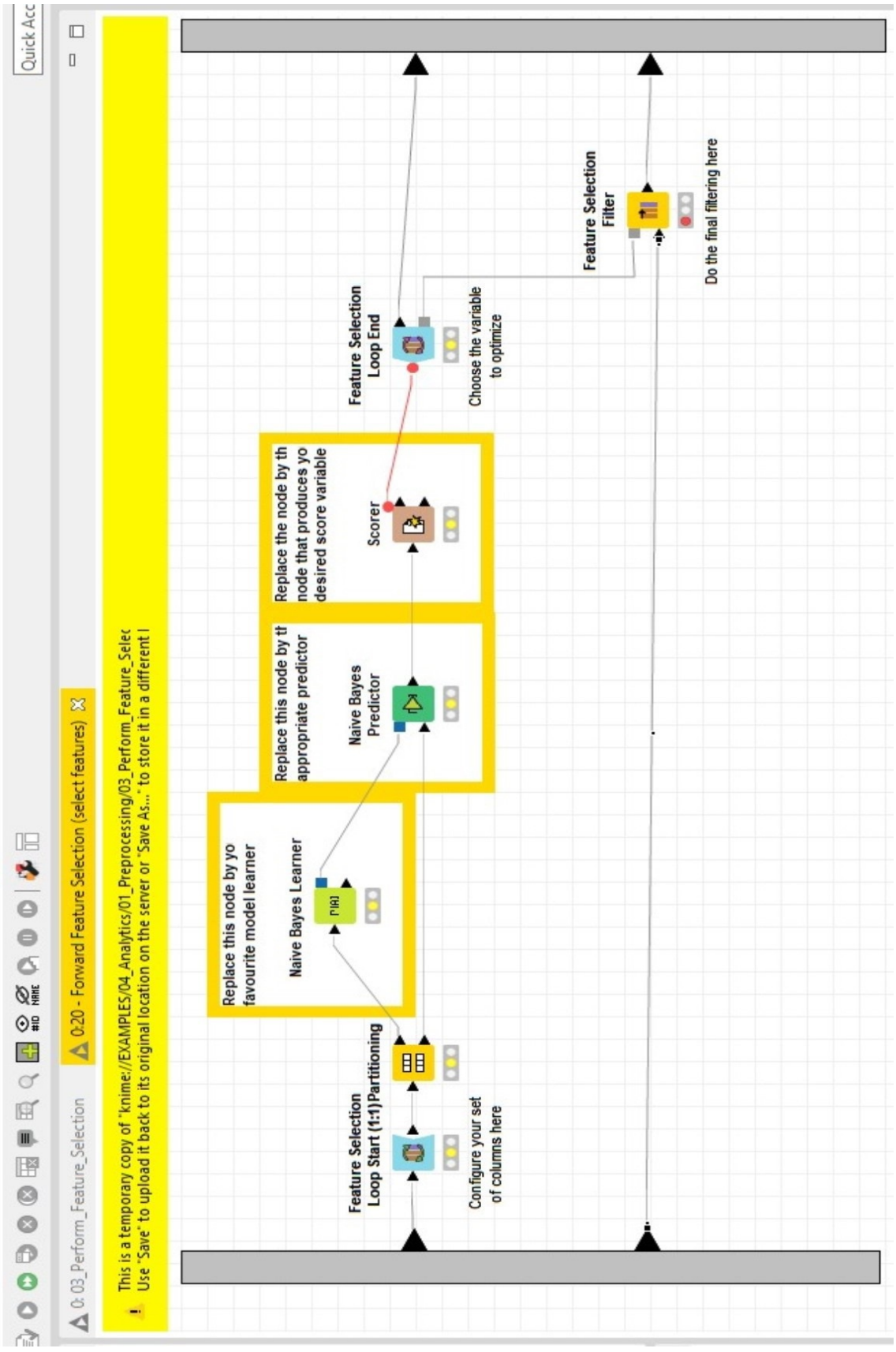

Figure 4.2: KNIME forward feature selection workflow (Berthold et al., 2009)

5. CLASSIFICATION METHODS

In this chapter, classification methods are applied for the division of mice to their groups, such as c-SC-m, t-CS-s. Four classification methods, DNN, gradient boosted tree, random forest and SVM are carried out. These classification methods are implemented by using Python and Scikit Learn package (Hao and Ho, 2019). For selecting the most suitable parameters for classification methods, grid search method (Bergstra et al., 2011) is applied.

Grid search is the technique of implementing hyper-parameter tuning for defining the ideal model. This is vital as the achievement of the whole procedure depends on the hyper-parameter values. The hyperparameter value has to be determined before the start of learning process. (For example, the number of hidden layers in NN, k (number of nearest neighbours) in k -Nearest Neighbors) Grid-search is used to discover the ideal hyperparameters of a model which produces the most true predictions. GridSearchCV of the sklearn library is utilized as shown in Figure 5.1.

```
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVR
gsc = GridSearchCV(
   estimator=SVR (kernel='rbf'),
   param_grid=
   'C': [0.1, 1, 100, 1000],
   'epsilon': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10],
   'gamma': [0.0001, 0.001, 0.005, 0.1, 1, 3, 5] ,
   cv=5, scoring='neg_mean_squared_error', verbose=0, n_jobs=-1)
grid_result = gsc.fit (X, y)
```

Figure 5.1: Hyper parameter tuning using GridSearchCV

First, it is needed to add GridSearchCV from the sklearn library which is a machine learning library for python. The parameter of GridSearchCV predictor needs the model for tuning hyper parameter. The rbf kernel of the Support Vector Regression model (SVR) is used. The param_grid parameter entails the range of parameter values. The most important parameters needed when operating with SVR model are c, γ and ε. A list of selected values should be provided to each hyper parameter. These values can be changed to observe which value gives higher performance. A cross validation procedure is done in order to identify the hyperparameter value set.

In our works, different classification algorithms are applied for different datasets. The parameters of these algorithms are selected using grid search method. The selected parameter range for DNN, gradient boosting, random forest and SVM are explained below.

The parameters of DNN are activation, hidden_layer_sizes, learning_rate_init and max_iter.

Example of DNN function : MLPClassifier (activation= 'relu', hidden_layer_sizes= 16, learning_rate_init= 0.01, max_iter= 160)

Grid search selects parameters for DNN. The user gives range of parameters and algorithm selects parameters that show best fit to function. The parameter range for DNN is shown below:

tuned_parameters = ['activation': ['identity', 'logistic', 'tanh', 'relu'], 'max_iter': [40,80,120,160,200], 'learning_rate_init': [0.01,0.05,0.1,0.5], 'hidden_layer_sizes': [8,12,16,20]]

Activation parameter represents activation function for the hidden layer. It can be identity, logistic, tanh orrelu. Identity is the no operation activation and returns f(x) = x. Logistic is the logistic sigmoid function and returns f(x) = 1 / (1 + exp(-x)). Tanh is the hyperbolic tan function and returns f(x) = tanh(x). Relu is the rectified

linear unit function and returns f(x) = max(0, x). The gradient of the tanh function is steeper as compared to the logistic sigmoid function. The tanh is preferred over the sigmoid function. It is zero centered and the gradients are not restricted to move in a one direction. The advantage of the ReLU function over other activation functions is that it does not activate all neurons at the same time. Thus, the ReLU function is more computationally efficient when compared to the sigmoid and tanh function.

Learning rate adjusts weight. Maximum number of iterations shows the number of iterations.

The parameters of random forest are max_depth, max_features, min_samples_split, min_samples_leaf, bootstrap, criterion.

Example of random forest function : RandomForestClassifier (n_estimators=100, criterion = 'gini', max_depth = None, min_samples_leaf = 1, max_features= 3, min_samples_split= 2, random_state=42)

The parameter range for random forest algorithm is shown below:

tuned_parameters = "max_depth": [3, None], "max_features": sp_randint(1, 11), "min_samples_split": sp_randint(2, 11), "min_samples_leaf": sp_randint(1, 11), "bootstrap": [True, False], "criterion": ["gini", "entropy"]

The parameter of maximum depth is the depth of tree. It controls over-fitting. Higher depth makes model learn relations very specific. If it is none, the nodes are expanded until all leaves contain less than min_samples_split samples. The parameter of min_samples_split shows the minimum number of samples required for splitting. The parameter of max_features represents the number of features are considered for the best split. Generally, square root of the total number of features works efficiently. The parameter of min_samples_leaf is the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of left and right branches. The

parameter of bootstrap shows the bootstrap samples when building trees. If it is false, the whole dataset is used to build each tree. The parameter of criterion is the quality measure of a split. Gini and entropy are Gini impurity and information gain, respectively.

The parameters of random forest are C, kernel and gamma.

Example of SVM function : SVC (C=10, kernel ="rbf", gamma=0.3)

The parameter range for SVM is shown below:

tuned_parameters = ['kernel': ['rbf', 'linear', 'poly'], 'gamma': [1e-3, 1e-4], 'epsilon': [0.0001,0.0005,0.001,0.005,0.01,0.05,0.1,0.5,1,5,10] ,'C': [1, 10, 100, 1000]]

The parameter of kernel shows the kernel type used in the algorithm. The gamma parameter defines kernel trick to handle nonlinear classification. When gamma is very small, the model is constrained to not capture the shape of the data. A small gamma causes low bias and high variance. A large gamma causes higher bias and low variance. The support vectors are the instances across the margin and the samples being penalized. The value of epsilon defines a margin of tolerance where no penalty is given to errors. The epsilon value which is larger than the range of the target causes a bad result. Epsilon must be chosen to reflect the data. It affects smoothness of the SVM's response. The C parameter is a trade off between correct classification of data and maximization of margin. A large C causes low bias and high variance. A smaller margin is accepted if the decision function is better at classifying all training points correctly. A small C causes higher bias and lower variance. It causes larger margin at the cost of training accuracy.

The parameters of gradient boosting classifier are n_estimators, min_samples_split, random_state, learning_rate.

Example of gradient boosting function : GradientBoostingClassifier (n_estimators=50,

min_samples_split= 3, random_state=42, learning_rate =0.5)

The parameter range for gradient boosting algorithm is shown below:

tuned_parameters = "n_estimators" : [50,100,200], "min_samples_split" : [2,3,5], "learning_rate" : [0.05,0.1,0.25,0.5,1.0]

Learning rate controls the magnitude of change in the estimates and determines the impact of each tree on the final outcome. Lower values make the model robust to the specific characteristics of tree. With lower values, higher number of trees is required for all relations and so it causes computationally expensive operation. The parameter of n_estimators shows the number of sequential trees to be modeled. At higher number of trees, algorithm is more roboust. However, it can be overfit at one point. Thus, cross-validation is required. The parameter of min_samples_split defines the minimum number of samples required in a node to be considered for splitting.

Also, for developing robust and reliable classification model, 5-fold cross validation is used. Cross-validation is a statistical process utilized to approximate the proficiency of machine learning models (Neale, 2019).

In k-fold cross validation (Wong, 2015), the data is splitted into k subsets. Only one of these subsets is utilized as the test set and the others are used as a training set at each time. This procedure is reiterated k times. The error approximation is the mean of all k trials to obtain total performance. This significantly diminishes bias because of utilizing most of the data for fitting. It also decreases variance as most of the data is also being utilized in validation set. The general methodology of $k-$ fold cross validation is given in Figure 5.2.

1. Mix the dataset randomly
2. Divide the dataset into k groups
3. For each particular group:
    - Take the group as test data set or hold out
    - Take the remaining groups as a training data set
    - Fit a model on the training set and valuate it on the test set
    - Keep the evaluation score and eliminate the model
    - Sum up the model skill using the model evaluation scores sample

Figure 5.2: Methodology of $k-$fold cross validation

Each observation is nominated to a group individually and remains in that group for the process time. This indicates that each sample is taken the chance to be utilized in the hold out set. Also, it is utilized to train the model $k-1$ times.

In the rest of this chapter, the four types of classification methods that are used in this thesis are described.

5.1 Deep Neural Network

Neural networks (NN) are developed to identify patterns. NN resemble to human brain as they gather knowledge through learning. Then, it constitutes artificial neurons which are equivalent to neurons in a brain. Each connection between neurons passes a signal to another neuron and can increase or decrease the power of the signal. As shown in Figure 5.3 (Huang, 2018), NN have three layers-input, hidden and output. NN analyze sensory data and assist to classify or cluster.

The perceptron performs on the following simple steps:

1. All inputs ($x\_1, x\_2, ..., x\_n$) are multiplied with their weights ($w\_1, w\_2, ..., w\_n$)
2. All multiplied values are added and called as weighted sum
3. The weighted sum is applied to the correct activation function

Weights demonstrate the power of the particular node. The activation functions are practiced to project the input between the required values like (0, 1) or (-1, 1).

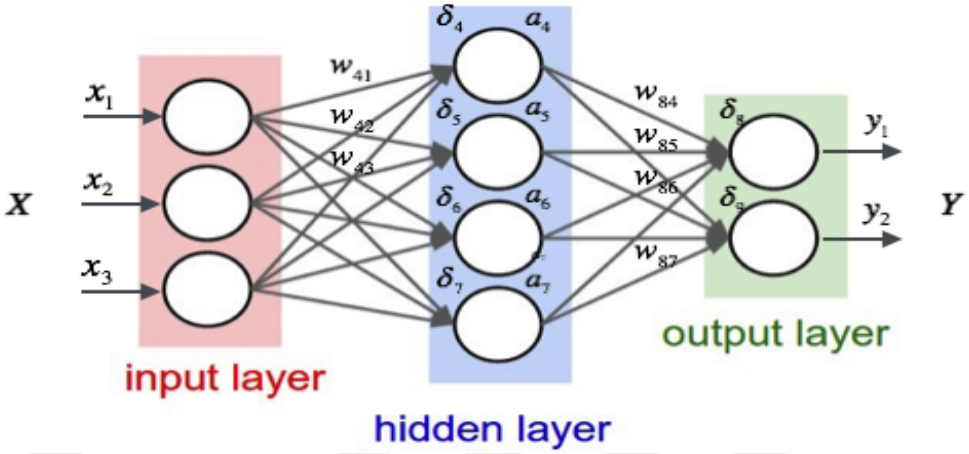

Figure 5.3: Neural network representation (Huang, 2018)

Perceptron is a single layer neural network as shown in Figure 5.4 below. (Rao, AS.; Avadhani, PS. and Chaudhuri, NB., 2016)

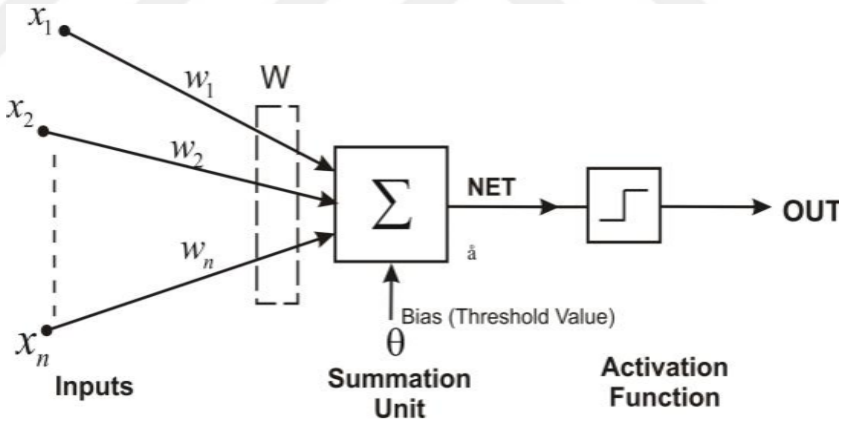

Figure 5.4: Percepton model (Rao, AS., Avadhani, PS. and Chaudhuri, NB., 2016).

DNN is type of a neural network with multiple layers between the input and output layers (Wang et al., 2017). A representation of a DNN is given in Figure 5.5 below. In order convert input into output, DNN tries to determine whether the relationship is linear or not. The network proceeds in the layers computing the probability of each output (Lecun, Bengio and Hinton, 2015).

A multi-layer perceptron utilized in this work is a subset of DNN (Pereira, 2006). It

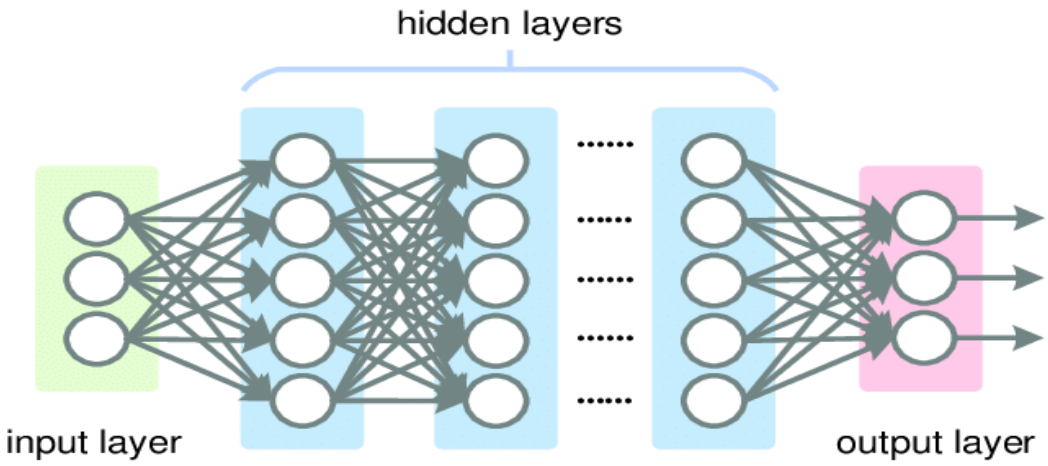

Figure 5.5: Deep neural network representation (Wang et al., 2017)

is the original perceptron model introduced by Rosenblatt (1958). It is also called as Feed Forward Neural Network (FFNN).

MLP (Multi Layer Perceptron) learning is the procedure to adjust the weights of connections. It obtains a smallest variation between the desired response and the network response. The commonly used algorithm is Back- Propagation Algorithm. The network outputs are measured by the following equations:

$$y_i = f(\sum_{k=1}^{n} w_{k,j}^{n} h_n^k), \tag{5.1}$$

$$Y = (y_1, ..., y_j, ..., y_{n+1}) = F(W, X) \tag{5.2}$$

where $w_{k,j}$, is the weight between the neuron k in the hidden layer i and the neuron j in the hidden layer i, n_i is the number of the neurons in the ith hidden layer. h_n^k represents neurons in the hidden layers and the output of h_i^j can be computed as follows:

$$h_i^j = f(\sum_{k=1}^{n_{i-1}} w_{k,j}^{i-1} h_{i-1}^k)\ i = 2, 3..., N\ and\ j = 1, 2, ..., n_i \tag{5.3}$$

where Y is the output layer vector, F is the transfer function.

The matrix of weights W is desribed as

$$W = [W^0, ..., W^j, ..., W^n] \tag{5.4}$$

$$W^i = (w^i_{j,k}) \tag{5.5}$$

where $0 <= i <= n, 1 <= j <= n_{i+1}\ and\ 1 <= k <= n_i$

A multi-layer perceptron is subset of DNN as shown in Figure 5.6.

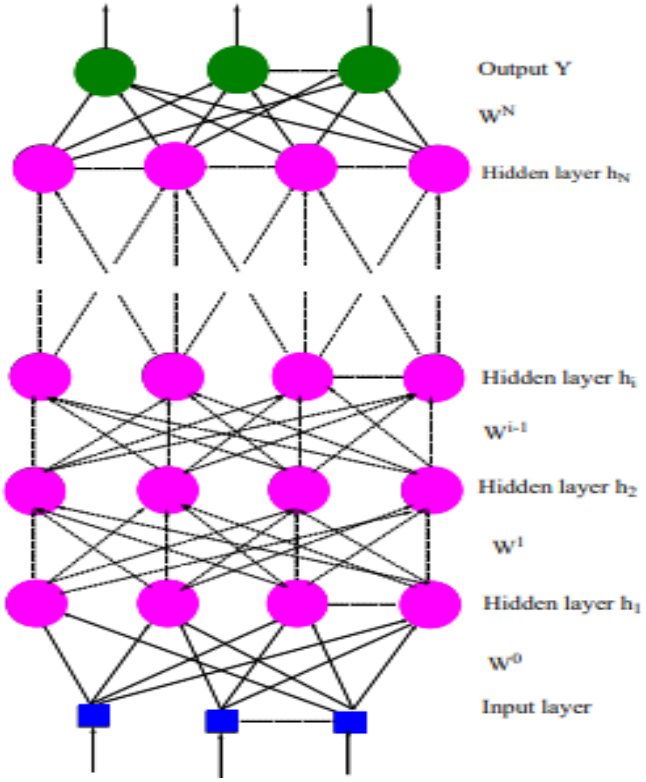

Figure 5.6: Multi-layer perceptron model (Pereira, 2006)

5.2 Gradient Boosted Tree

The fundamental reasons of variation in actual and predicted values can be noise, bias and variance. Ensemble assists to diminish these factors except noise since it is an irreducible error. It assembles predictors which stick together to give a final prediction.

Boosting is an ensemble method that the predictors are not produced separately, but successively. It transforms weak learners to a strong learner by boosting mislabeled

data with higher weight. Therefore, the subsamples of data have an unequal chance of emerging in subsequent models. The ones with the highest error are seen most (Shubham, 2018).

The idea on the observation by Leo Breiman is that boosting can be translated as an optimization algorithm on a proper cost function (Breiman, 1997). Regression gradient boosting algorithms were consequently built by Jerome H. Friedman (2001, 2002) and Mason et al. (1999). These two papers proposed the concept of boosting algorithms as repetitive functional gradient descent algorithms. This means that algorithms revise a cost function by repetitively selecting a function (weak hypothesis) that directs in the negative gradient direction.

Figure 5.7 (Johansson, 1995) shows the loss function minimization in gradient boosting algorithm when submodels are ensembled gradually. Gradient boosting adds submodels incrementally to minimize a loss function.

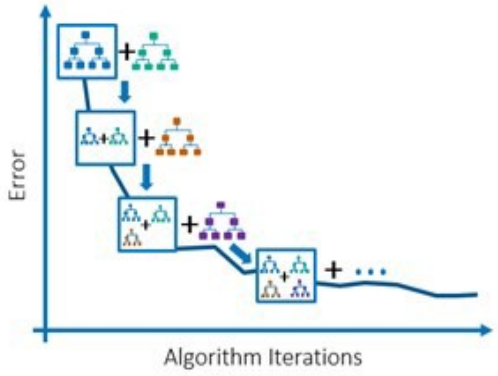

Figure 5.7: Minimization of a loss function in gradient boosting (Johansson, 1995)

Gradient boosting is a methodology of machine learning for classification and regression problems. It produces a forecast model in the form of a weak prediction models ensemble, typically decision tree (Natekin and Knoll, 2013). In pseudocode (Hastie et al., 2009), the method of generic gradient boosting is shown in Figure 5.8.

At each stage, the decision tree $h_m(x)$ that is base learner is choosen to minimize a loss function L given the current model $F_{m-1}(x)$, where m is the iteration number,

Input: Training set $(x_i, y_i)_{i=1}^n$, differentiable loss function $L(y, F(x))$,$L(y, F(x))$, number of iterations M.

Algorithm:

1. Start model with a constant value

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma). \tag{5.6}$$

2. For m = 1 to M:

- Calculate pseudo residuals

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n. \tag{5.7}$$

- Adjust a base learner $h_m(x)$ to pseudo-residuals, i.e. train it exploiting the training set

$$\{(x_i, r_{im})\}_{i=1}^{n} \tag{5.8}$$

- Calculate multiplier $\gamma_m \gamma_m$ by deciding the one dimensional optimization problem

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right). \tag{5.9}$$

- Update the model

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{5.10}$$

3. Output $F_M(x)$.

Figure 5.8: Methodology of gradient boosting algorithm

$F_m(x)$ is the model and γ is the rate of learning.

5.3 Support Vector Machines

SVM is a supervised machine learning classification technique that uses a d-dimensional Euclidean space data set (Cortes and Vapnik, 1995). The number of d represents the features quantity in the data set. SVM discovers an optimal $(d-1)$ dimensional hyperplane to split the data by class. Figure 5.9 demonstrates the possible hyperplanes. These hyperplanes discriminate classes. The range between the hyperplane and the nearest data point from each part of the hyperplane is known as the margin. In order to classify new data accurately, the distance between the hyperplane and data must be larger.

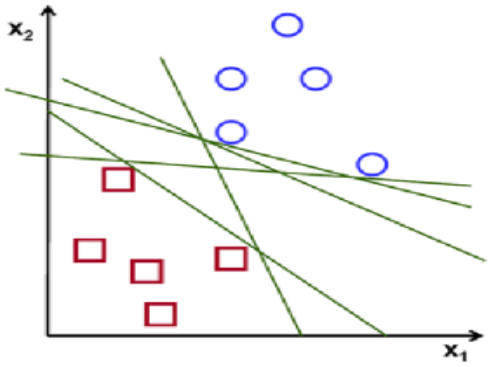

Figure 5.9: Possible hyperplanes in SVM (Cortes and Vapnik, 1995)

Figure 5.10 shows the optimal hyperplane in SVM.

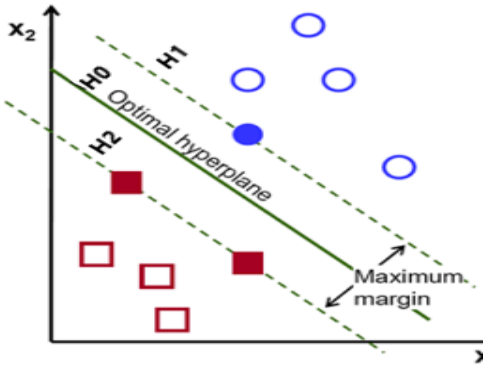

Figure 5.10: Optimal hyperplane in SVM.

As seen in Figure 5.10, the H1 and H2 are the two hyperplanes. If we are given a training dataset of n points of the form (x_1,y_1),.....,(x_n,y_n), y_i can be either 1 or -1, each indicating the class to which the point x_i belongs.

The two hyperplanes separate these two classes of data and can be described by the equations:

$$H1 : w * x_i + b = +1 \tag{5.6}$$

$$H2 : w * x_i + b = -1 \tag{5.7}$$

The plane H0 is the median between H1 and H2, where $w * x_i + b = 0$ in which b is bias, w is a weight vector and x is input vector.

5.4 Random Forest

Random forest is created from many decision trees that are selected from a random subset of training set. The representation of random forest is shown as Figure 5.11 below (Koehrsen, 2017).

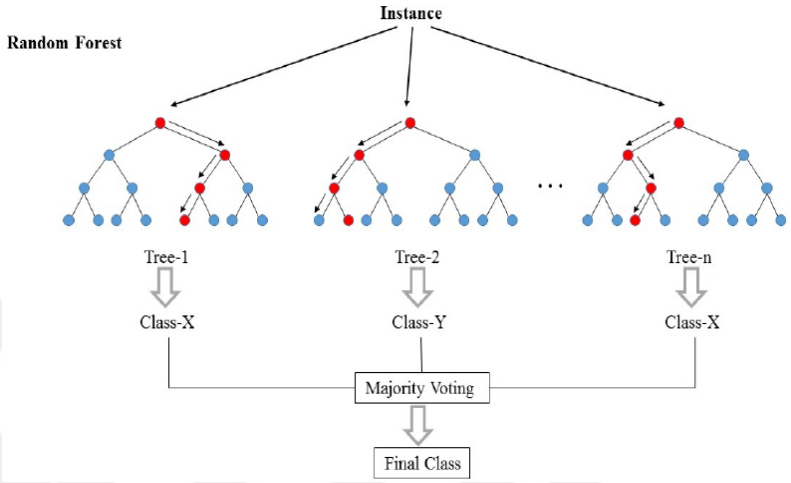

Figure 5.11: Representation of random forest (Koehrsen, 2017)

This method builds random forest by joining a great number of decision trees. The output is the class that is the classes mode or the individual tree mean prediction (Breiman, 2001). Random Forest is a strong learner and is built as an ensemble of decision trees. The decision trees are weak learners to perform different tasks such as classification and regression. Random Forest pseudocode is given in Figure 5.12.

1. Choose K features randomly from m features where $k << m$.
2. Measure the node d using the best split point among the K features.
3. Divide the node into daughter nodes based on the best split.
4. Reiterate the a to c steps until number of nodes has been obtained.
5. Construct forest by reiterating steps a to d for n number times to create n number of trees.

Figure 5.12: Random forest pseudocode (Koehrsen, 2017)

Random forest is tuned with *ntree* and *ntry* parameters to obtain optimized forest structure. The parameter *ntree* indicates how many trees are to be produced to generate the random forest. The parameter ntry indicates the number of variables that will be taken into account at any time in deciding how to divide the dataset.

The methodology of random forest classification is described as follows:

1. Select ntree samples from original data
2. For each sample, create an unpruned tree by following modification:
    - At each node, randomly sample ntry of the predictors and select the best split from those variables
3. Predict new data by combining the predictions of the ntree trees

6. CRITICAL PROTEINS ASSOCIATED WITH DS

In this chapter, the critical protein subsets associated with DS are obtained and presented. The forward feature selection method is applied to different datasets. The selected subsets of proteins differentiate healthy and unhealthy mice.

By comparing the classification accuracy, the importance of selected proteins can be understood. If accuracy result of selected proteins in this thesis is higher than what is found in previous work, it can be concluded that more critical protein subset is selected. This subset differentiates healthy and unhealthy mice more accurately.

KNIME tool is used. Naïve Bayes algorithm is applied in forward feature selection technique. In Naïve Bayes algorithm, Naïve Bayes learner creates a Bayesian model from the input training data. Naïve Bayes Predictor applies Bayesian modes to the input data. The scorer component is added at the end of the workflow and measures classifiers' performance. In addition, upper and lower number limit of selected feature, population size and iterations are specified in forward feature selection. After selecting the subset of features, different classification methods are applied for differentiating mice. The parameters of classification methods are selected based on grid search method.

The phyton code for grid search method applied on DNN for obtaining parameters in successful learning and the result of grid search method are seen in Listing 6.1 and Listing 6.2. The selected parameters are input to DNN method in Listing 6.3.

Listing 6.1: Grid search method for selecting parameters of DNN

```
from __future__ import print_function
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
import numpy as np
x=np.array([[]]) #specify data
y=[] # specify label
X_train, X_test, y_train, y_test = train_test_split(
    x, y, test_size=0.33, random_state=0)
# Set the parameters by cross-validation
tuned_parameters = [{'hidden_layer_sizes': [8,12,16,20],
   'activation': ['identity', 'logistic', 'tanh', 'relu'],
   'learning_rate_init': [0.01,0.05,0.1,0.5],
   'max_iter': [40,80,120,160,200]} ]
clf = GridSearchCV(MLPClassifier(), tuned_parameters, cv=5)
clf.fit(X_train, y_train)
print("Best parameters set found on development set:")
print(clf.best_params_)
print("Grid scores on development set:")
means = clf.cv_results_['mean_test_score']
stds = clf.cv_results_['std_test_score']
print()
for mean, std, params in zip(means, stds, clf.cv_results_['params'
   ]):
print("\%0.3f (+/-\%0.03f) for \%r"
              \% (mean, std * 2, params))
print("Detailed classification report:")
y_true, y_pred = y_test, clf.predict(X_test)
print(classification_report(y_true, y_pred))
```

Activation parameter in tuned_parameters shows activation function of a node. It describes the output of that node. The purpose of an activation function is to add some kind of non-linear property to the function.

Listing 6.2: Grid search result of DNN

```
Best parameters set found on development set:
{'activation': 'relu', 'hidden_layer_sizes': 16, '
    learning_rate_init': 0.05, 'max_iter': 120}
Grid scores on development set:
0.884 (+/-0.066) for {'activation': 'identity', 'hidden_layer_sizes
    ': 8, 'learning_rate_init': 0.01, 'max_iter': 40}
0.915 (+/-0.026) for {'activation': 'identity', 'hidden_layer_sizes
    ': 8, 'learning_rate_init': 0.01, 'max_iter': 80}
.
.
0.961 (+/-0.062) for {'activation': 'relu', 'hidden_layer_sizes':
    20, 'learning_rate_init': 0.5, 'max_iter': 200}
Detailed classification report:
             precision    recall  f1-score   support

          0       1.00      0.96      0.98        45
          1       0.96      1.00      0.98        44
          2       0.98      0.98      0.98        57
          3       0.98      0.98      0.98        43

avg / total       0.98      0.98      0.98       189
```

Listing 6.3: DNN for successful learning

```
import numpy as np
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import cross_val_score
x=np.array([[]])
y=[]
# build a classifier
clf = MLPClassifier(activation= 'relu', hidden_layer_sizes= 16,
    learning_rate_init= 0.05, max_iter= 120,random_state=42)
scores = cross_val_score(clf, x, y,cv=5)
scores=scores.mean()
print(scores)
```

The best parameters (relu activation function, 16 hidden sizes, 0.05 learning rate and 120 iterations) are selected by grid seach method as stated in Listing 6.1. Listing 6.2 shows grid search result of DNN. The selected parameters are input to DNN algorithm in Listing 6.3. The accuracy result is calculated based on the selected parameters. Thus, grid search method for hyperparameter optimization is required in order to obtain accurate and solid accuracy results.

6.1 Finding Important Proteins in DS

As a first step, important proteins in DS are found by applying the following different classification methods: NN, Random Forest and SVM. These methods are run for classifying mice. With grid search method, the parameters are optimized. The classification results are compared with the results of B.Feng et al. (2017) where AdaBoost algorithm was used for selecting 30 features. They used NN, SVM and Random Forest methods for mice classification.

Table 6.1 shows precision and 10-folds cross validation results of classification techniques. For building solid and stable classification models, a 10-folds cross-validation for each model ran to produce a systematic assessment. In this work, Random Forest and SVM generated from the selected proteins achieves higher accuracy than Feng et al. (2017). DNN gives highest accuracy results. Feng et al. (2017) did not apply DNN to the selected protein subset.

Table 6.1: Accuracy result comparison of successful learning.

Classifiers	**Accuracy Result of B.Feng's Work (Feng, 2017)**	**Accuracy Result of Our Work**
Deep Neural Network	-	0.993
Random Forest	0.977	0.991
SVM	0.956	0.986

Also, Furat and Ibrikci (2018) applied Bayesian Network, KNN (K Nearest Neighbor), Decision Table, Random Forest and SVM classification techniques. Compared to B.Feng et al. (2017), they did not reduce feature set and took into account all features. We also applied same classification methods and compare accuracy results

with Furat and Ibrikci (2018).

Comparison results of classification performance using five different algorithms with 10-fold cross validation and 50–50% train-test data partition can be seen Table 6.2 and Table 6.3, respectively.

Table 6.2: Comparison of accuracy result with 10-fold cross validation.

Classifiers	**Accuracy Result of Furat and Ibrikci (2018)**	**Accuracy Result of Our Work**
Bayesian Network	0.944	0.950
KNN	0.993	1
Decision Table	0.955	0.966
Random Forest	1	1
SVM	1	1

As can be seen in Table 6.2, the obtained accuracy results in our work are higher or equal for all classification techniques. The improvement in accuracy results can be arisen from the selection of appropriate parameters. For example, neighbors parameter of KNN is selected 3 in our work.

Table 6.3: Comparison of accuracy result with 50–50% train-test data partition.

Classifiers	**Accuracy Result of Furat and Ibrikci (2018)**	**Accuracy Result of Our Work**
Bayesian Network	0.954	0.931
KNN	0.983	0.982
Decision Table	0.983	0.993
Random Forest	1	1
SVM	1	1

As can be seen in Table 6.2, the obtained accuracy with Bayesian Networks is lower than Furat and Ibrikci (2018). The difference can be arisen from the train - test partition. With lower test size partition, the accuracy increases as algorithm can be learned with higher size train data.

6.2 Systematic Analysis of Finding Important Proteins in DS

The systematic analysis is done for determining feature subsets for three cases-succesful learning, rescued learning with drug and failed learning.

These feature subsets are compared with Higuera et al. (2015) where three feature subsets are highlighted for successful learning, rescued learning and failed learning. Higuera et al. (2015) evaluated control mice and trisomic mice separately and together in order to comprehend changes in protein expression. In the first case, all groups of normal mice were evaluated to understand which changes in protein expression level are required for successful learning. For determining critical proteins in rescued learning, trisomic mice exposed to CFC with and without memantine were inspected as the second case. The third case finds out important protein abnormalities in failed learning case by comparing normal and trisomic mice expression levels.

In this thesis, these three feature subsets are also selected to identify critical proteins in successful learning, in rescued learning and in failed learning cases. The number of features in subsets are chosen based on the number stated in Higuera et al. (2015). After resolving the different feature subsets for the three cases, classification is performed for differentiating classes of mice. DNN, Gradient Boosted Tree, Random Forest and SVM classification methods are implemented by using Python and Scikit Learn package (Hao and Ho, 2019).

The parameters of classifiers are resolved based on the hyper-parameter optimization technique, grid search. (Bergstra et al., 2011). The accuracy results of selected feature subsets are compared with the Higuera's accuracy results.

6.2.1 Feature Subset from Control Mice and Classification Result

Table 6.2 describes the selected features and their accuracy with successful learning. Accuracy is obtained when selected feature is inserted into the subset. In the first

case, feature subset is selected from control group mice. By comparing control group mice with and without memantine treatment and CFC stimulation (c-CS-m, c-CS-s, c-SC-m, c-SC-s), critical proteins in successful learning can be figured out.

Table 6.2 also presents Higuera et al.'s (2015) feature subset. When these two subsets are compared, there are 4 common proteins out of 11 proteins and they are shown as bold. Higuera et al. (2015) selected eleven proteins. The selected proteins play important roles in L/M, immune response, MAPK pathyway, mTOR pathway and AD. In order to compare this work with Higuera's work in a quantitative manner, eleven proteins are selected for successful learning as in Higuera's work.

Table 6.4: Feature subset of successful learning.

Feature No	**Feature Accuracy**	**Feature Subset**	**Feature Subset of Higuera et al. (2015)**
1	0.656	**SOD1**	DYRK1A
2	0.751	Ubiquitin	ITSN1
3	0.852	**pGSK3B**	pERK
4	0.873	**S6**	BRAF
5	0.905	**CaNA**	**SOD1**
6	0.921	IL1B	pNUMB
7	0.937	BAX	**pGSK3B**
8	0.942	pNR2A	CDK5
9	0.942	BDNF	**S6**
10	0.942	pJNK	GFAP
11	0.942	pCFOS	**CaNA**

SOD1 is located on chromosome 21 and causes immune problems in Amyotrophic lateral sclerosis (ALS) disease (Milani et al., 2011). Ribosomal Protein S6 and pGSK3B are components of mTOR pathway which play roles in learning (A. McCombe and D. Henderson, 2011). Also, in the literature it is noted that GSK3 inhibitors provide to inhibit excessive inflammation and ameliorate the autoimmune disease (Beurel,Grieco and Jope, 2015). CaNA and IL1B are known to be pathogenesis of AD (Nicoll et al., 2000; Dinarello, 2011). Also, it is known that IL1B is natural suppressor of innate inflammatory (Reese and Taglialatela,2011). BAX and ubiquitin play critical roles in apoptosis and immune response (Tano et al.,2011;Sujashvili,2016). BDNF takes action in L/M (Cunha, Brambilla and Thomas, 2010). Also, BDNF

bridges neuroplasticity and inflammation (Calabrese et al.,2014;Tu et al.,2016). pNR2A has well established roles in learning (Li et al., 2007). pJNK, component of MAPK pathway is associated with L/M (Shen, 2014). pCFOS is an IEG and significant in long term memory (Kidambi et al.,2010).

In the first case of systematic analysis, protein expression levels of control group mice are analyzed. It can be deduced that proteins related to the L/M pathway and the immune responses are critical in successful learning.

Table 6.5 shows the classification accuracies of selected feature subsets for successful learning in this thesis and Higuera et al. (2015). It can be seen that the feature subsets in this work give higher accuracy results for all classification techniques. SVM gives the highest accuracy.

Table 6.5: Accuracy result comparison of successful learning.

	Accuracy Result of Our Work	**Higuera et al. (2015) Accuracy Result**
Deep Neural Network	0.972	0.967
Gradient Boosted Tree	0.935	0.902
Random Forest	0.963	0.902
SVM	0.981	0.961

6.2.2 Feature Subset from Trisomic Mice and Classification Result

To understand the important proteins in rescued learning, features are selected from data consisting of trisomic mice which are exposed to CFC with and without memantine (t-CS-m, t-CS-s). When exposed to CFC, the trisomic mice are unsuccessful to learn if they are not injected with the drug memantine. Table 6.6 shows the selected features and the accuracy results for the rescued learning. There are 2 common proteins (BRAF, CDK5) with Higuera et al.'s work (2015) shown in bold. Accuracy of feature shows the accuracy of feature subset when the corresponding feature is added. 9 proteins are selected in order to compare result with Higuera et al. (2015). Higuera et al also selected 9 proteins for rescued learning.

Table 6.6: Feature subset of rescued learning.

Feature No	Feature Accuracy	Feature Subset	Feature Subset of Higuera et al. (2015)
1	0.762	**BRAF**	DYRK1A
2	0.838	S6	pERK
3	0.85	**CDK5**	**BRAF**
4	0.887	BDNF	**CDK5**
5	0.887	pCREB	RRP1
6	0.9	PKCA	GFAP
7	0.912	SOD1	GluR3
8	0.925	PSD95	P3525
9	0.925	pNR2A	Ubiquitin

BRAF and PKCA are associated with MAPK pathway and effective in learning (Lee et al., 2014; Zhang et al., 2009). CDK5 is synaptic protein and plays a important role in long-term memory (Pollonini,2008). Also, it regulates the escape of tumors from the immune system (Shupp,Casimiro and Pestell, 2017). PSD95 is a neuropathological indicator of AD observed in later stage of DS (Shao,2011). In addition, PSD95 colocalizes with major histocompatibility complex class I (MHCI) which is the marker of its expressed proteins. Also, it is significant for the immune system to differentiate self from nonself (Marin and Kipnis, 2013). CREB adjusts vital cell stages and participates in neuronal plasticity (Ortega-Martı'nez, 2015). Thus, it can be concluded that proteins which are important in rescued learning are related to the L/M and the immune response.

Table 6.7 shows the comparison of rescued learning results. DNN and SVM give highest accuracy. The accuracy results of the feature subset in this work are higher than previous work for all classification methods.

Table 6.7: Accuracy result comparison of rescued learning.

	Accuracy Result of Our Work	Higuera et al. (2015) Accuracy Result
Deep Neural Network	0.971	0.954
Gradient Boosted Tree	0.933	0.892
Random Forest	0.946	0.883
SVM	0.971	0.921

6.2.3 Feature Subset from Control and Trisomic Mice and Classification Result

To determine proteins that are critical in failed learning with trisomic mice, features are selected from the trisomic mice protein expression exposed to CFC without memantine (t-CS-s) and the control mice protein expression levels which are exposed to CFC with and without memantine (c-CS-m, c-CS-s).

Table 6.8 shows the selected features and accuracy results of feature subset in failed learning. There are 2 common proteins (P38, GluR3) out of 10 proteins with Higuera et al.'s work (2015) .

Table 6.8: Feature subset of failed learning.

Feature No	Feature Accuracy	Feature Subset	Feature Subset of Higuera et al. (2015)
1	0.636	**P38**	pNR1
2	0.713	pPKCAB	APP
3	0.775	CAMKII	mTOR
4	0.814	pCAMKII	**P38**
5	0.868	**GluR3**	NR2B
6	0.891	DSCR1	RAPTOR
7	0.907	nNOS	S6
8	0.915	BAX	Tau
9	0.93	pCFOS	**GluR3**
10	0.93	ERK	EGR1

Two of these proteins (BAX and pCFOS) were also highlighted in successful learning and described above. The remaining selected proteins are largely connected to MAPK signaling pathway, such as P38, pPKCAB, CAMKII, pCAMKII and ERK. GluR3 is related to glutamate receptors which cause memory deficit if excess amount of glutamate binds to receptor (Ahmed et al., 2009). DSCR1 is known to be over expressed in DS. It also affects signaling pathway (Lee et al., 2009). Failed learning case also shows us the importance of signaling pathway in the learning process.

Table 6.9 shows the comparison of classifications for failed learning. DNN and SVM

give highest accuracy results. The classification results of our feature subsets are higher than Higuera et al.'s work (2015) for all classification methods.

Table 6.9: Accuracy result comparison of failed learning.

	Accuracy Result of Our Work	**Higuera et al. (2015) Accuracy Result**
Deep Neural Network	0.926	0.921
Gradient Boosted Tree	0.879	0.844
Random Forest	0.892	0.859
SVM	0.926	0.910

6.3 Response Similarity of Different Drugs Treating Ts65Dn Mice

To understand whether different drugs-treated Ts65Dn mice exhibit similar response or not, the critical proteins expressed in response to memantine and RO4938581 are selected. By this way, the molecular pathway of rescued performance in DS can be understood and effective drugs can be developed.

Table 6. 10 shows the expressed proteins when mice are injected with memantine and RO4938581 drugs. 4 gene products shown as bold are in common with RO4938581 and memantine. In order to compare the results of two drugs, the same number of proteins are selected from the protein expression datasets of two drugs. The selected proteins have important roles in learning pathway and neural growth.

Table 6.10: Feature subsets of RO4938581 and memantine.

Feature Subset with RO4938581	**Feature Subset with memantine**
BRAF	**BRAF**
S6	**S6**
MEK	CDK5
ADARB1	**BDNF**
pBRAF	pCREB
PKCA	**PKCA**
NR1	SOD1
SHH	PSD95
BDNF	pNR2A

BRAF, PKCA and MEK (Mitogen Activated Protein (MAP) Kinase) are associated with MAPK pathway and important in learning (Gardiner, 2004). Ribosomal Protein S6 is component of mTOR pathway which takes action in learning (A. McCombe and D. Henderson, 2011). NR1 (N-Methyl-D-Aspartate Receptor Subunit) is the component of NMDAR receptor and plays an essential role in excitatory transmission and L/M process (Zorumski and Izumi, 2012). BDNF gene encodes a unit of the nerve growth factor family of proteins. The SHH (Sonic Hedgehog) gene produces instructions for generating a protein which is required for the progress of forebrain. This signaling protein helps to create the line that divides the right and left sides of the forebrain.

6.4 Protein Subsets which Display the Regional Fluctuation with Aging

The critical protein subsets which display regional fluctuation with aging are determined in this section. These subsets are obtained from two datasets which show the expression profiles of young and old mice at three different brain regions (CB, CR, HP). Using the selected subsets, the process of DS can be analyzed by inspecting molecular pathways where the selected proteins take rolez. Also, by looking into the selected feature subsets from old mice and young mice datasets, the aging process in DS can be understood.

Figure 6.1 shows accuracy dispersion of old mice protein expression. Max accuracy is obtained with 10 proteins. Thus, 10 proteins are selected from old mice protein dataset.

The most critical proteins in old mice dataset can be observed in Table 6.11 below.

The first selected protein is RCAN1 (Regulator of Calcineurin 1) and also called as Down Syndrome Critical Region 1 (DSCR1). It regulates calcineurin (CN) signaling in the brain (Lee et al., 2009). Errors in CN function were also associated with AD. Ubiquitin plays critical roles in apoptosis (Chen and Qiu, 2013). APP is AD related protein (Long et al., 2018). TH participates in the conversion of tyrosine to dopamine

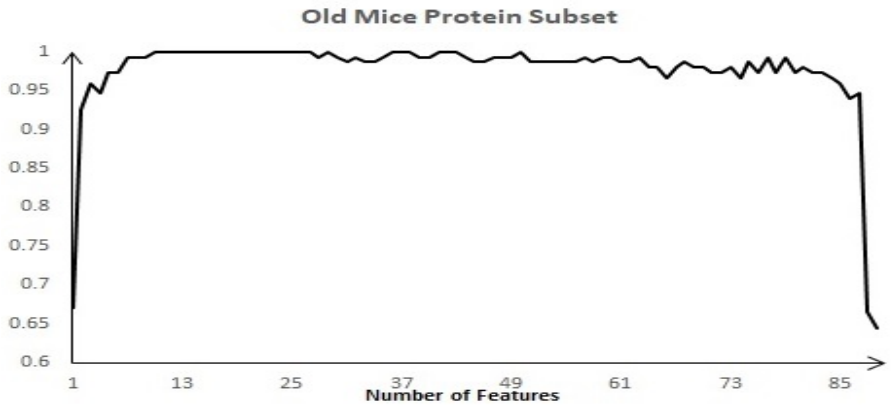

Figure 6.1: Accuracy result of old mice protein set across brain regions

Table 6.11: Feature subset from old mice dataset across brain regions

Accuracy of Subset	Selected Feature
0.67114094	RCAN1r
0.926174497	Ubiquitin
0.959731544	APP
0.946308725	TH
0.973154362	ARC
0.973154362	ERK
0.993288591	mTOR
0.993288591	ERBB4
0.993288591	H3MeK4
1	pNR2A

and has a key role in the physiology of adrenergic neurons (Nagatsu and Nagatsu, 2016).

ARC (Activity Regulated Cytoskeleton) is a unit of the immediate-early gene (IEG) family and a marker for plastic changes in the brain (Gallo et al., 2018). ERKs are protein kinase intracellular signaling molecules. Disruption of the ERK pathway causes cancer. mTOR is a component of mTOR pathway. Knockout of ERBB4 functions in synaptic plasticity. Also, ERBB4 affect the dendritic spine development (Cooper and Koleske, 2014). H3meK4 (Methylated Lysine 4 on Histone H3) is histone protein and has a role in memory formation (Peixoto and Abel, 2012). pNR2A which is the subunit of NMDAR has well established roles in learning (Li et al., 2007) .

Figure 6.2 shows accuracy dispersion of young mice protein dataset. Max accuracy

is obtained with fourteen proteins. Thus, fourteen proteins are selected from young mice protein dataset.

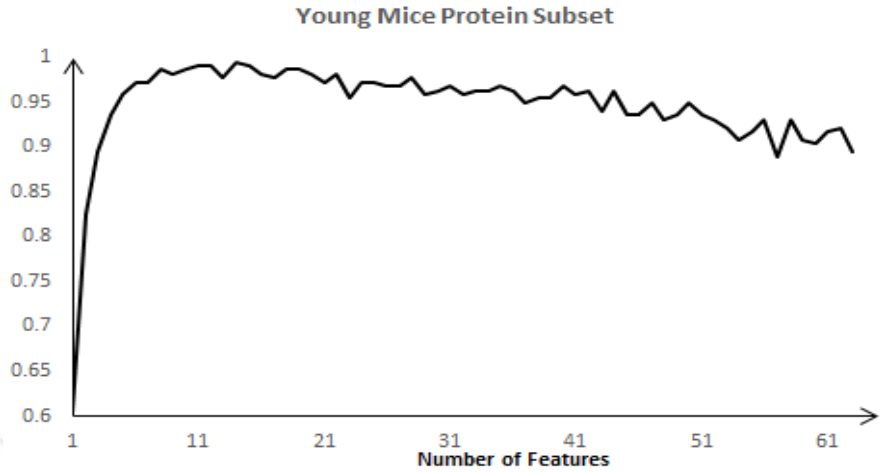

Figure 6.2: Accuracy result of young mice protein set across brain regions

Table 6.12 shows the accuracy of feature subset and the selected proteins across brain regions in young mice.

Table 6.12: Feature subset from young mice dataset across brain regions

Accuracy of Subset	Selected Feature
0.605504587	MEK
0.825688073	APP
0.894495413	ITSN1
0.935779817	GluR3
0.958715596	pGJA1
0.972477064	P3525
0.972477064	DYRKA1
0.986238532	AKT
0.981651376	pPKCG
0.986238532	CTTNB1
0.990825688	NUMB
0.990825688	PRMT2
0.97706422	BDNF
0.995412844	BCL2

MEK and pPKCG (Phospho Protein Kinase C Gamma) are components of MAPK pathway. APP and P3525 are AD related proteins which observed in later stage of DS. ITSN1 (Intersectin 1) and DYRKA1 (Dual Specificity Tyrosine Phosphorylation-Regulated Kinase 1A) are Hsa21 proteins. Their perturbations in pathways cause

L/M deficits. GluR3 and NUMB are the component of NMDAR receptor which plays an essential role in excitatory transmission and L/M process (Moore and Baleja, 2012). pGJA1 (Phospho Gap Junction Alpha-1 Protein) provides cell-to-cell communication by forming channels between cells. Also, it is involved in placenta development (Dbouk et al., 2009). AKT plays role in mTOR pathway. CTTNB1 (Catenin Beta-1) plays an essential role in neuro-development (Dong et al., 2016). BDNF encodes a member of the nerve growth factor family of proteins. BCL2 is playing critical roles in apoptosis (Adams and Cory, 2007).

When the selected proteins for old and young mice datasets are evaluated, it is seen that they play important roles in the processes like, MAPK signaling pathway, mTOR signaling pathway cell-to cell communication, apoptosis process, AD pathway. However, one-to-one comparison between two subsets can not be done as the proteins in young and old mice datasets are different. Young mice dataset contains only 3 proteins from 10 proteins in subset of old mice and only APP protein is common between two protein subsets.

In the literature, the datasets were examined with statistical techniques and results showed only the increase or decrease of protein expression in different parts of brains. Also, the protein fluctuations of old and young mice were compared. The general sketch of the protein expression profiles throughout the aging was obtained. Rather than the general picture, in this thesis, the protein subsets which are critical regionally are determined for old and young mice. By giving efforts to these subsets, the age related change in the mechanism of molecular pathways can be understood for age related drug treatment in DS.

6.5 Protein Subsets which Highlight the Importance of Mice Type (TS65Dn - Tc1)

Using Tc1 mice and littermate controls, Ahmed et al. (2014) measured 64 protein levels in cerebellum, 90 protein levels in hippocampus and cortex in order to identify

the molecular cause for the phenotypic features. It is stated that there are abnormal protein levels involved in immediate early gene (IEG), MAP kinase pathway, mTOR pathway, neuregulin signaling, and receptor proteins by comparing range of protein expression in three different brain regions, hippocampus, cortex and cerebellum. However, expression levels of proteins in one or two regions were not determined and proteins which differentiate brain regions were not identified.

In this thesis, critical proteins differentiate in three brain regions are selected by applying forward feature selection method. Figure 6.3 shows accuracy dispersion of Tc1 mice protein dataset. Max accuracy is obtained with seventeen proteins. Thus, seventeen proteins are selected.

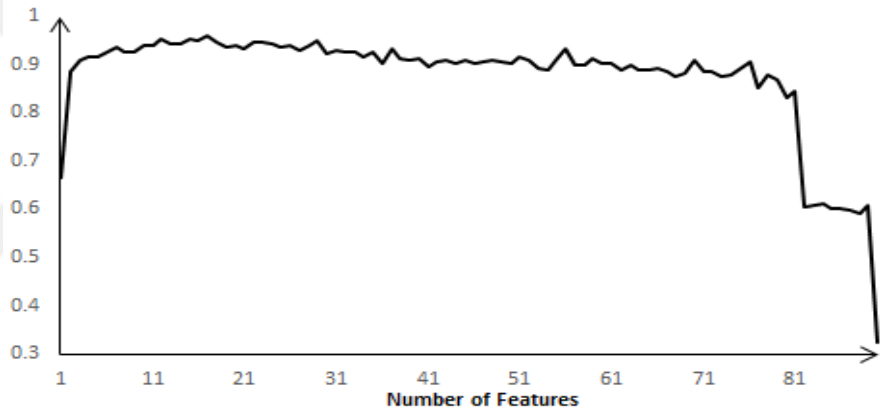

Figure 6.3: Accuracy result of Tc1 mice protein set across brain regions

Table 6.13 shows the selected proteins with Tc1 mice. 17 features are selected as maximum accuracy is obtained with 17 features.

Most of the selected proteins are same as in Ahmed et al. (2014). The selected four proteins (nNOS, DYRK1A, SOD1 and APP) change significantly in one or more brain regions. After literature review, it can be deduced that selected proteins are relevant to the L/M pathway.

nNOS participates in neurotransmission (Esplugues, 2002). DYRKA1 is H21 proteins and their perturbations in pathways cause L/M deficits. SOD1 found on chromosome

Table 6.13: Feature subset of Tc1 mice across brain regions

Accuracy of Subset	Selected Feature
0.6654676258992805	nNOS
0.8848920863309353	**DYRK1A**
0.9100719424460432	SOD1
0.9172661870503597	CHAF1B
0.9172661870503597	**AKT**
0.9244604316546763	BAX
0.935251798561151	SYP
0.9244604316546763	NR2A
0.9244604316546763	**CTTNB1**
0.9388489208633094	**PRMT2**
0.9388489208633094	NR1
0.9532374100719424	ADARB1
0.9424460431654677	**APP**
0.9424460431654677	**GluR3**
0.9532374100719424	Ubiquitin
0.9496402877697842	TRKA
0.960431654676259	NR2B

21 causes immune abnormalities in ALS (Milani et al., 2011). Also, it increases reactive oxygen in DS. CHAF1B (Chromatin Assembly Factor 1 Subunit B) takes action in chromatin assembly after replication (Duan et al., 2019). AKT plays role in mTOR pathway. BAX and ubiquitin play critical roles in apoptosis and immune response. SYP (Synaptophysin) encodes an integral membrane protein of tiny synaptic vesicles in brain (Leube, Wiedenmann and Franke, 1989). NR2A, NR1 and NR2B (N-Methyl D-Aspartate Receptor Subtype 2B) have well established roles in learning. CTTNB1 plays an essential role in neurodevelopment (Dong et al., 2016). PRMT2 (Protein Arginine N-Methyltransferase 2) has protein homodimerization activity and transcription coactivator activity. APP is AD related proteins observed in later stage of DS (Long et al., 2018). GluR3 is the component of NMDAR receptor which plays an essential role in excitatory transmission and L/M process (Moore and Baleja, 2012). TRKA (Tropomyosin Receptor Kinase A) plays a role in specifying sensory neuron subtypes (Lechner et al., 2009).

Figure 6.4 shows accuracy dispersion of Ts65Dn mice protein dataset. Max accuracy is obtained with fourteen proteins. Thus, fourteen proteins are selected.

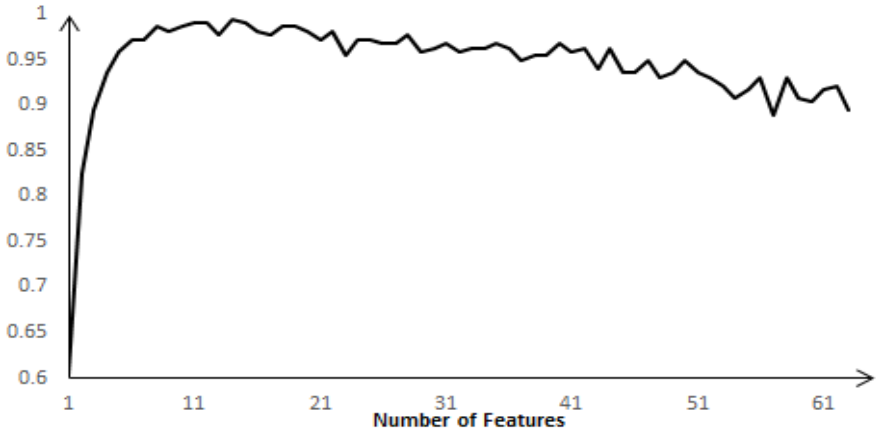

Figure 6.4: Accuracy result of Ts65Dn mice protein set across brain regions

Table 6.14 shows the selected fourteen proteins with Ts65Dn mouse model. The max accuracy is obtained with fourteen proteins. Therefore first fourteen proteins are considered for Ts65Dn mice when analyzing the protein fluctuation across different brain regions. When selected proteins from Tc1 mice are compared with selected proteins from Ts65Dn mice, it can be seen that there are 6 common proteins (APP, GluR3, DYRKA1, AKT, CTTNB1 and PRMT2) out of 14 proteins which are shown in bold.

Table 6.14: Feature subset of Ts65Dn mice across brain regions

Accuracy of Subset	Selected Feature
0.605504587	MEK
0.825688073	**APP**
0.894495413	ITSN1
0.935779817	**GluR3**
0.958715596	pGJA1
0.972477064	P3525
0.972477064	**DYRKA1**
0.986238532	**AKT**
0.981651376	pPKCG
0.986238532	**CTTNB1**
0.990825688	NUMB
0.990825688	**PRMT2**
0.97706422	BDNF
0.995412844	BCL2

In the literature, we can observe that selected proteins are related to important

processes. MEK and pPKCG are components of MAPK pathway. APP and P3525 are AD related proteins which observed in later stage of DS. ITSN1 and DYRKA1 are Hsa21 proteins and their perturbations in pathways cause L/M deficits. GluR3 and NUMB are the component of NMDAR receptor which plays an essential role in excitatory transmission and L/M process. pGJA1 provides cell-to-cell communication by forming channels between cells. Also, it is involved in placenta development. AKT (Protein kinase B) plays role in mTOR pathway. CTTNB1 plays an essential role in neurodevelopment. PRMT2 has protein homodimerization activity and transcription coactivator activity. BDNF encodes the nerve growth factor family of proteins. BCL2 (B-Cell Lymphoma 2) plays critical roles in apoptosis.

Expression levels of some proteins were not determined before. We believe that proteins selected in this thesis can be utilized to understand the process of DS as they potentially contribute to phenotypic features and influence drug responses.

6.6 Determine the Protein Subsets which Show Importance of Brain Region Fractions

Max accuracy are obtained with thirteen and twenty-six proteins for cytosolic and nuclear fractions. They are shown in Figure 6.5 and 6.6, respectively.

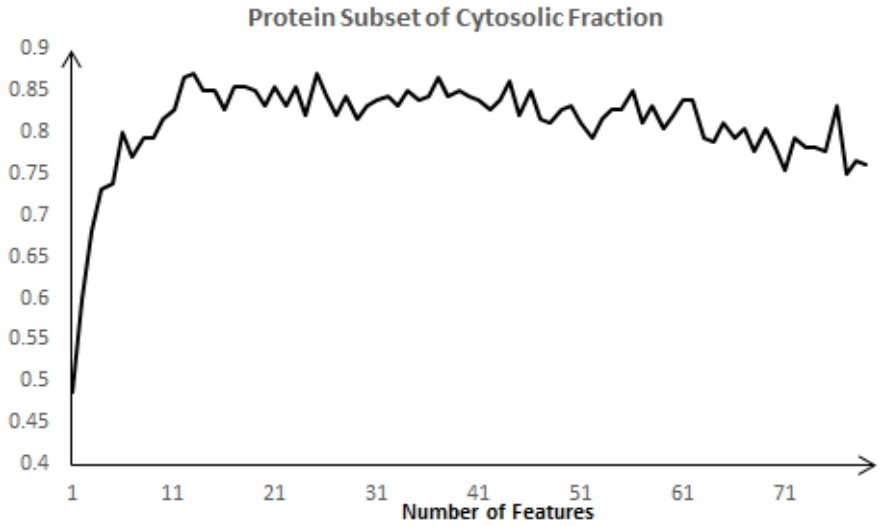

Figure 6.5: Protein subset accuracy of cytosolic fraction from Ts65Dn mice cortex

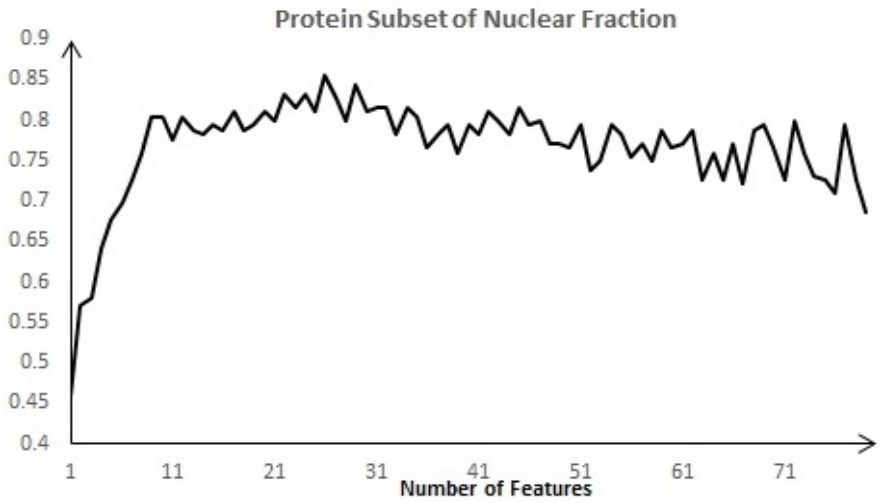

Figure 6.6: Protein subset accuracy of nuclear fraction from Ts65Dn mice cortex

Table 6.15 shows the critical proteins expressed in cytosolic and nuclear fractions of cortex.

The critical proteins are related to different pathways and processes, such as MAPK and MTOR signaling pathways, AD, neurotrophin signaling pathway and apoptosis. Four gene products (DSCR1, P3525, H3AcK18, GSK3B) are common to both cytosolic and nuclear fractions of cortex.

Cortex is the most developed region of brain. It takes roles in thinking, perceiving and understanding language. Human brain development is marked by gene expression across the lifespan. One factor governing the changes in development is the compartmentalization of the transcriptome by the nuclear membrane into nuclear and cytoplasmic fractions. pre-mRNA and longer genes take more time to be transcribed. Thus, genes are often overrepresented in the nucleus compared to cytoplasm. Also, nuclear membrane filter bursts of gene expression from the cytoplasm. Thus, analyzing the compositional differences can be helpful for understanding DS.

Table 6.15: Accuracy and feature subset of cytosolic and nuclear fractions from cortex

Accuracy	Cytosolic Fraction	Nuclear Fraction	Accuracy
0.48603352	**DSCR1**	P38	0.458100559
0.597765363	**P3525**	pGSK3B	0.569832402
0.681564246	P70S6	Ubiquitin	0.581005587
0.731843575	pNR1	pPKCAB	0.642458101
0.737430168	PKCA	NR2A	0.675977654
0.798882682	GFAP	Tau	0.698324022
0.770949721	MEK	BCL2	0.726256983
0.793296089	ARC	APP	0.759776536
0.793296089	**H3AcK18**	H3DMe4	0.804469274
0.815642458	SOD1	EGR1	0.804469274
0.826815642	**GSK3B**	pNR2B	0.776536313
0.865921788	pPKCG	pCAMKII	0.804469274
0.87150838	pERK	pRSK	0.787709497
		ERBB4	0.782122905
		SHH	0.793296089
		ELK	0.787709497
		pAKT	0.810055866
		MTOR	0.787709497
		P3525	0.793296089
		CAMKII	0.810055866
		AcetylH3K9	0.798882682
		nNOS	0.832402235
		DSCR1	0.815642458
		H3AcK18	0.832402235
		GSK3B	0.810055866
		BAX	0.854748603

7. PATHWAY ANALYSIS OF SELECTED PROTEIN SUBSETS

In this chapter, pathway analysis is performed to determine relevant proteins within a pathway or constructing pathway de novo from the interested proteins. In this thesis, for understanding the effects of the selected proteins in DS, the pathway analyses are done using reactome Pathway Browser. Pathway Browser is bioinformatics tool for analysis, interpretation and visualization of pathway knowledge (Fabregat et al., 2016).

Reactome is a database of pathways and reactions in human biology. Reactions can be thought as pathway stages. Reactome explains a reaction as any event in biology that alters the state of biological molecule. The Pathway Browser is the tool of interacting and viewing pathways in Reactome. It evaluates datasets and examines pathways. This tool provides numerous types of analysis such as:

- Comparison of a pathway with corresponding pathway in another species
- Over-representation pathway
- The protein-protein or protein-compound interaction data from external databases or user data onto a pathway
- The expression of user data onto a pathway

7.1 Pathway Analysis of Successful Learning

For successful learning case, eleven proteins (SOD1 (Superoxide Dismutase 1), Ubiquitin, pGSK3B (Phospho Glycogen Synthase Kinase 3 Beta) , S6, CaNA (Carbonic Anhydrase), IL1B (Interleukin 1 Beta), BAX (BCL2 Associated X), pNR2A (Phos-

pho N-Methyl D-Aspartate 2A), BDNF (Brain Derived Neurotrophic Factor), pJNK (Phospho c-Jun N-Terminal Kinases), pCFOS (Phospho FBJ Murine Osteosarcoma Viral Oncogene Homolog)) are selected. Figure 7.1 shows pathway visualization of the selected genes.

Figure 7.1: Pathway visualization of selected genes for successful learning.

Table 7.1: Specific pathways of selected genes for successful learning.

Pathway Name	Entities				Reactions
	found	ratio	p-value	FDR	found
Cellular responses to stress	11/690	0.048	4.16e-09	1.67e-06	41/227
Cellular responses to external stimuli	11/708	0.049	5.44e-09	1.67e-06	41/258
CLEC7A/Inflammasome pathway	3/8	5.54e-04	3.42e-07	7.00e-05	2/4
Signaling by interleukins	9/639	0.044	4.65e-07	7.12e-05	36/490
Transcription regulation by RUNX	5/147	0.01	3.96e-06	4.87e-04	16/84
Cytokine signaling in immune system	10/1261	0.087	1.61e-05	0.002	45/699

As can be seen in Figure 7.1, the selected genes of successful learning case take part in immune system, signal transduction and gene expression pathways extensively. Also, these genes play roles in programmed cell death, cellular responses to external stimuli, disease and DNA repair mechanisms.

Table 7.1 shows the specific pathways where the selected genes in succesful learning take actions. When Table 7.1 is inspected, it is seen that selected proteins play roles in cellular responses to stress and external stimuli. Also, signaling of CLEC7A (C-Type Lectin Domain Family 7 Member A) (Sun and Zhao, 2007), Interleukin-1 and cytokines (Zídek, Anzenbacher and Kmoníčková, 2009) are important in successful learning. CLEC7A is a pattern-recognition receptor and triggers direct innate immune responses. Interleukin-1 plays a crucial role in many auto inflammatory diseases (Nicoll et al., 2000; Dinarello, 2011). Cytokines are small-scale proteins that adjust and mediate immunity and inflammation. They are secreted in response to immune stimuli. As seen in Table 7.1, selected proteins in successful learning trigger immune system, imflammatory response extensively.

Selected pathways are ranked by the p-value calculated from over-representation analysis. p-value is the probability that would be greater than or equal to the actual observed results when the null hypothesis is true. Found entities shows the number of genes in gene list that take action in specified pathway. Entities ratio

Table 7.2: Specific pathways of selected genes for rescued learning with memantine.

Pathway Name	Entities				Reactions
	found	ratio	p-value	FDR	found
Transcriptional regulation by MECP2	5/100	0.007	4.53e-08	1.31e-05	12/77
MECP2 regulates neuronal ligands transcription	3/13	9.0e-04	3.30e-07	4.78e-05	4/8
Signaling by NTRKs	4/118	0.008	5.67e-06	5.45e-04	43/127
Signaling by Receptor Tyrosine Kinases	6/554	0.038	1.19e-05	8.54e-04	53/657
Activated NTRK2 signals through CDK5	2/10	6.9e-04	5.01e-05	0.002	6/6
MECP2 regulates transcription factors	2/10	6.9e-04	5.01e-05	0.002	2/8

is the proportion of Reactome pathway molecules. Over-representation analysis determines whether certain Reactome pathways are over-represented. It calculates a probability score that is corrected for false discovery rate (FDR). Entities FDR shows corrected over-representation probability. Reactions found is the number of reactions in the pathway that are represented by at least one molecule.

7.2 Pathway Analysis of Rescued Learning with Memantine

The trisomic mice fail to learn if not they are injected with the drug memantine that recovers the skill of learning. Table 7.2 shows the specific pathways where the selected genes in rescued learning with memantine take actions. Table 7.2 shows the specific pathways where the selected genes in rescued learning with memantine take actions. Methyl-CpG-binding protein 2 (MeCP2) regulation and NTRK2 (Neurotrophic Receptor Tyrosine Kinase 2) signaling play important roles in pathway of rescued learning. MeCP2 is a methylated-DNA-binding protein and errors lead to autism spectrum disorder (Nagarajan et al., 2006). CDK5 binds to BDNF-activated NTRK2 (TRKB). Signaling by TRKB and CDK5 plays a role in inflammation (Kumar Pareek, 2012). In addition, NTRK2 plays a vital role in LTP and learning (Minichiello, 2009). As seen in Table 7.2, signal transduction, neural system and disease pathways are crucial in rescued learning with memantine.

Figure 7.2: Pathway visualization of selected genes for rescued learning.

Figure 7.2 shows the pathways where selected genes in rescued learning with drug memantine take roles. The selected genes with drug memantine play roles in immune system, signal transduction, neural system, gene expression and disease pathways.

Table 7.3: Specific pathways of selected genes for rescued learning with RO4938581

Pathway Name	Entities				Reactions
	found	ratio	p-value	FDR	found
Negative feedback regulation of MAPK pathway	2/8	5.54e-04	2.78e-05	0.004	2/3
MECP2 regulates neural ligands transcription	2/13	9.01e-04	7.33e-05	0.006	2/8
Frs2-mediated activation	2/17	0.001	1.25e-04	0.008	6/13
Prolonged ERK activation events	2/20	0.001	1.73e-04	0.008	8/19
Signalling by NTRKs	3/118	0.008	1.86e-04	0.008	40/127
RAF activation	2/37	0.003	5.86e-04	0.019	6/10
Signaling to ERKs	2/42	0.003	7.53e-04	0.021	9/32
Negative regulation of MAPK pathway	2/46	0.003	9.61e-04	0.021	6/12
Signaling by high-kinase activity BRAF mutants	2/52	0.004	0.001	0.021	6/6
Signaling by moderate kinase activity BRAF mutants	2/54	0.004	0.001	0.021	7/7

7.3 Pathway Analysis of Rescued Learning with RO4938581

Like memantine, another drug- RO4938581 that is GABAA receptor negative allosteric modulator (NAM) is used for rescuing protein anomalies. Table 7.3 shows the pathway response of drug- RO4938581. Like pathway response of drug memantine, the selected proteins with drug- RO4938581 play roles in signal transduction extensively. Table 7.3 shows selected pathways of drug- RO4938581. Regulation by MeCP2 and NTRK2 signaling are common in pathways of memantine and RO4938581.

7.4 Pathway Analysis of Failed Learning

Failed learning differentiates successful learning from the absence of impulse to learn. Table 7.4 shows the specific pathways where the selected genes (P38, pPKCAB (Phospho Protein Kinase C Alpha/Beta), CAMKII (Ca 2+ /Calmodulin-Dependent Protein Kinase II), pCAMKII, GluR3 (Glutamate Receptor 3), DSCR1 (Down Syndrome Critical Region 1), nNOS (Neuronal Nitric Oxide Synthase), BAX, pCFOS, ERK (Extracellular Signal-Regulated Kinase) in failed learning play roles.

Table 7.4: Specific pathways of selected genes for failed learning.

Pathway Name	Entities				Reactions
	found	ratio	p-value	FDR	found
Trafficking of AMPA receptors	3/37	0.003	9.20e-06	0.001	4/4
Glutamate binding and synaptic plasticity	3/39	0.003	1.08e-05	0.001	9/9
Inhibition of nitric oxide production	2/5	3.46e-04	1.44e-05	0.001	2/5
Signaling by interleukins	6/639	0.044	4.10e-05	0.002	10/490
Interleukin-4 and interleukin-13 signaling	4/211	0.015	7.22e-05	0.002	2/46
Activation of the AP-1 family of transcription factors	2/12	8.31e-04	8.23e-05	0.002	3/5
Cytokine signaling in immune system	7/1261	0.087	2.18e-04	0.004	13/699
GluR2-containing AMPA receptors Trafficking	2/23	0.002	3.00e-04	0.005	3/3
NMDA receptors unblocking and glutamate binding	2/28	0.002	4.44e-04	0.006	5/5
Formation of the cornified envelope	3/138	0.01	4.46e-04	0.006	9/27

AMPA receptors, interleukin signaling are important factors in failed learning. AMPA receptors (AMPARs) moderate the vast of fast excitatory synaptic communication in the brain (Wang, Gilbert and Man, 2012). Interleukin plays a crucial role in many auto inflammatory diseases (Dinarello, 2011).

As can be seen in Figure 7.3, the selected genes of failed learning take part in immune system, signal transduction, neural system and gene expression broadly. Also, they take actions in programmed cell death and cellular response to external stimuli. It can be seen that like rescued learning, failed learning does not play role in DNA repair pathway.

Figure 7.3: Pathway visualization of selected genes for failed learning.

7.5 Pathway Analysis of Young Mice

By looking into the two different feature subsets obtained from old mice and young mice, the aging process in DS can be understood. These subsets are obtained from two datasets which show the expression profiles of young mice and old mice at three different brain regions.

Table 7.5: Specific pathways of selected genes from young mice.

Pathway Name	Entities				Reactions
	found	ratio	p-value	FDR	found
BH3-only proteins associate and inactivate anti-apoptotic BCL-2	2/11	7.6e-04	1.21e-04	0.024	3/4
MECP2 regulates transcription of neuronal ligands	2/13	9.1e-04	1.68e-04	0.024	2/8
Trafficking of GluR2-containing AMPA receptors	2/23	0.002	5.23e-04	0.042	3/3
Estrogen-dependent nuclear events downstream of ESR membrane signaling	2/29	0.002	8.27e-04	0.042	1/12
Neurodegenerative diseases	2/30	0.002	8.84e-04	0.042	2/22
Deregulated CDK5 triggers neurodegenerative pathways in AD	2/30	0.002	8.84e-04	0.042	2/22
Inflammasomes	2/33	0.002	0.001	0.044	5/28
Trafficking of AMPA receptors	2/37	0.003	0.001	0.046	4/4
Glutamate binding, activation of AMPA receptors	2/39	0.003	0.001	0.046	9/9
Beta catenin independent WNT signaling	3/166	0.012	0.002	0.046	6/51

Table 7.5 shows the specific pathways which the selected genes obtained from young mice take roles.

BH3-only proteins (BCL-2 homology domain 3) of the BCL-2 family are the orchestrating cell death, surveillants of cellular stress by means of apoptosis in neurons (Saleem et al., 2018). MeCP2 is a methylated-DNA-binding protein and errors cause autism spectrum disorder (Cheng and Qiu, 2014). AMPA receptors (AMPARs) interfered the numerous of fast excitatory synaptic transmission in the brain (Wang, Gilbert and Man, 2012).

As can be seen in Figure 7.4, the selected genes from young mice play roles in signal transduction, immune system extensively.

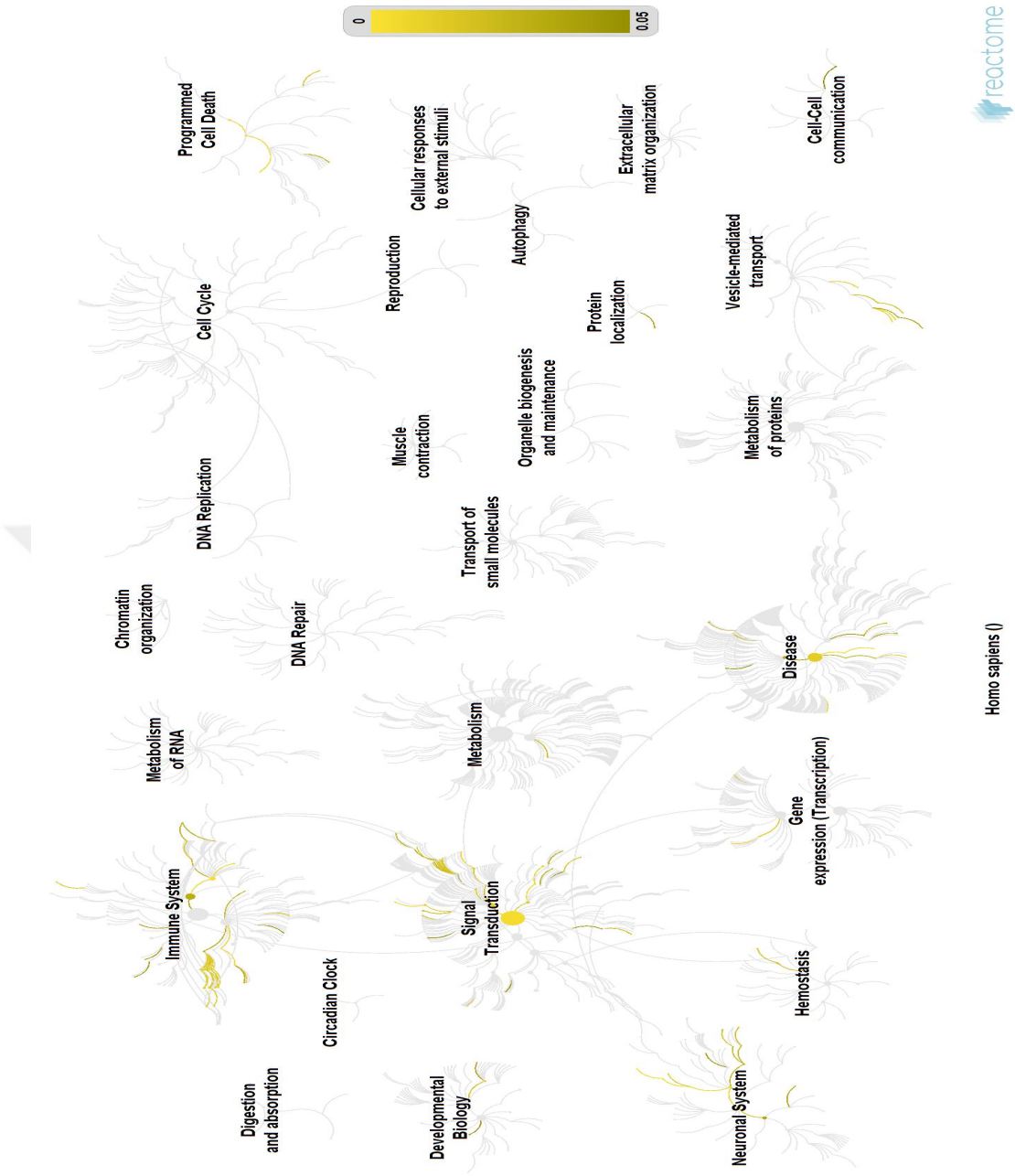

Figure 7.4: Pathway visualization of selected genes from young mice.

7.6 Pathway Analysis of Old Mice

Table 7.6 shows the specific pathways where the selected genes obtained from old mice play roles. ERBB4 (v-erb-a Erithroblastic Leukemia Viral Oncogene Homology 4), ERBB2 (v-erb-b2 Avian Erythroblastic Leukemia Viral Oncogene Homolog 2), PTK6 (Protein Tyrosine Kinase 6) receptors play roles in pathways of genes selected

Table 7.6: Specific pathways of selected genes from old mice.

Pathway Name	Entities				Reactions
	found	ratio	p-value	FDR	found
Downregulation of ERBB4 signaling	4/11	7.6e-04	2.00e-09	1.04e-06	5/5
Signaling by ERBB4	5/82	0.006	1.12e-07	2.15e-05	52/52
Long-term potentiation	4/31	0.002	1.24e-07	2.15e-05	1/7
Downregulation of ERBB2 signaling	4/36	0.002	2.24e-07	2.91e-05	9/14
Nuclear signaling by ERBB4	4/47	0.003	6.44e-07	6.70e-05	34/34
ERBB2 activates PTK6	3/18	0.001	2.54e-06	1.70e-04	2/2
ERBB2 modulates cell motility	3/19	0.001	2.98e-06	1.70e-04	2/2
Signaling by non-receptor tyrosine kinases	4/71	0.005	3.28e-06	1.70e-04	3/53
Signaling by PTK6	4/71	0.005	3.28e-06	1.70e-04	3/53
SHC1 events in ERBB4	3/21	0.001	4.02e-06	1.73e-04	4/4

from old mice. Changes in the ErbB4 signaling pathway lead to a variety of neurodevelopmental deficiencies including deficiencies in synaptic plasticity and neuronal migration (Perez-Garcia, 2015). Protein tyrosine kinase 6 (PTK6), also called as breast tumor kinase BRK (Breast Tumor Kinase), is a member of a specific family of kinases that is relevant to the SRC (Sarcoma) family of tyrosine kinases (Brauer and Tyner, 2010). PTK6 enhances growth factor signaling (Brauer and Tyner, 2010). As seen as Figure 7.5, genes selected from old mice play roles in signal transduction.

As can be seen in Figure 7.5, the selected genes from old mice play roles in disease, immune system, signal transduction, DNA repair, disease pathway, cellular responses to external stimuli and programmed cell death. Compared to selected genes from young mice, genes selected from old mice participate DNA repair and cellular responses to external stimuli pathways.

Figure 7.5: Pathway visualization of selected genes from old mice.

8. CONCLUSIONS

In this thesis, critical proteins in DS are determined using machine learning algorithms. The protein profiles of different datasets are analyzed by applying biochemical techniques in laboratory. However, the list of analyzed proteins is long and not all proteins in list are not related to DS. Thus, it is required to decrease the long list in a meaningful and important list. In order to obtain subset of critical proteins and differentiate healthy and unhealthy mice based on the selected subset, machine learning algorithms are applied in this thesis. In previous works, changes in protein expression are determined. In this thesis, the protein subsets which discriminate classes of mice more accurately are found for treatment of DS. These proteins are very important in order to understand causes and cure of DS. The biological processes can be understood by analyzing the pathways on which the selected proteins affect one by one or aggregately. The selected proteins can be effective in specific DS aspects such as ID and affects motor, cognitive, linguistic, personal or social skills. Thus, evaluation of proteins is important in order to understand the causes of different DS aspects. After understanding the cause of the DS, the treatments can be possible by developing the effective drugs.

In this thesis, different machine learning algorithms are applied to different datasets. These datasets contain protein expression profiles of mice that are trained in CFC with and without injection of memantine. In CFC experiment, CS mice have learning capacity and SC mice have not learning capacity. Protein responses after CFC have been reported. The trisomic (such as Ts6Dn, Tc1) CS group of mice cannot to learn. However, if the trisomic CS group of mice is injected with drug, learning can be rescued.

Different machine learning algorithms are used in steps of feature selection and classification of mice. In feature selection step, the critical protein subsets are selected using forward feature selection technique. In classification step, healthy and unhealty mice are differentiated based on selected protein subsets. Before applying different machine learning algorithms, preprocessing step that consists of handling missing values in datasets and normalization of dataset is carried out. The replacement method used in handling missing value is different from previous studies. In the previous studies, missing values were replaced with the average value of all protein expression levels in same class of mice and effect of dilution ratio did not consider. 15 tissue samples that are three replicates of a five-point dilution series were obtained per mouse. The effect of dilution ratio is considered in this work and missing values are replaced with the average expression value of equivalent sample in same class mice. In addition to replacing missing parts, all measurements are normalized with Z-score normalization. It prevents proteins with higher amounts influence on the classification result erroneously. Since Z-score normalization preserves range (maximum and minimum), Z-score normalization is applied rather than max-min normalization that is used in other works.

The feature selection method, named as forward feature selection, is applied using KNIME tool. It is the heuristic method which tries to detect the ideal feature subset by iteratively choosing features based on the classifier achievement. The method begins with empty feature subset and adds one feature at a time for each round. This one feature is taken from the all features pool that are not in the feature subset. When it is added into feature subset, best classifier result is obtained. The process is reiterated until the desired number of features are added. Forward feature selection is applied. For the learning process in KNIME, naive Bayes learner which is efficient for multi classification problem is used. In spite of the underlying assumption of conditional independence, naive Bayes performs well with more than two classes problem. In previous studies, the applied algorithms suffered from an efficient multiclass classification technique. In this thesis, this deficiency is eliminated with naive Bayes algorithm in forward feature selection.

After selecting features, classification methods are applied for differentiating mice in different subgroups. We carried out four classification methods: DNN, gradient boosted tree, random forest and SVM. These classification methods are implemented by using Python and Scikit Learn package. In order to select the most appropriate parameters of classification methods, grid search method is applied. For building robust and reliable classification model, $5-fold$ cross validation is applied. In $K-fold$ cross validation, the data is splitted into k subsets. Only one of these subsets is utilized as the test set and the others are constituted to a training set at each time. This procedure is repeated k times. The failure estimation is averaged over all k trials to obtain whole efficiency. This way greatly decreases bias since most of the data are utilized for fitting. It also greatly diminishes discrepancy as most of the data is also being utilized in validation set.

By applying feature selection method, the protein subset is determined. Based on critical proteins in selected protein subset, the healthy and unhealthy mice are separated by applying different classification methods. Compared to previous studies, the selected feature subsets in our works provide more accurate class separation of mice. This substantiates importance of the selected feature subsets in different cases. Also, some proteins selected in this thesis were not determined before. After selecting protein subsets, the pathway analyses are done using reactome Pathway Browser and importance and visualization of selected proteins within a pathway are observed.

In the classification step, firstly, important proteins in DS are found by applying different classification methods: NN, Random Forest and SVM. The classification results are compared with the results of B.Feng et al. work where AdaBoost algorithm was used for selecting 30 features and NN, SVM and Random Forest methods were used for mice classification. Random Forest and SVM classification models applied in our work achieved higher accuracy when compared to B.Feng's work. DNN gave the highest accuracy result. Results of DNN are not compared as B.Feng et al. did not apply DNN.

Also, the systematic analysis is done for determining feature subsets for three cases: succesful learning, rescued learning with drug, failed learning. DNN, Gradient Boosted Tree, Random Forest and SVM classification methods are implemented. To understand which changes in protein expression levels are required for successful learning, all groups of normal mice are inspected in the first case. For determining important proteins in rescued learning, trisomic mice exposed to CFC with and without memantine are analyzed in the second case. The third case finds out important protein abnormalities in failed learning case by comparing expression levels of normal and trisomic mice. In the first case, feature subset (SOD1, Ubiquitin, pGSK3B, S6, CaNA, IL1B, BAX, pNR2A, BDNF, pJNK, pCFOS) is selected from control group mice. Control group mice with and without memantine treatment and with and without CFC stimulation are analyzed. When compared with Higuera's work, there are 4 common proteins out of selected 11 proteins. The selected feature subset gives higher accuracy results for all classification techniques and SVM gives the highest accuracy. When the selected proteins are evaluated in pathway analysis, it is shown that they have important roles in immune system, signal transduction and gene expression extensively. Also, these gene products play roles in programmed cell death, cellular responses to external stimuli, disease and DNA repair mechanisms. In the second case, for understanding the important proteins in rescued learning, features are selected from data consisting of trisomic mice which are exposed to CFC with and without memantine. The selected proteins are BRAF, S6, CDK5, BDNF, pCREB, PKCA, SOD1, PSD95, pNR2A. There are 2 common proteins with previous work. The selected proteins play roles in immune system, signal transduction, neural system, gene expression and disease pathways. PSD95 is a scaffold protein and a regulator of synaptic strenght. Thus, it is inferred that drug can be effective for stabilizing synapse structure. DNN and SVM give highest accuracy. The accuracy results of the selected feature subset are higher than previous work for all classification methods. In the third case, to pinpoint proteins that are critical in unsuccessful learning, features are selected from protein expression levels of trisomic and control mice. Selected proteins are P38, pPKCAB, CAMKII, pCAMKII, GluR3, DSCR1, nNOS, BAX,pCFOS, ERK. Selected proteins take part in immune

system, signal transduction, neural system and gene expression broadly. Also, they take actions in programmed cell death and cellular response to external stimuli. DNN and SVM give highest accuracy results and classification results of the selected feature subsets are higher than previous work for all classification methods.

Later, in order to understand whether different drugs-treated Ts65Dn mice exhibit similar response or not, expression profiles of two drugs (RO4938581 and memantine) are analyzed. 4 gene products are expressed in common with RO4938581 and memantine. The selected first two proteins (BRAF and S6) are common with two drugs. When designing drugs for treatment of DS, special attention to these two genes must be considered. Also, MEK and SHH are selected from RO4938581 dataset. MEK is important in cell cycle and division. Neural cells can not divided. Thus, it is speculated that glial cells may be effective and increase of glial cells highlights the importance of MEK. Also, SHH can provide the settlement of glial cells.

After the analysis of different drug responses, the critical protein subsets which display the regional fluctuation with aging are determined. These subsets are obtained from two datasets which show the expression profiles of young and old mice at three different brain regions. The selected proteins from old mice dataset are RCAN1r, Ubiquitin, APP, TH, ARC, ERK, mTOR, ERBB4, H3MeK4, pNR2A. Ubiquitin is a turnover protein and a good indicator in aging process. Also, DS increases the risk of AD at later stage of life. Approximately all adults with DS display the neuropathological modifications of AD by the age of 40 years. Thus, APP builds up in the brain through the lifetime of people with DS and contributes a special change to grasp the temporal advancement of AD. The selected proteins from young mice dataset are MEK, APP, ITSN1, GluR3, pGJA1, P3525, DYRKA1, AKT, pPKCG, CTTNB1, NUMB, PRMT2, BDNF, BCL2. By looking into the two different subsets for old mice and young mice, the aging process of people with DS can be understood. The selected genes from young mice play roles in signal transduction, immune system extensively. The selected genes from old mice play

roles in disease, immune system, signal transduction, DNA repair, disease pathway, cellular responses to external stimuli and programmed cell death. Compared to selected genes from young mice, genes selected from old mice participate DNA repair and cellular responses to external stimuli pathways.

In addition to the importance of drug treatment and age, the protein expression profiles on different parts of brain are also crucial for mechanism of DS. All cells do not synthesize proteins even though all are brain cells. When some parts of brain in healthy mice synthesize specific protein, the other parts of unhealthy mice can synthesize this protein. Proteins cannot be effective in wrong place and cause problems. Thus, analysis of the protein expression profile on different parts of brain can be crucial. Cortex is the most extremely matured unit of the human brain and chargeable for understanding language, thinking and perceiving. The expression levels of proteins which change critically in subcellular fractions of cortex are determined as genes are differentially expressed in cytosolic and nuclear fractions of cortex. Thus, the compositional differences can be crucial for analyzing DS. The selected critical proteins are related to different pathways and processes, such as MAPK and MTOR signaling pathways, AD, neurotrophin signaling pathway.

Thanks to the different type of mice, expression profiles of different proteins can be analyzed. The different type of mice maps the different parts of chromosome 21. Thus, it is very important to inspect more than one mouse in order to analyze expressed genes in DS. In this thesis, expression profiles of Ts65Dn and Tc1 mice are inspected. Most of the selected proteins obtained from Tc1 mice do not change when compared to Ahmed et al.'s work (2013). The selected four proteins (nNOS, DYRK1A, SOD1 and APP) change significantly in one or more brain regions.

When the selected protein subsets are analyzed in pathway analyses, it is monitored that selected proteins have vital roles in the processes, such as apoptosis, learning and memory, signaling pathways, disease pathways, immune system and AD.

In conclusion, this thesis identifies the critical protein subsets that seperate healthy and unhealty mice more accurately. The contribution of this work is the application of different steps to protein expression datasets. The preprocessing steps, feature selection and classification techniques are applied in a different way. This difference provides to differentiate healthy and unhealthy mice more accurately. The obtained higher classification accuracies for all classification methods substantiate the efficieny of different processing steps. When the pathway analyses for selected protein subsets are done, it is realized that selected proteins have important roles in biological processes. Also, these protein subsets can be critical in specific DS aspects such as ID and affects motor, immunity, cognitive, linguistic, personal or social skills. Thus, the evaluation of proteins can be important to understand the causes of different aspects for DS. After understanding the causes, the treatments can be possible by developing the effective drugs. Recently, analyses of non-coding RNAs are important in theraphy of diseases. Thus, in the future, non-coding RNAs of selected proteins can be evaluated in order to understand the causes of DS.

REFERENCES

A. McCombe, P., and D. Henderson, R. (2011). The Role of Immune and Inflammatory Mechanisms in ALS. *Current Molecular Medicine*, 999(999), pp.1-9.

Abdi, H., and Lynne, J. (2010). Normalizing Data. *In: Encyclopedia of research design.* Thousand Oaks, pp.935-938.

Adams, J., and Cory, S. (2007). Bcl-2-regulated apoptosis: mechanism and therapeutic potential. *Current Opinion in Immunology*, 19(5), pp.488-496.

Ahmed, AH., Wang, Q., Sondermann, H. and Oswald, RE. (2009).Structure of the S1S2 Glutamate Binding Domain of GluR3. *Proteins*, 75(3), pp. 628–637.

Ahmed, M., Block, A., Tong, S., Davisson M.T. and Gardiner, K. (2017). Age exacerbates abnormal protein expression in a mouse model of Down syndrome. *Neurobiol Aging*,57, pp. 120-132.

Ahmed, M., Dhanasekaran, A., Block, A., Tong, S., Costa, A., Stasko, M. and Gardiner, K. (2015). Protein Dynamics Associated with Failed and Rescued Learning in the Ts65Dn Mouse Model of Down Syndrome. *PLOS ONE*, 10(3), p.e0119491.

Ahmed, M., Dhanasekaran, A., Block, A., Tong, S., Costa, A. and Gardiner, K. (2014). Protein Profiles Associated With Context Fear Conditioning and Their Modulation by Memantine. *Molecular & Cellular Proteomics*, 13(4), pp.919-937.

Ahmed, M., Dhanasekaran, A., Block, A., Tong, S., Costa, A. and Gardiner, K. (2013). Protein Profiles in Tc1 Mice Implicate Novel Pathway Perturbations in the Down Syndrome Brain. *Human Molecular Genetics*, 2013, Vol. 22, No. 9.

Alasadi, S. and Bhaya, W. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 12(16), pp.4102-4107.

Al-Moghyrah, H. (2017). Assistive Technology Use for Students with Down syndrome at Mainstream Schools in Riyadh, Saudi Arabia: Teachers' Perspectives. *Journal of Education and Practice*, 8(33).

Aly, M. (2005). Survey on Multiclass Classification Methods.

Antonarakis, S., Lyle, R., Dermitzakis, E., Reymond, A. and Deutsch, S. (2004). Chromosome 21 and Down syndrome: from genomics to pathophysiology. *Nature Reviews Genetics*, 5(10), pp.725-738.

Antonarakis, S., Lyle, R., Dermitzakis, E., Reymond, A. and Deutsch, S. (2019). Chromosome 21 and Down syndrome: from genomics to pathophysiology. [online] Semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Chromosome-21-and-Down-syndrome%3A-from-genomics-to-Antonarakis-Lyle/3a54d36cc7f7134d188e2b7670b08532c65b2bd7/figure/2 [Accessed 28 Sep. 2019].

Barbee C, Chicago.cbslocal.com. (2019). Big News: Toddler With Down Syndrome Stars In OshKosh Ad And More. [online] Available at: https://chicago.cbslocal.com/2016/12/05/big-news-toddler-with-down-syndrome-stars-in-oshkosh-ad-and-more/ [Accessed 28 Sep 2019].

Bergstra, J., Bardenet, R., Bengio, Y. and Kegl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, pp.2546-2554.

Bernal, J. and Briceno, I. (2006). Genetic and other diseases in the pottery of Tumaco-La Tolita culture in Colombia-Ecuador. *Clinical Genetics*, 70(3), pp.188-191.

Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Thiel, K. and Wiswedel, B. (2009). KNIME - the Konstanz information miner. *ACM SIGKDD Explorations Newsletter*, 11(1), p.26.

Beurel, E., Grieco, SF. and Jope, RS.(2015) Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases. *Pharmacology therapeutics*, pp.14–131.

Block, A., Ahmed, M., Rueda, N., Hernandez, M., Martinez-Cué, C. and Gardiner, K. (2018). The GABA A α5-selective Modulator, RO4938581, Rescues Protein Anomalies in the Ts65Dn Mouse Model of Down Syndrome. *Neuroscience*, 372, pp.192-212.

Braudeau, J., Delatour, B., Duchon, A., Pereira, P., Dauphinot, L., de Chaumont, F., Olivo-Marin, J., Dodd, R., Hérault, Y. and Potier, M. (2011). Specific targeting of the GABA-A receptor α5 subtype by a selective inverse agonist restores cognitive deficits in Down syndrome mice. *Journal of Psychopharmacology*, 25(8), pp.1030-1042.

Brauer, P. and Tyner, A. (2010). Building a better understanding of the intracellular tyrosine kinase PTK6 — BRK by BRK. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1806(1), pp.66-73.

Breiman, L. (1997). Arcing The Edge. *Technical Report 486*. Statistics Department, University of California, Berkeley.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5-32.

Busciglio, J., Capone, G., O'Byran, J. and Gardiner, K. (2013). Down Syndrome: Genes, Model Systems, and Progress towards Pharmacotherapies and Clinical Trials for Cognitive Deficits. *Cytogenetic and Genome Research*, 141(4), pp.260-271.

Calabrese, F., Rossetti, AC., Racagni, G., Gass, P., Riva, MA., Molteni, R. (2014). Brain-derived neurotrophic factor: a bridge between inflammation and neuroplasticity. *Frontiers in Cellular Neuroscience*, 8:430.

Chang, Q. and Gold, P. (2008). Age-related changes in memory and in acetylcholine functions in the hippocampus in the Ts65Dn mouse, a model of Down syndrome. *Neurobiology of Learning and Memory*, 89(2), pp.167-177.

Chapman, R. and Hesketh, L. (2000). Behavioral phenotype of individuals with Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, 6(2), pp.84-95.

Chen, H. and Lipton, S. (2005). Pharmacological Implications of Two Distinct Mechanisms of Interaction of Memantine with N-Methyl-d-aspartate-Gated Channels. *Journal of Pharmacology and Experimental Therapeutics*, 314(3), pp.961-971.

Chen, H., Chrast, R., Rossier, C., Morris, M., Lalioti, M. and Antonarakis, S. (1996). Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. *Genome Research*, 6(8), pp.747-760.

Chen, J., Huang, H., Tian, S. and Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3), pp.5432-5435.

Chen, Y. and Qiu, X. (2013). Ubiquitin at the crossroad of cell death and survival. *Chinese Journal of Cancer*, 32(12), pp.640-647.

Cheng, T. and Qiu, Z. (2014). MeCP2: multifaceted roles in gene regulation and neural development. *Neuroscience Bulletin*, 30(4), pp.601-609.

Choi, J. (2010). Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy: A Genetic Cause of Cerebral Small Vessel Disease. *Journal of Clinical Neurology*, 6(1), p.1.

Cooper, M. and Koleske, A. (2014). Ablation of ErbB4 from excitatory neurons leads to reduced dendritic spine density in mouse prefrontal cortex. *Journal of Comparative Neurology*, 522(14), pp.3351-3362.

Cooper, N. (2019). Robertsonian translocation. [online] Available at: https:

//mymds.bham.ac.uk/genetics/robtrans.htm [Accessed 28 Sep. 2019].

Corrales, A., Martínez, P., García, S., Vidal, V., García, E., Flórez, J., Sanchez Barceló, E., Martínez-Cué, C. and Rueda, N. (2013). Long-term oral administration of melatonin improves spatial learning and memory and protects against cholinergic degeneration in middle-aged Ts65Dn mice, a model of Down syndrome. *Journal of Pineal Research*, 54(3), pp.346-358.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp.273-297.

Costa, A. (2011). On the Promise of Pharmacotherapies Targeted at Cognitive and Neurodegenerative Components of Down Syndrome. *Developmental Neuroscience*, 33(5), pp.414-427.

Costa, A. and Grybko, M. (2005). Deficits in hippocampal ca1 ltp induced by tbs but not hfs in the ts65dn mouse: a model of down syndrome. *Neurosci Lett.*, 382, p.3.

Costa, A., Scott-McKean, J. and Stasko, M. (2007). Acute Injections of the NMDA Receptor Antagonist Memantine Rescue Performance Deficits of the Ts65Dn Mouse Model of Down Syndrome on a Fear Conditioning Test. *Neuropsychopharmacology*, 33(7), pp.1624-1632.

Cunha, C., Brambilla, R. and Thomas, KL.(2010). A Simple Role for BDNF in Learning and Memory? *Frontiers in Molecular Neuroscience*, 3:1.

Dahmane, N., Ghezala, G., Gosset, P., Chamoun, Z., Dufresne-Zacharia, M., Lopes, C., Rabatel, N., Gassanova-Maugenre, S., Chettouh, Z., Abramowski, V., Fayet, E., Yaspo, M., Korn, B., Blouin, J., Lehrach, H., Poutska, A., Antonarakis, S., Sinet, P., Créau, N. and Delabar, J. (1998). Transcriptional Map of the 2.5-Mb CBR–ERG Region of Chromosome 21 Involved in Down Syndrome. *Genomics*, 48(1), pp.12-23.

Das, I., Park, J., Shin, J., Jeon, S., Lorenzi, H., Linden, D., Worley, P. and Reeves, R. (2013). Hedgehog Agonist Therapy Corrects Structural and Cognitive Deficits in a Down Syndrome Mouse Model. *Science Translational Medicine*, 5(201), pp.201ra120-201ra120.

Daunhauer, L. and Fidler, D. (2011). The Down Syndrome Behavioral Phenotype: Implications for Practice and Research in Occupational Therapy. *Occupational Therapy In Health Care*, 25(1), pp.7-25.

Davies, J. (2019). RO4938581 | Ligand page | IUPHAR/BPS Guide to PHARMACOLOGY. [online] Available at: https://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=4299 [Accessed 28 Sep. 2019].

Davisson, M., Schmidt, C., and Akeson, E. (1990). Segmental trisomy of murine chromosome 16: a new model system for studying down syndrome. *Prog Clin Biol Res*, 360, pp. 263–280.

Davisson, M., Schmidt, C., Reeves, R., Irving, N., Akeson, E., Harris, B. and Bronson, R. (1993). Segmental trisomy as a mouse model for Down syndrome. *Progress in clinical and biological research*, 384(117), p.33.

Dbouk, H., Mroue, R., El-Sabban, M. and Talhouk, R. (2009). Connexins: a myriad of functions extending beyond assembly of gap junction channels. *Cell Communication and Signaling*, 7(1), p.4.

Dechant, G. and Barde, Y. (2002). The neurotrophin receptor p75NTR: novel functions and implications for diseases of the nervous system. *Nature Neuroscience*, 5(11), pp.1131-1136.

Demas, G., Nelson, R., Krueger, B., and Yarowsky, P. (1998). Impaired spatial working and reference memory in segmental trisomy (ts65dn) mice. *Behav Brain Res*, 90, pp.199–201.

Dinarello, C. (2011). Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. *Blood*, 117(14), pp.3720-3732.

Dong, F., Jiang, J., McSweeney, C., Zou, D., Liu, L. and Mao, Y. (2016). Deletion of CTNNB1 in inhibitory circuitry contributes to autism-associated behavioral defects. *Human Molecular Genetics*, p.ddw131.

Down, J. (1867). Observations on an Ethnic Classification of Idiots. *Journal of Mental Science*, 13(61), pp.121-123.

DrugCentral. (2019). memantine. [online] Available at: http://drugcentral.org / drugcard/1679 [Accessed 28 Sep. 2019].

DrugsDetails. (2019). Memantine | Drugs Details. [online] Available at: https: //drugsdetails.com/memantine/ [Accessed 28 Sep. 2019].

Dua, D. and Graff, C. (2017). Machine Learning Repository. [online] Archive. ics.uci.edu. Available at: http://archive.ics.uci.edu/ml [Accessed 28 Sep. 2019].

Duan, Y., Liu, T., Li, S., Huang, M., Li, X., Zhao, H. and Li, J. (2019). CHAF1B promotes proliferation and reduces apoptosis in 95-D lung cancer cells and predicts a poor prognosis in non-small cell lung cancer. *Oncology Reports*.

Duchon, A., Raveau, M., Chevalier, C., Nalesso, V., Sharp, A., and Herault, Y. (2011). Identification of the translocation breakpoints in the ts65dn and ts1cje mouse lines: relevance for modeling down syndrome. *Mamm Genome* 22, pp. 674–84.

Eicher, T. and Sinha, K. (2017). A support vector machine approach to identification of proteins relevant to learning in a mouse model of Down Syndrome. *In: International Joint Conference on Neural Networks*.

Escorihuela, R., Fernandez-Teruel, A., Vallina, I., Baamonde, C., Lumbreras, M., and Dierssen, M. (1995). A behavioral assessment of ts65dn mice: a putative down syndrome model. *Neurosci Lett.*, 199, pp. 143–146.

Escorihuela, R., Vallina, I., C Martinez-Cue, C. B., Tobena, A., and Dierssen, M. (1998). Impaired shortand long-term memory in ts65dn mice, a model for down syndrome. *Neurosci Lett.* 247, pp. 171–174.

Esplugues, J. (2002). NO as a signalling molecule in the nervous system. *British Journal of Pharmacology*, 135(5), pp.1079-1095.

Ezza, H. and Khadrawyb, Y. (2014). Glutamate Excitotoxicity and Neurodegeneration. *Journal of Molecular and Genetic Medicine*, (8), p.141.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., Rothfels, K., May, B. and Milacic, M. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research*, 44(D1), pp.D481–D487,.

Faizi, M., Bader, P., Tun, C., Encarnacion, A., Kleschevnikov, A., and Belichenko, P. (2011). Comprehensive behavioral phenotyping of ts65dn mouse model of down syndrome: activation of beta1-adrenergic receptor by xamoterol as a potential cognitive enhancer. *Neurobiol Dis*, 43, pp. 397–413.

Fanselow, M. (1990). Factors governing one-trial contextual conditioning. *Animal Learning & Behavior*, 18(3), pp.264-270.

Feng, B., Hoskins, W., Zhou, J., Xu, X. and Tang, J. (2017). Using Supervised Machine Learning Algorithms to Screen Down Syndrome and Identify the Critical Protein Factors. *In: International Conference on Intelligent and Interactive Systems and Applications*.

Fernandez, F., Morishita, W., Zuniga, E., Nguyen, J., Blank, M., and Malenka, R. (2007). Pharmacotherapy for cognitive impairment in a mouse model of down syndrome. *Nat Neurosci*, pp. 10, 411–413.

Fidler, D., Hepburn, S. and Rogers, S. (2006). Early learning and adaptive behaviour in toddlers with Down syndrome: Evidence for an emerging behavioural phenotype?. *Down Syndrome Research and Practice*, 9(3), pp.37-44.

Filippo, M. D., Tozzi, A., Ghiglieri, V., Picconi, B., Costa, C., and Cipriani, S. (2010). Impaired plasticity at specific subset of striatal synapses in the

ts65dn mouse model of down syndrome. *Biol Psychiatry.* 67, pp. 666–671.

Flavell, S. and Greenberg, M. (2008). Signaling Mechanisms Linking Neuronal Activity to Gene Expression and Plasticity of the Nervous System. *Annual Review of Neuroscience*, 31(1), pp.563-590.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5), pp.1189-1232.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), pp.367-378.

Furat, FG. and Ibrikci, T. (2018). Classification of Down Syndrome of Mice Protein Dataset on MongoDB Database. *Balkan Journal of Electrical and Computer Engineering.*, 6(2),vpp.112-117.

Galante, M., Jani, H., and Vanes, L. (2009). Impairments in motor coordination without major changes in cerebellar plasticity in the tc1 mouse model of down syndrome. *Hum Mol Gene*, 18, pp.1449–1463.

Gallo, F., Katche, C., Morici, J., Medina, J. and Weisstaub, N. (2018). Immediate Early Genes, Memory and Psychiatric Disorders: Focus on c-Fos, Egr1 and Arc. *Frontiers in Behavioral Neuroscience*, 12.

Garcia-Cerro, S., Martinez, P., Vidal, V., Corrales, A., Florez, J., and Vidal, R. (2014). Overexpression of dyrk1a is implicated in several cognitive, electrophysiological and neuromorphological alterations found in a mouse model of down syndrome. *PLoS One*, 9(9).

Gardiner, K. (2004). Building protein interaction maps for Down's syndrome. *Briefings in Functional Genomics and Proteomics*, 3(2), pp.142-156.

Gardiner, K. (2010). Molecular basis of pharmacotherapies for cognition in Down syndrome. *Trends in Pharmacological Sciences*, 31(2), pp.66-73.

Gardiner, K. (2014). Pharmacological approaches to improving cognitive function in Down syndrome: current status and considerations. *Drug Design, Development and Therapy*, p.103.

Gardiner, K. and Yaspo, M. (1998). Report of the seventh international workshop on human chromosome 21 mapping 1997. *Cytogenetic and Genome Research*, 82(1-2), pp.1-12.

Gardiner, K., Fortna, A., Bechtel, L. and Davisson, M. (2003). Mouse models of Down syndrome: how useful can they be? Comparison of the gene content of human chromosome 21 with orthologous mouse genomic regions. *Gene*, 318, pp.137-147.

Gardiner, K., Horisberger, M., Kraus, J., Tantravahi, U., Korenberg, J., Rao, V., Reddy, S. and Patterson, D. (1990). Analysis of human chromosome 21:

correlation of physical and cytogenetic maps; gene and CpG island distributions. *The EMBO Journal*, 9(1), pp.25-34.

Gardiner, K., Slavov, D., Bechtel, L. and Davisson, M. (2002). Annotation of Human Chromosome 21 for Relevance to Down Syndrome: Gene Structure and Expression Analysis. *Genomics*, 79(6), pp.833-843.

George, C., Ramaswami, G., Li, J. and Samuel, C. (2016). Editing of Cellular Self-RNAs by Adenosine Deaminase ADAR1 Suppresses Innate Immune Stress Responses. *Journal of Biological Chemistry*, 291(12), pp.6158-6168.

Ghani, S. (2019). Chromosome Abnormality: Down Syndrome | drbeen. [online] Available at: https://www.drbeen.com/blog/chromosome-abnormality-down-syndrome/ [Accessed 28 Sep. 2019].

Globaldownsyndrome.org.(2019). Katheleen Gardiner, PhD. [online] Available at: https://www.globaldownsyndrome.org/our-story/linda-crnic-institute/linda-crnic-institute/katheleen-gardiner-phd/

Han, C., Jeong, M. and Jang, S. (2017). Structure, signaling and the drug discovery of the Ras oncogene protein. *BMB Reports*, 50(7), pp.355-360.

Hao, J. and Ho, T. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, 44(3), pp.348-361.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). 10. Boosting and Additive Trees. *The Elements of Statistical Learning*, pp.1-51.

Hattori, M., Fujiyama, A., Taylor, T., Watanabe, H., Yada, T. and Park, H. (2000). Chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature*, (405), pp.311-319.

Head, E., T. Lott, I., M. Wilcock, D. and A. Lemere, C. (2015). Aging in Down Syndrome and the Development of Alzheimer's Disease Neuropathology. *Current Alzheimer Research*, 13(1), pp.18-29.

Hyde, L., Frisone, D., and Crnic, L. (2001). Ts65dn mice, a model for down syndrome, have deficits in context discrimination learning suggesting impaired hippocampal function. *Behav Brain Res.* 118, pp.53–60.

Heyn, S. (2005). Tc1 - a new down syndrome mouse model. *Stanford Medicine Down Syndrome Research Center News & Views Archives.*5.

Hickey, F., Hickey, E. and Summar, K. (2012). Medical Update for Children With Down Syndrome for the Pediatrician and Family Practitioner. *Advances in Pediatrics*, 59(1), pp.137-157.

Higuera, C., Gardiner, K. and Cios, K. (2015). Self-Organizing Feature Maps

Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome. *PLOS ONE*, 10(6), p.e0129126.

Huang, R. (2018). Age and Gender Classification using Convolutional Neural Networks - ppt download. [online] Slideplayer.com. Available at: https://slideplayer.com/slide/15749019/ [Accessed 28 Sep. 2019].

Hunter, C., Bimonte, H., and Granholm, A. (2003). Behavioral comparison of 4 and 6 month-old ts65dn mice: age-related impairments in working and reference memory. *Behav Brain Res.* 138, pp.121–131.

Johansson, R. (n.d.). An intuitive explanation of gradient boosting. [online] Cse.chalmers.se. Available at: http://www.cse.chalmers.se/ richajo/ dit866/lectures/l8/gbexplainer.pdf [Accessed 28 Sep. 2019].

Kamat, P., Rai, S., Swarnkar, S., Shukla, R., Ali, S., Najmi, A. and Nath, C. (2013). Okadaic acid-induced Tau phosphorylation in rat brain: Role of NMDA receptor. *Neuroscience*, 238, pp.97-113.

Kaushik, S. (2016). Feature Selection methods with example (Variable selection methods). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/ [Accessed 28 Sep. 2019].

Khalid, S., Khalil, T. and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Science and Information Conference.*

Kidambi, S., Yarmush, J., Berdichevsky, Y., Kamath, S., Fong, W. and Schianodi-Cola, J.(2010). Propofol induces MAPK/ERK cascade dependant expression of cFos and Egr-1 in rat hippocampal slices. *BMC Research Notes*,3:201.

Kishnani, P., Heller, J., Spiridigliozzi, G., Lott, I., Escobar, L., Richardson, S., Zhang, R. and McRae, T. (2010). Donepezil for treatment of cognitive dysfunction in children with Down syndrome aged 10-17. *American Journal of Medical Genetics Part A*, 152A(12), pp.3028-3035.

Kleschevnikov, A., Belichenko, P., Faizi, M., Jacobs, L., Htun, K., and Shamloo, M. (2012). Deficits in cognition and synaptic plasticity in a mouse model of down syndrome ameliorated by gabab receptor antagonists. *J Neurosci.* 32, pp. 9217–9227.

Kleschevnikov, A., Belichenko, P., Villar, A., Epstein, C., Malenka, R., and Mobley, W. (2004). Hippocampal long-term potentiation suppressed by increased inhibition in the ts65dn mouse, a genetic model of down

syndrome. *J Neurosci.* 24, pp.8153–8160.

Kliegma, R. (2011). Down Syndrome and Other Abnormalities of Chromosome Number. *Nelson textbook of pediatrics*, pp.Chapter 76.2.

Koehrsen, W. (2017). Random Forest Simple Explanation. [online] https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d. [Accessed 28 Sep. 2019].

Kumar Pareek, T. (2012). Cdk5: An Emerging Kinase in Pain Signaling. Brain *Disorders & Therapy*, 01(s1).

Kumar, V. and Minz, S. (2014). Feature Selection: A literature Review. Smart *Computing Review*, (4), pp.211-229.

Kumin, L. (1996). Speech and language skills in children with Down syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, 2(2), pp.109-115.

Lechner, S., Frenzel, H., Wang, R. and Lewin, G. (2009). Developmental waves of mechanosensitivity acquisition in sensory neuron subtypes during embryonic development. *The EMBO Journal*, 28(10), pp.1479-1491.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), pp.436-444.

Lee, J., Jang, H., Cho, E. and Youn, H. (2009). Down syndrome critical region 1 enhances the proteolytic cleavage of calcineurin. *Experimental and Molecular Medicine*, 41(7), p.471.

Lee, YS., Ehninger, D., Zhou, M., Oh, JY., Kang, M. and Kwak, C. (2014). Mechanism and treatment for the learning and memory deficits associated with mouse models of Noonan syndrome, *Nature neuroscience*,17(12), pp. 1736–1743.

Lejenue, J., Terpin, R. and Gautier, M. (1959). Chromosomic diagnosis of mongolism. *Archives Françaises de Pédiatrie*, (16), pp.962-963.

Leube, R., Wiedenmann, B. and Franke, W. (1989). Topogenesis and sorting of synaptophysin: Synthesis of a synaptic vesicle protein from a gene transfected into nonneuroendocrine cells. *Cell*, 59(3), pp.433-446.

Li, R., Huang, F., Abbas, A. and Wigström, H. (2007). Role of NMDA receptor subtypes in different forms of NMDA-dependent synaptic plasticity. *BMC Neuroscience*, 8(1).

Lin, S., Negulescu, A., Bulusu, S., Gibert, B., Delcros, J., Ducarouge, B., Rama, N., Gadot, N., Treilleux, I., Saintigny, P., Meurette, O. and Mehlen, P. (2017). Non-canonical NOTCH3 signalling limits tumour angiogenesis.

Nature Communications, 8(1).

Lipton, S. (2007). Pathologically-Activated Therapeutics for Neuroprotection: Mechanism of NMDA Receptor Block by Memantine and S-Nitrosylation. *Current Drug Targets*, 8(5), pp.621-632.

Lockrow, J., Boger, H., Bimonte-Nelson, H., and Granholm, A. (2011). Effects of long-term memantine on memory and neuropathology in ts65dn mice, a model for down syndrome. *Behav Brain Res.* 221, pp. 610–622.

Long, J., Maloney, B., Rogers, J. and Lahiri, D. (2018). Novel upregulation of amyloid-β precursor protein (APP) by microRNA-346 via targeting of APP mRNA 5′-untranslated region: Implications in Alzheimer's disease. *Molecular Psychiatry*, 24(3), pp.345-363.

Lysenko, L., Kim, J., Henry, C., Tyrtyshnaia, A., Kohnz, R., and Madamba, F. (2014). Monoacylglycerol lipase inhibitor jzl184 improves behavior and neural properties in ts65dn mice, a model of down syndrome. *PLoS One.* 9(12), p.e114521.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K. and Aboyans, V. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380(9859), pp.2095–128.

Malt, E., Dahl, R., Haugsand, T., Ulvestad, I., Emilsen, N., Hansen, B., Cardenas, Y., Skøld, R., Thorsen, A. and Davidsen, E. (2013). Health and disease in adults with Down syndrome. *Tidsskr Nor Laegeforen.*, 133(3), pp.290-4.

Maraka S, S. and Janku, F. (2018). BRAF alterations in primary brain tumors. *Discovery Medicine*, 26(141), pp.51-60.

Marin, I. and Kipnis, J. (2013). Learning and memory . . . and the immune system, *Learning and Memory.* 20(10), pp. 601–606.

Martinez-Cue, C., Martinez, P., Rueda, N., Vidal, R., Garcia, S., Vidal, V., Corrales, A., Montero, J., Pazos, A., Florez, J., Gasser, R., Thomas, A., Honer, M., Knoflach, F., Trejo, J., Wettstein, J. and Hernandez, M. (2013). Reducing GABAA 5 Receptor-Mediated Inhibition Rescues Functional and Neuromorphological Deficits in a Mouse Model of Down Syndrome. *Journal of Neuroscience*, 33(9), pp.3953-3966.

Mason, L., Baxter, J., Bartlett, P. and Frean, M. (1999). Boosting Algorithms as Gradient Descent. *Advances in Neural Information Processing Systems* 12. MIT Press., pp.512–518.

Milani, P., Gagliardi, S., Cova, E. and Cereda, C. (2011). SOD1 Transcriptional

and Posttranscriptional Regulation and Its Potential Implications in ALS. *Neurology Research International*, 2011, pp.1-9.

Minichiello, L. (2009). TrkB signalling pathways in LTP and learning. Nature *Reviews Neuroscience,* 10(12), pp.850-860.

Moore, F. and Baleja, J. (2012). Molecular remodeling mechanisms of the neural somatodendritic compartment. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1823(10), pp.1720-1730.

Morris, J., Mutton, D. and Alberman, E. (2002). Revised estimates of the maternal age specific live birth prevalence of Down's syndrome. *Journal of Medical Screening*, 9(1), pp.2-6.

Nadel, L. (2003). Down's syndrome: a genetic disorder in biobehavioral perspective. *Genes, Brain and Behavior*, 2(3), pp.156-166.

Nagarajan, R., Hogart, A., Gwye, Y., Martin, M. and LaSalle, J. (2006). Reduced MeCP2 Expression is Frequent in Autism Frontal Cortex and Correlates with Aberrant MECP2 Promoter Methylation. *Epigenetics*, 1(4), pp.172-182.

Nagatsu, T. and Nagatsu, I. (2016). Tyrosine hydroxylase (TH), its cofactor tetrahydrobiopterin (BH4), other catecholamine-related enzymes, and their human genes in relation to the drug and gene therapies of Parkinson's disease (PD): historical overview and future prospects. *Journal of Neural Transmission*, 123(11), pp.1255-1278.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. Fron*tiers in Neurorobotics*, 7.

Neale, C. (2019). Cross Validation: A Beginner's Guide. [online] Medium. Available at: https://towardsdatascience.com/cross-validation-a-beginners-guide-5b8ca04962cd [Accessed 28 Sep. 2019].

Neri, G. and Opitz, J. (2009). Down syndrome: Comments and reflections on the 50th anniversary of Lejeune's discovery. *American Journal of Medical Genetics Part A*, 149A(12), pp.2647-2654.

Nicoll, J., Mrak, R., Graham, D., Stewart, J., Wilcock, G., MacGowan, S., Esiri, M., Murray, L., Dewar, D., Love, S., Moss, T. and Griffin, W. (2000). Association of Interleukin-1 Gene Polymorphisms with Alzheimer's Disease. *Ann Neurol.*, 47(3), pp.365–368.

O'Doherty, A. (2005). An Aneuploid Mouse Strain Carrying Human Chromosome 21 with Down Syndrome Phenotypes. *Science*, 309(5743), pp.2033-2037.

Olivares, D., K. Deshpande, V., Shi, Y., K. Lahiri, D., H. Greig, N., T. Rogers, J. and Huang, X. (2012). N-Methyl D-Aspartate (NMDA) Receptor Antagonists and Memantine Treatment for Alzheimer's Disease, Vascular Dementia and Parkinson's Disease. *Current Alzheimer Research*, 9(6), pp.746-758.

Olson, L., Roper, R., Sengstaken, C., Peterson, E., Aquino, V., and Galdzicki, Z. (2007). Trisomy for the down syndrome 'critical region' is necessary but not sufficient for brain phenotypes of trisomic mice. *Hum Mol Genet.* 16, pp. 774–782.

Ortega-Martı'nez, S. (2015). A new perspective on the role of the CREB family of transcription factors in memory consolidation via adult hippocampal neurogenesis. *Frontiers in Molecular Neuroscience*, 8(46).

Pameer, A. (2019). Genetic abnormalities by Aamir khan pameer. [online] Slideshare.net. Available at: https://www.slideshare.net/AamirPameer/genetic-abnormalities-by-aamir-khan-pameer [Accessed 28 Sep. 2019].

Parker, S., Mai, C., Canfield, M., Rickard, R., Wang, Y., Meyer, R., Anderson, P., Mason, C., Collins, J., Kirby, R. and Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004-2006. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 88(12), pp.1008-1016.

Peixoto, L. and Abel, T. (2012). The Role of Histone Acetylation in Memory Formation and Cognitive Impairments. *Neuropsychopharmacology*, 38(1), pp.62-76.

Pereira, E. (2006). Image: Artificial Neuron models and its parts. Source: Adapted from .... [online] https://www.researchgate. net/publication/229036664. [Accessed 28 Sep. 2019].

Perez-Garcia, C. (2015). ErbB4 in Laminated Brain Structures: A Neurodevelopmental Approach to Schizophrenia. *Frontiers in Cellular Neuroscience*, 9.

Phelps, R. (2010). Helping Children With Down Syndrome Communicate Better: Speech and Language Skills for Ages 6–14. *Journal of Developmental & Behavioral Pediatrics*, 31(1), p.25.

Pollonini, G., Gao, V., Rabe, A., Palminiello, S., Albertini, G. and Alberini, CM. (2008). Abnormal Expression of Synaptic Proteins and Neurotrophin-3 in the Down Syndrome Mouse Model Ts65Dn. *Neuroscience*, 156(1), pp. 99–106.

Prost, E. and Nasreen, R. (2013). Downs: The History of a Disability. *Journal*

of the Canadian Academy of Child and Adolescent Psychiatry, 22(2), pp.180–181.

Rao, AS., Avadhani, PS. and Chaudhuri, NB. (2016) A Content-Based Spam E-Mail Filtering Approach Using Multilayer Percepton Neural Networks. *International Journal of Engineering Trends and Technology*, 41(1).

Rawale, S. (2018). Feature Selection Methods in Machine Learning. [Blog] Available at: https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc [Accessed 28 Sep. 2019].

Reese, LC. and Taglialatela, G. (2011) A Role for Calcineurin in Alzheimer's Disease. *Current Neuropharmacology*, 9(4), pp.685–692.

Reeves, R., Irving, N., Moran, T., Wohn, A., Kitt, C., and Sisodia, S. (1995). A mouse model for down syndrome exhibits learning and behaviour deficits. *Nat Genet.* 11, pp. 177–84.

Reymond, A., Camargo, A., Deutsch, S., Stevenson, B., Parmigiani, R., Ucla, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., de Souza, S., Iseli, C., Jongeneel, C., Bucher, P., Simpson, A. and Antonarakis, S. (2002). Nineteen Additional Unpredicted Transcripts from Human Chromosome 21. *Genomics*, 79(6), pp.824-832.

Reymond, A., Friedli, M., Henrichsen, C., Chapot, F., Deutsch, S., Ucla, C., Rossier, C., Lyle, R., Guipponi, M. and Antonarakis, S. (2001). From PREDs and Open Reading Frames to cDNA Isolation: Revisiting the Human Chromosome 21 Transcription Map. *Genomics*, 78(1-2), pp.46-54.

Rosenblatt, F. (1958). The Perceptron: A Theory of Statistical Separability in Cognitive Systems. Cornell Aeronautical Laboratory, (VG1196-G-1).

Rueda, N., Flórez, J. and Martínez-Cué, C. (2012). Mouse Models of Down Syndrome as a Tool to Unravel the Causes of Mental Disabilities. *Neural Plasticity*, 2012, pp.1-26.

Rueda, N., Llorens-Martin, M., Florez, J., Valdizan, E., Banerjee, P., and Trejo, J. (2010). Memantine normalizes several phenotypic features in the ts65dn mouse model of down syndrome. *J Alzheimers Dis*, 21, pp. 277–290.

Saleem, S., Saha, A., Akhter, R. and Chandra Biswas, S. (2018). Cooperation of BH3-only proteins in killing neurons. *Biomedical Research and Clinical Practice*, 3(2).

Salehi, A., Faizi, M., Colas, D., Valletta, J., Laguna, J., and Takimoto-Kimura, R. (2009). Restoration of norepinephrine-modulated contextual memory

in a mouse model of down syndrome. *Sci Transl Med*, 1(7), p.7ra17.

ScienceDirect. (2014). Forward Selection - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/computer-science/forward-selection [Accessed 28 Sep. 2019].

Shao, CY., Mirra, SS., Sait, HBR., Sacktor, TC. and Sigurdsson, EM.(2011). Postsynaptic degeneration as revealed by PSD−95 reduction occurs after advanced Aβ and tau pathology in transgenic mouse models of Alzheimer's disease. *Acta neuropathologica*, 122(3),pp. 285–292.

Shen, CP., Tsimberg, Y., Salvadore, C. and Meller, E. (2004). Activation of Erk and JNK MAPK pathways by acute swim stress in rat brain regions. *BMC Neuroscience*, 5:36.

Shubham, J. (2018). Ensemble Learning — Bagging and Boosting. [online] Medium. Available at: https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e [Accessed 28 Sep. 2019].

Shupp, A., Casimiro, MC. and Pestell, RG.(2017). Biological functions of CDK5 and potential CDK5 targeted clinical treatments. *Oncotarget*, 8(10), pp. 17373–17382.

Siarey, R., Stoll, J., Rapoport, S., and Galdzicki, Z. (1997). Altered long-term potentiation in the young and old ts65dn mouse, a model for down syndrome. *Neuropharmacology*, 36, pp. 1549–1554.

Silverman, W. (2007). Down syndrome: Cognitive phenotype. *Mental Retardation and Developmental Disabilities Research Reviews*, 13(3), pp.228-236.

Simpson, E. (2003). Sources of estrogen and their importance. *The Journal of Steroid Biochemistry and Molecular Biology*, 86(3-5), pp.225-230.

Sleigh, J. (2019). What Exactly is Down Syndrome?. [online] Camphill.org.za. Available at: http://www.camphill. org.za/ articles/what-exactly-is-down-syndrome [Accessed 28 Sep. 2019].

Stasko, M. and Costa, A. (2004). Experimental parameters affecting the morris water maze performance of a mouse model of down syndrome. *Behav Brain Res* 154, pp. 1–17.

Sturgeon, X. and Gardiner, K. (2011). Transcript catalogs of human chromosome 21 and orthologous chimpanzee and mouse regions. *Mammalian Genome*, 22(5-6), pp.261-271.

Sturgeon, X., Le, T., Ahmed, M. and Gardiner, K. (2012). Pathways to cognitive

deficits in Down syndrome. *Progress in Brain Research*, (197), pp.73-100.

Sujashvili, R. (2016) Advantages of Extracellular Ubiquitin in Modulation of Immune Responses. *Mediators of Inflammation*, 2016:4190390.

Sun, L. and Zhao, Y. (2007). The Biological Role of Dectin-1 in Immune Response. *International Reviews of Immunology*, 26(5-6), pp.349-364.

Tano, T., Okamoto, M., Kan, S., Nakashiro, K., Shimodaira, S., Koido, S. (2013). Prognostic Impact of Expression of Bcl-2 and Bax Genes in Circulating Immune Cells Derived from Patients with Head and Neck Carcinoma. *Neoplasia* (New York, NY),15(3), pp. 305–314.

Tassone, F., Xu, H., Burkin, H., Weissman, S. and Gardiner, K. (1995). CDNA selection from 10 Mb of Chromosome 21 DNA: efficiency in transcriptional mapping and reflections of genome organization. *Human Molecular Genetics*, 4(9), pp.1509-1518.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. and Kornblau, S. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics*, 5(10), pp.2512-2521.

Tsoumakas, G. and Katakis, I. (2007). Multi-Label Classification. *International Journal of Data Warehousing and Mining*, 3(3), pp.1-13.

Wang, G., Gilbert, J. and Man, H. (2012). AMPA Receptor Trafficking in Homeostatic Synaptic Plasticity: Functional Molecules and Signaling Cascades. *Neural Plasticity*, 2012, pp.1-12.

Wang, T., Wen, C., Wang, H., Gao, F., Jiang, T. and Jin, S. (2017). Deep learning for wireless physical layer: Opportunities and challenges. *China Communications*, 14(11), pp.92-111.

Weijerman, M. and de Winter, J. (2010). Clinical practice. *European Journal of Pediatrics*, 169(12), pp.1445-1452.

Winders, P., Wolter-Warmerdam, K. and Hickey, F. (2018). A schedule of gross motor development for children with Down syndrome. *Journal of Intellectual Disability Research*, 63(4), pp.346-356.

Wong, T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), pp.2839-2846.

Yu, T., Li, Z., Jia, Z., Clapcote, S., Liu, C., Li, S., Asrar, S., Pao, A., Chen, R., Fan, N., Carattini-Rivera, S., Bechard, A., Spring, S., Henkelman, R., Stoica, G., Matsui, S., Nowak, N., Roder, J., Chen, C., Bradley, A. and Yu,

Y. (2010). A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. *Human Molecular Genetics*, 19(14), pp.2780-2791.

Yu, X., Lu, L., Liu, Z., Yang, T., Gong, X., Ning, Y.(2016).Brain-derived neurotrophic factor modulates immune reaction in mice with peripheral nerve xenotransplantation. *Neuropsychiatric Disease and Treatment*, 12, pp.685–694.

Zhang, G., Liu, M., Cao, H., Kong, L., Wang, X. and O'Brien, JA.(2009). Improved spatial learning in aged rats by genetic activation of protein kinase C in small groups of hippocampal neurons, *Hippocampus,* 19(5), pp. 413–423.

Zídek, Z., Anzenbacher, P. and Kmoníčková, E. (2009). Current status and challenges of cytokine pharmacology. *British Journal of Pharmacology*, 157(3), pp.342-361.

Zorumski, C. and Izumi, Y. (2012). NMDA receptors and metaplasticity: Mechanisms and possible roles in neuropsychiatric disorders. *Neuroscience & Biobehavioral Reviews*, 36(3), pp.989-1000.

APPENDIX A: DATASET OF MICE PROTEIN EXPRESSION

Table A.1: The first twelve columns for two mice in dataset

MouseID	DYRK1A	ITSN1	BDNF	NR1	NR2A	pAKT	pBRAF	pCAMKII	pCREB	pELK	pERK
309	0.503643884	0.747193224	0.4301753	2.81632854	5.990151664	0.218830018	0.177565491	2.373744337	0.232223754	1.750935592	0.687906244
	0.51461708	0.689063548	0.411770344	2.789514042	5.685037861	0.211636155	0.172817023	2.292149909	0.226972108	1.596376881	0.69500623
	0.509183088	0.730246795	0.418308781	2.687201071	5.622058542	0.209010905	0.175722212	2.283336522	0.230246795	1.561316243	0.677348383
	0.442106692	0.61707615	0.358626307	2.466947197	4.97950319	0.222885842	0.176462604	2.152300801	0.207004208	1.595086195	0.583276775
	0.434940244	0.617429838	0.358802202	2.36578488	4.718678663	0.213105949	0.173626964	2.134013697	0.192157916	1.504229891	0.550960118
	0.447506385	0.62817583	0.36738809	2.38593897	4.807635435	0.218577766	0.176233365	2.14128243	0.195187525	1.442398172	0.566339562
	0.428032684	0.573695789	0.342708988	2.334223759	4.473130107	0.225172847	0.184003771	2.012413576	0.195788812	1.612036455	0.509899434
	0.416922604	0.564035627	0.327702703	2.260135135	4.268734644	0.214834152	0.179668305	2.007985258	0.18980344	1.424600737	0.501074939
	0.386310905	0.538428074	0.317720418	2.125725058	4.063950116	0.207221578	0.167778422	1.861513921	0.180684455	1.261890951	0.476653132
	0.380827447	0.499293643	0.362462159	2.096266398	3.598587286	0.22764884	0.188092836	1.717860747	0.188092836	1.414328961	0.455499495
	0.366511251	0.513277924	0.327792418	2.072572471	3.66105818	0.229474965	0.193796878	1.724508413	0.185282789	1.31867018	0.461179809
	0.364153907	0.499411072	0.355123675	2.006870828	3.466627405	0.216332941	0.204358068	1.672555948	0.190223793	1.166077739	0.447192776
	0.364873351	0.482101737	0.312539251	1.946200544	3.349591794	0.230898053	0.188821436	1.508687461	0.171027842	1.353359849	0.471844254
	0.381910612	0.485914001	0.31095107	1.958907011	3.349290405	0.225799619	0.188519381	1.50900233	0.179834781	1.219656852	0.463249312
	0.37440953	0.46231259	0.344629287	1.861162456	3.287122612	0.22181146	0.185459026	1.444239063	0.176422263	1.123023208	0.426987061
311	0.743117916	0.862652731	0.377741793	2.735757397	6.067569557	0.219049021	0.185337848	2.277491535	0.19446489	2.379508317	1.081554541
	0.711479945	0.807053942	0.351590595	2.546887967	5.595573997	0.199170124	0.165975104	2.118810512	0.174688797	2.050484094	1.07593361
	0.704633205	0.802537231	0.350110314	2.467733039	5.548400441	0.20532267	0.165057915	2.10728075	0.171400993	1.938913403	1.065637066
	0.677359194	0.770234987	0.356396867	2.563222678	4.975195822	0.228086535	0.186497576	2.259045132	0.190973517	2.167847818	0.977620291
	0.591572123	0.678768233	0.312479741	2.164181524	4.313938412	0.195786062	0.161102107	1.975688817	0.16191248	1.768719611	0.838573744
	0.618517229	0.716672468	0.319700661	2.285938044	4.571179951	0.206926558	0.171597633	2.127044901	0.174556213	1.767316394	0.871388792
	0.702608291	0.69958081	0.387750349	2.437587331	4.449231486	0.258500233	0.197019096	2.091057289	0.182114578	2.171401956	0.886353051
	0.598868778	0.690271493	0.349773756	2.308371041	4.229411765	0.221493213	0.187104072	2.045475113	0.176696833	1.922171946	0.850452489
	0.561866554	0.641891892	0.308488176	2.157728041	4.020692568	0.216427365	0.173775338	1.940878378	0.168074324	1.728040541	0.822212838
	0.55097021	0.561355562	0.320852692	2.19786827	3.558895873	0.237769882	0.194588685	1.783820716	0.178190763	1.831374693	0.711943154
	0.538412899	0.701865318	0.384445147	2.482137212	4.109705975	0.288017705	0.223205817	2.082832754	0.202023396	1.96522289	0.839076826
	0.521126761	0.583045442	0.304278501	2.053414829	3.381344672	0.251660909	0.189476482	1.800956683	0.178049429	1.51767207	0.733988839
	0.488372093	0.540085679	0.29130967	1.913402693	2.873623011	0.215422277	0.210526316	1.414932681	0.1624847	1.689412485	0.685740514
	0.514854614	0.564791403	0.316055626	1.957016435	2.978824273	0.250632111	0.175094817	1.487357775	0.162136536	1.678571429	0.686156764
	0.485099846	0.556374808	0.287250384	1.892165899	2.847004608	0.223041475	0.177572965	1.462365591	0.169892473	1.537941628	0.695852535

Table A.2: The last four columns of two mouse in dataset.

MouseID	Genotype	Treatment	Behavior	class
309	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
311	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m
	Control	Memantine	C/S	c-CS-m