PeerJ Computer Science

Genetic variations analysis for complex brain disease diagnosis using machine learning techniques: opportunities and hurdles

Hala Ahmed[1], Louai Alarabi[2], Shaker El-Sappagh[3,4], Hassan Soliman[1] and Mohammed Elmogy[1]

[1] Information Technology Department, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt
[2] Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia
[3] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[4] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

ABSTRACT

Background and Objectives: This paper presents an in-depth review of the state-of-the-art genetic variations analysis to discover complex genes associated with the brain's genetic disorders. We first introduce the genetic analysis of complex brain diseases, genetic variation, and DNA microarrays. Then, the review focuses on available machine learning methods used for complex brain disease classification. Therein, we discuss the various datasets, preprocessing, feature selection and extraction, and classification strategies. In particular, we concentrate on studying single nucleotide polymorphisms (SNP) that support the highest resolution for genomic fingerprinting for tracking disease genes. Subsequently, the study provides an overview of the applications for some specific diseases, including autism spectrum disorder, brain cancer, and Alzheimer's disease (AD). The study argues that despite the significant recent developments in the analysis and treatment of genetic disorders, there are considerable challenges to elucidate causative mutations, especially from the viewpoint of implementing genetic analysis in clinical practice. The review finally provides a critical discussion on the applicability of genetic variations analysis for complex brain disease identification highlighting the future challenges.
Methods: We used a methodology for literature surveys to obtain data from academic databases. Criteria were defined for inclusion and exclusion. The selection of articles was followed by three stages. In addition, the principal methods for machine learning to classify the disease were presented in each stage in more detail.
Results: It was revealed that machine learning based on SNP was widely utilized to solve problems of genetic variation for complex diseases related to genes.
Conclusions: Despite significant developments in genetic diseases in the past two decades of the diagnosis and treatment, there is still a large percentage in which the causative mutation cannot be determined, and a final genetic diagnosis remains elusive. So, we need to detect the variations of the genes related to brain disorders in the early disease stages.

Submitted 12 April 2021
Accepted 5 August 2021
Published 20 September 2021

Corresponding authors
Louai Alarabi, lmarabi@uqu.edu.sa
Mohammed Elmogy, melmogy@mans.edu.eg

Academic editor
Khalid Raza

Additional Information and Declarations can be found on page 35

DOI **10.7717/peerj-cs.697**

Subjects Bioinformatics, Artificial Intelligence, Data Mining and Machine Learning, Data Science
Keywords Genetic analysis, Brain disease, Machine learning, Deep learning, Single nucleotide polymorphism (SNP), Microarrays

INTRODUCTION

The genetic mechanisms of complex diseases are challenging to discover. The power to identify the set of genes responsible for complex diseases using current methods is often lacking. In the last few years, large quantities of biological data have been generated for genomics and proteomics with rapid developments (*Mezlini & Goldenberg, 2017*). These data need a complex mathematical analysis to interpret biological data using the interdisciplinary science composed of computer science and information technology, which is known as bioinformatics or computational biology. This new research area is more important and will grow rapidly as we continue producing and integrating large amounts of protein, genomic and other data.

Bioinformatics research is an active area to solve biological problems. It utilizes data mining (DM) techniques and its applications to analyze biological data. In medicine, DM is emerging of high importance for diagnosing neurodegenerative diseases and supporting prognosis to provide deeper understanding (*Joshi et al., 2010*). The data analysis includes many examples, such as protein structure prediction, gene classification, gene classification based on microarray data, and clustering gene expression data. Therefore, DM and bioinformatics interaction should be increased to provide great potential (*Association, 2019*).

The association of the structural changes with brain diseases, which occurs as the disease progresses, is unknown (*Sherif, Zayed & Fakhr, 2015*; *Kim et al., 2013*; *Bertram, Lill & Tanzi, 2010*; *Yokoyama et al., 2015*; *Hwang et al., 2014*; *Guerreiro & Hardy, 2012*). So, we need to identify many genetic markers about their association with brain diseases. In the 1950s, artificial intelligence (AI) appeared as an independent field because of the potentials to make machines intelligent like a human. In streamlining complex analytical workflows, bioinformatics with AI techniques play an essential role in performing multistep analysis within one analytical framework. The problems with biological data, such as the complexity of data and the growing exponential rate, can be solved with workflows that enable processing and analysis. Machine learning (ML), knowledge discovery, and reasoning deployed by AI techniques are continuously improving. A formidable combination is presented between bioinformatics and AI, where the analysis of complex biological systems is enabled by bioinformatics and human-like reasoning. We can perform complex tasks based on reasoning by using AI-based tools (*Sandraa et al., 2010*; *Bertram & Tanzi, 2012*).

In the past few years, intensive computational efforts have been performed to study single nucleotide polymorphisms (SNPs) structural and functional consequences. In this context, ML and DM techniques have been widely performed for SNP data analysis. Biological systems are complex, so most studies on large scales concentrate only on one specific aspect of the biological system. For example, genome-wide association studies (GWAS) concentrates on genetic variants associated with phenotypes measurement.

Table 1 Early symptoms comparison of various brain diseases.

Comparison between early symptom	Dementia of Lewy body	Parkinson's disease (PD)	Alzheimer's disease (AD)
Age of onset	>60 years old	>70 years old	>60 years old
Gender-specific	Men > Women	Conflicting	Men = women
Family history	No	Conflicting	Yes
Significant Loss of Memory	Possible	Possible Years After Diagnosis	Always
Problems of Language	Possible	Possible	Possible
Fluctuating Cognitive Abilities	Likely	Possible	Possible
Planning or Problem-solving Abilities	Likely	Possible	Possible
Decline in Thinking Abilities that Interfere with Everyday Life	Always	Possible Years After Diagnosis	Always
Difficulty with a Sense of Direction or Spatial Relationships between Objects	Likely	Possible	Possible

So, the feature should be chosen precisely from the provided dataset. The best possible subset of the feature sets is determined using search methods. Then, evaluation techniques are used to evaluate them.

Some examples of genetic diseases that affect the brain are Alzheimer's Disease (AD) and Parkinson's, as shown in Table 1. The progressive decline of cognition and memory is caused by AD, which is a degenerative disease. It causes the nerve cells' degeneration in the brain due to AD's side effects, which are related to language and memory (*Ang et al., 2015*). After 65 years, symptoms appear, and the spread of disease with age increases sharply. It is considered the most common form of dementia in the disease's onset genetic factors for specific genes. However, they are not the primary effect of the disease. Also, the suffering can be increased by other factors, such as age, smoking and alcohol (*Ang et al., 2015*; *Zuk et al., 2012*, *2014*; *Hormozdiari et al., 2015*). AD has many common symptoms, such as complete memory loss, impairments of movements, misplacing things, verbal communication, and abnormal moods (*Sherif, Zayed & Fakhr, 2015*). If AD is not initially diagnosed, the disease's severity increases (*Hemani, Knott & Haley, 2013*; *Pinto et al., 2014*; *Parikshak, Gandal & Geschwind, 2015*; *Nakka, Raphael & Ramachandran, 2016*).

One of the most demanding tasks in a post-genomic era is identifying disease genes from a vast amount of genetic data. Moreover, complex diseases present a very heterogeneous genotype that makes it difficult to identify biological markers. ML is widely used to identify these markers, but their performance relies heavily on the size and quality of the data available (*Asif et al., 2018*). Also, modern computational systems help researchers to analyze complex data like genetic information of humans and its underlying patterns. These patterns reveal the genes that cause diseases. Mathematical models are helpful to build robust ML models for analyzing gene expression. So, this paper introduces a comprehensive review of genetic variations analysis for discovering complex disease-related genes, especially genetic brain diseases. The SNPs data is commonly used as a

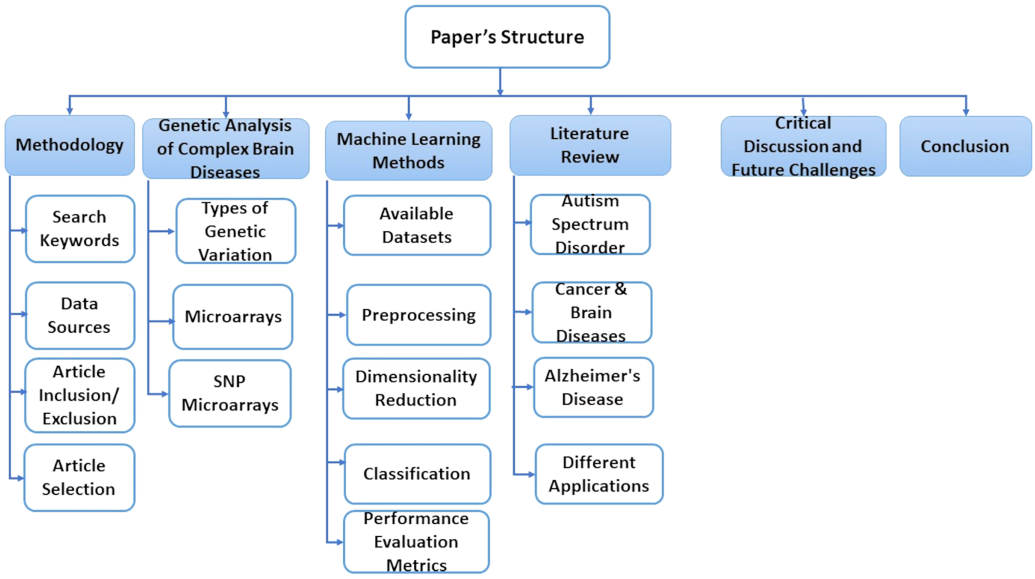

Figure 1 **The structure of the survey.** Full-size DOI: 10.7717/peerj-cs.697/fig-1

marker-based association or association study in a population. We focus on identifying a disease's biomarkers based on SNPs that alter phenotypes by altering some molecular functions. There are many challenges to defining so-called functional variants. First, the marker variants themselves may be unbalanced (or linkage, depending on the study) with the causal variant. Second, the challenge is identifying candidate genes related to disease to narrow down an area for SNP prioritization. Finally, the molecular function is poorly understood, so we need to increase how SNPs disrupt their functionality (*Hemani, Knott & Haley, 2013*; *Pinto et al., 2014*; *Parikshak, Gandal & Geschwind, 2015*; *Nakka, Raphael & Ramachandran, 2016*; *Guyon et al., 2002*; *Printy et al., 2014*).

This paper is consisting of six sections. In "Methodology", we explain some basic concepts of gene sequence and types of genetic variations. In "A Genetic Analysis of Complex Brain Diseases", we present some classification, preprocessing, and dimensionality reduction techniques. "ML Methods" discusses essential research areas about various diseases depending on discovering genetic mechanisms as shown in tables in this section. "Literature Review on Complex Disease & Applications" introduces current research topics, future directions, and challenges. Finally, we discuss the conclusion of this work in "Critical Discussion and Future Challenges". Also, Fig. 1 represents the structure of these sections.

METHODOLOGY

This section introduces the protocol used to survey various genetic variations to diagnose complex brain diseases using machine learning. We list the search keywords, data sources, inclusion/exclusion criteria, and article selection in this section. Also, we illustrated in Fig. 2 the structure of the systematic review analysis and how to remove the overlap articles among different databases. We remove overlap with two ways by Endnote by choosing the following elements: (1) Title, author and year. (2) Title, author and journal.

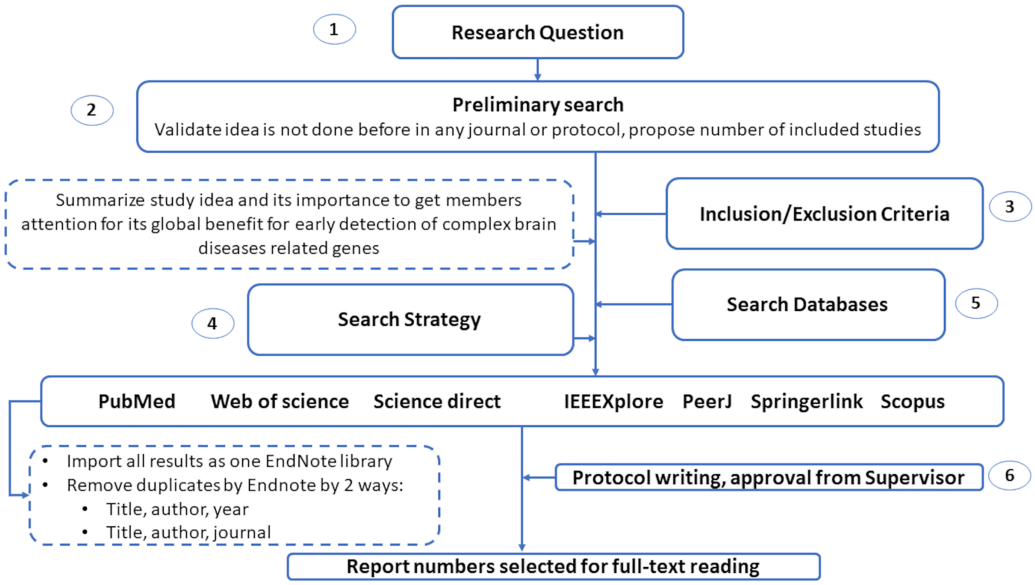

Figure 2 **The guidelines for systematic review analysis.** Full-size DOI: 10.7717/peerj-cs.697/fig-2

Table 2 **The used databases for selecting the academic articles in this article.**

Academic database	Source
Science direct	http://www.sciencedirect.com/
Springerlink	https://link.springer.com/
IEEEXplore	https://ieeexplore.ieee.org/
Web of science	https://apps.webofknowledge.com/
PubMed	https://pubmed.ncbi.nlm.nih.gov/
PeerJ	https://peerj.com/
Scopus	https://www.scopus.com/

Search keywords

Keywords have been carefully chosen for the initial search. After the initial search, new words found in several relevant articles have formulated multiple keywords. The study's keywords included "genetic analysis for brain disease using machine learning", "genetic variations based on single nucleotide polymorphism (SNP)", "Gene analysis for complex brain disease based on machine learning".

Data sources

Our survey used different academic database sources to obtain articles, as shown in Table 2.

Article inclusion/exclusion criteria

Inclusion/exclusion criteria have been developed to determine which papers are eligible for the next review phase. Relevance for research has been considered for articles that meet

inclusion criteria and excluded articles that do not fulfill inclusion criteria. The following criteria are defined for inclusion/exclusion.

Inclusion

1. Our paper only focuses on genetic variation for complex diseases related to genes.
2. Only articles that performed on genetic variation for complex diseases related to genes using machine learning.
3. For inclusion, only articles in English were taken into account.

Exclusion

1. Papers that do not focus on genetic variation for complex diseases related to genes were excluded.
2. Articles that do not concentrate on genetic variation for complex diseases related to genes using machine learning were excluded.
3. Articles that are not in English were excluded.

Article selection

Research has established criteria for inclusion and exclusion to choose which articles are eligible for the next review phase. Articles that satisfy the inclusion criteria were considered research-related, and items that did not meet the inclusion criteria were excluded. In the previous section, we present the list of inclusion/exclusion criteria. Three phases were followed to select an article for this research. The first step was to consider extracting only the titles and abstracts of the articles. The second stage was to analyze the abstract, introduction, and conclusion to refine the first stage's choice. At the end of the process, the articles were carefully read and the quality of the papers was then measured according to their relevance to the research.

A GENETIC ANALYSIS OF COMPLEX BRAIN DISEASES

The core of modern medical genetics is to enhance our understanding of the genetic mutations and possible factors in genetic risk, which cause or contribute directly to human disease. Evolutionary theory can describe selective forces that affect causal alleles and susceptibility to human genetic disorders. In the general population, complex disorders are common and result from the interaction of many sites of sensitivity and environmental factors (*Rahit & Tarailo-Graovac, 2020*; *Spataro et al., 2017*). In contrast, Mendelian disorders are usually rare and have predictable genotypes as they are generally caused by a single causative mutation in the gene. However, the classic distinction between Mendelian diseases and complex diseases is not always absolute due to its heterogeneity and incomplete penetration. There is a continuum between pure Mendelian diseases and more complex diseases.

Diseases that follow Mendelian inheritance patterns are called Mendelian disorders. About 80% of all rare diseases are hereditary, and most of these diseases are monogenic/

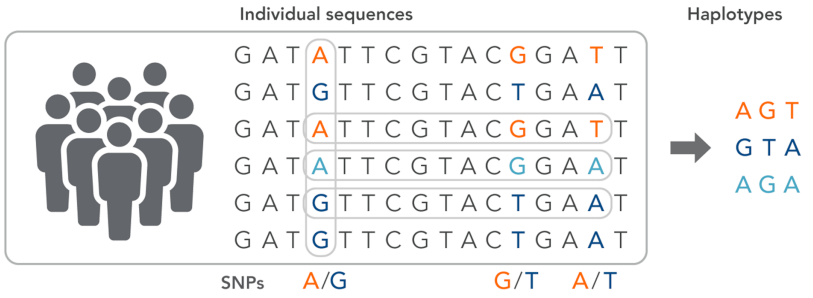

Figure 3 **The haplotype of SNPs for individuals.** Full-size DOI: 10.7717/peerj-cs.697/fig-3

Mendelian. According to an estimate, there are 400 million people worldwide with around 7,000 various rare diseases (*Rahit & Tarailo-Graovac, 2020*; *Spataro et al., 2017*). So, studying gene function by biomedical researchers should consider the impact of the variations between these diseases. Genetic information about these diseases can supply valuable insight into the critical regions and operating ranges, proteins, or regulatory elements. Genetic variants may differ in nature concerning single nucleotide variants, tandem repeats or deletions, and small insertions to massive copy number variants (CNVs) (*Siavelis et al., 2016*). Until lately, genetic variation information is limited to some extent. However, many large-scale surveys of diversity have provided numerous data on covariance. A picture is started from the driving forces in human evolution and population diversity that is emerged (*Siavelis et al., 2016*; *Mathur, 2018*; *Raza, 2012*; *Wodehouse, 2006*; *Mount & Pandey, 2005*; *Barnes, 2010*).

SNPs work as a pointer in the association and linkage studies to identify the genome's part in a particular disease (*Hasnain et al., 2020*; *Ng & Henikoff, 2003*; *Uppu, Krishna & Gopalan, 2016*; *Liang & Kelemen, 2008*; *Korani et al., 2019*). Polymorphs are found in the same coding and organizing regions, which may be involved in diseases. A non-synonymous SNP is called the SNP that causes amino acid substitution. The amino acid variations that lead to genetic lesions that cause diseases have a strong concentration and interest. Studies aimed at determining polymorphisms and analysis of mutation complement each other to identify substitutions for amino acids in protein code regions where any variation can change protein function or structure (*Clare & King, 2001*). Substitution of one nucleotide by another is the most simple form of DNA variation between individuals. As shown in Fig. 3, this type of change is known as SNP. SNPs are estimated to occur at a rate of 1 in 1,000 bps across the genome (*Shastry, 2007*). These simple changes can be of transitional or transitional type. Approximately 50% of the polymorphisms are found in non-coding regions, whereas 25% of them result in missense mutations (encoding SNPs or cSNPs). The remaining 25% are silent mutations that do not alter the code of the amino acid. These silent SNPs are known as synonymous, and most likely, they are not subject to natural selection.

On the other side, non-synonymous SNPs may cause pathology and may be natural for selection (nSNPs and codified amino acids). SNPs (both synonymous and unknown) influence promoter activity and pre-mRNA morphology (or stability) (*Shastry, 2007*). They also change the protein's ability to bind to its substrates or inhibitors and alter

subcellular proteins' localization (nSNPs). Therefore, they may be responsible for disease susceptibility, drug deposition, and genome evolution. However, many of them affect gene function. Besides, many of them are not harmful to living organisms and must have survived selection pressure (*Halushka et al., 1999*).

Due to the inheritance of certain diseases, a direct aim of the human genome project is to identify the genes which predispose individuals to different diseases. It also examines how sequential variation of a gene affects the functions of its product. SNPs often occur across the genome, as has already been mentioned. They can therefore be used as markers by studying correlation to discover disease-causing genes. Two closely related alleles (gene and marker) are supposed to be inherited together in such studies. So, a simple comparison of patterns of genetic differences between patients and normal individuals may facilitate a method for determining which sites are responsible for disease susceptibility.

Types of genetic variation

Candidate genes tend to decrease the number of SNPs in the study to some mutations where genes are intended to provide the genetic basis for the disease under study. Although overall candidate genes' data sets are considered to be, the various tests of hundreds or even thousands of polymorphisms in our situation make link detection troublesome. A conceivable method to overcome testing overpowering quantities of SNPs, mainly based on the studies of a candidate gene, is to order SNPs as per their priority of functional significance (*Clare & King, 2001*). Previous natural information can be used in current databases to reduce the number of SNPs by focusing on specific genomic regions, and computational methodologies and aptitude are used to separate the neutral polymorphisms from poly-morphisms of potential functional interest (*Mikhail et al., 2020*).

The critical biological determinant is a genetic variation that supports evolution and determines the phenotype's genetic basis. A researcher may wish to deal with the genetic data depending on the researcher's viewpoint (*Barnes, 2010*). Researchers of the biomedical view tend to focus on genetic or phenotypic directions. The function of a gene cannot be understood fully without realizing the possible variation within the gene. This means that the biomedical researchers, who study genetics, must know what variants present and what effect these variants are on the gene's function and thus the phenotype. The genetic variation comes in many forms, but every form arises from only two types of mutation events. The simplest type of different genetic variation results from a simple primary substitution that comes from mutation. This type of mutation event explains the most common form of variation (*Barnes, 2010*), SNP. However, it also represents rare mutations that may manifest Mendelian inheritance in families. Most other kinds of variation arise directly or indirectly from inserting or deleting a portion of DNA.

Microarrays

Microarray is a technique in which 1,000 nucleic acids are attached to a surface. They are used to measure DNA sequences' concentration in a mixture by hybridization and the

subsequent detection of hybridization events (*Bumgarner, 2013*). Millions of sequences in one reaction are simultaneously analyzed using microarrays. Microarrays come in three basic types: (A) spotted arrays on glass, (B) self-assembled arrays and (C) insitu synthesized arrays. Many types of microarrays depend on the type of data (*Berry et al., 2019*; *Mao, Young & Lu, 2007*; *Prince, 2017*; *Smith et al., 2007*; *Isik, 2010*; *Yin et al., 2017*). These types are DNA microarray, SNP microarray, cDNA microarrays, SNP microarrays, protein microarrays, MM chips, peptide microarrays, etc.

All studies on genome association aim to determine the genetic basis for traits and disease sensitivities using SNP microarrays. They carry the genetic variants, which are the most common in the human. Variants were identified for families at risk for several diseases (*Coelho et al., 2009*; *Chen et al., 2012*). However, with only a few notable exceptions, such as the related age of macular degeneration, risk variants usually describe only a small portion of the genetic risks known for their existence. Many factors are found that favor is contributing to this observation. Common variants may have minor effects on the phenotype or have variable penetration due to cognitive or epigenetic effects. Two other factors are CNVs and rare variants (*Barnes, 2010*). Genomic variation types have an essential effect. However, examine these variants on disease phenotypes should perform.

SNP microarrays

SNP microarray is a type of DNA microarrays, which is used to identify polymorphisms within a population. SNP is a variation of a single site in DNA, and it is the most common kind of variation in the genome. About 335 million polymorphisms have been specified in the human genome (*Berry et al., 2019*; *Mao, Young & Lu, 2007*). Microarrays of SNP have grown as a robust tool for the large-scale detection of epigenetic changes in genomes of predictive and/or predictive value. For genotyping in 1998; SNP matrix technology was developed (*Association, 2019*). Since then, this technique has been extensively developed and has become one of the most robust genetic analyses (*Latkowski & Osowski, 2015*; *González & Belanche, 2013*; *Xue, Zhang & Browne, 2012*; *Teng, Dong & Zhou, 2017*; *Kong, Mou & Yang, 2009*; *Saeys, Inza & Larranaga, 2007*). In this study, the very long execution times result in an enormous volume of data. Some preliminary tests were performed to select the three following constraints for the primary sequence of experiments (*Mao, Young & Lu, 2007*). First, it depends on the used data, which we could not generalize all data analysis. The second issue is the selection of SNP. Finally, the third issue is the used ML method.

ML METHODS

Microarray technology is a valuable tool for capturing information from biological, genetic data. An extensive set of genetic data and many computational techniques are needed to determine whether a human is normal or abnormal as we need to identify the biomarker in the gene responsible for the diseases. ML algorithms play an important role in distinguishing unhealthy and normal genes extracted from humans' genomes (*Karthik & Sudha, 2018*). ML methods contain steps to manage the classification process. ML methods

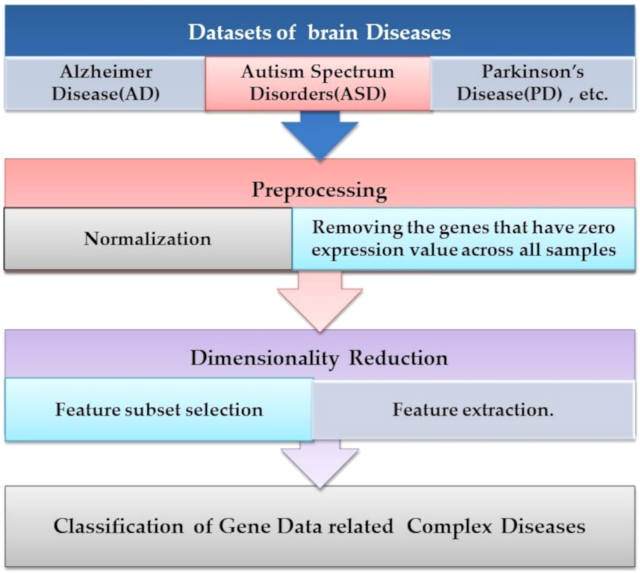

Figure 4 **The common ML procedure for the complex brain disease diagnosis.**
Full-size DOI: 10.7717/peerj-cs.697/fig-4

contain some stages to manage the diagnosis process and determine candidate genes that cause diseases, as shown in Fig. 4. Each step is discussed in the following subsections.

Available datasets

In this section, we provided several benchmark databases that can be used to test the researchers' proposed techniques and methods (*Guyon et al., 2002*) as shown in Table 3. Research in biology and medicine may benefit from higher-order gene screening to confirm recent genetic disease research discoveries or new ways to explore them. Therefore, the ADNI (*Rémi et al., 2011*) dataset's primary goal is to support researchers with the chance to combine genetics, imaging, and clinical data with supporting the investigation of disease mechanisms. Genotyping and sequencing data were generated for ADNI 1, ADNI GO, and ADNI 2 subjects and are available to ADNI investigators. ADNI was released in 2003 as a public-private partnership. ADNI phase 1 data were collected from 757 subjects (214 controls, 366 MCI and 177 AD cases).

Preprocessing

There is a lot of information in a raw dataset prepared by researchers. It depends entirely on the requirements and purposes if the information is valuable or not. Data preprocessing is the first step in ML. It is necessary to ensure that the dataset is fully adapted to the needs. First, we removed unwanted attributes, missing values, and correct arrangement, and the expression value is normalized. The Bioconductor package completes the two steps. This raw file must be processed to extract the correct attributes for the following study. R programming language is chosen as a preferred language for analyzing data. Therefore, R programming language packages were applied for the preprocessing of the dataset. The Bioconductor analyzed the value of the expression and set the data set further,

Table 3 Some of the benchmark datasets.

Dataset	Diseases	Source
GEO Database	ASD	https://www.ncbi.nlm.nih.gov/geo/
(KEGG) database	AD	https://www.genome.jp/kegg/
ADNI/Whole-genome sequencing (WGS) datasets	AD	http://adni.loni.usc.edu/
TCGA	Cancer	The Cancer Genome Atlas (TCGA): https://portal.gdc.cancer.gov/
UCI for Cancer	Cancer	UCI Machine Learning Repository: http://archive.ics.uci.edu/ml
NCBI Gene Expression Omnibus (GEO)	Cancer	NCBI repository: https://www.ncbi.nlm.nih.gov/geo/

using the data normalization that reduces the data range to be studied as shown in Eq. (1). Unwanted attributes and the samples that have missing values must be removed. Data rearrangement according to format requirements will be completed before the next phases proceed. For classification purposes, only desired attributes are kept.

Gene removal with no expression value through all samples is one of the most simple and straightforward preprocessing techniques used in *Karthik & Sudha (2018)* and *Daoud & Mayo (2019)*. Researchers applied this step by removing samples with 20% of the deleted features. After then, the data are normalized. Normalization is to remove the unimportant technical differences in data and to filter matrices with the p-value. Preprocessing is a step towards facilitating the use of data. The normalization was performed in the preprocessing stage by changing the scale or range of data from 0 to 1. Data normalization is necessary because microarray data has a significant range difference. The function of data normalization is presented by Eq. (1).

$$y' = \frac{y - y_{min}}{y_{max} - y_{min}} \tag{1}$$

where y' is the value of features in the normalization domain, while the data's value before the normalization process is y. At the same time, y_{min} and y_{max} refer to the smallest and largest values of all data in an attribute to be normalized, respectively. Algorithms of ML tend to be affected by noisy data. To avoid unnecessary complexity, noise should be decreased as much as possible in the inferred models (*Karthik & Sudha, 2018*; *Daoud & Mayo, 2019*; *Adiwijaya et al., 2018*). Two common types of noise can be known: (1) class noise and (2) attribute noise. Class noise is affected by samples classified as belonging to more than one class, which leads to wrong classifications. Simultaneously, the attribute causes attribute noise value errors, such as variables with wrongly measured and missing values.

The classification of patterns with missing data generally affects two issues: handling missing values and classifying patterns. The ability to handle missing data has become an essential requirement for pattern classification because inadequate data processing can lead to significant errors or false classification results. Most literature approaches can be grouped into four different types according to how both problems are resolved, as shown in Fig. 5.

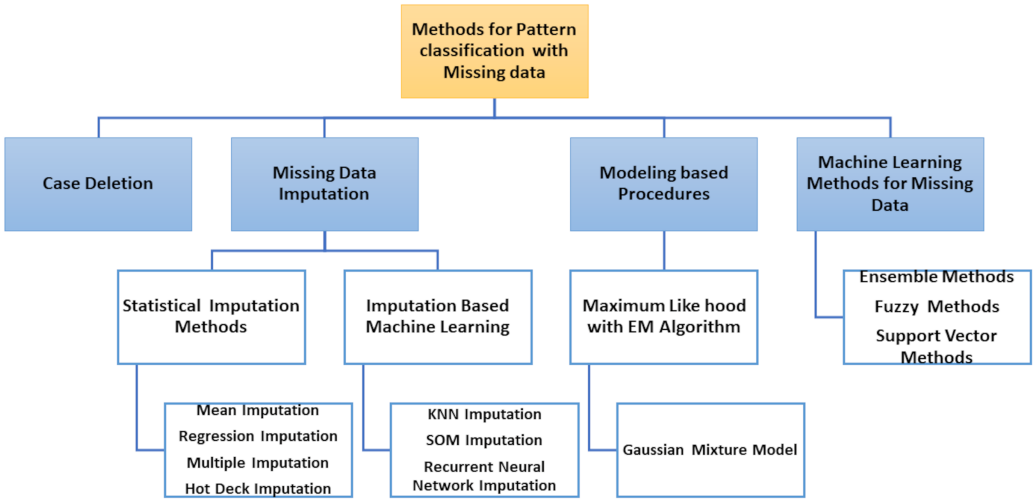

Figure 5 **Methods for pattern classification with missing data.**
Full-size DOI: 10.7717/peerj-cs.697/fig-5

1. Removal of incomplete cases and design of classifier using only the full data portion.
2. Imputation or estimation by editing set of missing data and the classification problem learning, *i.e.*, with imputed values for complete data portion and incomplete patterns.
3. Use of model-based processes where the distribution of data is modelled using certain procedures, *e.g.* by expectation–maximization (EM) algorithm.
4. Use of ML methods in which missing values are included in the classification system.

The two first types of approaches are solved separately: the handling of missing value (data deletion, imputation) and the pattern classification, but the third type of approaches model the probability density function (PDF), used by the theory of decision in Bayes to classify the input data (complete and incomplete). Finally, the classifier is designed to process incomplete input data without previous estimates of missing data in the last types of approaches.

Dimensionality reduction

As for increasing dimensions, the cost of computations also increases, usually exponentially. To get around this problem, it is essential to find a way to decrease the number of features in mind. Often two methods are used, which are feature selection and extraction (*Daoud & Mayo, 2019*; *Adiwijaya et al., 2018*).

Feature selection

Informative polymorphisms that are known as Tag SNPs and irrelevant ones are all found in the whole dataset. SNP data have high dimensionality, such as microarray gene expression, but it is larger several hundred times, which leads to the impractical of data analysis. Thus, removing the useless SNPs and extracting a small subset of those discriminants would help identify polymorphisms markers, which could be used as biomarkers. This process can be performed through feature selection, reducing a search

space to facilitate classification tasks. Feature selection is divided mainly into the filter, wrapper, and embedded types. Before classification, task filtering techniques are applied to a dataset and usually to simple statistics, such as the t-statistic or F-stat, or the p-value that is used as the trait evaluation (*Batnyam, Gantulga & Oh, 2013*).

Conversely, wrapper methods take advantage of the learning algorithm to define a feature subset by integrating it within the feature search and selection. In the third type, the search for an ideal subset of features is included in the classifier's construction and can search for the combined space of sub-feature sets and hypotheses. Table 4 provides a common classification of feature selection methods and shows most of the technique's advantages and disadvantages (*Batnyam, Gantulga & Oh, 2013*, *Van Rooij et al., 2019*). An example of feature selection techniques is the information gain (IG) measure. It is dependent on the concept of entropy (*Pereira et al., 2018*; *Joachims, 1998*; *Chandrashekar & Sahin, 2014*; *Khalid, Khalil & Nasreen, 2014*; *Kursa & Rudnicki, 2010*). It is commonly used to measure feature suitability in filtering strategies that assess features' individuality, and this method is quick. Suppose D (A_1,A_2,…,A_n, C), $n \geq 1$, be a dataset with $n + 1$ attributes, where C is the class attribute. Let m be the number of distinct class values. The class distribution of entropy in D, represented by Entropy(D), is defined by Eq. (2).

$$Entropy(D) = \sum_{i=1}^{m} p_i * \log_2 * p_i \tag{2}$$

where p_i is the probability that an arbitrary instance of D belongs to the class c_i. Eq. (2) defines the single-label strategy concepts, also known as IG attribute ranking (*Mezlini & Goldenberg, 2017*). It calculates the feature's ability to distinguish between class values. In *Daoud & Mayo (2019)*, For dealing with multi-label data, the C4.5 algorithm was adapted. This decision tree algorithm lets multiple labels at the tree leaves by calculating an entropy adaptation, as described by Eq. (3).

$$Entropy.ML(D) = -\sum_{i=1}^{l} p(\lambda_i) * \log_2 * p(\lambda_i) + q(\lambda_i) * \log_2 * q(\lambda_i) \tag{3}$$

where $p(\lambda_i)$ is the probability that an arbitrary instance in D belongs to the class label λ_i, $q(\lambda_i) = 1 - p(\lambda_i)$, and the number of labels is l in the dataset (*Kursa & Rudnicki, 2010*; *Shaltout et al., 2015*; *How & Narayanan, 2004*). They have adopted this formula to create an IG feature selection capable of dealing with multi-label data. By using IG as a filter approach, the feature selection can be performed with any multi-label classifier.

While embedded methods have the advantage of interacting with the classification algorithm, they have a lower computational cost than wrapper methods. It is referred to as the hybrid approach, which typically combines two or more feature selection algorithms of the various search strategies sequentially. For example, a less costly algorithm such as a filter might eliminate features first and use a more complex and costly algorithm afterward, such as a wrapper.

Table 4 The main feature selection techniques.

Feature selection	Advantages	Disadvantages
Filter	Univariate	
	Fast	Neglect dependencies with feature
	Scalable	Neglect the classifier interaction
	The selection of classifier is independent	
	Multivariate	
	Models feature dependencies	Slower
	Regardless of the classifier	Less scalable
	Better computational complexity than wrapper methods	Interaction is neglected with the classifier
	Deterministic	
Wrapper	Simple	Risk of overfitting
	Interacts with the classifier	More prone than randomized
	Feature of models is dependent	Algorithms to obtaining stuck in a local optimum (greedy search)
	Less computationally	The selection of classifier is dependent
	Intensive than randomized methods	
	Randomized	
	To local optima is less prone	Computationally intensive
	Models feature dependencies	The selection of classifier is dependent
		Overfitting with higher risk
		Than deterministic algorithms
Embedded	Interacts with the classifier	The selection of classifier is dependent
	Computational complexity is better than wrapper methods	
	Feature of models is dependent	

Feature extraction

Various techniques have been adopted to decrease the dimensions of gene data by choosing a subset of genes. There are several methods implemented to extract necessary information from microarrays and thus reduce their size. New variables are created by feature extraction as combinations of other variables to reduce the selected features' dimensions. The feature extraction algorithms are consisting of two classes, which are linear and nonlinear techniques (*Chu & Wang, 2005*; *Khodatars et al., 2020*).

Linear

linear features extraction techniques suppose that the data is located in a low-dimensional linear subspace. The matrix factorization is projected onto this subspace. The most popular algorithm for dimensional reduction is principal component analysis (PCA). PCA finds main components in the data representing uncorrelated eigenvectors, representing a percentage of the variance of the data, using the covariance matrix and its eigenvalues and eigenvectors. PCA and many of its variants have been performed to reduce data dimensions in microarray data (*Adiwijaya et al., 2018*).

The steps for dimensional reduction algorithm using PCA, according to *Chu & Wang (2005)*, are described below:

1. Let X be a matrix of input for PCA. X is training data consist of an n-vector with data dimension m.

2. For each dimension (X), calculate the mean data using Eq. (4).

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4}$$

where n = samples number or number of data observation, X_i = observation data.

3. The covariance matrix (CX) is calculated with Eq. (5), where (X) = mean data.

$$C_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \tag{5}$$

4. The eigenvectors (v_m) and eigenvalues (λ_m) of the matrix of covariance is calculated using Eq. (6).

$$c_x v_m = \lambda_m v_m \tag{6}$$

5. In descending order, the sorting of eigenvalues.

6. The principal component (PC) is the eigenvector set corresponding to the eigenvalues sorted in step 5.

7. Based on the eigenvalues, the PC dimension will be reduced. Several ways are found for reducing. PC dimension based on eigenvalues, such as:

a) Using a scree plot. The number of eigenvectors is determined based on the curve's point, which is no longer sharply decreasing.

b) Using the cumulative proportion of variance (eigenvalues) of the total variance (eigenvalues).

$$PVV = \frac{\lambda_i}{\sum \lambda_i} \times 100\% \tag{7}$$

Next, the number of eigenvectors was set by comparing the threshold with the cumulative proportion of variance. PC, which has reduced dimensionally ($\widehat{PC}$), is a matrix consisting of the largest k eigenvalues. Also, k eigenvectors are vectors corresponding to meet Eq. (8). The testing data (Y) dimensions are reduced by multiplying testing data with ($\widehat{PC}$) in Eq. (9).

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \times 100\% > Threshold \tag{8}$$

$$y' = Y \times PC \tag{9}$$

It was argued that there is no guarantee that PCs are linked to the class variable when computing PCs of data sets. Supervised principal component analysis (SPCA) has therefore been introduced, in which the PCs based on the class variables were selected. This

Table 5 A comparison between selection and extraction techniques.

Method	Advantages	Disadvantages
Selection	Data Preserving for data interpretability	Discriminative power
		Lower times of training
		Reducing overfitting
Extraction	Higher distinguishing power	Interpretability of data is lost
	Overfitting Controlled when it is unsupervised	Switching can be costly

extra step is named the gene screening step. Even though the supervised PCA's version works better than its unsupervised version, the PCA has a significant limitation: Nonlinear relationships, especially in complex biological systems, that often exist in data are not identified. SPCA works as follows (*Hira & Gillies, 2015*):

1. Calculate the relationship measurement between each gene using linear, logistic, or proportional hazard model results.
2. Using cross-validation of the models in step (1) select the most associated genes with the outcome.
3. Estimate PC scores using only the selected genes.
4. Using the model in step (1), fit regression with the outcome.

The method was extremely effective in determining essential genes, and in cross-validation tests, it was only outperformed by gene shaving, a statistical method for clustering, like hierarchical clustering. The term "shaving" is derived from the removal or having a percentage of the genes (normally 10%) that have the inner product with the smallest absolute with the leading PC. The main difference is that more than one cluster can be part of the genes.

Nonlinear

Reducing dimensions by using nonlinear techniques is applied by using many distinct ways. The low-dimensional surface can be mapped to high dimensional space to establish a nonlinear relation between the features. In theory, it is possible to use the lift function (x) to map the features over a space with higher dimensions. In a higher space, the relationship between the features can be shown as linear and easily discoverable. This is then sent to space with the lower dimensions, and their relevance can be considered nonlinear (*Daoud & Mayo, 2019*; *Adiwijaya et al., 2018*). Table 5 provides a comparison between techniques of feature selection and feature extraction.

Classification

After reducing the dimensional complexity of the data, the next step is the classification process. Classification is not only the primary purpose of this research, but also it is critical to detect biomarkers of complex diseases. At this stage, the data was diagnosed (classified) based on whether they are affected by a specific disease or not. Biomarkers that

Table 6 The confusion matrix elements.

		Predicted	
		Normal	Abnormal
Actual	Normal	*TP*	*FN*
	Abnormal	*FP*	*TN*

Table 7 Some of the performance evaluation metrics.

Metrics	Description	Formula
Accuracy	This is a relation between the sum of TP and TN divided by the total sum of the population	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity, Recall	This is a relation between TP divided by the total sum of TP, FN	$Recall = \frac{TP}{TP + FN}$
Specificity	This is a relation between TN divided by the total sum of TN, FP	$Specificity = \frac{TN}{TN + FP}$
AUROC	This metric is used to measure the average area under ROC	$TPR = \frac{TP}{TP + FN}\ FPR = \frac{FP}{FP + TN}$
DSC	This is a relation between TP divided by the total sum of TP, FN, FP	$DSC = \frac{2 * TP}{2 * TP + FP + FN}$
Precision	It is a relation between TP divided by the total sum of TP, FP	$Precision = \frac{TP}{TP + FP}$
MCC	An effective solution overcoming the class imbalance issue comes from MCC	$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN}}$

can detect and diagnose brain diseases accurately are urgently required. So, we introduce a comprehensive review of genome sequencing analysis for discovering complex genes related to genetic brain diseases. Most of the genetic variation is contributed by SNPs about the human genome. Many complex and common diseases are associated with SNPs like AD. Early diagnosis and treatment can improve by investigating SNP biomarkers at different loci of these diseases. The investigation of genetic variants in the human genome related to complex diseases is one of the most important study objectives (*Sherif, Zayed & Fakhr, 2015*; *Saeys, Inza & Larranaga, 2007*). So, we divided the review into three sections. Each section discusses a disease, which is summarized in Tables 8–10 and 12. Table 11 provides some of the available datasets that the researcher can use.

Support vector machine (SVM)

SVM is considered the well-known classification technique, which can be used in disease diagnosis. SVM is depend on statistical learning theory (*Hira & Gillies, 2015*; *Tang, Suganthan & Yao, 2006*; *Bansal et al., 2018*). It is an algorithm that identifies a specific linear model, which maximizes the hyper-plane margin. Maximizing the margin of hyperplanes will maximize classes separation (*Abd El Hamid, Omar & Mabrouk, 2016*; *Breiman, 2001*). The training points closest to the margin of the cloud are the support vectors. These vectors (points) are only used to define the boundaries between classes.

Table 8 An overview of ASD using different ML methods.

Authors	Application on diseases	Method	Results	Problem with method
Spencer et al. (2018)	ASD	FPM algorithms& contrast mining	Including 193 novel autism candidates as significant associations from connected 286 genes.	It is a challenge for FPM to store many combinations of items as a memory requirement problem.
Al-Diabat (2018)	ASD	Data mining and fuzzy rule	Fuzzy rules achieved accuracy up to 91.35% and 91.40% sensitivity rate.	Considering the assessment of features on the dataset in this study is not extensive and does not consider other target data sets like adults, adolescents, and infants.
Alzubi et al. (2017)	ASD	SVM, NB, linear discriminant analysis, and KNN	Classification accuracy up to 96%	Their work remains incomplete until the basis of genetic diseases, and traits well understand.
Krishnan et al. (2016)	ASD	ML approach	Area Under the Receiver–Operator curve (AUC) = 0.80.	The accuracy of the system needs to enhance
Cogill & Wang (2016)	ASD	SVM	A mean accuracy = 76.7%	-
Jiao et al. (2012)	ASD	DSs, ADTrees, and FlexTrees	DS and FlexTree With an accuracy = 67%,	One limitation of this work is that this study includes only 29 SNPs.

Table 9 An overview of cancer diseases using different ML methods.

Authors	Application on diseases	Method	Results	Problem with method
Ismaeel & Ablahad (2013)	Cancer disease	Classification by ANN	The best classifier with mean square error 0.0000001.	The system cannot apply as general for all diseases, so we need to propose a novel system to detect all diseases.
Jain, Jain & Jain (2018)	Cancer	NB classifier with cross-validation of stratified 10-fold	An accuracy up to 100%	In the microarray datasets, this approach is less reliable and having a small sample size.
Liu et al. (2008)	Periodontal Disease (PD), and Cardiovascular Disease (CVD)	NBSVM, and UDC.	The performance of NB and SVM better than (uncorrelated normal based quadratic Bayes classifier) UDC.	The number of features is limited, so the differences between the test's accuracy levels were not noticeable.
Hussain et al. (2019)	Cancer	ANN, k-NN, DTs, NB, RF, and SVM	RF provides the best performance for the classification	

Assuming the classes are linearly separable, they obtain the hyper-planes with maximum margin to separate them (*Chu & Wang, 2005*; *El-Gamal et al., 2018*; *Gayathri, Sumathi & Santhanam, 2013*; *Karthik & Sudha, 2018*; *Farhadian, Shokouhi & Torkzaban, 2020*).

Quadratic Programming (QP) methods are well Known methods for solving constrained problems to determine the optimum line for the data. When the data cannot be separated linearly, the data is mapped to a larger dimensional space using a kernel function to be divided linearly in this new space. Different kernel functions, like linear,

Table 10 An overview of AD diseases using different ML methods.

Authors	Application on diseases	Methods	Results	Problem with method
Tejeswinee, Shomona & Athilakshmi (2017)	AD	Classification techniques fed by each subset feature selection	Classification accuracy 93%	They cannot generalize all machine learning techniques are efficient in the classification of disease.
Park et al. (2018)	AD	Feature selection & ML &10-fold validation	Classification accuracy 91.6%	The system cannot apply as general for all diseases, so we need to propose a novel system to detect all diseases.
Abd El Hamid, Mabrouk & Omar (2019)	AD	Classification of ML techniques	Accuracy of NB 99.63%	Their system does not support integrating some metadata like gender, age, and smoking to check if these data are associated SNPs or not.
Shahbaz et al. (2019)	AD	ML & DM techniques	Accuracy of GLM is 88.24%	The model cannot be trained with unbalanced data and insufficient data for all classes of disease.
Zafeiris, Rutella & Ball (2018)	AD	ANN	The methodology can reliably generate novel markers	The power of the algorithms and speed are needed to be improved.
Bringas et al. (2020)	AD	CNN	The accuracy 90.91% and an F1-score of 0.897	This methodology needs to be used in a cloud architecture system to collect accelerometer data and a service that users can subscribe to for monitoring changes in the AD stage.

Table 11 Some publicly available datasets.

Authors	Application on diseases	Dataset
Al-Diabat (2018)	ASD	UCI data repository
Ismaeel & Ablahad (2013)	Breast Cancer disease	NCBI
Tejeswinee, Shomona & Athilakshmi (2017)	AD	(KEGG) database
Abd El Hamid, Mabrouk & Omar (2019)	AD	Phase 1 (ADNI-1)/Whole genome sequencing (WGS) datasets (*Rémi et al., 2011*).

radial, polynomial and sigmoid, can be used in this situation. The optimal hyperplane is the initial goal that should be found. The hyperplane is the border between classes. Not only separating between classes is the primary goal of the optimal hyperplane, but it also increases the margin. Margin is the longest distance between the hyperplane and the closest data (support vector) in each category (*Chu & Wang, 2005*; *Karthik & Sudha, 2018*; *Farhadian, Shokouhi & Torkzaban, 2020*).

1. Let $\{x_1,x_2,\ldots,x_n\}$ is a dataset of real value. It is a dataset of real value.
2. Assume classes is $y_i \in \{-1,1\}$ is the label of data, w is a weighted vector. Eq. (10) can be written as follow to estimate hyperplane:

$$f(x) = wx_i + b = 0 \tag{10}$$

Then, from Eqs. (10)–(12) are obtained:

$$wx_i + b \geq +1 \; for \; class + 1 \tag{11}$$

Table 12 Different used applications to detect biomarkers for diseases.

Authors	Application on diseases	Methods	Results	Problem with method
Narayanan et al. (2019)	CAD for malaria, brain tumors, etc	ML	Their approach provided reasonable performance for all these applications	Their research needs to further extend by studying the ROI determined by class activation mapping.
Carter, Dubchak & Holbrook (2001)	Bacterial and archaeal	ML approach using ANN and SVM	RNA genes could be recognized with high confidence using the ML	Future studies are necessary to characterize the extent of these elements and the accuracy of predection.
Narayanan, Hardie & Kebede (2018)	Lung cancer	SVM	CAD perform sensitivity with 82.82	They should optimize the feature set for SVM classification.
Yang et al. (2020)	AD	ML	The AUC to 0.84	
De Velasco Oriol et al. (2019)	LOAD	ML	Classification performance is 72%	The classification performance needs more enhancement.
Ahn et al. (2019)	Business big data business analytics	Fuzzy logic dependent on ML tools	Their improved version led to benefits of additional 10% accuracy.	Their tool need to handle further outliers and edge cases and explore factors that are more explicit and/or implicit in business processes.
Gao & Tembine (2016)	Network systems	Distributed Mean-Field	Mean Square Error =2.01	They use no more complex detectors based on vision. In addition to reducing measuring noise they don't use deep learned features.
Xu et al. (2020)	Smoking prediction Models	SVM and RF	AUC of SVM = 0.720, and AUC of RF = 0.667.	Their work in the future needs to identify the inner complex relations between these SNPs and smoking status.

$$wx_i + b \leq -1 \; for \; class - 1 \tag{12}$$

where x is the input data, w is the normal plane, and b is the position relative to the middle field coordinates.

3. The main goal of SVM is to find high levels between two classes that increase margins.

$$min_w \frac{1}{2}\|w\|^2 \tag{13}$$

$$y_i(wx_i + b) - 1 \geq 0 \tag{14}$$

The problem can be solved in quadratic programming using the Lagrange multipliers shown in Eq. (15):

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{I} \alpha_i(y_i((wx_i + b) - 1)) \tag{15}$$

$$L(\alpha) = \sum_{i=1}^{I} \alpha_i \frac{1}{2} \sum_{j=1}^{I} \alpha_i \alpha_j y_i y_j x_i x_j \tag{16}$$

where, α_i is the weight (parameter obtained from the Lagrangian Multipliers).

4. For making decisions, Eq. (17) is used for linear equations, while Eq. (18) is used for nonlinear equations:

$$f(x_d) = sign\left(\sum_{i=1}^{n} \alpha_i y_i (x_i, x_d) + b \right) \tag{17}$$

$$f(x_d) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x_d) + b \right) \tag{18}$$

where n is the number of support vectors and x_d is the test data and $K(x_i,x_d)$ is the kernel function used, as per Eq. (19).

Linear Kernel:

$$K(\vec{x}_i, \vec{x}_d) = (\vec{x}_i, \vec{x}_d) \tag{19}$$

The radial basis function kernel (RBF) is shown in Eq. (20):

$$K(\vec{x}_i, \vec{x}_d) = exp\left(\frac{-|\vec{x}_i - \vec{x}_d|^2}{2\sigma^2} \right) \tag{20}$$

K-nearest neighbors algorithm (KNN)

KNN is considered one of the most straightforward techniques of DM. It is also known as a memory-based classification because it is at run-time needed for the training examples to be in the memory (*Shouman, Turner & Stocker, 2012*; *Alpaydin, 1997*). However, the major drawback of the KNN classifier is the large memory requirements, which are needed to store the entire sample. When the sample is large, the response time on a serial computer is also significant. In the first step, the closest point to P is found, where P is the point for which the label needs to be predicted. Then, the label is assigned to P at the closest point. Second, the k nearest to P is identified, and the majority of its k neighbors classify points by vote. The most voting class of each object is predicted by its class and by the most votes class. The distances between these points, such as Euclidean, Hamming, Manhattan, and Minkowski, are calculated using distance measures for finding the nearest similar points. The algorithm has the following basic steps: distance calculation, find closest neighbors, and labels voting.

In order to calculate the distance between P and its closest neighbors, there are three most commonly used distances measures:

1. The difference between features is calculated using the Euclidean distance when dealing with continuous features. If the first instance is (a_1, a_2, a_3,…, a_n) and the second

instance is (b_1, b_2, b_3,…,b_n), by the following formula the distance between them is computed :

$$EuclideanDistance = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots (a_n - b_n)^2} \quad (21)$$

The main problem is that the frequencies of large values swamp into small values in dealing with the Euclidean distance formula.

2. Manhattan distance: It is the distance that is usually preferred over the more common Euclidean distance where data are of a high dimension.

$$Manhattandistance = \sum_{i=1}^{K} |a_i - b_i| \quad (22)$$

3. Minkowski distance:

$$Minkowskidistance = \left(\sum_{i=1}^{K} (a_i - b_i)^q \right)^1 / q \quad (23)$$

Logistic regression (LR)

LR is one of the ML classification algorithms for analyzing the dataset in which there are one or more independent variables that identify the outcome and the categorical dependent variable (*Bertram, Lill & Tanzi, 2010*). In many ways, LR is the natural complement of normal linear regression when the target variable is categorized (*Sa'id et al., 2020*). For output (dependent) variable Y to classify two class and input (independent) variable X, let $g(x) = pr(X = x) = 1 - pr(X = x)$, the LR model has a linear form for logit probability as follows:

$$logit[g(x)] = log\left(\frac{g(x)}{1 - g(x)}\right) = \alpha + \beta x \quad (24)$$

where $(\frac{g(x)}{1-g(x)})$ is called odd. The logit has a form of linear approximation. It is equated with the logarithm of the odds. The parameter β is the rate of increase or decrease of the S-shaped curve of g(x).

Performance evaluation metrics

Performance evaluation metrics are used to measure the classification model's performance and investigate how a model works well to achieve the goal. On the test dataset, performance evaluation metrics are used to estimate the classification model's performance and effectiveness, chosen correct metrics. It is essential, such as the confusion matrix in Table 6 to evaluate the model performance. Some commonly used performance metrics include accuracy, precision, and Matthews's correlation coefficient (MCC). which are listed in Table 7 (*Jain, Jain & Jain, 2018*; *Singh & Sivabalakrishnan, 2015*; *Raj & Masood, 2020*; *Le et al., 2020*; *Do & Le, 2019*).

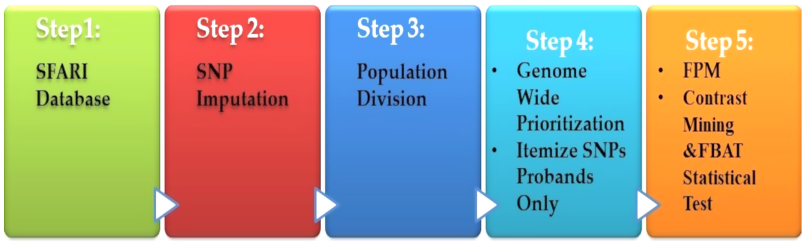

Figure 6 **Diagnosis system for ASD based on DM techniques.**
Full-size DOI: 10.7717/peerj-cs.697/fig-6

LITERATURE REVIEW ON COMPLEX DISEASE & APPLICATIONS

Autism spectrum disorder (ASD)

Spencer et al. (2018) presented a heritable genotype system. Their system consisted of five stages, as shown in Fig. 6. The preprocessing is the first step, which contained the missing value imputations for the SNPs, and then they selected the most SNPs significantly. The second stage is population division. Prioritization is used as a primary association for each genome-wide subgroup procedure. The third stage is the prioritization of genome-wide by turning to the question: what sets of SNPs are they to test? The answer is to use the minimum threshold support in the frequent models (FPM) algorithm. The filter is performed for SNP sets, according to its prevalence among the affected population. The fourth stage is FPM, one of the most DM techniques that excellently identifies the most common occurrence in feature combinations. The data must be interpreted in binary form with both reference states in one person. The presence or absence of the item in FPM shows potential interactions among variants. Finally, the stage is the UICsup, which is a contrast mining utilizing. In the end, the comparison with another application is carried out, and its system for all was superior.

Al-Diabat (2018) presented an autism screening process which comprises presenting some questions to parents, family members, and caregivers to answer the child's behalf to identify potential autism traits. The data mining and fuzzy rule are adapted to improve the screen efficiency and accuracy, which is considered one potential solution. To discover the features of ASD, the fuzzy unordered rule induction (FURIA) algorithm was evaluated. The automated method, FURIA, is designed to examine patterns from controls and historical cases. After then, they used the models to determine the possibility of autism traits in new individuals. The higher performance of fuzzy data mining models is evidenced by experimental results regarding predictive accuracy and sensitivity rates.

Alzubi et al. (2017) developed an accurate method based on hybrid feature selection for identifying the most SNPs that are informative and chosen an optimal SNP subset. Their method was based on the fusion of the wrapper and filter method. The system performance was evaluated against four feature selection methods and four classifiers on SNPs' five different data sets. The experiment results show the adopted feature selection approach's efficiency and achieve classification accuracy up to 96% for the used data set. Overall, they concluded that the whole genome could efficiently differentiate between

people with complex diseases and healthy individuals. Their method has been validated in an independent and large case and evidence study.

Krishnan et al. (2016) introduced a complementary ML approach based on a network for the human brain's specific genes to present a wide prediction genome of autism risk genes, comprising candidates who may be hundreds is minimal or no previous genetic evidence. Leveraging these genome- and network-specific predictions of the brain, we showed that a large group of ASD genes converge in fewer major pathways and developmental stages of the brain. Finally, they identified potentially pathogenic genes within autism-related CNVs and suggested genes and pathways that are likely mediators of ASD across multiple CNVs.

The model decision function output ranks the gene lists comprising an ASD risk gene and adjacent genes. *Cogill & Wang (2016)* developed SVM with training on brain developmental gene expression data for classification and prioritization of ASD risk genes. The pre-reduction process has been shown to improve the accuracy of the SVM classification. The results showed that each filter procedure identifies different gene sets with a specific gene repetition. Due to the high variance in ASD of several genes, further selection steps are necessary.

Jiao et al. (2012) presented an SNP-based predictive model which could predict the severity of ASD's symptoms. They divided 118 ASD children into a moderate group of autisms (n = 65) and a severe group of autisms (n = 53). 29 SNPs of 9 ASD-related genes were obtained for every child. They applied three ML techniques to create predictive models: Decision stumps (DSs), ADTrees, and FlexTrees. With an accuracy = 67%, DS and FlexTree produced modestly better classifiers. All models of the SNP rs878960 in GABRB3 were selected and linked to the CARS evaluation.

Cancer & brain diseases

Ismaeel & Ablahad (2013) developed a method that can predict the disease by mutations. Bioinformatics techniques were used to train and used back-propagation algorithms to test whether the patient has the disease or not on the collected data, using all expected mutations for genes of some diseases (*e.g.*, BRCA1 and BRCA2). They Implemented their method as the first way of predicting the disease based on mutations in the gene sequence causing this disease which showed two decisions were achieved successfully, as shown in Fig. 7. The first way was to diagnose whether a patient had cancer mutations or not by using bioinformatics techniques. Back-propagation is the second classification of these mutations.

Jain, Jain & Jain (2018) introduced a two-phase hybrid model for cancer classification. A low-dimensional set is selected for this model of prognostic genes. For biological samples classification into binary and multi-class cancers using naïve Bayes (NB) classifier with a cross-validation stratified 10-fold. Different cancer types have been evaluated in their system by using 11 microarray datasets. The experiment results are compared with seven methods. Their system achieved better results, showing the classification accuracy and the selected number of genes in most cases. It is obtained classification accuracy for seven

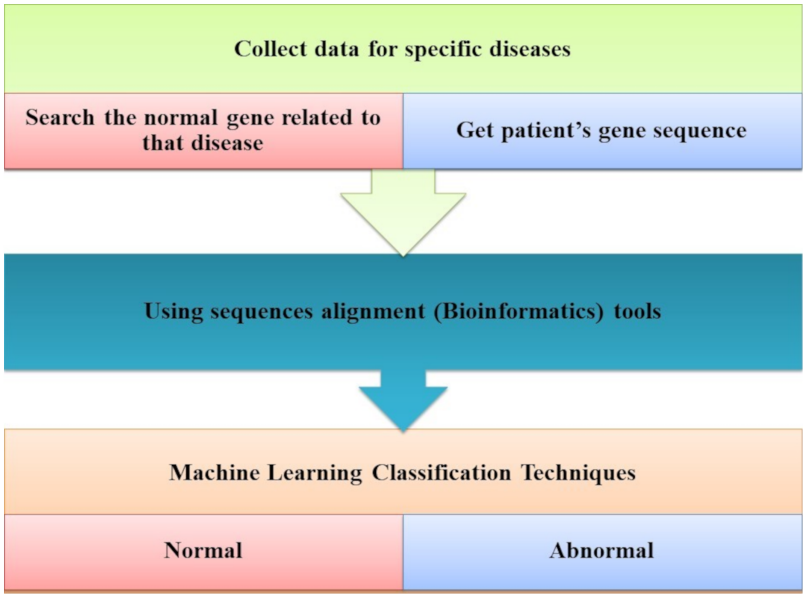

Figure 7 **The main tasks of the gene-based analysis system.**
Full-size DOI: 10.7717/peerj-cs.697/fig-7

datasets up to 100% with prognostic gene subset with very small sizes (up to <1.5%) for all datasets.

Liu et al. (2008) developed a supervised recursive feature addition (SRFA) method for feature selection. This approach chose candidate features/SNPs. Supervised learning and statistical measures are combined to control the redundancy information and enhance classification performance in association studies. Additionally, they have developed a support vector-based recursive feature addition (SVRFA) scheme in the SNP association analysis of disease.

Hussain et al. (2019) analyzed the DNA sequences of cancer patients using various classification techniques. Based on this study, it can be concluded that people's cancer diseases can be diagnosed easily based on their DNA sequences. Different classifiers have been used to analyze purposes, such as an artificial neural network (ANN), KNN, decision trees (DTs), and fuzzy classifiers, NB, random forest (RF), and SVM. Based on the analysis of results, it has been found that these classifiers provide sufficient performance in terms of accuracy, recall, specificity, and other parameters. By using the gene dataset, the paper used different classifiers to analyze the data. They found the RF provides the best performance for the classification of cancer patients.

Alzheimer's disease (AD)

The most common forms of dementia that degenerate neurons are AD and PD found in the brain cells. *Tejeswinee, Shomona & Athilakshmi (2017)* described a computational framework to investigate neurodegenerative disorders, as shown in Fig. 5. The collection of data that they performed has not been the computational explorations of earlier utilization. In the first stage, the genes related to AD and PD are collected as the dataset generation.

Figure 8 **The computational system for AD disorders.** Full-size DOI: 10.7717/peerj-cs.697/fig-8

Genes uniquely of AD are 74 and 38 genes uniquely of PD, representing 112 of genes collected. Both diseases have 95 common genes. Feature selection is the second step. Three feature selection methodologies are used in this study to select the optimal feature set. The result revealed that it is the best-optimized feature subset. This subset is used to feed the six classification algorithms individually to each of them in the classification phase, as shown in Fig. 8. The accuracy of all these algorithms was measured when predicting the correct diagnostic class.

Park et al. (2018) introduced genetic interactions based on ML from heterogeneous gene expression profiles, as presented in Fig. 9. Related gene determination of disease and dis-ease mechanisms is a research goal and necessary process; this problem has been approached with many studies problems by analyzing gene expression profiles and datasets interaction for genetic networks. Among pairs of genes, correlations or associations must be determined to construct a gene network. However, when heterogeneous data for gene expression is noisy with high levels for assigned samples to the same condition, it is difficult to specify whether a gene's pair represents a significant (GGI). To find a solution for this challenge, they introduced an RF-based method to classify data of gene expression of significant GGIs. The model is trained by defining sets of novel features and utilizing various datasets with high confidence interactome.

Abd El Hamid, Mabrouk & Omar (2019) developed a method for detecting biomarkers SNPs associated with the disease with high accuracy classification, leading to early disease prediction and diagnosis. For discovering new biomarkers, ML techniques are utilized for the dis-ease. Many common diseases have SNPs related to AD. SNPs are recognized for this disease as significant biomarkers. In its early stages, they allow in understanding and detecting the disease. They apply many techniques that are performed on all genetic data of AD. They achieved the highest accuracy of classification. The results detected that K2 learning and NB algorithms accomplished an accuracy of 98.40% and 98%, respectively.

Shahbaz et al. (2019) introduced an AD classification using ML techniques to classify AD. In their paper, they performed six different ML and DM algorithms. The first step in their work is a data processing and feature selection. The second step is to partition data partitioning using 70% of training data and the reset for testing. Finally, classification techniques are used. Five stages of AD's can efficiently classify using the generalized linear

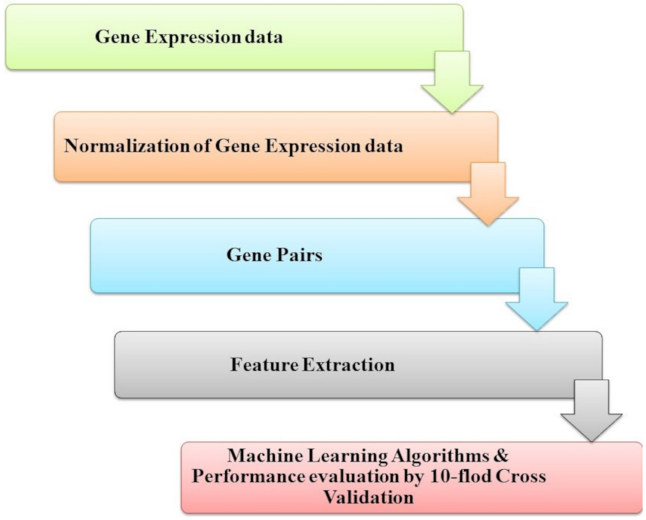

Figure 9 **The identification of genetic interactions system.**
Full-size DOI: 10.7717/peerj-cs.697/fig-9

model (GLM) described in the study results on the test dataset with 88.24% accuracy. In medicine and healthcare, these techniques can be successful for early disease detection and diagnosis.

Zafeiris, Rutella & Ball (2018) discussed, explored, and evaluated an integrated ANN pipeline in AD for biomarker discovery and validation. Over time, the most popular form of dementia and no specific cause and no treatment available. The developed system consists of analyzing data that are public with a categorical and predicting gene interactions. A continuous stepwise algorithm is used and further examined through network inference. Novel markers can be generated by their methodology and another well-known study that could be used to guide future research in AD.

Bringas et al. (2020) implemented deep learning model AD stage identification. Data from a daycare center collected for one week from 35 AD patients on smartphones were used. The data sequences for each patient recorded changes in the accelerometer during daily operations and labeled with the disease stage (early, middle or late). Their methodology uses the CNN model to identify the patterns which identify each step. Their methodology is used to process these time series. As a result, the 90.91% accuracy and an F1-score of 0.897 were achieved through the CNN-based method, which greatly improves the results achieved by classifying features.

Different applications

Narayanan et al. (2019) provided a computer-aided detection (CAD), and it is considered as a hot research area that is attracting significant interest in the past ten years. ML algorithms have been widely used for this application because they provide clinicians with valuable opinions to make the right decision. Regardless of the many ML models available for medical imaging applications, not much has been implemented in the real world due to the inexplicable nature of the network's decisions. Their research paper investigated

the results supported by deep neural networks to detect malaria, brain tumors, etc., in various imaging methods. They visualized category activation mappings for all applications to improve understanding of these networks. This work will help data science experts understand their models and assist clinicians in the decision-making process. The main drawback of their work should extend by studying the region of interest (RoI).

Carter, Dubchak & Holbrook (2001) developed an ML approach using ANN and SVM to extract standard features among known Ribonucleic acids (RNAs) to predict new RNA genes in regions not shown on them. Results showed RNA genetic differences in composition and structural parameters in known RNAs compared to non-coding sequences could be recognized with high confidence using M L-Approach in both bacterial and archaeal genomes. Besides cross-validation testing, highly predicted sequences in *E. coil*, M.geneitalium, etc., not involved in the training datasets, were experimentally distinguished and reported in the literature as expressed fRNAs.

Narayanan, Hardie & Kebede (2018) presented a study of lung cancer and explored SVM's performance based on a wide range of features. The SVM performance was studied according to the number of features. Their results demonstrated greater robustness, faster computation with a wide range of features, and less prone to overtraining than conventional classifiers. Besides, they also offered a computationally efficient approach to select SVM features. Results are presented to the publicly available 2016 lung nodule analysis dataset. Their results show that the SVM is superior to the Fisher linear discrimination classifier based on 10-fold cross-validation by 14.8%.

Yang et al. (2020) provided a review that introduced sequencing technology development and explains the structure of DNA sequence data and sequence similarity. Second, they analyzed the necessary DM process, summarized several of the significant ML algorithms, and highlighted the future challenges faced by ML algorithms in extracting biological sequence data and possible future solutions. Third, they reviewed four typical applications of ML in DNA sequencing data. Fourth, they analyzed the background and relevance of the corresponding biological application. Fifth, they have systematically summarized the evolution and potential problems in the field of DM for DNA sequencing in recent years. Finally, they summarized the content of the review and looked forward to the future for some research directions.

Bellinger et al. (2017) performed a systematic review of the application of DM and ML methods in the epidemiology of air pollution. A systematic review identified current trends and challenges and explored recent trends for applying data extraction methods to air pollution epidemiology. Their work presented that DM is increasingly performed in the epidemiology of air pollution. The potential is to enhance air pollution epidemiology continues to grow with DM's advances for temporal and geospatial mining and deep learning. New sensors and storage media, which allow greater and better information, also support this. This shows that a large number of successful applications can in the future be expected.

Bracher-Smith, Crawford & Escott-Price (2021) introduced a review for psychiatric disorders based on ML for genetic prediction. ML methods have been performed to make predictions in psychiatry from genotypes, with the potential to enhance the process of

prediction of outcomes in psychiatric genetics; however, their current performance is unclear. This paper aimed to systematically review ML methods for psychiatric disorders prediction from genetics alone and estimate their discrimination, bias, and implementation.

Romero-Rosales et al. (2020) compared three ML methods that have been shown to build robust predictive models (genetic algorithms, LASSO, and progressive wisdom) and provided the involvement of markers from misclassified samples to enhance the accuracy of overall prediction. Their results described that adding markers from a prototype plus model markers fitted to misclassified samples enhances the AUC about 5%, to ≈ 0.84, which is very competitive using only the genetic information. The computational strategy applied here can support good ways to enhance classification models for AD. Their work could have a positive effect on early AD diagnosis.

De Velasco Oriol et al. (2019) performed systematic comparisons of representative ML models to predict the late onset of AD (LOAD) from the ADNI cohort's genetic variance data. Their experimental results demonstrated that classification performance is the best-tested model resulting in 72% of the area under the receiver operating characteristic (ROC) curve.

Mishra & Li (2020) introduced the first intensive reviews for applying AI technology in medicine and the current genetic research state in AD. Next, the extensive review concentrated on applying AI in genetic research of AD, including diagnosis and prediction of AD based on genetic data, analysis of genetic variance, gene expression profile, and gene expression. AD analysis based on the knowledge base. Although several studies have yielded some meaningful results, it is still in the preliminary stage. Fundamental shortcomings include database limitations, failure to leverage AI to conduct a systematic biological analysis of multi-level databases, and a theoretical framework for analysis results. Finally, the future direction of development is to aspire to high quality, comprehensive sample size, and data sharing resources should be developed. An AI analysis strategy for multi-level system biology is a development trend. Computational innovation may play a role in constructing and validating the theoretical model and designing new intervention protocols for AD.

Ahn et al. (2019) introduced fuzzy logic dependent on ML tools for enhancing business big data business analytics in complex AI environments. Specifically, they have provided a suitable and extended the traditional C4.5 algorithm by incorporating fuzzy logic for business processes. To classify and predict the individual wage, government staff, help take actions appropriate to those in need and/or make better use of their available resources. The resulting tool supported analysts in this process. Additionally, they have also released an enhanced version, and this base copy has been reinforced by branch extension. The enhanced version speeded up the mining process and the business analytics and led to higher accuracy.

Gao & Tembine (2016) presented networked systems for the effective way of extraction and utilization and the availability of accurate prediction, proactive tools of mean-field type dynamical systems. They introduced large-scale networked systems based on the distributed mean-field filter (DMF). The filter exploits the network's topology and breaks it

down into highly independent components concerning marginal mean-field correlations. The experiments showed for two object tracking scenarios are performed to illustrate the performance of their algorithm. Results of the evaluation show that DMF superior to the existing filtering algorithms.

Yazdani et al. (2020) introduced simultaneously complex and common features that arise from multiple genes interacted and regulated. Therefore, to unravel the fundamental biological networks, it is vital to characterize genes' interconnectivity. They have systematically combined transcription, genotyping, and Hi-C data to determine the interconnections between individual genes as a causal network. They used various ML techniques to extract information from the network, determine differential regulatory patterns between cases, and control schizophrenia.

Yazdani et al. (2018) introduced a method which is known as Bounded Fuzzy Possibilistic Method (BFPM) that takes into account preserving the flexibility of the search space for higher accuracy cluster sampling (*Yazdani, 2020*). The method assessed samples for their movement from one cluster to another. This technique lets us finding important samples in advance of those with the potential to belong to other clusters in the near future. BFPM was performed on the metabolism of individuals in a lung cancer case-control study. Metabolism as close molecular signals of actual disease processes may be powerful biomarkers of the current disease process. They aimed to find out whether healthy human serum metabolites can be distinguished from those with lung cancer. With BFPM, some differences were noticed, pathology data were evaluated, and essential samples were identified.

Xu et al. (2020) used the methods of SVM and RF for developing smoking prediction models. For a model building of 10- fold cross-validation, they first used 1,431 smokers and 1.503 non-smokers and tested the model prediction models on independent datasets of 213 smokers and 224 non-smokers. AUC of 0.691, 0.721 and 0.720 for training, tests, and independent test samples were obtained by SVM with the 500 top SNPs, selected using LR methods ($p < 0.01$). AUCs 0.671.665 and 0.667 were obtained for the training, the tests, and independent test samples of 500 top SNPs selected using LR ($p < 0.01$).

Le & Nguyen (2019) applied DL in the prediction of SNARE proteins, which is one of life science's most essential molecular functions. Several human illnesses involve a functional loss of SNARE proteins (*e.g.*, neurodegenerative, mental illness, cancer, and so on). There is thus a critical problem in understanding these diseases and designing the drug targets by establishing a precise model to determine their functions. With 2D convolutional neural network (CNN) and position-specific scoring matrix profiles, their SNARE-CNN model could identify the accuracy reached with the SNARE proteins of 89.7%.

In many aspects of cellular life activities, antioxidant proteins are essential. Cell and DNA are protected against oxidative agents. In this study, *Ho Thanh Lam et al. (2020)* developed an ML model from a benchmark set of sequence data that was applied for this prediction purpose. The experiments have been performed through 10-fold cross-validation during the training and validated by three separate datasets. On the optimum set of sequence features, various ML and DL algorithms were evaluated. Among

Table 13 Different applications of ML and their accuracy.

Authors	Methods	Accuracy	Disease
Alzubi et al. (2017)	ML	96%	ASD
Tejeswinee, Shomona & Athilakshmi (2017)	ML	93%	AD
Al-Diabat (2018)	DM and Fuzzy Rules	91.35%	ASD
Jain, Jain & Jain (2018)	NB	100%	Cancer
Park et al. (2018)	ML	91.6%	AD
Narayanan, Hardie & Kebede (2018)	SVM	82.82%	Lung Cancer
Abd El Hamid, Mabrouk & Omar (2019)	NB	99.63%	AD
Shahbaz et al. (2019)	GLM	88.24%	AD
Bringas et al. (2020)	CNN	90.91%	AD
Yang et al. (2020)	ML	84%	AD

them, RF was identified as the best model for identifying antioxidant proteins with the highest performance. Their optimal model has achieved 84.6% high accuracy.

CRITICAL DISCUSSION AND FUTURE CHALLENGES

Microarray data's emergence poses many ML research challenges due to its large dimensional nature with small sample size. Aside from the notable disadvantage of having many features of a small number of samples, researchers also have to meet the unbalanced classes characterizing the data, testing datasets and extracting training in various cases, and the presence of outliers (dataset shift). For all these reasons, new technologies continue to emerge every year, aiming to improve the classification accuracy of previous approaches and help biologists discover and understand the underlying mechanism linking gene expression to diseases. So, we added Table 13, which displays the classification accuracy of previous approaches.

So, it is useful in biological research to determine individuals' phenotype and genotype characteristics. The phenotype connected with physical appearance, and the genotype is the genotype of the individual. SNPs make the individual different from others as well as help in determining genetic variation in the population. The main factor in identifying the relationship of an individual to a population is genomic variation. SNPs play an essential role in genome-based disease detection, drug design, drug and isolation reaction to an environmental factor, such as toxins and disease development risk in a population, so it is used for this purpose (*Tahir & Sardaraz, 2020*).

Biomarkers are urgently needed for the early detection of complex genetically related brain diseases. So, the main goal is to identify a set of genetic variants that happened together. Besides, it can determine individuals who must start screening at a young age (*Chen et al., 2016*). Determination of related genes to disease and disease mechanisms is a research goal and a necessary process. Some challenges are summarized in the following points, and also as shown in Fig. 10:

Data processing: The processing of large-scale DNA sequence data still presents efficiency challenges.

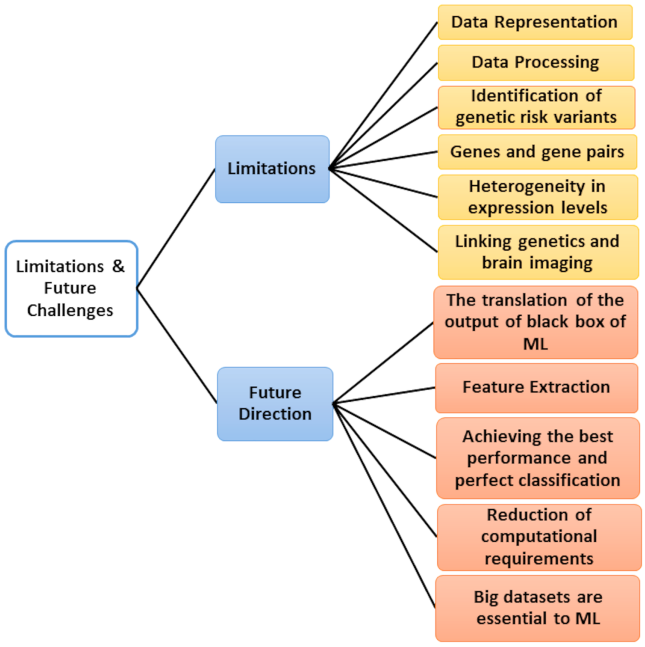

Figure 10 **Limitations and future directions.** Full-size DOI: 10.7717/peerj-cs.697/fig-10

Data Representation: Certain problems exist with the quantification of DNA sequence aspects. No one knows which representation in these nucleotides is best for encoding numeric values. However, we cannot avoid using the numeric representations of those biological units when applying ML to biological studies.

Genes and gene pairs: A new method is required to discover associated genes and gene pairs or to study the relation between genes with subgroups of disorders. Recent discoveries about the new sub-classification of AD patients make this disease a suitable target for this method. To verify their significance, autism candidate genes identified in this study should be integrated into future data collection. These genes are contributed by understanding the mechanisms of genes to develop autism subtypes (*Jain, Jain & Jain, 2018*).

Identification of genetic risk variants: Identifying genetic risk variants in ASD and their effects on the brain and morphology remains largely elusive. More recently, the advent of a gene imaging approach integrates genetic discoveries with technological advances in brain imaging. Researchers were strongly encouraged to analyze ASD pathophysiology mechanisms and to characterize neurological systems which are directly affected by risk gene variants (*Jain, Jain & Jain, 2018*; *Spencer et al., 2018*).

Heterogeneity in expression levels across large samples: Due to heterogeneity, there may be an asymmetry in this effect on sample quality among patients with genetic brain disease and the degree of affected gene expression. Recently, an attempt was made to identify heterozygous genes in the study of gene expression data with genetic brain disease patients (*Tejeswinee, Shomona & Athilakshmi, 2017* , *Escott-Price et al., 2014*; *Baker et al., 2019*; *Ruiz et al., 2014*; *Edwards, Stajich & Hansen, 2009*; *Staples et al., 2019*; *Pop & Salzberg, 2008*; *Yazdani et al., 2018*) .

Linking genetics and brain imaging: The genetic imaging techniques make it possible to improve our understanding of the etiology of genetic diseases by bridging the gap between genetic differences and their resulting biological effects on the brain (*Chu & Wang, 2005*; *Hussain et al., 2019*).

The capability of generalizing: The ability to generalize and adapt techniques to various data is considered the most challenging task. Also, the ability to reduce hardware power requirements is considered a critical task.

Genetic structure for rare diseases and common complex diseases: Rare and common diseases may be used for the study of Next Generation Sequencing (NGS). Causal variants can be found even in smaller sample sizes for single-gene rare diseases. It doesn't remain easy, however, above all to define causal alleles. Not all rare diseases have a simple genetic structure like single-gene diseases. Rare diseases are usually diagnosed or identified by symptoms, while different mechanisms can cause the same symptoms. In fact, some rare diseases are a group of diseases that exhibit similar symptoms (*Gao & Tembine, 2016*). Determination of causal variants of these diseases generally requires larger sample sizes than monogenic diseases. NGS technologies provide a huge quantity of biological data, presenting different problems, including high processing times and high memory requirements. Research is therefore aimed at detecting SNP in genome sequences which can solve these problems. Also, there are many problems with SNPs detection algorithms, *e.g.*, overhead computing costs, accuracy, and memory requirements. These issues are considered as another challenge for SNP identification (*Tahir & Sardaraz, 2020*; *Wang, Lu & Zhao, 2015*).

In the future, researchers should focus on identifying variations of rare diseases. They have great potential in the last few years, and genotyping and sequencing technologies have advanced. The researchers should also concentrate on these discoveries, the technologies that supported their discovery of rare variant detection and working with big data as challenges. Another challenge is imaging genetics, which addresses this question by bridging the gap between imaging and the genetic fields. Finally, finding the relationship between genes also helps discover complex diseases that are affected by risk genes.

In the future, we need to learn how to use biomarkers together in an integrated manner and learn to map them to specific necessary treatment actions or symptoms. Biomarkers may also be very helpful in explaining the significant variability associated with genetic diseases. The underlying pathophysiology symptoms of genetic diseases may vary significantly from patient to patient. ML is the DM core and the most widely used method of data processing. The main advantage of ML algorithms is that it can be used to filter large quantities of data to check patterns which could be ignored otherwise. ML plays a crucial role in discovering predictable patterns in biological systems and big data biomedical research. The current implementation of ML in biomedical data mainly faces the following issues:

1. Big datasets are essential to ML. Currently, most biological datasets are still too small to meet the requirements of ML algorithms. Although the total amount of biological data is

huge and increases daily, data collection is obtained from various platforms. Due to technology and biology differences themselves, it is challenging for different data sets to integrate.

2. Because of biological data differences itself, trained ML models on one dataset may not generalize well to other datasets. The results of the analysis of the ML model can be wrong if the new data is significantly different from the training data.
3. The black-box nature of ML models obtained a new challenge to biological applications. The translation of the output of a specific model from a biological viewpoint that limits the use of the model is usually difficult.
4. Reduction of computational requirements: ML models are often very complex and require a lot of training. It is often computer-intensive and memory-intensive to produce well-trained modeling and even to use the model productively. These requirements limit the deployment of ML seriously on computer-saving machines, especially in the bioinformatics and healthcare sectors, which are also data-intensive. Several methods were proposed for ML compression, which can at the very beginning reduce the computational requirements of these models, like parameter pruning, which reduces the redundant parameters that do not contribute to the performance of the model.
5. Achieving the best performance and perfect classification by implementing various optimization techniques.
6. Additional attempts could be made to extract other DNA features before feeding them to the models. Overall, this could lead to more accurate assessments.

Currently, many softwares can be helpful for ML, and knowledge discovery tools for different uses are available, such as the Waikato Environment for Knowledge Analysis (WEKA), RapidMiner, Clementine, Intelligent Miner, Rosetta, etc. These tools and software provide methods and algorithms that help users better use information and data, including data analysis methods and algorithms, cluster analysis, etc (*Borges, Marques & Bernardino, 2013*).

RapidMiner (RM): It is an ML and DM processing environment. It is an open-source, free Java project. It is a new way of designing all the complicated problems. For the input and output of data in various file formats, RM has flexible operators. It includes over 100 learning schemes for regression, classification, and clustering tasks (*Borges, Marques & Bernardino, 2013*).

WEKA: It is an open-source, non-commercial project. WEKA includes tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also ideal for developing new systems of ML (*Borges, Marques & Bernardino, 2013*).

R: It is one of the key data science languages. It offers excellent visualization features that are essential to explore the data in an automated learning process before it is submitted and the results of the learning algorithm assessed. A lot of R packages are available for ML, and R implements numerous modern methods for statistical learning (*Borges, Marques & Bernardino, 2013*).

So, ML models will in the future offer many opportunities for discovering new insights from the research gaps discussed. In particular, analysis of genetic variations that lead to diseases, such as ASD, AD, cancer, and other fatal diseases will help find out a way to cure new medicines and therapies permanently. As shown in Fig. 9, we present limitations and future directions to handle it in the future.

CONCLUSION

This review comprehensively reviews genetic variations analysis of gene expression analysis to discover complex genes associated with diseases and related genetic diseases. This survey describes various diseases, such as AD, ASD, and cancer, to identify genetic variations that cause diseases. Here, we describe variation identified for putative rare genetic risk. We observe that variants' rare identification can decrease disease susceptibility than generally seen with common risk variation and protein-coding changes that can be modeled efficiently. Finally, we have highlighted some open issues and future research directions for the progress of this field. Also, genetic mapping supplies a powerful method for identifying the presence of SNPs in genes. These SNPs are found in DNA, which is very similar to mutation. But this SNP is damaged DNA in humans, causing serious disease in the future. So, our focus is to detect the SNPs that lead to diseases. New genome-wide multiple gene testing or high-throughput genetic testing has come with new hope in the field, offering a fast, cost-effective, and intensive analysis of genetic variation. This is particularly interesting with high genetic disorders heterogeneity. Therefore, the interaction between data mining, machine learning, and bioinformatics should be increased as a great potential for analyzing gene sequences in the future.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Hala Ahmed conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Louai Alarabi conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.
- Shaker El-Sappagh conceived and designed the experiments, performed the experiments, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.

- Hassan Soliman conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Mohammed Elmogy conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This literature review has no data.

REFERENCES

Abd El Hamid MM, Mabrouk MS, Omar YM. 2019. Developing an early predictive system for identifying genetic biomarkers associated to alzheimer's disease using machine learning techniques. *Biomedical Engineering: Applications, Basis and Communications* **31(5)**:1950040.

Abd El Hamid MM, Omar YM, Mabrouk MS. 2016. Identifying genetic biomarkers associated to alzheimer's disease using support vector machine. In: *2016 8th Cairo International Biomedical Engineering Conference (CIBEC)*. Piscataway: IEEE, 5–9.

Adiwijaya, Wisesty UN, Lisnawati E, Aditsania A, Kusumo DS. 2018. Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. *Journal of Computer Science* **14(11)**:1521–1530 DOI 10.3844/jcssp.2018.1521.1530.

Ahn S, Couture SV, Cuzzocrea A, Dam K, Grasso GM, Leung CK, McCormick KL, Wodi BH. 2019. A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments. In: *2019 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. Piscataway: IEEE, 1–6.

Al-Diabat M. 2018. Fuzzy data mining for autism classification of children. *International Journal of Advanced Computer Science and Applications* **9(7)**:11–17 DOI 10.14569/issn.2156-5570.

Alpaydin E. 1997. Voting over multiple condensed nearest neighbors. In: *Lazy Learning*. Berlin: Springer, 115–132.

Alzubi R, Ramzan N, Alzoubi H, Amira A. 2017. A hybrid feature selection method for complex diseases SNPs. *IEEE Access* **6**:1292–1301 DOI 10.1109/ACCESS.2017.2778268.

Ang JC, Mirzal A, Haron H, Hamed HNA. 2015. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13(5)**:971–989 DOI 10.1109/TCBB.2015.2478454.

Asif M, Martiniano HF, Vicente AM, Couto FM. 2018. Identifying disease genes using machine learning and gene functional similarities, assessed through gene ontology. *PLOS ONE* **13(12)**:e0208626.

Association A. 2019. 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **15(3)**:321–387.

Baker E, Sims R, Leonenko G, Frizzati A, Harwood JC, Grozeva D, Morgan K, Passmore P, Holmes C, Powell J, Brayne C, Gill M, Mead S, Bossù P, Spalletta G, Goate AM, Cruchaga C, Maier W, Heun R, Jessen F, Peters O, Dichgans M, FröLich L, Ramirez A, Jones L, Hardy J, Ivanov D, Hill M, Holmans P, Allen ND, Morgan BP, Seshadri S, Schellenberg GD, Amouyel P, Williams J, Escott-Price V, GERAD/PERADES, CHARGE, ADGC, EADI, IGAP consortia. 2019. Gene-based analysis in hrc imputed genome wide association data identifies three novel genes for Alzheimer's disease. *PLOS ONE* **14(7)**:e0218111.

Bansal D, Chhikara R, Khanna K, Gupta P. 2018. Comparative analysis of various machine learning algorithms for detecting dementia. *Procedia Computer Science* **132**:1497–1502.

Barnes MR. 2010. Genetic variation analysis for biomedical researchers: a primer. *Methods in Molecular Biology* **628**:1–20 DOI 10.1007/978-1-60327-367-1_1.

Batnyam N, Gantulga A, Oh S. 2013. An efficient classification for single nucleotide polymorphism (SNP) dataset. In: *Computer and Information Science*. Berlin: Springer, 171–185.

Bellinger C, Jabbar MSM, Zaane O, Osornio-Vargas A. 2017. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health* **17(1)**:1–19 DOI 10.1186/s12889-017-4914-3.

Berry NK, Scott RJ, Rowlings P, Enjeti AK. 2019. Clinical use of SNP-microarrays for the detection of genome-wide changes in haematological malignancies. *Critical Reviews in Oncology/Hematology* **142(10)**:58–67 DOI 10.1016/j.critrevonc.2019.07.016.

Bertram L, Lill CM, Tanzi RE. 2010. The genetics of Alzheimer disease: back to the future. *Neuron* **68(2)**:270–281 DOI 10.1016/j.neuron.2010.10.013.

Bertram L, Tanzi RE. 2012. The genetics of Alzheimer's disease. *Progress in Molecular Biology and Translational Science* **107(6)**:79–100 DOI 10.1016/B978-0-12-385883-2.00008-4.

Borges LC, Marques VM, Bernardino J. 2013. Comparison of data mining techniques and tools for data classification. In: *Proceedings of the International C* Conference on Computer Science and Software Engineering*. 113–116.

Bracher-Smith M, Crawford K, Escott-Price V. 2021. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular Psychiatry* **26**:1–10 DOI 10.1038/s41380-020-0825-2.

Breiman L. 2001. Random forests. *Machine learning* **45(1)**:5–32 DOI 10.1023/A:1010933404324.

Bringas S, Salomón S, Duque R, Lage C, Montaña JL. 2020. Alzheimer's disease stage identification using deep learning models. *Journal of Biomedical Informatics* **109(9)**:103514 DOI 10.1016/j.jbi.2020.103514.

Bumgarner R. 2013. Overview of DNA microarrays: types, applications, and their future. *Current Protocols in Molecular Biology* **101(1)**:22.

Carter RJ, Dubchak I, Holbrook SR. 2001. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research* **29(19)**:3928–3938 DOI 10.1093/nar/29.19.3928.

Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* **40(1)**:16–28 DOI 10.1016/j.compeleceng.2013.11.024.

Chen L-F, Su C-T, Chen K-H, Wang P-C. 2012. Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. *Neural Computing and Applications* **21(8)**:2087–2096 DOI 10.1007/s00521-011-0632-4.

Chen J, Wu X, Huang Y, Chen W, Brand RE, Killary AM, Sen S, Frazier ML. 2016. Identification of genetic variants predictive of early onset pancreatic cancer through a population science analysis of functional genomic datasets. *Oncotarget* **7(35)**:56480–56490 DOI 10.18632/oncotarget.10924.

Chu F, Wang L. 2005. Applications of support vector machines to cancer classification with microarray data. *International Journal of Neural Systems* **15(6)**:475–484 DOI 10.1142/S0129065705000396.

Clare A, King RD. 2001. Knowledge discovery in multi-label phenotype data. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 42–53.

Coelho L, Goertzel B, Pennachin C, Heward C. 2009. Classifier ensemble based analysis of a genome-wide SNP dataset concerning late-onset Alzheimer disease. In: *2009 8th IEEE International Conference on Cognitive Informatics*. Piscataway: IEEE, 469–475.

Cogill S, Wang L. 2016. Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. *Bioinformatics* **32(23)**:3611–3618 DOI 10.1093/bioinformatics/btw498.

Daoud M, Mayo M. 2019. A survey of neural network-based cancer prediction models from microarray data. *Artificial Intelligence in Medicine* **97**:204–214.

De Velasco Oriol J, Vallejo EE, Estrada K, Peña JGT, The Alzheimer's Disease Neuroimaging Initiative. 2019. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinformatics* **20(1)**:1–17.

Do DT, Le NQK. 2019. A sequence-based approach for identifying recombination spots in saccharomyces cerevisiae by using hyper-parameter optimization in fasttext and support vector machine. *Chemometrics and Intelligent Laboratory Systems* **194**:103855.

Edwards D, Stajich J, Hansen D. 2009. *Bioinformatics: tools and applications*. Berlin: Springer Science & Business Media.

El-Gamal FE-ZA, Elmogy MM, Khalil A, Ghazal M, Soliman H, Atwan A, Keynton R, Barnes GN, El-Baz AS. 2018. A significant regional-based diagnosis system for early detection of Alzheimer's disease using smri scans. In: *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. Piscataway: IEEE, 407–412.

Escott-Price V, Bellenguez C, Wang L-S, Choi S-H, Harold D, Jones L, Holmans P, Gerrish A, Vedernikov A, Richards A, DeStefano AL, Lambert J-C, Ibrahim-Verbaas CA, Naj AC, Sims R, Jun G, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thornton-Wells TA, Denning N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin C-F, Schmidt H, Kunkle B, Dunstan ML, Vronskaya M, Johnson AD, Ruiz A, Bihoreau MT, Reitz C, Pasquier F, Hollingworth P, Hanon O, Fitzpatrick AL, Buxbaum JD, Campion D, Crane PK, Baldwin C, Becker T, Gudnason V, Cruchaga C, Craig D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Hernández I, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fiévet N, Huentelman MJ, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuinness B, Larson EB, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossù P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Garcia FS, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Scarpini E, Bonuccelli U, Mancuso M, Siciliano G, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Frank-García A, Panza F, Solfrizzi V, Caffarra P, Nacmias B, Perry W, Mayhaus M, Lannfelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH Jr, Bennett DA, Harris TB, Fratiglioni L, Holmes C, De Bruijn RFAG, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JSK, Boerwinkle E, Riemenschneider M, Boada M, Hiltunen M, Martin ER, Schmidt R, DRujescu R, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nöthen MM, Graff C, Psaty BM, Haines JL, Lathrop M, Pericak-Vance MA, Launer LJ, Van Broeckhoven C, Farrer LA, Van Duijn CM, Ramirez A, Seshadri S, Schellenberg GD, Amouyel P, Williams J, Cardiovascular Health Study (CHS), United

Kingdom Brain Expression Consortium. 2014. Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease. *PLOS ONE* **9(6)**:e94661.

Farhadian M, Shokouhi P, Torkzaban P. 2020. A decision support system based on support vector machine for diagnosis of periodontal disease. *BMC Research Notes* **13(1)**:1–6.

Gao J, Tembine H. 2016. Distributed mean-field-type filters for big data assimilation. In: *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. Piscataway: IEEE, 1446–1453.

Gayathri B, Sumathi C, Santhanam T. 2013. Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems* **4(3)**:105.

González F, Belanche LA. 2013. Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy. *arXiv*. *Available at https://arxiv.org/abs/1302.1733*.

Guerreiro RJ, Hardy J. 2012. Tomm40 association with Alzheimer disease: tales of apoe and linkage disequilibrium. *Archives of Neurology* **69(10)**:1243–1244.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* **46(1)**:389–422.

Halushka MK, Fan J-B, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics* **22(3)**:239–247.

Hasnain MJU, Shoaib M, Qadri S, Afzal B, Anwar T, Abbas SH, Sarwar A, Talha Malik HM, Tariq Pervez M. 2020. Computational analysis of functional single nucleotide polymorphisms associated with slc26a4 gene. *PLOS ONE* **15(1)**:e0225368.

Hemani G, Knott S, Haley C. 2013. An evolutionary perspective on epistasis and the missing heritability. *PLOS Genetics* **9(2)**:e1003295.

Hira ZM, Gillies DF. 2015. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics* **2015(5)**:1–13 DOI 10.1155/2015/198363.

Ho Thanh Lam L, Le NH, Van Tuan L, Tran Ban H, Nguyen Khanh Hung T, Nguyen NTK, Huu Dang L, Le NQK. 2020. Machine learning model for identifying antioxidant proteins using features calculated from primary sequences. *Biology* **9(10)**:325 DOI 10.3390/biology9100325.

Hormozdiari F, Penn O, Borenstein E, Eichler EE. 2015. The discovery of integrated gene networks for autism and related disorders. *Genome Research* **25(1)**:142–154 DOI 10.1101/gr.178855.114.

How BC, Narayanan K. 2004. An empirical study of feature selection for text categorization based on term weightage. In: *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*. Piscataway: IEEE, 599–602.

Hussain F, Saeed U, Muhammad G, Islam N, Sheikh GS. 2019. Classifying cancer patients based on DNA sequences using machine learning. *Journal of Medical Imaging and Health Informatics* **9(3)**:436–443 DOI 10.1166/jmihi.2019.2602.

Hwang M-L, Lin Y-D, Chuang L-Y, Yang C-H. 2014. Determination of the SNP–SNP interaction between breast cancer related genes to analyze the disease susceptibility. *International Journal of Machine Learning and Computing* **4(5)**:468–473 DOI 10.7763/IJMLC.2014.V4.456.

Isik AT. 2010. Late onset Alzheimer's disease in older people. *Clinical Interventions in Aging* **5**:307 DOI 10.2147/CIA.

Ismaeel AG, Ablahad AA. 2013. Novel method for mutational disease prediction using bioinformatics techniques and backpropagation algorithm. *arXiv*. *Available at https://arxiv.org/abs/1303.0539*.

Jain I, Jain VK, Jain R. 2018. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing* **62(16)**:203–215 DOI 10.1016/j.asoc.2017.09.038.

Jiao Y, Chen R, Ke X, Cheng L, Chu K, Lu Z, Herskovits EH. 2012. Single nucleotide polymorphisms predict symptom severity of autism spectrum disorder. *Journal of Autism and Developmental Disorders* **42(6)**:971–983 DOI 10.1007/s10803-011-1327-5.

Joachims T. 1998. Text categorization with support vector machines: learning with many relevant features. In: *European Conference on Machine Learning*. Berlin: Springer, 137–142.

Joshi S, Shenoy D, Simha GV, Rrashmi P, Venugopal K, Patnaik L. 2010. Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods. In: *2010 Second International Conference on Machine Learning and Computing*. Piscataway: IEEE, 218–222.

Karthik S, Sudha M. 2018. A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases. *International Journal of Engineering and Advanced Technology* **8(2)**:182–191.

Khalid S, Khalil T, Nasreen S. 2014. A survey of feature selection and feature extraction techniques in machine learning. In: *2014 Science and Information Conference*. Piscataway: IEEE, 372–378.

Khodatars M, Shoeibi A, Ghassemi N, Jafari M, Khadem A, Sadeghi D, Moridian P, Hussain S, Alizadehsani R, Zare A, Khosravi A, Nahavandi S, Rajendra Acharya U, Berk M. 2020. Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review. *arXiv*. *Available at https://arxiv.org/abs/2007.01285*.

Kim J, Sohn I, Kim DDH, Jung S-H. 2013. Snp selection in genome-wide association studies via penalized support vector machine with max test. *Computational and Mathematical Methods in Medicine* **2013(2)**:1–8 DOI 10.1155/2013/340678.

Kong W, Mou X, Yang B. 2009. Study DNA microarray gene expression data of alzheimer's disease by independent component analysis. In: *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*. Piscataway: IEEE, 44–47.

Korani W, Clevenger JP, Chu Y, Ozias-Akins P. 2019. Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. *The Plant Genome* **12(1)**:180023 DOI 10.3835/plantgenome2018.05.0023.

Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG. 2016. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience* **19(11)**:1454–1462 DOI 10.1038/nn.4353.

Kursa MB, Rudnicki WR. 2010. Feature selection with the boruta package. *Journal of Statistical Software* **36(11)**:1–13 DOI 10.18637/jss.v036.i11.

Latkowski T, Osowski S. 2015. Data mining for feature selection in gene expression autism data. *Expert Systems with Applications* **42(2)**:864–872 DOI 10.1016/j.eswa.2014.08.043.

Le NQK, Ho Q-T, Yapp EKY, Ou Y-Y, Yeh H-Y. 2020. DeepETC: a deep convolutional neural network architecture for investigating and classifying electron transport chain's complexes. *Neurocomputing* **375(1)**:71–79 DOI 10.1016/j.neucom.2019.09.070.

Le NQK, Nguyen V-N. 2019. SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data. *PeerJ Computer Science* **5(17)**:e177 DOI 10.7717/peerj-cs.177.

Liang Y, Kelemen A. 2008. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys* **2**:43–60 DOI 10.1214/07-SS026.

Liu Q, Yang J, Chen Z, Yang MQ, Sung AH, Huang X. 2008. Supervised learning-based tagSNP selection for genome-wide disease classifications. *BMC Genomics* **9(1)**:1–9 DOI 10.1186/1471-2164-9-S1-S6.

Mao X, Young BD, Lu Y-J. 2007. The application of single nucleotide polymorphism microarrays in cancer research. *Current Genomics* **8(4)**:219–228 DOI 10.2174/138920207781386924.

Mathur M. 2018. Bioinformatics challenges: a review. *International Journal of Advanced Scientific Research* **3(6)**:29–33.

Mezlini AM, Goldenberg A. 2017. Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases. *PLOS Computational Biology* **13(10)**:e1005580 DOI 10.1371/journal.pcbi.1005580.

Mikhail MN, Sayed AY, Mabrouk MS, Eldeib AM. 2020. Investigation of genome-wide association SNPs and Alzheimer's disease. *American Journal of Biomedical Engineering* **10(1)**:1–8.

Mishra R, Li B. 2020. The application of artificial intelligence in the genetic study of Alzheimer's disease. *Aging and Disease* **11(6)**:1567 DOI 10.14336/AD.2020.0312.

Mount DW, Pandey R. 2005. Using bioinformatics and genome analysis for new therapeutic interventions. *Molecular Cancer Therapeutics* **4(10)**:1636–1643 DOI 10.1158/1535-7163.MCT-05-0150.

Nakka P, Raphael BJ, Ramachandran S. 2016. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics* **204(2)**:783–798.

Narayanan BN, De Silva MS, Hardie RC, Kueterman NK, Ali R. 2019. Understanding deep neural network predictions for medical imaging applications. *arXiv*. *Available at https://arxiv.org/abs/1912.09621*.

Narayanan BN, Hardie RC, Kebede TM. 2018. Performance analysis of feature selection techniques for support vector machine and its application for lung nodule detection. In: *NAECON 2018—IEEE National Aerospace and Electronics Conference*. Piscataway: IEEE, 262–266.

Ng PC, Henikoff S. 2003. Sift: predicting amino acid changes that affect protein function. *Nucleic acids research* **31(13)**:3812–3814.

Parikshak NN, Gandal MJ, Geschwind DH. 2015. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics* **16(8)**:441–458.

Park C, Kim J, Kim J, Park S. 2018. Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles. *PLOS ONE* **13(7)**:e0201056.

Pereira RB, Plastino A, Zadrozny B, Merschmann LH. 2018. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review* **49(1)**:57–78 DOI 10.1007/s10462-016-9516-4.

Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z, Vorstman JAS, Thompson A, Regan R, Pilorge M, Pellecchia G, Pagnamenta AT, Oliveira B, Marshall CR, Magalhaes TR, Lowe JK, Howe JL, Griswold AJ, Gilbert J, Duketis E, Dombroski BA, De Jonge MV, Cuccaro M, Crawford EL, Correia CT, Conroy J, Conceição IC, Chiocchetti AG, Casey JP, Cai G, Cabrol C, Bolshakova N, Bacchelli E, Anney R, Gallinger S, Cotterchio M, Casey G, Zwaigenbaum L, Wittemeyer K, Wing K, Wallace S, van Engeland H, Tryfon A, Thomson S, Soorya L, Rogé B, Roberts W, Poustka F, Mouga S, Minshew N, McInnes LAlison, McGrew SG, Lord C, Leboyer M,

Le Couteur AS, Kolevzon A, Jiménez GP, Jacob S, Holt R, Guter S, Green J, Green A, Gillberg C, Fernandez BA, Duque F, Delorme R, Dawson G, Chaste P, Café Cátia, Brennan S, Bourgeron T, Bolton PF, Bölte S, Bernier R, Baird G, Bailey AJ, Anagnostou E, Almeida J, Wijsman EM, Vieland VJ, Vicente AM, Schellenberg GD, Pericak-Vance M, Paterson AD, Parr JR, Oliveira G, Nurnberger JI, Monaco AP, Maestrini E, Klauck SM, Hakonarson H, Haines JL, Geschwind DH, Freitag CM, Folstein SE, Ennis S, Coon H, Battaglia A, Szatmari P, Sutcliffe JS, Hallmayer J, Gill M, Cook EM, Buxbaum JD, Devlin B, Gallagher L, Betancur C, Scherer SW. 2014. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics* **94(5)**:677–694.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24(3)**:142–149 DOI 10.1016/j.tig.2007.12.006.

Prince M. 2017. Progress on dementia—leaving no one behind. *The Lancet* **390(10113)**:e51–e53 DOI 10.1016/S0140-6736(17)31757-9.

Printy BP, Verma N, Cowperthwaite MC, Markey MK. 2014. Effects of genetic variation on the dynamics of neurodegeneration in Alzheimer's disease. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Piscataway: IEEE, 2464–2467.

Rahit K, Tarailo-Graovac M. 2020. Genetic modifiers and rare mendelian disease. *Genes* **11(3)**:239 DOI 10.3390/genes11030239.

Raj S, Masood S. 2020. Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science* **167**:994–1004.

Raza K. 2012. Application of data mining in bioinformatics. *arXiv. Available at http://arxiv.org/abs/1205.1125*.

Romero-Rosales B-L, Tamez-Pena J-G, Nicolini H, Moreno-Treviño M-G, Trevino V. 2020. Improving predictive models for Alzheimer's disease using gwas data by incorporating misclassified samples modeling. *PLOS ONE* **15(4)**:e0232103.

Ruiz A, Heilmann S, Becker T, Hernández I, Wagner H, Thelen M, Mauleon A, Rosende-Roca M, Bellenguez C, Bis J, Harold D, Gerrish A, Sims R, Sotolongo-Grau O, Espinosa A, Alegret M, Arrieta JL, Lacour A, Leber M, Becker J, Lafuente A, Ruiz S, Vargas L, Rodríguez O, Ortega G, Dominguez M-A, Mayeux R, Haines JL, Pericak-Vance MA, Farrer LA, Schellenberg GD, Chouraki V, Launer LJ, Van Duijn C, Seshadri S, Antúnez C, Breteler MM, Serrano-Ríos M, Jessen F, Tárraga L, Nöthen MM, Maier W, Boada M, Ramírez A, IGAP. 2014. Follow-up of loci from the international genomics of Alzheimer's disease project identifies trip4 as a novel susceptibility gene. *Translational Psychiatry* **4(2)**:e358.

Rémi C, Gerardin E, Jérôme T, Auzias G, Séphane L, Habert M-O, Chupin M, Benali H, Colliot O. 2011. Automatic classification of patients with Alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *Neuroimage* **56(2)**:766–781.

Saeys Y, Inza I, Larranaga P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23(19)**:2507–2517.

Sandraa B, Mathieua R, Maguelonnec T, Pascala P, Ronzad AR, Jean-Micheld V, Ginad D. 2010. Discovering novelty in sequential patterns: application for analysis of microarray data on alzheimer disease. *Studies in Health Technology and Informatics* **160**:1314–1318.

Sa'id AA, Rustam Z, Wibowo VVP, Setiawan QS, Laeli AR. 2020. Linear support vector machine and logistic regression for cerebral infarction classification. In: *2020 International Conference on Decision Aid Sciences and Application (DASA)*. Piscataway: IEEE, 827–831.

Shahbaz M, Ali S, Guergachi A, Niazi A, Umer A. 2019. Classification of Alzheimer's disease using machine learning techniques. In: *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*. 296–303.

Shaltout N, Moustafa M, Rafea A, Moustafa A, ElHefnawi M. 2015. Comparing PCA to information gain as a feature selection method for influenza: a classification. In: *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. Piscataway: IEEE, 279–283.

Shastry BS. 2007. Snps in disease gene mapping, medicinal drug development and evolution. *Journal of Human Genetics* **52(11)**:871–880.

Sherif FF, Zayed N, Fakhr M. 2015. Discovering Alzheimer genetic biomarkers using bayesian networks. *Advances in Bioinformatics* **2015**:1–8 DOI 10.1155/2015/639367.

Shouman M, Turner T, Stocker R. 2012. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology* **2(3)**:220–223 DOI 10.7763/IJIET.2012.V2.114.

Siavelis JC, Bourdakou MM, Athanasiadis EI, Spyrou GM, Nikita KS. 2016. Bioinformatics methods in drug repurposing for Alzheimer's disease. *Briefings in Bioinformatics* **17(2)**:322–335 DOI 10.1093/bib/bbv048.

Singh RK, Sivabalakrishnan M. 2015. Feature selection of gene expression data for cancer classification: a review. *Procedia Computer Science* **50**:52–57 DOI 10.1016/j.procs.2015.04.060.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S. 2007. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **25(11)**:1251–1255.

Spataro N, Rodrguez JA, Navarro A, Bosch E. 2017. Properties of human disease genes and the role of genes linked to mendelian disorders in complex disease aetiology. *Human Molecular Genetics* **26(3)**:489–500 DOI 10.1093/hmg/ddw405.

Spencer M, Takahashi N, Chakraborty S, Miles J, Shyu C-R. 2018. Heritable genotype contrast mining reveals novel gene associations specific to autism subgroups. *Journal of Biomedical Informatics* **77(1)**:50–61 DOI 10.1016/j.jbi.2017.11.016.

Staples M, Chan L, Si D, Johnson K, Whyte C, Cao R. 2019. Artificial intelligence for bioinformatics: applications in protein folding prediction. In: *2019 IEEE Technology & Engineering Management Conference (TEMSCON)*. Piscataway: IEEE, 1–8.

Tahir M, Sardaraz M. 2020. A fast and scalable workflow for SNPs detection in genome sequences using hadoop map-reduce. *Genes* **11(2)**:166 DOI 10.3390/genes11020166.

Tang EK, Suganthan PN, Yao X. 2006. Gene selection algorithms for microarray data based on least squares support vector machine. *BMC Bioinformatics* **7(1)**:1–16 DOI 10.1186/1471-2105-7-1.

Tejeswinee K, Shomona GJ, Athilakshmi R. 2017. Feature selection techniques for prediction of neuro-degenerative disorders: a case-study with Alzheimer's and Parkinson's disease. *Procedia Computer Science* **115(1)**:188–194 DOI 10.1016/j.procs.2017.09.125.

Teng X, Dong H, Zhou X. 2017. Adaptive feature selection using v-shaped binary particle swarm optimization. *PLOS ONE* **12(3)**:e0173907.

Uppu S, Krishna A, Gopalan RP. 2016. A review on methods for detecting snp interactions in high-dimensional genomic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15(2)**:599–612.

Van Rooij JG, Meeter LH, Melhem S, Nijholt DA, Wong TH, Bank NB, Rozemuller A, Uitterlinden AG, Van Meurs JG, Van Swieten JC. 2019. Hippocampal transcriptome profiling

combined with protein–protein interaction analysis elucidates Alzheimer's disease pathways and genes. *Neurobiology of Aging* **74**:225–233.

Wang Q, Lu Q, Zhao H. 2015. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Frontiers in Genetics* **6**:149.

Wodehouse P. 2006. Bioinformatics and pattern recognition come together. *Journal of Pattern Recognition Research* **1**:37–41.

Xu Y, Cao L, Zhao X, Yao Y, Liu Q, Zhang B, Wang Y, Mao Y, Ma Y, Ma JZ, Payne TJ, Li MD, Li L. 2020. Prediction of smoking behavior from single nucleotide polymorphisms with machine learning approaches. *Frontiers in Psychiatry* **11**:416.

Xue B, Zhang M, Browne WN. 2012. Particle swarm optimization for feature selection in classification: a multi-objective approach. *IEEE Transactions on Cybernetics* **43(6)**:1656–1671 DOI 10.1109/TSMCB.2012.2227469.

Yang A, Zhang W, Wang J, Yang K, Han Y, Zhang L. 2020. Review on the application of machine learning algorithms in the sequence data mining of DNA. *Frontiers in Bioengineering and Biotechnology* **8**:1032 DOI 10.3389/fbioe.2020.01032.

Yazdani H. 2020. Bounded fuzzy possibilistic method. *Fuzzy Sets and Systems* **389(6)**:51–65 DOI 10.1016/j.fss.2019.07.011.

Yazdani H, Cheng LL, Christiani DC, Yazdani A. 2018. Bounded fuzzy possibilistic method reveals information about lung cancer through analysis of metabolomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **17(2)**:526–535 DOI 10.1109/TCBB.2018.2869757.

Yazdani A, Mendez-Giraldez R, Yazdani A, Kosorok MR, Roussos P. 2020. Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network. *BMC Bioinformatics* **21(1)**:1–19.

Yin Z, Lan H, Tan G, Lu M, Vasilakos AV, Liu W. 2017. Computing platforms for big biological data analytics: perspectives and challenges. *Computational and Structural Biotechnology Journal* **15(7)**:403–411 DOI 10.1016/j.csbj.2017.07.004.

Yokoyama JS, Bonham LW, Sears RL, Klein E, Karydas A, Kramer JH, Miller BL, Coppola G. 2015. Decision tree analysis of genetic risk for clinically heterogeneous Alzheimer's disease. *BMC Neurology* **15(1)**:1–11.

Zafeiris D, Rutella S, Ball GR. 2018. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. *Computational and Structural Biotechnology Journal* **16(1)**:77–87 DOI 10.1016/j.csbj.2018.02.001.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* **109(4)**:1193–1198 DOI 10.1073/pnas.1119675109.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* **111(4)**:E455–E464 DOI 10.1073/pnas.1322563111.