MACHINE LEARNING IN EARLY GENETIC DETECTION OF MULTIPLE SCLEROSIS DISEASE: A SURVEY

Nehal M. Ali[1], Mohamed Shaheen[2], Mai S. Mabrouk[3]
and Mohamed A. AboRezka[1]

[1]College of Computing and Information Technology, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt
[2]College of Computing and Information Technology, Arab Academy for Science Technology and Maritime Transport, Alexandria, Egypt
[3]Department of Biomedical Engineering, Faculty of Computer Science, Misr University for Science and Technology, Cairo, Egypt

ABSTRACT

Multiple sclerosis disease is a main cause of non-traumatic disabilities and one of the most common neurological disorders in young adults over many countries. In this work, we introduce a survey study of the utilization of machine learning methods in Multiple Sclerosis early genetic disease detection methods incorporating Microarray data analysis and Single Nucleotide Polymorphism data analysis and explains in details the machine learning methods used in literature. In addition, this study demonstrates the future trends of Next Generation Sequencing data analysis in disease detection and sample datasets of each genetic detection method was included .in addition, the challenges facing genetic disease detection were elaborated.

KEYWORDS

Multiple sclerosis, Machine learning, Microarray, Single Nucleotide Polymorphism, early disease detection, Next Generation Sequencing.

1. INTRODUCTION

Multiple Sclerosis (MS) is a demyelination disease, that is, the immune system attacks the myelin sheath causing fatal damages to the nerve cells in human Central Nervous System (CNS) and consequently fatal physical disabilities including partial or total blindness, double vision, muscle weakness, motor disabilities in addition to mental, and sometimes psychiatric impacts[1][2].
Myelin sheath is a fatty matter that encloses the axons of nerve cells, (i.e. the conductors of the nervous system). This myelin sheath permits electrical impulses to be transmitted efficiently and rapidly along the nerve cells. If this sheath is damaged, these impulses slow down causing serious physical and psychological impacts[3].

Initially, MS can commence as a Clinically Isolated Syndrome (CIS), where a patient suffers an attack indicative of demyelination, but does not attain the criteria for MS, 30-70% of CIS patients develop MS[4].

MS disease takes assorted forms, relapsing-remitting (RRMS), primary progressive (PPMS) and secondary progressive (SPMS) forms. Relapsing-remitting MS is repetitive separated attacks with

DOI: 10.5121/ijcsit.2020.12501

new symptoms arising with each attack whereas symptoms accumulate over time in progressive MS [5].

The primary difference between PPMS and SPMS is the presence or absence of relapses, that is, as a secondary phase of the disease course, SPMS develops in patients who originally had RRMS. On the other hand, PPMS in patients who have never experience relapses and develops a progressive disease course [6].

Pathophysiologically, the three primary features of MS that interact together causing symptoms arousal are the destruction of myelin sheaths of neurons, inflammation in addition to lesions formation within the CNS; Lesions are known as the scars that form within the CNS in the brain stem, basal ganglia, the white matter of the optic nerve, the spinal cord in addition to white matter adjacent to the lateral ventricles [1][7].

MS is diagnosed between the ages of 20 and 50, ratio of women to men with MS is 2:1. In 2012, MS affected more than 400,000 people in the United States with prevalence rate of149.2 per 100,000 individuals, while 2.1 million people were affected worldwide in 2016 [8]. Moreover, the estimated MS prevalence in the united states was 309.2 per 100,000 with female: male ratio 2.8 in 2019 [9].

In middle east, MS prevalence in Egypt is 1.41% or 14.1 per 1000 in Al Quseir City [10]. the ratio of female/ male MS patients has ranged from 0.8 in Oman to 4.3 in Saudi Arabia, while the disease prevalence of MS has ranged from 14.77/100,000 population in Kuwait (2000) to 101.4/100,000 in Turkey (2006). The overall MS prevalence in the region was 51.52/100,000. The mean age at disease onset ranged from 25.2 years in Kuwait to 32.5 years in Northeastern Iran, with an overall estimate of 28.54 years [11].

To our knowledge, there is no known treatment available for this disease, the treatment protocols are to reduce the impacts after an attack or to prevent more attacks.[12][13] Thus, detection of this disease is substantial to avert its serious impacts.

Several methods were introduced for MS detection including, lesion detection in central nervous system by image analysis, wavelet transform in addition to genetic detection methods. Machine Learning has played a significant role in MS detection by analyzing Magnetic Resonance Imaging (MRI) data, primarily by detecting the existing lesions caused by the MS in the central nervous system [14][15][16][17].

In MS detection, image analysis is mainly used to quantify and detect multiple sclerosis (MS) lesions in the central nervous system. Accordingly, the number and volume of lesions is used to assess the MS disease impact, to determine the disease progression and the treatment to be followed accordingly.

MS detection by analyzing images using machine learning methods can be categorized in terms of the images being analyzed to MRI data[14][15]; Functional Magnetic Resonance Imaging (FMRI)[18]; and Deep Range Imaging (DRI)[19][20].Moreover, some works have focused on using machine learning in MS lesion detection in only grey matter, white matter or cortex of the brain [16][21][22].

Under certain circumstances; it is challenging to detect MS by analyzing different types of MRI images. For instance, due to serious tissue damage, the abnormal white matter areas of the central nervous system appear normal in MRI images [23].

Another detection method for MS that have machine learning methods is analyzing different wavelet transform methods of MRI images such as two dimensional, biorthogonal, and Haar wavelet transform. [24][25][26][27].

Furthermore, some metabolic methods have used machine learning methods, that is, the imbalance of oxidant/antioxidant and inflammatory/anti-inflammatory molecules impacts the demyelination and axonal damage in MS; Thus, some studies have used machine learning methods, namely Support Vector Machine (SVM) and Naïve Bayes classifier in order to determine the MS patients given the plasma levels of tumor necrosis factor (TNF)-α, soluble TNF receptor (sTNFR)1, sTNFR2, advanced oxidation protein products (AOPP), hydroperoxides, adiponectin, total plasma antioxidant capacity using the total radical-trapping antioxidant parameter (TRAP), nitric oxide metabolites, , sulfhydryl (SH) groups, in addition to serum levels of zinc [28][29] [30][31]. Figure 1 summarizes MS disease detection methods that have used machine learning detective models.

Since diagnosis of MS typically depends on radiologically-determined central nervous system (CNS) lesions along with the presentation of nonspecific clinical symptoms, MS diagnosis is usually late [32].

Thus, Genetic detection of this disease can have significant impact on MS early detection and consequently avoiding its severe impacts. This survey introduces how machine learning methods were used in the early genetic detection of this autoimmune disease [33][34].

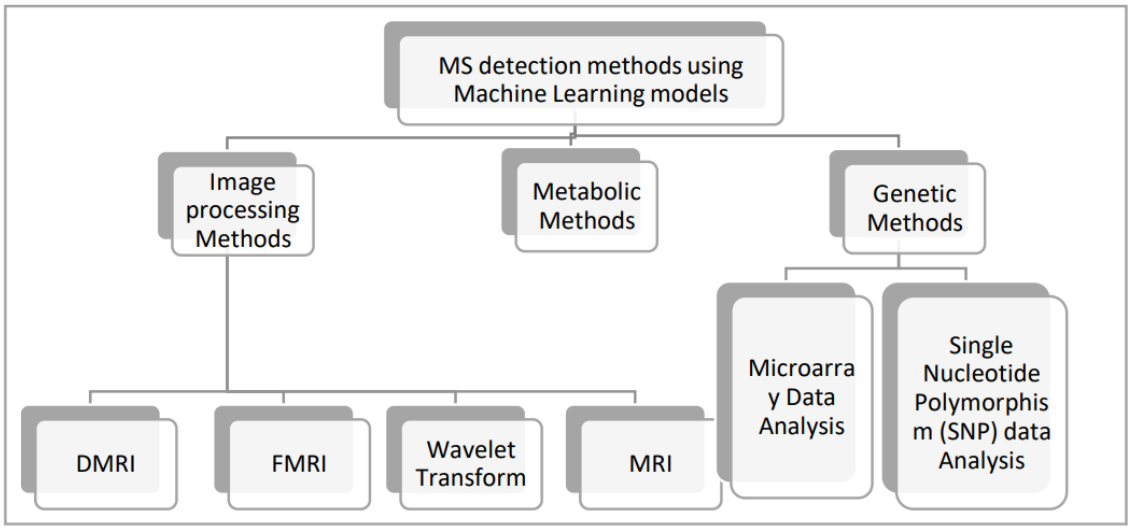

Figure 1: MS disease detection methods that uses machine learning detective models

2. DISEASE GENETIC DETECTION METHODS

2.1. Microarray Data Analysis

Microarray is typically a laboratory tool; fundamentally, a two-dimensional array on a solid substance that is used to examine massive amounts of biological material using multiplexed, high-throughput screening, with parallel processing and detection methods; Microarrays are primarily used for parallel detection of thousands of genes expressions [35][36] [37].

Genetically, gene expression is the most foundational level where the genotype is expressed as a phenotype. That is, the genetic information stored in DNA denotes the genotype, while the phenotype is denoted by the interpretation of that information (i.e. recognizable trait) [38][39]. Some studies have utilized supervised machine learning methods, primarily SVM and Random Forest,as predictive models given the gene expression resulting from microarrays.

Those predictive models have been utilized in studying the conversion of CIS patients into clinically definite MS (CDMS) patients by studying gene expression from CIS patients at time of diagnosis and 1year after and accordingly determine gene expressions that can be used in MS patients' discrimination or in order to classify the MS patients from healthy controls [15][40][41][42].

Furthermore, some studies have utilized the gene expressions resulting from microarrays in order to obtain genetic signatures related to MS diagnosis. That is, these studies' findings were in the direction of determining the defective pathways such as viral or bacterial infections that impacts the MS development [33][43].

The Procedures of MS detection by analyzing microarray gene expression can be formalized in figure 2. Given an obtained gene expression, preprocessing steps are then applied; these preprocessing steps can primarily include data validation and data normalization. Data validation can be obtained by assistant of practitioners or by validation against another dataset, that is, a dataset of gene expressions can be validated against an MRI dataset for the same cases and controls. Afterwards, the gene expression data values are mathematically or by means of tools such as robust multi-array normalization (RMA) normalized using [33][41].

Subsequently, feature extraction is obtained accordingly, by determining the discriminatory power of the genes, common methods in this step are Fisher's ratio and P-value and fold change [44][45].

Afterwards, predictive model is applied accordingly. SVM, Random Forest and Naïve Bayes are commonly used machine learning methods. Lastly, the model accuracy measures such as area under the curve (AUC) is determined and the model findings are subsequently analyzed.

Analyzing the model results denotes the most discriminant genes found by the developed algorithm. Based on this data a correlation tree between the most discriminant genes is created, which elaborates the relationships between the most discriminatory genes in the studied data and helps implying how these genes impacts the gene expression [33]. Moreover, the accuracy and the confusion matrix of machine learning model is obtained accordingly.

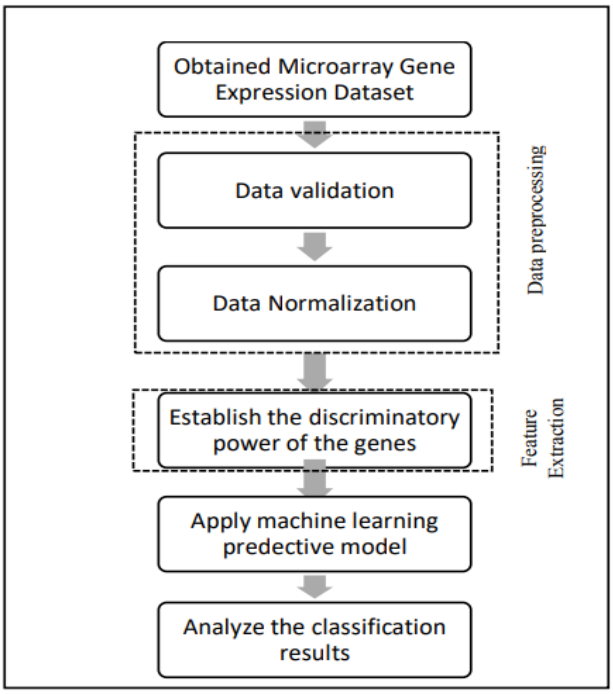

Figure 2: Procedures of MS detection by applying machine learning predictive model on microarray gene expression data

2.2. Single Nucleotide Polymorphism Data Analysis

A Single Nucleotide Polymorphism (SNP) results from the substitution of a single nucleotide of a genome at a specific position that emerges in more than 1% of a population [46][47].

The determination of the SNPs that impacts a given disease is one of the key method that is increasingly used in the recent studies in diseases prediction. Thanks to the significant improvement of computational power, the analysis of Genome Wide Association (GWAS) data has become more feasible [48][49].

Studies have used predictive machine learning models such as Random Forest, SVM and Naïve Bayes in order to analyze genetic data and determine SNPs correlated with MS disease. Studies have primarily investigated parts of human genome, that is, the informative regions that contains the studied SNPs [34][50] [51][52][53][54].

Figure 3 formalizes the general procedures followed in SNP data analysis using the machine learning models. Given an obtained genetic dataset of MS cases and healthy controls, data is first genotyped in all individuals, that is, the SNPs significance to the disease is identified. This step is commonly applied automatically using genotyping tools such as PLINK [55].Patients datasets primarily includes the MS type (the data point label), the patients' treatment protocols, personal information, and some RNA data might be included. A snippet of a simple MS patients' dataset is demonstrated in figure 4 the last five columns denote the SNPs to be studied[56].

Subsequently, and according to the determined SNPs (i.e. features), classification model is executed. This massive data requires high computational power; thus, libraries that implement GPU processing such as Keras are utilized [34].

Parameters tuning of the predictive model has major impact on the classification results, such as the number of epochs used in training phase and the number of levels within an ANN.

Afterwards, the model is tested against a validation dataset, could be a second dataset or a subset of the main dataset that was not utilized in the model training stage.

Finally, the model classification findings are analyzed to determine the impact of the studied SNPs on the disease risk in addition to denoting the SNPs that can prevent the disease [50]. In addition, based on the model predictions, the model prediction accuracy and confusion matrix are obtained to measure [34][52].

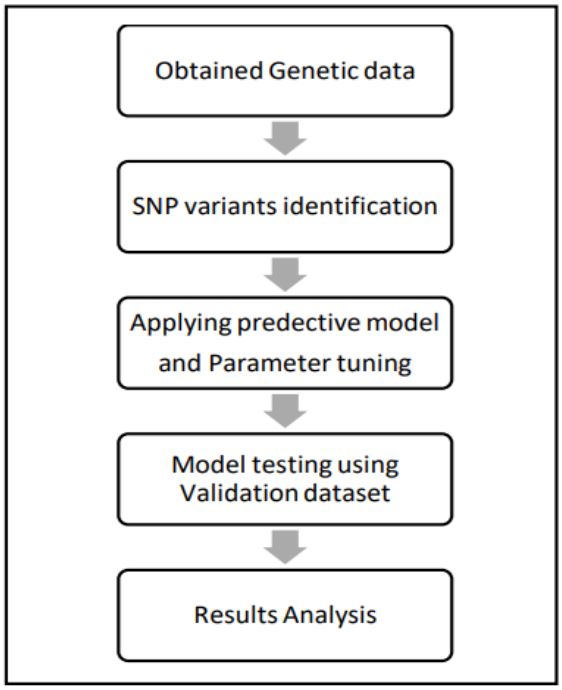

Figure 3: Procedures of MS detection by classifying genetic data using SNPs

					SNPs					
GENDER	AGE	MS TYPE	EDSS	TREATMENT	rs2910164	rs3027898	rs767649	rs57095329	rs2067079	rs1625579
					146a(C/G)	IL-1RK(A/C)	155 (A/T)	146a (A/G)	GAS5 (C/T)	micro 137 (G/T)
1	22	2	3.5	4	CG	AC	TT	AG	CC	TT
1	36	1	2.5	1	GG	AA	AT	AG	CC	GT
1	26	1	1	1	GG	AC	TT	AA	CT	GT
1	23	2	3.5	1	CG	AC	TT	AA	CC	GT
1	23	1	3.5	1	CG	CC	TT	AA	CT	TT
1	34	1	0	1	CG	AA	TT	AG	CC	GT
1	31	1	2	1	CG	AC	AT	AG	CT	TT
1	29	1	2.5	1	CG	AC	AT	AA	CT	TT
1	21	1	1.5	2	GG	AC	TT	AA	CC	TT
1	31	1	3.5	1	GG	CC	AT	AA	CT	GT
1	34	1	2	2	CG	AA	AT	AA	CC	GT
1	28	1	1	1	GG	AC	AT	AA	TT	GT
1	25	2	6.5	3	CG	AC	TT	AG	CC	TT
1	46	1	2.5	1	CC	AA	AT	AA	CC	GT
1	33	1	1.5	1	GG	AA	AT	AA	CC	GT
1	29	2	6	4	CG	AA	AT	AA	CC	GT
2	17	1	2	1	CG	CC	TT	AA	CT	TT
1	29	1	1	2	CG	AC	TT	AG	CT	TT
1	53	2	6	1	GG	AC	AT	AA	CT	GG
1	31	1	2	2	CG	AC	AT	AA	TT	GT

Figure 4: snippet of MS patients' raw dataset [56]

3. MACHINE LEARNING METHODS USED IN EARLY MS DETECTION

Due to the significant development of the computational power, the application of more advanced supervised machine learning methods to analyze the numerous genetic datasets became more feasible. As the conventional statistical methods can only indicate the main impacts of genetic variants on risk for disease, supervised machine learning methods are primarily suited to emerge higher order and non-linear impacts. According to our literature review, methods such as SVM, Random Forest and Naïve Bayes are the main machine learning methods used with MS genetic detection methods[33] [52][57][58][59].

3.1. Support Vector Machine (SVM)

Given a priori labeled dataset that has linearly separable data points, SVM models attempt to determine a discriminant function (i.e. a hyperplane) that can classify the given dataset such that, the distance between the data points (i.e. support vectors) on both of the hyperplane sides and the hyperplane is maximized [60][61].

For non-linearly separable data, kernel functions are utilized to transform the data into a higher dimension and then decision boundaries are determined accordingly. Radial Basis Function (RBF), Sigmoid and Polynomial are commonly used kernel types. Figure 5 formulates the flow chart of SVM algorithm using kernels. As shown, the two main parameters to be provided to the algorithm are C and Ȯ, C is the parameter that regulates the tradeoff between obtaining a low training error and a low testing error. While Ȯ determines the kernel function being applied; After training and testing the model the model performance and accuracy is obtained accordingly [62].

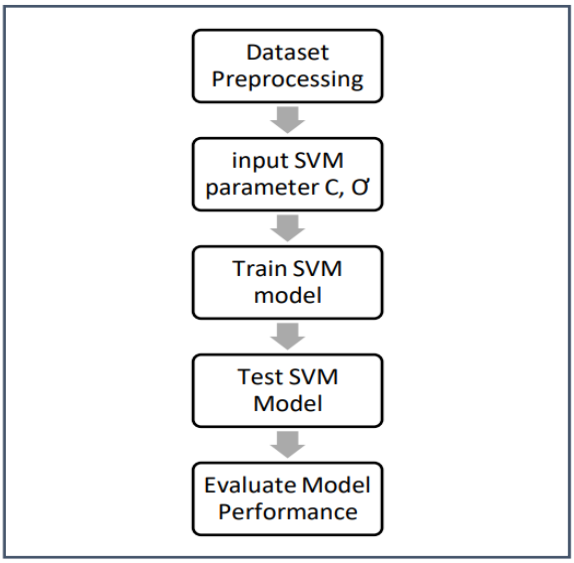

Figure 5: SVM processing steps

3.2. Decision Tree

Decision Tree is primarily a predictive modelling approach that is mainly used in data mining, machine learning and statistics. The main purpose of decision trees is to obtain a target value (tree leaves) given a set of observations (tree branches). Thus, in a classification problem, the obtained classes are given by the tree leaves.[63]

Algorithms for constructing decision trees applies top-down approach, by determining a variable at each tree level that best splits the given dataset. Information gain, Gini impurity and variance reduction are the most popular decision metrics that are used with decision trees. [64][65]

3.3. Random Forest

Random forest is an supervised method to solve classification and regression problems. by constructing a multitude of decision trees during the model training phase and obtains the class by obtaining the mode of the classes of the those individual decision trees [66][67]. Figure 6 summarizes the pseudocode of random forest classification algorithm.

For each data row in the dataset

1. Randomly select i features from m features, where i<<m
2. Given the i features, obtain a node using the split function.
3. Split the root node into leave nodes using the split function.
4. Repeat steps 1 to 3 until "max" number of nodes has been reached.
5. Repeat steps 1 to 4 for "n" times to construct a forest of ntrees.
6. Obtain the prediction for each tree in n
7. calculate the mode of the classification decision of the n trees
8. Obtain classification decision from step 7

Figure 6: Random Forest Algorithm

3.4. Naïve Bayes Classifier

Naïve Bayes classifier is a probabilistic classifier, primarily based on applying Bayes' theorem assuming that the given features are strongly independent. Hence, it requires high dimensional data [68].

Naive Bayes involves building the classification model rapidly as well as obtaining faster predictions. In addition, it requires a number of parameters linear to the number of features in a learning problem, that is, it is highly scalable [69] Figure 7 shows Naïve Bayes algorithm Pseudo code .

1. Read the input dataset S
2. For each class, Obtain the mean and the standard deviation of the predictor features
3. Repeat
    1. use the gauss density equation to calculate the probability of f_i
4. until the probability is calculated for all features (f_1,f_2,f_3…f_n)
5. Repeat steps 1 to 4 for "n" times to construct a forest of ntrees.
6. Calculate the likelihood for each class
7. Obtain classification decision from step 6 based on highest likelihood

Figure 7: Naive Bayes Algorithm Pseudocode

Logistic regression is a supervised machine learning algorithm that is based on statistical concepts. It is primarily used to classify a set of observations. Some of the examples of classification problems are online fraud transactions detection, tumor classification and email

spamming. Logistic regression obtains its output (i.e. the dependent variable) as a probability value by using the logistic sigmoid function. [74]
The use of sigmoid function in machine learning is mapping predictions to probabilities. In other words, mapping a real value into a value between 0 and 1. Machine learning method determination.

There are three main types of Logistic regression as follows:[75][76]

3.4.1. Binomial Logistic Regression

In this type of classification, the dependent variable has only two possible types either 1 and 0 (i.e. binary classification). For example, the value of a dependent variable could be represented as yes or no, success or failure, yes or no, case or control... etc. [75][77]

3.4.2. Multinomial Logistic Regression

In this type of classification, the dependent variable could have 3 or more possible types, such that these types do not quantitative significance, that is, these types are unordered. For example, these variables could represent “Type A” or “Type B” or “Type C”. [15]

3.4.3. Ordinallogistic Regression

In this type of classification, dependent variable can have 3 or more types, such that these types have a quantitative significance. That is, these types are unordered. For example, these variables may represent “strongly disagree”, “disagree”, “neutral”, “agree” and “strongly agree”. [15]

3.5. Artificial Neural Networks

Artificial neural network (ANN) is an artificial intelligence method that is meant to imitate the human brain functioning. ANN is consisted of nodes (i.e. neurons) that are Processing units of this network, neurons within an ANN are organized in layers, primarily input, output and hidden layers that are connected with edges. The inputs are what the ANN learns from (i.e. the data points of the studied dataset) to obtain the output according to a predefined activation function. [70][71]. An activation function of a node determines the output obtained by that node given an input or set of inputs; Sigmoid Linear Unit (SLI), Gaussian, Inverse square root unit (ISRU) and Rectified linear unit (RLU) are some types of activation functions that are commonly used.[72][73]

This activation function has to be determined in addition to the number of network layers and number of nodes in each layer on building the model. Moreover, after applying the needed preprocessing steps, the dataset has to be split into training and validation dataset, like other machine learning techniques, this validation part is used for testing the model after completing the training phase. Figure 8 summarizes the steps of building a model with an ANN.

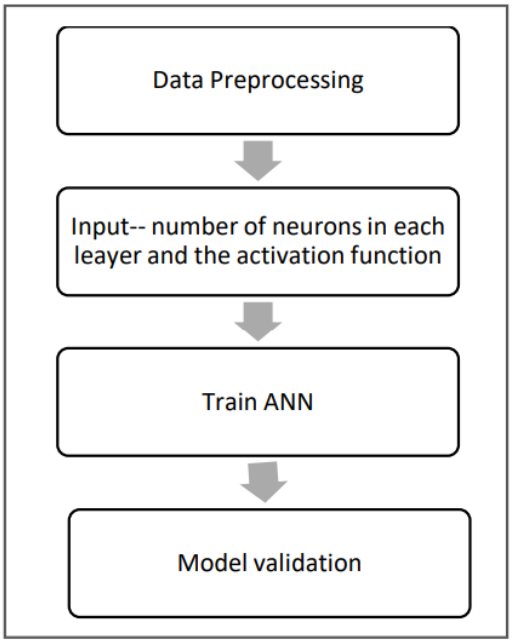

Figure 8: Artificial Neural Networks processing steps

4. MACHINE LEARNING METHOD DETERMINATION

Based on literature, there is no ultimately superior algorithm for solving a specific problem, depending on the problem being solved and the studied dataset, the algorithm to be used should be determined. For instance, when the number of the studied features is greater than the number of samples being fed to the model, the SVM does not perform well and it requires more work on feature engineering than the required for an ANN; Also, ANN can handle multi-class problems by producing probabilities for each class, while SVM solve this type of problems as multiple two-class problems. On the other hand, SVMs are better than ANNs in certain aspects, that is, ANNs are more probable to becoming trapped in local minima, in addition, while most machine learning algorithms can overfit given low number of training samples, ANNs can also overfit if training dataset were too large which is a drawback that SVMs do not have [61][70].

Moreover, Random forest works with both categorical and numerical data, on the other hand, SVM is highly dependent on the distance between the data points, as it maximizes the hyperplane margin, consequently, categorical data has to be encoded before applying SVM. In addition, for multiclass problems, random forest Random Forest is fundamentally proper, while SVM primarily performs better with two-class problems. That is, on using SVM with multiclass problem, the problem needs to be divided into several binary classification problems. On the other hand, random forest models tend to overfit more than SVM, so trees pruning and tuning is required to overcome this. Furthermore, SVM models generally perform better than random forest on sparse data. [18]

Furthermore, the primary advantages of Naïve Bayes algorithm are that, it performs well with small training datasets and independent predictors, it does not require pruning, and does not easily overfit. On the other hand, the main limitation of the Naïve Bayes is assuming that the data attributes are independent, this assumption is not very applicable on realistic datasets, so if this assumption is not valid, Naïve Bayes can be outperformed by other classification algorithm, whilst it can give very distinctive results if this assumption was valid.[78][79]

Finally, Decision trees' main advantage is its simplicity of implementation and understanding in addition to the ability to work on categorical and numerical data, whilst the main disadvantage is that this model is not robust enough and can be easily impacted with a minor change in the dataset. On the contrary of the Logistic regression algorithm that is quite robust with such changes and performs well with large datasets. [75] Table 1 summarizes the pros and cons of the different machine learning algorithms discussed in this paper.

Table 1: Pros and Cons of different machine learning algorithms

Algorithm	Pros	Cons
SVM	• Performs well in Higher dimension datasets • Outperforms other algorithms on separable datasets • Performs well on binary class problems • Performs well on sparse data	• Does not perform well with high number of features and small dataset or on overlapped classes • Categorical data has to be encoded • Selecting the appropriate kernel can be challenging
Decision Tree	• Easy to understand and visualize • Not impacted by missing data • Does not require data normalization	• Not robust, sensitive to data changes • Can be drawn into overfitting • Requires large training set
Random Forest	• Performs well with high-dimensional data • Performs well with on Imbalanced datasets • Not impacted by missing data • Can be applied on both categorical and numerical data • Performs well on multiclass problems	• The trees' output has to be uncorrelated • Needs features that have significance weight on the prediction results • functions as a black box • poor performance on sparse data
Naïve Bayes	• Fast performance • Robustness with irrelevant features • Performs well with small datasets	• Feature independence is not applicable in most of datasets • Can be easily outperformed by other models
Logistic Regression	• Easy to implement • No hyperparameter tuning needed • Can be applied for both classification and regression problems	• Does not perform well with highly correlated features. • Not robust with non-linear data. • Dependent on data representation
ANN	• Performs well with nonlinear and high dimensional data • Can be applied for both classification and regression problems • Does not require feature engineering	• Hardware dependent • Functions as a black box • Requires a lot of parameters tuning • can overfit if training dataset were too large

5. FUTURE TRENDS

Analyzing Next Generation Sequences (NGS) data is a future trend in disease detection thanks to the significantly developed GPU development platforms, analysis of such numerous amount of data became feasible. NGS data files provide massive DNA encoded sequences. These sequences are series of 4 basis proteins (Adenine, Guanine, Cytosine, and Thymine - A, G, C and T respectively-). Analyzing NGS files involves analyzing these massive data sequences to reveal disease biomarker sequences. Analyzing NGS files, is quite challenging due to the size of the file being processed, the size of one file exceeds 10-15GB, and there is considerable data noise [80].
One proposed method is to process these sequences as text in order to determine the disease biomarker sequences with promising results. This method requires considerable data preprocessing phase incorporating quality check to the studied files in addition to noise reduction by performing data trimming based on the performed quality check. Afterwards, K-mer count method is used for feature extraction, then applying a machine learning method for classification and consequently evaluate the model, and analyze the results to identify the sequences of a disease and accordingly define genetic biomarkers of a disease. Primarily, K-mers are the unique subsequences out of a length k sequence, K-mer counting is implemented as well to determine the count of each segmented k-mer, this obtains the "K-mer counts matrix" of M Sequences × 4k (i.e. given the 4 proteins A,C,T and G) [81]. Table 2 demonstrates some studies that have analyzed NGS data on different diseases.

Table 2: sample literature of NGS data analysis

Ref	Work Proposed	Pros	Cons
[82]	The research team have developed a software that works on high specs servers over a dataset of bacterial isolates, this software is based on K-mer frequency feature extraction method and then applying regression analysis. The proposed work aimed to-predict a phenotype from a sequencing data of a bacterial isolate. The method was validated on 167 Klebsiella pneumoniae isolates, 200 Pseudomonas aeruginosa isolates, and 459 Clostridium difficile isolates. The results have shown that the proposed model have reported accuracy of 88%	This work has reported that model building on a given dataset requires 3 to 5 hours per phenotype, whilst the phenotype prediction required less than one second on assembled genomes.	The used dataset was too specific to the proposed model, authors mentioned that more general dataset shall be used in their future work.
[83]	A model for detection of Colorectal Cancer (CRC) from metagenomics data - organism taxonomic units (OUT). The model has used a hybrid method of Taxonomy method and NLP methods for feature extraction and then SVM and Random Forest machine learning models were applied accordingly. The study has combined both Taxonomy and NLP methods, and compared the resulting accuracy of the proposed hybrid model with NLP method singularly and Taxonomy method singularly and showed no significant accuracy improvement.	The authors have reported that os using 149 features of the taxonomy table, random forest algorithm reported accuracy of 72%, while SVM algorithm has reported 74 %.whilst training all 256 features using 4-length k-mers, attained accuracy of 74 % using SVM and 65% using random forest.	The introduced hybrid approach did not outperform the K-mer approach nor the Taxonomy approach.
[84]	The authors have introduced a stand-alone application for retrieving and processing sets of NGS data. In order speed up file transfers, authors have used the *Asperahighspeed* file transfer protocol. FASTQC software was used to evaluate the quality of raw sequence data, while Trimmomatic tool was	The paper have introduced a promising NGS files processing pipeline	Further dataset is needed for more proof of concept verification

Ref	Work Proposed	Pros	Cons
	used for data trimming and alignment[85]. Results have shown that Octopus-toolkit can deliver the result faster than Galaxy or GenePattern products provided that a computer that is capable of NGS analysis is used.		

6. DATASETS

Samples of the literature datasets for Ms genetic detection techniques of the literatures are shown in Table 3 in addition to sample datasets of NGS datasets.

Table 3: samples of literature datasets

Reference	Dataset Description	Dataset Type
(33)	Gene expression profiles dataset obtained from naive CD4+ T cells consists of 54675 probes and 113 samples: 73 MS cases and 40 controls by the European Bioinformatic Institute E-GEOD 13732 microarray – used as training dataset. gene expression profiles dataset obtained from t cells of 20 samples: 10 ms cases, 10 controls and 54675 probes by the european bioinformatic institute e-geod-43592 microarray) – used as validation dataset	microarray expressions
(40)	Gene expression profiles dataset of 26 multiple sclerosis cases and 18 controls. Obtained from the transcriptome of peripheral blood mononuclear cells.	microarray expressions
(41)	microarrays study gene expression in naïve $CD4^{+}$T cells from 37 CIS patients	microarray expressions
(50)	Samples of 401 MS cases and 390 controls ACE (rs4359 and rs1799752), EVI5 (rs6680578, rs10735781 and rs11810217), CBLB (rs12487066) and VEGFA (rs3025039 and rs2071559), MALAT1 (rs619586 and rs3200401), ANRIL (rs1333045, rs1333048, rs4977574 and rs10757278), NINJ2 (rs11833579 and rs3809263), GRM7 (rs6782011 and rs779867), VLA4 (rs1143676), GAS5 (rs2067079 and rs6790), H19 (rs2839698 and rs217727),	SNP
(51)	DNA samples from 191 MS patientsand 25,482 SNPS consented via the Pennsylvania State University.	SNP
(52)	MS case-control study conductedby the International Multiple Sclerosis Genetics Consortium; consists of 931 MS cases and 2,431 controls (n = 3,362). of genotypes for a total of 325,807 SNPs	SNP
(76)	16S rRNA sequence files obtained by Bioproject: PRJNA280026 by The European Nucleotide Archive (ENA).	NGS
(79)	Data simulated with ART sequencing read simulator	NGS-Simulating Data

7. CHALLENGES

Disease detection by analyzing genetic data is a relatively recent field, and it faces many challenges on several aspects; primarily, the lack of biological knowledge for computer science researchers to understand the studied data and formulating it into a predictive model accordingly, that is, data understanding and gaining the required knowledge about the studied disease, consumes long time which does not encourage the researchers to participate in such field. In

addition, the studied datasets are relatively big, which requires high computational power, also, providing a verification dataset for some models is quite challenging.

For the studies that uses microarray datasets, the main challenge is to determine a set of discriminative genes associated with MS disease from a numerous set of genes given by the microarray expression. From the prospective of predicting MS disease methods based on SNP genotypes, the challenge is the precise genotypic analysis and determining the SNPs that directly impact the disease diagnosis [42][53].

Furthermore, studying NGS data is very challenging considering the sophisticated hardware specifications required for the data analysis; super computers are primarily required which is not easy to avail in most of the labs and universities. In addition, as whole genome sequences are considered under the big data umbrella, learning the analytics tools is an additional prerequisite for researchers in order to build their models properly. Moreover, significant data preprocessing is required for NGS data, for noise reduction and feature selection [86].

8. TOOLS

To our knowledge, many tools are commonly used in disease genetic analysis, including PLINK for genotypic analysis, which facilitate the genotypic analysis using wide genotypic analysis functions and capabilities [87].

For the supervised model development, libraries such as Keras and Tensor flow were launched in 2015 and have been applied in the recent studies, they provide support to various machine learning methods, models evaluation and, most importantly, it supports developing on GPU, which provides computational power that motivates studying and handling bigger amount of data and obtaining more sophisticated models [34][88][89].In addition, for sequence data, Trimmomatic and Fsatqc are commonly preprocessing tools[90] [85]; Moreover, Pubmed provides useful genetic tools that can support the researchers in data handling, nevertheless, needs lots of time and effort for non-biological experts to use it accordingly [91].

9. CONCLUSIONS

Multiple Sclerosis (MS) is an autoimmune disease that induces significant impact to human CNS and consequently causes severe physical disabilities in addition to psychological impacts. This study has elaborated in details the genetic approaches that use machine learning methods for early MS detection including microarray data analysis and SNP data analysis in addition to the future trends in genetic disease detection. Which excels the traditional detection methods of image processing methods and metabolic data analysis in disease early detection. In addition, machine learning methods were detailed, concluding that, there is no machine learning method can be defined as the best method, and that, the determination of the machine learning method depends on the type of data being studied. Moreover, this work has elaborated the challenges of the genetic detection methods of this disease including the size of the data, the required computational power and hardware specifications in addition to the needed biological background.

REFERENCES

[1] K. Berer and G. Krishnamoorthy, “Microbial view of central nervous system autoimmunity,” FEBS Letters, vol. 588, no. 22. Elsevier, pp. 4207–4213, 17-Nov-2014.

[2] M. Naghavi et al., "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013," Lancet, vol. 385, no. 9963, pp. 117–171, Jan. 2015.
[3] J. L. Salzer and B. Zalc, "Myelination," Current Biology, vol. 26, no. 20. Cell Press, pp. R971–R975, 24-Oct-2016.
[4] R. M. Van Der Vuurst De Vries et al., "Application of the 2017 Revised McDonald Criteria for Multiple Sclerosis to Patients with a Typical Clinically Isolated Syndrome," JAMA Neurol., vol. 75, no. 11, pp. 1392–1398, Nov. 2018.
[5] O. Olerup et al., "Primarily chronic progressive and relapsing/remitting multiple sclerosis: Two immunogenetically distinct disease entities," Proc. Natl. Acad. Sci. U. S. A., vol. 86, no. 18, pp. 7113–7117, Sep. 1989.
[6] K. Bashir and J. N. Whitaker, "Clinical and laboratory features of primary progressive and secondary progressive MS," Neurology, vol. 53, no. 4, pp. 765–771, Sep. 1999.
[7] A. Compston and A. Coles, "Multiple sclerosis," The Lancet, vol. 372, no. 9648. Elsevier, pp. 1502–1517, 25-Oct-2008.
[8] P. Dilokthornsakul, R. J. Valuck, K. V Nair, J. R. Corboy, R. R. Allen, and J. D. Campbell, "Multiple sclerosis prevalence in the United States commercially insured population," 2016.
[9] M. T. Wallin et al., "The prevalence of MS in the United States: A population-based estimate using health claims data," Neurology, vol. 92, no. 10, pp. E1029–E1040, Mar. 2019.
[10] N. El-Tallawy, W. M A Farghaly, R. Badry, N. A. Metwally, M. Abd El Hamed, and M. R. Kandil, "Prevalence of multiple sclerosis in al Quseir city, red sea governorate, egypt," 2016.
[11] P. Heydarpour, S. Khoshkish, S. Abtahi, M. Moradi-Lakeh, and M. A. Sahraian, "Multiple Sclerosis Epidemiology in Middle East and North Africa: A Systematic Review and Meta-Analysis," Neuroepidemiology, vol. 44, no. 4, pp. 232–244, 2015.
[12] M. Stangel, I. K. Penner, B. A. Kallmann, C. Lukas, and B. C. Kieseier, "Towards the implementation of 'no evidence of disease activity' in multiple sclerosis treatment: the multiple sclerosis decision model.," Ther. Adv. Neurol. Disord., vol. 8, no. 1, pp. 3–13, Jan. 2015.
[13] A. Zager, "Modulating the immune response with the wake-promoting drug modafinil: a potential therapeutic approach for inflammatory disorders," Brain. Behav. Immun., Apr. 2020.
[14] R. J. Ramteke and K. Monali, "Automatic Medical Image Classification and Abnormality Detection Using K-Nearest Neighbour," 2012.
[15] Y. Zhao et al., "Exploration of machine learning techniques in predicting multiple sclerosis disease course," 2017.
[16] M. J. Fartaria et al., "Automated detection of white matter and cortical lesions in early stages of multiple sclerosis," J. Magn. Reson. Imaging, 2016.
[17] S. H. Wang et al., "Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling," Front. Neurosci., 2018.
[18] V. Saccà et al., "Evaluation of machine learning algorithms performance for the prediction of early multiple sclerosis from resting-state FMRI connectivity data," Brain Imaging Behav., 2019.
[19] C. Cavaliere et al., "Computer-aided diagnosis of multiple sclerosis using a support vector machine and optical coherence tomography features," Sensors (Switzerland), 2019.
[20] A. P. del Palomar et al., "Swept source optical coherence tomography to early detect multiple sclerosis disease. The use of machine learning techniques," PLoS One, 2019.
[21] K. Bendfeldt et al., "Multivariate pattern classification of gray matter pathology in multiple sclerosis," Neuroimage, vol. 60, no. 1, pp. 400–408, Mar. 2012.
[22] M. F. Rachmadi et al., "Limited One-time Sampling Irregularity Map (LOTS-IM) for Automatic Unsupervised Assessment of White Matter Hyperintensities and Multiple Sclerosis Lesions in Structural Brain Magnetic Resonance Images," Comput. Med. Imaging Graph., 2020.
[23] Y. Zhang et al., "Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: Decision tree, k -nearest neighbors, and support vector machine," Simulation, 2016.
[24] M. Torabi, H. Moradzadeh, R. Vaziri, R. D. Ardekani, and E. Fatemizadeh, "Multiple sclerosis diagnosis based on analysis of subbands of 2-D wavelet transform applied on MR-images," in 2007 IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2007, 2007.
[25] D. R. Nayak, R. Dash, and B. Majhi, "Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests," Neurocomputing, 2016.

[26] S. H. Wang et al., "Multiple Sclerosis Detection Based on Biorthogonal Wavelet Transform, RBF Kernel Principal Component Analysis, and Logistic Regression," IEEE Access, 2016.
[27] X. Wu and M. Lopez, "Multiple Sclerosis Slice Identification by Haar Wavelet Transform and Logistic Regression," 2017.
[28] L. Mezzaroba et al., "Antioxidant and Anti-inflammatory Diagnostic Biomarkers in Multiple Sclerosis: A Machine Learning Study," Mol. Neurobiol., 2020.
[29] S. E Fiedler et al., "Analysis of IL-6, IL-1β and TNF-α production in monocytes isolated from multiple sclerosis patients treated with disease modifying drugs," J. Syst. Integr. Neurosci., 2017.
[30] E. Tönnies and E. Trushina, "Oxidative Stress, Synaptic Dysfunction, and Alzheimer's Disease," Journal of Alzheimer's Disease. 2017.
[31] A. M. Witkowska et al., "Serum Levels of Biomarkers of Immune Activation and Associations With Neurological Impairment in Relapsing-Remitting Multiple Sclerosis Patients During Remission," Biol. Res. Nurs., 2016.
[32] S. L. Andersen et al., "Metabolome-based signature of disease pathology in MS," Mult. Scler. Relat. Disord., 2019.
[33] E. J. deAndrés-Galiana, G. Bea, J. L. Fernández-Martínez, and L. N. Saligan, "Analysis of defective pathways and drug repositioning in Multiple Sclerosis via machine learning approaches," Comput. Biol. Med., vol. 115, Dec. 2019.
[34] S. Ghafouri-Fard, M. Taheri, M. D. Omrani, A. Daaee, and H. Mohammad-Rahimi, "Application of Artificial Neural Network for Prediction of Risk of Multiple Sclerosis Based on Single Nucleotide Polymorphism Genotypes," J. Mol. Neurosci., Mar. 2020.
[35] I. Barbulovic-Nad, M. Lucente, Y. Sun, M. Zhang, A. R. Wheeler, and M. Bussmann, "Bio-microarray fabrication techniques - A review," Critical Reviews in Biotechnology, vol. 26, no. 4. Taylor and Francis Inc., pp. 237–259, 01-Dec-2006.
[36] J. G. Duarte and J. M. Blackburn, "Advances in the development of human protein microarrays," Expert Review of Proteomics, vol. 14, no. 7. Taylor and Francis Ltd, pp. 627–641, 03-Jul-2017.
[37] A. Brazma et al., "Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data," Nature Genetics, vol. 29, no. 4. Nature Publishing Group, pp. 365–371, 2001.
[38] R. Edgar, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," Nucleic Acids Res., 2002.
[39] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression," Science (80-.)., vol. 270, no. 5235, pp. 484–487, Oct. 1995.
[40] P. Guo, Q. Zhang, Z. Zhu, Z. Huang, and K. Li, "Mining gene expression data of multiple sclerosis," PLoS One, vol. 9, no. 6, p. e100052, Jun. 2014.
[41] J.-C. Corvol et al., "Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event," 2008.
[42] R. Ulrich, A. Kalkuhl, U. Deschl, and W. Baumgärtner, "Machine learning approach identifies new pathways associated with demyelination in a viral model of multiple sclerosis Keywords: cholesterol • demyelination • immunohistology • microarray • multiple sclerosis • random forest machine learning algorithm • s," J. Cell. Mol. Med, vol. 14, no. 2, pp. 434–448, 2010.
[43] Y. D. Zhang, C. Pan, J. Sun, and C. Tang, "Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU," J. Comput. Sci., 2018.
[44] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," Proc. Natl. Acad. Sci. U. S. A., 2001.
[45] G. Stelzer et al., "GeneDecks: Paralog hunting and gene-set distillation with genecards annotation," Omi. A J. Integr. Biol., 2009.
[46] M. W. Nachman, "Single nucleotide polymorphisms and recombination rate in humans," Trends in Genetics. 2001.
[47] S. Srinivasan and J. Batra, "Single nucleotide polymorphism typing," in Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 2018.
[48] P. R. Burton et al., "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," Nature, 2007.
[49] L. K. Hoeffding, A. Rosengren, J. H. Thygesen, H. Schmock, T. Werge, and T. Hansen, "Evaluation of shared genetic susceptibility loci between autoimmune diseases and schizophrenia based on genome-wide association studies," Nord. J. Psychiatry, 2017.

[50] S. Ghafouri-Fard, M. Taheri, M. D. Omrani, A. Daaee, and H. Mohammad-Rahimi, "Application of Artificial Neural Network for Prediction of Risk of Multiple Sclerosis Based on Single Nucleotide Polymorphism Genotypes," J. Mol. Neurosci., 2020.
[51] C. Lopez, S. Tucker, T. Salameh, and C. Tucker, "An unsupervised machine learning method for discovering patient clusters based on genetic signatures," J. Biomed. Inform., 2018.
[52] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos, "An application of Random Forests to a genome-wide association dataset: Methodological considerations and new findings," BMC Genet., vol. 11, p. 49, Jun. 2010.
[53] J. Ostmeyer et al., "Statistical classifiers for diagnosing disease from immune repertoires: A case study using multiple sclerosis," BMC Bioinformatics, vol. 18, no. 1, p. 401, Sep. 2017.
[54] F. B. S. Briggs et al., "Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12 566 individuals," Hum. Mol. Genet., 2010.
[55] S. Purcell et al., "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," Am. J. Hum. Genet., vol. 81, no. 3, pp. 559–575, Sep. 2007.
[56] M. M. A. El Hamid, N. M. Ali, M. N. Saad, M. S. Mabrouk, and O. G. Shaker, "Multiple sclerosis: an associated single-nucleotide polymorphism study on Egyptian population," Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 9, no. 1, p. 48, Dec. 2020.
[57] S. L. Andersen et al., "Metabolome-based signature of disease pathology in MS," Mult. Scler. Relat. Disord., vol. 31, pp. 12–21, Jun. 2019.
[58] J. Ostmeyer et al., "Statistical classifiers for diagnosing disease from immune repertoires: A case study using multiple sclerosis," BMC Bioinformatics, 2017.
[59] R. Sun, K. L. Hsieh, and J. J. Sosnoff, "fall Risk prediction in Multiple Sclerosis Using postural Sway Measures: A Machine Learning Approach."
[60] C. Cortes, "Support-Vector Networks," 1995.
[61] V. Kecman, "Support Vector Machines – An Introduction," 2005, pp. 1–47.
[62] T. Hastie, R. Tibshirani, and J. Friedman, Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction. 2009.
[63] D. Ignatov and A. Ignatov, "Decision stream: Cultivating deep decision trees," in Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2018, vol. 2017-November, pp. 905–912.
[64] F. Wang Cynthia Rudin CSAIL, "Falling Rule Lists," 2015.
[65] M. Barsacchi, A. Bechini, and F. Marcelloni, "An analysis of boosted ensembles of binary fuzzy decision trees," Expert Syst. Appl., vol. 154, p. 113436, Sep. 2020.
[66] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell., 1998.
[67] B. Xu, J. Z. Huang, G. Williams, Q. Wang, and Y. Ye, "Classifying very high-dimensional data with random forests built from small subspaces," Int. J. Data Warehous. Min., 2012.
[68] R. E. Kass, "Statistical inference: The big picture," Stat. Sci., 2011.
[69] A. Pérez, P. Larrañaga, and I. Inza, "Bayesian classifiers based on kernel density estimation: Flexible classifiers," Int. J. Approx. Reason., 2009.
[70] C. C. Aggarwal and C. C. Aggarwal, "An Introduction to Neural Networks," in Neural Networks and Deep Learning, Springer International Publishing, 2018, pp. 1–52.
[71] C. C. Aggarwal and C. C. Aggarwal, "Advanced Topics in Deep Learning," in Neural Networks and Deep Learning, Springer International Publishing, 2018, pp. 419–458.
[72] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, "Learning activation functions to improve deep neural networks," in 3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings, 2015.
[73] D. Hendrycks and K. Gimpel, "Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units," arXiv, 2016.
[74] J. Tolles and W. J. Meurer, "Logistic regression: Relating patient characteristics to outcomes," JAMA - Journal of the American Medical Association, vol. 316, no. 5. American Medical Association, pp. 533–534, 02-Aug-2016.
[75] K. P. Murphy, Machine learning : a probabilistic perspective. MIT Press, 2012.
[76] B. U. Park, L. Simar, and V. Zelenyuk, "Nonparametric estimation of dynamic discrete choice models for time series data," Comput. Stat. Data Anal., 2016.

[77] E. M. Sweeney et al., "A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI," PLoS One, vol. 9, no. 4, p. e95753, Apr. 2014.
[78] A. S. Altheneyan and M. E. B. Menai, "Naïve Bayes classifiers for authorship attribution of Arabic texts," J. King Saud Univ. - Comput. Inf. Sci., 2014.
[79] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," in 2014 International Conference on Communication and Network Technologies, ICCNT 2014, 2015.
[80] L. Low, M. Tammi, L. Low, and M. T. Tammi, "Introduction to Next Generation Sequencing Technologies," in Bioinformatics, 2017.
[81] G. B. Han and D. H. Cho, "Genome classification improvements based on k-mer intervals in sequences," Genomics, vol. 111, no. 6, pp. 1574–1582, Dec. 2019.
[82] E. Aun Id, V. Kisand, T. Tenson, and M. Remm Id, "A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria," 2018.
[83] A. Kishk et al., "A Hybrid Machine Learning Approach for the Phenotypic Classification of Metagenomic Colon Cancer Reads Based on Kmer Frequency and Biomarker Profiling," in 2018 9th Cairo International Biomedical Engineering Conference, CIBEC 2018 - Proceedings, 2019, pp. 118–121.
[84] T. Kim, H. D. Seo, L. Hennighausen, D. Lee, and K. Kang, "Octopus-toolkit: A workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data," Nucleic Acids Res., vol. 46, no. 9, May 2018.
[85] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," Bioinformatics, vol. 30, no. 15, pp. 2114–2120, Aug. 2014.
[86] C. Chen, S. S. Khaleel, H. Huang, and C. H. Wu, "Software for pre-processing Illumina next-generation sequencing short read sequences," Source Code Biol. Med., 2014.
[87] T. Lencz et al., "PLINK: A tool Set for wholegenome ssociation and population based linkage analyses," Front. Genet., 2018.
[88] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, 2016.
[89] L. Rampasek and A. Goldenberg, "TensorFlow: Biology's Gateway to Deep Learning?," Cell Systems. 2016.
[90] S. Andrews, "FASTQC A Quality Control tool for High Throughput Sequence Data," Babraham Inst., 2015.
[91] A. Haroon, "PubMed (http://www.ncbi.nlm.nih.gov/PubMed)," The Lancet, 1998. .