International Journal of Engineering & Technology, 7 (4) (2018) 5533-5538

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.11016
Research paper

A survey of machine learning techniques for genomic diseases and data sets

Manu Phogat [1] *, Dr. Dharmender Kumar [2]

[1]Research Scholar, Deptt. Of CSE, GJUS&T, Hisar, India
[2]Professor, Deptt. of CSE, GJUS&T, Hisar, India
**Corresponding author E-mail: kunjean4181@gmail.com*

Abstract

From the very early age of Medical Science, medical practitioners have been concerned about visualizing and analyzing complex biological data which was not so easy. Today is the era of GWAS (genome-wide association studies), so the quest for understanding the genotype of various complex diseases is rapidly increasing day by day. Recently, high throughput molecular data have provided ample information about the whole genome, and have popularized the computational tools in genomics. Due to the humongous size and high dimensionality of genomic data, it is not possible to analyze it with conventional techniques, so machine learning tends to develop efficient computational techniques that will raise with experience, for analysis the vast complex data sets. This article give an outline of different machine learning techniques for examination of the genomics data of diseases and epigenetic, proteomic data.

Keywords: *Machine Learning; ANN, KNN, RF, SVM, Genomic; Mutation.*

1. Introduction

The field of machine learning is the examination of computational techniques that will improvise with experience. Machine learning strategies have been connected to an extensive variety of territories with in genomics and genetics. This survey describes that how machine learning techniques can be used to in diseases related to genomic data.
Genomics is a discipline in genetics that relates to the study of the functional and informational structure encoded in the DNA sequences of living cells. The diploid human genome has twenty-three pairs of chromosomes is made out of 20-25 thousand genes; and haploid set contains evaluated to be 3.2*109 base pairs [1], which contains huge measure of genomic information. A genome is known to be an instruction booklet for building of an organism, and each gene is a like a page in the instruction booklet. The physical medium of genetic information storage [2] are the DNA molecules since 1953 and by year 2000 the human genome is sequenced, assembled and annotated which produced a very large raw information content [3]. The bigger challenge is to interpret the function, structure and meaning of genetic information. The genes that build or which codes for any molecules or protein are called protein coding genes, whereas noncoding genes have regulatory functions and does not code any protein. The human genome nearly contains 20000 protein coding genes, and 25000 non coding genes [4]. Some genes are essential for life, some are essential for health, and some can be cut out in their entirety without apparent harm. In the entire structure of a gene the occurrence of alternating regions are known as introns and exons. To decide the limits between the regions we need to discover the examples in the nucleotide sequence. Many diseases caused by mutation are occurring due to disturbing in the nucleotide patterns.
Sickle cell anemia is a leading genetic disorder in African, Mediterranean, south and Central American, Caribbean and Middle Eastern regions [5]. A survey states that only 50% of patients having sickle cell anemia survived beyond fifth decade [6].
In figure.1 the single nucleotide change from adenine to thymine in a normal gene tends to produce an acid called Valine instead of Glutamic acid that causes the sickle cell disease. Though the Sickle cell disease is well studied and diagnosed by outward symptoms, but genetic testing is necessary for confirmation and Therapeutic development. This mechanism is more complex in other genetic diseases.

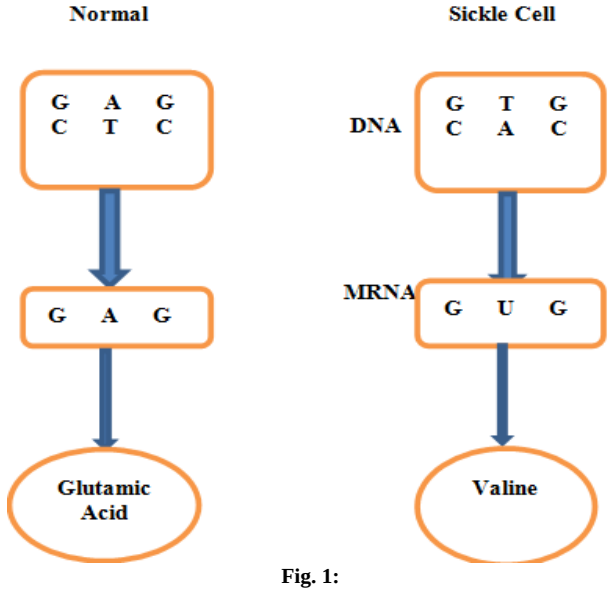

Fig. 1:

The most common heterogeneous genetic disease are autism, cystic fibrosis, cancer etc. many diseases shows similar symptoms but

require different treatments [7], like for autism, cancer, it is essential to have genomic data for more detailed diagnosis [8].
Currently the most understood region in genome are protein coding exons. The protein genetic code was confined over 50 years ago [9]. The mutation caused in coding region of genetic code is called as coding mutation. Coding mutation generally change in corresponding amino acid sequence. The disease causing mutations are also found outside the protein coding region that is also called non coding region of genetic code [10].
There is various kind of gene base mutation in human diseases or disorders. The table.1 shows the various types of mutation and their corresponding genes.

Table 1: Common Mutation and Related Diseases

Mutation	Definition	Example (Gene)	Disease/ condition	References
Point Mutation	Mutation in single base of DNA. It includes change of pyrimidine (C, T) to pyrimidine (T, C) or trans version: pyrimidine (T, C) to purine (A, G).	A>G, A>T		[11]
1. Missense (Non Synonymous)	Resulting different amino acid due to change in single nucleotide.	A82P (HSD3B2)	3βHSD deficiency	[12]
2. Nonsense	Due to change in single nucleotide resulting stop codon.	G23X (HBB)	Beta Thalassemia	[13]
3. Synonymous	Change in single nucleotide which changes a codon into amino acid with similar properties.	V153I (GJB2)	Hearing loss	[14]
4. Silent	It's a type of Missense, but change in third nucleotide of a codon doesn't change the amino acid.	I69I (GJB2)	Hearing loss	[14]
5. Neutral	A change in single nucleotide a neither beneficial nor harmful.			
Duplication	Segment of DNA breaks off and attached onto the homologous chromosome.	920dupTCAG (LDLR)	Familial hypercholesterolemia	[15]
Deletion	Segment of chromosome is deleted and loss of gene at that part of segment.	delE120 (GJB2)	Hearing loss	[16]
Insertion	Insertion of a nucleotide pair at some location on DNA.	3524insA (FBN1)	Marfan syndrome	[17]
Splice mutation	Change in the splice site of introns and exons.	IVS1+1G>A (GJB2)	Hearing loss	[14]
Dynamic mutation	Numerous quantities of copies of a sequence are changed in this mutation amid meiosis division.	(CGG)n>200 (FMR1)	Fragile X syndrome	[18]

Table 2:

Decade wise Developments in the field of Bioinformatics	
1970's	Sequence databases, similarity matrices, Molecular Evolution. PAM matrices used for protein sequence.
1980's	Sequence Alignment search (Smith-waterman algorithm), FASTA and BLAST Algorithms.
1990's	HMM (Hidden Markov Model), Ab inito Protien Structure prediction, Genomics and Comparative Genomics.
2000's	Human Genome is Sequenced, assembled and annotated, GWAS (Genome Wide Association studies), Bio image informatics,
2010's	High-Throughput technologies, Next generation sequencing, Transcriptome Sequencing, Gene prediction, RNA Secondary Structure Prediction etc.

To know the genotype and phenotype of a disease various tools of bioinformatics and computational biology have emerged over last 50 years. The decade wise advancement in the field of bioinformatics describe in table 2. Due to this rapid development in bioinformatics field the molecular data increasing very rapidly. The molecular data basically consist of:

- Gene and protein sequence
- Genome sequence
- Protein structure
- Chemical compounds

The main focus is on inferring properties of molecular data is like predict the function of a gene and its given sequence, predict the structure of a protein and predict the boundaries of a gene given in genome segment. To find out the needful information from this huge amount of data one has to use computational techniques such as machine learning rather than traditional bioinformatics techniques. This review incorporates some significant machine learning techniques and their application in genomics. The main application areas of machine learning in medicine or genomics are:

- Disease Identification/Diagnosis
- Drug Discovery/Manufacturing
- Clinical Trial Research
- Radiology and Radiotherapy
- Smart Electronic Health Records
- Epidemic Outbreak Prediction
- Personalized Treatment/Behavioral Modification

2. Machine learning in genomic data

The field of Machine learning came out from the artificial intelligence community. Machine learning becomes very popular in late 1990's. Machine learning can be characterized as the capacity of the computing machine to expand its performance based on previous outcomes. The main goal of machine learning is to find and learn and afterward adapt to the conditions that may change over the long run and hence enhancing the performance of the machine. This segment gave a review of various types of learning techniques and furthermore talks about machine learning strategies and their application in genomics. The survey; it additionally plots a portion of the real difficulties of applying machine learning strategies to genomics information pragmatic issues.

2.1. Supervised and unsupervised learning

Machine learning strategies usually falls into two classes: supervised and unsupervised techniques. Supervised methods are trained the model on labelled datasets, so it can predict the outcome of out-of –sample data, whereas the unsupervised methods finds patterns, structure and grouping in a datasets without labels. There is learning between supervised and unsupervised learning called as semi supervised learning, which utilizes a little measure of labelled information with huge unlabelled information.
The popular machine learning techniques used in genomic data are RF (Random Forest), SVM (support vector machine), GA (genetic algorithm), K means clustering, ANN (artificial neural network) etc.

2.2. Random forest

Random forest [19] is an ensemble classifier using many decision tree models. Due to its popularity as an ensemble learning method RF has very vast application in machine learning and data mining. The ensemble models combine the results from different models.

RF is a tree based nonparametric ensemble approach that consolidates the thoughts of versatile closest neighbor with bagging for valuable information versatile deduction. RF grows heaps of decision trees in light of random determination of information and random choice of variables; it also provides the class of dependent variables based on many trees.

RF is additionally used to distinguish and rank factors by exploiting factors essential measures. Such properties of RF make it a convenient tool for genomic data analysis. In genomic data analysis predication is a primary goal. For example, one needs to foresee diseases status, for example, cancer sub types utilizing genetic markers. RF is a suitable technique and has been for the most part used to predict clinical outcomes under various high throughput genomic stages. Chen et al. [20] used Random Forest to propose a pathway based predictors other than of each gene for cancer survival prediction. The method had performed well in both predication accuracy and interpretations. Random forest also successfully apply to predict the binding sites of DNAs [21], RNA-protein [22], and predict sites of protein-protein interactions [23]. It also usefully applied in building models for predicting drug responses for tumor cell lines [24].

Presently day's cutting edge genome-wide association (GWA) examines indicates us disease association with regular genetic varieties utilizing a thousands of SNPs across the human genome. Li –chung chung [25] used random forest for building a risk model for bipolar disorder form GWA data. The RF approach shows a significant performance in selecting the informative markers from huge GWA data. Shi et al. [26] fruitfully used RF for cancer class discovery based on immunohistochemical tumor marker expression.

The RF has proven to be a powerful and effective statistical learning tool for complex and high dimensional genomic data; yet it has possessed some disadvantages such as:

- RF is still not fully understood in non-standard small sample size and large feature space.
- RF is quite slow to create predictions once trained.
- Random forests have been seen to over fit for some datasets with noisy classification/regression tasks.

2.3. Support vector machine

SVM is a supervised machine learning technique developed by Vapnik and Cortes [27]. The primary objective of support vector machine is to locate the optimal and isolate the hyper plane which augments the margin of the training data. Margin is the maximum distance of the hyper plane to the closest data points from both classes and vectors that are that are nearest to this hyper plane called support vectors. The main goal is to minimize the classification error and maximize the margin.

SVM centre on identifying a linear separator to partition data points of two classes that is the reason SVM called as a non-probabilistic binary linear classifier. In contrast with other machine learning techniques SVM is capable at perceiving unpretentious patterns in complex data sets [28], because of this advantage SVM is highly used technique for genomic data and other complex computational biological problems. Shen [29] used a model selection method using SVM, and proposed a two stage technique to detect diseases associated SNP interactions. The SVM has been used in cancer classification after the availability of high throughput microarray gene expression data in early 2000's. Waddell et al. [30] used support vector machine for studying the special case of multiple myeloma, a tumour of antibody discharging plasma cell that develop and grows in the bone marrow. In a study Moler et al. [31] used support vector machine in classification of the tissues of colon cancer using the selected features. They used a collection of 20 normal colon tissue and 40 colon cancer tumours. Chen et al. [32] developed a netSVM (Network compelled support vector machine for amending biologically organize biomarkers utilizing cooperation of gene expression data and protein –protein interaction data).

SVM is most widely used technique to identify mutation which is specific to cancer. The classifier were trained to predict whether the mutation occur across complete genome [33] or as in specific class of proteins. The method shows huge accurate predictions base on cross validation. Bari et al. [34] develops a support vector machine model to reveal another class of tumour-related genes that are neither mutated or nor differentially expressed. Though SVM is widely used in computation of genomic data but it has some disadvantages also such as:

- SVM is slow to train, especially if the input data sets have a large number of features.
- The results and inner working of SVM is also difficult to understand because of the model which is based on complex mathematical system.

2.4. Artificial neural networks

ANN consider as most simplified model of biological network structure [35]. The main advantage of ANN is that it does not need any mathematical model. The learning procedure of ANN enlightens the interconnections between the processing components which constitute the network topology. ANN organized into various layers, with each layer consists number of respective neurons which constitute that layer. The number of hidden layer and neurons in each layer always depend on complexity of problem. The feed forward and back propagation are most common in artificial neural network.

The primary target of neural network is to build a model that accurately maps the quantity of inputs to their individual outputs using information with the objective that the model can be used to create the output when the desired output is unknown. Due to its ability to cope up with noisy, non-linear and high dimensional datasets, ANN used in many areas of computational biology and medicine. The first major application using ANNs for these complex data sets was perhaps by the seminal paper of Khan et al.[36]. In another study Catalogna et al. [37] use ANN for analysis of urine and blood samples of diabetic patients. Narayanan et al. [38] use ANN to analyze gene expression datasets of patients diagnosed with multiple myeloma and normal bone marrow cases and find that genes that were reliably positive or negatively expressed could be recognized from huge datasets. In another study ANN is used by karabulut et al. [39] for the early diagnosis of coronary artery Diseases (CAD). Though use ANN is very common in genomics data but it also possesses some disadvantages such as:

- ANN is unable to perform better on non- linear data.
- Artificial Neural network models are inclined to over fitting.

2.5. K- Nearest neighbours

KNN is an unsupervised machine learning algorithm and also very simple classification and regression algorithm. In KNN the classification case new information focuses get classified specifically class and in regression, new information gets labelled tie on average value of k nearest neighbour. KNN also called as, example based reasoning, lazy learning or memory based reasoning. The default method of measuring distance in KNN is Euclidean distance. KNN has huge application in genomic data as listed below in table 3, but it also have some limitations such as:

- KNN use lot of space when training set is large.
- KNN is also a lazy learner.

Table 3: Machine Learning in Diseases

Machine Learning technique	Disease	Scope	Datasets	References
ANN	Cardiac arrest	Predict Cardiac arrest risks	http://archive.ics.uci.edu/ml/ datasets/Heart+Disease, UCI Machine Learning Repository	[40]
K-NN	Heart diseases	To check weather patients at high risk of having heart disease	http://archive.ics.uci.edu/ml/ datasets/Heart + Disease, UCI Machine Learning Repository	[41]
SVM	Diabetes	To foresee patients at high danger of diabetes	Pima Indian Diabetes Dataset	[42]
ANN	Liver tumor	Predict liver tumor	Database of National Health Insurance Research Taiwan	[43]
ANN	Hepatitis	Recognize whether if patients persisting hepatitis are expected to be alive or not	UCI Machine Learning Repository	[44]
SVM	Breast Cancer	examine of Breast cancer	Breast Cancer Wisconsin	[45]
ANN	Hepatitis type C	examine of hepatitis C virus	UCI Machine Learning Repository	[46]
K-NN	Breast tumor	To distinguish as normal or abnormal	IRMA MIAS	[47]
SVM	Breast Tumor	Select the markers for tumor	Digital Database for Screening Mammography	[48]
K-NN	Leukemia	Analysis of microarray data	http://www.ncbi.nlm.nih.gov/ gds/, National Center of Biotechnology Information	[49]
RF	Colon cancer	Identify colon cancer	(HCT-116 colon cancer) TCF7L2 dataset	[50]
ANN	Breast tumor	To distinguish between normal and micro calcifications and then between benign & alignant	MIAS	[51]
SVM	Breast Cancer	Selecting gene responsible for tumor	tumor dataset Chung-Shan Medical University Hospital	[52]
RF	Kidney tumor	Finding kidney tumor	HEK293 (embryonic kidney cells) TCF7L2 dataset	[53]
ANN	Ovarian cancer	Diagnosis of ovarian tumor	. Asp, FDA-NCI Clinical Proteomics Program Databank	[54]
RF	liver cancer	Identify liver tumor	HepG2 (liver cancer cells) TCF7L2 dataset	[55]
K-NN	(SRBCT)	Locate a small set of most noteworthy instructive genes to characterize cancer	www.broadinstitute. org/cgibin/cancer/publications/ pub_paper.cgi?mode%20=% 20view&paper_id=43	[56]

Table 4: Some Popular Genomics Datasets

Types of Data	Source	Refrences
Genotype: SNP arrays , Exome and whole-genome sequencing Transcriptome: RNA-seq Phenotype: comprehensive profiles of subjects	Genotype Tissue Expression (GTEx)	[57]
Genotype: exome sequencing Transcriptome: m/miRNA microarray Phenotype :cancer cell lines(drug tested) Proteome: SWATH profiles	NCI-60 (National Cancer Institute Anticancer drug screen)	[58]
Genotype: whole genome for subset of cell lines Transcriptome: RNA-seq Epigenome: CHIP-seq, DNASE ,5C	Encyclopedia of DNA Elements(ENCODE)	[59]
Genotype: cancer whole – genomes Phenotype: pathology reports	International Cancer Genome Consortium(ICGC)	[60]
Genotype: cancer whole genome and exome sequencing Transcriptome: RNA –seq (m/miRNA) Phenotype: pathlogy reports(baseline and drug-tested) Proteome: expression levels of signalong pathways(reverse- phase protein arrays) some samples matched with TCGA Epigenome: methylation	TCGA (The Cancer Genome Atlas), TCPA (The Cancer Proteome Atlas)	[61][62]
Genotype: whole genome sequencing , high quality variant calls Transcriptome: RNA-seq for a large fraction of cell lines Phenotype: different populations, trios(parents and offsprings)	The 1000 Genomes Project	[63]
Genotype: whole genome (a subset of cell lines) Transcriptome: RNA-seq & smRNA-seq Phenotype:tens of cell lines and ex vivo differentiated cells Epigenome: comprehensive chIP - seq	NIH Roadmap Epigenomics Project	[64][65]
Genotype: SNP arrays Phenotype: body size and measure of obesity	Genetic Investigation of Anthropometric Traits (GIANT)	[66][67]

3. Discussion and conclusion

Today one of the most challenging problems in bioinformatics or computational biology is to analyze the tremendous volume of information. Machine learning has been important tool to analyze both small scale and large scale information. The machine learning techniques provide a robust, computationally efficient assumption for several computational biological problems that are difficult to solve out by traditional techniques. This review discussed the general philosophy for the utilization of machine learning

techniques in genomic data of various diseases. The review introduces some most useful techniques of machine learning such as SVM, RF, ANN, KNN etc. and shows their application in genomic data of diseases. Moreover despite the common use of machine learning in genomic data, it possesses many limitations such as in terms performing slow on massive data using big platforms; also over fitting is an also common mistake in ML techniques. So for the future perspectives the machine learning techniques combines with techniques such as Metaheuristics methods both single solution and population based like as swarm intelligence, tabu search, genetic algorithm etc. to find the optimum results in genomic data for various diseases.

References

[1] International human genome sequencing consortium: Finishing the euchromatic sequence of the human genome. Nature 2004; 431(7011):931- 45. https://doi.org/10.1038/nature03001.

[2] J. D. Watson and F. H. C. Crick (1953), ''Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid,'' Nature, vol. 171, no. 4356, pp. 737–738. https://doi.org/10.1038/171737a0.

[3] E. S. Lander et al. (2001), ''Initial sequencing and analysis of the human genome,'' Nature, vol. 409, no. 6822, pp. 860–921. https://doi.org/10.1038/35057062.

[4] J. Harrow et al. (2012), ''GENCODE: The reference human genome annotation for the ENCODE project,'' Genome Res., vol. 22, no. 9, pp. 1760–1774. https://doi.org/10.1101/gr.135350.111.

[5] Kevin Jarrett, Mary Williams, Spencer Horn, David Radford, and J. Michael Wyss (2016), "Sickle cell anemia: tracking down a mutation": an interactive learning laboratory that communicates basic principles of genetics and cellular biology" Advances in Physiology education, vol.40, pp. 110-115. https://doi.org/10.1152/advan.00143.2015.

[6] Gravitz L, Pincock S. (2014), "Sickle-cell disease" Nature, Vol. 515, Issue.7526. https://doi.org/10.1038/515S1a.

[7] D. Hanahan and R. A. Weinberg (2011), ''Hallmarks of cancer: The next generation,'' Cell, vol. 144, no. 5, pp. 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

[8] M. A. Rubin (2015), ''Make precision medicine work for cancer care,'' Nature, vol. 520, no.547, pp. 290–291. https://doi.org/10.1038/520290a.

[9] F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin (1961), ''General nature of the genetic code for proteins,'' Nature, vol. 192, pp. 1227–1232. https://doi.org/10.1038/1921227a0.

[10] L. A. Hindorff et al. (2009), ''Potential etiologic and functional implications of genome-wide association loci for human diseases and traits,'' Proc. Nat. Acad. Sci. USA, vol. 106, no. 23, pp. 9362–9367. https://doi.org/10.1073/pnas.0903103106.

[11] Rabbani B, Mahdieh N, Haghi Ashtiani MT, et al. (2011), "Molecular diagnosis of congenital adrenal hyperplasia in Iran:Focusing on CYP21A2 gene" , Iranian Journal of Pediatrics, vol.21, no.2, pp.139-50.

[12] Rabbani B, Mahdieh N, Haghi Ashtiani MT, et al. (2012), "In silico structural, functional and pathogenicity evaluation of a novel mutation:An overview of HSD3B2 gene mutations", Gene, vol.503, no.2, pp.215-219. https://doi.org/10.1016/j.gene.2012.04.080.

[13] Ghanem N, Girodon E, Vidaud M, et al. (1992), "A comprehensive scanning method for rapid detection of beta-globin gene mutations and polymorphisms", Human Mutation, vol.1, no.3, pp.229-239. https://doi.org/10.1002/humu.1380010310.

[14] Mahdieh N, Rabbani B, Wiley S, et al. (2010), "Genetic causes of nonsyndromic hearing loss in Iran in comparison with other populations", Journal of Human Genetics, vol.55, pp. 639-48. https://doi.org/10.1038/jhg.2010.96.

[15] Garcia-Garcia AB, Real JT, Puig O, et al. (2001), "Molecular genetics of familial hypercholesterolemia in spain:Ten novel LDLR mutations and population analysis", Human Mutation, vol.18, no.5, pp.458-469. https://doi.org/10.1002/humu.1218.

[16] Mahdieh N, Bagherian H, Shirkavand A, et al. (2010), " High level of intrafamilial phenotypic variability of non- syndromic hearing loss in a Lur family due to DELE120 mutation in GJB2 gene", International Journal of Pediatric Otorhinolaryngology, vol.74, no.9, pp.1089-91. https://doi.org/10.1016/j.ijporl.2010.06.005.

[17] Schrijver I, Liu W, Odom R, et al. (2002), "Premature termination mutations in FBN1: Distinct effects on differential allelic expression and on protein and clinical phenotypes", American Journal of Human Genetics, vol.71, no.2, pp. 223-37. https://doi.org/10.1086/341581.

[18] Madan K, Seabright M, Lindenbaum RH, et al. (1984), "Paracentric inversions in man", Journal of Medical Genetics, vol.21, no.6, pp. 407-412. https://doi.org/10.1136/jmg.21.6.407.

[19] Xi Chen, Hemant Ishwaran (2012), "Random forests for genomic data analysis", Genomics, vol. 99, pp. 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003.

[20] X. Chen, L.Wang, H. Ishwaran (2010), "An integrative pathway-based clinical-genomicmodel for cancer survival prediction", Statistics & Probability Letters. Vol.80 no.17–18, pp. 1313–1319. https://doi.org/10.1016/j.spl.2010.04.011.

[21] J.S. Wu, H.D. Liu, X.Y. Duan, Y. Ding, H.T. Wu, Y.F. Bai, X. Sun (2009), "Prediction of DNAbinding residues in proteins from amino acid sequences using a random forest model with a hybrid feature", Bioinformatics, vol.25, no.1, pp.30–35. https://doi.org/10.1093/bioinformatics/btn583.

[22] Z.P. Liu, L.Y. Wu, Y. Wang, X.S. Zhang, L. Chen (2010), "Prediction of protein–RNA binding sites by a random forest method with combined features", Bioinformatics, vol. 26, no.13, pp.1616–1622. https://doi.org/10.1093/bioinformatics/btq253.

[23] M. Sikic, S. Tomic, K. Vlahovicek (2009), "Prediction of protein–protein interaction sites in sequences and 3D structures by random forests", PLOS Computational Biology, vol.5, no.1, e1000278. https://doi.org/10.1371/journal.pcbi.1000278.

[24] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, H.A. Fine (2011), "Predicting in vitro drug sensitivity using random forests", Bioinformatics, vol. 27, no. 2, pp.220–224. https://doi.org/10.1093/bioinformatics/btq628.

[25] Li-ChungChuang, and Po-Hsiu Kuo (2017), "Building a genetic risk model for bipolar disorder from genomewide association data with random forest algorithm", Scientific Reports, Nature, vol.7, no. 39943, pp. 1-10.

[26] T. Shi, D. Seligson, A.S. Belldegrun, A. Palotie, S. Horvath (2005), "Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma", Mod. Pathol, vol. 18, no.4, pp.547–557. https://doi.org/10.1038/modpathol.3800322.

[27] Vapnik V (1963), "Pattern recognition using generalized portrait method", Automation Remote Control, vol. 24, pp.774-780.

[28] Shujun Huang et al. (2018), "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics", Cancer Genomics & Proteomics, vol.15, pp. 41-51.

[29] Y. Shen, Z. Liu, and J. Ott (2012), "Support Vector Machines with L 1 penalty for detecting gene–gene interactions," International journal of data mining and bioinformatics, vol. 6, pp. 463-470. https://doi.org/10.1504/IJDMB.2012.049300.

[30] Waddell M, Page D, Zhan F (2005), Predicting cancer susceptibility from single-nucleotide polymorphism data: A case study in multiple myeloma. Proceedings of the 5th ACM SIGKDD Workshop on Data Mining in Bioinformatics. Chicago, IL. https://doi.org/10.1145/1134030.1134035.

[31] Moler E, Chow M and Mian I (2000), "Analysis of molecular profile data using generative and discriminative methods", Physiological Genomics, vol. 4, no.2, pp. 109-126. https://doi.org/10.1152/physiolgenomics.2000.4.2.109.

[32] Chen L, Xuan J, Riggins RB, Clarke R and Wang Y (2011), "Identifying cancer biomarkers by network-constrained support vector machines," BMC Systems Biology, vol. 5, no.1, pp. 161. https://doi.org/10.1186/1752-0509-5-161.

[33] Capriotti E and Altman RB (2011), "A new disease-specific machine learning approach for the prediction of cancer-causing missense variants," Genomics, vol. 98, no.4, pp. 310-317. https://doi.org/10.1016/j.ygeno.2011.06.010.

[34] Bari MG, Ung CY, Zhang C, Zhu S and Li H (2017), "Machine Learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks," Scientific Reports, vol.7, pp. 6993. https://doi.org/10.1038/s41598-017-07481-5.

[35] Taghipour M1, Vand AA, Rezaei Aand Karim GR (2015), "Application of Artificial Neural Network for Modeling and Prediction of MTT Assay on Human Lung Epithelial Cancer Cell Lines," Journal of Biosensors & Bioelectronics, vol.6, no.2.

[36] Khan J, Wei JS, Ringner M, et al. (2001), "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol.7, pp.673–679. https://doi.org/10.1038/89044.

[37] Catalogna M, Cohen E, Fishman S, Halpern Z, Nevo U, Ben-Jacob E 92012), "Artificial neural networks-based controller for glucose monitoring during clamp test," Public Library of Science One, vol.7, no. e44587.

[38] Narayanan A, Keedwell EC, Gamalielsson J, et al. (2004), "Single-layer artificial neural networks for gene expression analysis," Neurocomputing, vol.61, pp.217–40. https://doi.org/10.1016/j.neucom.2003.10.017.
[39] Karabulut E, Ibrikçi T. (2012), "Effective diagnosis of coronary artery disease using the rotation forest ensemble method," Journal of Medical Systems, vol.36, pp.3011–3018. https://doi.org/10.1007/s10916-011-9778-y.
[40] Samuel, O.W., Asogbon, G.M., Sangaiah, A.K., Fang, P., Li, G. (2017), "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Systems with Applications, vol.68, pp.163–172. https://doi.org/10.1016/j.eswa.2016.10.020.
[41] Shouman, M., Turner, T., Stocker, R. (2012), "Applying k-nearest neighbour in diagnosing heart disease patients," Int. J. Inf. Educ. Technol, vol.2, no.3, pp. 220. https://doi.org/10.7763/IJIET.2012.V2.114.
[42] V. Anuja Kumari, R.Chitra (2013), "Classification Of Diabetes Disease Using Support Vector Machine," International Journal of Engineering Research and Applications, vol.3, no. 2, pp.1797-1801.
[43] Rau, H.-H., Hsu, C.-Y., Lin, Y.-A., Atique, S., Fuad, A., Wei, L.-M., Hsu, M.-H (2016), , "Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network," Computer Methods and Programs in Biomedicine, vol.125, pp. 58–65. https://doi.org/10.1016/j.cmpb.2015.11.009.
[44] Kaya, Y., Uyar, M. (2013), "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease," Applied Soft Computing, vol.13, no.8, pp.3429–3438. https://doi.org/10.1016/j.asoc.2013.03.008.
[45] Joshi J., Doshi R., Patel J. (2014), "Diagnosis and prognosis breast cancer using classification rules," International Journal of Engineering Research and General Science, vol.2, no.6, pp. 315–323.
[46] Jilani, T.A., Yasin, H., Yasin, M.M. (2011), "PCA-ANN for classification of Hepatitis-C patients," International Journal of Computer Applications, vol.14, no.7, pp. 1–6 (0975–8887).
[47] Gardezi, S.J.S., Faye, I., Bornot, J.M.S., Kamel, N., Hussain, M. (2017), "Mammogram classification using dynamic time warping," Multimedia Tools and Applications, pp.1–22.
[48] Abdelaal M.M.A., Farouq M.W., Sena H.A., Salem A.-B., M., "Using data mining for assessing diagnosis of breast cancer," International Multiconference on Computer Science and Information Technology; 2010 March 17–19; Hong Kong, China. p. 11–17.
[49] Kumar, M., Rath, N.K., Rath, S.K. (2016), "Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier," The Journal of Biomedical Informatics, vol.60, pp.395–409. https://doi.org/10.1016/j.jbi.2016.03.002.
[50] Gasiorek JJ, Blank V. (2015), "Regulation and function of the NFE2 transcription factor in hematopoietic and non-hematopoietic cells," Cell Mol Life Sci CMLS, vol.72, pp.2323–35. https://doi.org/10.1007/s00018-015-1866-6.
[51] Mohamed, H., Mabrouk, M.S., Sharawy, A. (2014), "Computer aided detection system for micro calcifications in digital mammograms," Computer Methods and Programs in Biomedicine, vol.116, no.3, pp. 226–235. https://doi.org/10.1016/j.cmpb.2014.04.010.
[52] Huang C.-L., Liao H.-C., Chen M.-C. (2008), "Prediction model building and feature selection with support vector machines in breast cancer diagnosis," Expert Systems with Applications, vol.34, pp.578–587. https://doi.org/10.1016/j.eswa.2006.09.041.
[53] Xin Wang, Peijie Lin1 and Joshua W. K. Ho1 (2018), "Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest," BMC Genomics, vol 19, no.1, pp.929. https://doi.org/10.1186/s12864-017-4340-z.
[54] Thakur, A., Mishra, V., Jain, S.K. (2011), "Feed forward artificial neural network: tool for early detection of ovarian cancer," Scientia Pharmaceutica, vol.79, no.3, pp.493–506. https://doi.org/10.3797/scipharm.1105-11.
[55] Babeu J-P, Boudreau F. (2014), "Hepatocyte nuclear factor 4-alpha involvement in liver and intestinal inflammatory networks," World J Gastroenterol WJG, vol.20, pp.22–30. https://doi.org/10.3748/wjg.v20.i1.22.
[56] Mahmoud, A.M., Maher, B.A., El-Horbaty, E.-S.M., Salem, A.B.M. (2013), "Analysis of machine learning techniques for gene selection and classification of microarray data," Proceedings of the 6th International Conference on Information Technology.
[57] T. G. Consortium, ''the genotype-tissue expression (GTEx) project. (2013)'' Nature Genetics, vol. 45, no. 6, pp. 580–585. https://doi.org/10.1038/ng.2653.
[58] R. H. Shoemaker (2006), ''The NCI60 human tumour cell line anticancer drug screen,'' Nature Rev. Cancer, vol. 6, no. 10, pp. 813–823. https://doi.org/10.1038/nrc1951.
[59] M. Kellis et al. (2014), ''Defining functional DNA elements in the human genome,'' Proc. Nat. Acad. Sci. USA, vol. 111, no. 17, pp. 6131–6138. https://doi.org/10.1073/pnas.1318948111.
[60] T. J. Hudson et al. (2010), ''International network of cancer genome projects,'' Nature, vol. 464, no. 7291, pp. 993–998. https://doi.org/10.1038/nature08987.
[61] K. Chang et al. (2013), ''the cancer genome atlas pan-cancer analysis project,'' Nature Genetics, vol. 45, no. 10, pp. 1113–1120. https://doi.org/10.1038/ng.2764.
[62] J. Li et al., ''TCPA: A resource for cancer functional proteomics data,'' Nature Methods, vol. 10, no. 11, pp. 1046–1047. https://doi.org/10.1038/nmeth.2650.
[63] G. Project et al. (2013), ''an integrated map of genetic variation from 1,092 human genomes,'' Nature, vol. 491, no. 7422, pp. 556–665, 2012.
[64] B. E. Bernstein et al. (2010), ''The NIH roadmap epigenomics mapping consortium,'' Nature Biotechnol., vol. 28, no. 10, pp. 1045–1048. https://doi.org/10.1038/nbt1010-1045.
[65] R. E. Consortium et al. (2015), ''Integrative analysis of 111 reference human epigenomes,'' Nature, vol. 518, no. 7539, pp. 317–330. https://doi.org/10.1038/nature14248.
[66] A. R. Wood et al. (2014), ''Defining the role of common variation in the genomic and biological architecture of adult human height,'' Nature Genetics, vol. 46, no. 11, pp. 1173–1186. https://doi.org/10.1038/ng.3097.
[67] A. E. Locke et al. (2015), ''Genetic studies of body mass index yield new insights for obesity biology,'' Nature, vol. 518, no. 7538, pp. 197–206. https://doi.org/10.1038/nature14177.