The Nucleus **55**, *No. 2 (2018) 85-92*

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)
ISSN 2306-6539 (Online)

Classification of Primary Open Angle Glaucoma through Genetic and Demographic Data

M. Nosheen[1]*, A. Anis[1], M.S. Malik[2] and M.A. Fahiem[1]

[1]*Department of Computer Science, Lahore College for Women University, Lahore, Pakistan*

[2]*Department of Computer Science & IT, The Islamia University, Bahawalpur, Pakistan*

ARTICLE INFO

Article history :
Received : 18 April, 2018
Accepted : 17 September, 2018
Published : 12 October, 2018

Keywords:
Classification,
Demographic data,
Genetic data,
Glaucoma,
POAG,
Prediction

ABSTRACT

Primary Open Angle Glaucoma (POAG) is considered to be one of the leading causes of irreversible blindness. More than 66 million people lie in this category from which 50% of people unaware of its adverse effect. To prevent its adverse effects like blindness there is an imperious need for automated technique to be developed for its detection. Recently, a genetic data has been explored with machine learning techniques for the detection and prevention of POAG. The genes along with other demographic data sets give an evident base for the detection of this disease. In this paper both the genetic and demographic data is used for the detection of this disease. Here an algorithm is proposed for preprocessing. The feature sets comprises of genes (i.e. MYOC, CYP1B1, NTF4, OPTN), SNP alleles, risk alleles, chromosomes, family history, race, age and gender of patients. For this paper, we use 590 patients' genetic and demographic data sets from various online repositories. For the performance evaluation of the proposed approach we have applied different types of classifiers (Naïve Bayes, J48, SMO, LWL, K). The classifiers were evaluated to understand their ability of predicting the desire results on sensitivity, accuracy and specificity parameters. The results revealed that Support Vector Machine (SMO) classifier meet high classification accuracy i.e. 98%.*

1. Introduction

In the past few decades researches in the field of bioinformatics flourish tremendously [1]. The term bioinformatics refers as applying informatics techniques (derived from multiple disciples like math's, computer science and statistics) on biological (molecules) data [2]. In computer science many fields like machine learning, neural networks, and data mining provide set of systematic algorithms and methods for the collection, analysis and interpretation of such data. The technological advancement increases the amount of bioinformatics data especially genetic data and opens many new horizons for bioinformatics researchers [3]. Gene expression, chromosome no's, single nucleotide polymorphisms (SNP), alleleomorphs (Alleles) are some of the example of genetic data. Now it is available in different sizes, formats and structures from different online repositories [4]. Integrating data from multiple sources become a challenge for the researchers [5]. For better and deep understanding of any disease like cancer, diabetes, stroke and eye diseases etc., practitioners combine genetic data with proteomic and demographic factors in their studies [6]. The demographic data contains the clinical details of the patients which contain age, race, family history, blood group, country etc.

The glaucoma is one of the complex eye diseases which comprises a group of clinically and genetically heterogeneous optic neuropathic characterized by a gradual and progressive loss of visual field [7, 8]. Besides other eye related diseases it is considered to be one of the leading causes of preventable blindness over 65 million people across the globe [8, 9]. Almost 50% affected people are unaware of its irreversible damaging facts [9]. Glaucoma is further categorized as primary and secondary based on their causation and aqueous fluid dynamics. In primary glaucoma, POAG is the most common form effecting 1 to 2% population over age 40 [9, 10]. POAG increases with age and is deeply correlated with other inherited diseases (e.g. diabetes etc.) [10, 11]. The demographic data like population, race and geographic location also gives varied genetic effects in POAG [12, 13]. The literature shows that POAG is attributed to multiple genes with different degree of interactions and environmental effects [14]. The genetic disorder and mutation in genes aggravated its effects [14].

In early studies different techniques have been used for the prediction of POAG by using demographic and genetic data sets. These techniques investigate, check, extract and predict the medical data in different disciplines like neural network [15, 16], data mining [17], machine learning [18, 19] etc. Rao et al. [20] defines that genetic heterogeneity is the bases of all types of glaucoma with multiple chromosomal loci associated with the disease. They elaborated that only a few genes are characterized for POAG i.e. MYOC, OPTN, WDR36, NTF4, CYPIBI and LTBP2.

Restrepo et al. [21] proposes a data mining algorithm, for the identification of POAG for the genetic data sets of Africana Americans. Elshazly et al. [22] compare the accuracy of different classifiers (i.e. neural network (NN), decision tree (DT), rotation forest tree (ROT and Fuzzy Logic classifiers.) on the POAG datasets. Rao et al. [23] focuses on the automatic classification of PAOG through fundus images. For that they apply various machine learning classifiers (e.g. GA, CFS etc.) and algorithms on

*Corresponding author : m_sufyan2000@yahoo.com

350 normal and 150 POAG patients data sets. Anitha et al. [8] tests the various classifiers for predicting the glaucoma associated with protein on 205 genetic data sets.

Rao et al. [24] defines the genes variation for Indian population and finds that NTF4, VAV2, and VAV3 are not involved in the POAG patients. They took 537 subjects as datasets where they took age, ethnicity and geographic region as demographic factors. Oracle Advanced Analytics also provide a support through broad range of machine learning algorithms, for the prediction of any types of genetic data sets [25].

To explore the pathogenesis and materialized effects of any disease affected by genes, various machine learning and data mining approaches are adopted by the researchers [8, 21]. These approaches facilitate the physicians not only assessing the risk factors associated with the disease of their patients, but also help in differentiating the similar clinical disorders. In this scenario, predicting the genes with demographic data sets that manifest POAG may play a vital role in its treatment and diagnosis [26].

In this research work we test various classifier methods to find their prediction ability in classifying genetic and demographic data of POAG patients. The prediction ability of all classifies are weighted by using genetic and demographic data gathered from different online repositories. These classification based approaches are grouped by similarity in terms of their functionalities (e.g. probabilistic, tree based methods, instance based etc.). The most widely adopted categories are Instance based methods, Decision trees, Bayesian (probabilistic); Optimization based algorithms and, Regression based methods etc. In this research work we use Locally Weighted Learning (LWL), Kstar (K*) as Instance based methods, J48 as Decision trees, Naïve Bayes as Bayesian method and Sequential minimal optimization (SMO) in Optimization based algorithms.

The next section describes the material and methods used in our proposed approach. The proposed approach section describes the flow of the approach in detail. The result section describes the final outcome of our proposed approach followed by a detailed discussion. At the end we conclude our paper with a comparison between our and other approaches. The future section describes the future prospects of the proposed approach.

2. Materials and Methods

In this paper we have used genetic and demographic data for POAG patients that are collected from multiple online repositories see section 3.1 for details. For this study we use Myocilin (MYOC) [20, 26], Cytochrome (CYP1B1) [20], Neurotrophin (NTF4) [20, 24], OptineurinOPTN [20, 26] as genes. These genes are involved in the onset of POAG disease. Race, gender and family history are selected as demographic factors [27-29]. The detailed description of the data sets is given in Table 1.

Table 1: Dataset details.

Attributes	Values
No. of patients	590
Age Group	40-80
Genetic Data Set	Genes
Allele	SNP, Risk, Chromosomes
Demographic Data	Family History, Age, Race,

3. Proposed Approach

For the prediction of POAG a classification based approach is used. The proposed approach used in this paper is sub divided in to four main phases named as Data extraction, Preprocessing, Feature set Formulation and Classifications. The flow diagram of the proposed approach is given in Fig. 1. The detailed description of each phase is given in next sections

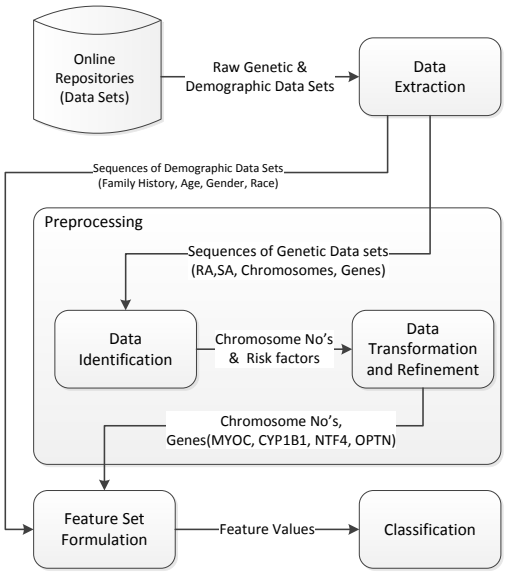

Fig 1: Block Diagram of the proposed approach.

3.1 Data Extraction

The human genome consists of a complete set of nucleic acid sequences. Each nuclei consist of DNA's embedded in to 23 chromosome. Approximately 19,000-20,000 human protein-coding genes are in these ranges which are revised periodically. The genetic repositories are the systematic software based resource where the researchers extract, add and retrieve the genetic data (genes, gene products, variants, phenotypes). In this research paper the data sets are extracted from various online databases i.e. dbSNP, Genes, ClinVar, etc. The data sets are retrieved by using keywords such as POAG, focal adhesion and other similar terms related to POAG. The details about the data sets used in this research work are given in Table 1. Whatever the data is extracted it is considered to be an unstructured data

(Raw genetic and demographic data) and is used for further processing.

3.2 Preprocessing

The basic purpose of this phase is to structure an unstructured data for further processing. In the proposed framework Preprocessing phase is further divided into two sub phases i.e. Data identification and Data transformation, refinement. In data identification phase the genetic cases are defined while data transformation, refinement phase transform these cases through chromosome no's and associated risk factors which are identified and stored in a tabular form for further processing. Generally the genetic cases are decomposed into multiple online databases and are not suitable for further analysis. NBCI (National Center for Biotechnology Information) is one of these databases [4]. It provides biomedical and genomic information from various resources, including GenBank, RefSeq, TPA and PDB. In NBCI different databases contain different type of information e.g. dbSNP contains the SNP Alleles, chromosome information of any disease while ClinVar contains the clinical variations of similar genes. These databases are not only varied according to its functionalities but also varied in terms of data structure and data formats, like dbSNP use ASN.1, XML and FASTA while ClinVar use XML to present their information. Such type of information is considered as unstructured information. The extraction and storage of genetic cases from such data sources is a very difficult task for the researchers. To overcome this difficulty, different tools [30-33], methods [34, 35] and algorithms [36] are proposed by the research community, but these algorithms, methods and tools are complex [36] and need heavy equipment for processing[30]. The comparison of different algorithm in the similar scenarios is given in Table 2.

Table 2: Comparison of different algorithms

Author	Method/Technique/Algorithm	Data Sets	Time Complexity
Daemen et al.[34]	Kernel based Method with LS-SVMs	Clinical and Microarray Data	$O(n^3)$
Troyanskaya et al. [35]	Bayesian Methods	Protein, Genes	$O(2^n)$
Kumar et al. [37]	SWIFT(Sorting tolerant Form Intolerant)	Protein Sequences form Various Databases	$O(n^2)$
Gevaert et al. [38]	Bayesian Networks	Clinical and Microarray Data	$O(2^n)$

An algorithm in Table 3 is designed to structure data for POAG disease. The proposed algorithm structures the genetic cases for further processing. The proposed algorithm takes the genetic data sequences with disease associated genes and population size as an input. The algorithm checks each disease associated gene existence and its location in chromosome. The algorithm repeats the process for the whole disease associated genes. For each gene type the algorithm checks the genetic data containing SNP allele (SA), risk allele(RA) and gene location in chromosome. In this paper we took POAG as a disease and selected its 4 genes (i.e. MYOC, CYP1B1, NTF4, and OPTN). For example in case one gene MYOC is given as an input and if RA, SA and chromosome related to this gene exists then it is extracted. The remaining cases checks similarly for other genes types (i.e. CYP1B1, NTF4, and OPTN). The SA is a variation in the single nucleotide which occurs at some specific positions in the gene while the RA sometimes influences alleles which is used to show the severity of disease in the effected patients. The demographic data taken from other databases like ClinVar of the similar cases has been inserted in the tabular form for further processing. Furthermore, if a genetic data set of POAG patient who do not have the complete information regarding the SA, RA and chromosome details get rejected, and that data are considered discarded. Out of 590 datasets total 415 datasets are structured. The sample data sets after structuring are given in Table 4.

Table 3: Algorithm:For POAG Genetic and Demographic Data Structuring

Input: GENETIC DATA SEQUENCES, GENE_TYPE[MYOC, CYP1B1, NTF4,OPTN], NO_OF_GENES, POPULATION_SIZE

Output: STRUCTURED DATA SETS

```
Set DATA_STRUCTURE [m,j] = 0, n= FACTORS_LENGTH, N= 590
    Set  Gen, RA, SA, Chro, Gender, Race, Age, FH,…….. = 0
    //Factors_length = columns of table, N = Row of table
    // RA= Risk Allele, SA= SNP Allele.
    // Chro= Chromosomes, Gen= Gene.
    For m (1 to N){
        For j (1 to n){
        If "Gen exists in chromosomes".
            {   If GENE_TYPE[a] exist in chromosomes"
                    Then DATA_STRUCTURE[m, j] =RA
                    DATA_STRUCTURE [m,j+1] = SA
                    DATA_STRUCTURE [m, j+…] = ….
                      a = a+1
                    else  a = a+1
                         Break
            }}
```

3.3 Feature Set Formulation

In classification based approaches, feature is considered to be an individual measurable thing or representative of a phenomenon being experiment. In this paper, features set comprises of genes (i.e. MYOC, CYP1B1, NTF4, and OPTN), SNP alleles, Risk alleles, chromosomes, family history, race, age and gender of POAG patients. The total 360 data sets from 415 found after the refinement. The sample of the final feature set is given for classification.

Table 4: Structured data after execution of proposed algorithm.

Patient#	MYOC	CYP1B1	NTF4	OPTN	SA	RA	Chro	Gen	Race	FH	Mutation	Other factors
1	1	1	1	0	C/T, C/G	T,A,C	1,2, 19	Male	Nigeria	Mother	1	·
2	0	1	0	0	C/T	A,T	1	Female	Kenya	Father	0	·
3	1	0	0	1	G/T, A/T, A/C	A,G,C	1,10,19	Male	China	Grandfather	1	·
4	1	1	0	1	G/T,C/T, A/G	A,G,T	2,18,10	Female	Nigeria	Mother	1	·
5	0	1	1	0	A/G	A	9	Female	Indian	Father	0	·

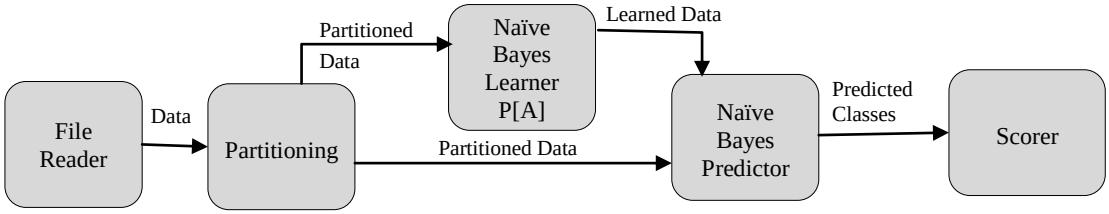

Fig. 2: Block diagram of Naïve Bayes classifier.

3.4 *Classification*

In machine learning discipline, classification is considered as a problem of identifying a set of classes which new instances belong. It's based on trained datasets containing observations (or instances) whose group membership is known. For the prediction we use different classifiers. The complete 350 feature sets, consisting of seven sub features are entered into the classifiers. In this research work we use Naïve Bayes, Locally Weighted Learning (LWL), Sequential minimal optimization (SMO), J48 and Kstar(K*) classifiers. The detailed description of each classifier is given below. There are two reasons of choosing these classifiers 1) is a variety of classification ways provide more dimensions to evaluate the proposed approach and 2) These classifies are widely adopted in literature and provide a more accurate results for multiple data sets. SMO is considered to be a highly scalable classifier.

3.4.1 *Naïve bayes*

The Naive Bayes is based on Bayesian theorem and belongs to the family of simple probabilistic classifiers. Bayesian theorem describes the probability of an event, which is based on the preceding information of the conditions that might be related to that event. For example, if cancer is linked with age, then, a patient's age can be used to more accurately assess the probability of the disease (i.e. cancer).

The Naive Bayes works on independence assumptions between the features sets. It absorbs from the train data sets the conditional probability of each attribute A_i on the given class label C see [eq.(1)]. It applies the classification by using the Bayes rule.

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \quad (1)$$

Where

$$P(C|A) = P(A1|C) \times P(A2|A) \ldots\ldots \times P(An|C) \times P(C)$$

The block diagram of the classifier is given in Fig. 2.

The algorithm takes input from the file reader and partitioned (split trained sets from test data) it, which are then transmitted for learning (through Naïve Bayes model) and prediction of the classes. At the end the classifier display the score result.

3.4.2 *Locally weighted learning (LWL)*

Locally weighted learning is also called as memory-based learning or instance-based learning classifiers. It adopts instance-based algorithm for assigning observations a weight which is then further utilized by a weighted observation handler. The instance based methods are typically used databases of example data and compare new data to the database using similarity measures for finding the best match for prediction. Fig. 3 illustrates the conceptual diagrammatical view of instances-based learning models.

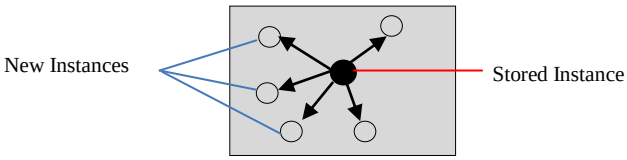

Fig. 3: Diagrammatical view of instances- based learning algorithms.

LWL did prediction by using an approximated local model around the desire point of interest.

3.4.5 *Sequential minimal optimization (SMO)*

Sequential minimal optimization (SMO) is used to train the Support Vector machines (SVMs), which requires a solution of very large quadratic programming optimization (QP) problem [39]. SMO iteratively breaks the large QP problem into a series of smallest possible QP problems [39]. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP

optimization as an inner loop. SMO required linear memory in the training set size, which allows SMO to handle very large training sets. It is more systematic and provides more accuracy rate as compare to the other classifiers. It divides the problem into a series of small sub problems and then solved them analytically.

3.4.3 J48

J48 is based on decision tree technique which is one of a predictive machine leaning techniques. The conceptual view of decision tree technique is given in Fig. 4.

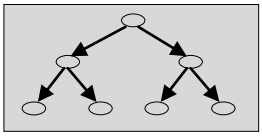

Fig. 4: Conceptual view of decision tree technique

This classifier decides the desire value of any sample through the number of attribute values from the available data. The internal nodes represents attributes the branches characterizes the possible values of the attributes and the terminal nodes indicate the dependent variables.

3.4.4 Kstar

Kstar also known as K* is also an instance-based classifier. It acts on the similarity bases and represents the class of a test instance which is based on the class of train instances similar to it. The similarity checking is done on some similarity function. It uses entropy-based distance function which differentiates it from other instance-based learning techniques.

4. Performance Evaluation

The performance of evaluation for the above mentioned classification algorithms are measured by the standard parameters i.e. Accuracy [eq. (2)], Sensitivity [eq. (3)] and Specificity [eq. (4)]. There formulations are

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FP} \times 100\% \quad (3)$$

$$Specificity = \frac{TN}{FP+TN} \times 100\% \quad (4)$$

Sensitivity (Sn) is used to predict the correct results while Specificity (Sp) predicts incorrect results. At the end Accuracy (Acc) is the degree of correctness between predicted and actual results.

5. Results and Discussion

The availability of complete genetic and demographic data of patients is a rich source of information to predict the glaucoma especially POAG. Due to its complexity many computational tools and machine leaning algorithms are proposed. To validate the results of proposed approach, total 590 cases are taken from which 70% are trained cases while 61% are used for testing. The results from all classifiers are illustrated in Table 5 In the performance measures among the five classifiers SMO shows the highest accuracy rate which is 0.98. These results make it strong optimization technique as compare to other classifiers. The highest accuracy rate implies the number of people predicted as POAG patient WEKA (Waikato Environmental for Knowledge Analysis) tool is used for experiment and prediction. Patient data is gathered from different online sources and given as an input. The statistical values of different classifiers are given in Table 6. Sensitivity indicates pure positive and Specificity represents pure negative rates which are 87% and 65% respectively. The overall system performance is satisfactory. The graphical representation of all classifiers in terms of performance factors is given in Fig. 5.

Table 5: Performance of different classifiers using proposed approach

Classifier	Specificity	Sensitivity	Accuracy
Naïve Bayes	0.73	0.67	0.78
LWL	0.34	0.45	0.56
SMO	0.87	0.65	0.98
J48	0.67	0.56	0.45
Kstar	0.78	0.67	0.56

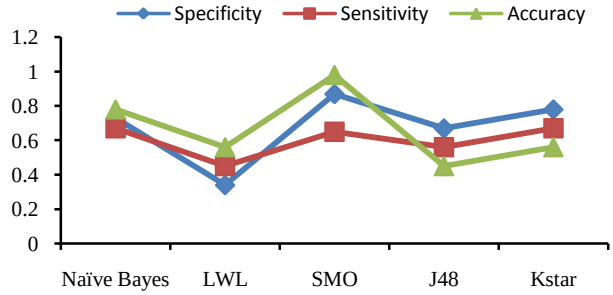

Fig. 5: Graphical representation of results.

6. Conclusion and Future Recommendation

POAG is one of the global eye related disease causes complete blindness. In this paper we classify POAG through genetic and demographic data. For the proposed approach genetic and demographic data sets have been collected from multiple databases. The data sets get structured by using the proposed algorithm. In this study different classifiers are evaluated for predicting the POAG disease using structured data sets as input. For the prediction perspective we use 350 patients data sets on multiple classifies with WEKA as a software tool. It is observed that SMO is providing better accuracy results as compare to other classifiers. The results are shown in table. The comparison of current approaches with our approach is given in Table 7. The accuracy rate of the proposed system is 98%. In this study we use selected genes and demographic information for the predictions but in future large genes with more specific demographic and environmental data sets e.g. accident cases, height, weight, diet etc. will be added for prediction. This research can further be extended by using the SVM, CMIM for feature selection techniques.

Table 6: Statistical values of various classifiers.

Classifiers	TP	FP	Precision	Recall	F-Measure	Roc-Area
Naïve Bayes	0.444	0.4	0.444	0.421	0.667	0.467
LWL	0.444	0.421	0.563	0.197	0.364	0.456
SMO	0.667	0.761	0.571	0.597	0.564	0.761
J48	0.667	0.543	0.341	0.438	0.465	0.634
KStar	0.444	0.533	0.438	0.381	0.431	0.643

Table 7: Comparison of proposed approach with existing techniques.

Approach	Method	Risk Factors	Age	Data Sets Size	Accuracy
Hoover and Goldbaum [40]	Fuzzy convergence of blood vessels	Digital images	-	50	89%
Green et al. [27]	Statistical based	Genetic data, Family history	> 40	467	86%
Zhang et al. [41]	Machine learning approach, Statistical Based	Genetic data	40-80	412	96%
Tokuda et al. [42]	Machine learning approach, Statistical based	Genetic data, Cytokine data	-	115	74%
Eng et al.[43]	Decision tree (DT), Fuzzy logic and Neural network	Genetic data, Demographic data (Age, Medical history, Clinical observations)	> 40	398	80%
Elshazly et al. [22]	Neural Network (NN), Decision tree (DT) and Fuzzy logic	Genetic data, Demographic data (Age, Medical, Family history, Clinical observations)	> 40	398	86%
Moore et al. [44]	Artificial intelligence, Logistic regression	Genetic data, Family history	>40	272	61%
Agarwal et al. [45]	Image processing, Statistical based	Fundus images	-	110	90%
Agurto et al. [46]	PLS	Fundus images, Diabetes, Hypertension, Cataracts, Age, Gender	-	104	95%
Huang et al. [47]	Genomic technique(WES), Statistical based	Genetic Data, Demographic data (Race, Family history)	<40	67	12%
Rao et al. [23]	Statistical based	Genetic data , Race	>40	537	83%
Apreutesei et al. [48]	Neural network	Genetic data, Diabetes, Age	>40	101	95%
Baboolal and Smit [49]	Statistical based	Genetic data , Demographic data (Race, Age, Gender, Diabetes, Hypertension)	18-94	402	87%
Gharahkhani et al. [50]	Pathway-based, Statistical based	Genetic data, Age	-	307	95%
Biggerstaff et al. [51]	Statistical Based	Demographic (Race, gender), Clinical data	>40	334	95%
Nitta et al. [52]	Statistical Based	Genetic data, Clinical data,	10-65	312	75%
Restrepo et al. [21]	Data mining	Genetic and clinical data (Race, Gender)	>40	267	76%
Proposed Approach	Classification	Genetic data, Demographic data (Race, Family history, Gender)	40-80	590	98%

References

[1] A. Youssef and A. Rich, "Exploring trends and themes in bioinformatics literature using topic modeling and temporal analysis", Proc. of the Int. Conf on Long Island Systems, Applications and Technology Conference (LISAT), IEEE, Farmingdale, NY, USA, 2018.

[2] N.M. Luscombe, D. Greenbaum and M. Gerstein, "What is bioinformatics? A proposed definition and overview of the field", Methods Archive, vol. 40, no. 4, pp. 346-358, 2001.

[3] K.A. Shakil and M.Aslam, "Cloud computing in bioinformatics and big data analytics: Current status and future research", Big Data Analytics, Springer, 2018.

[4] NCBI, “Welcome to NCBI”, 2018, Reterived on: 8th Feb., 2018, Url: https://www.ncbi.nlm.nih.gov.

[5] J.S. Hamid, P.Hu, N.M. Roslin, V. Ling, C.M.T. Greenwood and J. Beyene, “Data integration in genetics and genomics: Methods and challenges”, AGE Hindawi Access to Research Human Genomics and Proteomics, vol. 1, no 1, pp. 1-13, 2009.

[6] R.A. Mejía, C. Linares, J. Garrabou, A. Antunes, E. Ballesteros, E. Cebrian, D. David and J. Ledoux, “Combining genetic and demographic data for the conservation of a mediterranean marine habitat-forming species”, PLOS, pp.1-19, 2015.

[7] R.N. Weinreb and P.P.T. Khaw, "Primary open-angle glaucoma”, The Lancet, vol. 363, no. 9422, pp. 1711-1720, 2004.

[8] D. Anitha, M. Suganthi and T.S. Gnanendra, "Evaluation of data mining classifiers for prediction and classification of glaucoma associated proteins", Int. J. Pharma and Bio Science, vol. 9, no. 1, pp. 1-11, 2018.

[9] I. Hecht,A. Achiron, V. Man and Z.Burgansky-Eliash,"Modifiable factors in the management of glaucoma: a systematic review of current evidence", Graefe's Archive for Clinical and Experimental Ophthalmology, vol. 255, no. 4, pp. 789-796, 2017.

[10] E.M. Stone, J.H. Fingert, W.L.M. Alward, T.D. Nguyen, J.R. Polansky, S.L.F. Sunden, D. Nishimura, A.F. Clark, A. Nystuen, B.E. Nichols, D.A. Mackey, R. Ritch, J.W. Kalenak, E.R. Craven and V.C. Sheffield, "Identification of a gene that causes primary open angle glaucoma", Science, vol. 275, no. 5300, pp. 668-670, 1997.

[11] A. Desai, D. Patel, A. Sapovadia, P. Mehta and J. Brahmbhatt, "A study of relation between primary open angle glaucoma and type II diabetes mellitus", Int. J. Res. Med. Sci., vol. 6, no. 3, pp. 997-1001, 2018.

[12] A.M. Williams, W. Huang, K.W. Muir, S.S. Stinnett, J.S. Stone and J.A. Rosdahl, "Identifying risk factors for blindness from primary open-angle glaucoma by race: a case–control study", Clinical Ophthalmology, vol. 12, pp. 377-383, 2018.

[13] P.W.M. Bonnemaijer, C. Cook, A. Nag, C.J. Hammond, C.M.V. Duijn, H.G. Lemij, C.C.W. Klaver and A.A.H.J. Thiadens, "Genetic African ancestry is associated with central corneal thickness and intraocular pressure in primary open-angle glaucoma", Investigative Ophthalmology & Visual Science, vol. 58, no. 7, pp. 3172-3180, 2017.

[14] F. Wang, Y. Li, L. Lan, B. Li, L. Lin, X. Lu and J. Li, "Ser341Pro MYOC gene mutation in a family with primary open-angle glaucoma", Int. J. of Molecular Medicine, vol. 35, no. 5, pp. 1230-1236, 2015.

[15] H.I. Elshazly, M. Waly, A.M. Elkorany and A.E. Hassanien, "Chronic eye disease diagnosis using ensemble-based classifier", Proc. of the Int. Conf. on Engg and Tech. (ICET), IEEE, pp. 1-6, 2014.

[16] E.Oh, T.K.Yoo and S. Hong, "Artificial neural network approach for differentiating open-angle glaucoma from glaucoma suspect without a visual field test", Investigative Ophthalmology & Visual Science, vol. 56, no. 6, pp. 3957-3966, 2015.

[17] T.R. Kausu,V.P. Gopi, K.A. Wahid, W. Domaand S.I. Niwas, "Combination of clinical and multiresolution features for glaucoma detection and its classification using fundus images", Biocybernetics and Biomedical Engineering, vol. 38, no. 2, pp. 329-341, 2018.

[18] S.J. Kim, K.J. Cho and S.Oh, "Development of machine learning models for diagnosis of glaucoma", PloS One, vol. 12, No. 5, pp. 1-16, 2017.

[19] E. Long, P.Wan and Y. Zhuo, "Predicting the real-world future of glaucoma patients? Cautions are required for machine learning", Translational Vision Science & Technology, vol. 6, no. 6, pp. 1-2, 2017.

[20] K.N. Rao, S. Nagireddy and S. Chakrabarti, "Complex genetic mechanisms in glaucoma: an overview" Indian Journal of Ophthalmology, vol. 59, no. Suppl 1, pp. 31-42, 2011.

[21] N.A. Restrepo, E. Farber-Eger, R. Goodloe, J.L. Haines and D.C. Crawford, "Extracting primary open-angle glaucoma from electronic medical records for genetic association studies", PloS one, vol. 10, no. 6, pp. 1-15, 2015.

[22] H.I. Elshazly, M. Waly, A.M. Elkorany and A.E. Hassanien, “Chronic eye disease diagnosis using ensemble-based classifier”, Int. Conf. on Engg. and Tech. (ICET), IEEE, pp. 1-6, 2014.

[23] M.N. Rao, M. Rao and V. Gopala, "A fusion technique to classify glaucoma from fundus images", IIOAB JOURNAL, vol. 7, no. 9, pp. 812-824, 2016.

[24] K.N. Rao, I. Kaur, R.S. Parikh, A.K. Mandal, G. Chandrasekhar, R. Thomas and S. Chakrabarti, "Variations in NTF4, VAV2 and VAV3 genes are not involved with primary open-angle and primary angle-closure glaucomas in an indian population", Investigative Ophthalmology & Visual Science, vol. 51, no. 10, pp. 4937-4941, 2010.

[25] Oracle, "Oracle advanced analytics' machine learning algorithms sql functions", 2018, Reterived on: 8-Feb-2018, Url: http://www.oracle.com/technetwork/database/enterprise-edition/odm-techniques-algorithms-097163.html.

[26] S. Kumar, M.A. Malik, K. Sooraj, R. Sihota and J. Kaur, "Genetic variants associated with primary open angle glaucoma in Indian population", Genomics, vol. 109, no. 1, pp. 27-35, 2017.

[27] C.M. Green, L.S. Kearns, J. Wu, J.M. Barbour, R.M. Wilkinson, A. Maree, T.L. Wong, A.W. Hewitt and D.A. Mackey, "How significant is a family history of glaucoma? Experience from the Glaucoma Inheritance Study in Tasmania", Clinical & Experimental Ophthalmology, vol. 35, no. 9, pp. 793-799, 2007.

[28] B.T. Whigham, S.E.I. Williams, Y. Liu, R.M. Rautenbach, T.R. Carmichael, J. Wheeler, A. Ziskind, X.Qin, S. Schmidt, M. Ramsay, M. A.Hauser and R.R. Allingham, "Myocilin mutations in black South Africans with POAG", Molecular Vision, vol. 17, no. 1, pp. 1064-1069, 2011.

[29] F.S. Philomenadin, R. Asokan, N. Viswanathan, R. George, V. Lingam and S. Sarangapani, "Genetic association of SNPs near ATOH7, CARD10, CDKN2B, CDC7 and SIX1/SIX6 with the endophenotypes of primary open angle glaucoma in Indian population", PloS one, vol. 10, no. 3, pp. 1-12, 2015.

[30] V. Avagyan, N. Ma, J. Papadopoulos , K. Bealer and T.L. Madden, “BLAST+: architecture and applications",Bioinformatics, vol. 10, no. 421, pp. 1-9, 2009.

[31] DAVID, “DAVID Bioinformatics Resources”, 2018, Reterived on 20th Feb 2018, URL: https://david.ncifcrf.gov.

[32] J.Ostell, “The Entrez Search and Retrieval System”, NCBI, 2018.

[33] M. Safran, I. Dalah, J. Alexander, N.Rosen, T.I. Stein, M. Shmoish, N. Nativ, I. Bahir,T. Doniger, H. Krug, A. Sirota-Madi, T. Olender,Y.Golan,G. Stelzer, A. Harel and D. Lancet, “GeneCards Version 3: The human gene integrator”, The Journal of Biological Database and Curration, vol. 1, pp. 1-16, 2010.

[34] A. Daemen, O.Gevaert and B.D.Moor, “Integration of clinical and microarray data with kernel methods”, Proc. of the 29th Annual International Conference of the EMBS, IEEE, Lyon, France, 2007.

[35] O.G. Troyanskaya, K. Dolinski, A.B. Owen, R.B. Altman and D. Botstein, “A Bayesian framework for combining heterogeneous data sources for gene function prediction”, Proc. of the National Academy of Sciences of the United States of America, vol. 100, no. 14, pp.8349-8353, 2003.

[36] A. Abdiansah and R. Wardoyo, “Time Complexity Analysis of Support Vector Machines (SVM) in LibSVM”, Int. J. Comp. Appl., vol. 128, no. 3, pp. 0975 – 8887, 2005.

[37] P. Kumar, S.Henihoff and C.P.Ng, “Predicting the effects of coding non-synonymous variants on protein fucntion using the SIFT algorithm”, Nature Protocols, vol. 4, no. 8, pp. 1073-1082, 2009.

[38] O. Gevaert, F.D. Smet, D. Timmerman, Y.Moreau and B.D. Moor, “Predicting the prognosis of brest cancer by integrating clinical and microarray data with Bayesian Networks”, Bioinformatics, vol 22, no. 14, pp.185-190, 2006.

[39] R.P. Aharwal, “Evaluation of various classification techniques of weka using different datasets”, Int. J. Adv. Res. and Innov. Ideas in Education, vol. 2, no. 2, pp. 2395-4396 , 2016.

[40] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels", IEEE transactions on medical imaging, vol. 22, no. 8, pp. 951-958, 2003.

[41] Z. Zhang, J. Liu, C.K. Kwoh, X. Sim, W.T. Tay, Y. Tan and F. Yin, "Learning in glaucoma genetic risk assessment", Proc. of Annual Int. Conference of the IEEE on Engg. in Med. and Bio. Soc. (EMBC), pp. 6182-6185, 2010.

[42] Y. Tokuda, T. Yagi, K. Yoshii, Y. Ikeda, M. Fuwa, M. Ueno, M. Nakano, N. Omi, M. Tanaka, K. Mori, M. Kageyama, I. Nagasaki, K. Yagi, S. Kinoshita and K. Tashir, "An approach to predict the risk of glaucoma development by integrating different attribute data", SpringerPlus, vol. 1, no. 1, pp. 1-10, 2012.

[43] M. Eng, A.S. Waly and K.Wahba, "A comparison of different prediction models in the progression of ocular hypertension to primary open angle glaucoma", Int. J. Appl. Inform. Sys., vol. 5, no. 3, pp. 30-41, 2013.

[44] J.H. Moore, C.S. Greene and D.P. Hill, “Identification of novel genetic models of glaucoma using the “EMERGENT” genetic programming-Based artificial intelligence system, Genetic Programming Theory and Practice XII, Springer, pp. 17-35, 2015.

[45] A. Agarwal, S. Gulia, S. Chaudhary, M. K. Dutta, R. Burget and K. Riha, "Automatic glaucoma detection using adaptive threshold based technique in fundus image", Proc. of the 38th Int. Conf. on Telecommunications and Signal Processing (TSP), IEEE, pp. 416-420,2015.

[46] C. Agurto, S. Nemeth, G. Zamora, M. Vahtel, P. Soliz and S. Barriga, "Comprehensive eye evaluation algorithm", Proc. of the Medical Imaging 2016: Computer-Aided Diagnosis, Int. Society for Optics and Photonics, pp. 978518-7, 2016.

[47] C. Huang, L. Xie, Z. Wu, Y. Cao, Y. Zheng, C. Pang and M. Zhang, "Detection of mutations in MYOC, OPTN, NTF4, WDR36 and CYP1B1 in Chinese juvenile onset open-angle glaucoma using exome sequencing", Scientific reports, vol. 8, no. 1, pp. 1-8, 2018.

[48] N.A. Apreutesei, F. Tircoveanu, A. Cantemir, C. Bogdanici, C. Lisa, S. Curteanu and D. Chiselita, "Predictions of ocular changes caused by diabetes in glaucoma patients", Computer methods and programs in biomedicine, vol. 154, pp183-190, 2018.

[49] S. Baboolal and D. Smit "South African Eye Study (SAES): Ethnic differences in central corneal thickness and intraocular pressure", Eye, 2018.

[50] P. Gharahkhani, K.P. Burdon, J.N.C. Bailey, A.W. Hewitt, M.H. Law, Louis. R. Pasquale, J.H. Kang, J.L. Haines, E. Souzeau, T. Zhou, O.M. Siggs, J. Landers, M. Awadalla, S. Sharma, R.A. Mills, B. Ridge, D. Lynn, R. Casson, S.L. Graham, I. Goldberg, A. White, P.R. Healey, J. Grigg, M. Lawlor, P. Mitchell, J. Ruddle, M. Coote, M. Walland, S. Best, A. Vincent, J. Gale, G. RadfordSmith, D.C. Whiteman, G. W. Montgomery, N.G. Martin, D.A. Mackey, J.L. Wiggs, S. MacGregor and J.E. Craig, "Analysis combining correlated glaucoma traits identifies five new risk loci for open-angle glaucoma", Scientific Reports, vol. 8, no. 1, pp. 1-12, 2018.

[51] K.S. Biggerstaff, B.J. Frankfort, S. Orengo-Nania, J. Garcia, E. Chiao, J.R. Kramer and D. White, "Validity of code based algorithms to identify primary open angle glaucoma (POAG) in Veterans Affairs (VA) administrative databases", Ophthalmic epidemiology, vol. 25, no. 2, pp. 162-168, 2018.

[52] K. Nitta, R. Wajima, G. Tachibana, S. Inoue, T. Ohigashi, N. Otsuka, H. Kurashima, K. Santo, M. Hashimoto, H. Shibahara, M. Hirukawa and K. Sugiyama, "Prediction of Visual Field Progression in Patients with Primary Open-Angle Glaucoma, Mainly Including Normal Tension Glaucoma", Scientific Reports, vol. 7, no. 1, pp. 1-12, 2017.