Abyssinia Journal of Engineering and Computing Abyss. J. Engg & Comput., Vol.3, No.2, 2023, 27-33

Application of Data Mining Techniques to Developing A Classification Model for Glaucoma Type Identification

Belete Mamo [a*], Tadelech Tsegaw [b],

[a] Department of Information Systems, Kombolcha Institute of Technology, Wollo University, Kombolcha, Ethiopia.
[b] Developmental Policy Analysis, Bonga University, Ethiopia.

DOI:

ISSN : 2788-6239 (Print)
ISSN: 2788-6247 (online)

*Corresponding Author: gerado1273@gmail.com

ABSTRACT

Data mining, also known as Knowledge Discovery in Databases (KDD), is a process that entails extracting valuable, interpretable, and useful information from raw data. Glaucoma, characterized by an elevation in intraocular pressure (IOP), leads to glaucomatous optic neuropathy and subsequent loss of retinal ganglion cells and their axons, ultimately resulting in blindness. Those tasked with treating glaucoma patients may face challenges in accurately identifying the type of glaucoma and prescribing appropriate treatment, often due to subjective decision-making, limited knowledge, and reliance on instrument visualization. These challenges contribute to resource wastage and time-consuming processes. The primary goal of this research is not to completely eliminate the problem but to alleviate biased decisions made by ophthalmologists. This is accomplished by developing an easily accessible method for identifying glaucoma types through the creation of an improved classification model. In this study, data mining techniques are employed to unveil new knowledge based on the collected dataset. Among various data mining classification algorithms, this paper utilizes naïve Bayes, GRIP, J48, and PART algorithms, along with two test options involving complete and selected features. According to the empirical analysis conducted, the PART algorithm, with a 10-fold cross-validation test option using selected features, yielded the highest accuracy result, reaching 71.4%.

Keywords: Machine Learning, Data Mining, Classification, Eye diseases, Glaucoma, Knowledge Base system

1. INTRODUCTION

Glaucoma is a category of eye diseases characterized by progressive optic neuropathy and visual field defect [1] and considered by damage to the optic nerve with corresponding visual field loss. In other words, it is characterized by increased intraocular pressure (IOP), which is responsible for the glaucomatous optic neuropathy involving the death of retinal ganglion cells and their axons [2] which in turn causes damage of the optic nerve, resulting in blindness. In glaucoma, the drainage of aqueous humor through trabeculae is blocked, resulting in increased IOP. When the IOP is above 60 mm Hg, the optic nerve fibers at the optic disk are compressed.

Initially it decreases the visual field, which eventually leads to total blindness. Untreated glaucoma leads to permanent damage of the optic nerve and results in blindness [3]. Typically, patients initially lose the mid-periphery of their visual field, while central vision tends to be involved later. The patients become aware of a functional defect when visual field loss impinges upon or involves central vision [4, 5]. Patients with glaucoma may experience difficulty in identifying faces, steering, reading, noticing objects in their peripheral vision, and adapting to different levels of lighting. Moreover, they are also at an enlarged risk of falls and accidents [6]. Fifteen percent of the world's blindness is attributed due to glaucoma and around 600,000 people go blind annually [2]. In 2013, the number of people with glaucoma worldwide was projected to be 64.3 million [7]. This is expected to rise to 76 million in 2020 and 111 million in 2040 [7]. Africa accounts for 15% of the world's blindness burden due to glaucoma [7]. In Ethiopia, glaucoma is the fifth common cause of blindness which results in an irreversible sight loss for an estimated 62,000 Ethiopians [8, 9]. The increasing prevalence of glaucoma is expected to cause a significant economic burden and poor quality of life [10-12] High glaucoma morbidity among some African communities may be attributed to low trained professionals, low awareness, and under-utilization of eye care service [13–15]. It has been estimated that half of the glaucoma patients are already blind in at least one eye in presentation in Africa [16]. Data itself is nothing, but to process it, is very useful and motivating. The way of identifying glaucoma type is so difficult and to handle through data mining technology by discovering new and hidden knowledge.

2. LITERATURE REVIEW

According to Maıla C et.al. [17] conducted paper entitled automatic glaucoma detection based on optic disc segmentation and texture feature extraction with the major aim of developing an automatic detection method of glaucoma in the retinal image to check whether or not the eye become glaucomatous via classifying the retinal image. To achieve the final objective, the methods used by the researchers were the acquisition of image databases, optic disk segmentation, and texture feature extraction in different color modes. The classifier algorithms adopted by their paper were multiplying perceptron, random committee, random forest, and radial basis function. Finally, the researchers obtained the highest accuracy result of 93.03 % by the MLP. Kinjan Chauhan and Ravi Gulati [18] have conducted paper adopting a data mining approach with the major aim of detecting and diagnosing glaucoma disease using the Perimetry and Optical Coherence Tomography (OCT) images and predicting the diagnosis for progression of glaucoma. In Addition to the image processing techniques they have used for feature extraction, some of the data mining algorithms they adopted were neural network, decision tree, logistic regression, machine learning classifier (MLC), linear discriminant analysis, and support vector machine (SVM). Finally, the SVM and Naïve Bayes algorithms achieved the highest accuracy value, and accordingly, the model is developed for detecting glaucoma based on a decision of whether the glaucoma ailment is absent or present in the eye. The paper was done by Sadaf Malik et al. [19] aimed at developing a classification model for classifying the different types of eye diseases by following a data-driven approach and by adopting machine learning and data mining techniques such as decision tree, naive Bayes, random forest and neural network from which the random forest algorithm achieved the highest accuracy result (86.36%). Finally, the score points of 86.36% have been registered via the algorithm of random forest. The paper was focused on classifying diseases by considering glaucoma as one class label. In [20] paper entitled "classifying chief complaint in eye diseases using data mining techniques", has been conducted to discover new hidden knowledge for classifying the different types of eye disease using the intelligent capability of data mining technologies like

MLP, ANN and naïve baye from which the last two achieved the highest accuracy. Like all the literature reviewed, this paper also uses glaucoma as one of the class labels, and the last target is only classified the eye disease as one of the types. In [21] paper entitled "Data Mining Techniques for Diagnostic Support of Glaucoma using Stratus OCT and Perimetric data for the identification of glaucoma from other diseases in the case of retinal image processing. The Parameters obtained from the Perimetry and Stratus Optic Coherence Test (OCT) have been fed to each technique to find out their performance in terms of accuracy, sensitivity, and specificity by using the data mining technique. The data mining algorithms used for diagnostic classification were Decision Tree, Linear Regression, and Support Vector Machine (SVM). Among those algorithms, the Decision Tree and Linear Regression Model performs much better than other algorithms for the diagnosis of Glaucoma by achieving the highest accuracy of 92.56% and 70.25% respectively. The specificity of Linear Regression and SVM is 97.56% and 96.34% respectively. A paper done in [22] applied machine learning techniques such as fuzzy logic, decision tree based on ID3 algorithms, support vector machine, and k-nearest neighbor. Among those techniques, the last three achieved the highest accuracy result of 85, 80, and 86%. In [23] a paper has been done for glaucoma detection and prediction using data mining classification algorithms namely, KNN and Naïve Baye. The final analysis was decided on a 65% accuracy result scored through the Naïve Baye algorithm. The basic difference with this work is that the related paper is focused on detecting glaucoma disease.

3. MATERIALS AND METHODS

The methodology of the paper encompasses the design and implementation tools employed in research. It involves the utilization of these tools to gather pertinent data within a specific research study. For the execution of this study, the researcher conducted interviews, examined documents, and employed data analysis mechanisms. Primarily, the paper adopts a design science approach with a focal point on generating data to address issues within the field. This approach involves creating new artifacts from the provided dataset, and design science research achieves a deeper understanding of a problem by establishing designed artifacts, among other methods. The study utilizes data mining, classification algorithms, and tools such as RapidMiner for data preprocessing and the WEKA tool for knowledge generation, following the Knowledge Discovery in Databases (KDD) process model.

Data preprocessing:

Data preprocessing is a crucial aspect of data mining techniques used to structure data analysis and transform raw data into a useful and efficient format.

The primary concern of data preprocessing includes handling missing values through methods like ignoring or filling them through various approaches, data cleansing, and data transformation involving normalization and discretization. In this paper, the collected dataset contains missing values, imbalanced classes, and

requires feature selection. Missing values are addressed by replacing them with the most frequent value in the recorded dataset. Specifically, the missing values are either filled with the maximum frequent value in the dataset or are ignored. Feature selection is performed based on information gain values, calculated by assessing the entropy of classes (four classes, namely primary open-angle glaucoma, primary angle-closure glaucoma, secondary glaucoma, and congenital glaucoma). The gain values of each attribute are calculated by subtracting them from the entropy and arranging them in descending order. Another outstanding step in data cleansing involves the discretization of data, specifically applied to an age parameter. Addressing the issue of class imbalance, an enhancement has been implemented through SMOTE analysis. Following the application of SMOTE, the dataset now comprises 4166

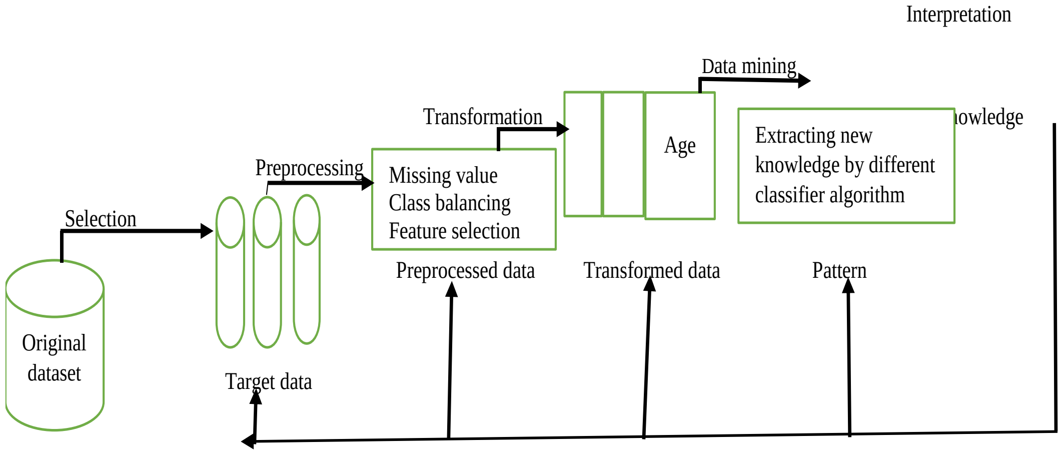

Fig 1. Knowledge discovery in databases (KDD) process

instances, featuring 18 attributes, and the selection of features was executed based on their information gain values. Table 1 presents the data gathered for different disease attributes.

Table 1.
Data acquired about characteristics

S.No	Attribute	Information gain value
1	Headache	0.127
2	Age	0.091
3	Nausea	0.074
4	Sudden eye pain	0.072
5	Frequent eye trauma	0.066
6	Sudden sight loss	0.058
7	Long sight effect	0.053
8	Itchy	0.047
9	Vomiting	0.040
10	Tear percolate	0.036
11	Living area	0.035
12	Eye domineer	0.025
13	Eye hoodness	0.023
14	Cloud cornea	0.011
15	Sex	0.006
16	Eye casing distant	0.006
17	Blurred vision	0.005
18	Light sensitivity	0.003

The suggested framework:

The initial step involves collecting raw data from the designated organization, namely BMH. However, this data is initially in manual format, and the researcher organizes these manual records into a Microsoft Word Excel-friendly format. Following data acquisition, preprocessing activities, such as handling missing values, feature selection, and data discretization, are necessary. Upon completing the pre-processing tasks, the model is developed through an empirical analysis of four distinct classification algorithms: Naïve Bayes, Jrip, J48, and the PART algorithm. Two basic test options, namely 10-fold cross-validation with default values and a percentage split of the dataset with an 80/20 ratio, are employed. The selection of these four classification algorithms is justified by their superior accuracy compared to other algorithms, as well as their simplicity and interpretability in outcome analysis and identification of hidden patterns. The model development attributes are chosen either in their entirety or through a selected attribute subset. From the four algorithms, the one with the highest accuracy result is chosen by comparing others based on two test options, considering either the entire set of attributes or selected attributes. Ultimately, a single model is selected based on the highest accuracy result, considering either 10-fold cross-validation or an 80/20 percentage split, using either the entire set of attributes or selected attributes.

3. RESULTS AND DISCUSSION

To execute the experiment effectively, a comprehensive understanding of features along with their information gain values is essential. The prioritization of attributes for classifying glaucoma is achieved through the WEKA tool after preprocessing the data with RapidMiner. The glaucoma dataset obtained from the specified area (Borumeda Hospital) was transformed into ARFF format, making it compatible with the WEKA tool.

Test options: For experimentation, the researcher opted for two fundamental test options involving four different algorithms. The dataset was converted into an ARRF (Attribute-Relation File Format) file for software analysis.

K-Fold cross-validation: Cross validation is a resampling process that evaluates machine learning models on data sections. The researcher employed a 10-fold cross-validation with its default value, dividing the dataset into ten folds for analysis.

Percentage split: This test option involves partitioning the dataset into training and test sets based on a percentage. In this paper, an 80/20 splitting test option was chosen, utilizing 80% of the dataset for training and the remaining 20% for testing.aset into ten folds for analysis.

Algorithms and Processing Actions Conducted in the Experiment:

Naive Bayes algorithm: Naive Bayes is a classification algorithm based on Bayes' theorem with strong, naive dependence-free assumptions. It is used for classification in machine learning, considering conditional probabilities of features attributed to a class, with feature selection methods influencing the chosen features.

PART algorithm: PART (Projective Adaptive Resonance Theory) is a separate and conquer rule learner that produces sets of decision list rules. It builds decision trees iteratively, assigning classes based on rule matching, and selects the best leaf into a rule.

J48 algorithm: J48 is an extension of ID3 with added features like handling missing values, decision tree pruning, continuous attribute value ranges, derivation of rules, etc. The WEKA tool provides tree trimming options to address potential overfitting, ensuring a balance between flexibility and accuracy.

Jrip algorithm: Jrip (Repeated Incremental Pruning to Produce Error Reduction - RIPPER) is a popular algorithm utilizing sequential covering algorithms to create ordered rule lists. It examines classes in increasing size, producing rules for each class, and repeats the process until all classes are covered. The comprehensive outcomes of the test mode under various algorithms are presented in Table 2.

Table 2.
Overall results of the experiments

Test mode with attribute	Classifier algorithms				
	Performance metrics	Jrip	Naive aye	J48	PART
10 fold cross-validation/ whole attribute	Accuracy	63.5%	50.3%	68.5%	68.3%
	Precision	65%	50%	68%	68%
	TPR	65%	50%	65%	68%
	FPR	12%	16%	10%	10%
	F-score	63%	49%	68%	68%
	ROC area	79%	74%	83%	80%
Percentage Split 80/20/ whole attribute	Accuracy	62.6%	47.6%	68%	66%
	Precision	64%	49%	68%	66%
	TPR	52%	47%	68%	66%
	FPR	12%	17%	10%	11%
	F-score	62%	46%	68%	66%
	ROC area	83%	77%	84%	82%
10-fold cross validation/ selected attribute	Accuracy	67.5%	55.6%	69%	74.7%
	Precision	68.5%	55.8%	71%	70.4%
	TPR	67.5%	55.7%	71%	70.4%
	FPR	11.2%	15.6%	9%	9.9%
	F-score	66.75%	54.4%	71%	72%
	ROC area	84%	71%	76%	79%
Percentage Split 80/20/ selected attribute	Accuracy	65.9%	56.4%	70%	69.4%
	Precision	76%	56.7%	71.2%	69.7%
	TPR	65.9%	56.5%	70.8%	69.4%
	FPR	11.5%	15.6%	9.5%	10%
	F-score	65.7%	55.5%	71.2%	69.7%
	ROC area	83%	72%	75.7%	81%

4. CONCLUSION

In conclusion, by leveraging the dataset acquired from the health organization (BMH), we have gained valuable insights, achieving a notable accuracy rate of 71.4% through the application of the PART algorithm in identifying various types of glaucoma. This identification process aligns with the treated types of glaucoma available in Borumeda Hospital, incorporating knowledge derived from experts and existing documents. The final model has received a favorable acceptance rate of 82.2% from end users, underscoring its potential effectiveness when implemented within the organization. Looking ahead, this research reflects the researcher's perspective and guidance. Subsequently, other researchers may refine and update this study by adopting a country-based approach, gathering datasets from diverse health organizations. They may choose to exclusively utilize the PART algorithm based on its final accuracy or explore the possibility of combining algorithms (ensemble) for optimal accuracy. It is worth noting that the research follows the Knowledge Discovery in Databases (KDD) process rather than opting for a hybrid approach, which may prove beneficial in future endeavors.

REFERENCES

[1] S.T.Simmons, G.Cioffi, R.Gross., "Basic and clinical science course, section 10 Glaucoma", American Academy of Ophthalmology, San Franscisco, 2017.

[2] World Health Organization, Global data on visual impairment, Geneva,, WHO, 2012. http://www.who.int/about/licensing/copyright_form/en/index.htm

[3] K.Sembulingam and P.Sembulingam,"Essentials of medical physiology", 6th edition, Jaypee Brothers Medical Publishers, 2012.

[4] H.D.Jampel, "Glaucoma patients' assessment of visual function and quality of life transactions of the American ophthalmological society", Trans Am Ophthalmol Society, 2001.

[5] P.Gutierrez, M.R.Wilson, C.Johnson, M.Gordon , G.A.Cioffi and R.Ritch, " Influence of glaucomatous visual loss on health-related quality of life", Arch Ophthalmol Society, 1997.

[6] A.Heijl, "Delivering a diagnosis of glaucoma: considering the patient eyes", Acta Ophthalmol Scand Suppl., 2001.

[7] Y.C.Tham, X.Li, T.Y.Wong, H.A.Quigley, T.Aung and C.Y.Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040, A systematic review and meta-analysis", Ophthalmology, 2014.

[8] Y.Berhane, A.Worku, A.Bejiga and L.Adamu, W.Alemayehu and A.Bedri, "Prevalence and causes of blindness and low vision in Ethiopia. Ethiop J Health Dev, 2007.

[9] A.T.Giorgis, "Raising public awareness of glaucoma in Ethiopia" Community Eye Health Journal, 2012.

[10] M.C.Cypel, N.Kasahara, D.Atique, M.P.Alcântara and F.S Seixas, "Quality of life in patients with glaucoma who live in a developing country", Int Ophthalmol. 2004.

[11] S.Do, L.Hans, "Report of the rapid assessment for avoidable blindness in Cambodia", National Program for eye health, 2007.

[12] F.Cristina, S.Kafi, C.Otavio, P.Dave, C.Jonathan and H.Marcelo, "Burden of disease in patients with glaucoma in Brazil: Results from 2011 –2012 National health and wellness survey", Milan, Italy; 2015.

[13] S.Nwosu, "Patients knowledge of glaucoma and treatment options", Niger J Clin Pract., 2010.

[14] B.O.Adegbehingbe, L.Bisiriyu , "Knowledge, attitudes, and self-care practices associated with glaucoma among hospital workers in Ile-Ife", Osun State, Nigeria. Tanzan J Health Res., 2008.

[15] A.Tenkir, B.Solomon and A.Deribew, "Glaucoma awareness among people attending ophthalmic

outreach services in southwestern Ethiopia", BMC Ophthalmol. 2010.

[16] U.Altangerel, H.S.Nallamshetty, T.Uhler, J.Fontanarosa, W.C.Steinmann and J.M.Almodin, "Knowledge about glaucoma and barriers to follow-up care in a community glaucoma screening program" ,Can J Ophthalmol. 2009.

[17] Ma´ıla Claro, Leonardo Santos, Wallinson Silva Fl´avio, Arau´jo Nayara Moura, "Automatic Glaucoma detection based on optic disc segmentation and texture feature extraction", clei electronic journal, vol. 19, No.2, August 2016.

[18] Kinjan Chauhan and Ravi Gulati, "A proposed framework for diagnosis of Glaucoma - A data mining approach" International Journal of Engineering Research and Development, vol. 3, Issue5, August 2015.

[19] Sadaf Malik, Nadia Kanwal and Mamoona Naveed Asghar, "Data driven approach for eye disease classification with machine learning", International Journal of Engineering Research and Development, July 2019.

[20] L.Archana Rane and P. D. Mahajan, "Classifying chief complaint in eye diseases using data mining techniques" International Journal of Engineering Research and Applications, March 2012.

[21] Kinjan Chauhan, Prashant Chauhan and Anand Sudhalkar, "Data mining techniques for diagnostic support of Glaucoma using stratus OCT and perimetric data" International Journal of Computer Applications, oct 2016.

[22] Samina Kahalid, Tehmina Khalil and M.Adeel, Syed, "Machine learning techniques for glaucoma detection and prediction", Science and Information Conference, Aug 2014.

[23] Ritu Sindhu, "Data mining techniques for glaucoma detection", International Journal of Advanced Research in Electronics and Communication Engineering, June. 2018.