BERKELEY PUBLICATION

06.30.2021 AND RESEARCH INTERNATIONAL

Journal of African Sustainable Development Vol. 16 No.2

INTELLIGENT OPTICAL HEALTH CARE ANALYSIS AND BLINDNESS PREDICTION SYSTEM USING TANIMOTO DISTANCE BASED CLUSTERING (TDB-C) MACHINE LEARNING MODEL.

A.D. ADENIYI[1], N. S. AJOGE[2]. A. N. ABDULLAHI[3]

[1,2]Department of Computer Science, College of Science and Technology, , P.M.B. 2021, Kaduna Polytechnic, T/Wada, Kaduna, Nigeria. [3]Department of Education (Tech), College of Technical and Vocation Education, Kaduna Polytechnic, Kaduna, Nigeria

ABSTRACT

In recent years, medical errors have become a universal matter of concern to the international society. The quantity of medical errors due to human factor is becoming incredible to the extent that it has already become the fifth lethal. To alleviate this challenge, this paper proposes the design and realization of an automated optical health decision support and blindness prediction system using a simple but promising hybride predictive model by suitably combining the Tanimoto distance measurement with clustering technique. The proposed system is capable of overcoming scalability, computational complexity noisy data and low dimensionality attribute challenges. The proposed predictive engine is

Introduction:

Glaucoma, Diabetes retinopathy, uncorrected refractive error; uncorrected cataract and age related degeneration are among the first five major global causes of moderate to severe vision impairment Worldwide. According to world health organization (WHO) report WHO(2018), an estimated 253 million people live with vision impairment worldwide. 36 million are blind and 217 million have moderate to severe vision impairment. Approximately, 26.3 million people in Africa region have a form of visual impairment,

ISSN: 1334-4977

implemented through the use of an in- house developed PHP program experiment with XAMP/Apache HTTP server as hosting sever with MySQL application at the back end. The performance of the proposed system is compared with three baseline methods which are: the Euclidean distance K-Nearest neighbor (ED-KNN), the Case Based Reasoning (CBR) and the Traditional Clustering algorithms (TCL). The result shows an excellent performance of our system, with precision rates and predictive quality of equal to or greater than 75%. Therefore the proposed system is capable of providing accurate and efficient predictions to patients and the physician online, real time consistently and at any time.

Keywords: *Tanimoto, clustering, predictive, machine learning, optical health, Decision Support , Blindness.*

Of these, 20.4 million have low vision and 5.9 millionare estimated to be blind. It is estimated that 15.3% of the world's blind population reside in Africa. Pascolini and Mariotti (2010). However, it is estimated that due to the population growth and ageing, there could be 115 million blind people worldwide by the year 2050. Bourne, Flaxman, Braithwaite, Cicinelli, Das and Jonas, (2017).

This study presents the design and implementation of an online, intelligent optical health care analysis and blindness prediction system using Tanimoto distance based clustering machine learning algorithm. The study aimed at developing an automated online application that will be accessible by individual and medical personnel to determine the risk factor for blindness and the level of risk of patients with different optical challenges. The contribution of this work focused on the following aspects; First, we proposed a novel decision support algorithm based on Tanimoto distance clustering (TDB-C) model, with attention focused on improving the performance of the traditional clustering algorithm, being an unsupervised learning algorithm. We proposed a Semi-supervised clustering technique using distance based technique.

Second, we proposed the construction of an intelligent online optical health decision support and blindness prediction system implemented using an experimental website developed with PHP programming language, with XAMP/Apache HTTP server as hosting sever and MYSQL for database management. The proposed optical risk calculation will aggregate risk factor data from an individual, build a personalized risk level to a varying degree to the user online and in real time basis. The proposed optical risk calculator and decision support system uses optical medical profile supplied by the individual user/ patients or physician (risk factor questions) such as diagnosis data on diabetic retinopathy, cataract, retinal errors, glaucoma, macular degeneration etc. to predict the likelihood of the patients becoming blind. The system is capable of answering complex "what if" question which is lacking in the traditional decision support system.

Finally, the novel Tanimoto distance based clustering (TDB-C) algorithm that serves as basis for the development of the machine learning and predictive system will be presented. More so, the performance evaluation of the entire system will be carried out with a thorough presentation of the experimental results. The performance of the proposed system will be compared with three baseline methods which are: the Euclidean K-Nearest neighbor (ED-KNN), the Case Based Reasoning (CBR) and the Traditional Clustering (TCL) algorithms. This is to justify the rationale behind the selection of the TDB-C model for the risk calculation and predictive system. The proposed Optical risk calculator, when implemented will assist the physician and optical patients with those symptoms to be aware of their risk level and to help the public health operators to set priority for public health interventions and to minimize blindness in those patients and in the society at large.

REVIEW OF RELATED WORKS

This section provides an insight into various studies conducted by various scholars in the field of machine learning and predictive system. The section also showcases a resume of the past and present status of the problem

delineated by concise review of previous studies into closely related problems. The review is organized systematically below:

Overview of related Machine learning techniques

Hurwitz and Kirsch (2018), described Machine learning as a form of AI that enables a system to learn from data rather than through explicit programming. However, Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes. Scholars in the field of machine learning, data mining, and predictive system have identified different techniques for machine learning operation. These techniques include, K-Means, Case Based Reasoning, Bayesian, SOM, ANN, Random Forest, to mention just a few (Bhosale and Ade 2014; Adeniyi, Wei and Yongquan 2015; Wang and Tan 2011; Yaghini, Khoshraftar, and Fallahi 2012; Hssina , Merbouna , Ezzikouri and Erritali 2014; Han J, Kamber 2006) The challenge posed by the availability of these numerous techniques is in the choice of most suitable technique for the prediction problem at hand Adeniyi, Wei and Yongquan (2016).

Soni, Ansari and Sharma(2011) adopted the naive Bayesian, KNN and Decision tree algorithm for analyzing their heart disease data sets. They do found out that, Decision tree can handle high dimensional data, require no domain knowledge with easy to read and interpret result. However, the Decision tree requires training data to residing in memory becomes ineffective as it increases memory consumption due to the swapping of training tuples in and out of the main memory and cache memory. Our approach tries to overcome this challenge, since it's capable of handling data set that are too large to fit into the main memory.

The Naive Bayesian techniques adopted by Sonie et.al(2011), shows capability in handling complex real word problem but with low degree of accuracy due to the use of class conditional independence assumptions and unavailability of probability data. This is usually not the case while using TDB-C model. Aha (1991) and Adeniyi, Wei and Yongquan (2018). in their work adopted the Case Based Reasoning but found out that CBR suffers

several setbacks due to noisy cases and poor similarity functions Bichindaritz (2015). These are overcome by the present TDB-C techniques. A review of the Random Algorithm by Biau and Scornet (2015) shows that the RF algorithm may over fit noisy data set, the modelers may lose control of the operation of models. The proposed KNDB-C algorithm has the capability to overcome these challenges, through increase robustness, tolerance to noise data, reduced storage requirement with high predictive accuracy and effectiveness.

A review of related predictive system

Poplin et al. (2018), carried out prediction of cardiovascular risk factor from retinal fundus photography using deep learning models, their experimental result shows that deep learning can extract new knowledge from retinal fundus images. However, their approach lacks an established technique for selecting these factors from retinal images. This challenge is overcome through the adoption of the proposed TDB-C techniques. Soni et.al. (2011), in their work carried out an overview of heart disease prediction using three different supervised machine learning algorithms ie. Naive Bayesian, K-NN and Decision tree. They found out that the performance of the algorithms improved after applying the genetic algorithm to reduce the actual data size for their predictive system. This approach is similar to the present work except that we used a more salable approach.

Rivers et. al. (2011), proposed the use of Nave Bayesian, SVM and Decision tree, for predicting accidents and incidents in two companies. Their work shows a better result but with low degree of accuracy due to the class conditioning independence assumption used in the Bayesian techniques. This is not the case when using the present system since no assumption is made. Anbarasi and Iyengar (2010), proposed a method of performing heart disease prediction using Decision tree, Nave Bayesian and Clustering technique's with genetic algorithm to reduce the actual data size. However, their method shows possibility of high value attribute dominating the low

value attributes, therefore the need to normalize the attribute values. This problem and others have been taken care of by our proposed TDB-C model.

METHODOLOGY

In this section, we briefly describe several different methods that generally come into consideration during the design and realization of a predictive model, specifically, the proposed automated optical risk calculator.

Experimental design

The proposed optical risk calculation and blindness predictive model is trained on historical optical medical record data extracted from the Nigerian National Eye Centre, Mando, Kaduna, Nigeria for a period of twelve years (2008-2019). The historical optical medical record of the patients extracted is kept anonymous and secured. The data sets is made up of cohorts of patients diagnosed of various eye diseases which included cataract, macular degeneration, glaucoma, diabetes retinopathy, refractive errors etc. The diagnostics records were studied and was discovered that some of the patients are totally blind, some with partial blindness, some with sever vision impairment, some with Moderate vision impairment and some with Normal vision. These are used to categorize the records into different risk level such as fatal risk, high risk, low risk and No risk. The data set is preprocessed to scale down unwanted features using the proposed PFRF scaling technique before applying the proposed TDB-C model for the estimation of potential risk factors for 'fatal risk' ,'high risk', 'low risk' and 'No risk' by computing stimulates between the patients optical medical records historical data.

Data collection

The dataset used for this experimental optical risk calculator model was randomly selected from the optical patients' medical history of 13,280 anonymous patients of the Nigerian National Eye Centre, Mando, Kaduna Nigeria. The records are of patients who were diagnosed of various optical ailments over a period of twelve (12) years (2008-2019). The sample data

set includes 4,520 diagnostic record of cataract, 1,192 of refractive errors, 1,170 0f age related macular degenerations, 2,128 Glaucoma and 2,150 diagnostic report of Diabetics retinopathy and 2,120 of other causes. The raw data extracted were cleansed to eliminate irrelevant/noise data, data mart was developed. The result of the experiment was stored in a database, using my MYSQL DBMS; the implementation of the proposed system was done using in house developed PHP program, with XAmp /Apache HTTP server as hosting server.

A patient/ physician can personally interact with the system online and on real time basis by entering the risk factor questionnaire to obtain risk level estimate for the patient. The proposed TDB-C model compare this information with the content of the data mart. The result was sorted in ascending order of their Tanimoto distance similarity value nearest to the current patients, in order to predict the current users risk level.

The proposed Tanimoto Distance Based Clustering (TDB-C) model

Tanimoto similarly is a statistics tool used for measuring the similarity and diversity of a sample set. It is defined as the size of the intersection divided by the size of the union of the sample set. (Zhang, Vogt, Maggiora and Bajorath ,2015; Bajusz, Rácz and Héberger 2015). The Tanimoto coefficient ranges from 0, when the tuple have no feature in common to 1, when the tuples are identical.

Clustering is a technique of grouping a set of physical or abstract data object into classes of similar object. Han and Kamber (2006). In the field of data mining and machine learning, clustering is considered to be an unsupervised learning, since it does not rely on predefined classes and class labeled training examples. It learns by observation rather than learning by examples. The traditional clustering algorithm are marred with the problems of clustering scalability, inability to deal with different types of attributes, inability to deal with noisy and high dimensionality data, inability to detect clusters with arbitrary shape etc.

To overcome some if not all of these challenges, the present system proposes a Semi-supervised clustering technique using distance based

technique. It uses the Tanimoto distance measure, these methods transform a cluster task into classification task, it classifies set of points to be clustered into one class label or the other.

To quantify the similarity of tuples on the basis of their attributes, we applied Tanimoto similarity functions;

Given two tuple X and Y;

Let X be a point in a training cluster with f features: f_{i1},f_{i2},f_{i3}, ……f_{in}

Let Y be a test point with f features: f_{j1},f_{j2},f_{j3}, ……f_{jn}

Let K be the total number of the training points in the clusters: i = 1,2,3……k

Let n be the number of features: j=1,2,3….k

The Tanimoto distance between a training point and a test point can be derived as follows:

$$TD_{(X,Y)}= \frac{X \wedge Y}{(X \vee Y)-(X \wedge Y)} \qquad \text{Equation (1)}$$

Expressing equation (1) in general term, The Tanimoto distance between two points such as;

$X_{fi} = (f_{i1},f_{i2},f_{i3}, \ldots\ldots f_{in})$ and $Y_{fj}= (f_{j1},f_{j2},f_{j3}, \ldots\ldots f_{jn})$ will be:

$$TD_{(Xi,Yj)}= \frac{Xi \wedge Yj}{(Xi \vee Yj)-(Xi \wedge Yj)} \qquad \text{Equation (2)}$$

Where: i= j = 1,2,3, …. N, $X \wedge Y$= number of attributes shared by tuple X and Y

$X \vee Y$= number of attributes present in both tuple X and Y

Equation (2) takes the number of attributes shared by X and Y, $(X \wedge Y)$ divide it by the number of attribute possessed by X and Y, $(X \vee Y)$ minus the number of attributes shared by X and Y,$(X \wedge Y)$, this gives us the Tanimoto distance between the two points X_i and Y_j. The equation(2) is mostly applicable to numerical attributes, for categorical attributes, the nominal scale technique can be adopted. See Adeniyi Wai and Yongquan,(2016) for detail description of nominal scale technique.

The TDB-C technique partition the given database of n object or tuple into k partitions, each partition represents a cluster $K \leq n$ i.e. and it simply classifies the data into the k group containing at least one object and each object belongs to exactly one group. The TDB-C predictive model computes

the Tanimoto distances between a given test tuple and the different point in the corresponding training tuples in each cluster-k. The process continues until all the training points closest to the test tuple /points are captured, predictions are then made based on the closest Tanimoto distance to the given test tuple/points.The TDB-C assign to the test point/tuple, the major category label of its k-nearest training clusters. Algorithm Listing 1, shows the algorithm for the proposed TDB-C model.

Algorithm Listing 1: TDB-C algorithm

Input:

Let K = Number of clusters: Let D= Data sets with n objects d_{i1},d_{12}, d_{13}, ……d_{in}

Output; A set of K clusters

*predictive results

//Section to partition the training data set into clusters by the similarities in their classes according to the data collected

Method:

Select K objects from D at random as initial clustering center

2: Repeat:

(re) Assign each object to the cluster with the most similar objects

Until the end of D

// Section to the cluster of an unknown tuple

Input D of unknown cluster (i.e. given a new data point)

Let k=1 ≤ k≤ n: Let i= j=1

Do until (nearest cluster found)

Select a given test point (di) from the data set

Retrieve a point at random from cluster K_i

Compute the Tanimoto distance between the given test Point (x_i) and the selected point K_i from the cluster Y_j Using the expression: $TD_{(Xi,Yj)}= \frac{Xi \wedge Yj}{(Xi \vee Yj)-(Xi \wedge Yj)}$

If (i<k) Then: Let x_i be a member of K-nearest cluster

Else: If(x_i is closer to k_i than any previous point) Then,

ISSN: 1334-4977

Swap x_i for the previous closest cluster i.e add x_i to the top of the member of k-closest clusters
End if: increment I by 1 (i.e. i=i+1)
End do:
Classify $x_{i(d)}$ in the cluster of minimum Tanimoto distance
Else: classify x_i (d) in the majority cluster
End if: End

Application of the Proposed Tanimoto Distance Based Clustering (TDB-C) Model To predict user's Optical Risk level On our Patient's Optical health Database

Example:
Given an experimental patient optical medical history of eye diseases diagnosis as a vector with ten (10) attributes representing different risk factor assessment: **MDG, GLC, CTR, DBP RFE ICE**, **WTR**, **TSN**, **DVS** and **MSE** with diagnosed patients represented by d_1,d_2,d_3,d_4, ….. d_{10} as class, as shown in Table 1:

Table 1: The Optical Patients Risk factors Case-Based class labels training tuples

Patients	MDG	GLC	CTR	DBP	RFE	ICE	WTR	TSN	DVS	MSE	Risk Function/Levl
D_1	1	0	0	1	0	1	0	0	0	0	Low Risk
D_2	1	0	1	1	1	0	1	1	1	1	Blindness Risk
D_3	1	0	0	1	1	1	1	0	1	0	High Risk
D_5	1	0	1	1	1	1	1	0	1	0	High Risk
D_6	1	0	1	1	1	1	1	0	1	1	Blindness Risk
D_7	1	0	1	0	1	0	0	1	0	0	Low Risk
D_8	1	0	1	1	1	1	1	0	1	0	High Risk
D_9	1	1	0	1	1	0	1	1	1	0	High Risk
D_{10}	0	0	0	1	0	0	1	0	0	0	No Risk
D_4	1	0	0	0	1	0	1	0	0	0	?

Assuming the vision function/ level of vision impairment of patient's d_4 is unknown.

ISSN: 1334-4977

To predict the eye impairment level of patient d_4, we have to compute the Tanimoto distance between the patient d_4 and all other patients in the clusters by applying the equation (2):

$$TD_{(Xi,Yj)} = \frac{Xi \wedge Yj}{(Xi \vee Yj) - (Xi \wedge Yj)}$$

Where: $X \wedge Y$= number of attributes shared by tuple X and Y

$X \vee Y$= number of attributes present in both tuple X and Y

Given that:

$d_4 = \{1,0,0,0,1,0,1,0,0,0\}$ and $d_1 = \{1,0,0,1,0,1,0,0,0,0\}$, d_2= [1,0,1,1,1,0,1,1,1,1]

The distance between patient d_4 and patient d_1

$$= \frac{10}{(20)-10} = \frac{10}{110} = 1.000$$

The distance between patient d_4 and patient d_2

$$= \frac{5}{(20)-5}, = \frac{5}{15}, \ 0.333$$

The same process will be repeated for patient's d_4 and d_3, d_4 and d_5, etc. These computations will produce a stem of data as shown in Table 2. The patient's cases are sorted in ascending order of their distance to patients' d_4.

Table 2: Patients cases sorted according to Tanimoto distance to patient d_4.

Patients'	Vision Function	Distance to D_4
D1	Low Risk	1.000
D7	Low Risk	0.818
D10	No Risk	0.818
D3	High Risk	0.539
D5	High Risk	0.429
D8	High Risk	0.429
D9	High Risk	0.429
D2	Blindness Risk	0.333
D6	Blindness Risk	0.333

ISSN: 1334-4977

The TDB-C model simply pick the patients with the maximum of Tanimoto distance to d_4 (the first from the top to the list) and use its cluster label of "Low Risk" to predict the vision function of patient d_4 and therefore, predict a similar vision impairment a level of "Low Risk", cluster as in patient D_1, because it shows to be the majority label in the top three Clusters.

SYSTEM EVALUATION AND ANALYSIS OF RESULT

In this section, we evaluate our proposed Tanimoto Distance based cluster (TDB-C) predictive system in term of predictive accuracy and speed, using the Receive operating characteristics (ROC) curves. The ROC curves shows the tradeoff between the positive rate (proportion of correctly predicted cases) and the false rate(proportion of incorrectly predicted as positive).

Evaluation Data set

The present system uses the internal data set for the purpose of evaluation. We adopted the collected optical patients' medical diagnosis data collected from the Nigeria National eye centre, mando, Kaduna. The collected data were divided into five parts; four parts were used as training set and the remaining one was used as a test set. We assume that the optical risk level of the testing set is unknown while that of the training set was considered known, the risk level of the training set is used to infer the unknown,

To access the performance of the predictive system using the offline evaluation method, we compare the performance of the present TDBC model with three other baseline methods which include: the traditional clustering method, the Case Based Reasoning (CBR) method and the Euclidean distance KNN technique ED-KNN. We adopted the F-measure evaluation technique to measure the predictive accuracy of the TDB-C method. Adeniyi et.al. (2016).

F- Measure is the harmonic means of precision and recall i.e.

$$F = 2\frac{\text{Precision.Recall}}{\text{Precision} * \text{Recall}}$$

Precision (P) $= \frac{tp}{tp+fp}$ i.e. number of correctly predicted divided by number of all returned predictions

ISSN: 1334-4977

Recall(R) $=\frac{tp}{tp+fn}$ I,e, number of correct prediction divided by the number of all known interest supposed to be discovered.

Where: Tp = True position, Fn = False negation Fp = False position, Tn=True negation

Figure 1 shows the result of experiment in F-Measure using our optical patients' data set.

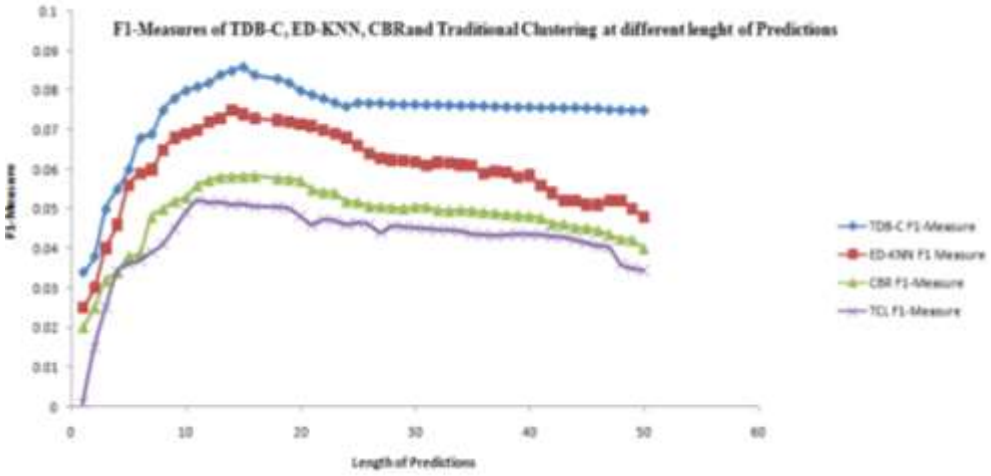

Figure 1: F1-Measures of TDB-C, ED-KNN, CBRand Traditional Clustering at different lenght of Predictions

Analysis of Results.

We have presented a novel algorithm that is efficient, intuitive, faster, scalable and less error phrone. Our experimental result shows that our proposed TDB-C algorithm significantly outperformed other machine learning algorithm such as the traditional clustering, the Case Based Reasoning (CBR) and the Euclidean distance KNN in term of speed and accuracy.

In the previous section, we established that classes with minimum Tanimoto distance to the unknown test tuple will the predicted as in our example. To this effect, an in-house software was developed using PHP programming language and MYSQL DBMS used at the back end for creating and managing the data base with Apache/XAMP HTTP server as hosting server to implement the TDB-C algorithm. Figures 2 and 3 shows sample interface from the optical risk calculator software developed to implement the TDB-C model.

ISSN: 1334-4977

Figure 2: The home page showing basic operations that can be performed by the user

Figure 3: Window presenting active user interface with response to optical risk factor questioning with the Analyze optical risk button

Figure 2is the home page showing basic operations that can be performed by the user such as evaluating optical risk level. Figure 3 is a window presenting active user interface with response to optical risk factor questioning with the Analyze optical risk button. When the user click the Analyze optical risk button, the system build the user profile and predict her risk level using the TDB-C algorithm, then present the risk level to the individual user online and in real time basis. The developed system can be implemented online by simply uploading the application on a web server and can be involved anytime with any web browser.

Discussion

The result of our experiment shows a significant improvement of our proposed TDB-C method over the baseline method studied. Generally, the value of the F-Measure rises rapidly before attaining a peak point after which they go down slowly. This shows that during the increase, the precision is nearly stable while recall increases and during the decrease, the precision decrease and while the recall almost constant. Figure 1 shows the performance of our proposed TDB-C and the baseline method i.e. Traditional Clustering algorithm, CBR and ED- KNN. The result shows that the TDB-C model significantly outperformed the baseline methods.

The result shows that the CBR, Traditional Clustering and the ED-KNN have lower F1-Measure when applied on our dataset; they performed poorly compared to

the TDB-C model. We experimented for about 60 different length of predictions under the same experimental settings. It was discovered that the proposed TDB-C has the highest F1-Measure on our data set compared with the baseline methods. At length of prediction lower than 20, the baseline method performed a little better, but the performances becomes poorer at higher length of prediction, therefore resulting to a limited number of accurate predictions. Our algorithm is able to perform better due to the approach adopted in building the TDB-C model i.e. the use of Tanimoto distance measurement etc. Our algorithm is able to overcome, error, computational complexity and scalability problems that plague other machine learning algorithms. Our algorithm is able to scale well for the present predictive system, with ability to handle noisy data and high dimensionally data common with the baseline technique.

To further demonstrate the excellent performance of our predictive system over the baseline method, we compared the run-time of our algorithm with the baseline methods; we recorded the run-time of each algorithm at different length of prediction. The result shows that the TDB-C have the lowest run-time, it executes faster than the baseline methods as shown in figure 4.This therefore makes the TDBC algorithm to be more computationally efficient than the baseline methods. Hence, this makes the TDBC model more suitable in domain with large, noisy and high dimensionally datasets.

Finally, the TDBC model is capable of producing an efficient, initiative, faster, predictable system with ability to overcome, noisy, inaccurate, and computations complexity challenges common to some machine learning algorithms. This makes the TDBC model to be able to outperform the CBR, Traditional clustering and ED-KNN techniques when it comes to a difficult predictive text in the large data set.

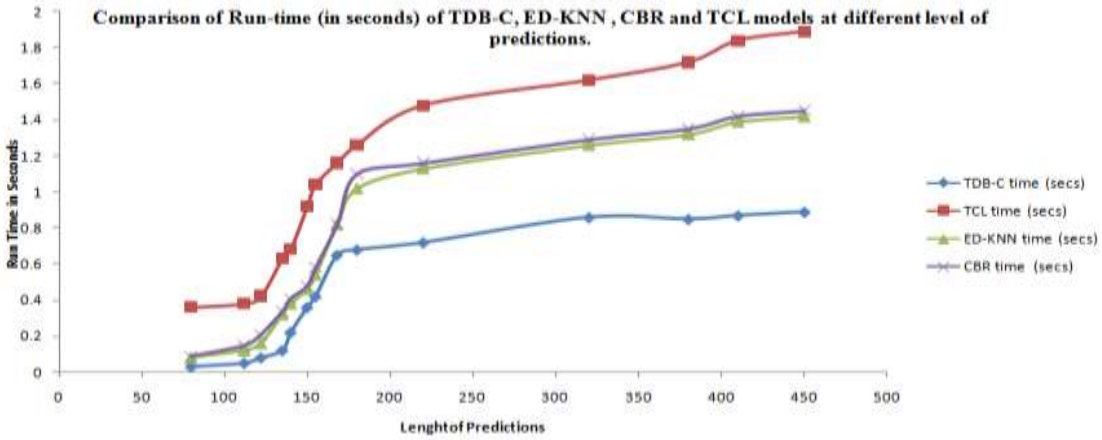

Figure 4: Comparison of Run-time (in seconds) of TDB-C, ED-KNN , CBR and TCL models at different level of predictions.

ISSN: 1334-4977

SUMMARY OF FINDINGS

In this study, efforts were made to construct an intelligent online optical health decision support and blindness prediction system using a hybrid predictive technique. We proposed a marriage of conveniences between the Tanimoto distance measurement technique and the clustering technique. We collected over 13,000 optical medical histories of anonymous patients of the Nigeria National eye centre, Mando, Kaduna. The collected data were preprocessed, data mart was developed. The proposed system was implemented using an in-house developed PHP program using MYSQL for database management and Apache/XAMP HTTP as hosting server. We carried out experiment on our designed experimental system, the results are presented and analyzed. Our experimental results shows that the adoption for the TDB-C model may lead to a more accurate, faster, scalable, efficient and useful predictions that can outperform the baseline methods i.e. the Traditional clustering(TCL), ED-KNN and the CBR predictive algorithms. Our experimental results shows that the precision rates of our predictive system is greater than 75%, this implies that over 75% of optical risk level predicted to clients are quite précised and accurate. The findings of this study can now be adopted by the governments, physicians and the public health system designed and administrators to improve on the optical health system of the society.

Conclusion

This work provides a basis for the design and implementation of an intelligent optical health care analysis and blindness prediction system. The system aggregates optical risk factor from the client, build a personalized risk factor estimate and predict a risk level to the user online and in real time basis.

We studied many approaches to creating online predictive systems; we discovered that some of these methods lack capability to handle large, noisy, high dimensionality datasets and irrelevant features. Our proposed system shows capability to overcome these challenges with ability to overcome computational complexity and scalability problem, common with most existing risk calculations and predictive systems. The results of our experiments shows that the proposed optical health risk calculator powered by the TDB-C model can outperform the baseline methods studied in terms of accuracy and speed. The proposed TDB-C model is able

to scale well for the present predictive system and can produce a useful, good, accurate, and faster risk level prediction to the clients consistently.

Recommendation for further work

In light of the findings of this work, we encourage other researchers to carry out more research on many other risk calculations and predictive techniques, compare the result with this model in order to validate the efficiency of this method and to determine more effective method of handling problems of this nature in future.

References

World Health Organization (WHO) (2018), Global data on vision impairment *Pascolini D, Mariotti SPM.(2010). Global estimates of visual impairment. British JournalOphthalmology.*

Bourne R.R.A., Flaxman S.R, Braithwaite T, Cicinelli M.V, Das A, Jonas J. B.(2017). Vision Loss Expert Group. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. Lancet Glob Health. 2017 Sep;5(9):e888–97. DOI:https://doi.org/10.1016/S2214-109X(17)30293-0

Judith Hurwitz and Daniel Kirsch (2018), Machine Learning For Dummies, IBM Limited Edition, USA, John Wiley & Sons, Inc. 111 River St. Hoboken, NJ 07030-5774. ISBN: 978-1-119-45495-3 (pbk); ISBN: 978-1-119-45494-6 (ebk)

Bhosale D, Ade R (2014). Feature selection based Classification using Naive Bayes, J48 and Support Vector Machine. Int J Comput Appl (0975–8887) 99 16.

Adeniyi DA, Wei Z, Yongquan Y (2015) Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. J Appl Comput Inform. https ://doi.org/10.1016/j.aci.2014.10.001

Wang K, Tan Y (2011). A New collaborative filtering recommendation approach Based on Naïve Bayesian method. ICSI, part II LNCS, 6729. pp 218–227

Yaghini M, Khoshraftar MM, Fallahi M (2012). A hybrid algorithm for artificial neural network training. Eng Appl Artif Intell (2012). https ://doi.org/10.1016/j.engap pai.2012.01.023, 1–9

Hssina B, Merbouna A, Ezzikouri H, Erritali M,(2014). A Comparative study of decision tree ID3 and C4.5. Int J Adv Comput Sci Appl https ://doi.org/10.14569 /Speci alIss ue.2014.04020 3 (Special issue on Advance in Vehicular Ad Hoc Networking.and Applications)

Han J, Kamber M (2006) Data mining concept and Techniques, 4111, 2nd edn. Morgan Kaufmann Publishers, Elsevier inc., San Francisco, pp 285–350.

Adeniyi DA, Wei Z, Yongquan Y (2016) Design and realization of online, real-time, web usage data mining and recommendation system using Bayesian classification method. Int J Comput Sci Eng Inf Technol Res (IJCSEITR). 6(3): 19–38 (**ISSN(P)): 2249–6831; ISSN(E): 2249–7943**).

Soni,J., Ansari, Z., Sharma, D. (2011).Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. International Journal of Advanced Trends in Computer Science and Engineering 3(6).

Aha DW (1991).Case based learning algorithms. In: Proceedings of the DARPA Case-Based Reasoning Workshop_ distributed by Morgan Kaufmann Publishers Inc. PP. 1–13.

Adeniyi DA, Wei Z, Yongquan Y (2018), Risk Factors Analysis and Death Prediction in Some Life-Threatening Ailments Using Chi-Square Case-Based Reasoning (χ2 CBR) Model. Interdisciplinary Sciences: Computational Life Sciences https://doi.org/10.1007/s12539-018-0283-6

Bichindaritz (2015). Data mining methods for case-based reasoning in health sciences. In: Proceedings of the ICCBR 2015 Workshops. Frankfurt, Germany. pp. 184–198.

Biau G, Scornet E (2015). A random forest Guide Tour., arxiv:1511.05741v\[maths.ST] 2015, pp. 1–42.

Ryan Poplin,R, . Varadarajan A.V. Blumer K., Liu,Y., McConnell, M.V., Greg S. Corrado, G.S., Peng,L. and Dale R. Webster, D.R.(2018).Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning Nature Biomedical Engineering, Macmillan Publishers Limited, part of Springer Nature. https://doi.org/10.1038/s41551-018-0195-0.

Rivas T, Paz M, Martins JE, Matias JM, Gracia JF, Taboadas J.,(2011). Explaining and predicting workplace accidents using data-mining techniques J Reliab Eng Syst Safety 96(*7*) 739–747. https ://doi.org/10.1016/j.ress.2011.03.006

Anbarasi M.,, Anupriya, E., Iyengar,S.N(2010). *Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,* International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376.

Zhang, B., Vogt, M. Maggiora, G.M. Bajorath, J.(2015) Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures J Comput Aided Mol Des. DOI 10.1007/s10822-015-9872-1

Bajusz,D., Rácz, A. and Héberger K. (2015). Why is Tanimoto index an appropriate choicefor fingerprint-based similarity calculations? Journal of Cheminformatics (2015) 7:20. DOI10.1186/s13321-015-0069-3.

ISSN: 1334-4977