International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Application of Information Gain Based Weighted LVQ for Heart Disease Diagnosis

Radhanath Patra[1], Bonomali Khuntia[2], Dhruba Charan Panda[3]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract:

To enhance the performance of Linear Vector Quantization (LVQ) for classification, an Information Gain based Weighted Linear Vector Quantization (IG-WLVQ) method is proposed in this paper. The information gain technique performs feature selection and provides informative attributes. So, information gain concept is embedded in winning vector calculation of the existing LVQ technique. It does dynamic features selection as well as calculates the winning vector in the run time. For analysing the performance of IG-WLVQ, Cleveland Heart disease dataset from UCI machine learning repository is used. Thus, the attributes having zero information are automatically wiped out in learning, and winning vectors are calculated from informative attributes. IG-WLVQ method not only performs dynamically future selection but also improves the classification performance of LVQ with a classification accuracy reaches to 100%.

Keywords: *LVQ, IG, IG-WLVQ, ML*

1. Introduction

Heart disease is a matter of serious concern worldwide. The mortality rate due to heart disease is increasing day by day in human beings. Therefore, effective measures are highly essential to control the disease. In this regard, role of computer aided diagnosis is highly needed for accurate disease prediction. Machine learning techniques have already proven as the most reliable and perfect platform for health care sector. The Cleveland heart dataset of UCI machine repository is considered for this analysis. Dataset is having 76 attributes and 303 instances. Most of the researchers used 13 attributes and one class level for the analysis neglecting the significance of remaining attributes [1][2]. The Majority of research which was carried out with limited attributes, provides good accuracy. Some of the authors use ensemble technique for feature extraction to improve classification accuracy [3][4]. In this paper, an attempt has been made to use the whole dataset without any information loss. LVQ combines the idea of competitive learning with supervised classification has the advantage of highest training speed with simple structure [5]. So, our idea is to boost the LVQ classification performance. This is done by Multiplying the Information gain of attribute in distance calculation formula of LVQ. In this paper Information gain used to find the distance which automatically performs feature selection. The attributes having zero information gain value are removed, and the informative attributes are used for finding winning vector. It is found that Information Gain based Weighted Linear Vector Quantization (IG-WLVQ) performance shows a better classification accuracy in compared to normal LVQ classifier.

2. Related Work

Abdullah Caliskan and Mehmet Emin Yuksel (2017) implemented deep neural network classifier model with two encoder and a SoftMax layer to analyze the coronary artery disease medical dataset by taking 303 records and 14 attributes including class level. The classification accuracy was 87.64% [6]. Ashraf et al. (2019) also used deep learning neural network for the same dataset and the classification accuracy was found to be 95%[7]. KaanUyar and Ahmet Ilhan (2017) developed genetic algorithm based recurrent fuzzy neural network (RFNN) and achieved classification accuracy of 96.63% [8]. Poornima v and Gladis D (2018) proposed orthogonal local preserving projection (OLPP) and hybrid classifier technique. In OLPP technique the authors implemented principal component analysis followed by orthogonal basis vector for dimension reduction of original dataset and afterwards a hybrid classifier composed of group search optimization algorithm along with Levenbergmarqartdt algorithm to get the best classification accuracy of 94% [9]. Amita Malav and Kalyani Kadam (2018) used knn clustering with multiplayer perceptron technique and achieved classification accuracy of 93.52% [10]. Mutasem Sh.Alkhasawneh (2019) proposed a hybrid neural structure, a combination of forward neural network with ELMAN neural network (HECFNN) for classification of the same dataset along with other four datasets of UCI machine learning repository. The neural network structure consists of input layer, hidden layer, context layer and an output layer. The classification accuracy achieved with the structure is 94.01% [11]. Ali et al. (2019) introduced chi2 ($\lambda 2$) test to remove noisy features and applied deep neural network (DNN) to the extracted dataset of 303 records with 76 attributes. The proposed system was designed to solve the over fitting and under fitting problem occurred due to various machine learning approach. The classification accuracy was found to be 93.3% [12]. Kathleen H Miao and Julia H Miao (2018) developed an enhanced deep learning model with the same dataset by taking 28 attributes which resulted in an accuracy of 83.6%.[13].

[1] *Radhanath Patra, India*
ORCID ID : 0000-0002-8047-2117
[2] *Bonomali Khuntia, India*
ORCID ID : 0000-0002-6223-0413
[3] *Dhruba Charan Panda,India*
ORCID ID : 0000-0003-1414-3400
** Corresponding Author Email: 1radhanath.patra@gmail.com, bonomalikhuntia@gmail.com*

3. Proposed Methodology/Algorithm

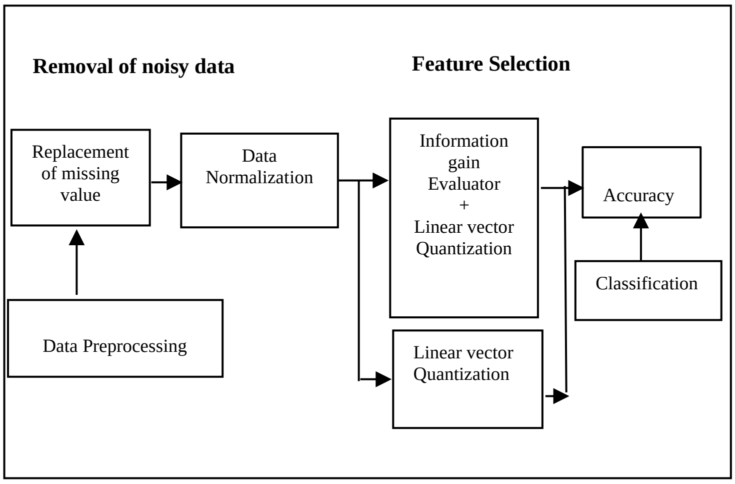
The UCI machine learning Cleveland heart dataset consisting of all 303 records having 76 attributes, is considered in the proposed system.

The basic function of the classification problem is presented in fig. 1. It involves three operations as follows:

a) Data pre-processing

b) Feature Selection/Information Gain Evaluation

c) Classification



Removal of noisy data
Feature Selection
Replacement of missing value
Data Normalization
Information gain Evaluator + Linear vector Quantization
Accuracy
Classification
Linear vector Quantization
Data Preprocessing

Fig 1: Schematic diagram representation of proposed IG-WLVQ for classification

3.1. Data Pre-processing

Pre-processing is essential for any dataset in machine learning application. It makes data smoother, reliable and fit for classification. Under pre-processing data imputation, feature selection or feature extraction have great roles. Undoubtedly, after pre-processing classifier improves performance.

3.2. Data Imputation

Heart dataset is considered with 73 input attributes and one class label. There are 6 records with missing values. It is 2% of the total sample values. The missing values of the attributes are replaced with mean values. Attributes are normalized. A uniform representation dataset is achieved. Normalization makes the data uniform and suitable for classification.

3.3. Feature selection

Feature selection is an important process by which redundant information are removed. It is basically categorized in to three parts i) Filter method ii) Wrapper Method and iii) Embedded method as displayed in figure 2.

Filter methods are much faster and simple for implementation. Here, information gain is proposed for attribute selection. Information technique is derived from entropy concept which is referred from equation(1) to (3). Information gain (IG) technique as mentioned in equation (4) processes all the attributes, and arranges it based upon information gain value.

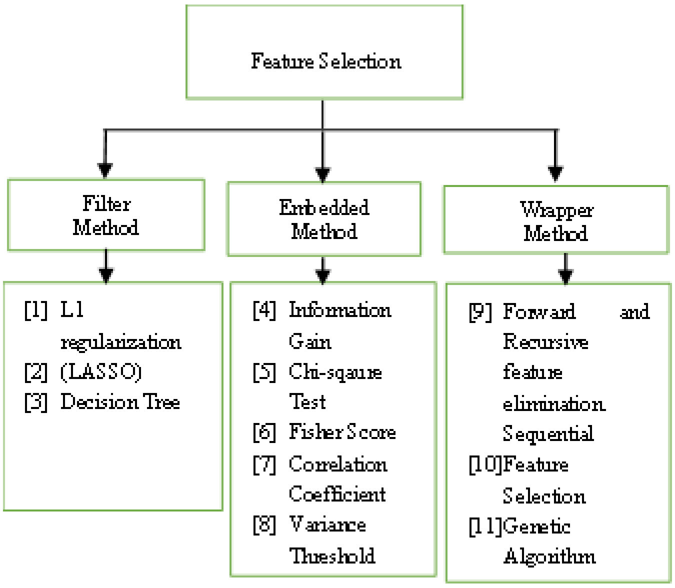


Fig2: Feature Selection Methods

3.4. Information gain evaluator

Information gain (IG) reduces the uncertainty of information, and calculates information gain value of attributes. While doing calculation of Information gain, entropy and conditional entropy of attributes are evaluated. And, it is explored that features those have high information gain values are more inclined to class label. So, Information gain values of attributes are sorted by ranker method that helps to distinguish the informative and non-informative attributes. It is proved that more the value of IG, lowers the value of entropy or uncertainty.

The information gain is calculated as follows:

Let

$x_i = \{x_1, x_2, x_3 \dots \dots x_n\}$

Entropy:

$$H(x) = E[I(X)] \quad (1)$$

$$H(x) = -\sum_i^n p(x_i) log_2 p(x_i) \quad (2)$$

conditional Entropy:

IF X and Y are represented two values of x_i and y_j,Then

$$H[X/Y]=\sum_{i,j} p(x_i, y_j) \log\left(\frac{p(x_i,y_j)}{p(j)}\right) \quad (3)$$

Information gain:

$$IG_i = \left(H(x) - H(X/Y)\right) \quad (4)$$

Using Information gain the features such as Age, Sex, Cp, TrestBps, Htn, Chol, Cigs, Years, Fbs, Dm, Famhist, Restecg, Ekgmo, Ekgday, EKgyr, Dig, Prop, Nitr, Pro, Diuretic, Rcaprox, Rcadist, Om1, Om2, Ramus, Cxmain, Laddist, Lmt, Cyr, Cday, Cmo, Thal:Thaldur, Thaltime, Met, Thalach, Thalrest, Tpeakbps, Tpeakbpd, Dummy, Trestbpd, Exang, ca, Xhypo, Odpak, Slope, lvx3, lvx4, lvf, , junk with information gain value are shown in figure 3 and figure 4. These can be easily distinguished from Smoke, Proto, Rld5, Rld5e, Restckm, Exerckm, Restef, Restwm, Exeref, Exerwm, Thalsev ,Thalpul, Earlobe, Ladprox, Diag, Lvx1, Lvx2, cathef e.t.c having zero "Information gain" value. Attributes with zero IG have no use for performance calculation.

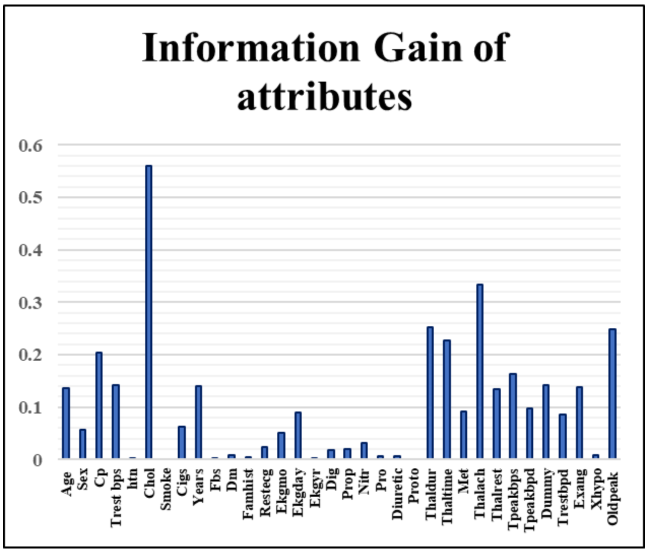


Fig3: Information gain of 1-35 attributes

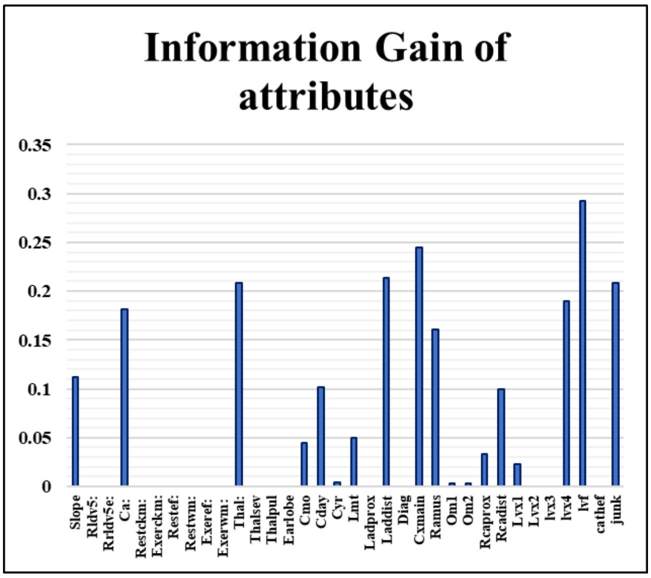


Fig4: Information gain of 36-70 attributes

3.5. Selected attributes

The selected attributes are 51 in numbers, 50 attributes are used as input and, one input is considered as class label. Moreover, 14 attributes used by the researchers belong to theses selected attributes. So, it is imperative to say that necessary and weightage attributes have been considered for our research and analysis.

3.6. Classification:

Linear vector quantization technique is based on the principle of winner take all algorithm concept. LVQ algorithm is a simple machine learning approach but its application and flexibility make it one of the powerful classification techniques under supervised learning. The power of LVQ lies on the Euclidean distance [18]. In our research paper, a weighted LVQ, Information Gain based Weighted Linear Vector Quantization (IG-WLVQ) is proposed for classification of heart disease. The distance is calculated based upon the weightage of the information gain of the corresponding attributes. IG of attributes are multiplied in distance calculation formula of LVQ. That enables to select the informative attribute at run time. So, feature selection process happens dynamically. In the learning process, multiplication of IG of attributes removes attributes having zero information gain. Also, it appears that the LVQ classification performance is improvised for selecting wining vector and its corresponding class value. Thus, LVQ performance gets enhanced.

The pseudo code for the IGW-LVQ is as described below and is explained in figure 5.

Pseudo code of IGWLVQ

```
$X_i$ =Training vector where i=1 to n, $\{X_1,X_2,X_3,X_4, \ldots X_n\}$

T=Class for training vector $X_i$

$w_j$=Weight vector for$j^{th}$output unit

$c_j$=Class associated with $j^{th}$output unit

Step1: Find the number of class label and consider as m.

Step2: Choose the weight vector corresponding to m unit and assign the class label.

Such that $w_j$ where j= 1to m.

Step3: Initialize the alpha value=α

Step 4: Find the information content of each attribute

Step5: Continue with step 6 to 9 till the stopping condition is not achieved.

Step5: Calculate the Euclidean distance for j=1 to m and I =1 to n.

    D (i, j) =$\sqrt{\left(x_i - w_j\right)^2 * IG_i}$.

Step6: Obtain the winning unit where d (j) has minimum value

Step7: Calculate the new weight of the winning unit by the following relation

    (i)     If T= $c_j$ Then $w_j = w_j(old) + \alpha\left[x - w_j(old)\right]$

    (ii)    if T≠ $c_j$ Then $w_j = w_j(old) - \alpha\left[x - w_j(old)\right]$

Step 8: Reduce the learning rate $\alpha$

Step9: Test stopping condition. The stopping condition is either maximum number of iterations or sufficiently small value of the learning rate $\alpha$.
```

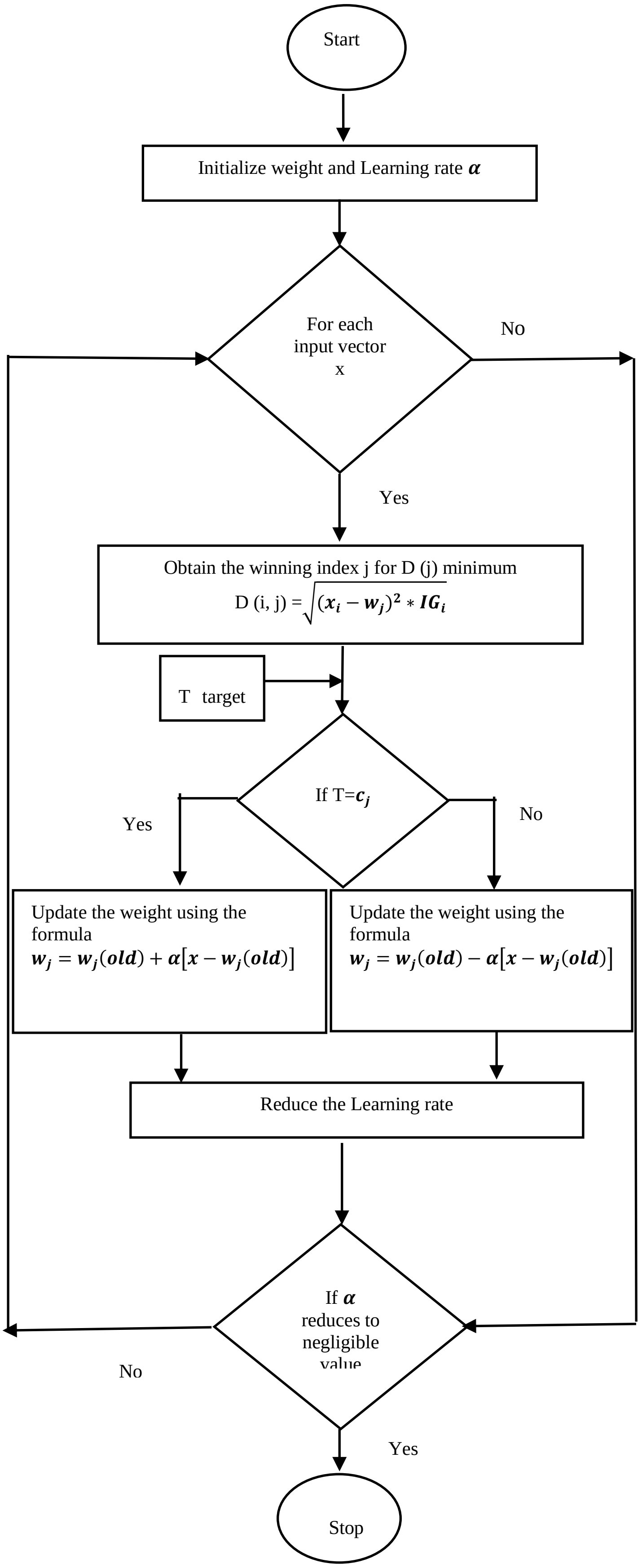

Start
Initialize weight and Learning rate α
For each input vector x
No
Yes
Obtain the winning index j for D (j) minimum
D (i, j) = $\sqrt{(x_i - w_j)^2 * IG_i}$
T target
If T=c_j
Yes
No
Update the weight using the formula $w_j = w_j(old) + \alpha[x - w_j(old)]$
Update the weight using the formula $w_j = w_j(old) - \alpha[x - w_j(old)]$
Reduce the Learning rate
If α reduces to negligible value
No
Yes
Stop

Fig5: Flowchart of IG- WLVQ Classification

4. Result Analysis

Dataset contains 303 records, 50 attributes and one class label. For implementation of LVQ, class level is categorized in to two parts, that are coded as "1" (presence of heart disease) and "0" (absence of heart disease). From the dataset two row vector are chosen randomly and assigned as weight vector to the two-output neuron. It is done in such a manner that two output neurons are assigned with two different class. Remaining dataset is divided into training part and testing part. The ratio of train dataset to test dataset is 9:1, and the updated weight found from train dataset is used for test dataset. In the proposed technique, information gain is multiplied to euclidean distance formula in LVQ for finding the code vector. Dataset having 73 attributes are processed with the above proposed methodology. From which, 50 attributes are selected in the run time and other attributes are automatically discarded due to zero information gain. In the forward computation, output is found out and compared to assigned class. The weight, bias input and learning rate updated in the learning process. The learning parameter alpha is set to 0.2. The process is learned for 100 epochs. This optimized calculated weight during raining process is calculated. In run time normal LVQ process, and IG-WLVQ option are chosen and compared. The observation is considered for 7 times as shown in table1. Thus, it is found that and IG-WLVQ outperforms to normal LVQ. So, the addition of information gain in distance formula, LVQ performance is enhanced. The calculated optimized weight is used for test dataset, IG-WLVQ gives a 100% classification accuracy. On the other hand, the classification accuracy performance by standard LVQ for the same dataset after pre-processing is found to be 90%. Thus, IG-WLVQ it simultaneously performs feature selection and improves the classification accuracy.

Table 1: Accuracy measurement at various iteration in-terms of percentage

Sl.No.	Standard LVQ	IG-LVQ (information gain with LVQ)	Improvement
1	70	76.66	6.66
2	73.33	83.33	10.00
3	76.66	80	3.34
4	80	86.66	6.66
5	83.33	86.66	3.33
6	86.66	90	3.34
7	90	100	10

Average increase in accuracy of 5.6% is achieved in implementation of IG-WLVQ over standard LVQ as displayed in figure 6.

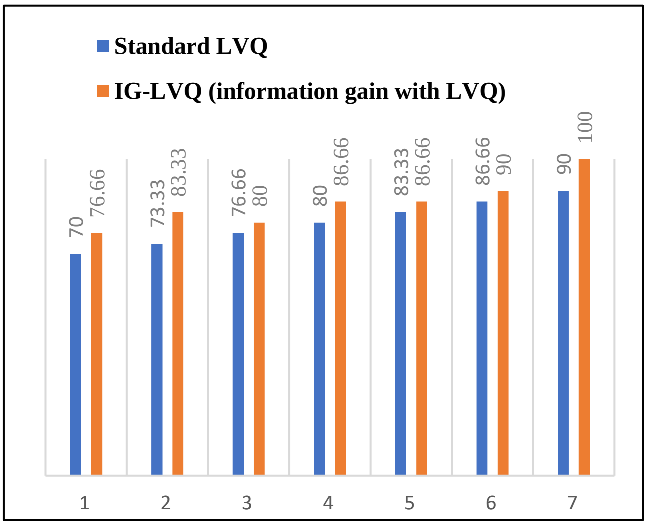


Fig 6. Comparison of LVQ and IG-WLVQ

Table 2: Heart Dataset Having 303 Rows And 73 attributes

Sl	Name of the classifier	Accuracy	Precision	Recall	F1_score	Auc
1	KNN	86.%	1	.78	.88	.86
2	SVM	92%	1	.86	.92	.92
3	RF	93%	1	.88	.93	.93
4	XGBOOST	95%	1	.91	,95	.95
5	NAÏVE BAYES	98.9	.97	1.0	.98	.98
6	DECISION TREE	98.9	.97	1.0	.98	.98
7	IG-WLVQ	100	1	1	1	1

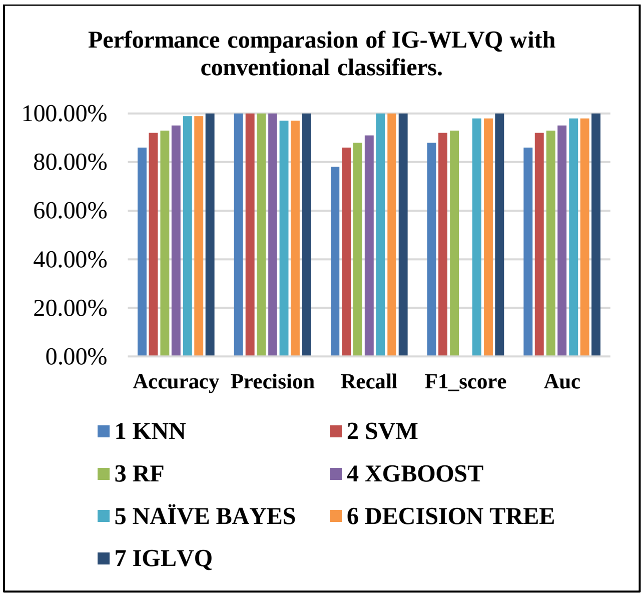


Fig 7. Comparison of IG-WLVQ with other classifiers

Thus, the proposed approach of LVQ(IG-WLVQ) performance is compared with other classifier performances for the same dataset. It is evident that the improvement in classification accuracy in the proposed method is significant. For Performance analysis various classifiers such as K-Nearest neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Extreme gradient Boosting (XG Boost) are used. Table 2.0 and Fig 7 shows the comparison of the proposed IG-WLVQ with other classifiers.

5. Conclusion

The IG based weighted LVQ (IG-WLVQ) performs both feature selection and accuracy enhancement. It not only improves the LVQ performance, but also introduces feature selection process at run time. The technique will be best suited for sequential data analysis and classification. The analysis can be done with many datasets, and addition of any optimization technique can further provide more robust approaches.

Author contributions

Radhanth Patra: Writing-Original draft preparation, Software, Validation., Field study, **Bonomali Khuntia:** Conceptualization, Methodology, Review, Data curation, Modification
Dhruba Charan Panda: Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554,10.1109/ACCESS.2019.2923707

[2] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access, 8, 107562-107582. 10.1109/ACCESS.2020.3001149

[3] Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. Computational and mathematical methods in medicine, 2017. https://doi.org/10.1155/2017/8272091

[4] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M. R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y., & Tayebi, R. M. (2019). New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. Health and Technology, 1-12. https://doi.org/10.1007/s12553-019-00396-3

[5] Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. Neural Computing and Applications, 25(3-4), 511-524. https://doi.org/10.1007/s00521-013-1535-3

[6] Caliskan, A., & Yuksel, M. E. (2017). Classification of coronary artery disease data sets by using a deep neural network. The EuroBiotech Journal, 1(4), 271-277 , doi:10.24190/ISSN2564-615X/2017/04.03

[7] Ashraf, M., Rizvi, M. A., & Sharma, H. (2019). Improved Heart Disease Prediction Using Deep Neural Network. Asian Journal of Computer Science and Technology, 8(2), 49-54.

[8] Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science, 120, 588-593.

[9] Poornima, V., & Gladis, D. (2018). A novel approach for diagnosing heart disease with hybrid classifier. 10.4066/biomedicalresearch.38-18-434.

[10] Krishnaveni, S. ., A. . Lakkireddy, S. . Vasavi, and A. . Gokhale. “Multi-Objective Virtual Machine Placement Using Order Exchange and Migration Ant Colony System Algorithm”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 6, June 2022, pp. 01-09, doi:10.17762/ijritcc.v10i6.5618.

[11] Malav, A., & Kadam, K. A. (2018). A hybrid approach for heart disease prediction using artificial neural network and K-means. Int J Pure Appl Math, 118(8), 103-110.

[12] Alkhasawneh, M. S. (2019). Hybrid cascade forward neural network with elman neural network for disease prediction. Arabian Journal for Science and Engineering, 44(11), 9209-9220. https://doi.org/10.1007/s13369-019-03829-3.

[13] Gupta, D. J. . (2022). A Study on Various Cloud Computing Technologies, Implementation Process, Categories and Application Use in Organisation. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 09–12. https://doi.org/10.17762/ijfrcsce.v8i1.2064

[14] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}} $ Statistical Model and Optimally Configured Deep Neural Network. IEEE Access, 7, 34938-34945. 10.1109/ACCESS.2019.2904800.

[15] Miao, K. H., & Miao, J. H. (2018). Coronary heart disease diagnosis using deep neural networks. Int. J. Adv. Comput. Sci. Appl., 9(10), 1-8.

[16] Andrie Dazlee, N. M. A., Abdul Khalil, S., Abdul-Rahman, S., & Mutalib, S. (2022). Object Detection for Autonomous Vehicles with Sensor-based Technology Using YOLO. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 129–134. https://doi.org/10.18201/ijisae.2022.276

[17] Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. IEEE Transactions on Knowledge and Data Engineering, 18(3), 304-319. https://doi.org/10.1109/TKDE.2006.46

[18] Rai, S. K. ., Rana, D. P. ., & Kashif, D. M. . (2022). Hotel Personnel Retention In Uttar Pradesh: A Study of HYATT Hotels. International Journal of New Practices in Management and Engineering, 11(01), 47–52. https://doi.org/10.17762/ijnpme.v11i01.173

[19] Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. IEEE Access, 6, 63279-63291.10.1109/ACCESS.2018.2877269

[20] Amuda, O. K., Akinyemi, B. O., Sanni, M. L., & Aderounmu, G. A. (2022). A PREDICTIVE USER BEHAVIOUR ANALYTIC MODEL FOR INSIDER THREATS IN CYBERSPACE. International Journal of Communication Networks and Information Security (IJCNIS), 14(1). https://doi.org/10.17762/ijcnis.v14i1.5208

[21] E. P. Wigner, “Theory of traveling-wave optical laser,” *Phys. Rev*., vol. 134, pp. A635–A646, Dec. 1965.

[22] Pratiwi, A. I. (2018). On the feature selection and classification based on information gain for document sentiment analysis. Applied Computational Intelligence and Soft Computing, 2018, https://doi.org/10.1155/2018/1407817

[23] Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. Neural Computing and Applications, 25(3-4), 511-524. https://doi.org/10.1007/s00521-013-1535-3