International Journal of Technical Innovation in Modern Engineering & Science, vol. 9, no. 4, 2023, pp. 142-155

A Hybrid Metaheuristic Optimization Approach for Improving K-Nearest Neighbor Prediction Accuracy

Allemar Jhone P. Delima*

College of Computing Education, University of Mindanao, Philippines

Received 06 August 2023; received in revised form 20 October 2023; accepted 16 November 2023

Abstract

The k-nearest neighbor (KNN) algorithm is vulnerable to noise, which is rooted in the dataset and has negative effects on its accuracy. Hence, various researchers employ variable minimization techniques before predicting the KNN in the quest so as to improve its predictive capability.

The genetic algorithm (GA) is the most widely used metaheuristics for such purpose; however, the GA suffers a problem that its mating scheme is bounded on its crossover operator. Thus, the use of the novel inversed bi-segmented average crossover (IBAX) is observed. In the present work, the crossover improved genetic algorithm (CIGAL) is instrumental in the enhancement of KNN's prediction accuracy. The use of the unmodified genetic algorithm has removed 13 variables, while the CIGAL then further removes 20 variables from the 30 total variables in the faculty evaluation dataset.

Consequently, the integration of the CIGAL to the KNN (CIGAL-KNN) prediction model improves the KNN prediction accuracy to 95.53%. In contrast to the model of having the unmodified genetic algorithm (GA-KNN), the use of the lone KNN algorithmand the prediction accuracy is only at 89.94% and 87.15%, respectively. To validate the accuracy of the models, the use of the 10-folds cross-validation technique reveals 93.13%, 89.27%, and 87.77% prediction accuracy of the CIGAL-KNN, GA-KNN, and KNN prediction models, respectively. As the result, the CIGAL carried out an optimized GA performance and increased the accuracy of the KNN algorithm as a prediction model.

Keywords: CIGAL-KNN, GA-KNN, IBAX operator, KNN algorithm, prediction models

1. Introduction

The prediction in the research serves as a powerful tool in the process of planning which can provide the researchers with a likelihood of future events. Technically, this data mining [1-2] approach is driven by using experiences and applying statistical, mathematical, or computational methods [3-5].

For most of the organizations, prediction helps the administration in analyzing the data which is needed for managerial decisions. The data can positively constitute an improvement in the excellence of the organization services [4]. Various researches in the different areas employ several machine learning and data mining algorithms for the prediction that helps management in the decision and policy-making undertakings. Among many predictive algorithms, the KNN is the simple, reliable, and one of the most commonly used algorithm for such prediction and classification purposes [6]. It has been used in

* Corresponding author. E-mail address: allemarjpd@yahoo.com
Tel.: +639661549000

crime mining [7], educational data mining [8], and healthcare services [9] and so on. However, the KNN algorithm is vulnerable to noise or irrelevant data features [10]. This problem is rooted in the dataset and leads to low prediction accuracy, which results in a flawed prediction tantamount to unreliable decision making.

To address the above-mentioned problem, various researchers have employed the data reduction methods to optimize the number of variables within the dataset to improve KNN's predictive accuracy. A technique has been used to increase the prediction accuracy of KNN through the integration of the genetic algorithm in the predictive model (GA-KNN) [11]. Among many metaheuristics, the genetic algorithm is one of the most competent indexes used for global optimizations, feature selections, and the data reduction algorithm that is widely used in the literature [12-14]. However, the accuracy result of the hybrid GA-KNN prediction model is still unsatisfactory. The genetic algorithm suffers a problem with premature convergence, a coupling-based problem is bounded on the crossover operator of the GA [13].

To address the above-mentioned issue, one of the suggested solutions is to prevent the premature convergence to design an efficient crossover operator; thus, the creation of the novel inversed bi-segmented average crossover (IBAX) [15]. There is an upcoming need of developing a new crossover operator for the genetic algorithm so as to carry out an optimized GA performance. Besides, it is necessary for variable minimization that will enhance the accuracy of the KNN prediction model [16]. The use of predictive models in an organization with erroneous data will lead to flawed predictions, which perpetuates the risk of additional harm which negatively affects the decision of the management in an organization.

This study aims to increase the accuracy of the KNN algorithm as a prediction model through the integration of the novel inversed bi-segmented average crossover (IBAX), a new crossover operator of the genetic algorithm applied in the faculty evaluation data. Specifically, the paper determines the reduction in the number of variables by using the unmodified genetic algorithm as against the crossover-improved genetic algorithm proposed by [15] and calculates the improvement in the accuracy of KNN through the integration with the data reduction techniques.

2. Related Literature

The KNN algorithm is one of the most effective nonparametric techniques due to its simplicity, suffers problems in its accuracy, and k sensitivity. Various researches were conducted to address the issue of k sensitivity. A robust generalized mean distance-based k-nearest neighbor classifier (GMDKNN) was proposed in the quest to prevent the degradation of KNN-based performance owing to the neighborhood k sensitivity. The GMDKNN observes the generalized mean distance as the preliminary that measures the distance similarity of the sample query and the k nearest neighbor. In general, the GMDKNN introduces the multi-generalized mean distances and the nested generalized mean of each class. The proposed method can employ much nearest neighbors for the favorable classification and has less sensitiveness to the values of k [17].

Further, another technique using the local mean representation-based KNN algorithm (LMRKNN) was proposed to solve the repetitive problem of the KNN. In the LMRKNN, the categorical k-nearest neighbors of a query sample are first chosen to calculate the corresponding categorical k-local mean vectors, and then the query sample is represented by the linear combination of the categorical k-local mean vectors; finally, the class-specific representation-based distances between the query sample and the categorical k-local mean vectors are adopted to determine the class of the query sample. The simulation results reveal that the LMRKNN model outperforms the other relative KNN-based methods used in the study [18].

Another modification on the KNN method was proposed by employing two locality constrained representation for k nearest neighbors. The method is called weighted representation-based k-nearest neighbor rule (WRKNN), while the other is termed as the weighted local mean representation-based k-nearest neighbor rule (WLMRKNN). In WRKNN, the linear combination of the k nearest neighbor from each class is represented as the test sample, and the localities of k-nearest neighbors per class as the weights constrain their corresponding representation coefficients. By using the representation

coefficients of k-nearest neighbors per class, the representation-based distance between the test sample and the class-specific k-nearest neighbors is calculated as the classification decision rule.

For the WLMRKNN, the k-local mean vectors of k-nearest neighbors per class are first calculated and then used for representing the test sample. In the linear combination of the class-specific k-local mean vectors to represent the test sample, the localities of k-local mean vectors per class are considered as the weights to constrain the representation coefficients of k-local mean vectors. The representation coefficients are employed to design the classification decision rule which is the class-specific representation-based distance between the test sample and k-local mean vectors per class. The simulation results reveals that the proposed methods perform better with less k sensitivity as against the local mean-based k-nearest neighbor (LMKNN) [19], collaborative representation-based nearest neighbor (CRNN) [20], and multi-local means-based nearest neighbor (MLMNN) [21-22].

Besides the extent on the nearest neighbor query, there have been numerous researches that were conducted. To name some, a novel data structure through buffer kd-tree for processing massive nearest neighbor queries on GPUs was introduced [23]. Similarly, the performance of the KNN algorithm was improved based on the revised buffer kd-tree integration. A fast neighbor search through the revised kd-tree integration was realized . however, the method is not suitable for high dimensional data [24]. With respect to high dimensional data problems, the scalable nearest neighbor method through the the k-means tree introduction for fast approximate matching of binary features along with the k-d forest was proposed and found to be effective in addressing that the issues arise when scaling to very large size data sets [25]. Other proposed studies include the implementation of the fast k-nearest neighbor search via dynamic continuous indexing (DCI) [26], prioritized dynamic continuous indexing (PDCI) [27], and a fast exact nearest neighbor search algorithm based on semi-convex hull tree over large scale data in the quest to find the k-nearest neighbor objects for a given point in the class and space [28].

Premised on improving the prediction rate of the KNN algorithm due to its known repetitive problem, various researchers employed the data reduction using optimization algorithms for such purpose. The use of PSO and CFS are instrumental for feature selection in the quest to improve KNN's predictive accuracy in predicting the occurrence of the malignant tumors in the skeletal bones called sarcoma. The simulation results reveal that the PSO-KNN method attains an 85% accuracy compared to the CFS-KNN model with 81% accuracy [29].

Further, the PSO-KNN model and CART algorithm, which is used in the water level estimation and water quality forecast in Poyang Lake in China, reveal a 86.68% prediction accuracy by using the PSO-KNN model that outperforms the average prediction of the CART algorithm with a 81.76% prediction accuracy [30]. Another effective technique to increase the accuracy of the KNN is to use the principal component analysis (PCA) before the prediction. The PCA-KNN hybrid model yields an accuracy of 61.34%. The PCA-KNN model outperforms other predictive models in heart disease diagnosis by using the heart disease dataset from the UCI machine learning repository [31].

The introduction of the gravitational search algorithm (GSA) serves as a feature reduction technique to increase the KNN's prediction accuracy for disease prediction by using biomedical data. By using the GSA on the heart dataset, there is a 64.61% reduction in the features while 57.64% and 77.77% of the features are removed through using the dermatology and breast cancer datasets, respectively. In general, an average of 66% of the removed variables considerably improves the prediction accuracy of the KNN algorithmand increases the accuracy from 64.81% to 82.96% for heart dataset to name one [32].

The genetic algorithm as a feature selection technique enhances the performance of the KNN and Naïve Bayes algorithms in diabetes detection. The GA-KNN model attains a 83.12% accuracy while a 81.82% prediction accuracy is depicted by using the GA-NB model making the GA-KNN as the optimal model for the prediction of the healthy or diabetic patients [11]. Furthermore, the use of GA-KNN and GA-SVM models was observed in disease diagnosis by using gene expression levels. The GA-KNN model obtains a 92.68% prediction accuracy on the prostate dataset [33].

3. Methodology

The enhanced prediction model is composed of two significant stages: the variable optimization stage and the prediction stage. The details are presented in Fig. 1.

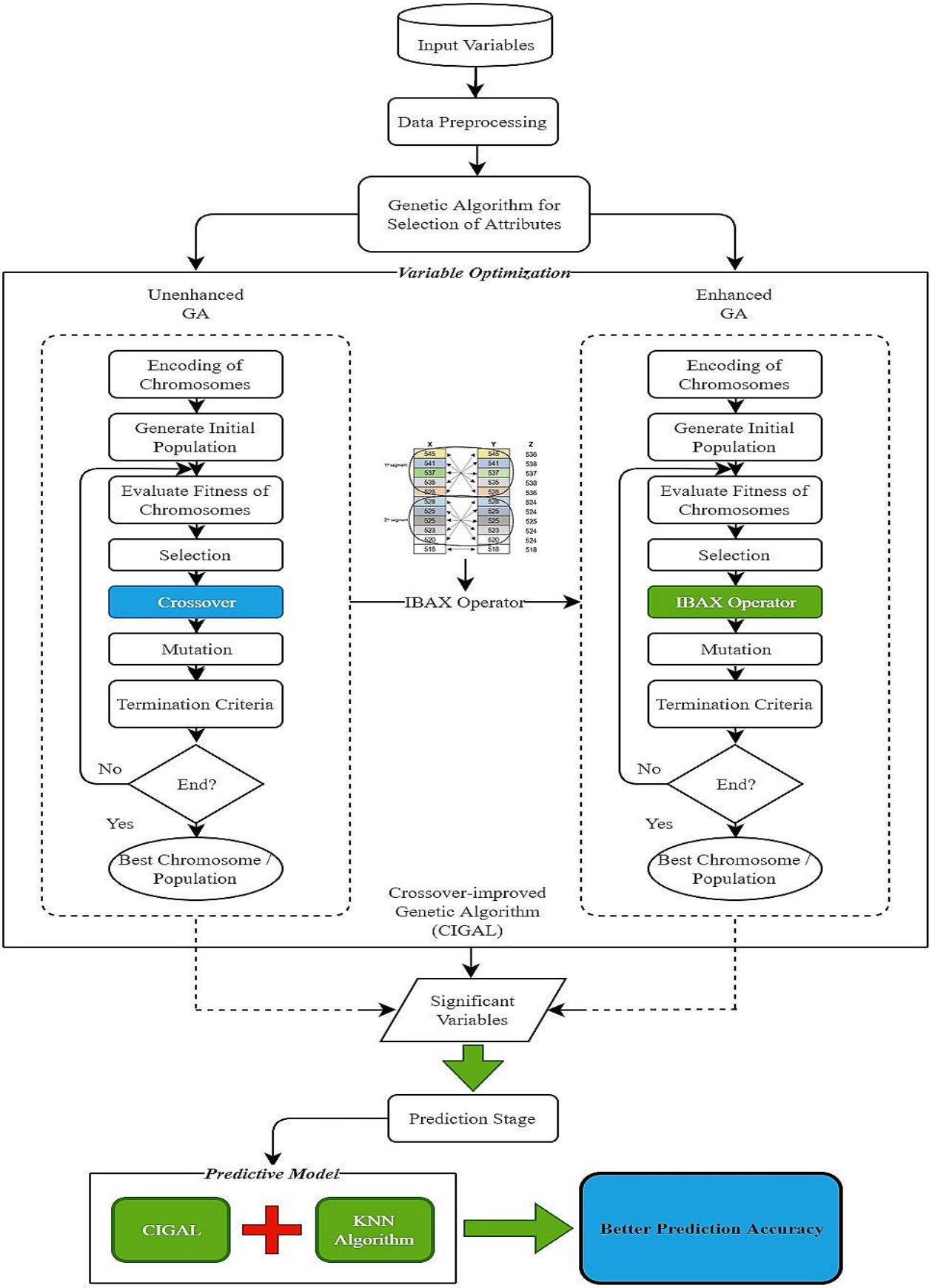

Fig. 1 Conceptual framework of the study

The study enhances the KNN predictive model by integrating the crossover-improved genetic algorithm with the IBAX operator. The use of the CIGAL is instrumental in the pre-processing step in optimizing variables within the dataset. The identified significant variables through the generation of the enhanced GA, which is the CIGAL, are used as the input for the prediction stage in the quest to obtain the better prediction accuracy of the model.

3.1. Data

The responses of 597 random student respondents are from four state universities and colleges (SUC) in Caraga Region, Philippines: the Surigao state college of technology (SSCT), Caraga state university (CSU), Surigao del Sur state university (SDSSU), and Agusan Sur sate college of agriculture and technology (ASSCAT) in the evaluation of the faculty instructional performance for the 2nd semester of S.Y. 2016-2017. Out of the total number of records, 70% and 30% data composition for training and testing are used for prediction. There are thirty variables that represent the instructional performance of the facultyand are divided into six (6) parts as to wit: methodology, classroom management, student discipline, assessment of learning, student-teacher relationship, and peer relationship as depicted in Table 1.

Table 1 Variables used in the study

Category	Description As a student, I have observed that my instructor/professor…	Variable
Methodology	Utilizes varied designs/ techniques/ activities suited to the different types of learners.	M1
	Explains learning goals and instructional procedures to the students.	M2
	Uses real-life examples in the class to sustain the student's interest in learning.	M3
	Creates a situation that encourages students to use critical thinking.	M4
	Delivers accurate/relevant/updated content knowledge.	M5
Classroom Management	Establishes routines to maximize the instructional time.	C1
	Organizes and assigns the daily cleaners.	C2
	Employs an effective system of the classroom set-up.	C3
	Employs strategies to maximize the use of resources in learning activities.	C4
	Implements rules/policies inside the classroom.	C5
Student Discipline	Handles behavior problems concerning the student's rights.	SD1
	Imposes disciplinary sanction(s) to the misbehaving student(s).	SD2
	Encourages students to submit requirements on time.	SD3
	Motivates students to respect each other.	SD4
	Allows students to exercise their creativity.	SD5
Assessment of Learning	Constructs a valid, reliable, formative, and summative test.	A1
	Uses appropriate non-traditional assessment techniques and tools (i.e. portfolio, journals, rubric, etc.)	A2
	Interprets and uses test results to improve teaching and learning.	A3
	Uses tools for assessing authentic learning.	A4
	Provides timely and accurate feedback to students.	A5
Student-teacher relationship	Encourages students to participate in class/school activities actively.	ST1
	Allows students to communicate directly to him/her.	ST2
	Provides equal opportunities for all students.	ST3
	Promotes teamwork among students.	ST4
	Makes him/herself available to students.	ST5
Peer relationship	Demonstrates appropriate behavior in dealing with students/peers/superiors.	P1
	Manifests flexibility, when deeming necessary.	P2
	Exhibits collegiality with colleagues.	P3
	Observes professionalism at all times.	P4
	Empathizes other needs and concerns.	P5

3.2. Modification on the crossover operator of genetic algorithm

In this study, the use of the new crossover operator of GA is called inversed bi-segmented average crossover, which is developed by [15]. It is the modification of the traditional average crossover operator. The existing genetic algorithm operator called average crossover chooses the first gene of the first and second chromosomes. An offspring is produced by calculating

the average of the mated genes. The process is performed repeatedly until the last genes of the two chromosomes have produced its offspring. The detailed flow of the average crossover mechanism is presented in Table 2.

Table 2 The flow of the average crossover operator of the genetic algorithm

Step No.	Steps
1	Select the first gene of the first chromosome (X) and the first gene of the second chromosome (Y).
2	Create one offspring (Z) out of the two selected genes by using the average formula
3	$Z = [X + Y]/2$
4	Repeat until the last gene of the first and second chromosomes are mated and have produced the offspring.

Meanwhile, the inversed bi-segmented average crossover operator works through segmenting the chromosomes (x and y) into two and inversely computing the average of genes within each created segment. For each segment, the process entails a repeated performance until the last gene of the first chromosome mates with the first gene of the second chromosome. The detailed flow of the new crossover mechanism is presented in Table 3, and the graphical representation of the IBAX operator is shown in Fig. 2.

Table 3 The flow of the IBAX operator of the genetic algorithm

Step No.	Steps
1	Count the number of genes found in the chromosomes. Identify if the variables are in odd or even count.
2	Segment the chromosomes (x and y) by dividing the total number of genes in the chromosomes into two. Make sure that both the first and second segments must contain an equal number of genes.
3	In the first segment, create offspring Z for each gene by inversely pairing the first gene from chromosome X to the last gene on chromosome Y. Repeat until the last gene of the chromosome X and the first gene of the chromosome Y have inversely mated and have produced an offspring using the average formula.
4	$Z = [X + Y]/2$
5	Execute the same process on the second segment until genes from all segments have produced offspring. In the case of odd datasets, the last genes of the chromosomes will not be combined in the second segment and will automatically be mated with each other to produce offspring.

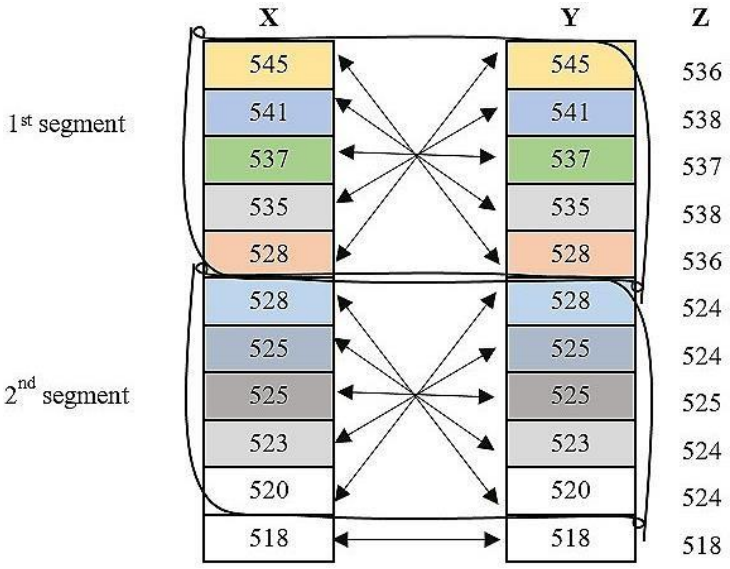

Fig. 2 Graphical representation of the IBAX operator

3.3. Variable optimization

The process identifies the significant variables in the dataset which are influential as the input on the KNN algorithm for the prediction stage in the quest to improve its prediction accuracy. The number of variables in the dataset is optimized with the help of the CIGAL having the IBAX mating scheme. The complete flow of the improved GA with the IBAX mating scheme is shown in Table 4.

To further test the effectiveness and the reduction rate of the CIGAL having the IBAX operator, the comparison on the reduction using the unmodified genetic algorithm which has the roulette wheel selection function, the original average

crossover as the crossover operator, and the swapping mutation function, are instrumental. For the CIGAL, the rank-based selection function, the inversed bi-segmented average crossover operator, and the swapping mutation function are also utilizedand execute both genetic algorithms for ten generations. In each generation, the new fitness value is calculated based on the result of the crossover function. The variables with the lowest fitness value after each generation for ten generations are aptly removed.

Table 4 The crossover-improved genetic algorithm integrating the IBAX operator

Step No.	Steps
1	Specify the number of chromosomes and generations, as well as the value of crossover and mutation rates
2	Generate the initial chromosome-chromosome number of the population and the initialization of the gene values based on the variables of an effective faculty instructional performance; calculate its fitness function
3	Evaluation of chromosome fitness value by calculating the objective function (Process steps 3-6 until the number of generations is met)
4	The use of rank-based selection function
5	Crossover having the IBAX operator
6	Mutation
7	Solution (Best Chromosomes)

3.4. Enhanced predictive model

The significant variables determined by the crossover-improved genetic algorithm with the IBAX operator and the unmodified genetic algorithm with the original average crossover are influential in the prediction which uses the KNN algorithm having a k value of 3. The simulation of the KNN employs the Waikato environment for knowledge analysis (WEKA) software with the version 3.8.2. The detailed flow of the K-Nearest Neighbor algorithm is illustrated in Table 5.

Table 5 The KNN algorithm

Step No.	Steps
1	For a training set $A = \{(a_1, b_1), \ldots, (a_T, b_T)\}$, the nth training sample is represented by $a_n \in A$, and $b_n \in \{w_1, w_2, \ldots, w_c\}$ represents the class label of the nth training sample; the total number of samples in the training set is represented as T, and the total number of classes is c.
2	Assign k value of 3
3	for all (Training samples (n=1, 2, ..., T)) do
4	
5	end for
6	The nearest k-training samples will be selected such as the minimum k distances.
7	Assign the class which has the most samples among the k-nearest samples to the testing sample.

3.5. Prediction accuracy evaluation

The allocation of 70% and 30% data composition for training and testing determine the accuracy of the KNN prediction model when integrated with the CIGAL having the IBAX operator (CIGAL-KNN) as against the model integrated with the unmodified genetic algorithm (GA-KNN). As tested, the predictive capability of the lone KNN algorithm further calculates the degree of the improvement in the accuracy of the data reduction techniques.

3.6. Model validation

The use of 10 folds cross-validation scheme is important in validating the accuracy of the model. To select the optimal model for the prediction, one must produce the highest accuracy rate with a lower statistical error value of the root mean squared error (RMSE). Along with it, the use of mean absolute error (MAE), precision, recall, and F-measure conform to the formula derived from [34] and can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

$$MAE = \sum_{t=T+1}^{T+h} \left| \hat{y}_t - \hat{y}_t \right| / h \tag{5}$$

$$RMSE = \sqrt{\sum_{t=T+1}^{T+h} \left(\hat{y}_t - \hat{y}_t \right) / h} \tag{6}$$

where TP, TN, FP, and FN mean true positive, true negative, false positive, and false negative, respectively. In the Eqs. (5) and (6), the forecast sample is represented by T+1,...,T+h, whereas the actual value is denoted by y_t along with $\hat{y}_t$ which represents the forecasted value in the period t. The lower the statistical error value is, the better the forecasting ability of the model.

4. Results and Discussions

4.1. Variable minimization results

The variable minimization results utilize the CIGAL and compare the generated results by using the unmodified GA. The index result is demonstrated in Table 6.

Table 6 The variable minimization result

Genetic algorithms	The number of variables	The number of removed variables	The number of left variables
Unmodified GA	30	13	17
CIGAL-IBAX	30	20	10

Table 7 Indexed simulation result using the CIGAL

The number of generations	The number of left variables	The number of removed variables	The removed variables	Percentage
0	30	0	-	-
1	30	2	C2, C3	6.66%
2	28	2	ST2, C1	6.66%
3	26	2	ST4, A2	6.66%
4	24	2	P2, A1	6.66%
5	22	2	ST3, A4	6.66%
6	20	2	C5, ST5	6.66%
7	18	2	SD5, SD2	6.66%
8	16	2	A5, M1	6.66%
9	14	2	A3, P3	6.66%
10	12	2	P4, M5	6.66%
10	-	-	-	-
Total Percentage of Variables Removed				66.66%

The simulation result shows that the crossover-improved genetic algorithm with the inversed bi-segmented average crossover outperforms the unmodified genetic algorithm in reducing the number of variables in the dataset. The use of the unmodified genetic algorithm has removed 13 of the variables from the dataset. From the 30 total number of variables, it was

reduced up to 17. Meanwhile, the crossover-improved genetic algorithm further reduces the number of variables in the dataset and leaves only ten variables in general. The indexed simulation result generated by the novel crossover operator of the genetic algorithm is illustrated in Table 7.

In the first generation, the variables C2 and C3 obtain the lowest fitness value among the variables in the groupand remove it from the chromosomes. The removed variables will have no chance to be included for the next generation. From the 30 variables, it is reduced to 28 from the first generation alone. The succeeding generations have removed variables ST2 and C1 for the second generation, variables ST4 and A2 for the third, variables P2 and A1 for the fourth generation, variables ST3 and A4 for the fifth generation, variables C5 and ST5 for the sixth generation, variables SD5 and SD2 for the seventh generation, variables A5 and M1 for the eighth generation, and variables A3 and P3 for the ninth generation; the variables P4 and M5 were removed on the tenth generation. From the 30 total variables, the CIGAL removes 20 variables from the dataset and retains variables M2, M3, M4, C4, SD1, SD3, SD4, ST1, P1, and P5 as these are identified to be instrumental for prediction. The graphical representation of the variable minimization using the genetic algorithms is shown in Fig. 3.

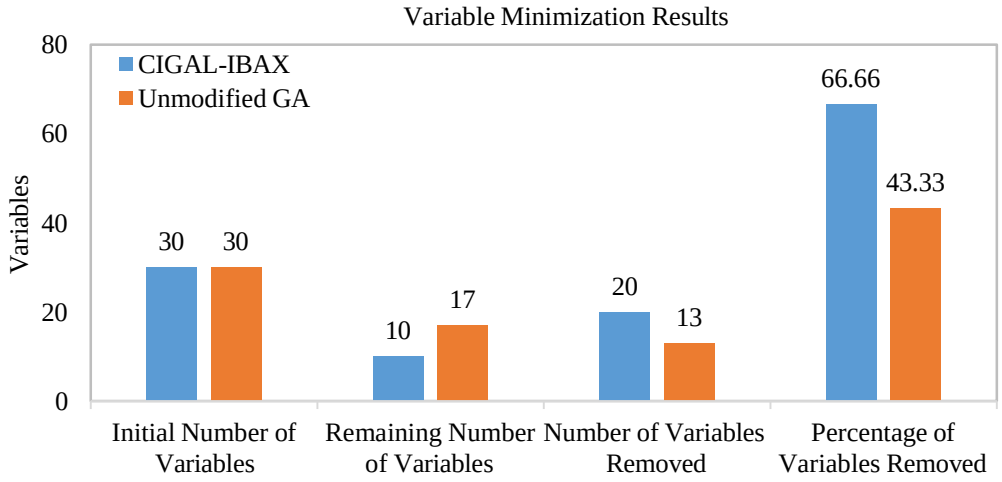

Fig. 3 The comparative result for variable minimization using genetic algorithms

The simulation results reveal that the amount of the reduction varies to the used genetic algorithm. The notion of dropping one or more variables within the dataset in the quest to help reduce dimensionality is certain. Therefore, the 66.66% removal of the variables is acceptable since the 60% ratio of feature reduction is suitable, as orchestrated by [35].

4.2. Prediction Model Accuracy Evaluation

The 70 and 30 percent data composition for training and testing is observed to evaluate the accuracy of the prediction models performed in WEKA software. The comparative results of the prediction model with the crossover-improved genetic algorithm (CIGAL-KNN) as against the prediction model with the unmodified genetic algorithm (GA-KNN) is shown in Table 8. The predictive capability of the KNN algorithm is also tested without the variable reduction stage.

Table 8 Indexed comparative accuracy results of KNN, GA-KNN, and CIGAL-KNN predictive models

Predictive Model	Accuracy
KNN Algorithm	87.1508%
GA-KNN	89.9441%
CIGAL-KNN	95.5307%

The simulation results shows that there is an increase in the accuracy of the models with the integration of GA; The phenomenon is particularly obvious on the models with the CIGAL. The CIGAL-KNN prediction model outperforms the GA-KNN model. It has the KNN algorithm alone with 95.53%, 89.94%, and 87.15% correctly classified instances,

respectively. By using the 70% and 30% data composition, the optimal model for predicting the accuracy of the responses on the faculty instructional performance evaluation in the four SUCs in the Caraga Region is the model with CIGAL having the IBAX operator integrated to the KNN. Fig. 4 shows the graphical representation of the obtained accuracies by using the three prediction models.

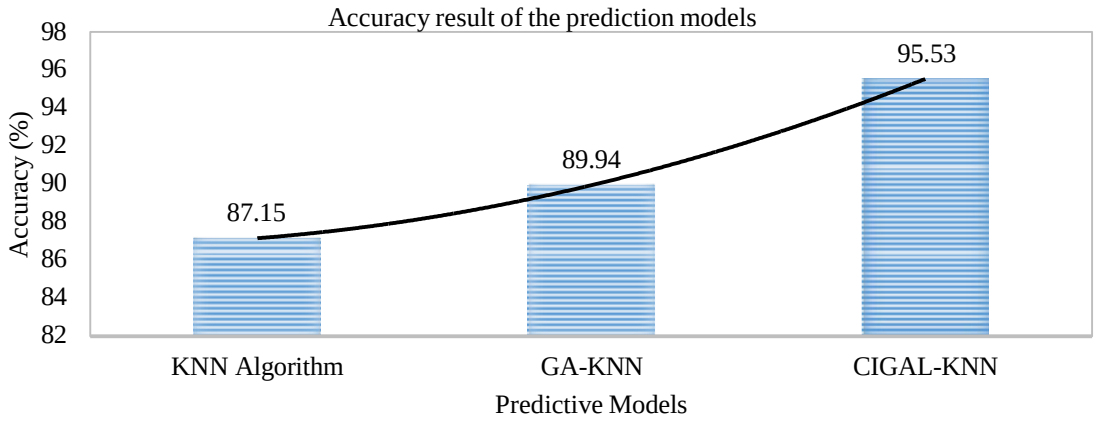

Fig. 4 The graphical representation of the accuracy results of the prediction models

4.3. Prediction model validation results

The predictive capability of the models is validated by using the ten folds cross-validation scheme performed in WEKA. The accuracy results of the model using the validation scheme are shown in Table 9.

Table 9 Indexed validation result of the predictive models

Predictive Models	Accuracy	RMSE	MAE	Recall	Precision	F- Measure
KNN Algorithm	87.7722%	0.2981	0.1385	0.878	0.876	0.877
GA-KNN	89.2797%	0.2595	0.1119	0.893	0.892	0.892
CIGAL-KNN	93.1323%	0.2146	0.0786	0.931	0.932	0.930

Based on the simulated validation result of the predictive models in Table 9, it can be seen that the CIGAL-KNN model still obtains the highest correctly classified instances of 93.13% as against the 89.27% and 87.77% prediction accuracies for GA-KNN and KNN predictive models respectively. The performance evaluation of the models makes use of the RMSE and MAE forecast statistical error tools. For the CIGAL-KNN model, both statistical error tests reveal a zero-based value for RMSE and MAE with the lowest error value of 0.21 and 0.07, respectively. The low estimated values reveal how concentrated the prediction is by using the crossover-improved genetic algorithm-based KNN prediction model. A low statistical error value depicts an ideal and desirable model for a good forecast.

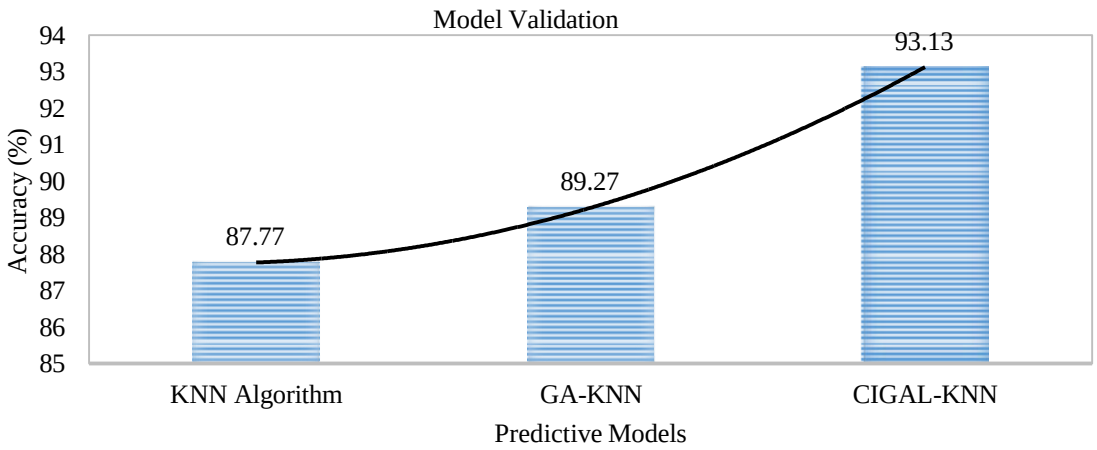

Fig. 5 Graphical representation of the model validation results

To further evaluate and compare the performance of the three models, the precision, recall, and F-measure performance metrics estimation are carried out. An overall precision score of 93.1% is depicted in the CIGAL-KNN model. The precision of the model in testing the accurateness of those predicted positives from the responses in the evaluation data is optimal. The recall metric denotes that 93% of the tested instances within the dataset are retrieved by the model correctly, and the F-measure determines the 93% balance performance of the model. The graphical representation of the model validation in terms of accuracy is demonstrated in Fig. 5.

5. Conclusions

The integration of the novel inversed bi-segmented average crossover to the genetic algorithm has paved the way for a more enhanced GA when it comes to optimization problems. Consequently, the incorporation of the crossover-improved genetic algorithm to the KNN has led to an increase in the accuracy of a prediction model. Since the prediction accuracy affected the decision of the organization management, increasing the accuracy of prediction models was viewed as necessary.

The variable minimization using the unmodified genetic algorithm removed 13 of the variables from the 30 total variables in the dataset and left 17 variables to be used for prediction. Meanwhile, the crossover-improved genetic algorithm having the novel IBAX operator outperformed the minimization capabilities of the unmodified GA, further removed 20 variables from the 30 total variables in the dataset, and left ten variables for prediction. In general, the integration of the crossover-improved genetic algorithm to the KNN predictive model (CIGAL-KNN) yielded an increase in prediction with a 95.53% accuracy as against the identified 89.94% prediction accuracy with the integration of the unmodified GA (GA-KNN), and 87.15% prediction accuracy utilizing the lone KNN algorithm.

Premised on the conclusions of the study, the use of other variable minimization, feature selection, and global optimization algorithms aside from the GA is recommended in the continued quest to improve the accuracy of the KNN predictive model and present a comparative analysis of the results. The utilization of the CIGAL on other predictive algorithms aside from the KNN is recommended to come up with a comprehensive literature review on the latest modifications to the different prediction models. Additionally, future researchers are encouraged to utilize the prediction model in other real-life datasets.

Author Contribution Statement

Allemar Jhone P. Delima: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Software, Validation, Visualization, Writing – original draft, Writing – review and editing.

Declaration of Competing Interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Data Availability

Data will be made available on request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

the content of the publication.

Ethics Approval and Consent

This study does not involve human participants, animals, or sensitive personal data. Ethical approval was therefore not required.

References

[1] A. J. P. Delima, "Predicting scholarship grants using data mining techniques," International Journal of Machine Learning and Computing, vol. 9, no. 4, pp. 513-519, August 2019.

[2] A. J. P. Delima, "Applying data mining techniques in predicting index and non-index crimes," International Journal of Machine Learning and Computing, vol. 9, no. 4, pp. 533-538, August 2019.

[3] M. J. Rezaee, M. Jozmaleki, and M. Valipour, "Integrating dynamic fuzzy C-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange," Physica A: Statistical Mechanics and its Applications, vol. 489, pp. 78-93, January 2018.

[4] U. O. Cagas, A. J. P. Delima, and T. L. Toledo, "PreFIC: predictability of faculty instructional performance through hybrid prediction model," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7, pp. 22-25, May 2019.

[5] A. J. P. Delima and M. T. Q. Lumintac, "Application of time series analysis for philippines' inflation prediction," International Journal of Recent Technology and Engineering, vol. 8, no. 1, pp. 1761-1765, May 2019.

[6] S. Fei, "The hybrid method of VMD-PSR-SVD and improved binary PSO-KNN for fault diagnosis of bearing," Shock and Vibration, vol. 2019, pp. 1-7, January 2019.

[7] V. Vishnupriya and M. Valarmathi, "An effective data mining techniques for analyzing crime patterns," IOSR Journal of Computing Engineering, vol. 1, pp. 26-30, 2017.

[8] M. Kumar, A. J. Singh, and D. Handa, "Literature survey on student's performance prediction in education using data mining techniques," International Journal of Education and Management Engineering, vol. 7, no. 6, pp. 40-49, November 2017.

[9] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," International Conference on Intelligent Computing and Control (I2C2), June 2017, pp. 1-8.

[10] D. García-gil, J. Luengo, S. García, and F. Herrera, "Enabling smart data: noise filtering in big data classification," Information Sciences, vol. 479, pp. 135-152, April 2019.

[11] R. N. Patil and S. C. Tamane, "Upgrading the performance of KNN and naïve bayes in diabetes detection with genetic algorithm for feature selection," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 3, no. 1, pp. 1371-1381, 2018.

[12] A. J. P. Delima, "An experimental comparison of hybrid modified genetic algorithm-based prediction models," International Journal of Recent Technology and Engineering, vol. 8, no. 1, pp. 1756-1760, May 2019.

[13] M. Y. Orong, A. M. Sison, and R. P. Medina, "A hybrid prediction model integrating a modified genetic algorithm to K-means segmentation and C4.5," TENCON 2018 - 2018 IEEE Region 10 Conference, October 2018, pp. 1853-1858.

[14] M. Mafarja, I. Aljarah, A. A. Heidari, A. I. Hammouri, H. Faris, A. M. Al-zoubi, et al., "Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems," Knowledge-Based Systems, vol. 145, pp. 25-45, April 2018.

[15] A. J. P. Delima, A. M. Sison, and R. P. Medina, "A modified genetic algorithm with a new crossover mating scheme," Indonesian Journal of Electrical Engineering and Informatics, vol. 7, no. 2, pp. 165-181, June 2019.

[16] A. J. P. Delima, A. M. Sison, and R. P. Medina, "Variable reduction-based prediction through modified genetic algorithm," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 356-363, 2019.

[17] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based K-nearest neighbor classifier," Expert Systems with Applications, vol. 115, pp. 356-372, January 2019.

[18] J. Gou, W. Qiu, Z. Yi, Y. Xu, Q. Mao, and Y. Zhan, "A local mean representation-based K-nearest neighbor classifier," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 3, pp. 29:1-29:5, April 2019.

[19] Y. Mitani and Y. Hamamoto, "A local mean-based nonparametric classifier," Pattern Recognition Letters, vol. 27, no. 10, pp. 1151-1159, July 2006.

[20] W. Li, Q. Du, F. Zhang, and W. Hu, "Collaborative-representation-based nearest neighbor classifier for hyperspectral imagery," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 2, pp. 389-393, February 2015.

[21] J. Gou, W. Qiu, Q. Mao, Y. Zhan, X. Shen, and Y. Rao, "A multi-local means based nearest neighbor classifier," 2017 IEEE 29th International Conference on Tools for Artificial Intelligence (ICTAI), June 2018, pp. 448-452.

[22] J. Gou, W. Qiu, Z. Yi, X. Shen, Y. Zhan, and W. Ou, "Locality constrained representation-based K-nearest neighbor classification," Knowledge-Based Systems, vol. 167, pp. 38-52, March 2019.

[23] F. Gieseke, J. Heinermann, C. Oancea, and C. Igel, "Buffer k-d trees: processing massive nearest neighbor queries on GPUs," Proc. 31st International Conference on Machine Learning, ICML 2014, January 2014, pp. 172-180.

[24] Y. Chen, L. Zhou, Y. Tang, J. P. Singh, N. Bouguila, C. Wang, et al., "Fast neighbor search by using revised k-d tree," Information Sciences, vol. 472, pp. 145-162, January 2019.

[25] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," IEEE Transactions on Pattern

Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2227-2240, November 2014.
[26] K. Li and J. Malik, "Fast K-nearest neighbour search via dynamic continuous indexing," International Conference on Machine Learning, June 2016, pp. 671-679.
[27] K. Li and J. Malik, "Fast K-nearest neighbour search via prioritized DCI," International Conference on Machine Learning, 2017, pp. 2081-2090.
[28] Y. Chen, L. Zhou, N. Bouguila, B. Zhong, F. Wu, Z. Lei, et al., "Semi-convex hull tree: fast nearest neighbor queries for large scale data on GPUs," 2018 IEEE International Conference on Data Mining (ICDM), November 2018, pp. 911-916.
[29] K. Baskaran, R. Malathi, and P. Thirusakthimurugan, "Feature fusion for FDG-PET and MRI for automated extra skeletal bone sarcoma classification," Materials Today: Proceedings,vol. 5, no.1, 2018, pp. 1879-1889.
[30] Y. Li, M. Y. A. Khan, Y. Jiang, F. Tian, W. Liao, S. Fu, et al., "CART and PSO + KNN algorithms to estimate the impact of water level change on water quality in poyang lake, China," Arabian Journal of Geosciences, vol. 12, no. 9, pp. 1-12, April 2019.
[31] R. S. El-Sayed, "Linear discriminant analysis for an efficient diagnosis of heart disease via attribute filtering based on genetic algorithm," Journal of Computers, vol. 13, no. 11, pp. 1290-1299, July 2018.
[32] S. Nagpal, S. Arora, S. Dey, and S. Shreya, "Feature selection using gravitational search algorithm for biomedical data," Procedia Computer Science, vol. 115, pp. 258-265, 2017.
[33] C. Gunavathi and K. Premalatha, "Performance analysis of genetic algorithm with KNN and SVM for feature selection in tumor classification," International Journal of Computer and Information Engineering, vol. 8, no. 8, pp. 1490-1497, 2014.
[34] E. Sugiyarti, K. A. Jasmi, B. Basiron, M. Huda, S. K, and A. Maseleno, "Decision support system for scholarship grantee selection using data mining," International Journal of Pure and Applied Mathematics, vol. 119, no. 15, pp. 2239-2249, 2018.
[35] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," Applied Soft Computing, vol. 74, pp. 634-642, January 2019.



Copyright© by the authors. Licensee IJTIMES India. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/).