Variable Reduction-based Prediction through Modified Genetic Algorithm

Allemar Jhone P. Delima[1], Ariel M. Sison[2], Ruji P. Medina[3]
Graduate Programs, Technological Institute of the Philippines, Quezon City, Philippines[1, 3]
Emilio Aguinaldo College, Manila, Philippines[2]

***Abstract*—Due to the massive influence in the use of prediction models in different sectors of society, many researchers have employed hybrid algorithms to increase the accuracy level of the prediction model. The literature suggests that the use of Genetic Algorithms (GAs) can sufficiently improve the performance of other prediction models; thus, this study. This paper introduced a new avenue of prediction integrating GA with the novel Inversed Bi-segmented Average Crossover (IBAX) operator paired with rank-based selection function to the KNN algorithm. The 70% of data from 597 records of student-respondents in the evaluation of the faculty instructional performance from the four State Universities and Colleges (SUC) in Caraga Region, Philippines were used as training set while the 30% was used for testing. The simulation result showed that the use of the proposed prediction model with the integration of the modified GA outperformed the KNN prediction model where GA with average crossover and roulette wheel selection function was used. The KNN where k value is three (3) was identified to be the optimal model for prediction with the 95.53% prediction accuracy compared to KNN with 1, 5, and 7 k values.**

Keywords—Enhanced prediction model; IBAX operator; modified genetic algorithm; prediction accuracy enhancement

I. Introduction

Data Mining (DM) is the process of extracting implicit information or knowledge from databases [1]-[3], that is drawn from the field of statistics [4] which uses mathematical and machine learning techniques and algorithms [5]. Knowledge Discovery in Databases (KDD) which is coined to data mining [6], represents the generally observed process in knowledge discovery where knowledge is the result of the data-driven discovery while data mining being the observed step in the process for efficiently automated discovery, employs diverse approaches of DM analysis [7].

The field of data mining has become standard practice in various disciplines such as business, finance, and marketing allowing to inadvertently impact social sciences and humanities in general [8]. The range of its application has also reached other sectors such as education [9], [10] and healthcare [11], [12]. DM is promising for researches applied in engineering, biomedical sciences, medical systems, web, sports, and shared market because of the accessibility to various vast datasets [13], [14].

There are several widely accepted major functions in data mining found in the literature such as association, classification, clustering, estimation, and prediction [15].

Prediction, as one of the optimal data mining approach, was defined by [16] as "a powerful tool in the process of planning that can provide the decision maker with a prediction about the future events according to using experiences and applying statistical, mathematical, or computational methods." It is commonly used in educational data mining (EDM) [17]-[19], crime mining [20]-[22], business and finance [23], [24], health [25], [26], and more.

Data preprocessing is one of the essential methods that are useful in data mining. It has led to the enhancement of the quality of data and improved the precision and accuracy level of a prediction model [22], [27]. Data reduction, as an important data preprocessing technique in DM, is performed through the selection and removal of the unneeded attributes in the dataset [28]. Reducing the training set or variables and retaining the most representative data is advisable. The goal is also to obtain nearly the same outcome or data-driven output [29]-[31]. Minimizing the size of the dataset aids in maximized accuracy [32], [28]. One of the widely used data reduction methods is the Genetic Algorithm [33] which was introduced by J.H. Holland. The average crossover, which is one of the crossover operators of GA, is modified in this study.

Due to the massive influence in the use of prediction methods in diverse fields such as weather and natural calamity, stock markets, telecommunication, transport organization, energy, economy and other sectors [16], researchers have employed models integrating algorithms for prediction as well as hybridizing algorithms and combining different techniques to elevate the accuracy level of a prediction model. To name some, the study of [34] employed a hybrid feature selection method integrating Weight by Relief and GA to select the best features in the dataset for myocardial infraction prediction using J48. An accuracy of 82.67% was depicted after applying the model to the imbalanced dataset. Also, a study of [35] used the K-Means segmentation technique and C4.5 algorithm to build a prediction model for customer loyalty in a multimedia service provider. The integration of K-Means and C4.5 algorithm have yielded an increase of 79.33% accuracy prediction from the identified 69.23% accuracy with the C4.5 algorithm alone.

Lastly, the prediction model of [36] used the K-Nearest Neighbor (KNN) algorithm to predict standard levels of OTOP's (One Tambon One Product) wood handicrafts product. Results showed that the model obtained the best prediction at the accuracy of 87.73%. The KNN algorithm is susceptible to noise and sensitive to irrelevant features [37]. Even though the prediction rate using KNN is already acceptable, but with the advent of combining genetic algorithm for variable reduction to address the problem of KNN, an increase of accuracy through the hybridization is hoped to be established.

With the advent of combining two or more models, an increase of prediction accuracy is evident [38] such that of [34] that obtained 82.67% and [35] with 79.33% after integrating hybridization than employing prediction with one algorithm alone.

Therefore, the quest of this study is not only to modify the genetic algorithm and introduce a new avenue of crossover mating scheme but also to increase the accuracy of the prediction model of [36] who used KNN algorithm through the integration of the modified genetic algorithm for feature selection and variable minimization before prediction. The rest of the paper is arranged as follows: Section II discusses the literature review of genetic algorithm and other prediction models. Section III includes the design and methodology used in the study. Section IV discusses the results and discussions while Section V highlights the conclusion and recommendation.

II. Literature Review

A. *Genetic Algorithm*

The genetic algorithm is one of the many evolutionary algorithms anchored on the biological adaptation in the quest for global optimization. GA is deliberately one of the famous technique used in the search for the optimal solution for problems with a large search space. GA produces and controls some individuals by assigning optimal operators on its three fundamental operations namely the selection, crossover, and mutation functions. In this study, a modified genetic algorithm with the integration of novel Inversed Bi-segmented Average Crossover (IBAX) is used. This novel crossover is a modified version of the traditional average crossover of GA.

B. *Genetic Algorithm-based Prediction Models*

The literature suggests that the genetic algorithm can efficiently increase the performance of other prediction models [39], [40]. The most significant benefit of the genetic algorithm is its ability to avoid being confined in local optima, and the use of GA or a hybrid GA gives the chance to select the best appropriate objective functions freely [41].

A recent study improved the accuracy of the self-organizing map (SOM), a type of unsupervised ANN, in predicting robotic manipulation failures for force-sensitive tasks using a genetic algorithm. The proposed hybrid GA-SOM model exhibited an increased accuracy prediction and improved the predictive capability of the SOM algorithm when used alone [42]. Moreover, the use of evolutionary technique like the genetic algorithm in enhancing ANN was observed along with SVM-Linear (L), SVM-Polynomial (P), SVM-Radial Basis Function (RBF), and CART in predicting the shape of carbon black reinforced rubbers. With the advent of the genetic algorithm, the prediction accuracy of each model has increased, and the most accurate model was obtained using GA-ANN hybrid model with those obtained using the GA-CART, GA-SVML, GA-SVM-P, and GA-SVM-RBF [43].

TABLE I. Indexed GA-Based Prediction Models

Algorithms/ Procedure	**Authors/ Year**	**Purpose**	**Significant Results**
Genetic Algorithm-based Self Organizing Map (GA-SOM) prediction model	Parisi & RaviChandran, (2018)	To enhance the performance of SOM using GA in predicting robotic failures	The hybrid GA-SOM yielded 91.95% prediction accuracy compared with the 84.96% prediction of the standalone SOM algorithm
Hybrid GA-ANN, GA-CART, GA-SVML, GA-SVMP, GA-SVM-RBF prediction model	Martinez et al., (2018)	To use GA, ANN, SVM, and CART to characterize rubber blends.	The GA-ANN model exhibited the finest classification accuracy of 75.75% improving the 74.80% accuracy attained without GA.
Genetic Algorithm-based Back Propagation (GA-BP) neural network prediction model	Zheng, Qian, Liu, & Liu, (2018)	Hybrid GA-BP neural network was used to model skid resistance of epoxy asphalt mixture	The optimized GA-BP neural network hybrid model was able to give an effective and accurate forecast of long term skid resistance with 99% accuracy.
Combination of Genetic Algorithm, Levenberg-Marquardt algorithm, and Back Propagation neural network as a prediction model	Zhou et al., (2018)	Application of GA-LM-BP Neural network in fault prediction of drying furnace equipment.	The GA-LM-BP hybrid prediction model obtained the decision coefficient R^2 of 0.97511 which is higher than the BP and GA-BP models.
The use of Genetic algorithm in least squares-support vector machine (LS-SVM), Back Propagation Neural Network (BPNN), and Random Forest (RF)	Liu et al., (2018)	Analyze the origin of extra virgin olive oils.	Simulation results showed that GA-LS-SVM model obtained 96.25% prediction accuracy and a prediction of 86.25% for GA-BPNN while a prediction accuracy of 82.5% for GA-RF was identified.
Genetic Algorithm-based Random Forest (RF) prediction model	Kumar & Sahoo, (2017)	To propose a hybrid GA-RF prediction model for cardiovascular disease diagnosis	Hybrid GA-RF prediction model outperformed the principal component analysis-based random forest (PCA-RF), Relief F-based random forest (Relief-F-RF), sequential forward floating search-based random forest (SFFS-RF), and sequential backward floating search-based random forest (SFBS-RF) having 93.2%, 84.8%, 85.4%, 79.1%, and 85.8% prediction accuracy, respectively.
Genetic Algorithm-based Artificial Neural Network (GA-ANN) prediction model	Armaghani et al., (2016)	To enhance the prediction rate of ANN in predicting AoP from blasting operation in granite quarry site.	GA-ANN model obtained 0.965 coefficient of determination, variance account for (VAF) value of 96.380 and RMSE of 0.049 than the ANN with those statistical function values of 0.857, 84.257, and 0.117 respectively.

Another study used the hybrid GA-BP neural network in predicting long-term skid resistance of epoxy asphalt mixture. The GA-BP model produced a great accuracy result when tested using the training set, validation set, and test set [44]. Meanwhile, the application of genetic algorithm, Levenberg-Marquardt (LM) algorithm, and backpropagation neural network were observed in fault prediction of drying furnace equipment. The hybrid GA-LM-BP model showed an increased prediction accuracy compared to both BP neural network and GA-BP neural network models [45].

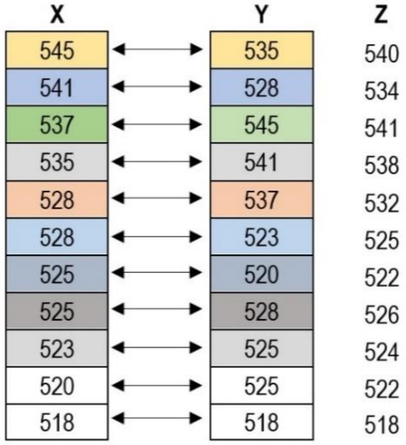
Further, the hybrid genetic algorithm-based least squares-support vector machine (GA-LS-SVM), genetic algorithm-based back propagation neural network (GA-BPNN), and genetic algorithm-based random forest (GA-RF) were employed in identifying the topographical origin of extra-virgin olive oils. The simulation results showed that GA-LS-SVM obtained the highest prediction accuracy for features selection methods compared to GA-BPNN and GA-RF models [46]. To further prove the superiority of GA as variable minimization algorithm, the genetic algorithm was used to perform feature selection where the extracted features are taken as an input to random forest (RF) classifier in accomplishing cardiovascular diagnostic problem. The outcome shows that the GA-RF model obtained the highest prediction accuracy rate when compared to other feature selection algorithms [47].

Lastly, an artificial neural network (ANN) and genetic algorithm-based ANN (GA-ANN) were proposed and evaluated to predict air overpressure from blasting operation in a granite quarry site in Penang, Malaysia. Simulation results proved the superiority of GA-ANN model in predicting air overpressure than using ANN algorithm alone [39]. The indexed GA-based prediction models are shown in Table I.

III. Methodology

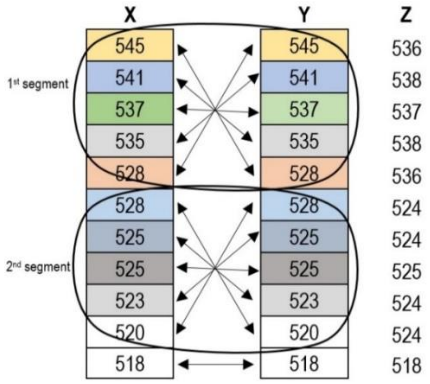
A. *Modified Genetic Algorithm for Variable Reduction*

To achieve the purpose of the study, the average crossover which is one of the crossover operators in the genetic algorithm as shown in Fig. 1, is modified. The modified crossover will be called Inversed Bi-segmented Average Crossover (IBAX) as depicted in Fig. 2. The use of rank-based selection function was observed in the simulation process.



X	Y	Z
545	535	540
541	528	534
537	545	541
535	541	538
528	537	532
528	523	525
525	520	522
525	528	526
523	525	524
520	525	522
518	518	518

Fig. 1. Average Crossover with Roulette Wheel Selection Function.



X	Y	Z
545	545	536
541	541	538
537	537	537
535	535	538
528	528	536
528	528	524
525	525	524
525	525	525
523	523	524
520	520	524
518	518	518

Fig. 2. Inversed Bi-Segmented Average Crossover with Rank-Based Selection Function.

For the IBAX operator to be realized, the following steps must be executed:

Step 1: Take the parents from the selection pool.

Step 2: Count the number of genes found in the chromosomes. Identify if the dataset is in odd or even numbers.

Step 3: Segment the chromosomes (x and y) by dividing the total number of genes in the chromosomes into two and make sure that both first and second segments must contain an equal number of genes in an even count.

Step 4: On 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 formula:

$$z = [x + y]/ 2 \tag{1}$$

Step 5: Execute the same process on the second segment until genes from all segments have produced offspring. In 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.

B. *K-Nearest Neighbor Algorithm*

Another recognized data mining algorithm for classification and prediction introduced by Fix and Hodges is the k-Nearest Neighbor (k-NN). This method adopts instance-based learning for prediction. The famous classifier is known as a non-parametric algorithm since it does not produce assumptions on the input data distribution; therefore, it is widely used in various applications [48], [49]. K-Nearest Neighbor (KNN) algorithm is simple and can be implemented through the following steps:

Step 1: Assign k values of the nearest neighbor of an instance in the algorithm.

Step 2: Perform the Euclidian distance calculation of each instance.

Step 3: Choose K neighboring attributes that have the lowest Euclidian distance.

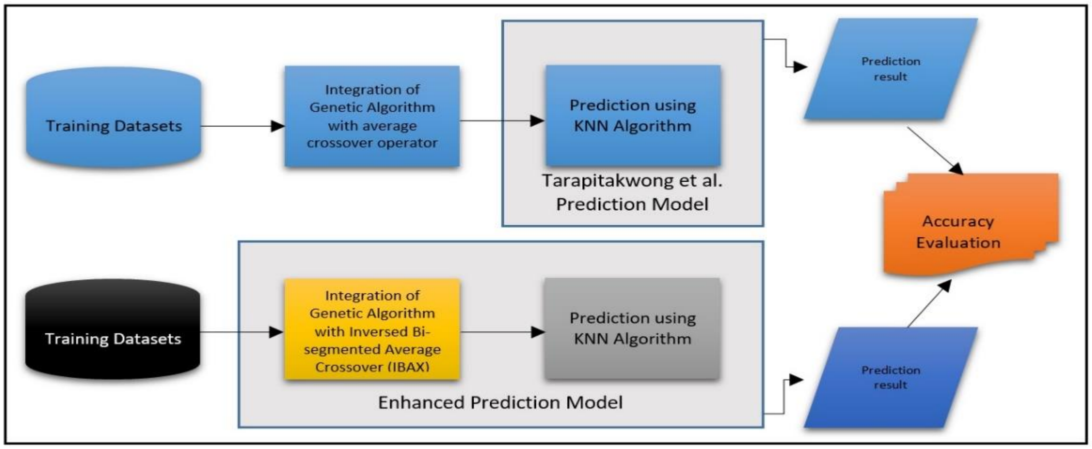

Training Datasets
Integration of Genetic Algorithm with average crossover operator
Prediction using KNN Algorithm
Tarapitakwong et al. Prediction Model
Prediction result
Accuracy Evaluation
Training Datasets
Integration of Genetic Algorithm with Inversed Bi-segmented Average Crossover (IBAX)
Prediction using KNN Algorithm
Enhanced Prediction Model
Prediction result

Fig. 3. Conceptual Framework of the Study.

A prediction model using K-Nearest Neighbor (KNN) algorithm was utilized in the study of [36] along with the many studies found in the literature.

C. Enhanced KNN Prediction Model

The study evaluated the accuracy level of [36] prediction model when integrated with GA having AX operator and with the modified GA with IBAX operator having 1, 3, 5, and 7 k values. The Waikato Environment for Knowledge Analysis (WEKA) version 3.8.2 was instrumental in the simulation of KNN prediction model. The simulation results of both existing and enhanced prediction models were compared to check the improvement rate of the accuracy level of the prediction model. The conceptual framework of the study is presented in Fig. 3.

D. Datasets

The datasets used in this study were the 597 records of student-respondents in the evaluation of the faculty instructional performance from the four State Universities and Colleges (SUC) in Caraga Region, Philippines. The thirty (30) variables that represent the faculty instructional performance (IP) having divided into six (6) parts viz., methodology, classroom management, student discipline, assessment of learning, student-teacher relationship, and peer relationship are reduced before the prediction to aid maximized accuracy. The 70% of the data were used as the training set while the remaining 30% were used for testing.

E. Prediction Evaluation

An optimal model is selected once the model with the highest prediction rate is identified granted that the model has the lowest root mean squared error and mean absolute error values. Countless forecasting and prediction models found in the literature are evaluated using the various forecast error statistical tools. The following tools listed below will be used along with Precision, Recall, and F-Measure:

Root Mean Squared Error (RMSE)

$$R.M.S.E. = \sqrt{\sum_{t=T+1}^{T+h}(\hat{y}_t - y_t)^2/h} \qquad (2)$$

Mean Absolute Error (MAE)

$$M.A.E. = \sum_{t=T+1}^{T+h}|\hat{y}_t - y_t|/h \qquad (3)$$

IV. Results and Discussion

A. Variable Minimization using GA with AX and IBAX Operators

The simulation on the genetic algorithm was done for ten generations utilizing the existing traditional average crossover and roulette wheel selection function. To generate new offspring from the two chromosomes (IP and Y), the average crossover was used where the average of the two chromosomes/parents was calculated. The new fitness values are then calculated based on the new offspring produced after the crossover function. Variables having the lowest fitness value were removed from the dataset. The sample simulation on the genetic algorithm having the original AX operator and roulette wheel selection function is presented in Table II.

First Generation: Variable C2 is removed from the chromosome since it obtained the lowest fitness value of 171396 as evident in Table II.

On the other hand, the simulation on the genetic algorithm with the novel Inversed Bi-segmented Average Crossover (IBAX) operator and rank-based selection function was done on the same dataset and number of generations.

First Generation: Variable C2 was removed from the list of variables after applying the rank-based selection. The variable C2 obtained the lowest fitness value in the rank-based selection. Hence, it does not have any chance to be selected. Moreover, after applying the inversed bi-segmented average crossover (IBAX) operator and obtained the fitness value of the offspring, variable C3 was removed from the chromosomes since it obtained the lowest fitness value of 224676 that will not warrant for the next generation. Thus, in the first generation, there were two variables removed from the list as shown in Table III.

Prior to prediction, the variables were minimized using GA with AX operator having roulette wheel selection function and GA with IBAX operator having rank-based selection function performed for ten generations.

TABLE II. GENERATION 1 USING AN AX OPERATOR WITH RWS FUNCTION

IP	X	Fitness	Rank	Pool		Off-spring	Fitness	Decision
				Y	IP			
M1	546	298116	22	552	M5	549	301401	
M2	565	319225	30	565	M2	565	319225	
M3	558	311364	27	548	SD1	553	305809	
M4	559	312481	28	546	SD5	552.5	305256.3	
M5	552	304704	24	474	C3	513	263169	
C1	490	240100	3	546	A3	518	268324	
C2	354	125316	1	474	C3	414	171396	Remove
C3	474	224676	2	556	ST1	515	265225	
C4	542	293764	18	490	C1	516	266256	
C5	528	278784	12	531	ST3	529.5	280370.3	
SD1	548	300304	23	500	A2	524	274576	
SD2	512	262144	5	542	C4	527	277729	
SD3	565	319225	29	546	A3	555.5	308580.3	
SD4	556	309136	26	558	M3	557	310249	
SD5	546	298116	21	528	C5	537	288369	
A1	513	263169	6	534	ST2	523.5	274052.3	
A2	500	250000	4	526	P3	513	263169	
A3	546	298116	20	513	A1	529.5	280370.3	
A4	518	268324	8	565	M2	541.5	293222.3	
A5	516	266256	7	516	A5	516	266256	
ST1	556	309136	25	556	ST1	556	309136	
ST2	534	285156	16	559	M4	546.5	298662.3	
ST3	531	281961	14	546	M1	538.5	289982.3	
ST4	541	292681	17	556	SD4	548.5	300852.3	
ST5	527	277729	11	552	M5	539.5	291060.3	
P1	531	281961	13	565	SD3	548	300304	
P2	533	284089	15	541	ST4	537	288369	
P3	526	276676	10	518	A4	522	272484	
P4	526	276676	9	565	SD3	545.5	297570.3	
P5	544	295936	19	565	SD3	554.5	307470.3	

The variable minimization process using the genetic algorithm with AX operator and RWS function has depicted a decrease after the ten generations. From the 30 variables, it was minimized to 17 with a total reduction of 43%. Meanwhile, the variable minimization process using the genetic algorithm with the proposed novel mating scheme called inversed bi-segmented average crossover operator, and rank-based selection function has depicted a noticeable decrease after the ten generations. From the 30 variables, the numbers were minimized to 10 variables after the generations. A total of 66.66% of variables were removed as depicted in Table IV.

The simulation result showed that the modified genetic algorithm with a new crossover mating scheme outperformed the average crossover of genetic algorithm in reducing variables prior to prediction. Since dropping one or more variables helps reduce dimensionality, predictions using the dataset having 17 and 10 variables were conducted using the KNN algorithm.

Meanwhile, in the extent of fitness function, the proposed IBAX operator of the genetic algorithm has increased and outperformed the rate of the fitness functions generated using the genetic algorithm with the existing AX operator. The variables who obtained the lowest fitness function in each generation for ten generations were removed. It is evident in Fig. 4 and Table V that the fitness functions that were removed using the new crossover operator is higher compared to the fitness functions that were removed using the existing average crossover. This denotes that the modified genetic algorithm has managed to increase the fitness function of the variables compared to the genetic algorithm with traditional AX operator.

TABLE III. GENERATION 1 USING IBAX WITH THE RANK-BASED SELECTION FUNCTION

			Rank-based			IBAX			
IP	X	Fitness	Rank	New Fitness		Parent 1	Parent 2	Offspring	Fitness
M2	565	319225	30	3629986.8		565	541	553	305809
SD3	565	319225	29	3504670.8		565	542	553.5	306362.3
M4	559	312481	28	3379354.8		559	544	551.5	304152.3
M3	558	311364	27	3254038.8		558	546	552	304704
SD4	556	309136	26	3128722.8		556	546	551	303601
ST1	556	309136	25	3003406.8		556	546	551	303601
M5	552	304704	24	2878090.8		552	548	550	302500
SD1	548	300304	23	2752774.8		548	552	550	302500
M1	546	298116	22	2627458.8		546	556	551	303601
SD5	546	298116	21	2502142.8		546	556	551	303601
A3	546	298116	20	2376826.8		546	558	552	304704
P5	544	295936	19	2251510.8		544	559	551.5	304152.3
C4	542	293764	18	2126194.8		542	565	553.5	306362.3
ST4	541	292681	17	2000878.8		541	565	553	305809
ST2	534	285156	16	1875562.8		534	490	512	262144
P2	533	284089	15	1750246.8		533	500	516.5	266772.3
ST3	531	281961	14	1624930.8		531	512	521.5	271962.3
P1	531	281961	13	1499614.8		531	513	522	272484
C5	528	278784	12	1374298.8		528	516	522	272484
ST5	527	277729	11	1248982.8		527	518	522.5	273006.3
P3	526	276676	10	1123666.8		526	526	526	276676
P4	526	276676	9	998350.8		526	526	526	276676
A4	518	268324	8	873034.8		518	527	522.5	273006.3
A5	516	266256	7	747718.8		516	528	522	272484
A1	513	263169	6	622402.8		513	531	522	272484
SD2	512	262144	5	497086.8		512	531	521.5	271962.3
A2	500	250000	4	371770.8		500	533	516.5	266772.3
C1	490	240100	3	246454.8		490	534	512	262144
C3	474	224676	2	121138.8		474	474	474	224676
C2	354	125316	1	-4177.2					

TABLE IV. VARIABLE MINIMIZATION SIMULATION RESULT FOR GENETIC ALGORITHMS WITH AX AND IBAX OPERATORS

Basic GA with AX Operator						**Proposed GA with IBAX operator**				
Number of Generations	Number of Variables Left	Number of Variables Removed	Variables Removed	Percentage		Number of Generations	Number of Variables Left	Number of Variables Removed	Variables Removed	Percentage
1	30	1	C2	3.33%		1	30	2	C3, C2	6.66%
2	29	2	M5, A2	6.66%		2	28	2	ST2, C1	6.66%
3	27	1	C3	3.33%		3	26	2	ST4, A2	6.66%
4	26	2	C4, A5	6.66%		4	24	2	P2, A1	6.66%
5	24	1	C1	3.33%		5	22	2	ST3, A4	6.66%
6	23	1	P3	3.33%		6	20	2	C5, ST5	6.66%
7	22	1	A1	3.33%		7	18	2	SD5, SD2	6.66%
8	21	1	SD1	3.33%		8	16	2	M1, A5	6.66%
9	20	1	SD2	3.33%		9	14	2	A3, P3	6.66%
10	19	2	C5, A3	6.66%		10	12	2	M5, P4	6.66%
10	17	-	-	-		10	10	-	-	-
Total Percentage of Variables Removed				**43.33%**		**Total Percentage of Variables Removed**				**66.66%**

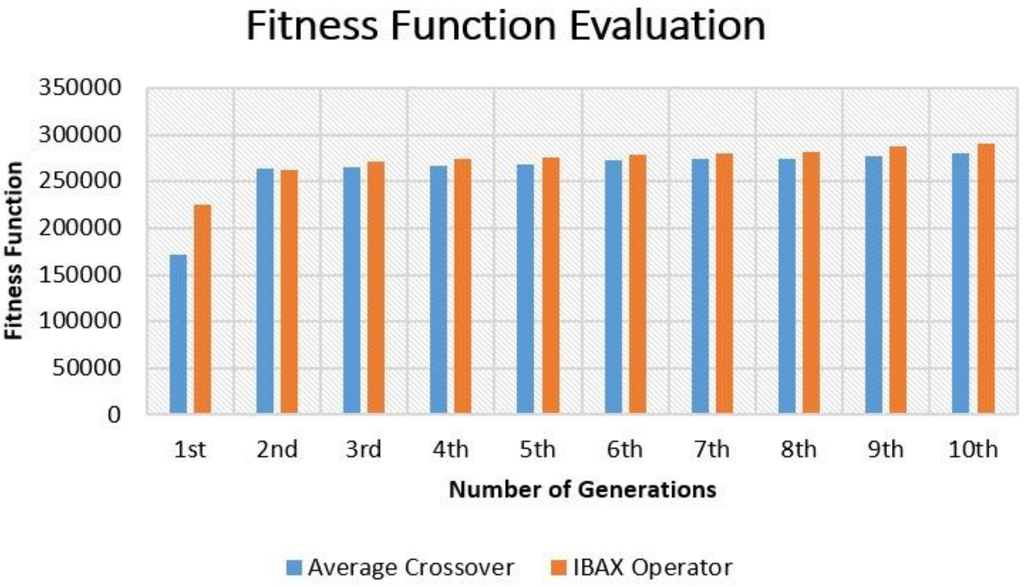


Fig. 4. Comparison of the Fitness Function of the Removed Variables in Every Generation.

B. Prediction Model Accuracy Evaluation

To evaluate the accuracy level of KNN as a prediction model, thirty percent (30%) of the data were used for testing while seventy percent (70%) were used as the training set. Table VI shows the comparison of results when GA with AX operator and roulette wheel selection function is used, and GA with IBAX operator with rank-based selection function are integrated prior to the prediction using KNN. The predictive capability of the KNN algorithm was also tested without the variable reduction stage and obtained a 90.50% prediction accuracy rate with a k value of 1.

The results showed that the prediction model gained an increase in the accuracy when integrated with genetic algorithm especially with the modified GA. The optimal model for predicting the instructional performance of the faculty in the four SUCs in the Caraga Region, Philippines is the KNN with a k value of 3. The model obtained a high 95.53% prediction accuracy. Meanwhile, the second best model that has 94.97% accuracy is the KNN with k=5 where its MAE and RMSE values are 0.08 and 0.20, respectively.

TABLE V. VALUE OF THE FITNESS FUNCTIONS REMOVED IN EVERY GENERATION

Number of Generations	AX Operator	IBAX Operator
1	171396	224676
2	263169	262144
3	265225	270920.3
4	266256	273529
5	268324	275100.3
6	272484	278256.3
7	274052.3	279841
8	274576	281961
9	277729	287296
10	280370.3	290521

TABLE VI. INDEXED KNN AND GA-BASED KNN PREDICTION MODELS

Model	K Value	Accuracy %	MAE	RMSE	Precision	Recall	F- Measure
KNN algorithm alone	1	90.5028%	0.0978	0.3055	0.907	0.905	0.906
	3	87.1508%	0.1379	0.3241	0.872	0.872	0.872
	5	84.3575%	0.1587	0.3311	0.840	0.844	0.841
	7	84.9162%	0.1649	0.321	0.846	0.849	0.846
GA-KNN with original AX and Roulette Wheel Selection	1	93.8547%	0.069	0.2434	0.938	0.939	0.938
	3	89.9441%	0.1196	0.261	0.899	0.899	0.899
	5	89.9441%	0.1293	0.2557	0.899	0.899	0.899
	7	89.3855%	0.1377	0.2652	0.893	0.894	0.894
GA-KNN with new IBAX and Rank Based Selection	1	94.4134%	0.0586	0.2364	0.944	0.944	0.943
	3	**95.5307%**	**0.0662**	**0.1963**	**0.956**	**0.955**	**0.955**
	5	94.9721%	0.0828	0.2067	0.953	0.950	0.948
	7	94.9721%	0.0956	0.2175	0.953	0.950	0.948

V. CONCLUSION AND RECOMMENDATION

With the integration of the genetic algorithm, the prediction model using the KNN algorithm has increased its prediction accuracy. The modified genetic algorithm with a new crossover mating scheme called Inversed Bi-segmented Average Crossover (IBAX) showed a considerably high prediction percentage than the genetic algorithm with average crossover having the roulette wheel as the selection function. Along with the GA-based prediction models found in the literature, the enhancement on the KNN as prediction model integrated with the modified genetic algorithm was a success and is added to the body of knowledge. Future researchers may consider using the modified GA-based KNN on different datasets as a prediction model.

REFERENCES

[1] Savaliya, A. Bhatia, and J. Bhatia, “Application of Data Mining Techniques in IoT: A Short Review,” Int. J. Sci. Res. Sci. Eng. Technol., vol. 4, no. 2, pp. 218–223, 2018.

[2] C. Suresh, K. T. Reddy, and N. Sweta, “A Hybrid Approach for Detecting Suspicious Accounts in Money Laundering Using Data Mining Techniques,” Int. J. Inf. Technol. Comput. Sci., vol. 8, no. 5, pp. 37–43, 2016.

[3] K. Rajalakshmi, S. S. Dhenakaran, and N. Roobini, “Comparative Analysis of K-Means Algorithm in Disease Prediction,” Int. J. Sci. Eng. Technol. Res., vol. 4, no. 7, pp. 2697–2699, 2015.

[4] I. A. Khan and J. T. Choi, “An application of educational data mining (EDM) technique for scholarship prediction,” Int. J. Softw. Eng. its Appl., vol. 8, no. 12, pp. 31–42, 2014.

[5] E. Sugiyarti, K. A. Jasmi, B. Basiron, M. Huda, K. Shankar, and A. Maseleno, “Decision support system of scholarship grantee selection using data mining,” Int. J. Pure Appl. Math., vol. 119, no. 15, pp. 2239–2249, 2018.

[6] E. Susnea, “Using data mining techniques in higher education,” in The 4th International COnference on Virtual Learning ICVL 2009, 2009, vol. 1, no. 1, pp. 371–375.

[7] E. Petrova, P. Pauwels, K. Svidt, and R. L. Jensen, “Advances in Informatics and Computing in Civil and Construction Engineering,” in 35th International Council for Research and Innovation in Building Construction W78 2018 Conference, 2019, pp. 19–26.

[8] R. Kitchin, “Big Data, new epistemologies and paradigm shifts,” Big Data Soc., vol. 1, no. 1, pp. 1–12, 2014.

[9] B. M. M. Alom and M. Courtney, “Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia,” Int. J. Inf. Technol. Comput. Sci., vol. 10, no. 2, pp. 1–9, 2018.

[10] A. Ikhwan et al., “A Novelty of Data Mining for Promoting Education Based on FP-Growth Algorithm,” Int. J. Civ. Eng. Technol., vol. 9, no. 7, pp. 1660–1669, 2018.

[11] M. M. Malik, S. Abdallah, and M. Ala’raj, “Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review,” Ann. Oper. Res., vol. 270, no. 1–2, pp. 287–312, 2018.

[12] R. Wadhawan, “Prediction of Coronary Heart Disease Using Apriori algorithm with Data Mining Classification,” Int. J. Res. Sci. Technol., vol. 3, no. 1, pp. 1–15, 2018.

[13] S. A. Aljawarneh, O. Bayat, and M. Essaaidi, “Introduction to the special section on new trends in data mining, games engineering and database systems,” Comput. Electr. Eng., vol. 66, pp. 420–422, 2018.

[14] P. Kaur, M. Singh, and G. S. Josan, “Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector,” Procedia Comput. Sci., vol. 57, pp. 500–508, 2015.

[15] M. Brilliant, DwiHandoko, and Sriyanto, “Implementation of Data Mining Using Association Rules for Transactional Data Analysis,” 3rd Int. Conf. Inf. Technol. Bus., pp. 177–180, 2017.

[16] 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, vol. 489, pp. 78–93, 2018.

[17] M. Zaffar, M. Ahmed, K. S. Savita, and S. Sajjad, “A Study of Feature Selection Algorithms for Predicting Students Academic Performance,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 5, pp. 541–549, 2018.

[18] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, “Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil,” J. Bus. Res., vol. 94, no. August 2017, pp. 335–343, 2019.

[19] R. Ahuja, A. Jha, R. Maurya, and R. Srivastava, Analysis of Educational Data Mining, vol. 741. Springer Singapore, 2019.

[20] S. Prabakaran and S. Mitra, “Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning,” J. Phys. Conf. Ser., vol. 1000, no. 1, pp. 1–10, 2018.

[21] O. Vaidya, S. Mitra, R. Kumbhar, S. Chavan, and R. Patil, “Comprehensive Comparative Analysis of Methods for Crime,” Int. Res. J. Eng. Technol., pp. 715–718, 2018.

[22] P. Vrushali, M. Trupti, G. Pratiksha, and G. Arti, “Crime Rate Prediction using KNN,” Int. J. Recent Innov. Trends Comput. Commun., vol. 6, no. 1, pp. 124–127, 2018.

[23] V. Ravi Jain, M. Gupta, and R. Mohan Singh, “Analysis and Prediction of Individual Stock Prices of Financial Sector Companies in NIFTY50,” Int. J. Inf. Eng. Electron. Bus., vol. 10, no. 2, pp. 33–41, 2018.

[24] P. Carmona, F. Climent, and A. Momparler, “Predicting failure in the U.S. banking sector: An extreme gradient boosting approach,” Int. Rev. Econ. Financ., pp. 1–54, 2018.

[25] A. Kiruthika, P. Deepika, S. Sasikala, and S. Saranya, “Predicting Ailment of Thyroid Using Classification and Recital Indicators,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 3, no. 3, pp. 1481–1485, 2018.

[26] K. Lakshmi, D. I. Ahmed, and G. Siva Kumar, “A Smart Clinical Decision Support System to Predict diabetes Disease Using Classification Techniques,” 2018 Ijsrset, vol. 4, no. 1, 2018.

[27] S. García, J. Luengo, and F. Herrera, “Tutorial on practical tips of the most influential data preprocessing algorithms in data mining,” Knowledge-Based Syst., vol. 98, pp. 1–29, 2016.

[28] A. Baldominos, P. Isasi, and U. C. I. I. I. De Madrid, “Feature Set Optimization for Physical Activity Recognition Using Genetic Algorithms,” Proc. Companion Publ. 2015 Genet. Evol. Comput. Conf. - GECCO Companion ’15, pp. 1311–1318, 2015.

[29] C.-F. J. Kuo, C.-H. Lin, and M.-H. Lee, “Analyze the energy consumption characteristics and affecting factors of Taiwan’s convenience stores-using the big data mining approach,” Energy Build., vol. 168, pp. 120–136, 2018.

[30] I. M. El-hasnony, H. M. El Bakry, and A. A. Saleh, “Comparative Study among Data Reduction Techniques over Classification Accuracy,” Int. J. Comput. Appl., vol. 122, no. 2, pp. 9–15, 2015.

[31] F. Herrera and S. Garc, “Prototype Selection for Nearest Neighbor Classification Taxonomy and Empirical Study,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 3, pp. 417–435, 2012.

[32] Y. Cheng, K. Chen, H. Sun, Y. Zhang, and F. Tao, “Data and knowledge mining with big data towards smart production,” J. Ind. Inf. Integr., vol. 9, pp. 1–13, 2018.

[33] M. Mafarja, I. Aljarah, A. A. Heidari, A. I. Hammouri, H. Faris, and A. M. Al-zoubi, “Evolutionary Population Dynamics and Grasshopper Optimization Approaches for Feature Selection Problems,” Knowledge-Based Syst., vol. 154, no. 7, pp. 25–45, 2017.

[34] A. Daraei and H. Hamidi, “An efficient predictive model for myocardial infarction using cost-sensitive J48 model,” Iran. J. Public Health, vol. 46, no. 5, pp. 682–692, 2017.

[35] S. Moedjiono, Y. R. Isak, and A. Kusdaryono, “Customer Loyalty Prediction In Multimedia Service Provider Company With K-Means Segmentation And C4 . 5 Algorithm,” in 2016 International Conference on Informatics and Computing (ICIC), 2016, pp. 1–6.

[36] J. Tarapitakwong, B. Chartrungruang, and N. Tantranont, “A Classification Model for Predicting Standard Levels of OTOP ’ s Wood Handicraft Products by Using the K-Nearest Neighbor,” Int. J. Comput. Internet Manag., vol. 25, no. 2, pp. 135–141, 2017.

[37] D. García-gil, J. Luengo, S. García, and F. Herrera, “Enabling Smart Data : Noise filtering in Big Data classification,” Inf. Sci. J., vol. 479, pp. 135–152, 2019.

[38] U. O. Cagas, A. J. P. Delima, and T. L. Toledo, “PreFIC : Predictability of Faculty Instructional Performance through Hybrid Prediction Model,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 7, pp. 22–25, 2019.

[39] D. J. Armaghani, M. Hasanipanah, A. Mahdiyar, M. Z. A. Majid, H. B. Amnieh, and M. M. D. Tahir, “Airblast prediction through a hybrid genetic algorithm-ANN model,” Neural Comput. Appl., vol. 29, no. 9, 2018.

[40] V. Rashidian and M. Hassanlourad, “Predicting the Shear Behavior of Cemented and Uncemented Carbonate Sands Using a Genetic Algorithm-Based Artificial Neural Network,” Geotech. Geol. Eng., vol. 31, no. 4, pp. 1231–1248, 2013.

[41] L. D. Chambers, Practical handbook of genetic algorithms: complex coding systems. CRC Press, 2010.

[42] L. Parisi and N. RaviChandran, “Genetic Algorithms and Unsupervised Machine Learning for Predicting Robotic Manipulation Failures for Force-Sensitive Tasks,” in 2018 4th International Conference on Control, Automation and Robotics (ICCAR), 2018, pp. 22–25.

[43] R. F. Martinez, P. Jimbert, J. Ibarretxe, and M. Iturrondobeitia, “Use of support vector machines, neural networks and genetic algorithms to characterize rubber blends by means of the classification of the carbon black particles used as reinforcing agent,” Soft Comput., pp. 1–10, 2018.

[44] D. Zheng, Z. Qian, Y. Liu, and C. Liu, “Prediction and sensitivity analysis of long-term skid resistance of epoxy asphalt mixture based on GA-BP neural network,” Constr. Build. Mater., vol. 158, pp. 614–623, 2018.

[45] W. Zhou, D. Liu, and T. Hong, “Application of GA-LM-BP Neural Network in Fault Prediction of Drying Furnace Equipment,” in MATEC Web of Conferences, 2018, vol. 232, pp. 1–5.

[46] W. Liu et al., “Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics,” Food Chem., vol. 251, pp. 86–92, 2018.

[47] S. Kumar and G. Sahoo, “A random forest classifier based on genetic algorithm for cardiovascular diseases diagnosis,” Int. J. Eng., vol. 30, no. 11, pp. 1723–1729, 2017.

[48] A. Tharwat, H. Mahdi, M. Elhoseny, and A. E. Hassanien, “Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm,” Expert Syst. Appl., vol. 107, pp. 32–44, 2018.

[49] M. Huang, R. Lin, S. Huang, and T. Xing, “A novel approach for precipitation forecast via improved K-nearest neighbor algorithm,” Adv. Eng. Informatics, vol. 33, pp. 89–95, 2017.