Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods

Phauk Sokkhey[1]
Graduate School of Engineering and Science
University of the Ryukyus
Senbaru, Nishihara, Okinawa, 903-0123, Japan
Institute of Technology of Cambodia, Phnom Penh
Cambodia

Takeo Okazaki[2]
Department of Computer Science and Intelligent Systems
University of the Ryukyus
1 Senbaru, Nishihara,
Okinawa, 903-0123,
Japan

***Abstract*—All educational institutions always try to investigate the learning behaviors of students and give early prediction toward student's outcomes for interventing and improving their learning performance. Educational data mining (EDM) offers various effective prediction models to predict student performance. Simultaneously, feature selection (FS) is a method of EDM that is utilized to determine the dominant factors that are needed and sufficient for the target concept. FS method extracts high-quality data that reduce the complexity of the prediction task that can increase the robustness of decision rule. In this paper, we provide a comparative study of feature selection methods for determining dominant factors that highly affect classification performance and improve the performance of prediction models. A new feature selection CHIMI based on ranked vector score is proposed. Analysis of feature sets of each FS method to get the dominant set is executed. The experimental results show that by using the dominant set of the proposed CHIMI method, the classification performance of the proposed models is significantly improved.**

Keywords—Educational data mining; dominant factors; feature selection methods; prediction models; student performance

I. Introduction

The development of developing countries mainly relies on the potential education system that can produce human resources. The success of human resource development depends on long-term investment in education from primary schools, secondary schools, and higher educations. Student performance in high school plays an important role that maximizes or minimizes student success in the secondary school national exam, higher education, and their future careers [1]. Student academic performance can be measured and monitored effectively by using methods of educational data mining (EDM).

Various EDM techniques are effectively used to predict student performance, identify their learning behaviors and progress, and many more [2-5]. The results of these tasks are helpful for students themselves, academic institutions, and related individuals to follow up academic performance, improve student performance, and use as information for planning and scheduling in education systems. In EDM, feature selection (FS) is used in many research work [6-8].

The prediction of student performance highly depends on the choice of selection of most relevant variables. In the educational domain, several factors were concerned to influence academic performance, mainly considers school environment factors, domestics environment factors, demographic background, attitudinal factors, and academic records. Various factors lead to have higher dimensions. Hence, many studies have focused on determining the related factors that affect student performance and predict their academic progress by using FS methods and applying predictive models of EDM [11-16]. In educational research, FS methods aim at determining important factors that are in need and sufficient to report the academic performance; we call it as *dominant factors*.

The dominant set was considered for two main contributtions. Interm of gving intervention to poor-performing students, dominant set is known as the set of important factors that affect student performance. Another most common contribution is to raise the performance of prediction models. FS methods are categorized into 3 classes: filter-based methods, wrapper-based methods, and embedded/hybrid methods [5]. wrapper method and hybrid-based features selection methods are effective, yet computational expesnive to detect the optimal sets in big data content [9]. Filter-based is a simple FS method, yet effective for all types of datasets. In addition, Filter is independence of classifiers and more scalable comparing to other FS methods [10]. The main objective of FS is to select optimal subsets consisting of relevant and informative features.

II. Literature Review

A. Feature Selection Methods and Prediction Models in EDM

This part presents a brief of previous works that have used FS techniques to enhance the performance of the prediction models of EDM.

Estrera et al. [11] gave analysis on high school record of student enrolled for a university. The analysis of the study proposed decision tree (DT), naive Bayes (NB), and k-nearest neighbor (KNN). The decision tree algorithm generates affective results in this classification and prediction problem. Several FS methods were utilized to improved the performance of proposed models and to detect student learning patterns. The proposed FS methods are: Chi-square Statistics (CHS) test, Information Gain (IG) test, and Information Gain Ration (IGR) test in the study. Experimental results indicated the decision tree produced the most satisfied accuracy.

Ramaswami et al. [12] developed a comparative study of six filter-based feature selection methods for improving academic performance prediction. The used algorithms are correlation-based feature evaluator (CB), Chi-square feature evaluator (CH), gain ration feature evaluator (GR), information gain feature evaluator (IG), Relief (RF), and Symmetric Uncertainty (SU). The results indicated an increase in prediction performance and reduced time consumption.

Febro [13] utilized feature selection methods to improve prediction models and extract important features that affect student retention in higher education. Three filter-based selection methods (Information Gain Ratio (IGR), Correlation-based Feature Selection (CFS), and Chi-square (CHS)) are introduced. The optimal subset of 14 features was extracted from an original set of 29 features. The accuracy result jumps to 92.09%.

Zaffar et al. [14] proposed a study of feature selection techniques to enhance the prediction performance of academic performance. The FS techniques utilized in the study are: correlation based feature evaluator (CFS), Chi-squared test, the fiiltered, gain ration (GR), principal component aanlysis (PC), and Relief method. To comfirm the performance of the proposed FS techniques, the study utilized fifteen prediction models and make comparision of the models on each FS methods. The experiment indicates the improvement of accuracy when applying feature selection.

Alhassan et al. [15] proposed a study of analyzing student learning behaviors and predicting their academic performance in web-based learning management systems (LMS). The study observed the student learning patterns on the online study platform using five machine learning classifiers: J48 of decision tree, random forest (RF), the logistic regression (LR), sequential minimal optimization (SMO), and multilayer perceptron (MLP). Analysis of all feature sets and subsets of features are conducted by using six feature selection methods: Correlation Attribute Evlaition, Information Gain, CFS Subset, Wrapper-J48, Wrapper-NB, Wrapper-IBK. The RF algorithm combined with the feature selection methods outperformed the rest models.

Mythili et al. [16] proposed an analysis of student performance by applying data mining algorithms. The various classification algorithms used in the study are J48, Random Forest (RF), Multilayer Perceptron (MLP), Artificial Neural Network (ANN), IBI of the nearest neighbor classifier, and Decision Table. The experiment was conducted by filtering the important features using Information Gain (IG), and Gain Ratio (GR). With the merit of selecting only high ranking important features, it is discovered that RF performance is the best than that of different algorithms employed in the study.

B. *The Current Study*

Even if several works of literature have studied using FS algorithms in EDM, however, a lot of attention and consiseration are needed to build academic performance prediction model with the analysis and help of FS methods. The primary purpose of our study is to present an analysis of feature selection methods to extract the *dominant factors* that are necessary and sufficient to evaluate the success of students' performance. The primary purpose of this study is to introduce a study of analysis on feature selection methods on a set of classifiers and then determine the performance of each algorithm on each classifier. The study proposed a novel FS algorithm to improve the performance of predictive models. We search for an optimal and effective subset that improves the classification performance of classifiers. Consequently, we can obtain the potential prediction model and the dominant factors that maximize and control the evaluation of student performance. The proposed framework of this study is shown in Fig. 1.

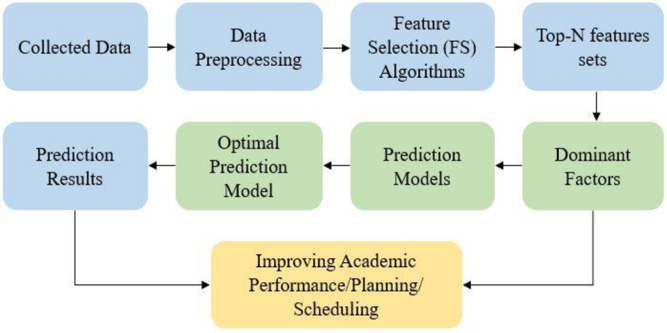

Fig. 1. The Proposed Framework of this Study.

III. Methodology

A. *Participants and Data*

The target of the study is to utilized related factors to predict student performace in high school.The proposed data in the study was collected from serval high schools in Cambodia. The questionnaire concerning with any related factors that affect (weak or strong) the student performance and it was designed into five main parts. The first part consist of six questions conecerning with student performanal information. The second part related to domestics or home factors, which has 17 questions. The third part has 15 questions including any information related with student learning behaviors and materials in study. The forth part consists of 14 questions in total concerning with school factors. The last part is the student score for output variables. The describption is illustrated in Table I.

B. *Data Preprocessing*

Preprocessing task is treated as a necessary step in every modeling. The utilized data mining models usually require data cleansing, data encoding, and data transformation to convert the data into an executable format and enhance model performance. The tool that is used in preprocessing and experiment in this study is R, a powerful tool for machine learning and statistical computation.

C. *Model Evaluation Metrics*

Evaluating model is a core part of EDM work, two standard model evaluation metrics are utilized in this study. Accuracy and root mean square error are the two commonly used metrics evaluating predction models.

- Accuracy (ACC): ACC is a common model evaluation metric used to evlauce the performance of prediction model by computing the percentage of coreectly prediction [15]. It is calculated as in (1).

$$\text{ACC} = \frac{\text{Correctly predicted values}}{\text{Total values}} \tag{1}$$

- Root Mean Square Error (RMSE): RMSE is a standard evaluation metric used to evaluate prediction error by computing the error or difference between actual output and predicted output [15]. It can be calculated using (2).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^a - y_i^p)^2} \tag{2}$$

D. Feature Selection Methods and Dominant Set

Analysis of student's information, their learning behavior, and factors affecting students' academic performance is still a challenging task in educational institutes [13]. Many cognitive and non-cognitive factors affect the academic performance of children and adolescents. Several related domains weakly or highly influence on results and achievements of high school students. Various factors lead to have higher dimensions. Hence, this study focus on determining the related factors that affect student performance and improved the proposed EDM with the merit of data set determining by FS techniques. This feature set is called the dominant set.

TABLE I. FEATURES AFFECTING STUDENTS PERFORMANCE

Factors	ID	Predictors (number of questions)	Data types
		Student personal information (6)	
Domestic	PEDU	Parents' educational levels (2)	Nominal
	POCC	Parents' occupational status (2)	Nominal
	PSES	Parents' socioeconomic levels (3)	Ordinal
	PI	Parents' involvement (4)	Ordinal
	PS	Parenting styles (4)	Ordinal
	DE	Domestic environment (2)	Ordinal
Student	SELD	Self-regulation on study (5)	Ordinal
	SIM	Interest and motivation (4)	Ordinal
	ANXI	Students' anxiety toward their classes and exams (3)	Ordinal
	POSS	Possession materials for study (3)	Nominal
School	CENV	School and class environment (1)	Ordinal
	CU	Curriculum (2)	Nominal
	TMP	Teaching methods and practices (4)	Ordinal
	TAC	Teachers' attribute & characteristics (4)	Ordinal
	ARES	Academic resource (3)	Nominal
SCORE	PL	Student's performance level based on their mark or score	Ordinal

The dominant set was considered for two main contributtions. Interm of gving intervention to poor-performing students, dominant set is known as the set of important factors that affect student performance. Another most common contribution is to improve the performance of prediction mod. From the literature reviews, filter-based feature selection methods is the most popular method in the research of educational domain. Filter-based is a simple FS method, yet effective for all types of datasets. In addition, Filter is independence of classifiers and more scalable comparing to other FS methods [17]. In this study, we propose comparative approach of experiement of a proposed feature selection method to three existing baseline methods.

1) Information Gain (IG): IG is one of the popularly used feature selection methods in data mining. IG utilized the entropy-based method to capture the importance of features [17]. Entropy class C prior to feature F is expressed as:

$$H(C) = -\sum p(c) \times \log_2 p(c), \tag{3}$$

where $p(c)$ is a marginal function of density probability for class C. The conditional entropy of class C for a feature F is denoted as:

$$H(C \mid F) = \sum_{i=1}^{m} \frac{|C_i|}{|C|} H(C_i), \tag{4}$$

where $C = \{C_1, C_2, ..., C_m\}$ is the m partition of class C. The IG of class C as acquired from a feature F is given by the following equation:

$$IG(C,F) \ = H(C) - H(C \mid F). \tag{5}$$

2) Symmetric Uncertainty (SU): SU is one of the leading feature selection techniques [17]. SU determines the correlation between a feature and target variable using entropy and information gain theory as in equation (6):

$$SU(A,D) = \frac{IG(D,A)}{H(D) + H(A)}, \tag{6}$$

where $H(D)$ and $H(A)$ are entropies of based on probability of class associated with the example set D and attribute A, respectively; $IG(D,A)$ is the information gain as shown in equation (5).

3) Mutual information (MI): MI of two random variables or features is a measure quantifies the dependence measurement between those variables [18]. It is asymmetric measurement such that $I(X,Y) = I(Y,X)$ that can recognize non-linear relationships between variables. MI of two discrete variables X and Y can be described as:

$$I(X,Y) = \sum_{x,y} p_{xy}(x,y) \log \frac{p_{xy}(x,y)}{p_x(x) \times p_y(y)}, \tag{7}$$

where $p_x(x)$ and $p_y(y)$ are marginal probability such that

$$p_x(x) = \sum_y p_{xy}(x,y), \;\; p_y(y) = \sum_x p_{xy}(x,y). \tag{8}$$

If the variables X and Y are independent, then the joint probability $p_{xy}(x,y) = p_x(x) \times p_y(y)$ and $I(x,y) = 0$.

4) Chi-square (CHI): CHI is a statistical method that is utilized for measuring the dependency of each input feature to the target class. The technique utilized the feature score from Chi-square test to get the rank list of all input features [9]. The list of an informative feature set can be computed using the equation below:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}}, \tag{9}$$

where l denotes the number of classes or determined intervals of a particular feature; c represents the number of classes in target variable; n_{ij} is the observed (actual) frequency of sample of i^{th} interval and j^{th} class; ε_{ij} indicate the expected frequency of n_{ij}.

5) The Proposed FS Method (CHIMI): Many studies have confirmed the effectiveness of information gain (IG), symetric uncertainty (SU), and mutual infromation (MI) in many applications. The techniques use the concepts of information theory [18]. These techniques are information-based methods. Chi-square is considered as one of the top methods utilized in many applications and works best with categircal data type [18]. CHI and MI are known as the two outstanding methods [19]; however, many studies have confrimed that working on combined-FS methods is better than trusting on a single method. The CHIMI: CHIMI is a proposed combined-FS method which is the combination of CHI and MI methods by computing the new feature scores.

CHI and MI methods have demonstated its merits of FS methods. Hence, this study come up with the concepts of calucating a new feature scores based on the score of CHI and MI algorithms. Since, it is broadly known that different FS methods compute scores of feature sets differently, therefore, produced different scales. Hence, before computing the new feature scores, the original scores of CHI and MI are first normalized. The normalization of MI and CHI scores can be done as in (10).

$$\overline{CHI} = \frac{CHI_i - CHI_{\min}}{CHI_{\max} - CHI_{\min}}, \; \overline{MI} = \frac{MI_i - MI_{\min}}{MI_{\max} - MI_{\min}} \tag{10}$$

Once, the normaliuzation of two score vectors are done, then it is passed to combined a new vector of feature score as formulated in (11).

$$CHIMI = \begin{pmatrix} \overline{CHI} \\ \overline{MI} \end{pmatrix} \tag{11}$$

The score vector indicated in (11) store the information of feature score of CHI and MI in form of vector. To get the absolute value or magnitude of the combined-FS method, Euclidean norm need to be computed. Hence, the new feature score of the combined-FS method is calculated using (12).

$$|CHIMI_i| = \sqrt{\left(\overline{CHI_i}\right)^2 + \left(\overline{MI_i}\right)^2} \tag{12}$$

To filter the redundant feature, Correlation Feature Selection (CFS) [13] is then introduced to filter the features of CHIMI. The CFS method evaluates the performance of feature subsets by evaluating the predictive ability of individual feature along with the degree of redundancy between input features.

This implies that the score of a feature in the CHIMI method containing the score vector generated by the CHI and MI algorithms with the different predictive ability of each feature. The new feature rearranges the order of importance of feature, feature with with bigger value of $|CHIMI_i|$ will be ranked higher. Unlike other previous methods of combining scores from different techniques such as AND and OR, our proposed approach gives a true metric on the space for score vector [20]. Some experimental results in earlier works reported a minor improvement or no improvement in classification performance when more than three feature selection methods were combined [21]. This method conducts a mathematical structure for examing the vector space of combined scores.

The normalization of CHI and MI scores is to introduce a new rank of input features based on the computed scores. This method may place the input features within their true rank and improves the higher possibility of certain significant features to being identified for selecting the dominant feature set.

E. Classification Algorithms

Several EDM techniques from many works of literature [11]-[16] were considered. The comparative study of prediction models on predicting student performance was conducted in [23]. The improvement version of the comparative study was conducted in [24]. The experimental results of both works indicated that k-nearest neighbor (KNN), two tree-based models: C5.0 and random forest (RF) are the optimal models. The developed EDM classifiers were proposed in earlier works [25][26][27]. The study of this work utilized the four prediction models as follows:

1) K-nearest neighbor (KNN): KNN KNN is known as an popular non-parametric EDM models utilized in many classicaiton problems. The KNN is confirmed to be a succesful classifier in our classifcation problem as proposed in the previous work [24]. Similarly to other classifiers, the KNN is noise-sensitive classifier. Its performance highly depends on the quality of the training data. The noise of data and mislabeled data, outliers, and overlaps regions between the data of different classes or targets lead to inaccurate classification [22].

2) Hybrid C5.0 and Hybrid RF: Hybrid C5.0 and Hybrid RF are the developed models that were studied in our earlier work [25]. The study gave the development and improvement version of [23] [24] for prediction academic performance. Four baseline models, naïve Bayes (NB), support vector machine (SVM), C5.0, and random forrest (RF) were utilized. The concept of principal component analysis (PCA) and k-fold cross validation (10-fold CV) were appied to baseline models. The Hybrid C5.0 (C5.0 + PCA + 10-fold CV) and Hybrid RF (RF + PCA + 10-fold CV) are the two standout models.

3) Improved Deep Belief Networks (IDBN): The IDBN is the optimization version of deep belief network (DBN) model. In our previous work, we gave a study of an optimization approach of DBN concerning (i) feature selection method, (ii) optimization of hyper-parameter, and (ii) regularization method [26]. The proposed IDBN successfully achieves the high prediction performance when applying with larger datasets.

IV. Experimental Results of Prediction Models

This section reported the performance evaluation of feature selection methods in selecting the dominant factors for predicting academic performance. We executed the proposed optimal classifiers using subsets that were obtained from each FS method. After applying the FS algorithms to the original datasets, each algorithm captures a subset of top *N* features. The FS algorithm selects the relevant factors to the target variables, then we rank the feature weight denoting the importance of features from sets selected by each FS algorithm decreasingly. We defined the dominant set as a set of input features containing top-n features that provide the highest prediction performance. The framework of the study is illustrated in Fig. 2.

The experiment was carried out with two phases. The first experiment was executed with the dataset ADS comprised of 1204 samples. The second experiment was with dataset GDS4 comprises 10000 samples. The second experiment was carried out with the context of a larger dataset to confirm the performance of IDBN and other proposed models. The experiment was made a minimal subset of five features to a fully selected set. To evaluate and compare the performance of prediction models, ACC and RMSE are measured. Recall that the higher value of ACC, the better model is. In contrast to ACC, the smaller value of RMSE, the better model is.

A. *Experimental Results with ADS Dataset*

This section illustrate the experiemental results of the proposed FS methods with the developed classfiers using ADS dataset. Table II describe the experimental results of the proposed method regarding with orginal dataset. Table III to Table VI illustrate the computational results of each classifier on subsets selected by IG, SU, CHI, MI, and the proposed CHIMI. The experiement aim to detect the dominant of each FS methods.

Table II illustrate the staitistcal results of the two metrics of KNN, Hybrid C5.0, Hybrid RF, and IDBN. The average of ACC and RMSE from serveral iteration run are recorded. The two developed tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE.

From Fig. 3 and Fig. 4, we can obtain the results of ACC and RMSE of KNN models using a selected set of each FS method, and it is summarized in Table III.

From Fig. 5 and Fig. 6, we can obtain the results of ACC and RMSE of the Hybrid C5.0 model using a selected set of each FS method, and it is summarized in Table IV.

From Fig. 7 and Fig. 8, we can obtain the results of ACC and RMSE of the Hybrid RF model using a selected set of each FS method, and it is summarized in Table V.

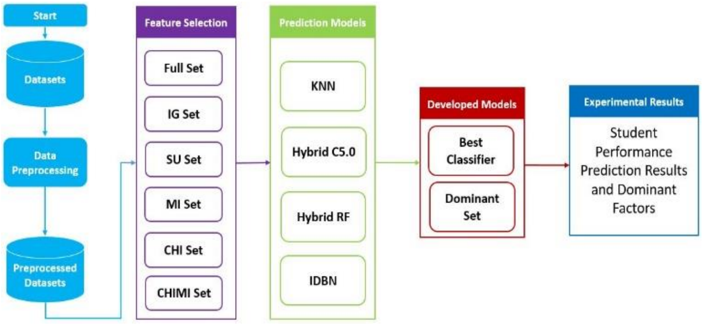

Fig. 2. Flowchart for the Experiment in this Study.

TABLE II. RESULTS OF PROPOSED MODELS ON ORIGINAL DATASETS

Proposed Models	**KNN**	**Hybrid C5.0**	**Hybrid RF**	**IDBN**
ACC (%)	94.95	99.25	99.72	83.14
Std. of ACC	0.801	0.601	0.357	0.640
RMSE	0.261	0.073	0.041	0.759
Std. of RMSE	0.041	0.045	0.029	0.031

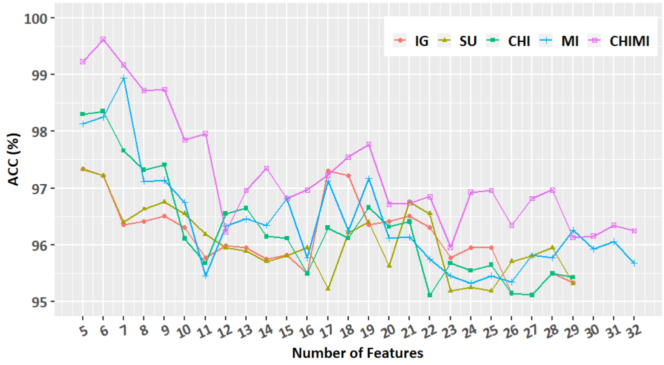

Fig. 3. ACC of KNN using ADS Dataset.

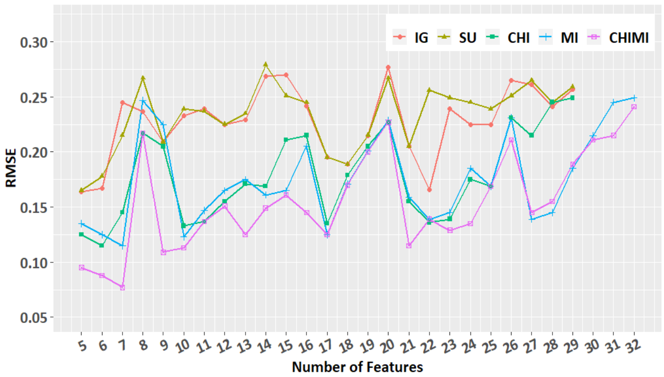

Fig. 4. RMSE of KNN using ADS Dataset.

TABLE III. THE PERFORMANCE EVALUATION ON KNN MODEL

Models	**Selected set**			**Dominant set**		
	N	***ACC***	***RMSE***	***N***	***ACC***	***RMSE***
IG	29	95.35	0.257	5	97.35	0.163
SU	29	95.32	0.259	5	97.33	0.164
CHI	29	95.43	0.249	6	98.35	0.115
MI	32	95.683	0.241	7	98.94	0.077
CHIMI	32	96.25	0.179	6	99.62	0.063

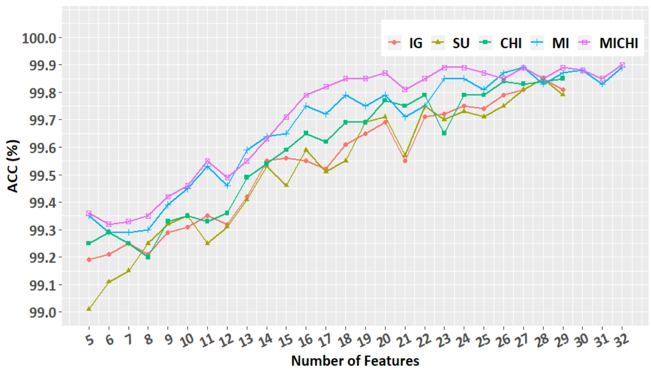

Fig. 5. ACC of Hybrid C5.0 using ADS Dataset.

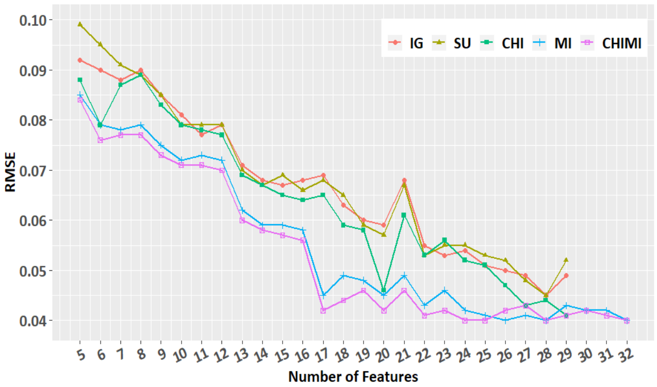

Fig. 6. RMSE of Hybrid C5.0 using ADS Dataset.

TABLE IV. THE PERFORMANCE EVALUATION OF HYBRID C5.0 MODEL

Models	**Selected set**			**Dominant set**		
	N	***ACC***	***RMSE***	***N***	***ACC***	***RMSE***
IG	29	99.81	0.049	28	99.85	0.045
SU	29	99.79	0.051	28	99.85	0.045
CHI	29	99.85	0.041	29	99.85	0.041
MI	32	99.89	0.035	32	99.89	0.035
CHIMI	32	99.90	0.033	32	99.90	0.033

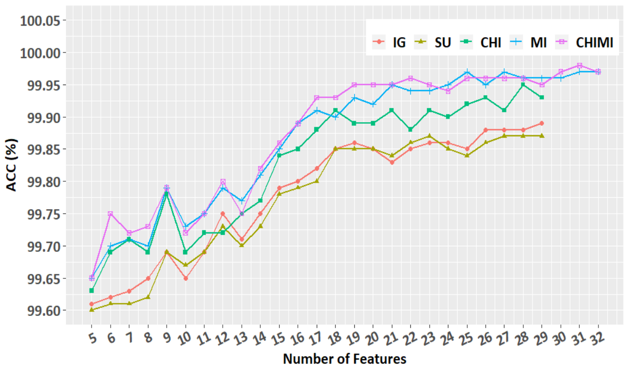

Fig. 7. ACC of Hybrid RF using ADS Dataset.

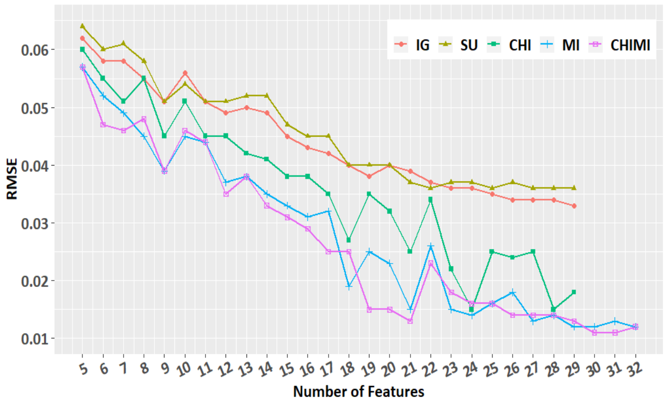

Fig. 8. RMSE of Hybrid RF using ADS Dataset.

TABLE V. THE PERFORMANCE EVALUATION OF THE HYBRID RF MODEL

Models	**Selected set**			**Dominant set**		
	N	***ACC***	***RMSE***	***N***	***ACC***	***RMSE***
IG	29	99.89	0.033	28	99.89	0.033
SU	29	99.87	0.036	28	99.87	0.036
CHI	29	99.95	0.015	29	99.95	0.015
MI	32	99.97	0.012	32	99.97	0.012
CHIMI	32	99.97	0.012	31	99.98	0.011

From Fig. 9 and Fig. 10, we can obtain the results of ACC and RMSE of the IDBN model using a selected set of each FS method, and it is summarized in Table VI.

The results presented in Table III demonstrate the performance of the KNN model on feature sets and dominant sets selected by IG, SU, CHI, MI, and CHIMI methods. The models work best with the low dimension of the most important features. The KNN model with the dominant set of the proposed CHIMI achieves the highest performance. The ACC and RMSE are improved to 99.62 and 0.063, respectively.

From Table IV, the performance of the Hybrid C5.0 is significantly improved when using the dominant set of the CHIMI. The ACC and RMSE of the Hybrid C5.0 are improved to 99.90 and 0.033, respectively.

The results of ACC and RMSE of the Hybrid RF are shown in Table V. The proposed CHIMI method outperforms the rest FS methods in achieving the highest ACC and lowest RMSE. The ACC and RMSE of the Hybrid RF are improved to 99.98 and 0.011, respectively.

Table VI demonstrates the performance of the developed IDBN classifier with the input feature sets selected by the five FS methods. The ACC and RMSE of the IDBN are improved to 87.32 and 0.514, respectively.

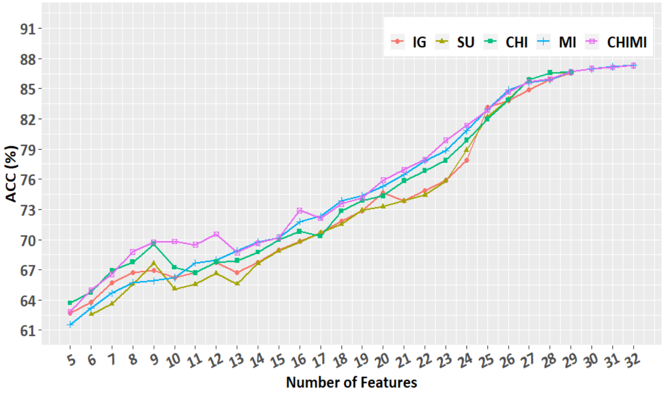

Fig. 9. ACC of IDBN using ADS Dataset.

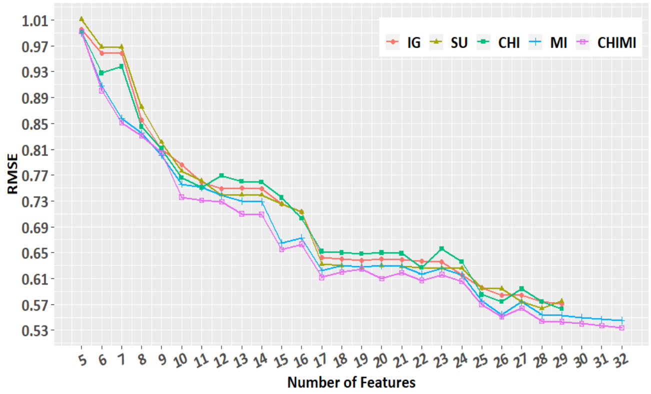

Fig. 10. RMSE of IDBN using ADS Dataset.

TABLE VI. THE PERFORMANCE EVALUATION OF THE IDBN MODEL

Models	Selected set			Dominant set		
	N	*ACC*	*RMSE*	*N*	*ACC*	*RMSE*
IG	29	86.55	0.571	28	86.55	0.571
SU	29	86.54	0.575	28	86.54	0.575
CHI	29	86.67	0.563	29	86.67	0.563
MI	32	87.11	0.545	32	87.11	0.545
CHIMI	32	87.32	0.514	32	87.32	0.514

B. Experimental Results with GDS4 Dataset

In this subsection, we do our experiment with an artificial dataset of a larger size of 10000 samples, dataset GDS4. We want to compare the performance of IDBN with other models using subsets of FS algorithms. The performance of the proposed models with the original dataset is shown in Table VII.

Table VII illustrate the staitistcal results of the two metrics of KNN, Hybrid C5.0, Hybrid RF, and IDBN. The average of ACC and RMSE from serveral iteration run using GS4 dataset are recorded. The two developed tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE, follow by the IDBN model.

From Fig. 11 and Fig. 12, we can obtain the results of ACC and RMSE of KNN models using a selected set of each FS method, and it is summarized in Table VIII.

From Fig. 13 and Fig. 14, we can obtain the results of ACC and RMSE of the Hybrid C5.0 model using a selected set of each FS method, and it is summarized in Table IX.

TABLE VII. RESULTS OF PROPOSED MODELS ON ORIGINAL DATASETS

Proposed Models	KNN	Hybrid C5.0	Hybrid RF	IDBN
ACC (%)	95.12	98.55	98.88	97.01
Std. of ACC	0.942	0.578	0.312	0.666
RMSE	0.193	0.163	0.161	0.195
Std. of RMSE	0.016	0.049	0.028	0.015

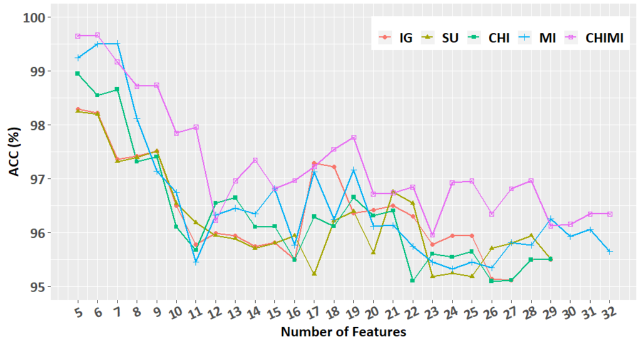

Fig. 11. ACC of KNN using GDS4 Dataset.

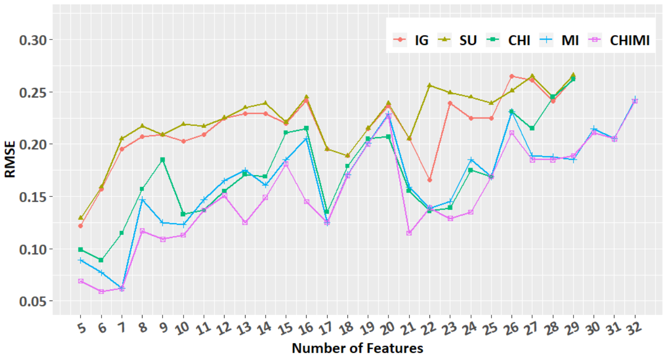

Fig. 12. RMSE of KNN using GDS4 Dataset.

TABLE VIII. THE PERFORMANCE EVALUATION OF THE KNN MODEL

Models	**Selected set**			**Dominant set**		
	N	***ACC***	***RMSE***	***N***	***ACC***	***RMSE***
IG	29	95.50	0.263	5	98.30	0.122
SU	29	95.52	0.266	5	97.25	0.129
CHI	29	95.51	0.262	6	98.95	0.089
MI	32	95.65	0.243	7	99.52	0.062
CHIMI	32	96.35	0.175	6	99.67	0.059

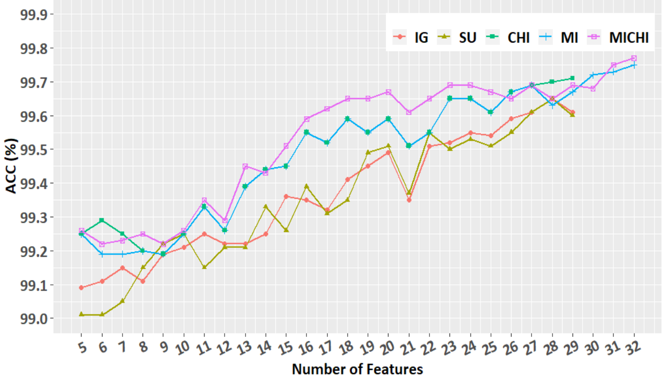

Fig. 13. ACC of Hybrid C5.0 using GDS4 Dataset.

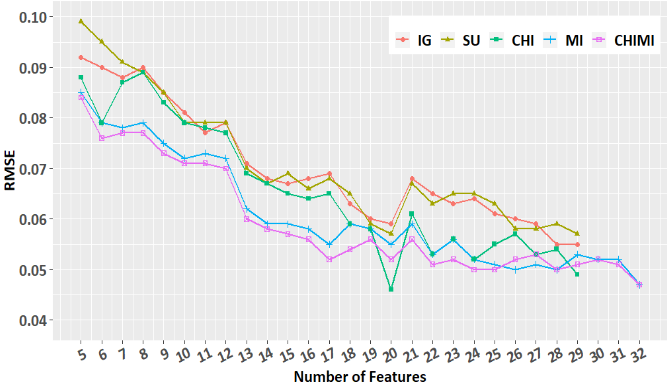

Fig. 14. RMSE of Hybrid C5.0 using GDS4 Dataset.

TABLE IX. THE PERFORMANCE EVALUATION OF HYBRID C5.0 MODEL

Models	**Selected set**			**Dominant set**		
	N	***ACC***	***RMSE***	***N***	***ACC***	***RMSE***
IG	29	99.61	0.059	28	99.65	0.055
SU	29	99.60	0.058	28	99.63	0.057
CHI	29	99.71	0.047	29	99.71	0.051
MI	32	99.75	0.045	32	99.75	0.047
CHIMI	32	99.75	0.045	32	99.75	0.045

From Fig. 15 and Fig. 16, we can obtain the results of ACC and RMSE of the Hybrid RF model using a selected set of each FS method, and it is summarized in Table X.

From Fig. 17 and Fig. 18, we can obtain the results of ACC and RMSE of the IDBN model using a selected set of each FS method, and it is summarized in Table XI.

The results presented in Table VIII demonstrate the performance of the KNN models on feature sets and dominant sets selected by IG, SU, CHI, MI, and CHIMI methods on the GDS4 dataset. The KNN model with the dominant set of the proposed CHIMI achieves the highest performance. The ACC and RMSE of the KNN are improved to 99.67 and 0.059.

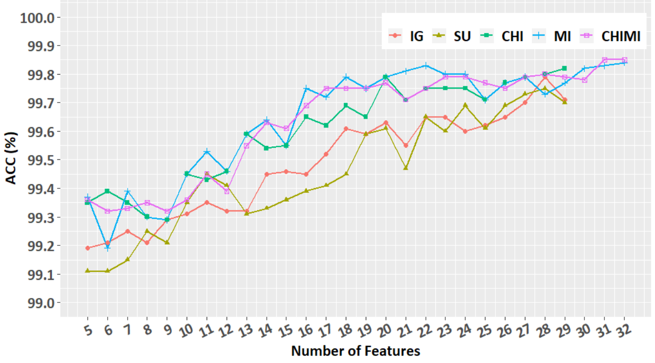

Fig. 15. ACC of Hybrid RF using GDS4 Dataset.

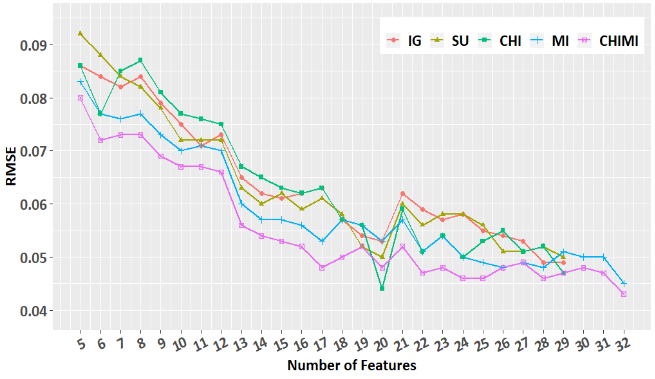

Fig. 16. RMSE of Hybrid RF using GDS4 Dataset.

TABLE X. THE PERFORMANCE EVALUATION OF THE HYBRID RF MODEL

Models	Selected set			Dominant set		
	N	*ACC*	*RMSE*	*N*	*ACC*	*RMSE*
IG	29	99.73	0.051	28	99.79	0.049
SU	29	99.73	0.052	28	99.75	0.050
CHI	29	99.82	0.047	29	99.82	0.047
MI	32	99.84	0.045	32	99.84	0.045
CHIMI	32	99.85	0.043	32	99.85	0.043

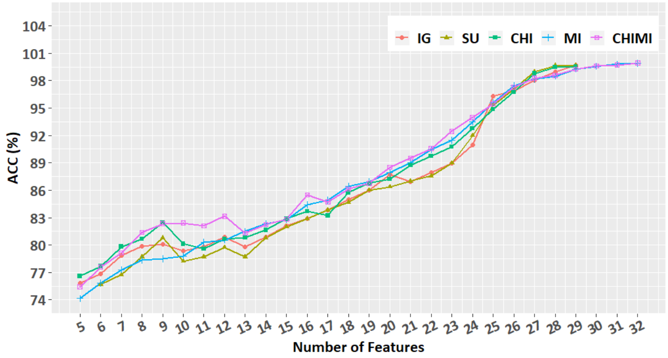

Fig. 17. ACC of IDBN using GDS4 dataset

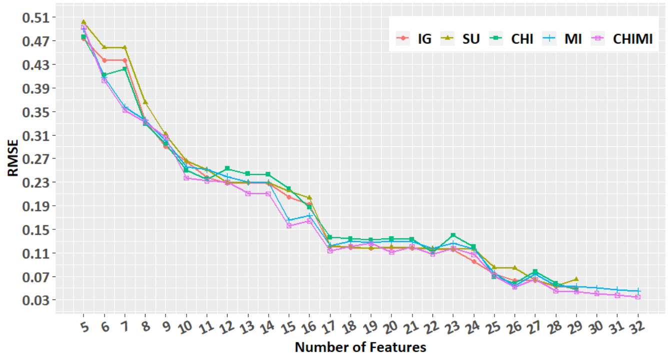

Fig. 18. RMSE of IDBN using GDS4 dataset

From Table IX, the performance of the Hybrid C5.0 is significantly improved when using the dominant set of the CHIMI. The ACC and RMSE of the Hybrid C5.0 are improved to 99.75 and 0.045, respectively.

The results of ACC and RMSE of the Hybrid RF are shown in Table X. The proposed CHIMI method outperforms the rest FS methods in achieving the highest ACC and lowest RMSE. The ACC and RMSE of the Hybrid RF are improved to 99.85 and 0.043, respectively.

TABLE XI. THE PERFORMANCE EVALUATION OF THE IDBN MODEL

Models	Selected set			Dominant set		
	N	*ACC*	*RMSE*	*N*	*ACC*	*RMSE*
IG	29	99.65	0.052	28	99.67	0.050
SU	29	99.65	0.052	28	99.67	0.050
CHI	29	99.77	0.047	29	99.77	0.047
MI	32	99.81	0.045	32	99.81	0.045
CHIMI	32	99.82	0.044	32	99.82	0.044

Table XI demonstrates the performance of the developed IDBN classifier with the input feature sets selected by the five FS methods. The ACC and RMSE of the IDBN are improved to 99.82 and 0.044, respectively.

C. Summary and Discussion

This study aims to boost up the performance of the proposed classifiers to reach the most classification results. Hence, the optimal models are then combined with dominant sets, which is belive to significantly improve the performance of prediction models and selected the highly influencing factors in academic performance.

From Table II to Table XI, we can summarize the performance of KNN, Hybrid C50, Hybrid RF, and IDBN on dominant sets of IG, SU, CHI, MI, and CHIMI methods. Fig. 19 and Fig. 20 summarize the value of ACC and RMSE of each classifier on each FS method on data ADS and GDS4, respectively.

Fig. 19 represents the values of ACC and RMSE of the four classifiers with the five FS methods using the ADS dataset. Concerning prediction models, Hybrid C5.0 and Hybrid RF are comparatively better than the other two models. Regarding the FS methods, the performance of IG and SU methods are not statistically different. The performance of CHI and MI algorithms standout the performance of IG and SU. The figure reports that the proposed CHIMI successfully improves the performance of the four prediction models and standout the performance of the four FS methods.

Fig. 20 graphically demonstrate the performance of the four classifiers with the five FS methods using the GDS4 dataset. In the context of a larger dataset, the performance of IDBN is significantly improved. The performance of IDBN and Hybrid are not statistically different and the two models standout the performance of Hybrid C5.0 and KNN. The proposed CHIMI roles as the best FS method in selecting the dominant factors for improving the performance of the prediction models.

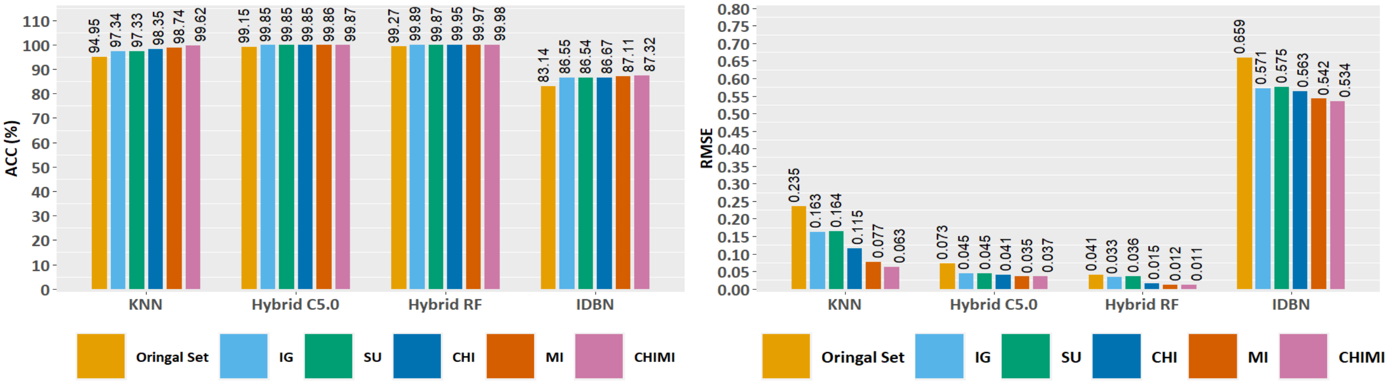

Fig. 19. ACC and RMSE Comparison using ADS Dataset.

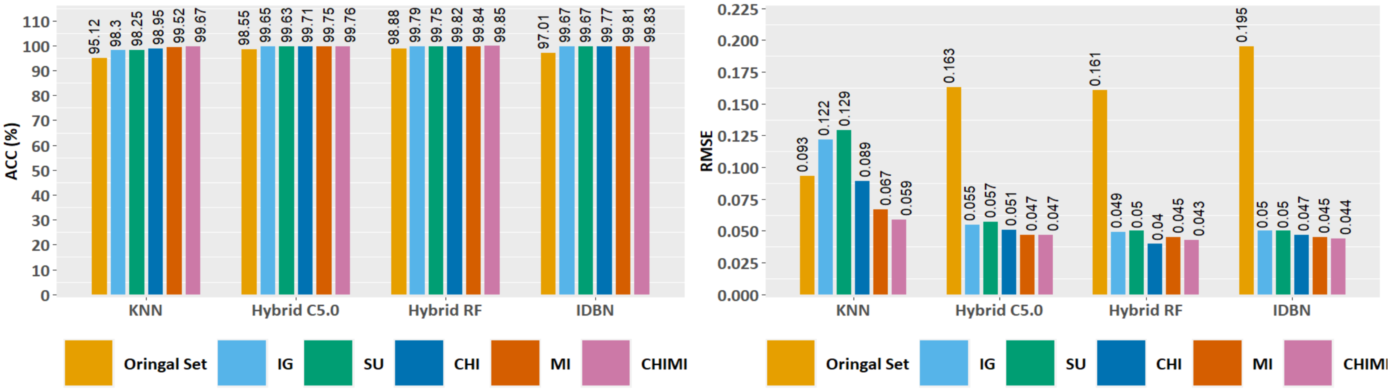

Fig. 20. ACC and RMSE Comparison using GDS4 Dataset.

V. Conclusion

Education is a crucial factor in the development of any country. Predicting student performance and mining their learning behaviors are challenging tasks in an educational environment. This paper presents a study of the analysis of dominant factors using feature selection methods and propose a novel feature selection algorithm for improving prediction performance to get the most successful classification results. The proposed CHIMI of the FS method significantly improves the performance of the proposed prediction models. The dominant set of the CHIMI method enhances the accuracy of prediction models from 1 to 5% increase and the RMSE from 0.06 to 0.2 decrease. The performance of the proposed prediction models reaches the superior classification results that can be effectively used to predict student performance. Once performance levels can be effectively predicted, hence, the at-risk group of students with poor-performing is identified, then they can be timely given intervention and additional assistance.

Acknowledgment

The authors would like to thank JSPS Grant-in-Aid for Scientific Research(S) for supporting this work (Grant Number JP17H06115).

References

[1] A. A. Saa, "Educational data mining & student's performance prediction," International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016.

[2] S. Slater, S. Joksimovic, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining," Journal of Educational and Behavioral Statistics, vol. 10, Issue 3, pp. 85-106, 2017.

[3] C. Romero and S. Ventura, "Data mining in education," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, issue 1, pp. 12-27, 2013.

[4] P. Thakar, A. Mehta, and Manisha, "Performance analysis and prediction in educational data mining," International Journal of Computer Application, vol. 110, no. 15, pp. 60-68, 2015.

[5] A. Pena-Ayala, "Educational data mining: Survey and a data mining-based analysis of recent works," Expert Systems with Application, vol. 41, pp. 1432-1462, 2014.

[6] E. Sosmangegovic, M. Suljic, and H. Gic, "Determining Dominant Factor for Student Performance Prediction by Using Data Mining Classification", Tranzicija, Vol. 6, 2014, pp.147-158.

[7] H. M. Hard and M.A. Moustafa, "Selecting Optimal Subset of Features for Predicting Student Performance Model", International Journal of Computer Science, Vol. 9, Issue 5, No. 1, pp. 253-262, 2012.

[8] Jindal P., and Kumar D.., "A Review on Dimensionality Reduction Techniques", International Journal of Computer Application, Vol. 173, No. 2, pp.42-46, 2017.

[9] L. Ma et al., "Evaluation of Feature Selection Methods for Object-Based Land Cover Mapping of Unmanned Aerial Vehicle Imagery Using Random Forest and Support Vector Machine Classifiers", International Journal of Geo-Information, Vol. 6, No. 51, 21 pages, 2017.

[10] P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risk in Hepatitis Disease", International Journal of Machine Learning and Computing, Vol. 5, No. 4, pp. 258-263, 2015.

[11] P. J. M. Estrera, et al, "Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School", International Journal of Engineering & Technologies, Vol. 3, No. 5, pp. 147-153, 2017.

[12] R amaswami M., Bhaskaran R., “A Study of Feature Selection Techniques in Educational Data Mining”, Journal of Computing, Vol. 1, Issue 1, pp. 7-11, 2009.

[13] J. D. Febro, “Untilizing feature selection in identifying predicting factors of student retention,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, 2019.

[14] M. Zaffar and K.S. Savita, “A study of feature selection algorithms for predicting students’ academic performance,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, 2018.

[15] A. Alhassan, B. Zaffar, and A. Mueen, “Predict students’ academic performance based on their assessment grades and online activity data,” International Journal of Advanced and Computer Science and Applications, vol. 11, no. 4, 2020.

[16] Mythili M.S., Mohamed S.A.R., An Analysis of Student Performance Using Classification Algorithms, IOSR Journal of Computer Engineering, Vol. 6, Issue 1, pp. 63-69, January 2014.

[17] Phyu T.Z. and Oo N.N, “Performance Comparison of Feature Selection Methods”, MATEC Web of Conference 42, 2016.

[18] A. Bommert et al., “Benchmark for filter methods for feature selection in high-dimensional classification data”, Journal of Computational Statistics and Data Analysis, vol. 143, 2020.

[19] Mazumder D.H, and Veilumuthu R., “An enhanced feature selection filter for classification of microarray cancer data”, Wiley ETRI Journal, pp.358-370, 2018.

[20] C. F. Tsai and Y. C. Hsiao, “Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches,” Decis Support System, vol. 50, no. 1, 258-269, 2010.

[21] A. Thubaity, N. Abanumay, S. Al-Jerayyed, A. Alrukban, Z. Mannaa, “The effect of combining different feature selection methods on Arabic text classification,” IEEE: The 14th ACIS International Conference Software Engineering, Artificial Intelligent, Networking and Parallel/distributed Computing (SNPD), 211-216.

[22] S. Oulianroglou and G. Evangelidis, “Dealing with noisy data in the context of k-NN Classification,”. Proceeding of the 7th Balkan Conference on Information Conference, Article ID. 28, pp. 1-4, 2015.

[23] P. Sokkhey and T. Okazaki, “Comparative study of prediction models for high school student performance in mathematics,” Journal of IEIE Transactions on Smart Processing and Computing, vol. 8, no. 5, pp. 394-404, 2019.

[24] P. Sokkhey and T. Okazaki, “Multi-models of educational data mining for predicting student performance: A case study of high schools in Cambodia,” vol. 9, no. 3, pp. 217-229, 2020.

[25] P. Sokkhey and T. Okazaki, “Hybrid machine learning algorithms for prediction academic performance,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, pp. 32–41, 2020.

[26] P. Sokkhey and T. Okazaki, “Development and optimization of deep belief networks for academic prediction with larger datasets,” Journal of IEIE Transactions on Smart Processing and Computing, (Accepted 20-April-2020).

[27] P. Sokkhey and T. Okazaki, “Developing Web-based Support System for Predicitng Poor-performing students Using Educational Data Mining Techniques,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, pp. 23–32, 2020.