An Intelligent Approach with Fused PCA and RAE Feature Selection in Education Data Mining

Dilshad Kaur
Research Scholar,
Guru Gobind Singh College Of Modern Technology,Kharar(PB),India
dilshadladdi@yahoo.com

***Abstract*— Data mining technique is widely used in the field of education for evaluating the student's performance. Student data mining or education data mining is done for evaluating the performance of students of an institute. EDM plays most important role for high quality universities and institutes to set their criteria or level among other universities. The record of the students of a university sets the academic achievement level of its own. Various techniques are used for Education data mining such as classification, regression, association and clustering etc. In this thesis a new technique is proposed for student's data mining.**

In this work, the hybrid data mining technique i.e. HPR-F-MLP is done for grading the student's performance on the basis of the various factors such as name, age, sex etc of the students. On the basis of the evaluated grades, the status of the student is evaluated that whether the student falls under the category of failed students or passed students. For this purpose the Principal Component and relief Attribute mechanism id applied for feature extraction and after extracting the features, the discrete wavelet fusion is applied to fuse them. The classification is performed by using the Multi Layered Perceptron classification technique.

The simulation is done in MATLAB by using various classification algorithms and feature extraction techniques. On the basis of the results it is observed that the HPR-F-MLP outperforms the traditional classification techniques in the terms of precision, recall and F-measure.

Keywords—Gateway nodes.

I. INTRODUCTION

Education data mining is done for evaluating the performance of students of an institute. EDM plays most important role for high quality universities and institutes to set their criteria or level among other universities. The record of the students of a university sets the academic achievement level of its own. Education data mining is also known as student data mining. The term education data mining or student data mining has various meanings according to previous work or literature survey. Student data mining can be performed by measuring the performance of the students and in order to measure the performance of a student there are various ways such as to calculate the learning assessment of the student and calculate the co-curriculum. Most of the previous work focuses on the performance evaluation of graduate students and their success rate. For example in Malaysia, most of the universities and colleges use final grades of the students for the process of student data mining. These final grades of the student is the combination of his performance in various activities like marks obtained in exams, assessment scores, scoring in various co-curriculum activities and Performa of the syllabi or course. The estimation of student performance is an important task to be performed by the universities. The idea behind the evaluation of the student's performance is to keep the record of a student's performance and efficiency of learning process. On the basis of the data collected by analyzing the student's performance various strategies and plans or policies can be created for student's welfare.

For evaluating the performance of the students, universities have to maintain their records in various activities. Then these records are used to analyze the student's performance. There are various techniques and methods which are used to analyze the recorded data. Data mining is one of the techniques. Data mining process in field of education is known as education data mining. Data mining technique is widely used in the field of education for evaluating the student's performance.

EDM or student data mining is a process which is used to haul out the useful and meaningful information from large datasets or database of the student's information. Then the extracted data or information is used for predicting the performance of the students. EDM is an application which is an aid to students' data mining process. EDM is an application of data mining.

II. PROBLEM FORMULATION

Previously, a lot of work is done to predict the performance of student using different feature selection techniques. In recent studies, researchers use different feature selection techniques and the combination of classifiers to produce efficient prediction models. A research is required to identify the performance analysis in terms of prediction accuracy in combination of different feature selection algorithms with differently classifiers. Moreover, advanced classifiers are required to be used for classifications. The techniques employed in the existing work provide less prediction accuracy, efficiency and effectiveness. Considering these issues in the existing work, a novel approach is required to

identify the prediction accuracy of different available feature selection algorithm in the context of classifiers being used on educational data.

III. PROPOSED WORK

After reviewing the issues in the existing work, a novel approach is proposed. Initially, feature selection will be performed using two techniques such as Principal Component Analysis and Relief Attribute Eval. The collaboration of these techniques can perform effectively in terms of prediction accuracy. Furthermore, the fusion of extracted features in the proposed work will be evaluated using Discrete Wavelet approach which is an effective approach in capturing both frequency and location information (location in time). Finally, MLP classifier is employed to classify the dataset. The review performed has concluded that shown that MLP classifier performed slightly better than other classifiers on student data set.

1. The proposed model has initial the collection of the dataset that is taken from the UCI machine learning website the link of the dataset is as follow
   https://archive.ics.uci.edu/ml/datasets/student+performance
2. The next step for the work is to normalization of dataset as the dataset available in the link is in raw format from which the useful information is to extract.
3. After the normalization of the dataset the dataset is divided into two groups that are
   a. Training dataset
   b. Testing dataset
4. Once the dataset is split into two groups next step is to extract the multiple feature for the dataset selected
5. In proposed model two feature extraction techniques are used those are
   a. Principle component analysis (PCA)
   b. Relief Attribute (RA)
6. As there are two set of feature extraction technique the number of features are going to be increased, therefore in the proposed methodology the features are fused to further process.
7. Wavelet based feature fusion is the next step for the proposed methodology.
8. After the fusion of the dataset a defined set of information is available those will be take as input to the MLP classifier.
9. The final grades are taken as the target for the input information of the MLP classifier's training.
10. Once the training is done the next step is to test the testing dataset.
11. For this step the step number 4 to 7 are repeated for the feature extraction then the features set is given to the MLP trained classifier
12. MLP classifier will predict the output in form of the grade level.
13. Next step is to evaluate the final results, the proposed model is validate using the performance matrix of parameters
    a. Precision
    b. Recall
    c. F-measure

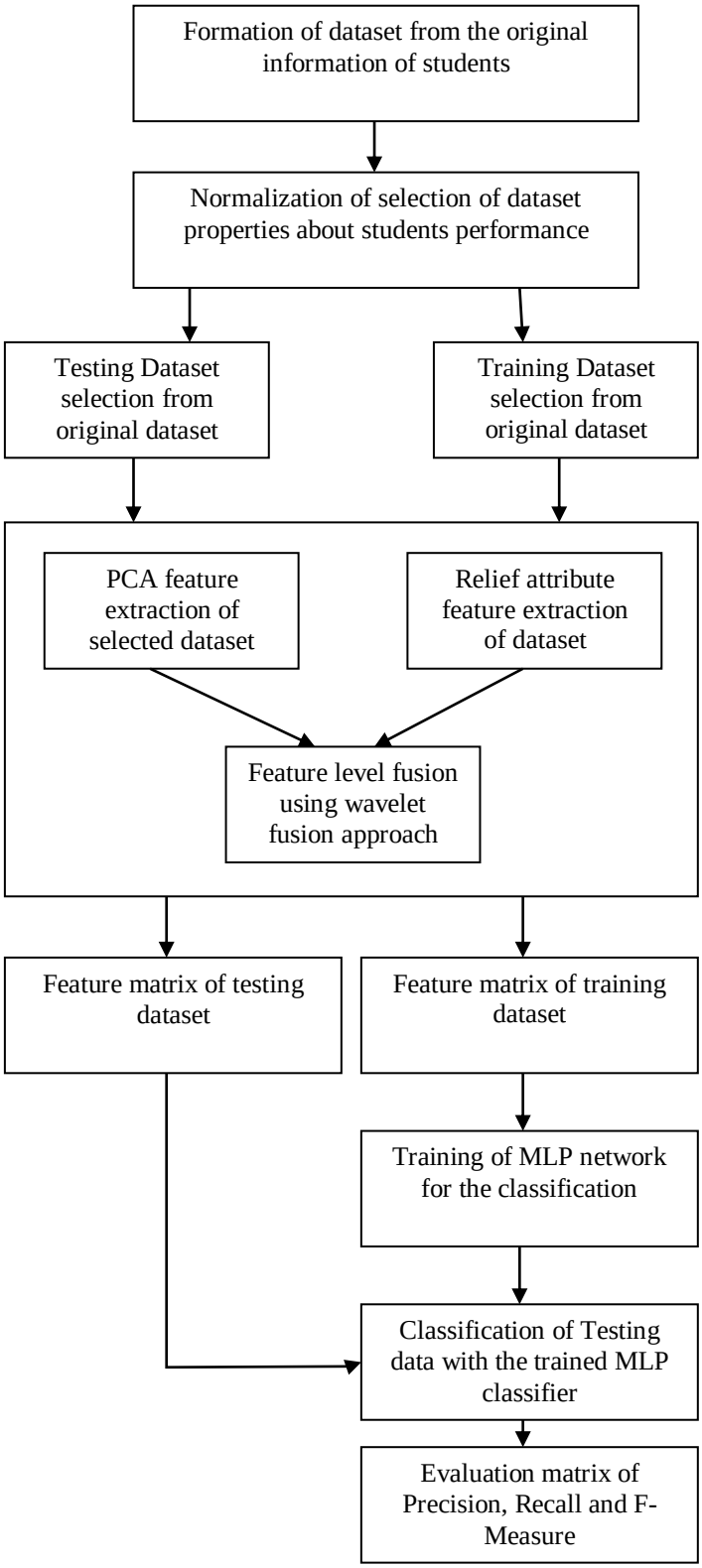

Figure 1 Flowchart of HPR-F-MLP

IV. RESULTS

This section serves the results that are attained afterwards implementing the proposed work. The implementation is done in MATLAB. There are some graphs in this section that confirm the adequacy of proposed technique. The performance of the proposed work is evaluated in the terms of MSE, Precision, Recall and F-measure. Precision is known as positive predictive value. It is measured as follows:

$$Precision = \frac{TP}{(TP + FP)} \ldots\ldots (1)$$

Where TP is true positive, FP is false positive.

MSE is mean square error that is a performance matrix which is used to measure the mean square error in the observed output. The MSE should be always low to assure the quality of the observed results or output. The following is the formulation of MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(AC - PR)^2 \ldots\ldots (2)$$

Where AC refers to the Actual Results and PR refers to the Predicted Results. The N is the total number of samples.

Recall defines that how many relevant items are elected. The formulation given below is used for evaluating the recall for proposed work:

$$Recall = \frac{TP}{(TP + FN)} \ldots\ldots (3)$$

F-Measure is a performance matrix that is used to evaluate the harmonic mean of precision and recall. The formulation is as follows:

$$F - Measure = 2 * \frac{Precision * Recall}{(Precision + Recall)} \ldots\ldots (4)$$

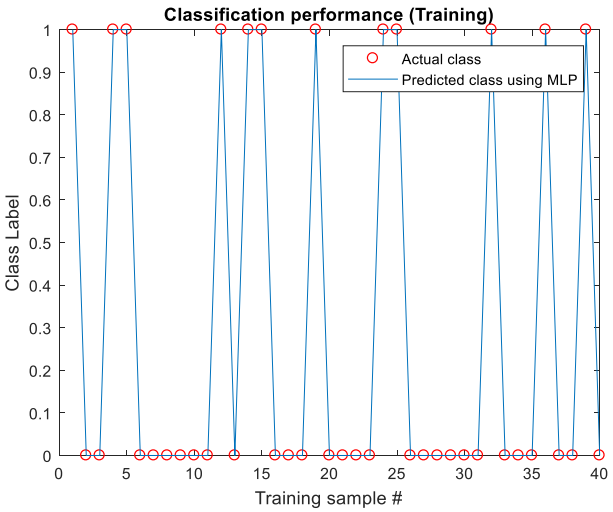

Figure 2 Classification Performance (training)

The graph in figure 2 and 3 shows the training sample and testing samples of classification performance. The x axis in both graphs calibrates the data in the form of training samples and the y axis in the graph defines the data for class labels. The class labels vary from 0 to 1 and the training samples varies from 0 to 40 for figure 2 and 0 to 50 for figure 3.

The graph in figure 3 is classification performance of testing of the proposed work. The blue line in the graphs pretends the predicted class using MLP and the red circle shows the related actual class.

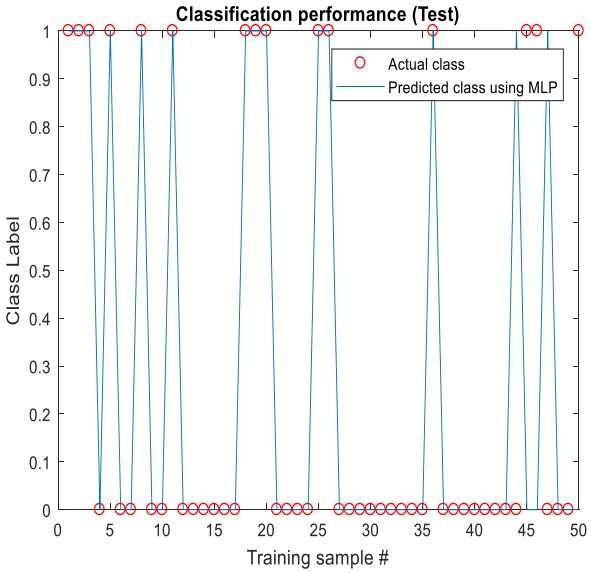

Figure 3 Classification Performance (testing)

The graph in figure 4 is the comparison analysis of the precision, recall and F-Measure for Cfs Subset Eval. The x axis in the graph shows the classification algorithms and the y axis in the graph shows the values for precision, recall and F-Measure. The bar in Blue refers to the values for the precision; the bar in yellow defines the results for F-measure and the bar in green denoted the performance of the recall for respective algorithms. The facts and figures that are gathered from the comparison graph are presented in the table 1.

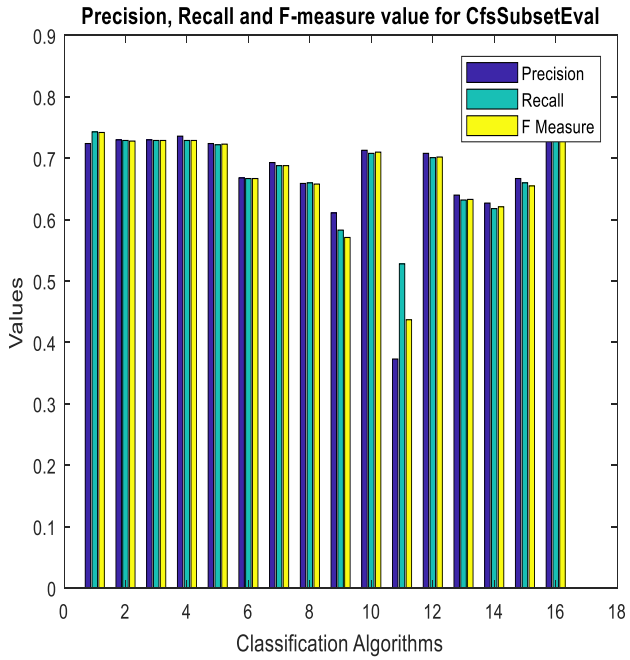

Figure 4 Analysis of Cfs SubsetEval

Table 1 Precision, Recall and F-Measure for CfssubsetEval

S.No	Classification Algorithm	Precision	Recall	F-Measure
1.	'BN'	0.724	0.743	0.742
2.	'NB'	0.73	0.729	0.728
3.	'NBU'	0.73	0.729	0.729
4.	'MLP'	0.736	0.729	0.729
5.	'SL'	0.724	0.722	0.723
6.	'SMO'	0.668	0.667	0.667
7.	'DT'	0.693	0.688	0.688
8.	'Jrip'	0.659	0.66	0.658
9.	'OneR'	0.611	0.583	0.571
10.	'PART'	0.713	0.708	0.71
11.	'DS'	0.373	0.528	0.437
12.	'J48'	0.708	0.701	0.702
13.	'RF'	0.64	0.632	0.633
14.	'RT'	0.627	0.618	0.621
15.	'RepT'	0.667	0.66	0.655
16.	HPR-F-MLP'(proposed Work)	0.857143	0.8	0.827586207

Similarly, the graph in figure 5 presents the comparison analysis of the respective performance measures for Chi Squared Attributes Eval. The Chi squared Attributes in another feature selection technique that evaluates the attribute by measuring the value for chi squared statistic related to the class set. The table 2 is organized to support the facts that are pretended by the graph of figure 5.

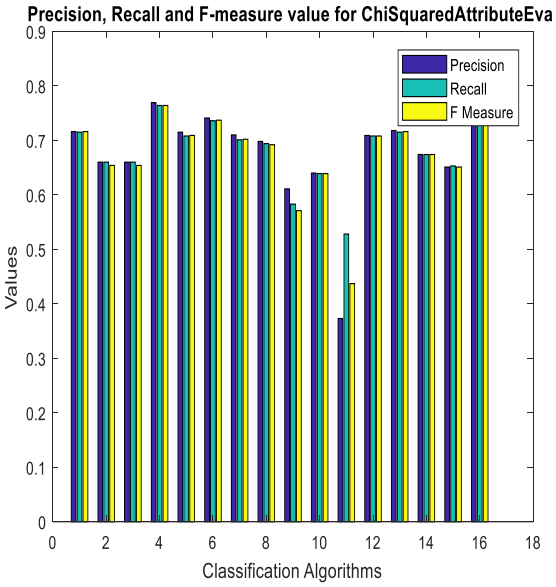

Figure 5 Analysis of Chi Squared Attributed Eval

Table 2 Precision, Recall and F-Measure for Chi Squared Attributed Eval

S.No	Classification Algorithm	Precision	Recall	F-Measure
1.	'BN'	0.716	0.715	0.716
2.	'NB'	0.66	0.66	0.654
3.	'NBU'	0.66	0.66	0.654
4.	'MLP'	0.769	0.764	0.764
5.	'SL'	0.715	0.708	0.709
6.	'SMO'	0.741	0.736	0.737
7.	'DT'	0.71	0.701	0.702
8.	'Jrip'	0.698	0.694	0.692
9.	'OneR'	0.611	0.583	0.571
10.	'PART'	0.64	0.639	0.639
11.	'DS'	0.373	0.528	0.437
12.	'J48'	0.709	0.708	0.708
13.	'RF'	0.718	0.715	0.716
14.	'RT'	0.674	0.674	0.674
15.	'RepT'	0.651	0.653	0.651
16.	HPR-F-MLP'(proposed Work)	0.857143	0.8	0.827586207

The graph in figure 6 delineates the comparison of the precision, recall and F-Measure for feature extraction technique i.e. Filtered Attribute Eval. The graph explains that the precision, recall and F-measure of proposed technique are higher than the rest of the mechanisms.

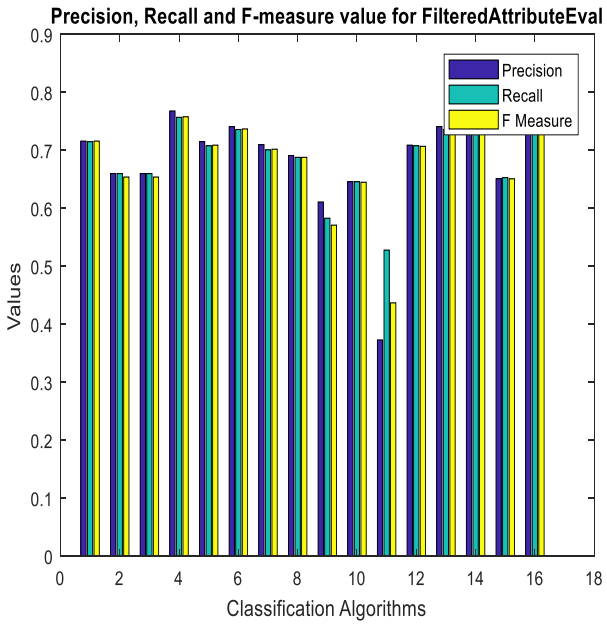

Figure 6 Analysis of Filtered Attribute Eval

Thus, it can be said that the proposed work is quite efficient in producing the qualitative results of decision. The values for respective parameters are shown by the table 3. The table also assures the results of the graph and proves that the proposed work outperforms the rest of the classification techniques in terms of precision, recall and F-measure.

Table 3 Precision, Recall and F-Measure for Filtered Attribute Eval

S.No	Classification Algorithm	Precision	Recall	F-Measure
1.	'BN'	0.716	0.715	0.716
2.	'NB'	0.66	0.66	0.654
3.	'NBU'	0.66	0.66	0.654
4.	'MLP'	0.768	0.757	0.758
5.	'SL'	0.715	0.708	0.709
6.	'SMO'	0.741	0.736	0.737
7.	'DT'	0.71	0.701	0.702
8.	'Jrip'	0.691	0.688	0.688
9.	'OneR'	0.611	0.583	0.571
10.	'PART'	0.646	0.646	0.645
11.	'DS'	0.373	0.528	0.437
12.	'J48'	0.709	0.708	0.707
13.	'RF'	0.741	0.736	0.737
14.	'RT'	0.738	0.729	0.73
15.	'RepT'	0.651	0.653	0.651
16.	HPR-F-MLP' (Proposed Work)	0.857143	0.8	0.827586207

The graph in figure 7 shows the comparison analysis of performance parameters by using various classification techniques and proposed work in the terms of Gain Ratio Attribute Eval feature extraction mechanism.

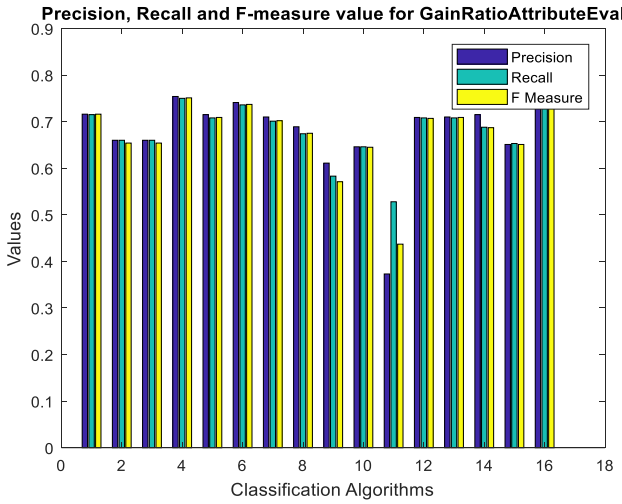

Figure 7 Analysis of Gain Ratio Attribute Eval

The gain Ratio is defined as non-symmetrical feature extraction technique which in specifically considered to compensate the biases of information gain. The table 4 defines the values for precision, recall and f-measure.

Table 4 Precision, Recall and F-Measure for Filtered Attribute Eval

S.No	Classification Algorithm	Precision	Recall	F-Measure
1.	'BN'	0.716	0.715	0.716
2.	'NB'	0.66	0.66	0.654
3.	'NBU'	0.66	0.66	0.654
4.	'MLP'	0.754	0.75	0.751
5.	'SL'	0.715	0.708	0.709
6.	'SMO'	0.741	0.736	0.737
7.	'DT'	0.71	0.701	0.702
8.	'Jrip'	0.689	0.674	0.675
9.	'OneR'	0.611	0.583	0.571
10.	'PART'	0.646	0.646	0.645
11.	'DS'	0.373	0.528	0.437
12.	'J48'	0.709	0.708	0.707
13.	'RF'	0.71	0.708	0.709
14.	'RT'	0.715	0.688	0.687
15.	'RepT'	0.651	0.653	0.651
16.	HPR-F-MLP'	0.857143	0.8	0.82758621

Similarly the graph in figure 8 and 9 depict the comparison analysis of the classification algorithms on the basis of the principal component and relief attributes eval feature extraction technique respectively.

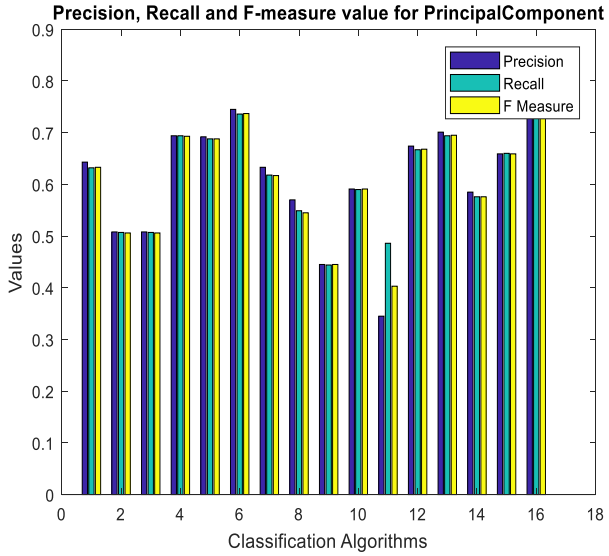

Figure 8 Analysis of Principal Component

On the basis of the observation it can be defined that the precision, recall and f-measure of the proposed work is higher in both of the cases i.e. Principal Component and Relief Attribute eval. The respective observations from both of the graphs are shown in table 5 and 6 respectively.

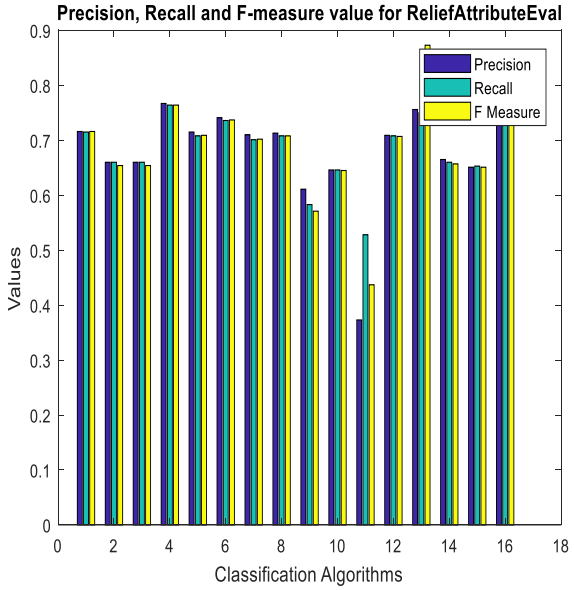

Figure 9 Analysis of Relief Attribute Eval

Table 5 Precision, Recall and F-Measure for Principal Component

S No	Classification Algorithm	Precision	Recall	F-Measure
1.	'BN'	0.643	0.632	0.633
2.	'NB'	0.508	0.507	0.506
3.	'NBU'	0.508	0.507	0.506
4.	'MLP'	0.694	0.694	0.693
5.	'SL'	0.692	0.688	0.688
6.	'SMO'	0.745	0.736	0.737
7.	'DT'	0.633	0.618	0.617
8.	'Jrip'	0.57	0.549	0.545
9.	'OneR'	0.445	0.444	0.445
10.	'PART'	0.591	0.59	0.591
11.	'DS'	0.345	0.486	0.403
12.	'J48'	0.674	0.667	0.668
13.	'RF'	0.701	0.694	0.695
14.	'RT'	0.585	0.576	0.576
15.	'RepT'	0.659	0.66	0.659
16.	HPR-F-MLP' (proposed work)	0.857143	0.8	0.827586207

Table 6 Precision, Recall and F-Measure for Relief Attribute Eval

S No	Classification Algorithm	Precision	Recall	F-Measure
1.	'BN'	0.716	0.715	0.716
2.	'NB'	0.66	0.66	0.654
3.	'NBU'	0.66	0.66	0.654
4.	'MLP'	0.767	0.764	0.764
5.	'SL'	0.715	0.708	0.709
6.	'SMO'	0.741	0.736	0.737
7.	'DT'	0.71	0.701	0.702
8.	'Jrip'	0.713	0.708	0.708
9.	'OneR'	0.611	0.583	0.571
10.	'PART'	0.646	0.646	0.645
11.	'DS'	0.373	0.528	0.437
12.	'J48'	0.709	0.708	0.707
13.	'RF'	0.756	0.75	0.873
14.	'RT'	0.665	0.66	0.657
15.	'RepT'	0.651	0.653	0.651
16.	HPR-F-MLP'(proposed work)	0.857143	0.8	0.827586207

V. CONCLUSION

Data mining is an automatic process which is used to remove the meaningful information from the data storage and further use this removed information for various purposes. The extraction of meaningful data can be performed by matching pattern. As the size of the data is increased various methods have been proposed for the data mining. The new hybrid approach i.e. HPR-F-MLP is developed by collaborating Principal component and Relief Attribute mechanisms for feature extraction. Along with this the Discrete Wavelet Fusion is applied for fusing the features that are extracted by principal component and relief attribute mechanism. After this, the MLP (Multi Layered Perceptron) is applied for the purpose of classification. The simulation is done in MATLAB by using various classification algorithms with respect to the various feature extraction techniques. From the results obtained it is concluded that the HPR-F-MLP method is better and efficient than the traditional systems. The designed system helps in decision making by implementing the process of extracting useful data from the data set.

As it is concluded from the results that HPR-KF-MLP is an effective approach for the data mining process to evaluate the

grades of the student on the basis of the various parameter such as name, age, sex etc. In future more advancements can be done in this work by working on various datasets and hybridization of the classifiers is also possible.

REFERENCES

[1] Surjeet Kumar Yadav, "Data Mining Applications: A comparative Study for Predicting Student's performance", ijitce, vol 1(12), Pp 13-20,

[2] Surjeet Kumar Yadav, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", WCSIT, vol 2(2), Pp 51-56, 2012,

[3] Brijesh Kumar Baradwaj, "Mining Educational Data to Analyze Students‟ Performance", ijacsa, vol 2(6), Pp 63-70, 2011,

[4] SAYALI RAJESH SUYAL," Quality IMPROVISATION OF STUDENT PERFORMANCE USING DATA MINING TECHNIQUES", vol 4(4), Pp 1-4, 2014,

[5] Paulo Cortez," USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE

[6] Amirah Mohamed ShahirI ," A Review on Predicting Student's Performance Using Data Mining Techniques ", ELSEVIER, vol 72, Pp 414-422, 2015,

[7] Alaa Mustafa,"Mining Student Data to analyze Learning behaviour", Rsearch gate, 2009,

[8] Muluken Alemu Yehuala, "Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University)", vol 4(4), Pp 91-95,

[9] Jennifer Sabourin," Student Privacy and Educational Data Mining: Perspectives from Industry", IEEE, Pp 164-170

[10] George Siemens," Learning Analytics and Educational Data Mining: Towards Communication and Collaboration", Pp 252-254, 2012

[11] Cristóbal Romero," Educational Data Mining: A Review of the State-of-the-Art", IEEE, vol 20, Pp 1-19, 2010

[12] Luis Talavera," Mining Student Data To Characterize Similar Behavior Groups In Unstructured Collaboration Spaces",CSCL, Pp 17-23, 2004,

[13] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) Vol. 2(2), pp.686-690, 2011.

[14] Pimpa Cheewaprakobkit," Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program", IMECS, vol 1, Pp 1-5, 2013

[15] M. R. Beikzadeh and N. Delavari, "A New Analysis Model for Data Mining Processes in Higher Educational Systems,"citesserx, pp 5-9, 2004,

[16] Ajinkya Kunjir," Recommendation of Data Mining Technique in Higher Education", IJCER, vol 5(3), Pp 29-35, 2015,

[17] Jayashree M Kudari," Survey on the Factors Influences the Students‟ Academic Performance", IJERMT, vol 5(6), Pp 30-37, 2016,

[18] RYAN S.J.D. BAKER," The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of education data mining, vol 1(1), Pp 1-14, 2009,

[19] Dr. Varun Kumar," An Empirical Study of the Applications of Data Mining Techniques in Higher Education",IJACSA, vol 2(3), Pp 80-85, 2011,

[20] K. Amarendra "Research on Data Mining Using Neural Networks" Special Issue of International Journal of Computer Science & Informatics (IJCSI), , Vol.- II, Issue-1, 2, Pp2231–5292,

[21] Anand V. Saurkar,(2014) , "A Review Paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engi neering, Volume 4, Issue 4,Pp 98-101,

[22] Dr. T. Karthikeyan(2012) " A Study on Ant Colony Optimization with Association Rule" International Journal of Advanced Research in Computer Science and Software Engineering

[23] Neelamadhab Padhy (2012), "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3,Pp 43-58