# An Efficient Data Mining Approach to Improve Students' Employability Prediction

Nancy Kansal
Research Scholar
Mewar University
Chhitorgarh, India

Vineet Kansal, PhD
Professor,
CSE Department
IET Lucknow, India

## ABSTRACT
Students' Employability Prediction is a major concern for the institutions offering higher education. A method for early prediction of employability of the students is always desirable to take timely action. In this paper, an efficient data mining approach is proposed to improve students' employability prediction. For Students' Employability Prediction, there are some steps which include data acquisition, data pre-processing, feature subset selection and decision and ranking. For feature subset selection, Chi-Square, Gini index, Information Gain and Correlation Coefficient methods are used and for best feature selection Crow Search based Feature Selection Algorithm is used. Hybridized Hidden Markov Model and Support Vector Machine (HMM-SVM) is used for the prediction of employability. The experimental results are carried out and compared with some existing methods which includes the classification based on Support Vector Machine (SVM), Hidden Markov Model (HMM), k-Nearest Neighbour (kNN) and Artificial Neural Network (ANN). When comparing with the existing methods, the experimental outcomes showed that 93.4% accuracy was obtained by using HMM-SVM classifier.

## General Terms
Educational Data Mining, Employability, Prediction

## Keywords
Gini Index, Chi-Square, Crow Search based Feature Selection Algorithm, Hidden Markov Model, Support Vector Machine

## 1. INTRODUCTION
Data mining is a computer based system which helps to analyse or extract data from large data storage, generate the information and discover knowledge. This data mining technique is used in many fields including marketing, management, engineering, web mining, and industries [1]. Different fields store large huge supply of data in web based systems. This will drive in the direction of evolution of new methodologies in data analysis [2]. The large datasets are discovered by the concepts and techniques of data mining [3]. In recent times the data mining techniques are developed such as classification, clustering, association, pattern matching, data visualization and meta-rule guided mining [4].

Now a days, the education system is computerized for providing effective and efficient education to the students. The institutions stores large amount of data like student enrolment, attendance and also the academic results of the students. The educational data mining is used to resolve issues in analysing the stored data by applying methods such as machine-learning, statistics, information retrieval and recommender systems [5]. For effective learning process innovation is important to maintain student performance [6]. There are some initial approaches in educational data mining such as prediction, clustering, relationship mining, discovering models and the refinement of data for human decisions [7].

In education field, the colleges and universities faces dropout rates of students. The machine learning technique in data mining helps to control the dropout rate of students in universities [8]. The dropout behaviour is based on the student academic and demographic variables such as gender, employment status, chosen subject, etc. The bottom-up approach using data mining is used to handle with these variables [9]. Logistic regression method is used to predict whether the student will continue the course or not. Obtained grades of students from the traditional courses are considered for this approach [10].

In this decade, unemployment is major crisis because of increasing graduates. The graduates with ability of working skill only get placed in industries. The graduate employability model implies, test the model which learning transfer from university to working place [11]. Analysis about the placement success of students helps potentially to improve the academic achievement. By adopting the k-fold cross method, the predictive power of the models can be measured [12]. The effective early warning system for prediction is needy to avoid the issues in the appropriated fields [13].

The use of Data mining technique in the education field is increasing now a days. There are many techniques used in educational data mining like Decision tree, Neural network, Naive Bayes, K-nearest neighbour, etc. By using these techniques some tasks discover knowledge such as association rules and clustering [14]. The decision tree is a type of classification technique it builds regression model in the form of tree structure [15]. The prediction technique in learning management system (LMS) data is a tool to find the at-risk students and provide governance for them. The tracking variables of the students correlated with student performance [16]. The goal of prediction is to obtain the unknown value. Training set of data is used to guide the learning process and test set is used to analyse the performance of students [17].

The recommender system approaches a new method to discover the student performance. The matrix factorization is not used yet in this era but it applies recommender algorithm to predict the performance of students [18]. Application of data mining techniques to educational data has provided support to the institutions in many activities. Predicting student employability can help to identify the students who are at risk of unemployment and thus management can intervene timely and take essential steps to train the students to improve their performance.

Employability of Students in any HEI needs an early prediction so that the corrective measures can be taken by the Institution. In earlier researchers have tried to establish links

between academic performance, socioeconomic conditions, job skills of the students and employability, however, mostly by deploying statistical method. Apart from above stated parameters, it also explores the link between and emotional skills like assertion, empathy, decision making, leadership, drive, stress management to predict employability using data mining techniques. The contribution in this research work includes the following.

- A novel natural inspired optimization algorithm named CSFS algorithm is used to select the best subset of features from an original set of feature of dataset without downgrading the performance.
- For prediction purpose, hybridization of HMM and SVM is used, so that the best accuracy has been achieved.

The rest of the paper is organized as: Section 1 provides the introduction. Section 2 contains the related work. Section 3 comprises the proposed methodology. Section 4 shows the experimental analysis. Section 5 contains the conclusion of the paper and Section 6 provides the references used in the paper.

# 2. RELATED WORK

Some of the recent research works related to students' employability prediction are given below:

Tripti Mishra *et al.* [19] applied various classifiers to find the employability of students and develop employability model based on the suitable classifier. It was found that J48 algorithm which was the implementation of pruned C4.5.Based on the experimental results they conclude that the Decision Tree algorithm of WEKA was the most suitable for the employability prediction of students.

Yogesh Bharambe *et al.* [20] proposed a methodology using data mining technique such as classification that was used to predict the employability status of the students. Classification analysis was done and revealed that Random Forest has the highest accuracy over other classification algorithms such as decision tree, KNN, Random forest, Naive Bayes, logistic regression, SVM (Linearsvc), Multi-class Ada Boosted and Quadratic Discriminant Analysis (QDA). Thus Random Forest classification algorithm was used on students' skillset data to predict the students' employability. Also students' strengths and detailed weaknesses were analysed so that they can overcome their weaknesses in order to get their expected company. Students were suggested the list of suitable companies having similar requirements for the job profile as per their skill sets. Pooja Thakar *et al.* [21] presented an empirical study that compared various classification algorithms on two datasets of MCA students collected from various affiliated colleges of a reputed state university in India. One dataset included only primary attributes, whereas other dataset was feeded with secondary psychometric attributes in it. The results showed that solely primary academic attributes didn't lead to smart prediction accuracy of students' employability, once they square measure within the initial year of their education. The study analyzed and stressed the role of secondary psychometric attributes for better prediction accuracy and analysis of students' performance. Timely prediction and analysis of students' performance could help the Management, Teachers and Students to work on their gray areas for better results and employment opportunities.

Bangsuk Jantawan and Cheng-Fa Tsai [22] built the Graduates Employment Model using classification task in data mining, and compared data-mining approaches such as Bayesian method and the Tree method. The Bayesian method included 5 algorithms, including AODE, BayesNet, HNB, NaviveBayes, WAODE. The Tree method included 5 algorithms, including BFTree, NBTree, REPTree, ID3, C4.5. The experiment used a classification task in WEKA, and compared the results of each algorithm, where several classification models were generated. To validate the generated model, the experiments were conducted using real data collected from graduate profile at the Maejo University in Thailand. The model was intended to predict whether a graduate was employed, unemployed, or in an undetermined situation.

Keno C. Piad *et al.* [23]predicted the employability of IT graduates using nine variables. First, different classification algorithms in data mining were tested making logistic regression with accuracy of 78.4 was implemented. Based on logistic regression analysis, three academic variables directly affected; IT_Core, IT_Professional and Gender identified as significant predictors for employability. The data were collected based on the five year profiles of 515 students randomly selected at the placement office tracer study.

# 3. PROPOSED METHODOLOGY

In this paper, an efficient data mining approach is proposed to improve students' employability prediction. The proposed approach comprises three phases: data acquisition, data pre-processing and prediction. Data Acquisition phase is responsible for two tasks: data acquisition and data aggregation. These tasks are performed prior to the core computation and processing of data. Data Computation and Processing Phase is the core processing phase and responsible for the overall computation such as pre-processing and robust feature extraction. Prediction is responsible for decision making and prediction of student's employability based on identified features. For efficient feature selection, feature selection methods used are Information Gain, Chi-Square, Gini index and Correlation Coefficient. Further, *Crow Search based Feature Selection (CSFS)* based solution is introduced for the purpose of feasible feature selection. Finally, HMM-SVM is employed to predict the employability according to their selected feasible features through prediction phase.

# 4. EXPERIMENTAL ANALYSIS

The proposed model is implemented in the platform MATLAB 2017a version, which is a high-performance language for technical computing and integrates computation, visualization, programming in an easy-to-use environment.

## 4.1 Dataset Description

The dataset used for this work is created manually based on the structured questionnaire method. The dataset is created with the aim of understanding the students' ability and skills which are necessary for predicting the employability of every student. Nearly ten thousand students' details are gathered for related work. The questionnaire comprised of both the option type and open ended questions on the basis of information being elicited. The collected data was recorded on structured forms and it is used for the proposed method.

## 4.2 Performance Metrics

When classifying, four categories are considered such as True positives (TP), True negatives (TN), False positives (FP) and False negatives (FN). True Positives are the examples which are correctly labeled as positives. False positives refer to

negative examples which are incorrectly labeled as positives. True negatives correspond to the negatives which are correctly labeled as negatives and the False negatives refer to the positive examples which are incorrectly labeled as negative. Based on the four categories we have to evaluate the performance metrics such as Accuracy, Error, F1-score, Kappa, Precision, Sensitivity and Specificity.

## 4.3 Analysis

The performances of the proposed method, which is HMM-SVM are evaluated and compared with the existing methods such as ANN, HMM and SVM classifiers. The existing methods are the similar machine learning algorithms, therefore, they have been considered for the comparison.

By hybridizing the HMM and SVM classifiers, better results are attained than the existing methods. Using the experimental results, the values of the existing methods and the proposed method are compared in Table 1.

**Table 1. Comparison values of Existing and Proposed Methods**

| Performance Metrics | Proposed Method (HMM-SVM) % | ANN % | HMM % | SVM % |
|---|---|---|---|---|
| **Accuracy** | **93.4** | **83.8** | **90.2** | **83.6** |
| **Error** | **6.6** | **16.2** | **9.8** | **16.4** |
| **Sensitivity** | **95.2** | **88.3** | **92.4** | **88** |
| **Specificity** | **90.1** | **75.8** | **86.3** | **75.8** |
| **Precision** | **94.3** | **86.4** | **92.1** | **86.4** |
| **F1-score** | **94.8** | **87.3** | **92.3** | **87.2** |
| **Kappa** | **85.7** | **64.7** | **78.8** | **64.3** |

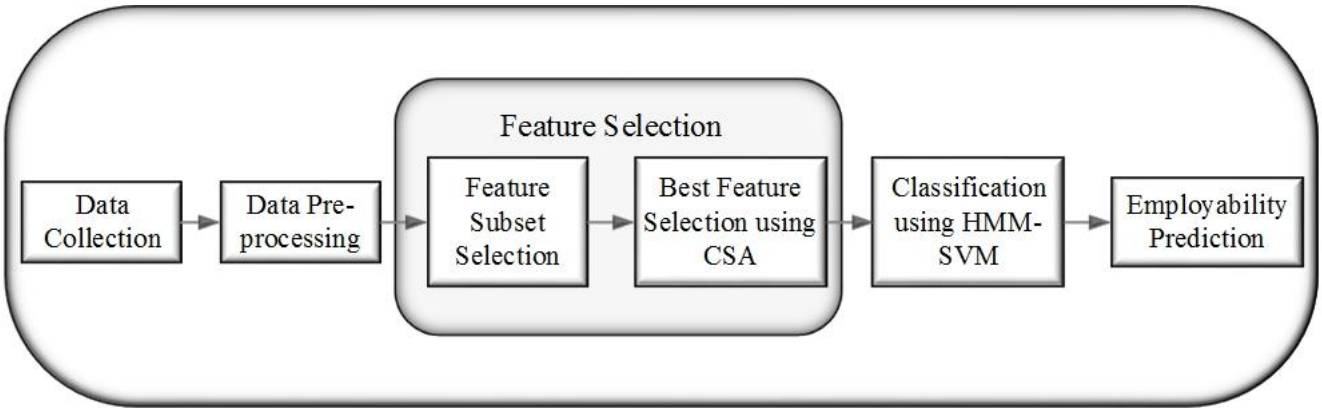


**Figure 1: Proposed Architecture**

**ANN:** ANN is efficient for large datasets and the number of hidden nodes in the network is considered as free parameter. Once a network has been structured for a particular application, that network is ready to be trained. There are two approaches to training, supervised and unsupervised. The most often used ANN is a fully connected, supervised network with backpropagation learning rule. This type of ANN is excellent at prediction and classification tasks.

**HMM:** HMM is a powerful statistical technique for modeling complex sequences of data. This classifier is a special kind which aims to find the posterior probability of each state given a sequence of observations and predicts the state with the highest probability.

**SVM:** An SVM is a discriminative classifier formally defined by a separating hyperplane. It can be efficient for the nonlinear datasets. The number of support vectors depends on the optimization problem that occurs and each support vector is always a subset of data. It utilizes kernel functions for formulating the linearity concept in the nonlinear datasets.

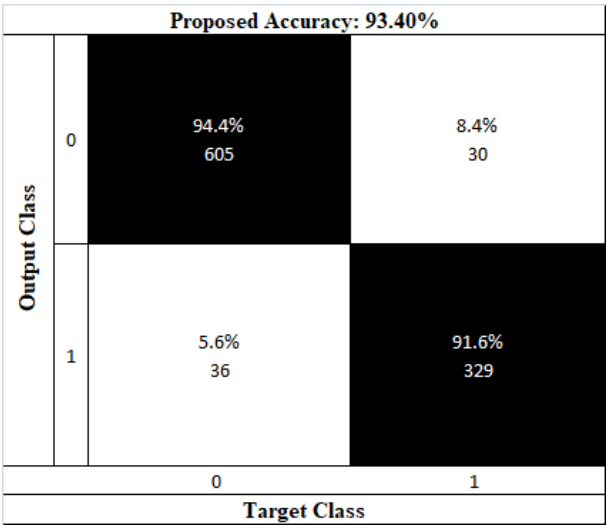


**Figure 2: Confusion Matrix**

Figure 2 shows the confusion matrix of the HMM-SVM model. The confusion matrix contains true positive, true negative, false positive and false negative values. From the confusion matrix, the values are represented in the table 2. Using these values the performance metrics are evaluated.

**Table 2: Positive and Negative values**

| | |
|---|---|
| **True Positive** | **605** |
| **True Negative** | **329** |
| **False Positive** | **36** |
| **False Negative** | **30** |
| **Positives** | **641** |
| **Negatives** | **359** |

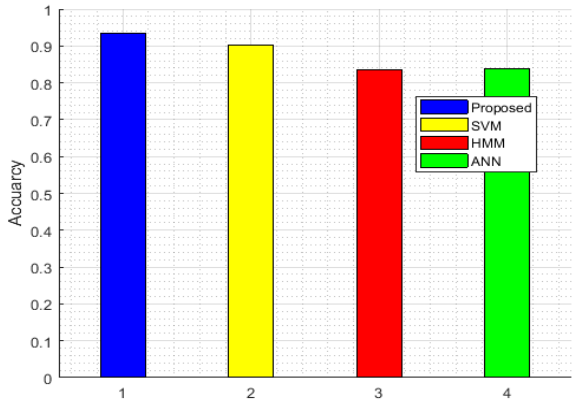


**Figure 3: Comparison of Accuracy**

Figure 3 represents the accuracy of the proposed method as well as the existing methods. The accuracy is attained highly as 93.4% for the HMM-SVM based classification. When SVM is used for classification, the accuracy obtained is 83.6%. The accuracy obtained by using the HMM classifier is 90.2% and when using the ANN classifier, the accuracy obtained is 83.8%.

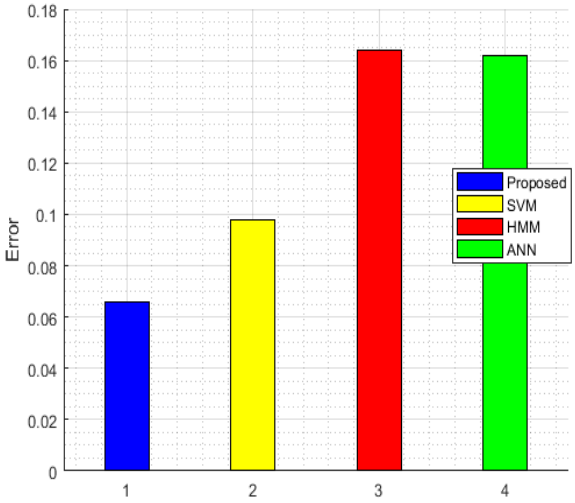


**Figure 4: Comparison of Error**

Figure 4 shows the comparison graph of error. The graph represents that the error is low for our proposed method when compared with the existing methods. Here, the error is high when we are using the SVM classifier, which is 16.4% and the error is low when using the HMM-SVM, which is only 6.6%. When using the ANN classifier, the obtained error is 16.2% and also when using the HMM classifier, the error attained is 9.8%.

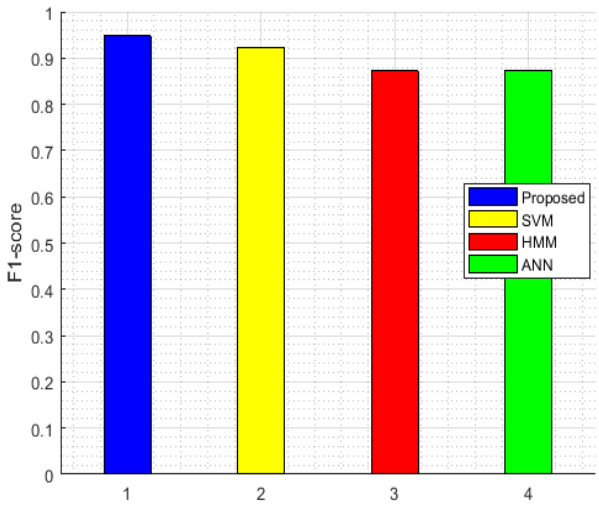


**Figure 5: Comparison of F1-Score**

Figure 5 represents the comparison graph of F1-score. The graph shows that the F1-score is high for our proposed method when comparing with the existing methods. Here, the F1-score is low for SVM classifier, which is 87.2%. And also the F1-score is high when using the HMM-SVM, which is 94.8%. When using the ANN classifier, the obtained F1-score is 87.3% and when using the HMM classifier the F1-score attained is 92%.

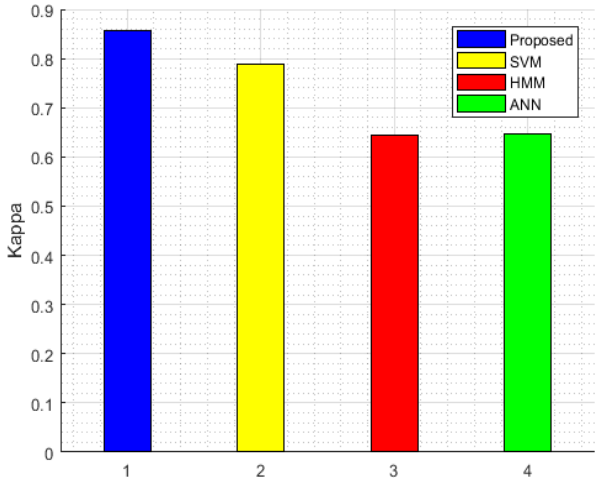


**Figure 6: Comparison of Kappa**

Figure 6 shows the comparison graph of kappa. The graph represents that the kappa is high for our proposed method when comparing with the existing methods. Kappa is low when we are using the SVM classifier, which is 64.3% and it is high as 85.7% when using our proposed approach. When using the HMM classifier, the obtained kappa is 78.8% and also when using the ANN classifier, the kappa attained is 64.7%.

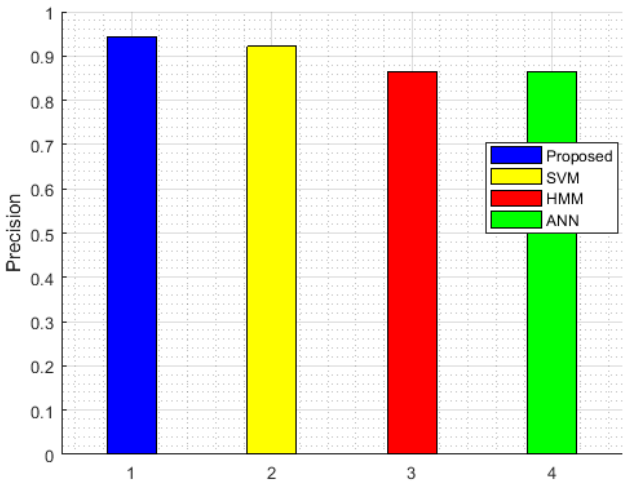


**Figure 7: Precision Comparison**

Figure 7 shows the graphical comparison of precision of the proposed method as well as the existing methods. The precision is attained highly as 94.3% for the HMM-SVM based classification. When SVM is used for classification, the precision obtained is 86.4%. The precision obtained by using the HMM classifier is 92.1% and when using the ANN classifier, the precision obtained is 86.4%.

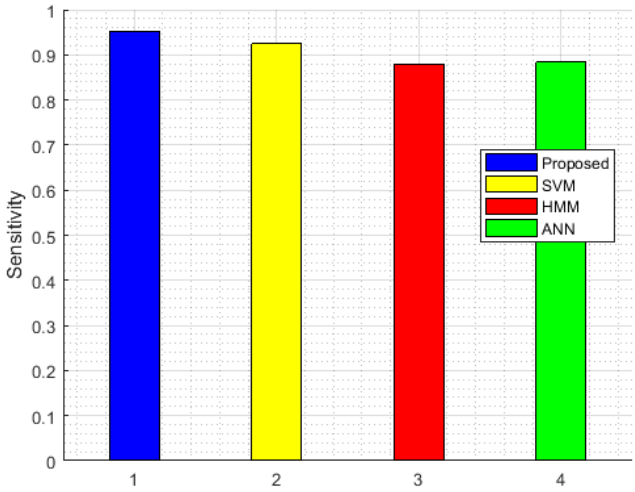


**Figure 8: Sensitivity Comparison**

Figure 8 represents the comparison of sensitivity of the proposed method and the existing methods. The sensitivity is high for the HMM-SVM based classification. When ANN is used for classification, the sensitivity obtained is 88.3%. The sensitivity obtained by using the HMM classifier is 92.4%.When using the SVM classifier, the sensitivity obtained is 88% and high sensitivity of 95.2% is attained by using the HMM-SVM classifier.

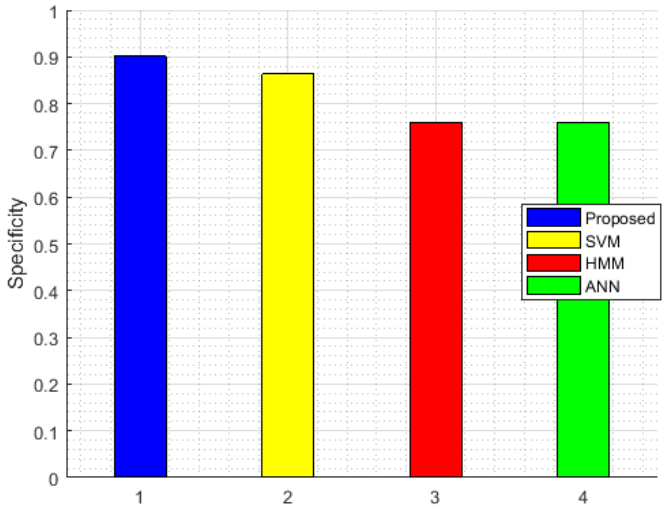


**Figure 9: Specificity Comparison**

Figure 9 represents the comparison of specificity of the proposed method and the existing methods. The specificity is high for the HMM-SVM based classification. When ANN is used for classification, the specificity obtained is 75.8%. The specificity obtained by using the HMM classifier is 86.3%. When using the SVM classifier, the specificity obtained is 75.8% and high specificity of 90.1% is attained by using the HMM-SVM classifier.

In Figure 10(a), the graph is the training accuracy with respect to the number of data. Training accuracy is estimated for the classifiers which includes SVM, HMM and HMM-SVM. The graph depicts that the proposed method has the highest training accuracy. In Fig. 10(b), the graph represents the testing accuracy of the classifiers with respect to the number of data. Testing accuracy also high for the proposed method when comparing with the other existing classifiers.

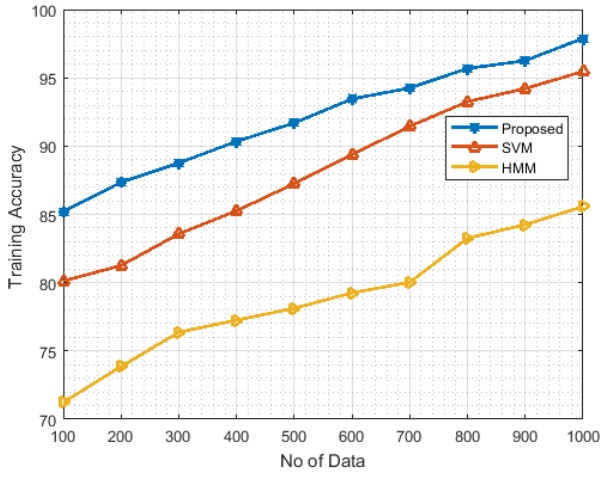


**Figure 10: a) Training accuracy**

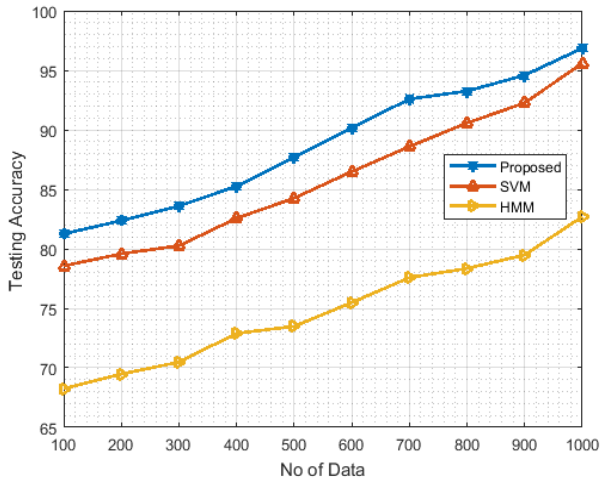


**Figure 10: b) Testing accuracy**

From the experimental analysis, it is proved that the performance metrics which includes accuracy, precision, sensitivity, sensitivity, kappa, F1 measure are improved when compared with the existing classifiers such as ANN, SVM and HMM. The proposed method provides better outcome by using the best feature selection method.

## 5. CONCLUSIONS

In this paper, employability of the students is classified and predicted using the hybridized HMM and SVM classifier. The HMM-SVM is useful for calculating employability smoothness of students in a simple manner. Through the hybridized HMM-SVM, employer can easily filter the best applicants based on their education skill and personnel development skills. The experimental results proved that the students' employability prediction is enhanced by applying our proposed HMM-SVM model. This provides a generalized solution to students' employability prediction and is scalable, so that, it can act as a base for developing unified decision support system in education domain. It plays a significant role in predicting students' employability in terms of selected or not by considering the gathered details. The proposed CSFS and the HMM-SVM employed on the obtained features through the collected data using the questionnaire manner. For selecting the features we are using Chi-Square, Gini Index, information gain and correlation coefficient methods and the CSFS algorithm is used for the best feature selection. Hybridization of HMM and SVM classifier is utilized for the classification process and attained 93.4% accuracy. The experimental results are evaluated and compared with the existing classifiers such as SVM, HMM and ANN.

## 6. REFERENCES

[1] Peña-Ayala, Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." *Expert systems with applications* 41.4 (2014): 1432-1462.

[2] Campagni, Renza, et al. "Data mining models for student careers." *Expert Systems with Applications* 42.13 (2015): 5508-5521.

[3] Ahmed, Shibbir, Rajshakhar Paul, and Abu Sayed Md Latiful Hoque. "Knowledge discovery from academic data using association rule mining." *Computer and Information Technology (ICCIT), 2014 17th International Conference on*. IEEE, 2014.

[4] Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." *Expert systems with applications* 39.12 (2012): 11303-11311.

[5] Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. "A systematic review on educational data mining." *IEEE Access* 5 (2017): 15991-16005.

[6] Shahiri, Amirah Mohamed, and Wahidah Husain. "A review on predicting student's performance using data mining techniques." *Procedia Computer Science* 72 (2015): 414-422.

[7] Asif, Raheela, et al. "Analyzing undergraduate students' performance using educational data mining." *Computers & Education* 113 (2017): 177-194.

[8] Kotsiantis, Sotiris B., C. J. Pierrakeas, and Panayiotis E. Pintelas. "Preventing student dropout in distance learning using machine learning techniques." *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, Berlin, Heidelberg, 2003.

[9] Yasmin, Dr. "Application of the classification tree model in predicting learner dropout behaviour in open and distance learning." *Distance Education* 34.2 (2013): 218-231.

[10] Burgos, Concepción, et al. "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout." *Computers & Electrical Engineering*(2017).

[11] Jackson, Denise. "Business graduate employability–where are we going wrong?." *Higher Education Research & Development* 32.5 (2013): 776-790.

[12] Şen, Baha, Emine Uçar, and Dursun Delen. "Predicting and analyzing secondary education placement-test scores: A data mining approach." *Expert Systems with Applications* 39.10 (2012): 9468-9476.

[13] Geng, Ruibin, Indranil Bose, and Xi Chen. "Prediction of financial distress: An empirical study of listed Chinese companies using data mining." *European Journal of Operational Research* 241.1 (2015): 236-247.

[14] Bunkar, Kamal, et al. "Data mining: Prediction for performance improvement of graduate students using classification." *Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on*. IEEE, 2012.

[15] Keramati, Abbas, et al. "Improved churn prediction in telecommunication industry using data mining techniques." *Applied Soft Computing* 24 (2014): 994-1012.

[16] Macfadyen, Leah P., and Shane Dawson. "Mining LMS data to develop an "early warning system" for educators: A proof of concept." *Computers & education* 54.2 (2010): 588-599.

[17] Xing, Wanli, et al. "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory." *Computers in Human Behavior* 47 (2015): 168-181.

[18] Thai-Nghe, Nguyen, et al. "Recommender system for predicting student performance." *Procedia Computer Science*1.2 (2010): 2811-2819.

[19] Mishra, Tripti, Dharminder Kumar, and Sangeeta Gupta. "Students' Employability Prediction Model through Data Mining." International Journal of Applied Engineering Research 11, no. 4 (2016): 2275-2282.

[20] Y. Bharambe, N. Mored, M. Mulchandani, R. Shankarmani and S. G. Shinde, "Assessing employability of students using data mining techniques," Advances in Computing, Communications and Informatics, Udupi, 2017, pp. 2110-2114.

[21] Thakar, Pooja, and Anil Mehta. "Role of Secondary Attributes to Boost the Prediction Accuracy of Students Employability Via Data Mining." arXiv preprint arXiv:1708.02940 (2017).

[22] Jantawan, Bangsuk, and Cheng-Fa Tsai. "The application of data mining to build classification model for predicting graduate employment." arXiv preprint arXiv:1312.7123 (2013).

[23] Piad, Keno C., Menchita Dumlao, Melvin A. Ballera, and Shaneth C. Ambat. "Predicting IT employability using data mining techniques." In Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), pp. 26-30. IEEE, 2016.