Boosting classifiers for Intrusion Detection

Richa Rawat1 , Anurag Jain2

1, 2 Computer Science Dept., Radharaman Institute of Technology & Science, RGTU Bhopal (M.P.) India

[1] er.richa22 @gmail.com

[2]anurag. akjain @gmail.com

***Abstract* — IDS detect intrusions using data mining techniques & other software techniques. The intrusion detection technique can efficiently expand the scope of defense of network and system. In this work we aim to improve efficiency for intrusion detection system. We proposed host based IDS model. There are two phases in certain ways, in the first phase we are using decision tree and SVM classifiers for classification of data and the second phase we boost both the decision tree and SVM classifiers, and detect intrusion more than a single class classifier system. We are using boosting classifier for miss-classification data sets to improve detection rate. The kddcup99 dataset is used as a simulation set. The result shows that our proposed approach achieves better precision and detection rate by using boosting.**

***Keywords*— Intrusion Detection System (IDS), Boosting, decision tree and support vector machine (SVM)**

I. INTRODUCTION

Information about ensuring safety in the private sector or the government has become a need. A vulnerability in the system, and valuable information to attract the most attention attacked [7]. In essence, the intrusion detection search by abnormal data from normal data to divide, which is a classification problem Intrusion Detection System (IDS) has been applied to detect intrusion [1].

Intrusion detection technology identifies and deals with the use of contaminated computers that the system can be defined as [2]. In this work we have presented of boosted of j48 decision tree and support vector machine classifiers for intrusion detection based on machine learning. We have first used a decision tree for classification of five-class data. We are also conducting experiments with support vector machines (SVM) & then boosting of multiple classifiers. Decision tree using as a binary classifier and svm is a single class classifier we are using one-against-one method in svm for removing multiple classification problem. We are applying boosting on misclassification data set to improve detection rate. KDDcup 1999 benchmark dataset is used for testing the proposed algorithm and the results are promising and more important, especially low false alarm rate and high detection rate with take less time to create a model to achieve, that outperforms the existing methods are presented [8]. Our motivation for IDS developing absolutely secure systems is not possible because most existing systems have security flaws, Abuses by privileged insiders are possible & not all kinds of intrusions are known. Quick detection of intrusions can help to identify intruders and limit damage. IDS serves as a deterrent.

II. INTRUSION DETECTION SYSTEM

Intrusion Detection System (IDS) to detect various types of attacks that play a crucial role. The main function of IDS data to find out intrusions among normal audit, and this classification can be a problem. Overhead is one of the problems of ids, which could be too high [9]. Unauthorized access or use of a computer system resource is intrusions. Intrusion detection system, can be used as an identifier is a software which is used to identify and target system responds to unauthorized or unusual activities. The main purpose performed by an intrusion detection system are: (1) monitor and analyze user and system activities, (2) to assess the integrity of critical system and data files, (3) to identify the activities that are known to respond automatically detect patterns reflecting the activity, and report the results of the identification process [10].

This paper is structured as follows: Section 2 gives the details of intrusion detection system. Section 3 gives the detail of related work. Section 4 gives the detail of data mining method which is including details of machine learning Decision Tree and SVM classifiers, ensemble method, boosting. Section 5 gives the details of the experiment and the results. Section 6 Conclusion.

III. RELATED WORK

In the following, we review the related work, focusing on different types of method that are described in the literature.

Jingbo Yuan et. Al. [1] proposed a svm classification (HTSVM) a conventional SVM classification theory is hypothesis test. Classification process, a soft edge version of the update that is compatible with the conventional SVM, but especially in the determination of the boundaries of the soft edge of the attribute data for hypothesis testing in previous training. CSVM than classification, HTSVM classification capability and enhanced generalization ability of learning. Simulation experiment results show that the intrusion detection system performance can be improved.

Mohhamadreza Ektefa (2010) [2] proposed Classification tree and two leading data mining methods as support vector machine techniques to detect network intrusions. As Experimental results show, C4.5 algorithm is

better Detection and false alarm performance of the SVM on a data sample rate.

P Amudha et al.[3] Proposed Classification and measurement in relation to the attack and hybrid attribute selection and ensemble of classifiers in order to analysis the efficiency of a series of experiments on KDD Cup'99 dataset. The experimental results, NBTree have small training data and a better detection rate and false alarm rate for R2L dataset and U2R datasets that allows for better accuracy, while the random forest, good accuracy, and detection rate, false alarm rate for DOS. They also build the model to the time taken by NBTree is further observed that compared to other classifiers. They conclude the random forest for DOS and probe dataset and NBTree for U2R and R2L dataset gives better performance.

Bayesian network intrusion detection system proposed by **Farah Jemili** [4] using an adaptive framework emphasized. Bayesian networks provide an automated search capability; they learn from the audit data and can detect both normal and abnormal connections. From audit information and can detect both normal and abnormal connections. Such Intrusions showed the high performance of their system. The system is able to provide recommendations based on attack types, which can be improved by integrating expert system. Another alternative to represent a qualitative assessment of the risk of infiltration of possibility networks, Bayesian networks are used.

neural network for a network anomaly detection method has been proposed by **A.S. Aneetha et. Al. [5]** and combined with the use of clustering algorithms has been proposed. Last modified self-organizing map of the SOM but for k-means that the implementation of the 1.5% increase it more than 2% higher detection rate has improved. The rate of increase in the number of output nodes, reducing the rate of learning, and learning to play a major role in the spread of the observed map found. DOS attacks, the detection rate of 98.5%, the proposed work is to identify the most effective.

H. Günes Kayacık et. Al. [6] Comprehensive analysis of the relevance of a feature of the machine is used by education researchers, who are on the KDD 99 training set. As a discriminating feature to feature high on the relevance of the information is expressed in terms. Training for all categories of feature sets in order to measure relevance, information gain per class gain as a result of separate data for each feature, binary classification is calculated on. Recent research decision trees, artificial neural networks, and the potential for functional classification and the remote user root and local attacks are very difficult to classify, search terms, and the false alarm rate, report.

IV. Proposed work

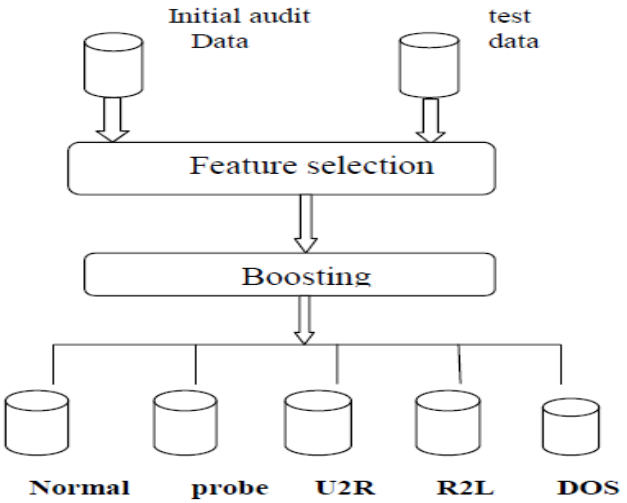

Figure 1. Experimental Flow

There is a set of records (training set) where each record contains a set of attributes, one of the features are class. In general, some of the data used in the model is set to training set and test set is used to authenticate, separated into training and test sets. We plan to select the various features of the different layers of the training. Trained using the same training data for each classification. Ensemble classifier is used for classification. Here we use the decision tree and SVM ensemble classification by boosting. Classification of knowledge presented by the training committee. This training dataset are classified into five classes, they are Normal, Probe, DOS, and U2R and R2L [15].

V. DATAMINING MACHINE LEARNING TECHNIQUES

Education statistics compact machine that uses artificial intelligence is a field. Machine learning algorithms, based on a data set that enables computers to make predictions. Machine learning; learn how to make a machine that enables a computer. In case of intrusion detection, normal or perverted action of the algorithm on the host should be able to predict. Many applications include machine learning such as Bioinformatics, brain-machine interface to find credit card fraud, stock market analysis, and so on [2].

A. Decision tree

Technology decision tree is a common method of classification, intuitionist and quickly. The Construction process is top-down allocation and based on dividing and rule mechanism. It is important greedy algorithms. Start from the root node for each node is non leaf node, first select an attribute to try to set an example; Second, a model train set into several sub-sample set, according to the experimental results, a set of samples creates a new leaf

node. Third, repeat the above divisional process, until having reached specific end conditions. Decision tree of the different algorithms using different technology. In practice, because the size of the sample set of training is usually large tree branches and layers of more. In addition, abnormal noise occurs in the sample set of training will also cause some branches are unusual, so we need to prune the decision tree. One of the greatest advantages of the decision of the algorithm, the classification tree is that: it does not require the user to know that a lot of background knowledge in the learning process. At present, many of the algorithms, such as a decision is: ID3, SLIQ, carts, CHAID, and more. But J48 algorithm is the most representative and widely used. It was proposed by Quinlan in 1993. J48 is a one-class classifier so it gives best result for one-class. We are using C4.5 algorithm which is updated algorithm of j48.

B. SVM

Support Vector Machines applied as a novel intrusion detection. A Support Vector Machine (SVM) maps to evaluate the subject gives instructions input point to be made in higher space to the point measures through some nonlinear mapping. SVM is a powerful tool for solutions to the problem of evaluating classification, regression and density. The framework builds on the principle of reducing risk. They speculated that a low probability of error approach in order to try to reduce the risk [15].

We solve multiclass classification problem in svm by using one-against-one method. In this method we construct n (n-1)/2 classifiers where each classifier is trained on data from two classes. "One against one" strategy, which is also known as "Pair wise Coupling", "all pairs" or "round-robin", each pair consists of the construction of an SVM Classes. Thus, the problem of class n, n (n-1)/ 2 Preparation of samples for SVMs to distinguish one class from the samples of the second class. Usually, the classification of an unknown sample is done by the maximum voting, where each SVM classifiers Votes for one class.

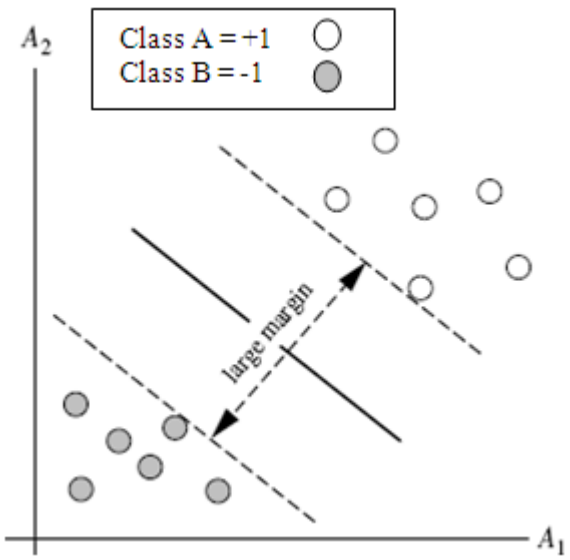

Figure 2 Two possible separating hyper lanes

VI. ENSEMBLE LEARNING

In ensemble method multiple learners are trained to resolve the identical problems, where is the machine learning paradigm. Usually referred to as an ensemble of base learners, which includes a number of learners. The generalization capability of an ensemble is regularly stronger than the base learners. It can make very accurate predictions of the strong learners slightly better than a random guess is; the weak learners are able to increase, because really, the ensemble learning appeals. Therefore, the basis of learners and "weak learners" are called. Base learners in general, the decision tree, neural network or other types of machine learning algorithms may be based learning algorithm that is generated from the training data. There are many ensembles learning technique which is Boosting, Bagging, Stacking. We are using BOOSTING ensemble learning technique.

A. Boosting

We can use boosting learning with individual classifier or multiple classifiers. Boosting used with multiple classifiers can improve efficiency of detection. Assign a weight to all training example, initially; each example is assigned a weight 1/n. According to the sampling distribution of the training examples drawn samples to obtain the new training data set. Trained classifier used to classify all examples in the original data set. Examples of weight training are updated at the end of each boosting round. Which is misclassified or correctly classified data. If data are misclassified then weight increased or correctly classified then weight decreased. It can build a strong classifier.

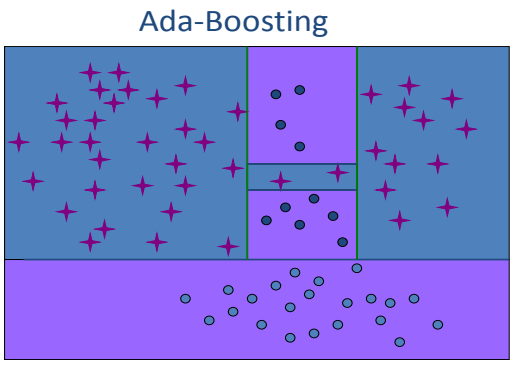

Figure 3 classification of data with different classifiers (j48 & svm)

VII. ALGORITHM

Step 1. We trained j48 classifier and apply testing on 5-class data which is classified or misclassified data.

Step 2. J48 is a one-class classifier; it gives better results for dos class.

If data class ~= best class (j48)
Then svm classify.

Step 3. Now we trained svm classifier and apply testing on 5-class data, it is classifies data which are misclassified by j48.

Step 4. For multiple classifications we are using one-against-one method in svm, it creates multiple classifiers and using voting for classification.

Step 5. We are using boosting for misclassifiers.

- Assign a weight to each training example.
  - Initially, each example is assigned a weight 1/n
- According to the sample distribution of the training examples drawn sample and obtain the new training data set.
- As a result classifier is trained used to classify all examples in the original data set.
- Weight of training examples are updated at the end of each boosting round.
- Adaptively change the weight at the end of each boosting round.

– The weight of an example correctly classified decreases.

– The weight of an example incorrectly classified increases.

- Each round generates a base classifier.

VIII. EXPERIMENT AND RESULT

The decision tree and support vector machine method are to be implemented within the mathematical computer language MATLAB. We have used KDDCup'99 intrusion detection dataset, which contains 31146 records with .8 training ratio. We have carried out our experiment on a Pentium 4 CPU 2.20GHz with 4GB RAM. We have first used a decision tree for classification of five-class data which is (normal, dos, u2r, r2l, and probe). This did classified or misclassified. We also conducted experiments with support vector machines (SVM) using one-against-one method, which created multiple classifiers and then classified data using a voting technique. SVM classified data which were misclassified by j48 & then apply boosting on multiple classifiers. This was focused on misclassified classifiers. Boosting creates a number of rounds and continues the process until all data sets are correctly classified. Attack detection can be measured by the following metrics.

False positive (FP): Or false alarm, Number corresponds to a detected attack, but it is a fact General.

False negative (FN): Corresponds to the number found in normal instances, but it is actually Attack, the target of these attack intrusion detection systems.

True positive (TP): The number of detecting attacks and that the attack is identified.

True negative (TN): Is the number of normal cases detected and it is actually normal.

An intrusion detection system is the accuracy of the detection rate and false alarm rate-related measures.

A. Detection rate comparison (TPR):

Detection rate refers to the percentage of attacks detected in all the information, and are defined as follows:

$$\text{Detection rate} = (TP/P+N)*100$$

B. False alarm comparison (FPR):

Incorrectly recognized as a false alarm rate of the attack, which refers to the percentage of normal data, and are defined as follows:

$$\text{False alarm rate} = (FP/P+N)*100$$

- Recall: The total percentage of relevant documents retrieved from a database Search. This is a database of relevant documents and the relevant documents, 100 were recovered Search 1000 is known, and then the recall will be 10%. It is calculated as below.

$$\text{Recall} = TP/(TP+FN)$$

- Precision: The number of documents related to the percentage of relevant documents retrieved. Search for and retrieve 100 documents are relevant to the 20, the precision is 20%. It is calculated as below.

$$\text{Precision} = P/(TP+FP)$$

- The overall success rate classifications divided by the total number of correct classifications.

$$\text{Success rate} = (TP+TN)/(TP+TN+FP+FN)$$

Error Rate = 1- Success rate

Table-1 Accuracy of different attacks through j48, SVM and Boosting

Method	Dos	Normal	Probe	R2l	U2r
J48	99.65	99.17	99.64	99.83	99.99
SVM	99.49	99.42	99.94	1	99.99
Ensemble Approach [21]	99.92	99.70	100	97.16	68.00
boosting	1	99.94	99.96	1	99.99

Table-1 shows the performance of accuracy of two classifiers and ensemble approach (boosting). From above result we can conclude that boosting gives better accuracy other than single classifier.

Table-2 Shows the result for j48, svm & boosting.

parameter	J48	svm	boosting
TPR	0.9914	0.9942	0.9994
FPR	0.0357	0.0291	0.0035
ACCURACY	0.9869	0.9943	0.9995
PRECISION	0.9861	0.9942	0.9994
RECALL	0.9914	0.9942	0.9994

From table 1 & 2. It is clear that boosting gave the best result.

Now we compare the result of the j48, svm & boosting. Firstly we run the j48 classification algorithm on kdd99 data. Secondly we run svm algorithm with one-against-one method & create multiple classifiers then we run boosting algorithm on misclassified classifiers to improve detection accuracy for intrusion detection.

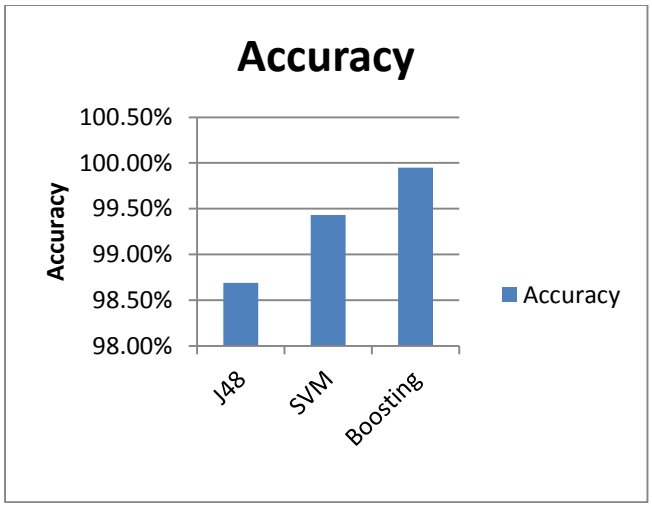

Figure 4 Comparison accuracy for j48, SVM & BOOSTING

From above fig. 4. It is clear that our proposed method gives the better accuracy which is desirable for good intrusion detection.

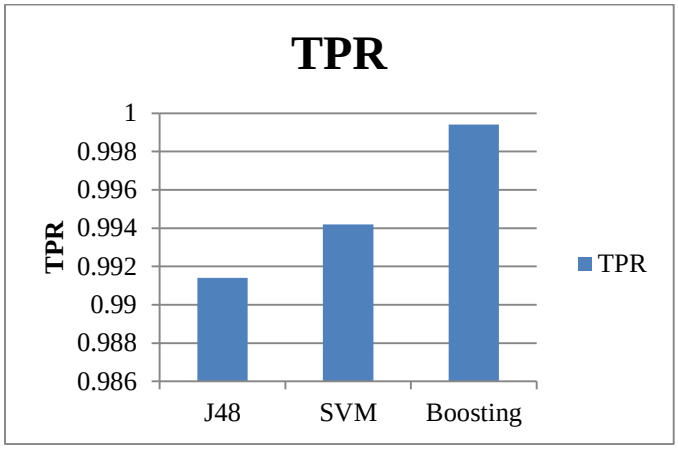

Figure 5 TPR comparison of J48, SVM & BOOSTING

Now we compare the TPR of the J48, SVM & boosting. For a good IDS TP rate should be high. Above figure 5. Shows that the TP rate of boosting classifiers is a high comparison of single classifiers.

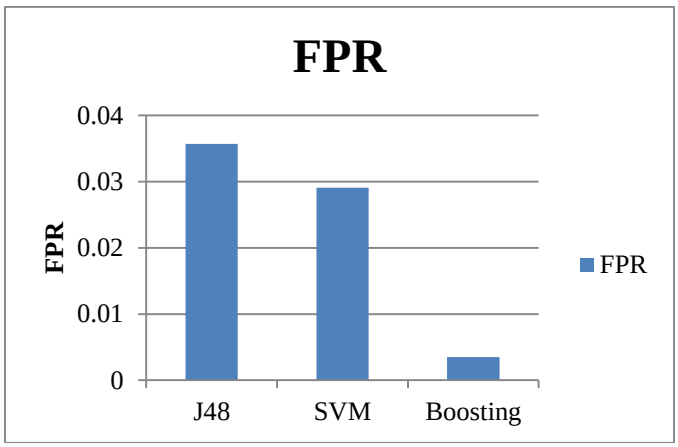

Figure 6 FPR comparison of J48, SVM & BOOSTING

Figure 6 shows that FPR of boosting classifiers is near about zero which is best for desirable intrusion detection. For good IDS FPR should be Low. From above fig. 4, 5 and 6 it is clear that boosting classifier algorithm accuracy, TPR, FPR is better than single classifier.

IX. CONCLUSION

We have proposed BOOSTING CLASIFIERS for intrusion detection. In the algorithm, J48 and SVM are used as weak classifiers. And evaluated their performance on the

benchmark KDD Cup 99 Intrusion data. We have first used a decision tree for classification of five-class data. We also conducted experiments with support vector machines (SVM) & then boosted of multiple classifiers. We boosted misclassified data to improve detection rate. The empirical results show the result of decision tree & SVM and finally boosting gives better result than both. So we have better accuracy 99.95% by boosting and 99.94% detection rate & 0.003 false positive rates also. The results also show that testing time and training time of the classifiers are slightly better than SVM. To remove multiple classification problems in svm we used "One-Against- One" method. It also helps increasing detection rate. Our Future work would include the Ensemble approach be investigated with various combinations of classifiers for network based IDS model to improve the performance.

REFERENCES

1. Jingbo Yuan, Haixiao Li," *Intrusion detection Model based on Improved support Vector Machine*" pp. 465-469 IEEE-2010.
2. Mohammadreza Ektefa," *Intrusion Detection Using Data Mining Techniques*" pp.200-203 IEEE-2010.
3. P Amudha, H Abdul Rauf," *Performance Analysis of Data Mining Approaches in Intrusion Detection*", IEEE-2011.
4. Farah Jemili," *A Framework for an Adaptive Intrusion Detection System using Bayesian Network*", pp. 66-70, IEEE-2007.
5. A.S. Aneetha and Dr. S. Bose, "*The Combined approach for Anomaly Detection using Neural Networks and Clustering Techniques*", Vol.2, No.4, pp. 37-46, CSEIJ-2012.
6. H. Günes Kayacık," *Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets*".
7. Z. Muda,"*Intrusion Detection based on K-Means clustering and oneR Classification*", pp.192-197, 2011, IEEE.
8. Mrutyunjaya Panda1 and Manas Ranjan Patra," *Ensemble Voting System for Anomaly Based Network Intrusion Detection*", Vol 2, No. 5, ACEEE-November 2009.
9. You Chen," *Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System*", pp. 153 – 167, 2006.
10. Wun-Hwa Chen," *Application of SVM and ANN for intrusion detection*", pp. 2617–2634, ELSEVIER-2004.
11. Zhi-Song Pan,"*Hybrid Neural Network And C4.5 For Misuse Detection*", pp. 2463-2467, IEEE-2003.
12. Weiming Hu," *AdaBoost-Based Algorithm for Network Intrusion Detection*", Vol. 38,pp. 577-583,IEEE-2008.
13. Albert Hung-Ren Ko," *Single Classifier-based Multiple Classification Scheme for weak classifiers: An experimental Comparison*", pp.3603-3622, Elsevier-2008.
14. Emna Bahri," *A Multiple Classifier System Using an Adaptive Strategy for Intrusion Detection"*, pp.124-128, ICICS'2012.
15. Richa, Anurag jain," *Review: Boosting Classifiers For Intrusion Detection System,*" IJSER, Vol. 4, No. 7, July-2013, ISSN 2229-5518.
16. Priyanka, S. Dongre,"*Intrusion Detection through Ensemble Classification Approach*," 2011, pp.11-15.
17. A. Jain, S. Sharma,"*Network Intrusion Detection by using Supervised and Unsupervised Machine Learning Techniques: A Survey*," Vol. 1, pp. 14-20.
18. Koshal,M. Bag, "*Cascading of C4.5 Decision Tree and Support Vector Machine For Rule Based Intrusion Detection system*," Vol. 8,2012, pp. 8-20
19. Ali Borji,"*Combining Heterogeneous Classifiers for Network Intrusion Detection," Springer* – 2007, pp. 254-260.
20. M. Khalilian, A. Mamma, "*Intrusion Detection System with Data Mining"*, Vol. 11, 2011, pp. 29-34.
21. Ajith Abraham, "Modeling Intrusion Detection System using Hybrid Intelligent Sytems," Journal of Computer and network aplications, pp. 1084-8045 Elsevier.