Efficient Electricity Theft Detection Using Machine Learning Algorithms

1)	Hrishikesh Mohan Dabir	M.E.S. College of Engineering,Pune
2)	Aditya Suresh Kadam	M.E.S. College of Engineering,Pune
3)	Gaurav Hadge	M.E.S. College of Engineering,Pune
4)	Ayushman Singh Rathore	M.E.S. College of Engineering,Pune
5)	Prof. Shubhangi Ingale	M.E.S. College of Engineering,Pune

Abstract:- Electricity theft is one of the major problems of electric utilities. Such electricity theft produce financial loss to the utility companies. It is not possible to inspect manually such theft in large amount of data. For detecting electricity theft introduces a gradient boosting theft detector (GBTD) which utilizes three gradient based classifiers also known as (GBCs) which can be boosted that are extreme gradient boosting (XGBoost), categorical boosting (Cat Boost), and method as (LightGBM).XGBoost is one machine learning algorithm which gives high accuracy in less time.In this we apply preprocessing on smart meter data then does feature selection.Various application of the given GBTD is for electricity theft detection by reducing time taken to generate results of the GBTD model which detects nontechnical loss (NTL) detection.

Unlike other ML algoritms which are aimed towards manipulating the hyperparameters of classifiers,our GBTD is focused on lowering the time complexity and feature engineering based preprocessing.Our GBTD improves the factors such as and false positive rate(FPR) by generating features such as standard deviation, mean values of daily electricity consumed. Emphasis has been laid upon various application of the given GBTD for electricity theft detection by reducing time taken to generate results of the GBTD model.More importantly, this model proposes an revision version of the six theft cases to simulate real world theft patterns for numerical processing of the given algorithm.

Keywords:- *Artificial Intelligence(AI), Artifical Neural Network(ANN), XGBoost, CatBoost, Light GBM, electricity theft detection, gradient boosting,.*

I. INTRODUCTION

A. *Many electric utilities have financial loss due to electricity theft.* Here are various types of electrical power theft, including Tapping a line .One could easily bypass the electricity meter in the surrounding area. If we try to detect the theft manually then it's not possible as large amount of data will be there. So here we are applying machine learning algorithm to detect the theft. Theft can be detected by monitoring the various patterns or abnormalities in a person's electricity consumption period. From user fundamental data it is easy task to analyze user behavior. We take the advantage of various algorithms like XGBoost algorithm for nontechnical loss (NTL) detection.

B. The objective of this algorithm is to identify fradulent works in the consumption of electricity .Analyzing behavioral patterns from historical data of the user using the concept like machine learning (ML).We have to create a supervised and efficient ML-based model for theft detection of electricity in a particular area that detects an fraudulent usage pattern has been dealt in smart meter of that particular user.Various surveys have shown that various AI approaches which have been developed a computational model for detecting the theft.Efficient techniques like time series approaches,artificial neural networks, and support vector machines (SVM). The writers of [2] have investigated the usage of XGBoost for theft detection.

C. In this model we will compare the three mentioned gradient boosting algoritms, and to propose a model for electricity theft detection.

D. We can get hold of the historical electricity consumption information.We compensate in our model by manipulating the use of the customer electricity consumptiom on mathematical formulae. We keep in mind these factors and give an updated version of theft cases [1] to simulate theft patterns in real-world for practical evaluation or numerical comparison.

Various surveys have shown that different ML approaches that are used to develop a computational model for detection of the theft of a particular user.The objective of theft detection is to identify fraudulent activities in the electricity usage of a smart grid (SG) meter in a particular area .

Analyzing and monitoring the user's electicity usage behavior from historical data is the fundamental concept of a data science approach like AI. We have implemented a supervised ML- based model for theft detection that detects whether an abnormal/fraudulent usage pattern has produced in the smart meter.

II. NUMERICALEVALUATION ANDPROPOSED ALGORITHM USING OLD THEFT CASES

Classifiers consists of XGBoost, LightBoost and CatGBM also known as GBCs and the given method uses the following steps:

1) The given dataset [6] consists of 30 min reported usage of each user(in kWh) for approx. 420 days, of which 361 were used in the training set and 59 days in testing set.Vector of 48 features to describe the user's daily consumption. Figure 1: the daily usage of a consumer (*t* being feature (or sample) index).

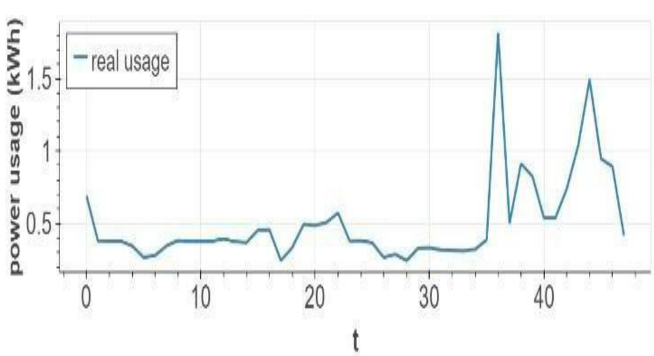


Fig. 1 Example of a single day usage

2) We used the same already existing six theft cases from reference [1] to produce our malicious/fraudulent samples of a user's comsumption.
3) The minority class, i.e., benign class, was oversampled to balance out the six theft cases generated using the synthetic minority oversampling technique (SMOT).
4) Now, we can train the model and compared it with the results to get the confusion matrix.

III. INTRODUCTION OF REVISEDTHEFT CASES PROPOSED ALGORITHM USING FEATUREENGINEERING

A. Generation of revised theft cases

Based on the given reference [1], we produced the updated version of existing and new six theft cases. The basic concept here is to generate more practical behavioral patterns and label them for the supervised ML algorithms. If the real usage of the consumer is xt, then the following are the revised theft patterns ($t \in [1,48]$):

1. $t1(xt) = xt * \text{rand}(0.1,0.9)$,
2. $t2(xt) = xt * rt$ ($rt = \text{rand}(0.1,1.0)$), 3. $t3(xt) = xt * \text{rand}[0,1]$,
4. $t4(xt) = \text{mean}(x) * \text{rand}(0.1,1.0)$, 5. 5( ) = ( ),
6. $t6(xt) = xT-t$, (T:sample size)

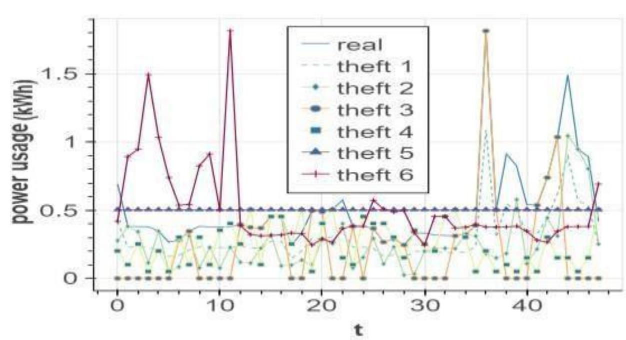


Fig. 2 Data generation and visualization

B. Evaluation of given algorithm along with newer theft cases

Here we used the new theft cases to generate malicious samples.For

The new theft cases, Table 1 confirms that in terms of average DR (94 to 97% for GBTD vs 88% for CPBETD) and FPR (5 to 7% for GBTD vs 15% for CPBETD), the former is more dominant than the latter. It also implies the new theft cases are harder to detect for the SVM-based CPBETD algorithm.

Old/New	CPBETD classifier	GBTD classifiers		
	SV M	XGBoost	CatBoos t	LightGB M
DR	94 /	95 / 94	96 / 97	96 /
(%)	88	7 / 6	8 / 5	97
FPR	11 /			7 / 7
(%)	15			

Table 1:- Average Detection Performance

USED (SEPARATED BY A SLASH '/')

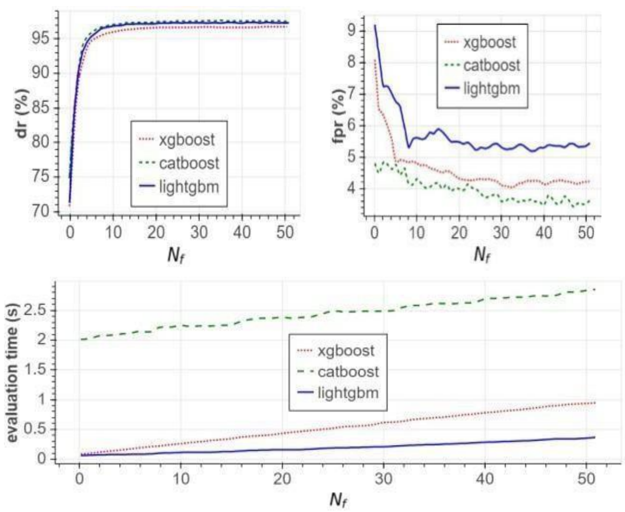


Fig. 3 DR, FPR and evaluation time vs the no. of features selected

IV. THE OLD AND THE NEW SYSTEMS

Existing System

Disadvantages

1. Require man power.
2. Getting theft chances are less.

PROPOSEDSYSTEM:

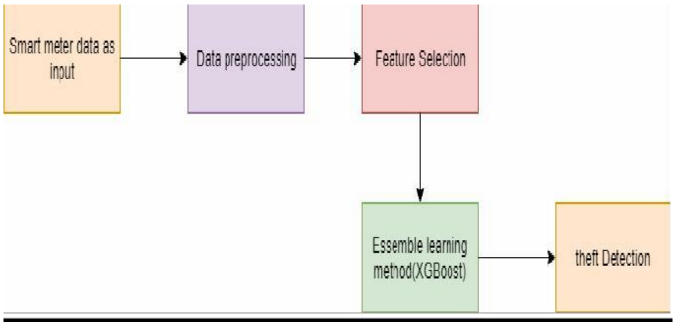
India ranks top in losing more money to electricity theft than any other country. The state of Maharashtra loses $2.8 billion per year. In this proposed system we use dataset having electricity consumption of a smart grid (SG) meter. Using this dataset we does feature selection and preprocessing on dataset. When we have large number of features in dataset then feature selection is very important part in our Machine Learning. As we use feature selection it gives us most important feature and this feature selection gives us more accuracy. Then we perform the preprocessing on that data. After that we take the advantage of given algorithm, for nontechnical loss (NTL) detection.

ADVANTAGES OF PROPOSED SYSTEM:

- We detect electricity theft to avoid financial loss of electric utilities.
- Required less time for this XGboost algorithm to detect electricity theft.
- Improve the accuracy.

In existing system utilities have to send their employee to check the smart meter of users and when employee goes to check and that time if he get any users meter is off then only they get the theft.

SYSTEM ARCHITECTURE



Explanation:

(a) SG meter data as input
(b) Next step is preprocessing on dataset.
(c) Feature selection method to select features.
(d) Essemble learning method(Xgboost) applied on data.
(e) Result theft detection.

V. LITERATURE SURVEY

1. Jokar, N. Arianpoo, V.C.M. Leung,[1] "Electricity theft detection in AMI using customers" consumption patterns", IEEE Trans. Smart Grid, 2016. As one of the major factors of the nontechnical losses (NTLs) in distribution networks, the electricity theft causes significant harm to power grids, which influences power supply quality and reduces operating profits. In order to help utility companies solve the problems of inefficient electricity inspection and irregular power consumption, a novel hybrid convolutional neural network-random forest (CNN-RF) model for automatic electricity theft detection is presented in this paper. In this model, a convolutional neural network (CNN) firstly is designed to learn the features between different hours of the day and different days from massive and varying smart meter data by the operations of convolution and downsampling. In addition, a dropout layer is added to retard the risk of overfitting, and the backpropagation algorithm is applied to update network parameters in the training phase. And then, the random forest (RF) is trained based on the obtained features to detect whether the consumer steals electricity. To build the RF in the hybrid model, the grid search algorithm is adopted to determine optimal parameters. Finally, experiments are conducted based on real energy consumption data, and the results show that the proposed detection model outperforms other methods in terms of accuracy and efficiency.

2. Patrick Glauner et al.[5] "Large-scale detection of non-technical losses in imbalanced data sets". In: Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society. IEEE. 2016:Non-technical losses (NTL) such as electricity theft cause significant harm to our economies, as in some countries they may range up to 40% of the total electricity distributed. Detecting NTLs requires costly on-site inspections. Accurate prediction of NTLs for customers using machine learning is therefore crucial. To date, related research largely ignore that the two classes of regular and non-regular customers are highly imbalanced, that NTL proportions may change and mostly consider small data sets, often not allowing to deploy the results in production. In this paper, we present a comprehensive approach to assess three NTL detection models for different NTL proportions in large real world data sets of 100Ks of customers: Boolean rules, fuzzy logic and Support Vector Machine. This work has resulted in appreciable results that are about to be deployed in a leading industry solution. We believe that the considerations and observations made in this contribution are necessary for future smart meter research in order to report their effectiveness on imbalanced and large real world data sets.

3. . T. Chen and C. Guestrin, "XGBoost [3] Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

4. S.S.S.R. Depuru, L. Wang, and V. Devabhaktuni,[2] "Electricity theft: overview,issues, prevention and a smart meter based approach to control theft,"Feb. 2011. Non-technical loss (NTL) during transmission

of electrical energy is a major problem in developing countries and it has been very difficult for the utility companies to detect and fight the people responsible for theft. Electricity theft forms a major chunk of NTL. These losses affect quality of supply, increase load on the generating station, and affect tariff imposed on genuine customers. This paper discusses the factors that influence the consumers to steal electricity. In view of these ill effects, various methods for detection and estimation of the theft are discussed. This paper proposes an architectural design of smart meter, external control station, harmonic generator, and filter circuit. Motivation of this work is to deject illegal consumers, and conserve and effectively utilize energy. As well, smart meters are designed to provide data of various parameters related to instantaneous power consumption. NTL in the distribution feeder is computed by external control station from the sending end information of the distribution feeder. If a considerable amount of NTL is detected, harmonic generator is operated at that feeder for introducing additional harmonic component for destroying appliances of the illegal consumers. For illustration, cost-benefit analysis for implementation of the proposed system in India is presented.

5. Patrick Glauner et al.[3] "The Challenge of Non-Technical Loss DetectionUsing Artificial Intelligence: A Survey". In: International Journal of Computational Intelligence Systems 10.1 (2017) Detection of non-technical losses (NTL) which include electricity theft, faulty meters or billing errors has attracted increase attention from researchers in electrical engineering and computer science. NTLs cause significant harm to the economy, as in some countries they may range up to 40 of the total electricity distributed. The predominant research direction is employing artificial intelligence to predict whether a customer causes NTL. This paper first provides an overview of how NTLs are defined and their impact on economies, which include loss of revenue and profit of electricity providers and decrease of the stability and reliability of electrical power grids. It then surveys the state-of-the-art research efforts in a up- to-date and comprehensive review of algorithms, features and data sets used. It finally identifies the key scientific and engineering challenges in NTL detection and suggests how they could be addressed in the future.

6. McLaughlin , B. Holbert, A. Fawaz, R. Berthier, and
S. Zonouz [6], "A multi-sensor energy theft detection framework for advanced metering infrastructures," IEEE J. Sel. Areas Commun Jul. 2013.The advanced metering infrastructure (AMI) is a crucial component of the smart grid, replacing traditional analog devices with computerized smart meters. Smart meters have not only allowed for efficient management of many end-users, but also have made AMI an attractive target for remote exploits and local physical tampering with the end goal of stealing energy. While smart meters posses multiple sensors and data sources that can indicate energy theft, in practice, the individual methods exhibit many false positives. In this paper, we present AMIDS, an AMI intrusion detection system that uses information fusion to combine the sensors and consumption data from a smart meter to more accurately detect energy theft. AMIDS combines meter audit logs of physical and cyber events with consumption data to more accurately model and detect theft-related behavior. Our experimental results on normal and anomalous load profiles show that AMIDS can identify energy theft efforts with high accuracy. Furthermore, AMIDS correctly identified legitimate load profile changes that more elementary analyses classified as malicious.

7. I. Guerrero , C. Leon, I. Monedero, F. Biscarri, and
J. Biscarri [9] , "Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection." 2014. The objective of the MIDAS project is the detection of Non-Technical Losses (NTLs) on power utilities. The NTLs represent the non-billed energy due to faults or illegal manipulations in clients ' facilities. Initially, research lines study the application of techniques of data mining and neural networks. After several researches, the studies are expanded to other research fields: expert systems, text mining, statistical techniques, pattern recognition, etc. These techniques have provided an automated system for detection of NTLs on company databases. This system is in test phase and it is applied in real cases in company databases.

8. R. Filho, E. M. Gontijo, A. C. Delaiba, E. Mazina, J. E. Cabral, and J. O. P. Pinto, [10] "Fraud Identification in Electricity Company Customers Using Decision Trees" in Proc. of 2004 IEEE International Conference on Systems, Man and Cybernetics,Oct. 2004. Electricity consumer dishonesty is a problem faced by all power utilities worldwide. Finding efficient measurements for detecting fraudulent electricity consumption has been an active research area in recent years. This thesis presents a new approach towards Non-Technical Loss (NTL) detection in power utilities using a combination of data mining and artificial intelligence (AI) based techniques, namely: Support Vector Machine (SVM) and the Fuzzy Inference System (FIS). The main motivation of this study is to assist Tenaga Nasional Berhad (TNB) in peninsular Malaysia to reduce its NTLs in the distribution sector. The intelligent system developed in this research study preselects suspicious customers to be inspected onsite by TNBD SEAL (Strike Enforcement Against Losses) teams for detection of fraud activities. This approach provides a method of data mining, which involves feature selection and extraction from historical customer consumption data. The Support Vector Classification (SVC) technique applied in this research study uses customer load profile information in order to expose abnormal behavior that is known to be highly correlated with NTL activities. The FIS is employed as a data postprocessing scheme, which uses knowledge of human expertise combined with the results from the SVC, in order to shortlist potential fraud suspects for onsite inspection. The proposed SVC and FIS model is trained using TNB Distribution's (TNBD's) historical kWh consumption data for the Kuala Lumpur (KL) Barat station, which is recorded with one of the highest rates of fraud activities in the state of Selangor in Malaysia. Model testing and validation is performed using customer data from three cities in the state of Kelantan in Malaysia. Feedback from TNBD for onsite inspection indicates that the fraud detection system developed is more effective as compared to the current actions taken by them. With the implementation of this new fraud detection system, TNBD's average hitrate for onsite customer inspection will become 40%, which increases their current inspection hitrate 35-37% from a mere 3-5%.

9. Soma Shekara Sreenadh Reddy Depuru, Lingfeng Wang, Vijay Devabhaktuni, and Robert C Green.
[11] High performance computing for detection of electricity theft. International Journal of Electrical Power & Energy Systems, 2013. Transmission and distribution of electricity involve technical as well as Non-Technical Losses (NTLs). Illegal consumption of electricity constitutes a major portion of the NTL at distribution feeder level. Considering the severity and devastating effects of the problem, illegal consumption of electricity has to be detected instantly in real-time. To this end, this paper investigates the possibility and role of High Performance Computing (HPC) algorithms in detection of illegal consumers. This paper designs and implements an encoding procedure to simplify and modify customer energy consumption data for quicker analysis without compromising the quality or uniqueness of the data. This paper parallelizes overall customer classification process. The parallelized algorithms have resulted in appreciable results as displayed in the results section of the paper.

10. Breno C Costa, Bruno LA Alberto, Andre M Portela,
´ W Maduro, and Esdras O Eler. [12] Fraud detection in electric power distribution networks using an annbased knowledge-discovery process. International Journal of Artificial Intelligence & Applications, 2013. Nowadays the electric utilities have to handle problems with the non-technical losses caused by frauds and thefts committed by some of their consumers. In order to minimize this, some methodologies have been created to perform the detection of consumers that might be fraudsters. In this context, the use of classification techniques can improve the hit rate of the fraud detection and increase the financial income. This paper proposes the use of the knowledge-discovery in databases process based on artificial neural networks applied to the classifying process of consumers to be inspected. An experiment performed in a Brazilian electric power distribution company indicated an improvement of over 50% of the proposed approach if compared to the previous methods used by that company.

11. AH Nizar, ZY Dong, and Y Wang[13]. Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems,2008. This paper presents a new approach to nontechnical loss (NTL) analysis for utilities using the modern computational technique extreme learning machine (ELM). Nontechnical losses represent a significant proportion of electricity losses in both developing and developed countries. The ELM-based approach presented here uses customer load-profile information to expose abnormal behavior that is known to be highly correlated with NTL activities. This

approach provides a method of data mining for this purpose, and it involves extracting patterns of customer behavior from historical kWh consumption data. The results yield classification classes that are used to reveal whether any significant behavior that emerges is due to irregularities in consumption. In this paper, ELM and online sequential-ELM (OS-ELM) algorithms are both used to achieve an improved classification performance and to increase accuracy of results. A comparison of this approach with other classification techniques, such as the support vector machine (SVM) algorithm, is also undertaken and the ELM performance and accuracy in NTL analysis is shown to be superior.

VI. CONCLUSION:

A. This proposed system detects the electricity theft using xgboost machine learning method.LightGBM can be considered as the fastest classifier. This proposed system helps to electricity utilities to detect electricity theft and they will not have to bare loss. This is most important application of this project.

B. This paper presented a SG(smart grid)meter for theft detection algorithm, called GBTD. LightGBM and CATBoost where efficient in terms of DR.We have also show that GBTD with feature engineering module lowers FPR rate and helps to improve time-space complexity.

REFERENCES

[1]. P. Jokar, N. Arianpoo, V.C.M. Leung, "Electricity theft detection in AMI using customers‟ consumption patterns", *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216-226, 2016.

[2]. M. Buzau, J. Aguilera, P. Romero, and A. Expósito, “Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning,” IEEE Trans. Smart Grid, Feb. 2018. [DOI: 10.1109/TSG.2018.2807925]

[3]. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, San Francisco, CA, USA [DOI 10.1145/2939672.2939785].

[4]. Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support,” Workshop on ML Systems at Neural Information Processing Systems (NIPS), 2017. 5. G. Ke, Qi Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." 31st Conference on Neural Information Processing Systems (NIPS), 2017

[5]. Patrick Glauner et al. “Large-scale detection of non-technical losses in imbalanced data sets”. In: Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society. IEEE. 2016, pp. 1–5

[6].]S.S.S.R. Depuru, L. Wang, and V. Devabhaktuni, “Electricity theft: overview,issues, prevention and a smart meter based approach to control theft,” Energy Policy, vol. 39, pp. 1007–1015, Feb. 2011.

[7]. Patrick Glauner et al. “The Challenge of Non- Technical Loss DetectionUsing Artificial Intelligence: A Survey”. In: International Journal of Computational Intelligence Systems 10.1 (2017), pp. 760–775

[8]. S. McLaughlin , B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz , “A multi-sensor energy theft detection framework for advanced metering infrastructures,” IEEE J. Sel. Areas Commun., vol. 31, no. 7, pp. 1319–1330, Jul. 2013.

[9]. J. I. Guerrero , C. Leon, I. Monedero, F. Biscarri, and J. Biscarri , “Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection.” Knowl.- Based Syst. , vol. 71, pp. 376–388, 2014.

[10]. J. R. Filho, E. M. Gontijo, A. C. Delaiba, E. Mazina, J. E. Cabral, and J. O. P. Pinto, "Fraud Identification in Electricity Company Customers Using Decision Trees" in Proc. of 2004 IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, pp. 3730-3734, Oct. 2004.