IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

Analyzing Network Security and Anomaly Detection using Ensemble Learning Techniques

Pruthvi Adekar, Pushpa Tandekar, Ashish Deharkar

Student, Department of Computer Science & Engineering, Shri Sai College of Engineering & Technology, Chandrapur, India

Assistant Professor, Department of Computer Science & Engineering, Shri Sai College of Engineering & Technology, Chandrapur, India

Assistant Professor, Department of Computer Science & Engineering, Shri Sai College of Engineering & Technology, Chandrapur, India

ABSTRACT: The operation of machine literacy models to network security and anomaly discovery problems has largely increased in the last decade; still, there's still no clear best- practice or tableware pellet approach to address these problems in a general environment. While deep- literacy is moment a major advance in other disciplines, it is di cult to say which is the stylish model or order of models to address the discovery of anomalous events in functional networks. We present an implicit result to all this gap, exploring the operation of ensemble literacy models to network security and anomaly discovery. We probe different ensemble- learning approaches to enhance the discovery of attacks and anomalies in network measures, following a particularly promising model known as the Super Learner. The Super Learner performs asymptotically as well as the stylish possible weighted combination of the base learners, furnishing a veritably important approach to attack multiple problems with the same fashion. We test the proposed result for two different problems, using the well- known MAWILab dataset for discovery of network attacks, and a semi-synthetic dataset for discovery of trac anomalies in functional cellular networks. Results con € rm that the Super Learner provides better results than any of the single models, opening the door for a conception of a best- practice fashion for these specific disciplines.

KEYWORDS: Ensemble Learning, Network Security, Anomaly Detection.

I. INTRODUCTION

Network security and anomaly discovery represent both a cornerstone to ISPs, who need to manage with an adding number of unanticipated events that put the network's performance and integrity at threat. The high- dimensionality of network data handed by current network monitoring systems opens the door to the massive operation of machine literacy approaches to ameliorate the discovery and classification of anomalous events. still, opting the stylish machine literacy model for a specific problem is a complex task- it is generally accepted that there's no tableware pellet for addressing problems contemporaneously. Indeed, indeed if multiple models could be veritably well suited for a particular problem, it may be veritably difficult to one which performs optimally for different data distributions and statistical composites. The ensemble literacy proposition permits to combine multiple models to form a (hopefully) better one. Ensemble styles use multiple literacy algorithms to gain better prophetic performance than could be attained from any of the constituent literacy algorithms alone. In principle, if no single model covers the true vaticination behind the data, an ensemble can give a better approximation of that mystic, true vaticination model. In addition, an ensemble of models exhibits advanced robustness with respect to misgivings in training data, which is largely beneficial. Ensemble literacy has in principle an advanced computational cost and complexity than single grounded literacy approaches. nonetheless, current big data platforms and of- the- shelf data processing technology is mature enough to allow a fast and resemblant operation of multiple algorithms (2), easing this constraint. In this paper we concoct a new discovery fashion for network security and anomaly discovery using the Super Learner ensemble literacy model (1). The Super Learner is a supervised literacy system that finds the optimal combination of a collection of base vaticination algorithms. The Super Learner performs asymptotically as well as the stylish possible weighted combination of the base learners, furnishing a veritably important approach to attack multiple problems with the same fashion. In addition, its denes an approach to minimize overfitting liability during training, using a variant of cross-validation. The proposed result is estimated on two different scripts and using five different ensemble combination algorithms for the Super Learner, using the well- known MAWILab dataset for discovery of network attacks, and a semi-synthetic dataset for

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

discovery of traffic anomalies in functional cellular networks. Evaluations con € rm that ensemble ways have the capability to perform as well as the stylish available base literacy model, achieving indeed better results in utmost scripts and using different combination approaches. To the stylish of our knowledge, this paper is the first attempt to apply the Super Learner approach to network anomaly discovery problems, aiming at discovering a new order of machine literacy models which could be applied in a more methodical fashion to networking problems. We believe that this study would enable a broader operation of ensemble literacy approaches to network security and anomaly discovery, with veritably promising results. 2 brie y reviews the affiliated work.Sec. 3 describes the main generalities behind the Super Learner ensemble approach, and presents the different literacy models used in the study, both at the individual position and at the ensemble position. Sec. 4 reports benchmarking results for the proposed ensemble literacy approaches, comparing their performance to that achieved by the individual models in the discovery of both network attacks and network traffic anomalies in two distinct datasets. Eventually, Sec. 5 concludes this work.

II. METHODOLOGY

There are a couple of expansive checks on general sphere anomaly discovery ways (11) as well as network anomaly discovery (12,), including machine literacy- grounded approaches. The operation of literacy ways to the problems of network security and anomaly discovery is largely extended in the literature. There is a particularly expansive literature in the operation of learning-based approaches for automatic traffic analysis and classification. We relate the interested anthology to (10) for a detailed check on the different ML ways applied to automatic traffic classification. The specific operation of ensemble literacy approaches to network security and anomaly discovery is by far more limited, and indeed if it's generally observed in the practice that ensembles tend to yield better results when there's a significant diversity among the models, only many papers have applied them to network security (15) and network anomaly discovery (14).

III. ENSEMBLE LEARNING

In the environment of supervised literacy there are several styles to train algorithms with available data to use them for vaticination purposes. The performance of a particular algorithm or predictor depends on how well it can assimilate the being information to approximate the mystic predictor, i.e., the ideal optimal predictor defined by the true data distribution. still, knowing a priori which algorithm will be the stylish suited for a given problem is nearly insolvable in practice. One could say that each algorithm learns a different set of aspects of reality from the training datasets, and also their separate vaticination capability also differs between problems. According to(7), single thesis algorithms or simple literacy models may suffer from three different backups statistical problem- arises when the space of thesis is too large for the quantum of available training data, performing in several algorithms with analogous delicacy and threat of choosing one that won't prognosticate unborn data points well; computational problem- several algorithms aren't guaranteed to find the global optimum; representation problem- results from the thesis space not containing any model that's a good approximations of the true distribution. Rather than finding the stylish model to explain the data, ensemble styles construct a set of models and also decide between them with some combinatorial approach, seeking complementarity in the sense that the literacy limitations of each predictor compensate for the others. therefore, the prosecution of several of these algorithms in resemblant provides diversity of prognostications. Several papers have studied styles exploiting this diversity to enhance the overall vaticination capability by combining the labors of multiple algorithms (5, 7). Basically, this is made by using a scheme known as Ensemble Learning that uses the affair of the predictors as inputs for a new algorithm that receives the name of alternate position or meta learner. A notable illustration of this approach is the well-known Random Forest algorithm. Classical ensemble literacy approaches include bagging, boosting, and mounding (7). General ensemble learning approaches might be prone to overfitting the data. In (1) a simple ensemble learning algorithm named Super Learner is proposed as a possible result for this overfitting limitation. It proposes a system to minimize the overfitting liability using a variant of cross-validation. In addition, the Super Learner provides performance bounds, as it performs asymptotically as good as the stylish available single thesis predictor. The Super Learner algorithm makes aggressive use of cross confirmation the available labeled dataset conforming of n samples is resolve in K roughly equal sets. As usual, each of these sets is used as a confirmation set, while its complement, $K - 1$ sets are used as the training set. For each split, the J first position learners are fitted with the training dataset and also do prognostications for the samples of the confirmation set. By incorporating the prognostications done for every pack we gain a new dataset Z of size $n \times J$, containing the prognostications done by each first position learner for every sample in the disjoint confirmation sets. This new dataset Z is used as design input matrix to train the meta learner algorithm, which will also be used to perform the final prognostications. In the original paper (1), the meta learner can be arbitrarily complex, yet a simple direct retrogression

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

model is used for the presented retrogression script. The paper presents a formal proof caching that this Super Learner is optimal in the sense that it can perform at least asymptotically as well as the stylish first position learner available.

3.1. Binary Classification:

The sense expressed in (1) can be acclimated for use on double classification problems similar as the bone we attack in this paper. Basically, suppose there are ni.i.d. compliances(Xi yi) ∼ P0 with i = 1,..., n that induce empirical probability distributions Pn, and that the thing is to estimate the classification function ψ0 similar that ψ0(X) = arg min ψ ∈ Ψ E(L(y, ψ(X)))(1) where L(y, ψ(X)) is a given loss function that measures the distinction between vaticination and real value-e.g., square loss in(1), for all possible point vector X ∈ X and its corresponding marker y ∈Y. ψ0 is also a mapping function from the point space into the marker space and Ψ the parameter space of all possible functions similar that X → Y. Now let {ψˆ j} j = 1, J be the collection of first position learners, which represent mappings from the empirical distribution Pn into parameter space Ψ.

IV. EVALUATION AND DISCUSSION

In this section we show that the Super Learner approach can end hance the results attained on the double classification problems studied in (3) and (4) for discovery of network attacks and anomalies independently. In those studies, different standard first- position literacy models are used. An exception to this is the operation of Random timbers RF) (10), which actually represent an ensemble literacy approach, by combining the affair of multiple decision trees through plain maturity voting. In a nutshell, each decision tree of a RF is erected on different subsets of input features, aimlessly named. The RF algorithm is more sophisticated than the Super Learner itself, in the sense that it formerly performs point selection during training.

Algorithm	DDoS	mptp-la	netscan-ACK	netscan-UDP	ping-flood
Decision Tree	0.715	0.864	0.921	0.931	0.928
Naive Bayes	0.752	0.654	0.879	0.937	0.903
Neural Net	0.914	0.993	0.968	0.985	0.989
SVM	0.897	0.998	0.941	0.996	0.967
kNN	0.838	0.918	0.950	0.955	0.952
Random Forest	0.820	0.915	0.946	0.919	0.931
logreg	0.927	0.999	0.964	0.997	0.990
MVaccuracy	0.929	0.993	0.966	0.992	0.990
MVexp	0.930	0.996	0.971	0.997	0.993
MVuniforme	0.925	0.993	0.965	0.993	0.990
CART	0.878	0.983	0.947	0.982	0.976

Table 1: ROC AUC on XYZ Dataset

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

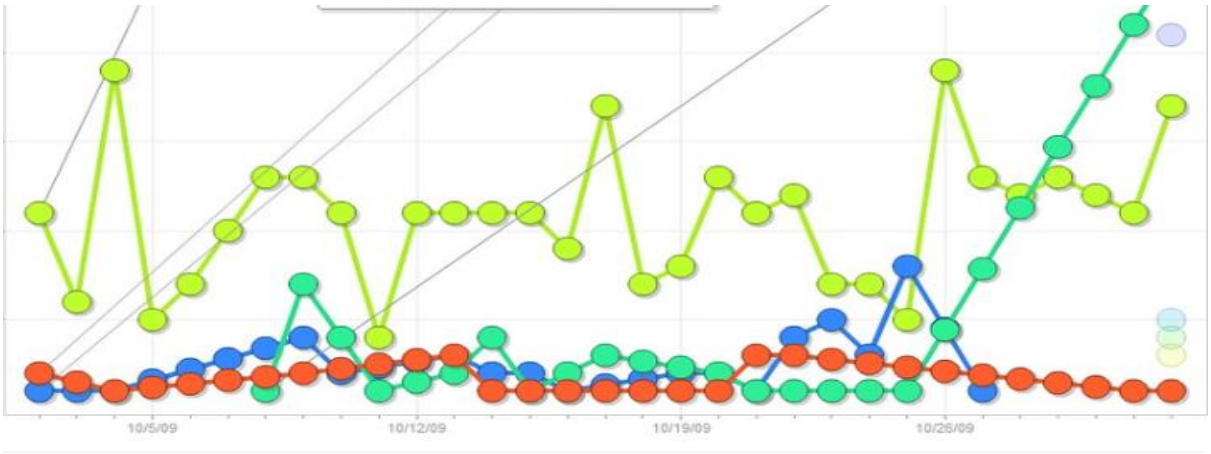

Figure 1: Detection performance per type of attack on the MAWI dataset. Except for the CART-based Super Learner, all Super Learners generally outperform base predictors.

We thus compare the performance achieved by each single, first- position sensor against that achieved by the proposed Super Learners. To have similar results to (3) and (4), we also add a RF- grounded sensor, trained on the same training set as the rest of the models, and having as numerous internal decision trees as first position learners has the Super Learner; i.e., 5 in this paper. Comparison is performed on the base of true positive vs false alarm rates (TPR/ FPR independently) through ROC angles, as well as by calculating the area under these ROC angles (AUC).

4.1 Data Description:

Two different datasets are used to test the performance of the proposed approaches the MAWI dataset for network security (9), and a semi-synthetic dataset for traffic anomalies in cellular networks, conceived in (4). Each dataset is resolve in two sets a training set, with roughly 20 of the samples, and a testing set with the remainder 80. We perform the literacy procedures on the training set and also estimate the performance of the predictors on the testing set using ROC angles and AUC values.

4.1.1 MAWI Dataset:

MAWI is a public collection of 15- nanosecond real network traffic traces captured every day on a backbone link between Japan and the US since 2001. structure on this depository, the MAWILab design uses a combination of four traditional anomaly sensors to incompletely label the collected traffic (9). From the labeled anomalies and attacks, we concentrate on a specific group which are detected contemporaneously as “anomalous” by the four MAWILab sensors to achieve a high quality on the attained markers. We consider five types of attacks anomalies in particular(i) DDoS attacks DDoS), (ii) HTTP ash crowds (mptp- la), (iii) Flooding Attacks (Ping Flood), (iv) UDP and(v) TCP probing traffic. The considered algorithms were trained to descry each of these attack types independently and in parallel, in the same fashion and using the same features as (3). As a result, each discovery approach can descry the circumstance of an attack and also classify its nature. The dataset spans a full week of MAWILab traffic traces collected in late 2015; traces are resolve in successive time places of one second each, and a high- dimensional set of 245 features describing the traffic in each of these places is used, see (3) for further details

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

4.2 Discovery of Network Attacks:

Fig. 1 depicts the ROC angles attained by each sensor and for each attack type, and Tab. 1 reports the corresponding AUC values.

Table 2: ROC AUC on semi-synthetic dataset.

Algorithm	E1	E2	E3
Decision Tree	0.988	0.879	0.992
Naive Bayes	0.994	0.862	0.990
Neural Network	0.997	0.949	0.997
SVM	0.998	0.943	0.996
k-Nearest Neigh.	0.992	0.865	0.961
Random Forest	0.998	0.882	0.999
Logistic Regression	0.997	0.953	0.997
Majority Voting (Accuracy)	0.998	0.945	0.997
Majority Voting (Exponential)	0.998	0.951	0.997
Majority Voting (Uniform)	0.998	0.942	0.997
CART (Classification and Regression Trees)	0.995	0.922	0.993

Besides many exceptions, utmost Super Learner strategies achieve better TPRs at the same FPR than both most first position learners and the Random Forest algorithm. Indeed, except for the wain- grounded Super Learner, all Super Learners generally outperform base predictors. Majority Voting with both exponential or direct delicacy grounded weights and the Logistic retrogression Super Learner strategies actually achieve the stylish results in all attack types; in particular, the MV exp Super Learner performs the stylish for all attack types, calling for a presumptive generalizable and particularly € t model for this testing script. Still, it's worth mentioning that, analogous to (3), performance is rather poor on the discovery of DDoS attacks; the stylish algorithms descry about 80 of the attacks with a FPR of 10. Exactly the contrary happens with the HTTP ash crowds, where all Super Learners and some of the base predictors are above 95 discovery rates with a FPR below 5. The benefits of the Super Learner approach are less apparent in this case, as there isn't important room left for enhancement, it’s worth mentioning that the Random timber sensor no way achieves an AUC score advanced than any of the Super Learner strategies.

4.3 Detection of Network Anomalies:

Fig. 2 depicts the ROC angles attained by each sensor for each anomaly type, and Tab. 2 reports the corresponding AUC values. Results are resolve by First Level Learners and Super Learners in figures. analogous to (4), achieved results for anomalies of type E1 and E3 differs fully from E2. Every predictor achieves an AUC over 99 and discovery rates above 97 at a FPR of 5 for E1 anomalies. therefore, there's little room for enhancement, which leads to only veritably subtle differences between the performances of Super Learners and base learners. analogous compliances can be drawn from the discovery of E3 anomalies. The attempt to descry E2 anomalies shows relatively different results. Not only all predictors performed fairly poor the stylish bones achieve nearly 80 TPR at a FPR of 5, but also numerous of them achieve veritably low performance; e.g., the ensemble algorithm Random Forest gets a TPR just above 60 for a 5 FPR and only 80 TPR for a high FPR of 20, easily worse than any other ensemble fashion. Still, this script is the one that better highlights the advantages of the Super Learner approach for anomaly discovery, as all the Super Learners- anticipate from CART, outperform the first position learners.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

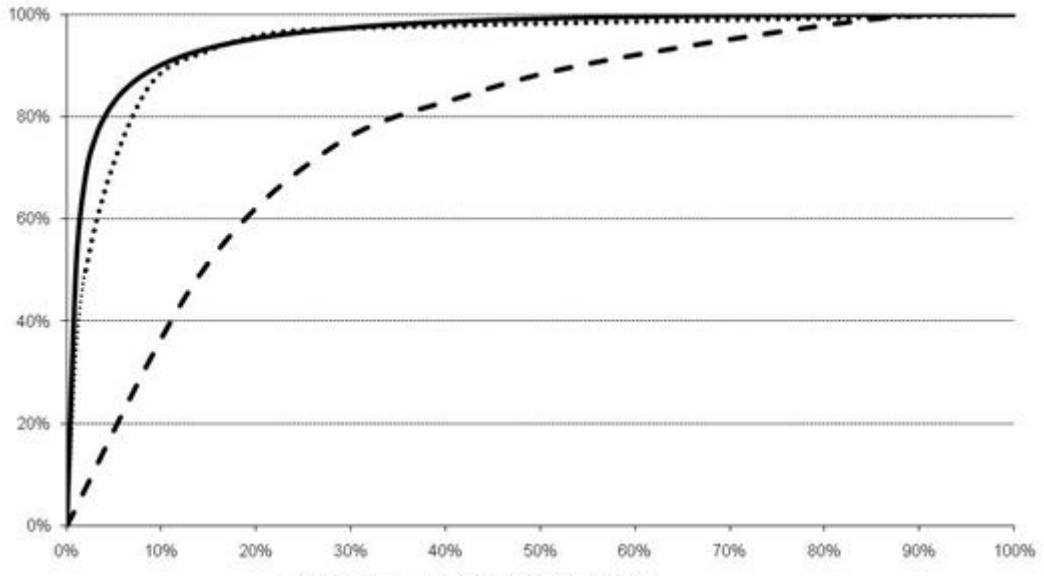

Figure 2: Detection Performance per type of anomaly on the semi-synthetic dataset

V. CONCLUSION

The advantages of ensemble literacy ways for discovery of attacks and anomalies, and particularly of the Super Learner approach, could be confirmed through evaluation. Not only we set up that Super Learner grounded predictors have the capability to perform as well as the stylish available first position learner, but frequently achieve better results. The performance advancements are advanced in scripts where the performance of the first position predictors were fairly downward; when first learners' performance is formerly high, there's not enough room for enhancement and the fresh computational cost of the Super Learner may not be justified. The different evaluated Super Learner schemes achieved veritably analogous performances. still, the MV exp Super Learner performs the stylish for all attack and anomaly types in both datasets, suggesting a potentially good approach to go for by dereliction in analogous double classification problems. The simplicity and veritably low computational costs of MV Exp maturity voting makes also a veritably nice case for similar type of models. We believe that this study would enable a broader operation of ensemble literacy approaches to network security and anomaly discovery, with veritably promising results.

REFERENCES

[1] M. Van der Laan, E. C. Polley and A. E. Hubbard, "Super learner", in Statistical applications in genetics and molecular biology, vol. 6 (1), pp. 1-21, 2007.
[2] P. Casas, A. D'Alconzo, T. Zseby and M. Mellia, "Big-DAMA: Big Data Analytics for Network Tra"c Monitoring and Analysis", in ACM SIGCOMM LANCOMM Workshop, 2016.
[3] P. Casas, A. D'Alconzo, G. Settanni, P. Fiadino and F. Skopik, "POSTER:(Semi)- Supervised Machine Learning Approaches for Network Security in High Dimensional Network Data", in ACM CCS, 2016.
[4] P. Casas, P. Fiadino and A. D'Alconzo, "Machine-learning based approaches for anomaly detection and classi cation in cellular networks", in TMA, 2016.
[5] Y. Freund, R. E. Schapire, Y. Singer and M. K. Warmuth, "Using and combining predictors that specialize", in ACM STOC, 1997.
[6] J. Hansen, "Combining predictors: Some old methods and a new method", available online at Citeseer, 1998.
[7] T. Dietterich, "Ensemble learning", in The handbook of brain theory and neural networks, vol. 2, pp. 110-125, MIT Press, 2002.
[8] P. Sollich and A. Krogh, "Learning with ensembles: How over tting can be useful", in Advances in neural information processing systems, pp. 190-196, 1996.
[9] R. Fontugne, P. Borgnat, P. Abry and K. Fukuda, "MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking", in ACM CoNEXT, 2010

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305499 |

[10] T. T. T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning", in IEEE Comm. Surv. & Tut., vol. 10 (4), pp. 56-76, 2008.
[11] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey", in ACM Comput. Surv., vol. 41 (3), pp. 1-58, 2009.
[12] M. Ahmed, A. Naser Mahmood and J. Hu, "A Survey of Network Anomaly Detection Techniques", in J. Netw. Comput. Appl., vol. 60, pp. 19-31, 2016.
[13] W. Zhang, Q. Yang and Y. Geng, "A Survey of Anomaly Detection Methods in Networks", in CNMT Symposium, 2009.
[14] R. Ravinder Reddy, Y. Ramadevi and K. V. N. Sunitha, "Real Time Anomaly Detection Using Ensembles", in ICISA International Conference, 2014.
[15] M. Ozdemir and I. Sogukpinar, "An Android Malware Detection Architecture based on Ensemble Learning", in Trans. on Machine Learning and Arti cial Intelligence, vol. 2 (3), pp. 90-106, 2(3), pp. 90-106, 2014.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details