IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 **UGC CARE Listed ( Group -I) Journal Volume 07, Iss 2, 2018**

# Intrusion system detection system using Decision tree compared to Linear Regression

**Chandrashekara A C**
Assistant Professor of Mathematics, Maharani's Science College for Women, (Affiliated Mysore University) J.L.B Road, Mysore, Karnataka- 570005, India

**Abstract:**

**Aim:** This work offers and compares two machine learning strategies for developing an efficient intrusion detection system, namely enhanced Novel Decision Tree and Linear Regression. **Materials and Methods:** Utilizing Intrusion Detection in conjunction with Novel Decision Trees and Linear Regression allows for the identification of network intrusions. With 20,000 records used for training and 8,000 records for testing, the total number of records utilized in this study is 28,000. By adjusting the Gpower settings to 0.05 and 0.85, the test achieves an average Gpower of almost 85%. **Result:** The two methods that were investigated for intrusion detection have a statistically significant difference, as shown by a significance value of 0.001 ($p<0.05$). When comparing accuracy, Novel Decision Tree (94.90%) performs better than Linear Regression (93.72%). The Linear Regression Classifier is among the most popular and easy-to-understand classification algorithms used today. The system's incapacity to react or halt assaults upon discovery is one of its flaws. **Conclusion:** The novel decision tree is more accurate than Linear Regression.

**Keywords:** Anaconda Navigator, Decision tree, Intrusion detection system, Linear Regression,Statistical analysis

## Introduction :

An Intrusion Detection System (IDS) is a network monitoring system that detects and warns potential security breaches by analyzing network traffic for any abnormal or suspicious behavior. It is an application that allows network or system scans to detect malicious activities or violations of policies(Saranya, S, Sridevi, and C, Deisy 2018). Typically, any harmful activity or breach is reported to an administrator or collected centrally via a security information and event management (SIEM) system. Security information and event management (SIEM) systems combine data from many sources and use alert filtering methods to distinguish between legitimate and malicious activities .There is a risk of false alarms with intrusion detection systems, despite their best efforts to monitor networks for hostile activities(Iwendi et al. 2018). Consequently, after first installing their IDS solutions, businesses need to adjust them to their own needs. It comes down to making sure that the intrusion detection devices are configured correctly so they can distinguish between legitimate network traffic and malicious activities

 

(Besharati, Naderan, and Namjoo 2018).In order to detect malicious activity and immediately provide warning signals, systems to prevent intrusions also monitor incoming network packets.An intrusion detection system's primary function is to analyze the frequency and nature of criminal activity.

The use of intrusion detection systems (IDS) in networks has been the subject of an average of 374 academic publications published in IEEE Xplore and 56 articles published in sciencedirect. An effective intrusion detection solution for networks that use ensemble-based ML.The researcher has studied how to improve anomaly detection by using the grey wolf optimization (GWO) technique. Experimental results showed that the approach has high accuracy for DoS (93.64%), Probe (91.01%), R2L (57.72%), and U2R (53.7%) (Alamiedy et al. 2018).A novel intrusion detection system (IDS) approach based on choosing features and grouping utilizing a decision tree algorithm has been suggested in a study. Findings reveal a detection accuracy of 95.03% (Mohammadi et al. 2018).An improved model for deep neural network-based computer network intrusion detection that accounts for connection quality across different network systems is the goal of this research. What makes an IDS unique are the characteristics of invasions as well as typical user activities. After that, the ideal measurement parameters are found by exploring the whole measurement space using an immunological genetic algorithm(Zhang and Bao n.d.).In contrast, a different study surveyed researchers on intrusion detection systems that used unsupervised and hybrid approaches.The findings demonstrate that modern IDS have progressed beyond basic detection to include correlation and attribution(Magán-Carrión et al. 2017). Accordingly, sophisticated data analytics tools allow the perpetrator to be identified.In addition, the study suggests adding three additional types of attacks targeting outbound network traffic to the current attack classifications. However, I am confused about the meaning of the three new classifications. The dataset and methodologies utilized determine the outcomes of each approach performed in IDS (Thakallapelli, Ghosh, and Kamalasadan 2016).

An area requiring investigation is the current approach due of its inaccuracy. Improving classification accuracy is the goal of this study, which contrasts Linear Regression with Novel Decision trees. The suggested paradigm brings an improvement to IDS. In order to identify intrusions in computer networks, this study aims to develop a novel deep neural network approach that takes into account the connection quality of various network systems. Based on their usual actions and the traits of invasions, a user may set up an intrusion detection system. Compared to Linear Regression, the Novel Decision Tree has higher accuracy.

## MATERIALS AND METHODS

The research was conducted at the Saveetha School of Engineering's User Interface Design Lab at the Saveetha Institute of Medical and Technological Sciences. The Gpower program determined the sample size by comparing the two controllers. We choose two groups at random to compare their methods and outcomes. Ten samples were selected for this inquiry, with ten



 

pairs of samples taken from each group. Two algorithms—Newton's Decision Tree and Linear Regression—are applied using technical analysis software. The sample size was determined using the GPower 3.1 tool, with 10 participants in each group (Gpower setup parameters: =0.05 and power=0.85). With 20,000 data used for training and 8,000 records for testing, the total number of records utilized in this study is 28,000. The sample size was determined using ClinCalc.com to be 10 individuals(Chen, Li, and Li 2018).

Python OpenCV is used for both planning and executing the intended task. For the purpose of testing deep learning, Windows 10 was used. An 8 GB RAM and an Intel Core i5 CPU were the components of the hardware setup. We used a 64-bit mechanism to sort the data. The code was implemented utilizing the Java platform. While the function is running, the dataset is handled discreetly to ensure an accurate output. The IDS's ML and DL dataset is obtained from www.kaggle.com. With a significance threshold of 0.048 ($p<0.05$), an average size of 10 was determined using ClinCalc.com.

**Decision Tree**

To run the Decision Tree algorithm, you must first construct a structure that looks like a tree. A simple tree structure with nodes, branches, and leaves is the basic building element of an algorithm. Following direction from the root node—representing the first feature or choice—the data is divided by internal decision nodes according to certain criteria. A Decision Tree further splits the data by constructing child nodes. The result might be a numerical value or a label for a categorization. The last stage of a tree's life cycle is the production of the leaves that ultimately bear the fruit.One of the most important parts of making Decision Trees was picking the right splitting criterion. The method determines the most informative feature or characteristic to use at each decision node before dividing the data. For classification jobs, this choice is often made using evaluation metrics like Gini impurity and data gain, whereas for regression problems, mean squared error is employed. Decision Trees excel at capturing intricate decision boundaries via iteratively choosing the best splitting criteria.

Step 1: Ascertain the root of the tree

Step 2: Compute the entropy for each class

Step 3: Determine the entropy after splitting for each variable

Step 4: Identifying the data gain for each split

Step 5: Execute the Split

Step 6: Form the Decision Tree

 

**Linear Regression**

Implementing a linear equation to actual data is what linear regression is all about when trying to predict the connection between two variables. An independent variable is one that does not rely on any other variables; the opposite kind is known as a dependent variable. For forecasting purposes, linear regression is often used. Looking at two things is the key to regression. The first question to ask is whether a given collection of predictor variables accurately forecasts the dependent variable. Second, we need to identify the factors that significantly impact the result variable.A correlation coefficient is a statistical measure of the degree of association between any two variables. The coefficient may take on values between -1 and +1. The correlation coefficient indicates the degree to which two variables are associated with the observed data.One reasonable way to explore and evaluate the connection between two continuous variables is to use a regression line to depict the behavior of the data. This is subsequently used in quantitative applications such as ML models, analysis of mathematics, the statistics area, and forecasting.

Step 1: Analyzing and comprehending the information

Step 2: Data visualization (Experimental Data Analysis)

Step 3: Gathering Information

Step 4: Creating two sets of data: one for training and one for testing

Step 5: Creating a straight line model

Step 6: Evaluation of the remaining train data

Step 7: Forecasting future outcomes based on the refined model and assessment

**Statistical Analysis**

Linear regression and decision trees may be statistically studied using SPSS. Image, items, distance, frequency, modulation, loudness, and dB are all variables that may be thought of as standing on their own. They depend on things, which are variables. We use an independent T test to determine the validity of each hypothesis (Chittora et al. 2017).

**RESULTS**

Anaconda Navigator was used to run the suggested Novel Decision tree and Linear Regression on a sample of 10 individuals. Table 1 displays the estimated loss and accuracy of the Novel Decision tree. Table 2 displays the expected loss and accuracy of the linear regression. We may compare the statistical results and loss values obtained from these ten data samples for each approach. The results of comparing the two approaches revealed that the Logistic Regression



 

technique had an average accuracy of 93.72% and the Novel Decision Tree method of 94.90%. Table 3 displays the mean accuracy ratings for both Linear Regression and Novel Decision Tree. On average, Novel Decision Tree performs better than Linear Regression, with a standard deviation of 0.5485 as opposed to 0.61818. Table 4 displays the findings of the independent sample T test with a value of 0.001 ($p<0.05$) for both the Novel Decision tree and Linear Regression models. Table 5 presents a comparison between Linear Regression and Novel Decision Tree in terms of accuracy. At 94.90%, Novel Decision Tree attains a greater accuracy level than Linear Regression. When it comes to forecasting intrusion detection, the Novel Decision Tree strategy fared better than the Linear Regression one. Figure 1 displays the average accuracy and loss for both Novel Decision Trees and Linear Regression.

For the Novel Decision tree, the corresponding mean, standard deviation, and standard error are 94.9020, 0.54857, and 0.17347. In addition, the linear regression mean, standard deviation, and standard error are, respectively, 94.72, 0.61818, and 0.19548. On the other hand, 5.0980 at the mean, 0.54857 at the standard deviation, and 0.17347 at the standard error mean are the loss values for the Novel Decision tree. For Linear Regression, the values of mean, standard deviation, and standard error mean loss are 6.280, 0.61818, and 0.19548, correspondingly.

The average, standard deviation, standard error mean, and collective statistics value are given for each strategy. Using Linear Regression and Novel Decision Tree, the two approaches' loss means are visually compared and categorized. Thus, Novel Decision Tree performs much better than Linear Regression, which only achieves 93.72% accuracy rate, with a rate of 94.90%.

## DISCUSSION

This paper suggests an effective IDS by choosing significant futures from the data set generated by the NSL-KDD using a classification system derived from Novel Decision Tree (DT). A feature selection strategy is necessary to reduce the amount of time spent learning and the amount of memory needed. Using selective multinomial Naive Bayes and filtering experiments, this paper offers a network IDS. The NSL-KDD dataset, which is a revised edition of the KDDCup 1999 malware detection benchmark dataset, was used to develop our suggested approach. Ten-fold cross-validation and two class classifiers are included in the dataset. Testing shows that the suggested approach reduces the amount of time required to build a network intrusion detection system (IDS) that is efficient and has a low positive error rate (Besharati, Naderan, and Namjoo 2018). Deep learning is employed in many safety-critical applications because to its rapid growth and success. Well-designed adversarial input samples damage deep neural networks. Human-invisible adversarial perturbations confuse DNNs during testing/deploying. Adversarial examples may affect safety-critical DNNs. Oppositional examples are popular for attacks and countermeasures. We review DNN adversarial instance discoveries, generation methods, and classification in this work(Yuan et al. n.d.).

 

An analysis of research studies indicates that the LR model may be used to intrusion detection problems with complexity, highly correlated features, and large amounts of network data stream. By adopting the LR data mining model, the conditional independence assumption of the LR technique is loosened.(Chen, Li, and Li 2016). Anomaly-based intrusion detection systems (IDS) learn to identify proper network activity in order to detect network intrusions. After that, when unusual network behaviors happen outside of its training sets, it notifies users. The challenge of distinguishing between typical and unusual network traffic patterns may lead to a high number of false positives from anomaly-based intrusion detection systems, raising security issues. In this work, anomaly-based IDS models are constructed using LR (Subba, Biswas, and Karmakar 2015).The JRip algorithm, REP Tree, and Forest PA classifiers based on rules and decision trees are used in this work to provide a unique intrusion detection system (IDS). Network traffic is categorized as Attack/Benign in two ways, depending on the features of the data stream. beginning data set characteristics, first and second classifier outputs, and third classifier inputs. The suggested IDS outperforms cutting-edge systems in terms of accuracy, detection rate, false alarm rate, and time overhead, according to CICIDS2017 dataset study (Ahmim et al. n.d.).

A significant value of 0.001 ($p<0.05$) was reached, suggesting that the Novel Decision tree outperforms Linear Regression. In contrast to its assessed accuracy of 94.90%, the Novel Decision tree's efficiency is 93.72%. The substantial time needed to create a Novel Decision tree is an inherent limitation of our approach, especially when working with large datasets. In the future, the research hopes to shorten the training period for the dataset and increase the system's item diversity.

## CONCLUSION

In order to detect intrusions, this research used two separate machine learning classifiers. If you want to lessen the occurrence of unusual system traffic patterns, this research recommends employing the DTs method. In comparison to Linear Regression's accuracy rating of 93.72%, the Novel Decision tree achieves a rate of 94.90%. After comparing the results, it is clear that this Novel Decision tree outperforms Linear Regression, which achieved an accuracy of 93.72%.

## REFERENCES

Ahmim, Ahmed, Leandros Maglaras, Mohamed Amine Ferrag, Makhlouf Derdour, and Helge Janicke. n.d. “A Novel Hierarchical Intrusion Detection System Based on Decision Tree and Rules-Based Models.” Accessed December 29, 2018. https://doi.org/10.1109/DCOSS.2017.00059.

Alamiedy, Taief Alaa, Mohammed Anbar, Zakaria N. M. Alqattan, and Qusay M. Alzubi. 2017. “Anomaly-Based Intrusion Detection System Using Multi-Objective Grey Wolf Optimisation Algorithm.” *Journal of Ambient Intelligence and Humanized Computing* 11 (9): 3735–56.

Besharati, Elham, Marjan Naderan, and Ehsan Namjoo. 2018. "LR-HIDS: Logistic Regression Host-Based Intrusion Detection System for Cloud Environments." *Journal of Ambient Intelligence and Humanized Computing* 10 (9): 3669–92.

Chen, Pengtian, Fei Li, and Jiatian Li. 2017. "Research on Intrusion Detection Model Based on Bagged Tree." In

Chittora, Pankaj, Tulika Chakrabarti, Papiya Debnath, Amit Gupta, Prasun Chakrabarti, S. Phani Praveen, Martin Margala, and Ahmed A. Elngar. 2018. "Experimental Analysis of Earthquake Prediction Using Machine Learning Classifiers, Curve Fitting, and Neural Modeling," September. https://doi.org/10.21203/rs.3.rs-1896823/v2.

Iwendi, Celestine, Suleman Khan, Joseph Henry Anajemba, Mohit Mittal, Mamdouh Alenezi, and Mamoun Alazab. 2016. "The Use of Ensemble Models for Multiple Class and Binary Class Classification for Improving Intrusion Detection Systems." *Sensors* 20 (9): 2559.

Magán-Carrión, Roberto, Daniel Urda, Ignacio Díaz-Cano, and Bernabé Dorronsoro. 2017. "Towards a Reliable Comparison and Evaluation of Network Intrusion Detection Systems Based on Machine Learning Approaches." *NATO Advanced Science Institutes Series E: Applied Sciences* 10 (5): 1775.

Mohammadi, Sara, Hamid Mirvaziri, Mostafa Ghazizadeh-Ahsaee, and Hadis Karimipour. 2018 "Cyber Intrusion Detection by Combined Feature Selection Algorithm." *Journal of Information Security and Applications* 44 (February): 80–88.

Saranya, T., S, Sridevi, and C, Deisy. 2017. "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review." *Procedia Computer Science* 171 (January): 1251–60.

Subba, Basant, Santosh Biswas, and Sushanta Karmakar. 2015. "Intrusion Detection Systems Using Linear Discriminant Analysis and Logistic Regression." In *2015 Annual IEEE India Conference (INDICON)*. IEEE. https://doi.org/10.1109/indicon.2015.7443533.

Thakallapelli, Abilash, Sudipta Ghosh, and Sukumar Kamalasadan. 2016. "Real-Time Frequency Based Reduced Order Modeling of Large Power Grid." In *2016 IEEE Power and Energy Society General Meeting (PESGM)*. IEEE. https://doi.org/10.1109/pesgm.2016.7741877.

Yuan, Xiaoyong, Pan He, Qile Zhu, and Xiaolin Li. n.d. "Adversarial Examples: Attacks and Defenses for Deep Learning." Accessed December 29, 2017.

ISSN PRINT 2319 1775 Online 2320 7876
 **UGC CARE Listed ( Group -I) Journal Volume 07, Iss 2, 2018**

## TABLES AND FIGURES

**Table 1.** Accuracy and Loss Analysis of Novel Decision Tree

| Iterations | Accuracy (%) | Loss(%) |
|---|---|---|
| 1 | 94.26 | 5.74 |
| 2 | 94.20 | 5.8 |
| 3 | 94.56 | 5.44 |
| 4 | 94.59 | 5.41 |
| 5 | 94.75 | 5.25 |
| 6 | 94.93 | 5.07 |
| 7 | 95.12 | 4.88 |
| 8 | 95.23 | 4.77 |
| 9 | 95.40 | 4.6 |
| 10 | 95.98 | 4.02 |

ISSN PRINT 2319 1775 Online 2320 7876
 **UGC CARE Listed ( Group -I) Journal Volume 07, Iss 2, 2018**

**Table 2.** Accuracy and Loss Analysis of Linear Regression

| **Iterations** | **Accuracy(%)** | **Loss(%)** |
|---|---|---|
| 1 | 93.7 | 6.3 |
| 2 | 94.45 | 5.55 |
| 3 | 93.76 | 6.24 |
| 4 | 93.59 | 6.41 |
| 5 | 93.28 | 6.72 |
| 6 | 93.78 | 6.22 |
| 7 | 92.96 | 7.04 |
| 8 | 93.95 | 6.05 |
| 9 | 93.93 | 6.07 |
| 10 | 93.68 | 6.32 |

ISSN PRINT 2319 1775 Online 2320 7876
 **UGC CARE Listed ( Group -I) Journal Volume 07, Iss 2, 2018**

**Table 3.** Group Statistical Analysis of Novel Decision Tree and Linear Regression. Mean, Standard Deviation and Standard Error Mean are obtained for 10 samples. Novel Decision Tree has higher mean accuracy and lower mean loss when compared to Linear Regression.

| | Group | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| **Accuracy** | Novel Decision Tree | 10 | 94.9020 | 0.5487 | 0.17347 |
| | Linear Regression | 10 | 93.72 | 0.61818 | 0.19548 |
| **Loss** | Novel Decision Tree | 10 | 5.098 | 0.54857 | 0.1734 |
| | Linear Regression | 10 | 6.280 | 0.61818 | 0.19548 |

**Table 4.** Independent Sample T-test: Novel Decision tree is significantly better than Linear Regression with p value 0.001 ($p<0.05$)

| **Group** | | **Levene's Test for Equality of Variances** | | **t-test for Equality of Means** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **F** | **Sig.** | **t** | **df** | **Sig. (2-tailed)** | **Mean Difference** | **Std. Error Difference** | **95% Confidence Interval (Lower)** | **95% Confidence Interval (Upper)** |
| **Accuracy** | **Equal variances assumed** | .082 | .778 | 6.401 | 18 | .001 | 1.67300 | .26136 | 1.123919 | 2.222091 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Equal variances not assumed** | | | 6.401 | 17.749 | .001 | 1.673000 | .261361 | 1.12335 | 2.222656 |

**Table 5.** Comparison of the Novel Decision Tree and Linear Regression with their accuracy. Novel Decision Tree (94.90 %) outperforms Linear Regression (93.72%) in terms of accuracy. When predicting intrusion detection, the Novel Decision Tree approach was more accurate than the Linear Regression classifier

| CLASSIFIER | ACCURACY(%) |
|---|---|
| Novel Decision Tree | 94.90 |
| Linear Regression | 93.72 |

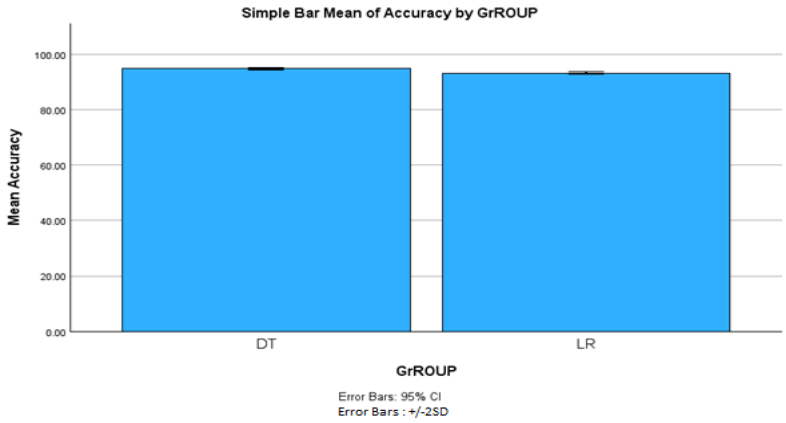


**Fig 1.** Comparison of Novel Decision Tree and Linear Regression. Classifier in terms of mean accuracy and loss. The mean accuracy of Novel Decision Tree is better than Linear Regression. Classifier; Standard deviation of Novel Decision Tree is significantly better than Linear Regression. X Axis: Novel Decision Tree Vs Linear Regression Classifier and Y Axis: Mean accuracy of detection with +/-2SD.