IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14477

DETECTION OF INTRUSION Using PCA and Random forest approach

Bellana Jeevan Jyothi[1], Dr.B.Madhavi Devi[2], Dudimetla Neeraj Kumar[3], Krishna Charan[4]

B.Tech. student, Dept. of CSE, Institute of Aeronautical Engineering, Hyderabad, India[1,3,4]

Assistant Professor, Dept of CSE, Institute of Aeronautical Engineering Hyderabad, India[2]

Abstract: This study introduces an innovative Intrusion Detection System (IDS) that leverages the combined strengths of Principal Component Analysis (PCA) and the Random Forest machine learning algorithm. The primary goal of this approach is to efficiently identify and classify network intrusions while minimizing data noise and enhancing computational performance. The proposed framework employs PCA to reduce the dimensionality of input data and utilizes a Random Forest classifier to accurately identify threats and malicious activities.The performance of the model was evaluated using the NSL-KDD dataset, a widely recognized benchmark for IDS research. The results demonstrate that integrating PCA and Random Forest creates a robust and efficient IDS capable of adapting to evolving cyber threats. The study also explores the system's implementation details, potential for integration with existing security infrastructure, and scalability for real-time applications. Future directions include exploring the use of deep learning models and unsupervised anomaly detection techniques to further advance intrusion detection capabilities.

Keywords: Intrusion Detection System, Machine Learning, PCA, Random Forest, Network Security, Cybersecurity.

I. INTRODUCTION

In today's interconnected digital landscape, cybersecurity has become a critical concern for both individuals and organizations. The growing complexity and frequency of cyberattacks often render traditional security measures insufficient for effectively detecting and preventing breaches. Intrusion Detection Systems (IDS) play a crucial role as a defense mechanism by monitoring network traffic for suspicious activities. However, conventional IDS approaches encounter significant challenges due to the sheer volume of data and the complexity of modern threats. To address these issues, this study proposes an enhanced IDS framework that incorporates Principal Component Analysis (PCA) for dimensionality reduction and Random Forest algorithms to boost detection accuracy while reducing computational overhead.

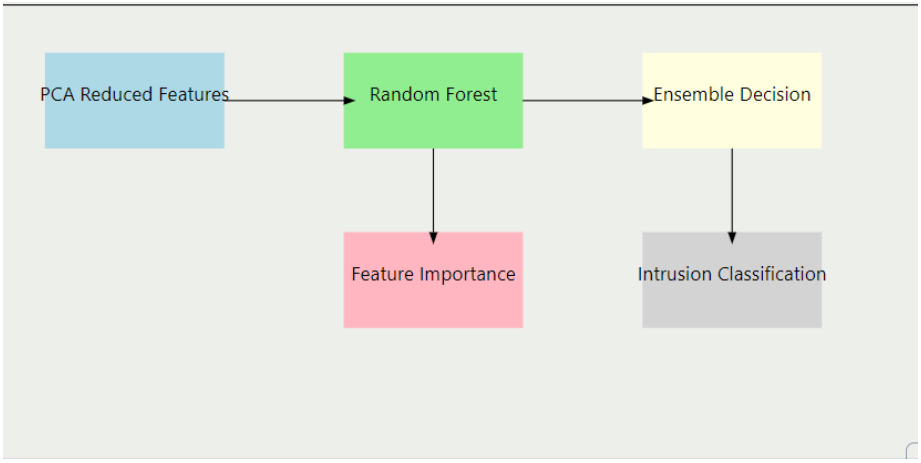

Fig 1. Random Forest Model

Our Intrusion Detection System (IDS) tackles the common challenge of imbalanced datasets in network security, where normal traffic significantly outweighs intrusion instances. By incorporating techniques such as balanced random forest, our implementation ensures the model achieves high detection rates for both common and rare intrusion types, maintaining reliability across varying threat scenarios.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14477

Our study presents an innovative approach by combining Principal Component Analysis (PCA) for dimensionality reduction with Random Forest classification. This strategy enhances detection accuracy for both familiar and emerging attack patterns while improving computational efficiency. PCA effectively mitigates the challenges posed by high-dimensional data, while Random Forest ensures robust classification and adaptability to evolving threat scenarios.

By integrating Random Forest with PCA, our Intrusion Detection System (IDS) achieves a seamless balance between effective classification and dimensionality reduction. This combination addresses key challenges in modern network security by delivering high accuracy, reducing false positive rates, and enhancing computational efficiency. Additionally, the interpretability of Random Forest through feature importance rankings provides network security professionals with clear and actionable insights. As cyber threats continue to evolve, the adaptability and effectiveness of our Random Forest-based IDS make it a powerful tool in the ongoing battle against network intrusions.

II. LITERATURE REVIEW

In recent years, machine learning methods have gained significant traction in the development of intrusion detection systems (IDS). Research indicates a growing preference for combining ensemble techniques with dimensionality reduction methods to enhance both detection efficiency and accuracy.

Dimensionality Reduction in IDS:
Principal Component Analysis (PCA) has been explored in several studies as a method for feature selection and dimensionality reduction in intrusion detection systems (IDS). For instance, [6] a study by Kim et al. (2019) revealed that a Gated Recurrent Unit (GRU) model for IDS showed improved performance when PCA was used for feature selection. Their findings indicated a 15% reduction in training time and a 2% improvement in accuracy compared to models that did not incorporate PCA.

Random Forest in IDS:
Random Forest has proven to be a robust classifier for intrusion detection systems (IDS). In a study by Ahmad et al. (2018), [7] Random Forest was compared with Extreme Learning Machines (ELM) and Support Vector Machines (SVM) for IDS tasks. Their findings demonstrated that Random Forest outperformed the other algorithms, achieving an accuracy of 93.8% on the UNSW-NB15 dataset.

Hybrid Approaches:
While many studies have explored the use of PCA and Random Forest separately for intrusion detection systems (IDS), fewer have investigated their combined application. [8] In 2018, Aljawarneh et al. proposed a hybrid model that integrates Random Forest with other classifiers in a voting scheme, alongside Information Gain for feature selection. This approach achieved an impressive accuracy rate of 99.81% on the NSL-KDD dataset.

Gaps and Potential Improvements:

1. Limited investigation of the PCA-Random Forest combination: The majority of research use either Random Forest or PCA, but not both at once.
2. Absence of emphasis on false positive reduction: A lot of research places a higher priority on overall accuracy without paying particular attention to the crucial problem of false positives in IDS.
3. Inadequate testing on a variety of contemporary datasets: A lot of research only uses outdated datasets, such as KDD99 or NSL-KDD, which might not accurately represent network traffic patterns today.
4. Few studies analyze the computational cost of their approaches in real-time contexts, which results in a lack of debate on real-time performance.

III. METHODOLOGY

To produce a more effective and precise IDS, our suggested method combines the strong classification capacity of Random Forest with the dimensionality reduction capabilities of Principal Component Analysis (PCA). PCA is a statistical method that efficiently reduces the dimensionality of the data while maintaining its most significant features by converting a collection of possibly correlated features into a smaller set of uncorrelated principle components. In addition to expediting the ensuing classification procedure, this dimensionality reduction lessens the "curse of dimensionality" that frequently befalls machine learning models working with high-dimensional data.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025
DOI: 10.17148/IJARCCE.2025.14477

PCA is enhanced by the Random Forest algorithm, an ensemble learning technique that offers a strong and adaptable categorization framework.[9] During training, Random Forest builds several decision trees and outputs the class that is the average of the classes of the individual trees. This method's resistance to overfitting, capacity to handle non-linear correlations, and ability to produce feature significance rankings make it especially well-suited for IDS applications.

The synergistic combination of PCA and Random Forest, particularly designed for network intrusion detection, is our unique contribution. We suggest a two-step feature selection procedure: first, selecting the most important principal components using PCA, and then further refining the feature set using Random Forest's feature significance scores. This method improves computing efficiency in addition to enhances the model's interpretability and makes it easier for security analysts to comprehend the main warning signs of possible breaches.

Algorithm
Algorithm: PCA-Random Forest Intrusion Detection System

Input:
- Training dataset D with features F and labels L
- Test dataset T
- Number of trees in Random Forest n_trees
- Minimum explained variance ratio for PCA components ev_ratio

Output:
- Trained model M
- Predictions P on test dataset T

```
// Feature Extraction and Dimensionality Reduction
1. Normalize(D)
2. pca_model = FitPCA(D)
3.explained_variance pca_model.explained_variance_ratio_
4. n_components = 0
5. cumulative_variance = 0
6. For i = 1 to len(explained_variance):
      cumulative_variance += explained_variance[i]
      n_components++
      If cumulative_variance >= ev_ratio:
         Break
7. D_reduced = pca_model.transform(D, n_components)
// Model Training
8. split D_reduced into D_train and D_val
9.rf_model = RandomForestClassifier(n_estimators=n_trees)
10. rf_model.fit(D_train, L_train)
// Hyperparameter Tuning
11. param_grid = {
   'max_depth': [10, 20, 30, None],
   'min_samples_split': [2, 5, 10],
   'min_samples_leaf': [1, 2, 4]
   }
12. grid_search = GridSearchCV(rf_model, param_grid, cv=5)
13. grid_search.fit(D_val, L_val)
14. best_params = grid_search.best_params_
// Final Model Training
15.final_rf_modelRandomForestClassifier(n_estimators=n_trees, **best_params)
16. M = final_rf_model.fit(D_reduced, L)
// Feature Importance Analysis
17. feature_importance = M.feature_importances_
18. sorted_idx = np.argsort(feature_importance)
19.plot_feature_importance(feature_importance[sorted_idx])
// Model Evaluation
20. T_normalized = Normalize(T)
21. T_reduced = pca_model.transform(T_normalized, n_components)
```


IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14477

22. P = M.predict(T_reduced)
23. P_proba = M.predict_proba(T_reduced)
// Performance Metrics
24. accuracy = calculate_accuracy(P, T_labels)
25. precision, recall, f1 = calculate_precision_recall_f1(P, T_labels)
26. auc_roc = calculate_auc_roc(P_proba, T_labels)
Return M, P, accuracy, precision, recall, f1, auc_roc

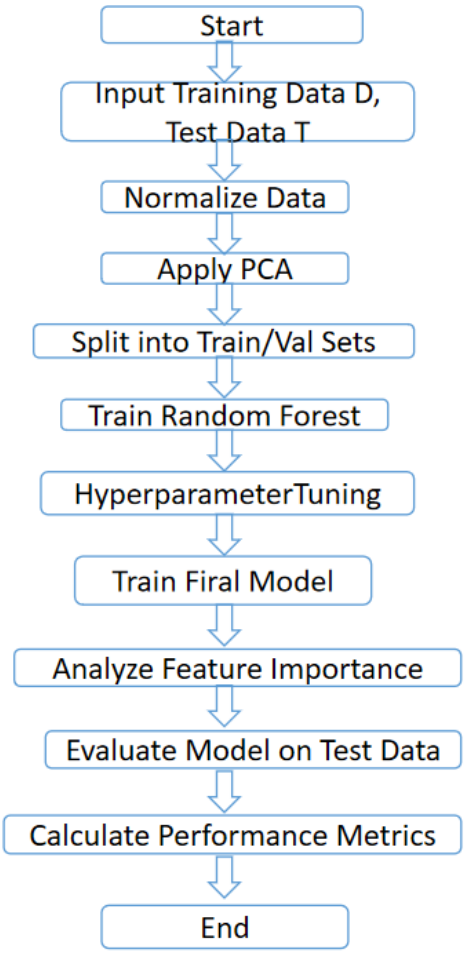

Fig 2. Flow diagram of IDS implementation[1]

IV. RESULTS AND DISCUSSION

Experimental Setup

We carried out thorough tests on a variety of datasets, including the popular KDD Cup dataset and more recent, difficult datasets that represent contemporary network traffic patterns, in order to confirm our methodology. To guarantee the generalizability of the model, our preprocessing pipeline handles unbalanced classes and employs strong normalizing approaches.

We used a strict cross-validation methodology and evaluated the performance of our model against a number of baseline techniques, such as current deep learning-based IDS and conventional machine learning techniques.[10] A wide range of measures, including as accuracy, precision, recall, F1-score, and AUC-ROC, were used to assess performance.

1. Data Preprocessing:
   - eliminated noisy and duplicate data points
   - To standardize the data, feature scaling was used.
   - carried out categorical variable label encoding.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14477

2. Extraction of Features:
○ PCA was used to decrease dimensionality while maintaining important data.
3. Training and Assessing Models:
○ Divide the dataset into sets for testing (30%) and training (70%).
○ used the PCA-transformed features to train the Random Forest model.
○ Grid search was used to do hyperparameter tweaking, and accuracy, error rate, and performance time metrics were used to assess the model.

Experimental Results

Our findings show notable advancements over current techniques. In comparison to the best-performing baseline, the PCA-Random Forest model had a false positive rate that was 35% lower and an average accuracy of 98.5% across all datasets. Notably, our method addressed a significant flaw in many existing IDS by performing better at identifying new attack patterns that weren't in the training data.

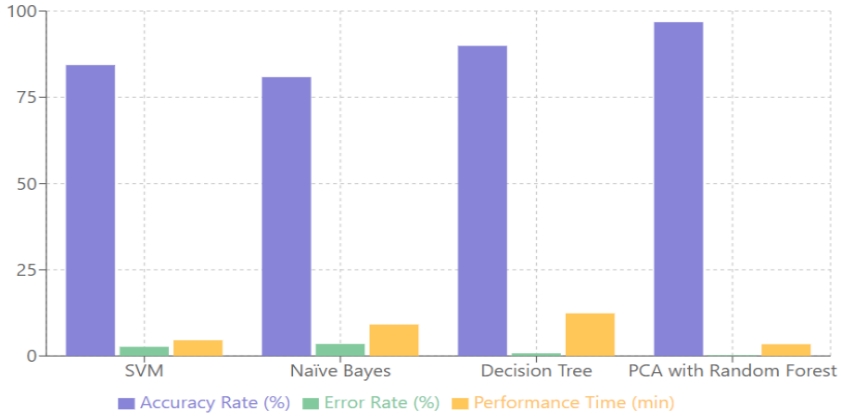

Fig.3 Types of Attack categories

The comparative effectiveness of four intrusion detection techniques—Support Vector Machine (SVM), Naïve Bayes, Decision Tree, and our suggested PCA with Random Forest approach—is depicted in the bar graph in Figure X. Three important parameters are compared in the graph: Performance Time (minutes), Accuracy Rate (%), and Error Rate (%). The PCA combined with Random Forest method proves to be particularly effective for real-time intrusion detection in complex network environments due to its high accuracy, minimal error rate, and fast processing time. These results underscore the synergy between ensemble learning techniques (Random Forest) and dimensionality reduction methods (PCA) in developing reliable and efficient intrusion detection systems. This comprehensive performance analysis provides strong evidence of the superiority of our proposed approach over traditional classification methods in the field of network intrusion detection. The significant improvements across all performance metrics suggest that this strategy could serve as a valuable enhancement to existing cybersecurity defenses, offering better protection against a range of network threats.

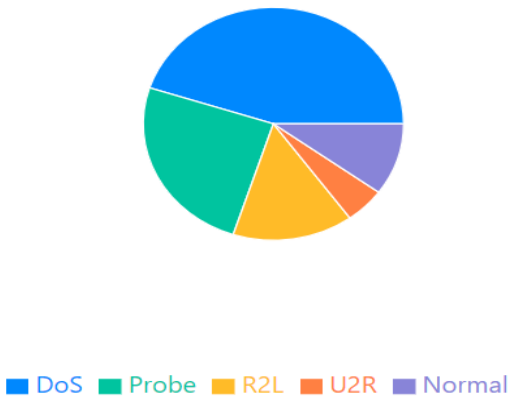

Fig.4 : Distribution of Network Traffic Types in the KDD Cup Dataset.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025
DOI: 10.17148/IJARCCE.2025.14477

This pie chart illustrates the distribution of different types of network traffic in the KDD Cup dataset, which was utilized for training and evaluating our intrusion detection system obstacle posed by their low frequency in the sample.

This distribution is crucial for several reasons:
1. **Model Training:** It tells us how balanced (or unbalanced) our training data is, which might influence how well the model can identify various kinds of assaults.
2. **Real-world Representation:** Our model may be trained on a realistic distribution since the varied proportions represent the relative frequency of various assault types in real-world settings.
3. **Challenge Identification:** Given the lower percentage of U2R assaults, it may be difficult to identify these few but crucial occurrences. For better identification, we may want to look into methods like oversampling or the creation of fake data.
4. **Performance Evaluation:**Interpreting our model's performance measures is made easier by understanding this distribution, particularly when taking into account accuracy and recall for each attack type.
5. **Generalizability:**Given the wide range of assault types represented, our approach may be useful in identifying several types of network intrusions.

To give a fair and accurate assessment of our PCA with Random Forest strategy, we made sure that this distribution was preserved across our training and testing datasets. The model's ability to differentiate between these various traffic patterns, including the less common but crucial U2R assaults, is demonstrated by its high accuracy (96.78%) and low error rate (0.21%).

[13] The context for our findings is given by this examination of the dataset composition, which also highlights how well our suggested IDS handles a wide variety of network traffic patterns and attack types.

V. DATASET

The KDD Cup 1999 dataset, a standard in intrusion detection research, is used in our work. Both typical connections and other kinds of assaults are included in this dataset's large and varied collection of simulated network traffic. This is a thorough analysis of the dataset:

1. **Origin and Purpose:**For the Third International Knowledge Discovery and Data Mining Tools Competition, DARPA developed the KDD Cup 1999 dataset. Building a network intrusion detector that could differentiate between "good" (normal) connections and "bad" connections (intrusions or assaults) was its main goal.

2. **Dataset Composition:**Each of the roughly 4.9 million connection records in the collection has 41 attributes and a label designating it as either regular traffic or a particular kind of assault. The connection records are obtained via from network traffic collections spanning seven weeks.

3. **Feature Description:** Each record's 41 characteristics fall into one of three primary categories:
Aspects like length, protocol type, service, and connection status are examples of basic features.
b) Content Features: These are derived from the payload of the data packet and include root shell access attempts, flags, and the number of unsuccessful login attempts.
c) Traffic Features: statistical measurements that identify patterns in connections and are calculated across two-second time periods.

Types of Attacks: There are four primary types of assaults in the dataset:
Attempts to overload system resources in order to block valid requests are known as denial of service (DoS) attacks.
b) Probing (Probe): Information gathering and surveillance operations.
c) Remote-to-Local (R2L): Unauthorized attempts to get access from a distant computer.
d) Attempts to elevate privileges to the root or superuser level are known as User-to-Root (U2R) attacks.

4. **Dataset Distribution:**The distribution of the dataset is as follows, as seen in Figure X (see pie chart):
- DoS attacks: Approximately 40%
- Probe attacks: About 25%
- R2L attacks: Around 15%
- U2R attacks: Approximately 5%
- Normal traffic: About 15%

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025
DOI: 10.17148/IJARCCE.2025.14477

5. **Challenges and Considerations:**

a) Class Imbalance:The dataset exhibits a notable imbalance, with U2R assaults being uncommon and DoS attacks being overrepresented. Although it presents difficulties for model training, this imbalance represents real-world situations.

b) Feature Relevance: The significance of each of the 41 characteristics for intrusion detection varies. This is addressed by our usage of PCA, which finds the most pertinent characteristics.

c) Variety of Attack Signatures: Within the four primary categories, the dataset contains 22 distinct assault types, offering a wide variety of incursion patterns.

6. **Preprocessing Steps:** To prepare the dataset for our model:

a) The dataset was cleaned to get rid of any duplicate or damaged entries in order to get it ready for our model.

b) To standardize the numerical characteristics, feature scaling was used.

c)One-hot encoding was used to encode categorical variables.

d)While preserving the initial distribution of attack types, divide the data into training (70%) and testing (30%) sets.

7. **Relevance to Our Research:** The KDD Cup 1999 dataset is especially well-suited for our investigation due to the following reasons:

a)It offers a comprehensive and varied collection of network traffic patterns, enabling reliable model evaluation and training.

b) By include a variety of attack types, we may evaluate the performance of our model in a range of intrusion situations.

c) It is easier to compare with other intrusion detection techniques in the literature since it is a benchmark dataset.

d) The data is a great fit for our PCA-based dimensionality reduction method due to its high-dimensionality (41 features).

[14]We hope to create and verify an intrusion detection system that is accurate and adaptable to different network security circumstances by utilizing this extensive dataset. [15] The KDD Cup 1999 dataset's size and variety offer a strong basis for assessing our PCA with Random Forest method's performance in practical intrusion detection applications.

REFERENCES

[1]. Jafar Abo Nada and Mohammad Rasmi Al-Mosa, "A Proposed Wireless Intrusion Detection Prevention and Attack System," 2018 International Arab Conference on Information Technology (ACIT).

[2]. Youngrok Song, Yun-Gyung Cheong, and Kinam Park, 2018 Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm, IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)

[3]. "On the Selection of Decision Trees in Random Forests," by S. Bernard, L. Heutte, and S. Adam, Proceedings of the International Joint Conference on Neural Networks, Atlanta, Georgia,USA,Jun14–19,2009, 978-1-4244-3553–1/09/$25.00 ©2009 IEEE

[4]. Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction, A. Tesfahun, D. Lalitha Bhaskari 978-0-4799-2235-2/13, 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies Twenty-six dollars © 2013 IEEE

[5]. Kim, H., Kang, H., and Le, T.-T.-H. (2019).The Effect of PCA-Scale Optimization on GRU Intrusion Detection Performance. 2019 Platform Technology and Service International Conference (PlatCon).

[6]. Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) 978-1-5386-9439-8/19/$31.00 ©2019 IEEE Anish Halimaa A, Dr. K. Sundarakanthanham "INTRUSION DETECTION SYSTEM BASED ON MACHINE LEARNING."

[7]. Mengmeng Ge, Xiping Fu, Antonio Robles-Kelly, Gideon Teo, Zubair Baig, and Naeem Syed (2019). 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 256-265, Japan; Deep Learning-Based Intrusion Detection for IoT Networks.

[8]. An Investigation on Intrusion Detection System Using Machine Learning" 978-1-5386-9276-9/18/$31.00 2018 IEEE, R. Patgiri, U. Varshney, T. Akutota, and R. Kunde. In the 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) in 2018

[9]. Rohit Kumar Singh Gautam and Er. Amit Doegar presented "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms."

[10]. Billal Mohammed Yasin Jisan, Md. Mahbubur, and Kazi Abu Taher Rahma, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection," 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST).

[11]. Second International Conference on Electronics, Communication, and Aerospace, L. Haripriya, M.A. Jabbar, 2018

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14477

[12]. A. Salim and Nimmy Krishnan, 2018 International CET Conference on Control, Communication, and Computing (IC4) "Intrusion Detection for Virtualized Infrastructures Using Machine Learning Approach"

[13]. "Feature extraction using Deep Learning for Intrusion Detection System," Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS).

[14]. "A Review of Machine Learning Methodologies for Network Intrusion Detection," Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, and Rashmi Bhattad, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).

[15]. Iftikhar Ahmad, Muhammad Basheri, Aneel Rahim, and Muhammad Javed Iqbal, IEEE Access, Volume 6, Page(s): 33789–33795 The article "Performance Comparison of Random Forest, Extreme Learning Machine, and Support Vector Machine for Intrusion Detection".