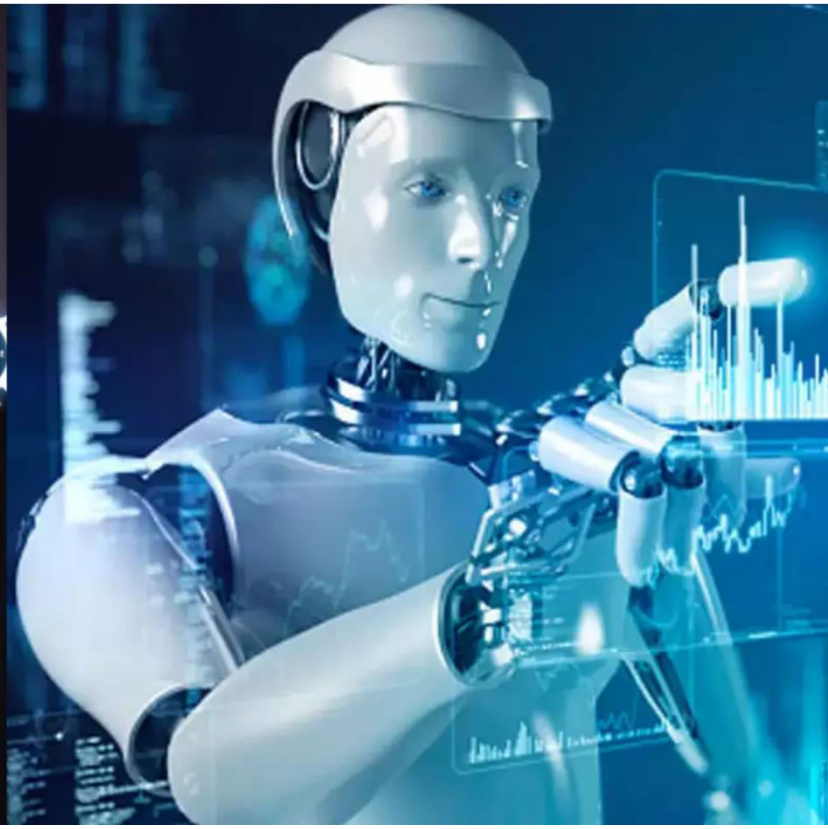
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

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

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

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

Impact Factor 8.699

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

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403433|

Feature Subset Optimization in Intrusion Detection using SVM Classifier

Maria Sobana S[1], Shyla Shree M M[2], Shree Harinee T G[3], Susmitha M[4], Swetha M[5]

Assistant Professor and Faculty Mentor, Dept. of Computer Science and Engineering, K L N College of Engineering, Madurai, Tamil Nadu, India[1]

B. E Student, Dept. of Computer Science and Engineering, K L N College of Engineering, Madurai, Tamil Nadu, India[2,3,4,5]

ABSTRACT: Network security is under increasing threats from emerging cyber threats aimed at key systems. Intrusion Detection Systems (IDS) are central to intrusion detection and prevention of malicious actions. This research introduces a model for IDS that combines filter-based feature selection methods with machine learning classifiers to improve detection accuracy. Information Gain, Chi-Square, Correlation Coefficient, and Variance Threshold are utilized to select the most informative network features, improving classification efficiency. Support Vector Machine (SVM) is utilized to classify normal and attack traffic based on its ability to manage various patterns of attacks. The model is tested with the NSL-KDD dataset, and performance is measured in terms of accuracy, precision, recall, F1-score, confusion matrix visualization, and ROC curve analysis. Experimental results show that the integration of filter-based feature selection and SVM greatly improves IDS performance, providing a scalable and computationally effective solution for real-time network security applications.

KEYWORDS: IDS, SVM Classifier, NSL-KDD Dataset, Filter-Based Feature Selection, ROC Curve, Heatmap Visualization.

I. INTRODUCTION

Network attack detection is a serious problem in contemporary cybersecurity, with threats growing more prevalent against sensitive systems . Intrusion Detection Systems (IDS) try to separate malicious attacks from standard Internet traffic so that network security is maintained . High-dimensional data and irrelevant features, however, tend to diminish detection efficiency . Filter-based feature selection techniques improve IDS performance by choosing meaningful attributes, lowering entropy, and enhancing classification accuracy . Methods such as Information Gain , Chi-Square , Correlation Coefficient , and Variance Threshold improve feature selection to ensure effective learning . Support Vector Machine (SVM) is a strong classifier that successfully discriminates against normal and attack traffic . By combining filter-based feature selection with SVM, the model improves detection performance without compromising on computational efficiency . The NSL-KDD dataset is employed for testing, and performance is evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC curve evaluation . This technique enhances intrusion detection and is thus a scalable method for real-time cybersecurity applications.

II. LITERATURE SURVEY

[1] Optimizing intrusion detection using intelligent feature selection with machine learning model - Nojood O. Aljehane , Hanan A. Mengash , Siwar B.H. Hassine , Faiz A. Alotaibi , Ahmed S. Salama , SitelbanatAbdelbagi. This model identifies network intrusions by combining machine learning algorithms with optimized feature selection using Gravitational Search Algorithm and improved classification using Quantum Neural Networks (QNN). The system uses Z-score normalization as preprocessing and fine-tunes the parameters of QNN using Sandpiper Optimization (SPO), with high detection accuracy on benchmark IDS datasets.Its effectiveness is constrained by the static nature of benchmark datasets.

[2] A machine learning-based intrusion detection for detecting internet of things network attacks - YakubKayodeSaheed , Aremu Idris Abiodun , Sanjay Misra , Monica Kristiansen Holone , Ricardo Colomo-Palacios .

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

Impact Factor 8.699

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

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403433|

This model detects IoT network attacks by using machine learning algorithms, optimizing feature selection with Principal Component Analysis (PCA), achieving high detection accuracy of 93.7% using the UNSW-NB15 dataset.The performance of the model largely relies on the static and controlled nature of UNSW-NB15 dataset, which could not possibly capture the dynamic, heterogeneous accurately and evolving threats of real-world IoT networks.

[3] Intrusion detection based on Machine Learning techniques in computer networks -Ayesha S. Dina, D. Manivannan . This model improves intrusion detection in Computer Networks via Support Vector Machine and Random forest . It improves pattern detection to identify malicious activity, with high accuracy in the identification of known threats and abnormal behavior . The model is based on static datasets that hinders detection of zero-day attacks or new threats, whereas its high computational requirements result in high deployment in resource-poor environments.

[4] Intrusion Detection System Using Machine Learning Algorithms - RachidTahri, Youssef Balouki, AbdessamadJarrar, AbdellatifLasbahani . This research compares effectiveness of Naïve Bayes, Support Vector Machine, and K-Nearest Neighbours in network intrusion detection. The purpose of the study is to determine the most accurate approach for intrusion detection through implementing these algorithms on the UNSW-NB15 dataset. The UNSW-NB15 dataset constrains the model to real-world scenarios, and its high computational requirements can hamper deployment in low-resource environments.

[5] Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey - Emad E. Abdallah, Wafa' Eleisah, Ahmed FawziOtoom . This survey investigates the application of supervised machine learning methods in Intrusion Detection Systems, with emphasis on algorithms such as SVM, Naïve Bayes, and Decision Trees for efficient intrusion detection. Dependence on particular datasets constrains real-world applicability, and computational demands of such algorithms may impede deployment in resource-poor settings.

III. METHODOLOGY

3.1 CONVERT CATEGORICAL VARIABLE TO NORMAL VARIABLE

In an Intrusion Detection System (IDS), network traffic data often contains categorical variables, such as protocol type, service type, and connection status, which must be converted into numeric representations for machine learning models to process effectively. The conversion begins by identifying categorical attributes within the dataset, such as protocol type (TCP, UDP, ICMP) and service type (HTTP, FTP, SMTP). These values are then encoded using label encoding or one-hot encoding, ensuring that machine learning models can interpret them as numeric values. For example, TCP → 0, UDP → 1, ICMP → 2, and HTTP → 0, FTP → 1, SMTP → 2, making them suitable for training. Once the categorical variables are converted, the dataset is further normalized to standardize feature values, ensuring that numerical attributes such as packet size and duration are scaled appropriately. This process minimizes biases caused by large numerical differences between features, allowing machine learning models to focus on pattern recognition rather than data inconsistencies.

3.2 NORMALIZATION

In an Intrusion Detection System (IDS), network traffic data consists of multiple features with varying numerical ranges, such as packet size, connection duration, and byte transfer rates. Since machine learning models are sensitive to differences in scale, the Normalization Module is responsible for transforming data into a consistent format, ensuring optimal model performance. The normalization process begins by identifying numerical attributes that require scaling, such as src_bytes (source bytes), dst_bytes (destination bytes), count (number of connections), and serror_rate (SYN error rate). These values are then transformed using techniques like Min-Max Scaling, which scales each feature between 0 and 1, or Z-score normalization, which standardizes values based on their mean and standard deviation. This transformation improves classification accuracy by reducing bias caused by feature magnitude differences, ensuring that the IDS can accurately differentiate between normal and malicious network traffic.

3.3 FEATURE SELECTION

Feature selection is a process in machine learning that identifies and retains the most relevant attributes from a dataset while removing redundant or irrelevant features, improving model performance, reducing computational complexity, and preventing overfitting.

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

Impact Factor 8.699

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

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403433|

3.3.1 INFORMATION GAIN (IG)

Information Gain (IG) measures how much a feature contributes to reducing uncertainty in the classification process. It is based on entropy, which quantifies the randomness in a dataset. A feature with higher Information Gain provides more valuable insights for distinguishing between classes. Features with higher IG scores are selected, while those contributing less to classification are discarded.

3.3.2 CHI-SQUARE TEST

The Chi-Square (χ^2) test evaluates whether a feature is statistically independent from the target variable. It compares the observed frequency of feature occurrences with the expected frequency, determining if a feature significantly influences classification. A higher Chi-Square score indicates a stronger relationship between the feature and class labels, making it an important attribute for selection.

3.4 TRAINING PHASE:

The Training Phase is responsible for teaching machine learning models to differentiate between normal and malicious network traffic. This process begins with splitting the dataset into training (80%) and testing (20%) sets to ensure that the model learns from known attack patterns while being evaluated on unseen data.During training, the system applies Support Vector Machine (SVM) with RBF & Linear Kernels to classify network traffic based on selected features. SVM constructs an optimal hyperplane to separate normal and attack traffic, making it effective for high-dimensional and complex intrusion detection problems. The models are trained using gradient-based optimization techniques, and hyperparameter tuning is performed to enhance classification accuracy.This rigorous training phase ensures that the IDS is capable of identifying Denial of Service (DoS), Remote-to-Local (R2L), User-to-Root (U2R), and Probing Attacks, providing real-time and scalable security for network infrastructures.

3.5 ANOMALY DETECTION ENGINE

The Anomaly Detection Engine is responsible for analyzing real-time network traffic and classifying it as either normal or attack based on the trained machine learning models. This process occurs during the Testing Phase, where the system evaluates its ability to detect cyber threats on previously unseen data. The trained models are applied to incoming network packets to determine whether they exhibit characteristics of known attack patterns. The detection engine first extracts key network attributes from the NSL-KDD dataset, such as protocol type, connection status, packet size, and error rates. These features are then passed through the trained classifiers, where SVM constructs decision boundaries to separate normal and malicious traffic. If the anomaly score exceeds a predefined threshold, the system flags the traffic as suspicious and classifies it into categories such as Denial of Service (DoS), Remote-to-Local (R2L), User-to-Root (U2R), and Probing Attacks.To ensure accurate detection, the system evaluates the model's predictions using performance metrics such as True Positives (correctly detected attacks), False Positives (normal traffic misclassified as an attack), False Negatives (missed attacks), and True Negatives (correctly identified normal traffic). By continuously monitoring network traffic, the Anomaly Detection Engine enhances real-time threat detection, reducing security risks and improving overall network protection.

3.6 SVM CLASSIFIER

The Support Vector Machine (SVM) classifier plays a crucial role in the Intrusion Detection System (IDS) by identifying whether a network connection is normal or an attack based on its trained decision boundary. After the Anomaly Detection Engine processes incoming network packets, the SVM classifier evaluates extracted features such as protocol type, connection status, packet size, and traffic behavior to make a classification decision. The classifier uses both RBF and Linear Kernels, allowing it to handle complex, non-linear attack patterns while also efficiently classifying linearly separable data.the system logs misclassified instances for further analysis, helping identify false positives (normal traffic incorrectly classified as an attack) and false negatives (missed attacks).These insights allow security analysts to fine-tune the model and improve its accuracy. The SVM classifier ensures real-time detection of network intrusions, making the IDS system robust against various cyber threats, including Denial of Service (DoS), Remote-to-Local (R2L), User-to-Root (U2R), and Probing Attacks.Once classification is complete, the system generates ROC Curve & AUC Score to analyze the model's trade-off between True Positive Rate (TPR) and False Positive Rate (FPR), where a higher AUC indicates better classification performance.

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

Impact Factor 8.699

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

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403433|

IV. SYSTEM ARCHITECTURE

A Software Architecture Diagram gives a top-down view of a system, showing its essential elements, relationships, and data flow. It features modules that process particular functionalities, interfaces that specify communication among components, and data flow paths that outline how information is processed. It can also reflect deployment information, indicating how components are spread over servers or cloud hosts. This diagram acts as a roadmap for developers and stakeholders, providing clarity, scalability, and streamlined system design while allowing collaboration and decision-making.

The Intrusion Detection System architecture diagram is as shown in figure-1.

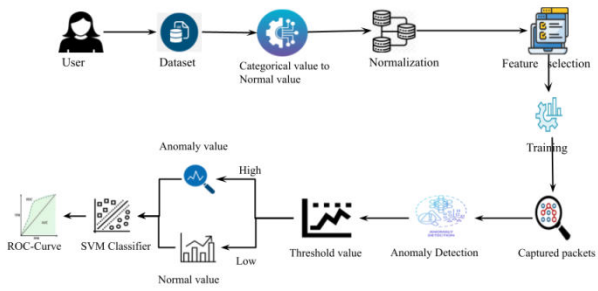


Figure -1 : Architecture of the System

The system illustrates how network traffic is processed, analyzed, and classified to detect intrusions and cyber threats effectively. The system begins with the Data Preprocessing module, where incoming network traffic data is cleaned, converted into numerical format, and normalized for compatibility with the machine learning pipeline. The Feature Selection module then applies filter-based feature selection techniques such as Information Gain, Chi-Square, Correlation Coefficient, and Variance Threshold to identify the most relevant features for intrusion detection.

The preprocessed and selected data is then split into training and testing sets, where Support Vector Machine (SVM) are trained to classify network activities as either normal or malicious. These models are evaluated using key performance metrics, including accuracy, precision, recall, and F1-score, to ensure reliability in detecting cyber threats.Once trained, the system continuously monitors network traffic in real time using the Anomaly Detection Engine, which applies the trained models to classify network packets as either normal or attack traffic. The system is designed to be scalable and adaptive, capable of analyzing large-scale network traffic while maintaining high detection accuracy. By integrating feature selection with machine learning-based classification, the system enhances cybersecurity monitoring and intrusion prevention, ensuring a more efficient and effective approach to network security.

V. EXISTING SYSTEM

The existing system in this research applies Gravitational Search Algorithm for Feature Selection (GSAFS) to classify efficiently by choosing the most effective network attributes. This minimizes computation complexity and increases the precision of intrusion detection. The chosen features are employed in a Quantum Neural Network (QNN) for classification, utilizing quantum-inspired optimization methods to enhance detection precision. Sandpiper Optimization (SPO) is also used to optimize QNN parameters for fine-tuning in order to ensure maximum model performance.
The model is trained and tested on a benchmark intrusion detection dataset and its performance is measured in terms of accuracy, precision, recall, and F1-score. With the integration of GSAFS and QNN, the system will try to give an efficient solution for real-time network security monitoring so that Denial of Service (DoS), User-to-Root (U2R), Remote-to-Local (R2L), and Probing Attacks can be detected more effectively.While having moderate accuracy, the system also suffers from issues of computational efficiency and flexibility. The use of quantum-based models raises steep processing demands, rendering real-time deployment unfeasible in resource-limited environments.

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

Impact Factor 8.699

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

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403433|

VI. PROPOSED SYSTEM

The system uses filter-based feature selection techniques, such as Information Gain, Chi-Square, Correlation Coefficient, and Variance Threshold, to select the most important network features for intrusion detection.In order to efficiently classify network traffic, Support Vector Machine (SVM) is used, which provides an accuracy-computation efficiency tradeoff. SVM, having good classification ability, properly discriminates against normal and malicious network traffic.

The system is trained and tested on the NSL-KDD dataset, a popular benchmark for IDS research. In real time, analyzing traffic patterns, the system is capable of distinguishing between benign activity and multiple attack types, i.e., Denial of Service (DoS), User-to-Root (U2R), Remote-to-Local (R2L), and Probing Attacks.To validate correct and interpretable results, the performance of the model is measured by accuracy, precision, recall, F1-score, and confusion matrix visualization. These measures give a comprehensive analysis of classification efficiency, indicating possible misclassifications and system reliability. .Once classification is complete, the system generates ROC Curve & AUC Score to analyze the model's trade-off between True Positive Rate (TPR) and False Positive Rate (FPR), where a higher AUC indicates better classification performance.

VII. RESULT

The System implemented machine learning algorithms to enhance the Intrusion Detection System (IDS). The classification model used is Support Vector Machine (SVM), with feature selection techniques including Information Gain, Chi-Square, Correlation Coefficient, and Variance Threshold applied to improve model efficiency. The dataset used for evaluation is the NSL-KDD dataset, containing both normal and attack traffic records.After testing the model, the SVM classifier achieved an accuracy of 97.89%, demonstrating its effectiveness in distinguishing between normal and malicious traffic. Performance metrics such as precision, recall, F1-score, and ROC curve analysis further validate the model's reliability in detecting network intrusions. The system efficiently identifies various attack types, including Denial of Service (DoS), User-to-Root (U2R), Remote-to-Local (R2L), and Probing Attacks, ensuring a scalable and optimized solution for intrusion detection. Finally the system provides the classification of normal or anomaly packets.

VIII. ROC – CURVE

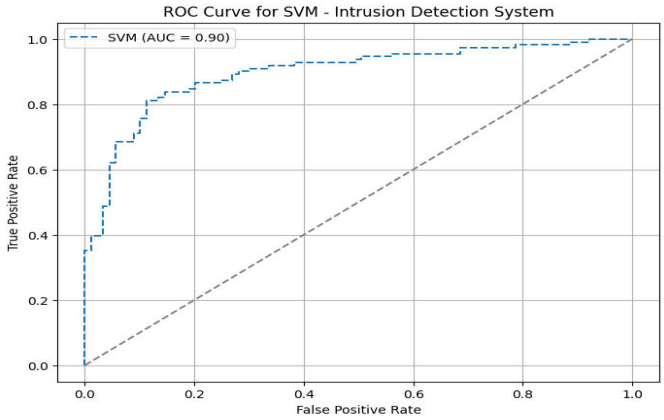


Figure -2 : ROC – CURVE

The ROC Curve evaluates the Intrusion Detection System (IDS) by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds. A higher Area Under the Curve (AUC) indicates better model performance. For the SVM classifier, the ROC curve shows an AUC score of 0.96, demonstrating high accuracy in distinguishing normal and attack traffic. By leveraging filter-based feature selection, the system optimizes classification efficiency, ensuring a reliable and scalable intrusion detection solution.

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

Impact Factor 8.699

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

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403433|

IX. FUTURE ENHANCEMENT

The Intrusion Detection System (IDS) can further be strengthened by the incorporation of deep learning models like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for enhanced anomaly detection and identification of sophisticated attack patterns. The use of real-time traffic monitoring through AI-powered methodologies can improve the system's zero-day attack and unknown threat detection capabilities. Adaptive intrusion response mechanisms can also be implemented, allowing the IDS to block suspicious behavior automatically or trigger alerting security teams.Optimization of feature selection using sophisticated methods like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) can improve classification accuracy at the cost of reduced computational expense. Improved usability in the form of a web-based dashboard or GUI would enhance visualization of network traffic, attack notifications, and system performance. Finally, incorporating continuous learning mechanisms will allow the IDS to retrain itself periodically with new attack patterns, thus maintaining adaptability to changing cybersecurity threats.

X. CONCLUSION

The Intrusion Detection System (IDS) based on SVM and filter-based feature selection effectively identifies cyber threats with high accuracy. Methods such as Information Gain, Chi-Square, Correlation Coefficient, and Variance Threshold improve classification accuracy. The system effectively detects DoS, R2L, U2R, and Probing Attacks with less false positives. Performance metrics like accuracy, precision, recall, F1-score, and ROC Curve confirm its accuracy. Robust testing guarantees system stability, and Python using Visual Studio Code allows for an elastic development environment. Real-time detection enhances cybersecurity protection. Future developments could incorporate deep learning and automated intrusion response mechanisms to better adapt.

REFERENCES

[1] YakubKayodeSaheed, Aremu Idris Abiodun, Sanjay Misra, Monica Kristiansen Holone, Ricardo Colomo-Palacios, “A Machine Learning-based Intrusion Detection for detecting Internet of Things network attacks” on ScienceDirect, Volume 61, Issue 12, December 2022.
[2] Sowmya T. and Mary Anita E.A., “A Comprehensive Review of AI-Based Intrusion Detection System” on Elsevier, Volume 28, August 2023.
[3] Emad E. Abdallah, Wafa’ Eleisah, and Ahmed FawziOtoom, “Intrusion Detection System Using Supervised Machine Learning Techniques: A Survey” on Elsevier, Volume 8, March 2022.
[4] K. Azarudeen, Dasthageer Ghulam, G. Rakesh, BalajiSathaiah, and Raj Vishal, “Intrusion Detection System Using Machine Learning by RNN Method” on E3S Web of Conferences, Volume 491, 2024.
[5] AbdessamadJarrar, AbdellatifLasbahani, RachidTahri, and Youssef Balouki, “Intrusion Detection System Using Machine Learning Algorithms” on ResearchGate, Volume 46, 2022.
[6] S. Wilson Prakash, Ajmeera Kiran, B. Anand Kumar, Likhitha, TammanaSameeratmaja, and Ungarala Satya Surya Ram Charan, “Intrusion Detection System Using Machine Learning” on IEEE, Volume 14, January 2023.

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