NATIONAL BOARD FOR TECHNICAL EDUCATION
NATIONAL JOURNAL OF TECHNICAL EDUCATION
Volume 24 Nos. 1 2025
ISSN No. 2992-3522

ACCELERATED NETWORK ANOMALY IDENTIFICATION WITH GENETIC ALGORIMTHS (GA) METHOD, SUPPORT VECTOR MACHINE (SVM) & RECURSIVE FEATURE REMOVAL

Abubakar Ibrahim Muhammad
Department of Mechatronics Engineering,
Kaduna Polytechnic Nigeria
ibngarko@yahoo.com

Kabir Mustapha Muhammad
Department of Electrical and Electronics Engineering,
Kaduna Polytechnic Nigeria
mkabir337@gmail.com

Ibrahim Shamsuddeen
Department of Electrical and Electronics Engineering,
Kaduna Polytechnic Nigeria
shamo4real65@gmail.com

Abstract

Finding odd trends and possible dangers in network data is the goal of network anomaly identification, a crucial cybersecurity activity. Support vector machines (SVMs) and other machine learning (ML) algorithms are frequently used in existing solutions. However, complicated, non-linear correlations in network data may be difficult for typical SVMs with fixed kernels to grasp, which could result in less-than-ideal detection performance. Furthermore, choosing the appropriate hyperparameters is frequently a difficult task. In this study, we examine the use of SVMs for network anomaly detection using several kernels (linear, sigmoid, polynomial, and radial basis function (RBF)). We expand our investigation to include genetic algorithm (GA)-based RBF kernel optimization for fine-tuning crucial hyperparameters (C and Gamma). Simultaneously, we incorporate recursive feature elimination (RFE) to achieve optimal feature selection and improve the discriminatory ability of the model. Our study uses a standard dataset with more than 40 features and over 40,000 training examples that has been carefully reprocessed in order to assess our suggested technique. A thorough assessment utilizing 10-fold cross-validation was carried out, and the result was a final validation on a different test dataset. Our empirical findings demonstrate that, when combined with RFE and GA-optimized hyperparameters, the GA-RBF-SVM model outperforms the other SVM variations. With 99.46% accuracy on the test dataset and 99.39% accuracy on validation data, this arrangement performs remarkably well. Our research also includes sophisticated (ensemble) ML methods, all of which exhibit outstanding performance benchmarks. This study shows that RBF-kernel based SVMs can outperform sophisticated machine learning algorithms in network anomaly detection when systematic parameter optimization and suitable feature selection are applied.

Keywords—*Anomaly Detection, Support Vector Machines, Genetic Algorithm, Radial Basis Function (RBF) Kernel, Recursive Feature Selection*

I. INTRODUCTION

In today's rapidly evolving world, effectively detecting and addressing anomalous activities is crucial to safeguarding digital ecosystems. This is due to the rapid progress experienced in the internet and communications sectors, leading to a substantial expansion in network size and the accompanying data volume [1]. Also, ensuring the security and stability of computer networks has become a paramount concern as organizations and individuals become increasingly dependent on networked systems for communication, data storage, and various essential operations, the potential impact of network anomalies and intrusions becomes more significant. These anomalies encompass a wide range of irregular behaviors, including but not limited to cyberattacks, hardware failures, and software glitches. Timely and accurate detection of such anomalies is crucial to maintaining the integrity and functionality of networked systems.

Traditional signature-based intrusion detection systems (IDS) have been widely employed to safeguard networks from known attack patterns [2, 3]. However, their effectiveness is limited by the fact that they can only detect attacks for which they possess predefined signatures. With the rise of sophisticated and adaptive attack techniques, there is a growing need for more robust and adaptive intrusion detection approaches [4].

Support vector machines (SVM) have shown promise in addressing this need [5]. SVM is a powerful machine learning algorithm that can effectively classify data points into different categories by identifying optimal decision boundaries [6]. Its application in network anomaly detection involves training on a labelled dataset of normal and anomalous network traffic, enabling it to discern patterns that differentiate regular behavior from potentially malicious activities. Despite its potential, SVM's performance heavily depends on the selection of appropriate features and hyper parameters.

This paper proposes an integrated approach that combines SVM, Genetic Algorithms, and Recursive Feature Elimination to develop an optimized network anomaly detection system. The primary objective is to create a model that can accurately identify anomalous network behavior while maintaining efficiency and adaptability. Genetic Algorithms (GA) are heuristic search algorithms inspired by the process of natural selection. GAs can efficiently explore large solution spaces and discover optimal or near-optimal solutions by evolving a population of potential solutions over successive generations [7-9]. Integrating GA and RFE into the SVM-

based detection framework aims to optimize the selection of hyperparameters and features respectively, thus improving the system's ability to discriminate between normal and anomalous network behavior. This study goes beyond the typical use of SVMs as we leverage the power of GA-guided fine-tuning to optimize key parameters like Regularization parameter (C) and Gamma (the variance of the Gaussian bell around the support vectors), specifically for the RBF kernel. Additionally, Recursive feature elimination (RFE) is used to intelligently select features, resulting in a collection of finely-tuned and robust models.

Our proposed approach was tested on a Network Intrusion Detection dataset, consisting of over 47,000 instances and 40+ features. We ensure data quality and compatibility through a number of preprocessing pipeline, including data cleaning, scaling, normalization, and encoding. The effectiveness of our models is assessed through 10-fold cross-validation on validation data, followed by an evaluation on a separate unseen test dataset.

II. RELATED WORKS

Prior research in network intrusion detection encompasses a variety of techniques ranging from traditional to modern machine learning approaches. The work of Faker & Dogdu (2019) [10] integrates Big Data and Deep Learning, employing Deep Feed-Forward Neural Network (DNN), Random Forest, and Gradient Boosting Tree (GBT) classifiers for network intrusion detection. The approach achieves high accuracy, with DNN excelling in binary and multiclass classification on the UNSW NB15 dataset (99.16% and 97.01%, respectively), and GBT achieving the best binary accuracy (99.99%) on the CICIDS2017 dataset, while DNN has the highest multiclass accuracy (99.56%). Lee et al. (2019) [11] present an AI technique for cyber-threats detection, based on artificial neural networks including FCNN, CNN, and LSTM, to convert and analyze security events for improved accuracy in discriminating between true and false positive alerts. Experimental evaluation on benchmark and real-world datasets demonstrates the effectiveness of the proposed method, surpassing conventional machine-learning approaches in network intrusion detection. Zhang et al. (2019) [12] propose an intrusion detection model based on Convolutional Neural Network (CNN). The paper addresses the challenge of cyber threats to Intrusion Detection Systems (IDS) in the context of integrating the Internet with social life. Traditional machine learning-based IDS performance falls short, leading to the proposal of a Convolutional Neural Network (CNN) IDS model. SMOTE-ENN algorithm was

applied to balance network traffic and using the NSL-KDD dataset. The proposed CNN IDS achieves an 83.31% accuracy, notably improving detection rates for User to Root (U2R) and Remote to Local (R2L) attacks, surpassing previous IDS models. Taking advantage of the robust NSL-KDD dataset, Negandhi et al. (2019) [13] employ the supervised learning algorithm random forests to train a model to detect various networking attacks. Aided by Gini-based feature selection, the approach demonstrates improved efficiency and accuracy in identifying network attacks. The work of Miah et al. (2019) [14] introduce a new method for improving detection rate to classify minority-class network attacks/ intrusions using cluster-based under-sampling with Random Forest classifier. The work of Nagaraja et al. (2020)

[15] studied similarity-based feature transformation for network anomaly detection. The paper introduced a novel gaussian distance function for similarity determination and proposed feature transformation for dimensionality reduction, and demonstrates improved anomaly detection performance compared to recent methods on KDD and NSL-KDD datasets. Ayub et al. (2020) [16] demonstrates the historical significance of Intrusion Detection Systems (IDS) in network defense and how machine learning (ML) has improved their accuracy. The paper introduces an ML approach using Multilayer Perceptron (MLP) for intrusion detection, highlighting its effectiveness on various datasets, but also reveals vulnerability to a model evasion attack using the Jacobian-based Saliency Map Attack (JSMA) method, emphasizing the potential for attackers to evade IDS effectively. The work of Pathak & Pathak (2020)[17] employs Decision Tree and KNN machine learning techniques on an IDS, assessing their performance using Univariate feature selection with ANOVA on the NSL-KDD dataset. Performance metrics including accuracy, recall, precision, and F-score are used to compare the algorithms' effectiveness. (Liu et al. (2020) [18] addresses the challenge of detecting malicious cyber-attacks hidden within imbalanced network traffic by proposing a novel Difficult Set Sampling Technique (DSSTE) algorithm. DSSTE divides the training set into difficult and easy sets using Edited Nearest Neighbour (ENN), compresses majority samples in the difficult set using KMeans, and synthesizes new minority samples to balance the classes. Experimental results on classic and comprehensive intrusion datasets show that DSSTE improves classification performance compared to other methods, benefiting classifiers like RF, SVM, XGBoost, LSTM, AlexNet, and Mini-VGGNet. Kumar et al. (2021) [19] employs

supervised (Naive Bayes) and unsupervised (Self Organizing Maps) machine learning techniques, along with deep learning methods like CNN for feature extraction, to Enhance Network Intrusion Detection. The best detection rate is achieved for User to Root attacks (93.0%), while Denial of Service attacks show the lowest rate (0.02%). Almaiah et al. (2022) [20] investigates an intrusion detection model using Principal Component Analysis for feature selection and various Support Vector Machine kernels (linear, polynomial, Gaussian radial basis function, Sigmoid) to enhance intrusion detection system performance. Evaluation on KDD Cup'99 and UNSW-NB15 datasets reveals the superiority of the Gaussian radial basis function kernel in terms of accuracy, Sensitivity, and F-measure, achieving 99.11%, 98.97%, and 99.03% on KDD Cup'99, and 93.94%,

93.23%, and 94.44% on UNSW-NB15 datasets, respectively.

III. METHODOLOGY

This study involves a systematic approach including dataset preparation and preprocessing, model selection, hyperparameter tuning, feature selection, and performance evaluation. This is depicted in Figure 1:

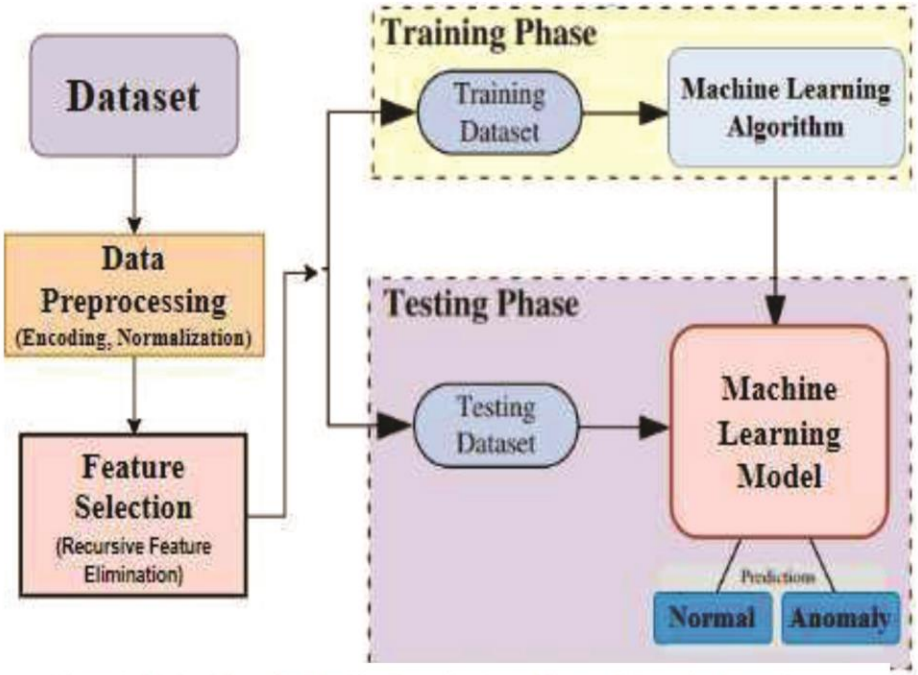

Figure 1. Methodology for Machine Learning-Based Network Intrusion Detection system

A. Dataset Preparation/Preprocessing

The utilized dataset comprises 47,736 instances (25,192 train and 22,544 test) and KDD dataset features with a binary classification label (Normal/Anomaly) which serves as the foundation for training, validation, and testing. Missing values are managed through mean imputation to ensure data integrity. To mitigate issues caused by varying scales, all features are scaled and normalized to a uniform range of 0 to 1, enhancing model convergence and performance. Since SVMs and other models require numerical input, categorical features are encoded into numerical values.

B. Model Selection

Several SVM Variants including Linear, Polynomial, Sigmoid, and Radial Basis Function (RBF) kernels are chosen for their varied capabilities

in capturing complex relationships within the data. In order to provide a comprehensive overview of the SVMs' performance against state-of-the-art algorithms. The models developed using SVM variants are compared with advanced ensemble learning techniques including LightGBM, XGBoost, CatBoost, and Random Forest.

C. Hyperparameter Tuning

We leverage GA optimization to fine-tune hyperparameters in our SVM-based NIDS. The scheme iteratively evolves the hyperparameters C and gamma, while constraining them within ranges (C=0.1 to 10.0 and gamma= 0.01 to 1.0) and tailoring the SVM models for improved performance. We utilize tournament-based selection, two-point crossover, and Gaussian mutation as genetic operators, and iteratively generate offspring, evaluate accuracy, and select top performers as depicted in Figure 2. The population size and number of generations were each set to 10. We extracted optimal C and gamma values from the fittest individual, and achieve refined SVM accuracy while reinforcing network security against intricate threats in the NID task with the ultimate goal to maximize accuracy. The optimal values of C and Gamma were obtained to be C= 8.568 and Gamma= 0.722 and these were used to train the GA-RBF-SVM model

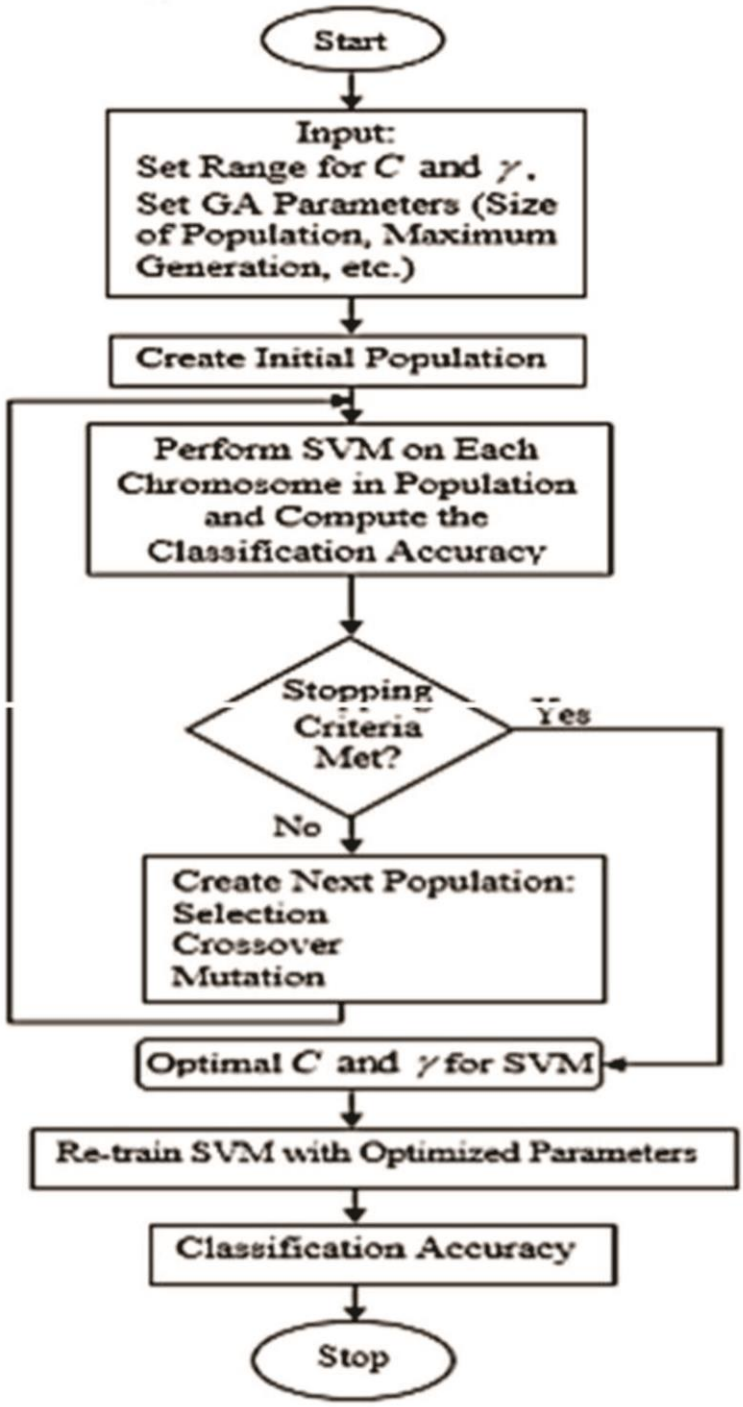

Figure 2. Flowchart of SVM System Optimized by GA

D. Feature Selection

To enhance model efficiency and mitigate noise, RFE is employed to intelligently select the most relevant features. This process contributes to the creation of a refined and robust model. The RFE technique was seamlessly integrated into the pipeline to enhance the quality of feature representation and systematically evaluate the significance of each feature's contribution to the model's performance. Starting with the entire feature set, the algorithm iteratively pruned the least influential features based on their impact on model accuracy. The process continued until the optimal subset of features was identified. This selection procedure not only improved the model's interpretability by focusing on the most relevant attributes but also contributed to the reduction of computational complexity. Consequently, RFE aided in crafting an optimized feature set that augmented the anomaly detection capabilities of our NID system. A visual representation showing the relative importance of each feature in the dataset is shown in figure 2.

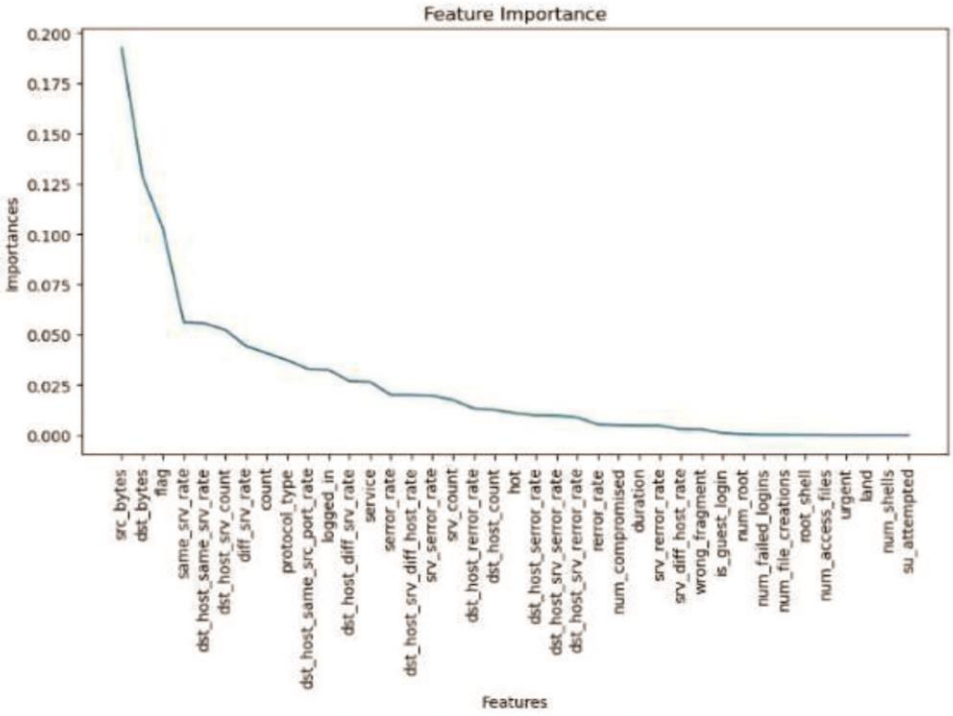

Figure 2: Feature Importance Plot of the Dataset *E. Model Evaluation*

The effectiveness of the model is evaluated using 10-fold cross-validation technique on the validation dataset. The final model's performance is assessed on an independent test dataset

to ascertain its real-world applicability and generalization capabilities. The models' performance is assessed using key metrics such as Accuracy, Precision, Recall, and F1-score.

IV. RESULTS AND DISCUSSION

The following section presents the outcomes of our study on the NID dataset using SVMs with various kernels, GA-optimized hyperparameters, and Recursive Feature Elimination (RFE) for feature selection. The study further encompasses a comprehensive evaluation of different SVM variants, including linear, polynomial, sigmoid, and Radial Basis Function (RBF) kernels. Additionally, the performance of ensemble techniques such as LightGBM, XGBoost, CatBoost, and Random Forest is compared to provide a holistic view of the NID model landscape. Our experimental results (Tables 1 and 2) showcase the performance of each algorithm in both validation and test data scenarios.

TABLE 1: RESULTS FOR MODEL TESTING ON VALIDATION DATA USING 10-FOLD CROSS VALIDATION

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1-score
Linear SVM	0.9448 ± 0.071	0.9625 ± 0.081	0.9173 ± 0.013	0.9394 ± 0.079
Sigmoid SVM	0.8532 ± 0.024	0.8415 ± .0243	0.8443 ± 0.060	0.8428 ± 0.026
Poly SVM	0.9815 ± 0.027	0.9867 ± 0.047	0.9735 ± 0.051	0.9801 ± 0.036
Default-RBF-SVM	0.9884 ± 0.023	0.9853 ± 0.047	0.9899 ± 0.025	0.9876 ± 0.025
GA-RBF-	0.9939	0.9951	0.9919	0.9935

SVM	± 0.022	± 0.043	± 0.033	± 0.023
LightGBM	0.9976 ± 0.080	0.9987 ± 0.070	0.9962 ± 0.019	0.9974 ± 0.095
XGBoost	0.9972 ± 0.070	0.9985 ± 0.013	0.9955 ± 0.017	0.9977 ± 0.088
CatBoost	0.9969 ±0.080	0.9983 ±0.011	0.9951 ±0.013	0.9967 ±0.090
Random Forest	0.9972 ± 0.010	0.9987 ±0.014	0.9954 ±0.090	0.9967 ±0.012

TABLE2: PREDICTION RESULT ON TEST DATA

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Linear-SVM	94.68	94.83	94.52	94.64
Sigmoid-SVM	85.35	85.28	85.28	85.28
Poly-SVM	98.12	98.12	98.12	98.12
Default RBF-SVM	98.84	98.80	98.86	98.83
GA-RBF-SVM	99.46	99.45	99.46	99.45
LightGBM	99.75	99.75	99.75	99.75
XGBoost	99.80	99.81	99.80	99.80
CatBoost	99.75	99.75	99.74	99.75
Random Forest	99.81	99.82	99.81	99.81

As can be seen from Tables 1 and 2, The evaluation of SVMs spanned diverse kernels, including linear, polynomial, sigmoid, and RBF. The Optimized GA-RBF-SVM outshone all other kernels, achieving remarkable values of 99.46%, 99.45%, 99.46%, and 99.45% for Accuracy, Precision, Recall, and F1-score, respectively, during 10-fold cross-validation on

the test data. On the validation data, its performance remained consistently high, further attesting to the robustness of this approach.

The ensemble techniques, including LightGBM, XGBoost, CatBoost, and Random Forest, showcased their prowess in NID displaying high levels of accuracy and precision, demonstrating their capability to effectively capture intricate patterns within network data. Random Forest demonstrated impressive results, achieving highest accuracy levels above 99.8% on the test data.

V. CONCLUSION

In summary, our study has extensively explored NID using SVMs with various kernels, including linear, polynomial, sigmoid, and RBF, integrating GA-guided optimization of key parameters and recursive feature elimination feature selection. After a rigorous evaluation on a substantial dataset, our methodology showcases exceptional performance, with the optimized RBF-SVM achieving a test accuracy of 99.46%. Additionally, we compare our approach with notable ensemble learning techniques like LightGBM, XGBoost, CatBoost, and Random Forest, further expanding the scope of our analysis.

While the proposed approach demonstrates exceptional performance on the specified dataset, scalability remains a concern when deploying such techniques in real-world cybersecurity environments. The process of training SVM models with varying kernels and optimizing hyperparameters using genetic algorithms can be computationally intensive, particularly as the size of the dataset increases. As the volume of network traffic grows exponentially in many organizations, scalability becomes a critical consideration. Future research should focus on developing efficient algorithms and parallel processing techniques to enable the timely training of anomaly detection models on large-scale datasets, ensuring that computational resources are utilized effectively without compromising detection accuracy.

Another aspect of scalability pertains to the deployment of anomaly detection models in production environments. While achieving high accuracy during the training and validation stages is crucial, the practicality of deploying these models in real-time network monitoring systems is equally important. The computational resources required to process incoming network traffic, apply anomaly detection algorithms, and generate alerts in real-time must be carefully managed to ensure

scalability and responsiveness. Furthermore, as network environments evolve and grow more complex, the ability to adapt and scale anomaly detection systems to accommodate changing conditions becomes paramount. Future research should address the scalability challenges associated with deploying anomaly detection models in dynamic network environments, including the development of lightweight and efficient algorithms that can operate effectively in high-speed, high-volume network traffic scenarios.

REFERENCES

Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Transactions on Emerging Telecommunications Technologies, vol. 32, no. 1, p. e4150, 2021.

C. Kruegel and T. Toth, "Using decision trees to improve signature-based intrusion detection," in International workshop on recent advances in intrusion detection, 2003: Springer, pp. 173-191.

K. Hwang, M. Cai, Y. Chen, and M. Qin, "Hybrid intrusion detection with weighted signature generation over anomalous internet episodes," IEEE Transactions on dependable and secure computing, vol. 4, no. 1, pp. 41-55, 2007.

X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," Ieee Access, vol. 7, pp. 82512-82521, 2019.

J. Gu, L. Wang, H. Wang, and S. Wang, "A novel approach to intrusion detection using SVM ensemble with feature augmentation," Computers & Security, vol. 86, pp. 53-62, 2019.

M. Gauthama Raman et al., "An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm," Artificial Intelligence Review, vol. 53, pp. 3255-3286, 2020.

Y. Ibrahim, M. B. Mu'azu, E. A. Adedokun, and A.Yusuf, "A Framework for Fingerprint Liveness Detection Using Support Vector Machine Optimized by Genetic Algorithm," i-manager's Journal on Pattern Recognition, vol. 5, no. 2, p. 1, 2018.

Y. Ibrahim, E. Okafor, and B. Yahaya,

"Optimization of RBF-SVM hyperparameters using genetic algorithm for face recognit," Nigerian Journal of Technology, vol. 39, no. 4, pp. 1190-1197, 2020.

I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 14, no. 4, pp. 1502-1509, 2016.

O. Faker and E. Dogdu, "Intrusion detection using big data and deep learning techniques," in Proceedings of the 2019 ACM Southeast conference, 2019, pp. 86-93.

J. Lee, J. Kim, I. Kim, and K. Han, "Cyber threat detection based on artificial neural networks using event profiles," Ieee Access, vol. 7, pp. 165607-165626, 2019.

X. Zhang, J. Ran, and J. Mi, "An intrusion detection system based on convolutional neural network for imbalanced network traffic," in 2019 IEEE 7th international conference on computer science and network technology (ICCSNT), 2019: IEEE, pp. 456-460.

P. Negandhi, Y. Trivedi, and R. Mangrulkar, "Intrusion detection system using random forest on the NSL-KDD dataset," in Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, Volume 2, 2019: Springer, pp. 519-531.

M. O. Miah, S. S. Khan, S. Shatabda, and D. M. Farid, "Improving detection accuracy for imbalanced network intrusion classification using cluster-based under-sampling with random forests," in 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT), 2019: IEEE, pp. 1-5.

A. Nagaraja, U. Boregowda, K. Khatatneh, R. Vangipuram, R. Nuvvusetty, and V. S. Kiran, "Similarity based feature transformation for network anomaly detection," IEEE Access, vol. 8, pp. 39184-39196, 2020.

M. A. Ayub, W. A. Johnson, D. A. Talbert, and A. Siraj, "Model evasion attack on intrusion detection systems using adversarial machine learning," in 2020 54th annual conference on information sciences and systems (CISS), 2020: IEEE, pp. 1-6.

A. Pathak and S. Pathak, "Study on decision tree and KNN algorithm for intrusion detection system,"

International Journal of Engineering Research & Technology, vol. 9, no. 5, pp. 376-381, 2020.

L. Liu, P. Wang, J. Lin, and L. Liu, "Intrusion detection of imbalanced network traffic based on machine learning and deep learning," Ieee Access, vol. 9, pp. 7550-7563, 2020.

P. Kumar, A. A. Kumar, C. Sahayakingsly, and A. Udayakumar, "Analysis of intrusion detection in cyber attacks using DEEP learning neural networks," Peer-to-Peer Networking and Applications, vol. 14, pp. 2565-2584, 2021.

M. A. Almaiah et al., "Performance investigation of principal component analysis for intrusion detection system using different support vector machine kernels," Electronics, vol. 11, no. 21, p. 3571, 2022.