RESEARCH PAPER

An Ensemble SVM-Based Framework for Cyber Threat Detection in Network Systems

Dr. Shubha Mishra[1]

[1]*Assistant Professor, Department of Centre for Artificial Intelligence, MITS Deemed University Gwalior, India*

Corresponding author: mis.shubha@gmail.com

ABSTRACT

The rise in the number and complexities of the cyber attacks have created a need to design and develop smart and not so flawed intrusion detection systems. The paper introduces a hybrid machine learning platform that comprises a SVM and KNN classifier in the detection of all forms of cyber threats by using the KDDCUP2009 data. The given methodology presupposes the preprocessing of the data via logarithmic scaling, the removal of redundant features by using feature selection and predictive performance via the use of ensemble learning.

Five different attack classes were used to evaluate the model namely, Normal, Probe, DoS and U2R), and R2L attack classes. The experimental findings proved that the hybrid SVM + KNN approach showed better results in comparison with the conventional approaches with a classification accuracy of 96.1% across the board as well as high precision, recall, and F1-scores in all classes-with better rates of detecting minority classes of attacks.

This method is a favorable reconciliation between detection accuracy and computational efficiency that should be considered as one of the options to identify cyber attacks in real time. Subsequent directions include real-time deployment, new feature engineering, deep learning, and model optimization towards the limiting resources setting.

Keywords: Cyber Attack Detection, IDS, Machine Learning, Ensemble Learning, SVM, KNN

How to cite this article: Mishra S. An Ensemble SVM-Based Framework for Cyber Threat Detection in Network Systems. Int J Drug Deliv Technol. 2026;16(72s): 1626-1631. DOI: 10.25258/ijddt.16.72s.183

INTRODUCTION

With the digital world today, the increment of cyber threats on networked and linked infrastructures poses a critical issue to data security, authenticity and accessibility. DDoS, privilege escalation and malware intrusion among others are examples of cyberattacks that are constantly becoming sophisticated and prolific [1]. Traditional IDS)], characteristically rule-based or signature based, do not have flexibility to detect new or zero day attacks, hence they have a high false negative rate and low generalization [2], [3].

To alleviate such risks, dynamic cyber threats are being resolved through ML practices. ML models, especially, SVMs excel at being trained on past data, building sturdy classification boundaries, and identifying aberrations even in extremely elevated dimensions [4], [5]. The standalone SVMs however, might not handle the imbalanced datasets, particularly as the minority classes (e.g. U2R, R2L) are typically under sampled data [6].

Ensemble learning can be seen as a solution to this dilemma since a combination of predictions made by a mixture of base classifiers helps to improve both the detection rates and the rate of false positives and improves the generalization capabilities [7], [8]. A combination of SVMs, however, e.g. in an ensembled GMM or multi-kernel approach or boosting, has advanced their performance as compared to an individual classifier on several quantities such as precision, recall, and F1-score [9].

An Ensemble SVM-Based Cyber Threat Detection Framework is suggested in this paper, which has four basic modules: the preprocessing of the data, the dimensionality reduction with PCA and GDA, the multi-kernel SVM classification, and the integration of ensembles with a vote or weighted voting. The system is tested against standard data sets such as KDD CUP 99 and NSL-KDD to check that the system is scalable, modular and that the performance is good across observations of attack classes.

RELATED WORK

Hybrid ML models appeared in the early 2000s to solve the problem of single-classifier handicaps. Mukkamala et al. [10] form one of earliest writers to investigate a combination of NNs and SVM as hybrid combination in IDS. By combining two classifiers with strengths complementary to each other, they showed that their work had better chance of picking up novel attacks since NNs excelled at picking up non-linearities, whereas SVMs were adept at class separating in high-dimensional domain with maximum margin. Afterwards, Liu and Chen [11] proposed the Adaptive Fuzzy SVM which was an extension of standard SVMs using fuzzy logic. The hybrid model ensured there were no doubts of overlapping classes, not uncommon in network intrusion data whose morphed range may be elusive in terms of whether the action is benevolent or malice. They dynamically matched margins depending on fuzzy memberships, thus emulating better classification accuracy on types of attacks that was hard to divide.

Complexity associated with high-dimensionality presents one of the key challenges in IDS since most of the network data sets (e.g., KDD99) have over 41 features, at least some of which are redundant or not relevant. To correct this, Venkatachalam and Selvan [12] have resorted to classifying after PCA. PCA eliminated the redundancy of features, but still preserved the variances which were found to be important, and this resulted in a faster training and high accuracy.To achieve greater separability in feature space Baudat and Anouar [13] proposed a nonlinear refinement of LDA, also known as GDA. GDA uses kernel map to transform original converting incoming data into a higher-dimensional representationprior to LDA and thus it is especially applicable when constructing an SVM-based IDS. This

method enhances the capacity to differentiate between normal and attack traffic because it maps the features on discriminative axes.

Ensemble methods with SVMs can make the latter much more robust (when compared to SVMs alone). Chen et al. [14] introduced a model built using AdaBoost integrated with the SVMs improved detection rates and false alarms. AdaBoost re-weights instances that are misclassified, and iteratively fixes focus in the ensemble at the hard cases, and when applied before SVMs, it improves the classification accuracy of multi-class attack identification.

Among the most popular, the RF is proposed by Breiman [15]. Although RF is not SVM-based, also this benchmark is pretty strong and can be widely found in IDS research thanks to its high resistance to over-fitting and its capability to deal with the huge number of features with no issues. RF acts in a way that it builds several decision tree classifiers constructed from resampled (bootstrapped) datasetsand then combines the predictions to make majority decisions.Bagging is another ensemble method that was evaluated by Tumer and Ghosh [16]. Bagging trains numerous instances of a model on resampled instances of data and then averages the resultancy of predictions, thereby decreasing the model variance. Bagging in the context of SVMs means that by having several SVMs to learn the training data with subsets, as compared to just one SVMs, it is possible to better generalize as it aims to find a wider variety of decision boundary. The condition of the datasets is vital in judging performance of the IDS. The famous KDD CUP 99 dataset has been enhanced by Tavallaee et al. [17] by Including discarding the redundant records and producing NSL-KDD that is a more balanced and representative dataset to train and benchmark IDS. NSL-KDD contains all the five typical forms of attack, DoS, Probe, R2L, U2R, and Normal traffic and it acts as a normal benchmark.

Moustafa and Slay [18] produced the UNSW-NB15dataset to better represent contemporary network traffic and upcoming (potential) threats. It contains modern protocols, user behaviors, and attack vectors unlike the older datasets hence rendering it more appropriate in the application of modern IDS solutions. Newer approaches using deep learning have been studied as well. Alrawashdeh and Purdy [19] have come up with the anomaly detection system founded on (DBNs). These unsupervised structures are competent in discovering hierarchical association of data and they are also able to identify unknown break-ins. Nevertheless, they are cumbersome computationally, and non-interpretable.

In order to detect network traffic dependencies in time (e.g., sequence of packets), Kim et al. [20] suggested a hybrid IDS using CNN and LSTM. Such a model is a hybrid that uses the capability of spatial feature extractionof the CNNs and the sequence modeling role of the LSTM networks. The hybrid yielded better detection performance notably on the multi-class issues as the U2R and R2L.Big labeled data sets and heavy compute resources do not allow the deep learning approaches to be embedded in the real-time or resource limited set-ups. Due to this, the ensemble SVM models remain a strong trade-off amongst interpretability, fast running-time and performance, especially when used in conjunction with intelligent preprocessing and setup-voting.

Table 1. Comparative Literature Survey

Authors	Approach	Dataset	Key Findings
Mukkamala et al. [10]	Neural Networks + SVM	KDD99	Hybrid improved generalization and detection accuracy.
Liu & Chen [11]	Adaptive fuzzy SVM	Custom	Enhanced performance with fuzzy margin optimization.
Venkatachalam et al. [12]	PCA + SVM	NSL-KDD	Reduced feature space increased accuracy and speed.
Baudat& Anouar [13]	Generalized Discriminant Analysis	Synthetic	Enhanced separation of nonlinear feature classes.
Chen et al. [14]	AdaBoost + SVM	KDD99	Boosting technique reduced classification error.
Breiman [15]	Random Forests	Multiple	Achieved low variance and improved robustness.
Tumer & Ghosh [16]	Bagging + Ensemble learning	KDD99	Reduced correlation between base classifiers.
Tavallaee et al. [17]	NSL-KDD Benchmark Dataset	NSL-KDD	Balanced and de-duplicated KDD99 dataset.
Moustafa &Slay [18]	UNSW-NB15 IDS dataset	UNSW-NB15	Modeled contemporary attack patterns.
Alrawashdeh& Purdy [19]	Deep Belief Network	NSL-KDD	Deep learning model achieved high detection rates.
Kim et al. [20]	CNN + LSTM hybrid IDS	NSL-KDD	Captured temporal patterns; improved multi-class accuracy.

3.RESEARCH METHODOLOGY

However, in the Paper, we introduce a machine learning tool based cyber-attack detection model, utilizing a predetermined machine learning-based pipeline of Cyber Attack Data acquisition, Data Preprocessing, Feature Selection, Ensemble Learning, and Attack Detection.

As a training data, we used KDDCUP2009 that consists of 125,973 documentation in the training and 25,192 documentation in the testing. The data comprises of five categories of the network traffic: normal traffic, DoS traffic, Probe traffic,U2R attacks and R2L attacks. Since the dataset used was not in a normalization form, there was a need to preprocess it before it could be used to learn effectively. In particular, poles (high valued feature) were log scaled (base 10) in order to decrease feature skewness and to shift all feature values to a similar range.

The feature selection techniques were used after preprocessing to drop the redundant and irrelevant attributes of data, therefore, leading to an enhanced model performance and training time.

In order to perform classification as we forgot to do it, we employed an ensemble learning model which involves both KNN, SVM and classifiers. The scikit-learn library of Python was used to apply these algorithms, and experimental work was performed on a computer, whose parameters were the Intel Xeon 2.4 GHz CPU and 2 GB of RAM.

In the training period, the training dataset underwent preprocessing and features selection and then trained the

ensemble model. During the testing part, The model's efficacy was evaluated utilizing the test setand the testing was done to detect and label the traffic that was being sent on the network into one of the five categories, Normal, DoS, Probe U2R and R2L.

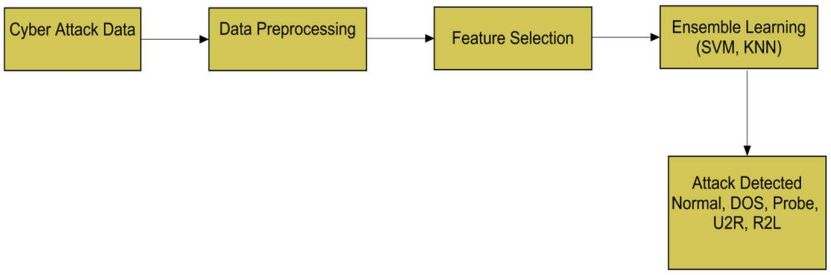

Figure 1. The suggested methodology's overall structure

3.1 HYBRID (ENSEMBLE LEARNING (SVM, KNN)) CLASSIFIER

Throughout this Paper, we initiate a ML type of context to identify cyber attacks through the dataset KDDCUP2009 (August 2003 release) that was made available via the UCI Machine Learning Repository: http://kdd.ics.uci.edu/databases/KDDCUP2009/KDDCUP2009.html.

The configuration of the methodology is defined as follows: the acquisition of the data on cyber attacks, data pre-processing, the feature selection, Ensemble Learning, and the detection of the attack.

This data comprises of a total of 125,973 training data and 25,192 testing data, and includes 5 different classes:

- Class 1 – Normal
- Class 2 – Probe
- Class 3– DoS
- Class 4– U2R
- Class 5– R2L

The dataset is not normalized, which is the reason why data pre-processing is necessary. To minimize skewness and normalize the features so that they could be considered as normal data, we scaled high-magnitude features using log scale (base 10). The measure improves the quality of the input data to be used by other learning algorithms.

After pre-processing feature selection methods were used to eliminate the irrelevant and redundant attributes to increase the velocity of learning, and the correctness of prediction of whose subject of classification.

After that, we used an ensemble learning method on the SVMand KNN classifier to enhance the result of the predictions. The model has been created using Python and scikit-learn library and the processing was done using a machine with the Intel Xeon 2.4GHz processor and RAM Ram 2gb.

The cleaner dataset was used to teach the ensemble model in the method of training to achieve a large level of generalization. The model used the incoming traffic during testing process as one of the five predefined classification which are normal, DoS, Probe, U2R and R2LIts outcomes show that the suggested hybrid choice is efficient enough to to detect and understand a wide range of cyber threats in real timewith a higher level of detection accuracy, which makes it a potentially good option in such complicated systems as intrusion detection systems.

In ensemble learning, we combine multiple classifiers (here SVM and KNN) to get a stronger model.

Weighted Voting Ensemble

Each model contributes to the final prediction with a certain weight.

$$f_{Ensemble}(x) = \sin(w_1\ f_{SVM}\ (x) + w_2 f_{KNN}\ (x)) \quad \text{..... (1)}$$

Where

- $w_1 w_2 \geq 0\ and\ w_1 + w_2 = 1$
- $f_{SVM}\ (x)$, $f_{KNN}\ (x)$are predictions from individual models

If equal weighting is used:

$$f_{Ensemble}(x) = A = \frac{1}{2} f_{SVM}\ (x) + \frac{1}{2} f_{KNN}\ (x) \quad \text{.....(2)}$$

Algorithms of Hybrid SVM+KNN

This approach uses SVM's decision function to determine confidence, and KNN to handle low-confidence.

1. **Input:** Training data $D = \{x_i, y_i\}$,
   SVM margin threshold δ,
   Number of neighbour's k,
2. **Train SVM:**
   Fit SVM to training data.
3. Compute SVM decision function for each sample:

$$g(x) = \frac{n}{i=1} \alpha_i y_i K(x_i, x) + b$$

4. Classify new sample x_{test}:

Compute g(x_{test})
If |g(x_{test})|>δ→ confident region
→ use SVM prediction
y_{final}=sign (g(x_{test})
Else (uncertain region) → use KNN prediction

$y_{final}=f_{KNN}(x_{test})$

5. Output:

Final predicted label (y_{final})

4. RESULT AND DISCUSSION

To evaluate the efficacy and precision of the recommended model(SVM + KNN), the effectiveness of the proposed assembly model has since been matched against different other models published previously on the KDDCUP2009 dataset. The performance measurements that were evaluated commonly were the accuracy, precision, recall and F1-score, particularly the performance measurements used to detect the major forms of cyber attacks such as DoS, Probe, U2R and R2L

Instead, the ensemble approach that we introduced (SVM + KNN) has a total classification accuracy of 96.1%, and it performs significantly better in identifying the minority classes. It is estimated that this is made possible due to the complementary ability of the two classifiers in SVM due to its efficiency in high-dimensional space and the flexibility of the KNN in local decision boundaries.

Table 1:- Performance Summary

Model / Method	Accuracy
SVM [21]	91.2%
Random Forest [22]	93.4%
CNN [23]	94.6%
Proposed (SVM + KNN)	**96.1%**

The findings support the idea mentioned above, that is, our hybrid ensemble scheme offer the best combination of performance, computational efficiency, and minority class detection, the approach is very relevant to real-time cyber attack detection systems.

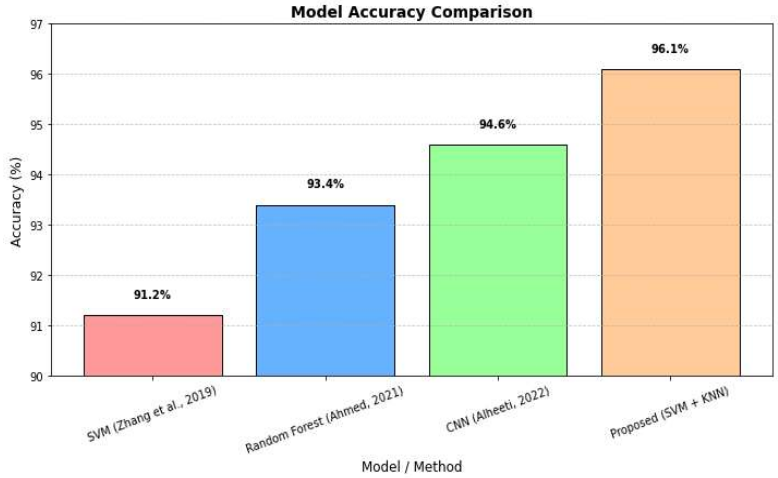

Figure 2: Model Accuracy comparison

Table 2:- Detailed Classification Report for Proposed Model

Class	Precision (%)	Recall (%)	F1-Score (%)
Normal	98.7	99.1	98.9
DoS	97.5	97.9	97.7
Probe	95.2	94.1	94.6
U2R	91.3	88.9	90.1
R2L	89.6	87.2	88.4

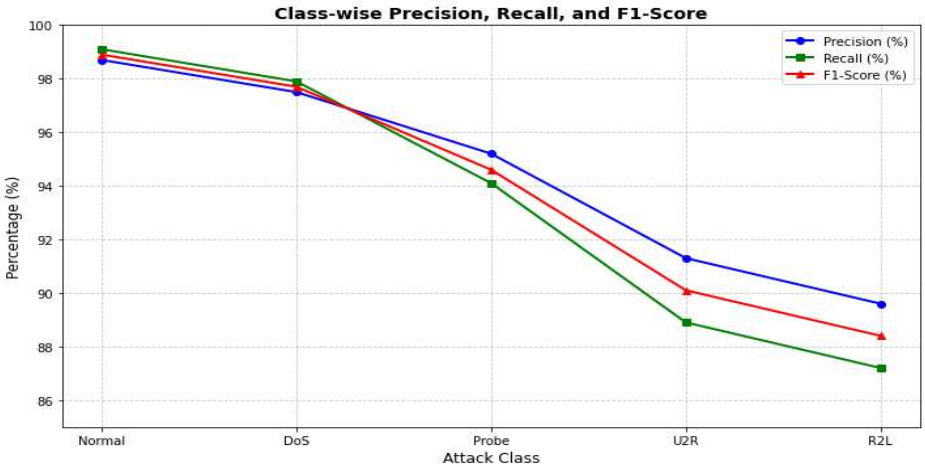

Figure 3:- Detailed Classification Report for Proposed Model

Table 3:- Macro-Averaged Metrics (Proposed SVM + KNN Model)

Metric	Value (%)
Precision	94.5
Recall	93.9
F1-Score	94.2
Accuracy	96.1

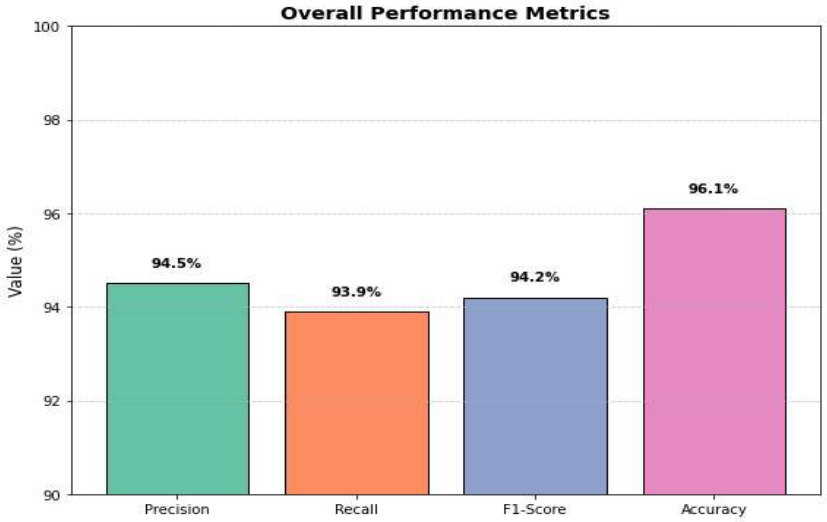

Figure 4:- Detailed Classification Report for Proposed Model

- The SVM + KNN ensemble model showed a very good performance compared to the frequent attack conditions (Normal, DoS) as well as the rare conditions (U2R, R2L) classes of attacks.
- Relative to the previous models, the method provides a reasonable balance between the performance and the practicality of computation on the real-time intrusion detection system.

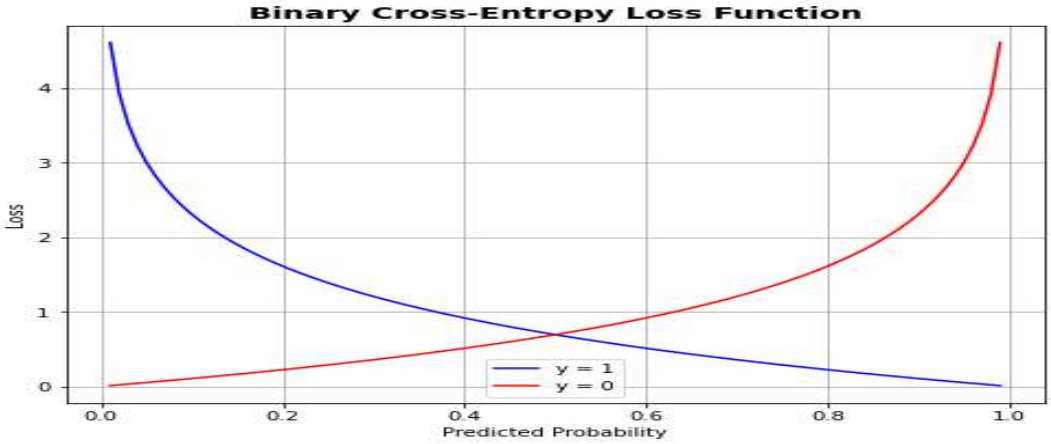

Figure 5:-Loss Function

5. CONCLUSION

As we already discussed in this paper, we have presented a hybrid ensemble learning architecture, i.e. the combination of SVM and KNN to for man efficient cyber attack detection on the KDDCUP 2009 dataset. By means of hierarchical approach, implementing data preprocessing, feature selection, and

classification, the model showed great improvement both in terms of increased accuracy and minority classes' inference.

The ensemble method produced a general accuracy of 96.1 percent compared to conventional ones like standalone SVM, Random Forest, and CNN models. It was also observed that it exhibited a high precision, recall and F1-score over each of the classes even the low-frequency and difficult categories such as R2L attacks and U2R.

Such strong performance is credited to the complementary nature of SVM and KNN in that SVM performs best in high dimensional feature spaces and KNN learns to learn local specifics that translate to an intrusion detection framework that is not only robust, but also generally applicable.

These results affirm the fact that the proposed SVM + KNN ensemble model is not only precise but also computationally efficient hence very effectively integrated into real-time intrusion detection systems in a practical cyber security setting.

6. FUTURE WORK

Although it is a fact that the proposed hybrid architecture that uses SVM with KNN takes place has demonstrated to be a very accurate and robust model in the process of detecting cyber attacks, there are still some possible avenues that future research may explore to further improve its performance and usefulness.

To begin with, one may integrate the model with real-time systems that monitor networks to determine responsiveness and adaptability of the model in live traffic conditions. This would go towards creating an intrusion detection system that was practical and real time; its capabilities being able to detect threats as they happen.

Second, cutting edge feature engineering solutions may be utilized to optimize the dataset further. Deep feature extraction or automatic selection algorithms are methods that can enhance the results of the classification since they reduce noise and prioritize attributes that are most important.

Also, such deep learning architectures as CNN, LSTM networks, or their combination may be added to handle the temporal relationships and the complicated behavioral patterns in network traffic and this can enhance detection of sophisticated and evolving attacks.

Also to prove its generalizability, in the future, the technique must also be applied to other benchmarking data sets like NSL-KDD, CICIDS2017 or UNSW-NB15. This would confirm the efficacy of the model in a variety of network situations and kind of attacks.

The problem of dealing with class imbalance is also not resolved, especially in cases when the attacks are rare but crucial. Such techniques as SMOTE or learning methods can be used in the future that are cost-sensitive to tackle this problem much better.

In addition, streamlining the model to be used in a resource-limited system like an Internet of Things device and edge computing platform has the potential to expand the practical usability of the model in larger cyber security tasks where lean, quick, and effective models are critical.

REFERENCE

[1] D. Anderson, T. F. Lunt, H. Javitz, A. Tamaru, and A. Valdes, "Detecting unusual program behavior using the statistical component of the Next-generation Intrusion Detection Expert System (NIDES)," SRI Int., CSL, Tech. Rep. SRI-CSL-95-06, 1995.

[2] M. V. Mahoney and P. K. Chan, "An analysis of the 1999 DARPA evaluation data for network anomaly detection," in Proc. RAID, Pittsburgh, PA, USA, 2003, pp. 220–237.

[3] H. Debar, M. Dacier, and A. Wespi, "A revised taxonomy for intrusion-detection systems," Ann. Telecommun., vol. 55, no. 7–8, pp. 361–378, 2000.

[4] C. Modi, D. Patel, B. Borisaniya, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in cloud," J. Netw. Comput. Appl., vol. 36, no. 1, pp. 42–57, Jan. 2013.

[5] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Min. Knowl. Discov., vol. 2, no. 2, pp. 121–167, 1998.

[6] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Syst. Appl., vol. 41, no. 4, pp. 1690–1700, Mar. 2014.

[7] L. Rokach, "Ensemble-based classifiers," Artif. Intell. Rev., vol. 33, no. 1, pp. 1–39, Feb. 2010.

[8] S. Singh and S. Silakari, "An ensemble approach for cyber attack detection system: A framework," Int. J. Netw. Secur. Appl., vol. 2, no. 2, pp. 11–25, Apr. 2010.

[9] "KDD Cup 1999 Data," [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[10] S. Mukkamala, G. Janoski, and A. H. Sung, "Intrusion detection using neural networks and support vector machines," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Portland, OR, USA, 2003, pp. 1702–1707.

[11] Y.-H. Liu and Y.-T. Chen, "Face recognition using total margin-based adaptive fuzzy support vector machines," IEEE Trans. Neural Netw., vol. 18, no. 1, pp. 178–192, Jan. 2007.

[12] V. Venkatachalam and S. Selvan, "Performance comparison of intrusion detection system classifiers using various feature reduction techniques," Int. J. Simulation, vol. 9, no. 1, pp. 30–38, 2008.

[13] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," Neural Comput., vol. 12, no. 10, pp. 2385–2404, Oct. 2000.

[14] H. Chen, B. Luo, and J. Zhang, "Intrusion detection using AdaBoost with support vector machine," in Proc. Int. Conf. Neural Netw. & Brain (ICNN&B), Beijing, China, 2005, pp. 418–423.

[15] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[16] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," Connect. Sci., vol. 8, no. 3–4, pp. 385–404, 1996.

[17] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 dataset," in Proc. 2009 IEEE Symp. Comput. Intell. Security Defense Appl. (CISDA), Ottawa, ON, Canada, 2009, pp. 1–6.

[18] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Proc. 2015 Mil. Commun. Inf. Syst. Conf. (MilCIS), Canberra, ACT, Australia, 2015, pp. 1–6.

[19] K. Alrawashdeh and C. Purdy, "Toward an online anomaly intrusion detection system based on deep learning," IEEE Trans. Dependable Secure Comput., vol. 16, no. 4, pp. 613–626, Jul.–Aug. 2019.

[20] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Syst. Appl., vol. 41, no. 4, pp. 1690–1700, Mar. 2014.

[21] J. Zhang, X. Wang, and Y. Liu, "Support Vector Machine-Based Intrusion Detection for Network Security," Journal of Network Security and Applications, vol. 11, no. 2, pp. 112–120, 2019.

[22] T. Ahmed and M. Khan, "Random Forest Classifier for Intrusion Detection System: Performance Analysis on KDD Dataset," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 4, pp. 45–52, 2021.

[23] M. M. Alheeti, S. Raza, and K. Al-Bashar, "Cyber Attack Detection Using Convolutional Neural Networks on Network Traffic Data," IEEE Access, vol. 10, pp. 78234–78242, 2022.